Synthetic Control Methods: The Good, the Bad, and the Ugly

Timo Kuosmanen[a], Xun Zhou[a,b], Juha Eskelinen[a], Pekka Malo[a]

[a] *Aalto University School of Business, Finland*
[b] *University of York, United Kingdom*

Abstract

The synthetic control method (SCM) is a major innovation in the estimation of causal treatment effects of policy interventions and programs in a variety of settings. However, the commonly used algorithms are ill-equipped for solving the SCM problem, which turns out to be a NP-hard bilevel optimization problem. In this paper we show how the original SCM problem can be reliably solved using iterative algorithms based on the Tykhonov descent approximations. Unfortunately, the true optimal solution to the original SCM is typically a corner solution where all weight is assigned to a single predictor, contradicting the intended purpose of predictors. To address this flaw, we propose to determine the predictor weights and donor weights separately. We show how the donor weights can be optimized when the predictor weights are given, and consider alternative data-driven approaches to determine the predictor weights. Re-examination of the two original empirical applications to Basque terrorism and California's tobacco control program demonstrates the failure of the existing SCM algorithms and illustrates our proposed remedies.

Keywords: Causal effects; Comparative case studies; Policy impact assessment; Treatment effect models

JEL Codes: C54; C61; C71

1. Introduction

The synthetic control method (SCM) is an appealing tool for estimating causal treatment effects of policy interventions and programs in a variety of settings. Abadie & Gardeazabal (2003) originally introduced SCM to examine the economic impacts of terrorism in the Basque Country. Abadie et al. (2010) further examined the statistical foundations of the method in their study of California's tobacco control program. Subsequently, SCM has been used in a large number of influential applications, including Acemoglu et al. (2016) (political connections), Cavallo et al. (2013) (natural disasters), Gobillon & Magnac (2016) (enterprise zones), Kleven et al. (2013) (taxation of athletes), and Abadie et al. (2015) (German reunification). Recently, Cole et al. (2020) apply SCM to study the impact of the Covid-19 lockdown on air pollution and health in Wuhan, China. Athey & Imbens (2017) refer to SCM as "arguably the most important innovation in the policy evaluation literature in the last 15 years."

Email addresses: timo.kuosmanen@aalto.fi (Timo Kuosmanen), xun.zhou@york.ac.uk (Xun Zhou), juha.p.eskelinen@aalto.fi (Juha Eskelinen), pekka.malo@aalto.fi (Pekka Malo)

Technically, SCM estimates the causal treatment effect by constructing a counterfactual of the treated unit using a convex combination of similar units not exposed to the treatment. The convex combination requires non-negative weights that sum to one to avoid extrapolation. In SCM, the weights are determined to ensure that the treated unit and the synthetic control resemble each other as closely as possible prior to the treatment, both with respect to the outcome of interest and some observed economic predictors. Since there are typically multiple predictors, the predictors are also weighted using another set of non-negative weights. Abadie & Gardeazabal (2003) and Abadie et al. (2010) discuss several alternative approaches to specify the predictor weights, including the use of subjective weights. In practice, a majority of published SCM applications resort to the data-driven procedure where the weights of predictors and control units are jointly optimized to minimize the mean squared prediction error of the synthetic control over the pre-treatment period, applying the *Synth* package described in Abadie et al. (2011).

Interestingly, a number of recent studies report that the synthetic control weights produced by *Synth* are numerically unstable and inaccurate (e.g., Klößner et al., 2018; Becker & Klößner, 2017, 2018; Becker et al., 2018; Albalate et al., 2021; Kuosmanen et al., 2021). A related but even more serious concern is that the predictors often turn out to have little impact on the synthetic control, as noted by several authors (e.g., Doudchenko & Imbens, 2017; Ben-Michael et al., 2021; Kaul et al., 2021). This is a serious problem because the statistical properties of the SCM estimator critically depend on the ability of the synthetic control to reproduce the observed and unobserved characteristics of the treated unit (Abadie et al., 2010; Abadie, 2021). If most predictors are assigned negligibly small weights, then the ability of SCM to reproduce the observed characteristics and the latent factors is seriously compromised.

Recently two parallel lines of research by Malo et al. (2020) and Albalate et al. (2021) have shed further light on the computational problems of SCM. These two groups of authors independently developed the first explicit mathematical formulations of the standard SCM problem where the predictor weights and the donor weights are jointly optimized, and find that the original SCM problem is in fact a NP-hard bilevel optimization problem. This finding not only explains the numerical instability of *Synth*, but also suggests that *Synth* and other SCM algorithms known in the literature are clearly ill-equipped for solving the original SCM problem. Consequently, several thousands of SCM applications published thus far are based on different weights than claimed by the authors, which may influence the qualitative conclusions.

The main contributions of the present paper[1] are threefold:

1) We develop a new iterative algorithm for solving the original SCM problem, based on the Tykhonov regularization, and formally prove that the proposed algorithm is guaranteed to converge to the optimal solution. Revisiting the classic SCM applications to the Basque terrorism (Abadie & Gardeazabal, 2003)

[1] This article consolidates and updates the main contributions of the two non-reviewed working papers Malo et al. (2020) and Kuosmanen et al. (2021) by the authors.

and the California tobacco control program (Abadie et al., 2010), we demonstrate that the computational algorithms currently in use fail to converge to the true optimum.

2) We point out that the true optimal solution of the original SCM problem is typically a corner solution where all weight is assigned to a single predictor. We show that this is also the case in the classic SCM applications. In our interpretation, the numerical instability of SCM is a symptom, but the tendency towards corner solutions is a more serious flaw of the original SCM method, caused by the joint optimization of donor weights and predictor weights. Developing better algorithms to solve the NP-hard bilevel optimization problem does not address the root cause of the problem.

3) As a remedy, we propose to determine the predictor weights and donor weights in two separate stages, in agreement with Albalate et al. (2021) who similarly propose to decouple the two nested optimization problems. To this end, we develop a simple two-step algorithm to optimize the donor weights when the predictor weights are given *a priori*. This proves a non-trivial task, in fact, we find that the *Synth* algorithm fails to produce optimal donor weights even when the predictor weights are given by the user. We also briefly explore alternative data-driven approaches to determine the predictor weights, complementing the SHAP approach proposed by Albalate et al. (2021). These include the use of regression-based weights, which is also the default option of the Stata implementation of *Synth* (Abadie et al., 2011), and has been used in some empirical studies (e.g., Bohn et al., 2014). Another possibility is to apply equal weights to standardized predictors, analogous to Bloom & Van Reenen (2007) approach to aggregate management survey indicators. We compare two variants of the regression-based approach and the uniforms weights for standardized predictors in case of the two classic SCM applications to the Basque terrorism and California tobacco control program.

The rest of the paper is organized as follows. Section 2 introduces the original SCM method and formulates the data-driven approach to compute the predictor and donor weights as a bilevel optimization problem. Section 3 develops an iterative algorithm that is guaranteed to converge to the optimal solution. Section 4 demonstrates empirically that the classic SCM applications to the Basque terrorism and the California tobacco control program both have corner solutions, and that the existing *Synth* and MSCMT algorithms fail to converge to the optimum. Section 5 explores alternative data-driven approaches to determine the predictor weights, proposes a simple two-step approach to optimize the donor weights when the predictor weights are given *a priori*, and revisits the two classic SCM applications to illustrate the proposed approaches. Section 6 presents our concluding remarks and discusses avenues for future research. Proofs of Theorems are presented in Appendix A of Online Supplement and the implementation of the descent algorithm is discussed in Appendix B. To allow readers reproduce our iterative algorithm to check for the feasibility of the unconstrained optimum and the possibility of corner solution and to reproduce our empirical results, the R code is provided in Appendix C. Appendix D of Online Supplement describes the imputation of missing values of the Basque terrorism application for applying an

alternative data-driven approach to determine the predictor weights.

2. Synthetic control method

2.1. Preliminaries

Following the usual notation (e.g., Abadie, 2021), suppose we observe units $j = 1, \ldots, J+1$, where the first unit is exposed to the intervention and the J remaining units are control units that can contribute to the synthetic control. The set of J control units is referred to as the donor pool. For the sake of clarity, we denote the number of time periods prior to treatment as T^{pre} and the number of time periods after the treatments as T^{post}. The outcome of interest is denoted by Y: column vectors Y_1^{pre} and Y_1^{post} with T^{pre} and T^{post} rows, respectively, refer to the time series of the pre-treatment and post-treatment outcomes of the treated unit. Similarly, matrices Y_0^{pre} and Y_0^{post} with J columns refer to the pre-treatment and post-treatment outcomes of the control group, respectively.

Ideally, the impact of treatment could be measured as

$$\alpha = Y_1^{\text{post}} - Y_1^{\text{post,N}}, \tag{1}$$

where $Y_1^{\text{post,N}}$ refers to the counterfactual outcome that would occur if the unit was not exposed to the treatment. If one could observe the outcomes $Y_1^{\text{post,N}}$ in an alternative state of nature, where the unit was not exposed to the treatment, then one could simply calculate the elements of vector α. The main challenge in the estimation of the treatment effect is that only Y_1^{post} is observable, whereas the counterfactual $Y_1^{\text{post,N}}$ is not.

The goal of SCM is to construct a synthetic control group to estimate the counterfactual $Y_1^{\text{post,N}}$. The key idea of the SCM is to use the convex combination of the observed outcomes of the control units Y_0^{post} as an estimator of $Y_1^{\text{post,N}}$. Formally, the SCM estimator is defined as

$$\hat{\alpha} = Y_1^{\text{post}} - Y_0^{\text{post}} W, \tag{2}$$

where the J elements of column vector W are nonnegative and sum to one. The weights W characterize the synthetic control, that is, a counterfactual path of outcomes for the treated unit in the absence of treatment.

To set the weights W, the simplest approach considered by Abadie & Gardeazabal (2003) is to track the observed path of pre-treatment outcomes as closely as possible to minimize the mean squared prediction error (MSPE). That is, one could apply the weights W that solve the following constrained least squares problem

$$\min_{W \in \mathcal{W}} L(W) = \frac{1}{T^{\text{pre}}} \left\| Y_1^{\text{pre}} - Y_0^{\text{pre}} W \right\|^2, \tag{3}$$

where

$$\mathcal{W} = \left\{ W \in \mathbb{R}^J : \sum_{j=2}^{J+1} W_j = 1, \; W_j \geq 0, \; j = 2, \ldots, J+1 \right\} \tag{4}$$

is the set of admissible weights for control units and $\|\cdot\|$ denotes the usual Euclidean norm. The constraints on the weights W ensure that the synthetic control is a convex combination of the control units in the pool of donors. The fact that SCM does not involve extrapolation is considered as one of its greatest advantages over regression analysis (e.g., Abadie, 2021). Note that if we relax the constraints on weights W, then the unconstrained minimization problem reduces to the classic OLS problem without the intercept term. In that case, one could simply regress the time series Y_1^{pre} on the parallel outcomes of the J donors in the control group, and set the weights W equal to the corresponding OLS coefficients. While the OLS problem has the well-known closed form solution that satisfies the first-order conditions, however, the optimal solution to the constrained least squares problem stated above is typically a corner solution where at least some of the constraints on weights W are binding. The constrained least squares problem can be efficiently solved by quadratic programming (QP) algorithms such as CPLEX, which are guaranteed to converge to the global optimum.

In addition to the outcome of interest, an integral part of SCM is to utilize additional information observed during the pre-treatment period. Suppose we observe K variables referred to as predictors (also known as growth factors, characteristics, or covariates), which are observed prior to the treatment or are unaffected by the treatment, which can influence the evolution of Y. These predictors are denoted by a $(K \times 1)$ vector X_1 and a $(K \times J)$ matrix X_0, respectively.[2] Abadie et al. (2010) prove unbiasedness and consistency of the SCM under the condition that the synthetic control yields perfect fit to the predictors, that is, $X_1 = X_0 W$. Abadie (2021) notes that "*In practice, the condition $X_1 = X_0 W$ is replaced by the approximate version $X_1 \approx X_0 W$. It is important to notice, however, that for any particular data-set there are not ex-ante guarantees on the size of the difference $X_1 - X_0 W$. When this difference is large, Abadie et al. (2010) recommend against the use of synthetic controls because of the potential for substantial biases.*"

Since the K predictors included in X do not necessarily have the same effect on the outcomes Y, Abadie & Gardeazabal (2003) introduce a $(K \times K)$ diagonal matrix V where the diagonal elements are weights of the predictors that reflect the relative importance of the predictors. The diagonal elements of V must be non-negative,[3] and are usually normalized to sum to unity.[4] That is

[2] A common practice in SCM is to include some convex combinations of the pre-treatment outcomes also as predictors (see Abadie et al., 2010, 2015, for discussion). However, Kaul et al. (2021) demonstrate that including all pre-treatment outcomes as predictors is not a good idea because the predictors become completely redundant in that case.

[3] While Abadie et al. (2010) assume that the diagonal elements must be positive, a positive real number can be arbitrarily close to zero, and therefore, the distinction between positive and non-negative model variables has no real meaning in optimization unless one imposes some explicit lower bound, e.g., $V_{kk} \geq 0.01$. Becker & Klößner (2018) set a lower bound $V_{kk} \geq 0.00000001$, which is so low that it has no practical meaning.

[4] Of course, other normalizations are possible, but we here restrict attention to the most standard normalization that allows one to interpret the diagonal elements of V as shared weights that sum to one.

$$V \in \left\{ \text{diag}(V) : \; V \in \mathbb{R}^{K \times K}, \; \sum_{k=1}^{K} V_{kk} = 1, \; V_{kk} \geq 0 \right\} =: \mathcal{V}, \tag{5}$$

which is a sub-set of all non-negative diagonal matrices.

Both Abadie & Gardeazabal (2003) and Abadie et al. (2010) suggest that weights V could be subjectively determined. However, most known applications of SCM resort to the data-driven procedure suggested by the authors. Unfortunately, these seminal papers do not explicitly state the required optimization problem. A closer examination of the SCM problem in the next section reveals that the SCM problem is far from trivial from the computational point of view.

2.2. Bilevel optimization problem

Since Abadie & Gardeazabal (2003) and Abadie et al. (2010) only state the SCM problem implicitly, to gain a better understanding of the data-driven approach, the first step is to formulate the SCM problem explicitly. Recently Malo et al. (2020) and Albalate et al. (2021) show that the optimal weights $V^\star$, $W^\star$ must be obtained as an optimal solution to the following optimistic bilevel optimization problem

$$\min_{V,\, W} L_V(V, W) = \frac{1}{T^{\text{pre}}} \|Y_1^{\text{pre}} - Y_0^{\text{pre}} W(V)\|^2 \tag{6}$$

$$\begin{aligned} \text{s.t.} \qquad & W(V) \in \Psi(V) := \operatorname*{argmin}_{W \in \mathcal{W}} L_W(V, W) = \|X_1 - X_0 W\|_V^2, \\ & V \in \mathcal{V}, \end{aligned} \tag{7}$$

where $\|\cdot\|_V$ is a semi-norm parametrized by V and $\Psi : \mathcal{V} \rightrightarrows \mathcal{W}$ denotes the solution set mapping from upper-level decisions to the set of global optimal solutions of the lower-level problem. For any $(K \times 1)$ real vector Z, we define $\|Z\|_V = (Z^\top V Z)^{1/2}$. This becomes a proper norm only when V is positive-definite. If we denote the diagonal elements of V by $v_1, \ldots, v_K$, we can write the lower level objective as

$$L_W(V, W) = \sum_{k=1}^{K} v_k \left(X_{k,1} - \sum_{j=2}^{J+1} X_{k,j} W_j \right)^2,$$

which allows the lower-level to be interpreted as an importance-weighted least squares with weight constraints. As pointed out by Klößner & Pfeifer (2015), this original setup can be easily extended to allow treatment of predictor data as time series, while maintaining the original structure of the optimization problem.

The explicit formulation of the optimization problem reveals several points worth noting. First, the SCM problem is a bilevel optimization problem, which is far from trivial from the computational point of view. The minimization problem (7) refers to the lower-level problem, and problem (6) is called the upper-level problem; the SCM literature commonly uses the terms inner and outer problems, but the meaning is the same. The problem is solvable, when it is interpreted as an optimistic bilevel problem, but the global optimum is not necessarily unique.

Proposition 1. *The synthetic control problem defined by (6)–(7) has a global optimistic solution* $(\bar{V}, \bar{W}) \in \mathcal{V} \times \mathcal{W}$.

Unfortunately, the bilevel optimization problems are generally NP-hard (Hansen et al., 1992; Vicente et al., 1994). In particular, the hierarchical optimization structure can introduce difficulties such as non-convexity and disconnectedness (e.g., Sinha et al., 2013), which are also problematic in the present setting, as will be demonstrated in the next section.

Second, the explicit statement of the optimization problem makes it more evident that the optimal solution will typically be a corner solution where at least some of the first-order conditions do not hold. This causes a serious problem for the usual derivative-based optimization algorithms. This observation can help to explain at least partly the numerical instability of the SCM results, observed by Becker & Klößner (2017), Klößner et al. (2018), and Albalate et al. (2021), among others. The general-purpose algorithms are simply ill-equipped for the task at hand.

3. Iterative algorithm

The purpose of this section is to discuss a general algorithm for solving the original SCM problem (4)–(5) where the predictor weights V are jointly optimized with the donor weights W. Since the general algorithm proves computationally demanding, we start by checking whether the unconstrained SCM problem (3) is a feasible solution, and also check the possibility of corner solutions. In case the optimal solution is not found through these feasibility checks, we suggest continuing search for an optimal solution using a descent-algorithm based on Tykhonov regularization technique or Karush-Kuhn-Tucker (KKT) approximations.

To highlight the importance of coordination between the upper-level and lower-level problems, we can rephrase the lower-level problem (7) as

$$\min_{W \in \mathcal{W}} L_W^{\varepsilon}(V, W) = \|X_1 - X_0 W\|_{V^\star}^2 + \varepsilon \|Y_1^{\text{pre}} - Y_0^{\text{pre}} W(V)\|^2 \tag{8}$$

where $\varepsilon > 0$ denotes an infinitesimally small non-Archimedean scalar.[5] Introducing the upper-level objective as a part of the lower-level QP problem in (8) makes a subtle but important difference compared to problem (7): the primary objective of both (7) and (8) is to minimize the loss function L_W with respect to predictors X. However, if there are alternate optima $W^\star$ that minimize the loss function L_W, problem (8) will choose the best solution for the upper-level problem.

Proposition 2. *For a given set of weights* $V^\star$, *let* $W_\varepsilon(V^\star)$ *denote the unique optimal solution to problem (8) for any* $\varepsilon > 0$. *Then, we have that*

$$\lim_{\varepsilon \to 0+} W_\varepsilon(V^\star) \in \operatorname*{argmin}_W \{L_V(V^\star, W) : \ W \in \Psi(V^\star)\}.$$

[5] The use of non-Archimedean ε was introduced by Charnes (1952) to avoid degeneracy in linear programming.

The proof of the proposition is simple and can be omitted. Having ensured that constraint (5) holds, it is important to note that the optimal weights W that minimize $\|X_1 - X_0 W\|^2_{V^\star}$ need not be unique. This is particularly relevant when there exist W that satisfy $\|X_1 - X_0 W\|^2_{V^\star} = 0$. In such cases, the non-Archimedean ε plays an important role by allowing us to select among the alternate optima for (5) the optimal weights W to minimize the upper-level objective (6).

Proposition 2 provides a useful result for SCM applications where the weights V are given. Recall that weights V might be subjectively determined, as Abadie & Gardeazabal (2003) and Abadie et al. (2010) suggest. Proposition 2 also demonstrates the critical importance of introducing an explicit link between the lower-level problem and the upper-level problem. In general, there can be many alternate optima where the loss function goes to zero, $L_W = 0$. Without coordination, there is no guarantee that the SCM algorithm would converge to the optimum. The lack of an explicit link between the upper-level and the lower-level problem is the most fundamental reason why the *Synth* algorithm fails to reach the optimum.

3.1. Checking the feasibility of an unconstrained solution

Consider first the situation where no predictors are used (i.e., $K = 0$). In this case, the bilevel optimization problem (6)–(7) reduces to the constrained regression problem (3). Problem (3) has a quadratic objective function and a set of linear constraints, which guarantees existence of a unique global optimum, when the usual assumptions of regression analysis hold (i.e., no rank deficiency). Such quadratic programming problems are considered straightforward from the computational point of view. While general-purpose derivative based algorithms may struggle with the constraints, the simplex-based algorithms (e.g. the CPLEX solver) will converge to the global optimum.

Let $L(W^{\star\star}) = \min_{W \in \mathcal{W}} L(W)$ denote the optimal solution to the problem (3), which is unique when no rank deficiency is present. As Kaul et al. (2021) correctly note, this solution is the lower bound for the optimal solution to the problem (6):

$$L_V(V, W) \geq L(W^{\star\star}) \text{ for all } V \in \mathcal{V},\ W \in \mathcal{W}. \tag{9}$$

Intuitively, imposing additional constraints can never improve the optimal solution. To test if there exist importance weights $V \in \mathcal{V}$ such that $W^{\star\star}$ is a feasible solution to the lower level problem (7), we next solve the following linear programming (LP) problem

$$\min_{V \in \mathcal{V}} L_W(V, W^{\star\star}) = (X_1 - X_0 W^{\star\star})^\top V (X_1 - X_0 W^{\star\star}). \tag{10}$$

While the objective function of problem (10) is the same as that of the lower level problem (7) in that both problems minimize the same loss function, problem (7) is minimized with respect to weights W, whereas problem (10) is minimized with respect to weights V, taking $W^{\star\star}$ as given. This LP problem finds the optimal predictor weights V to support the relaxed problem (3). Denote the optimal solution to problem (10) as $V^{\star\star}$.

If $L_W(V^{\star\star}, W^{\star\star}) = 0$, the optimal solution has been found. In other words, there exists matrix $V^{\star\star} \in \mathcal{V}$ such that $W^{\star\star}$ is a feasible solution to the lower level problem (7), i.e. $W^{\star\star} \in \Psi(V^{\star\star})$. Hence, this is also the optimal solution to the bilevel optimization problem (6)–(7).

3.2. Establishing an upper bound for L_V

In the context of SCM, the domain of predictor weights V has K basic solutions, with the following diagonal elements: $V_1 = (1, 0, \cdots, 0), V_2 = (0, 1, \cdots, 0), \cdots, V_K = (0, 0, \cdots, 1)$. That is, we assign all weight to just one of the predictors, and leave zero weight to all other predictors. We can insert the basic solution $V_k, k = 1, \cdots, K$ as the weights V in problem (8), and solve the QP problem to find the optimal W_k for each $k = 1, \cdots, K$. For each candidate weights W_k, $k = 1, \cdots, K$, we calculate the value of the upper-level loss function L_V stated in (6). Finally, we select the basic solution s in $1, \cdots, K$ that minimizes L_V. If $L_W(V_s, W_s) = 0$ and $L_V(V_s, W_s) = L(W^{\star\star})$, then the corner solution (V_s, W_s) is one of the optimal solutions. If only $L_W(V_s, W_s) = 0$ but $L_V(V_s, W_s) > L(W^{\star\star})$, the corner solution can be viewed as an upper bound for the optimal value.

Proposition 3. *If there exist weights $(\tilde{V}, \tilde{W}) \in \mathcal{V} \times \mathcal{W}$ satisfying $X_{0k}\tilde{W} = x_{1k}$ for some predictor k, then there exists another feasible solution $(V_k, \tilde{W})$ for the SCM problem (6)-(7), where $V_k \in \mathcal{V}$ is a corner solution satisfying $L_W(V_k, \tilde{W}) = 0$. If $(\tilde{V}, \tilde{W})$ is an optimal solution, then also $(V_k, \tilde{W})$ is an alternative optimal solution for the SCM problem.*

This result demonstrates that whenever the donor weights W satisfy the basic condition required for the consistency of the SCM, $X_1 = X_0 W$, even just for a single predictor k, then it is easy to generate feasible solution candidates that are obtained by considering corner solutions with respect to predictor weights V. Intuitively, when the number of predictors is large, it is practically impossible to construct a convex combination of control units that matches the treated unit, in other words, no matrix W that satisfies $X_0 W = X_1$ exists. But if we use weights V to reduce the dimensionality of X by assigning some of the predictors a zero weight, then it becomes considerably easier to find vectors W that satisfy $x_{0k}W = x_{1k}$ at least for some predictor k (note x_{0k} is the kth row of matrix X_0 and x_{1k} is a scalar). Consequently, the set of feasible solutions for the SCM problem often contains several candidate solutions that "switch off" the constraints concerning predictors X by assigning zero weight, except for a single predictor k for which a perfect fit is possible. Therefore, it is understandable that many algorithms attempting to solve the SCM problem (6)–(7) may end up assigning all weight to the most favorable predictor and discard all other predictors by assigning the zero weight. These observations can help to explain the empirical observation that the predictors often turn out to have little impact on the synthetic control, which has been noted by several authors (e.g., Kaul et al., 2021; Doudchenko & Imbens, 2017; Ben-Michael et al., 2021). While these solutions may not necessarily be optimal for the SCM problem, they can still provide good approximations for the optimal value of the upper-level objective. Note that the previous iterations provide us the corner solution (V_k, W_k) and the unconstrained solution $W^{\star\star}$, which can be used for

constructing the following bounds for the loss function of the true optimum $(V^\star, W^\star)$:

$$L_V(V_s, W_s) \geq L_V(V^\star, W^\star) \geq L(W^{\star\star}).$$

If the margin of L_V is small and $W_s \approx W^{\star\star}$ by reasonable tolerance, there is no need to iterate further. But if there is a significant gap, the following iterative procedure is guaranteed to find the optimum.

3.3. Finding an optimal solution using Tykhonov regularization

Building on Proposition 2, the basic idea is to construct an iterative descent algorithm to find the bilevel optimal solution by using the following regularized lower level problem:

$$\min_{W \in \mathcal{W}} L_W^\varepsilon(V, W) = L_W(V, W) + \varepsilon L_V(V, W), \tag{11}$$

where $\varepsilon > 0$. Note that problem (11) is just a re-stated version of the QP problem (8) above. When the optimal solution to the upper-level problem is uniquely defined, the regularized lower-level problem has considerably better regularity properties than the original formulation. In the literature on bilevel programming, this approach is known as Tykhonov regularization (Dempe, 2010). By requiring positive definiteness in the upper-level problem, we can make relatively strong claims regarding the properties of the optimal solutions for the regularized problem. Specifically, it can be shown that the unique optimal solution function to the problem (11), denoted by $W^\star_{\varepsilon_k}(V)$, is Lipschitz continuous and directionally differentiable.

Definition 1 (Lipschitz continuity). A function $z : \mathbb{R}^n \to \mathbb{R}^m$ is called locally Lipschitz continuous at a point $x^0 \in \mathbb{R}^n$ if there exists an open neighborhood $U_\varepsilon(x^0)$ of x^0 and a constant $l < \infty$ such that

$$||z(x) - z(x')|| \leq l||x - x'|| \ \forall x, x' \in U_\varepsilon(x^0).$$

Definition 2 (Directional differentiability). A function $z : \mathbb{R}^n \to \mathbb{R}$ is directionally differentiable at x^0 if for each direction $r \in \mathbb{R}^n$ the following one-sided limit exists:

$$z'(x^0; r) = \lim_{t \to 0+} t^{-1}[z(x^0 + tr) - z(x^0)].$$

The value $z'(x^0; r)$ is called the directional derivative of z at $x = x^0$ in direction r.

Proposition 4. *Consider the synthetic control problem in (6)-(7) and let the upper-level cross-product matrix $Y_0^\top Y_0$ be positive definite. Take any sequence of positive numbers $\{\varepsilon^k\}_{k=1}^\infty$ converging to 0+. Then,*

1. *the optimal value of the regularized bilevel problem converges to the optimal value of the original problem as $k \to \infty$ i.e.*

$$\min_{V,W} \{L_V(V, W) : \ W \in \Psi_{\varepsilon_k}(V), \ V \in \mathcal{V}\} \to L_V^\star,$$

where

$$\Psi_{\varepsilon_k}(V) = \operatorname*{argmin}_{W \in \mathcal{W}} L_W^\varepsilon(V, W),$$

$$L_V^\star = \min_{V,W} \{L_V(V,W) : \ W \in \Psi(V), \ V \in \mathcal{V}\}$$

denote the optimal solution set mapping for (11) *and the upper-level optimal value of the original problem, respectively.*

2. *for each* ε_k*, the unique optimal solution to the regularized lower-level problem* (11)*, denoted by* $W_{\varepsilon_k}^\star(V) \in \Psi_{\varepsilon_k}(V)$*, is directionally differentiable and*

$$\lim_{k\to\infty} \{W_{\varepsilon_k}(V)\} = \operatorname*{argmin}_W \{L_V(V,W) : \ W \in \Psi(V)\}$$

for every fixed $V \in \mathcal{V}$*.*

Based on this result, solving the synthetic control problem is equivalent to considering a sequence of problems

$$\min_V \{L_{\varepsilon_k}(V) : \ V \in \mathcal{V}\} \text{ for } \varepsilon_k \to 0+, \tag{12}$$

where the implicitly defined objective function $L_{\varepsilon_k}(V) = L_V(V, W_{\varepsilon_k}^\star(V))$ is directionally differentiable with respect to V. The implementation of the descent algorithm is discussed in Appendix B.1. As an alternative for the Tykhonov algorithm, the problem can be also solved using a recently developed approach based on KKT-conditions for bilevel problems (Dempe & Franke, 2019). This alternative is briefly described in Appendix B.2.

To summarize this section, the good news is that the SCM problem (6)–(7) is solvable. A bad news is that the required computations prove much more demanding than the original SCM studies assumed. Worse yet, the optimal solution is often a corner solution where most predictors are assigned a zero weight, or have a negligible impact. We stress that imposing some small bounds for V (e.g., $V_{kk} \geq 0.01$) would have little impact in practice, the corner solution would simply assign the minimum weight to all predictors, except for the most favorable predictor that would get the maximum weight (= 1–0.01(K–1)).

4. Comparison of *Synth*, MSCMT, and the global optimum

Applying the iterative algorithm proposed in Section 3 to the data of the two original SCM applications to Basque terrorism (Abadie & Gardeazabal, 2003) and the California tobacco control program (Abadie et al., 2010), we empirically verify that the optimal solution in both cases is indeed a corner solution. The corner solution is found superior to the solutions obtained by *Synth* and the MSCMT algorithm proposed by Becker & Klößner (2018). This observation demonstrates that the existing SCM algorithms fail to find the optimal solution even in the two original applications of SCM, which are also used as illustrative examples for *Synth*.

We compare the results of the following three algorithms: the standard implementation of *Synth* described

in Abadie et al. (2011),[6] the MSCMT package described in Becker & Klößner (2018), and the iterative algorithm proposed by Section 3, which ensures the true global optimum.[7] Tables 1 and 2 report the donor weights (W), the predictor weights (V), and the loss function values of the upper-level problem (L_V) and the lower-level problem (L_W) estimated by different algorithms in R for the Basque terrorism application and the California tobacco control application, respectively. For convenience, we discuss the results of both tables in parallel.

Recall that the value of L_V measures how well the synthetic control matches the pre-treatment outcomes of the treated unit, and this is the upper-level objective to be minimized. In this respect, all algorithms come relatively close to the global optimum. Note that L_V depends on the measurement units of outcomes: for example, multiplying Y_1^{pre} and Y_0^{pre} by 1 Thousand would increase L_V by a factor of 1 Million. Therefore, it is helpful to measure empirical fit with respect to the pre-treatment outcomes in terms of the coefficient of determination (R^2)—after all, the upper-level problem is just constrained least squares regression without intercept. Such a comparison reveals that the differences in empirical fit are rather marginal, the R^2 statistic varies between 0.96866 (*Synth*) to 0.98541 (optimum) in the Basque example and between 0.97518 (*Synth*) and 0.97878 (optimum) in the California example. In contrast, the differences in weights W and V are rather dramatic. The results of Tables 1 and 2 help to illustrate that good empirical fit may be achieved with a wide variety of weights W and V, but there is only one unique global optimum.

The loss function L_W measures how well the synthetic control matches the predictors X_1. Minimization of L_W is the lower-level objective, but the consistency of SCM depends on the (nearly) perfect match with the predictors. In this regard, the relatively high value of L_W given by the standard *Synth* command in both applications indicates that *Synth* fails to converge to the global optimum. Furthermore, the MSCMT procedure greatly improves L_W, but the performance varies between the two empirical examples: L_W converges to the global optimum in the California case but not in the Basque case. In contrast, the value of L_W at the global optimum goes to zero, suggesting a perfect match in terms of the weighted predictors. However, this is an illusion because the optimal solution is a corner solution that assigns all weight to a single predictor: real per capita GDP in the Basque terrorism application and cigarette sales per capita in 1980 in the California tobacco control application (see Tables 1 and 2). The MSCMT algorithm comes close to the corner solution in the former application, but fails to converge to the corner solution in the latter. The *Synth* algorithm appears to use more balanced weights for predictors, however, note that *Synth* also assigns almost 90% of the predictor weight to cigarette sales per capita (the outcome variable) during two years of the pre-treatment period. Unfortunately, *Synth* fails to solve the optimization problem it is supposed to solve; its predictor weights are not what they

[6] In addition to the standard *Synth* command, we have also considered the genoud() option available in *Synth*, as noted in Abadie et al. (2011). However, the use of the genoud() option does not improve the matter; in fact, the solution is only worse.

[7] The R code to implement this algorithm is documented in Appendix C. The latest updates to the R code are available on the GitHub page: `https://github.com/Xun90/SCM-Debug.git`.

Table 1: Basque terrorism application revisited: donor weights, predictor weights, loss functions, and empirical fit by different algorithms.

	Synth	MSCMT	Optimum
W			
Catalonia	0.8508	0.6328	0.0000
Madrid	0.1492	0.1479	0.4405
Baleares	0.0000	0.2193	0.3700
La Rioja	0.0000	0.0000	0.1895
V			
Schooling of working age population (%)			
Illiterates	0.0156	0.0000	0
Up to primary school	0.0018	0.0000	0
With some high school	0.0442	0.0000	0
With high school or above	0.0341	0.0003	0
Investment ratio	0.0001	0.0003	0
Real GDP per capita	0.2010	0.9993	1
Sectoral shares (%)			
Agriculture, forestry, and fishing	0.0948	0.0000	0
Energy and water	0.0077	0.0000	0
Industry	0.1339	0.0000	0
Construction and engineering	0.0087	0.0000	0
Marketable services	0.0097	0.0000	0
Non-marketable services	0.1081	0.0000	0
Population density	0.3403	0.0000	0
L_V	0.00886	0.00429	0.00413
L_W	0.24670	0.00034	0.00000
R^2	0.96866	0.98485	0.98541

Table 2: California tobacco control application revisited: donor weights, predictor weights, loss functions, and empirical fit by different algorithms.

	Synth	MSCMT	Optimum
W			
Utah	0.3432	0.3351	0.3939
Nevada	0.2358	0.2356	0.2049
Montana	0.1820	0.2019	0.2318
Colorado	0.1747	0.1595	0.0148
Connecticut	0.0624	0.0679	0.1091
New Hampshire	0.0000	0.0000	0.0454
V			
Income per capita	0.0006	0.0000	0
Retail price of cigarettes	0.0312	0.3333	0
Population aged 15–19 (%)	0.0034	0.3333	0
Beer consumption per capita	0.0124	0.0000	0
Cigarette sales per capita 1988	0.0682	0.0000	0
Cigarette sales per capita 1980	0.3917	0.0000	1
Cigarette sales per capita 1975	0.4925	0.3333	0
L_V	3.20908	3.07666	2.74366
L_W	0.00170	0.00000	0.00000
R^2	0.97518	0.97621	0.97878

are claimed to be, but just artifacts of a computational failure.

Of course, the most important piece of information for SCM are the donor weights W, which are used to form the synthetic control. As noted above, a marginal improvement in the empirical fit leads to rather dramatic changes in the composition of the synthetic control. Consider first the synthetic control for Basque. The *Synth* algorithm identifies Catalonia and Madrid as the benchmarks, with 85% weight assigned to Catalonia. The solution found by the MSCMT algorithm reassigns 22 percentage points of Catalonia's weight to the Balearic Islands, maintaining the weight of Madrid. In sharp contrast, the global optimum assigns no weight to Catalonia, whereas the largest weights are assigned to Madrid (44%) and the Balearic Islands (37%), but also the neighboring region of La Rioja enters the synthetic control with the 19% weight. Consider next the synthetic control for California. *Synth* and MSCMT yield almost the same donor weights despite their different estimates of the loss function values. However, the global optimum reassigns nearly all of Colorado's weight and 4 percentage points of Nevada's weight to Utah (consolidating as the largest weighting state), Montana, Connecticut, and New Hampshire (a new state entering the synthetic California).

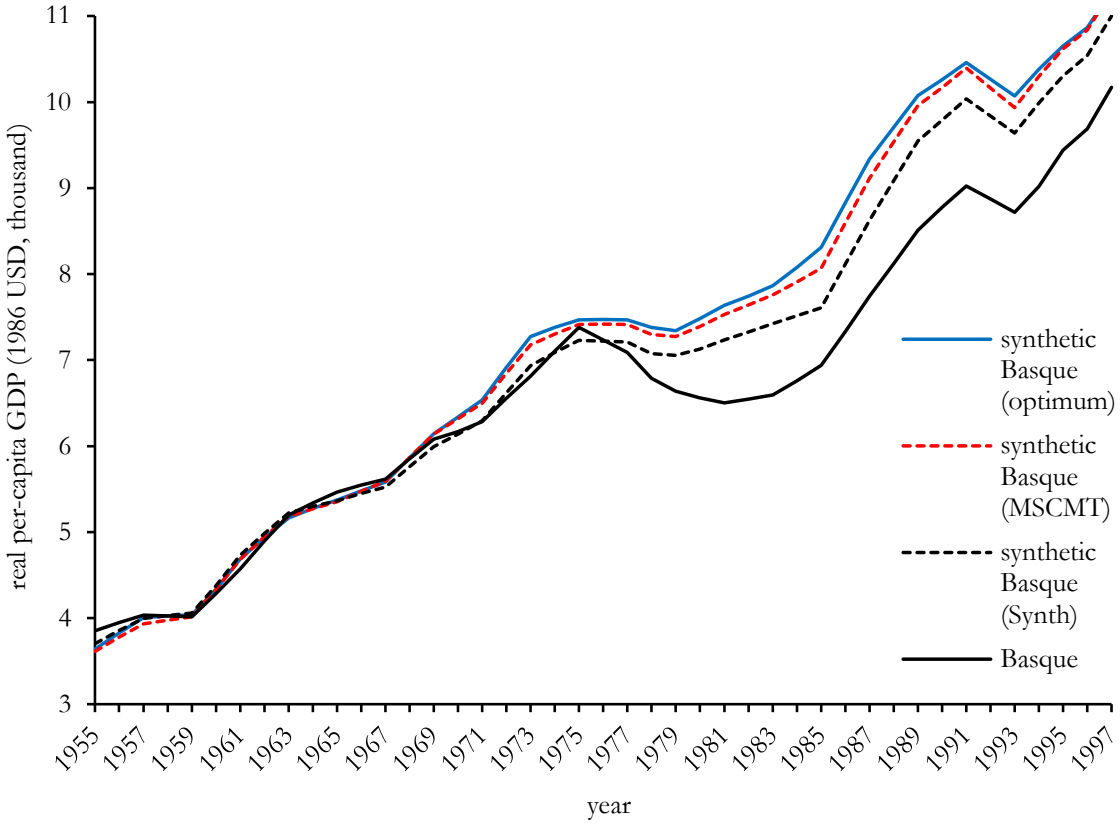
Figure 1 illustrates the impact of suboptimal donor weights on the evolution of the synthetic Basque (panel 1a) and the synthetic California (panel 1b). Fortunately the qualitative conclusions of these two original and highly influential applications remain, but the suboptimal weights lead to a lower treatment effect in both cases, particularly in the Basque terrorism application. We stress that the globally optimal weights minimize the MSPE of the pre-treatment outcomes Y_1^{pre}, but there is no guarantee that the weights are optimal to minimize the MSPE of the counterfactual because the good empirical fit to pre-treatment outcomes was achieved by disregarding all predictors except for one. We compare the solutions produced by the *Synth* and MSCMT algorithms to the global optimum just to illustrate the computational failure, but the practical use of this global optimum is not the approach that we advocate. We agree with Albalate et al.'s (2021) recent proposal to determine the predictor weights and donor weights separately.

5. Alternative data-driven approaches

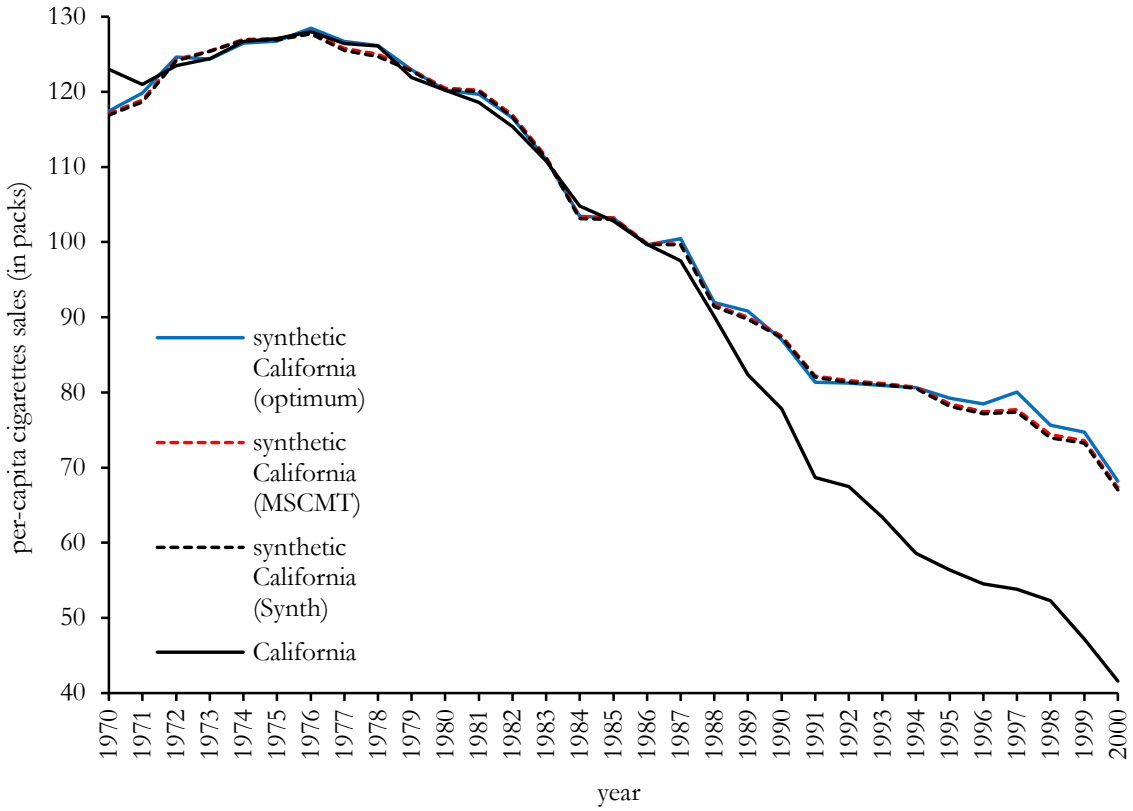
5.1. Optimizing donor weights when predictor weights are given

In the previous section we found that the original SCM problem is solvable, but unfortunately, the solution is not nice. In light of the arguments presented in the previous section, we would strongly recommend the users of SCM to determine the predictor weights V separately, before optimizing the donor weights W.

In this sub-section we develop a simple iterative procedure to compute the optimal weights W when the predictor weights $V^\star$ are given *a priori*. Recall from Section 3 the Tykhonov descent approach where the non-Archimedean ε is gradually decreased towards zero. In practice, it is difficult to ensure that ε is sufficiently close to zero to give the priority to the lower-level objective function L_W, but high enough to achieve coordination with the upper-level objective L_V. To operationalize the theoretical idea of Section 3, we propose to optimize the weights W using the following two-step procedure when the predictor weights $V^\star$ are predetermined:



(a) Basque terrorism



(b) California's tobacco control program

Figure 1: The impact of suboptimal W weights on the evolution of synthetic controls.

Step 1: Solve the QP problem

$$\min_W L_W = (X_1 - X_0 W)^\top V^\star (X_1 - X_0 W)$$

subject to

$$\mathbf{1}^\top W = 1$$

$$W \geq \mathbf{0}$$

Step 2: Given the optimal $L_W^\star$ from Step 1, solve the convex programming problem

$$\min_W L_V(V, W) = (Y_1^{\text{pre}} - Y_0^{\text{pre}} W)^\top (Y_1^{\text{pre}} - Y_0^{\text{pre}} W)$$

subject to

$$(X_1 - X_0 W)^\top V^\star (X_1 - X_0 W) = L_W^\star$$

$$\mathbf{1}^\top W = 1$$

$$W \geq \mathbf{0}$$

Breaking the problem into two separate stages allows to eliminate the non-Archimedean ε in (8). In Step 1 we minimize the lower-level objective function L_W, and its optimal value is subsequently inserted as a constraint to the optimization problem in Step 2. This establishes an explicit link between the upper-level and the lower-level objectives. The two-step procedure explicitly considers the possibility of alternate optima in Step 1. Since the *Synth* algorithm does not take the possibility of alternate optima into account, there is no guarantee that it finds the optimal donor weights W even when the predictor weights V are defined by the user (see Appendix B of Kuosmanen et al., 2021 for a numerical demonstration). In the next sub-sections we explore and demonstrate alternative data-driven strategies to determine the weights V empirically.

Before proceeding to the predictor weights, it is worth to note the recent study by Abadie & L'Hour (2020), which similarly takes the predictor weights V as given. The authors deviate from the original SCM approach in that they focus solely on the lower-level objective of optimizing the fit with respect to the predictors, ignoring the upper-level objective of optimizing the fit with respect to the pre-treatment outcomes. The authors introduce an additional penalty to minimize the sum of pairwise matching discrepancies, which ensures that the optimal donor weights are unique in this new setting. The additional penalty term to improve matching is a valuable extension, which could be readily combined with the developments of our study. However, omitting the upper-level objective function would typically result as poor fit to the pre-treatment outcomes. Of course, one might incorporate pre-treatment outcomes among the predictors, but this would quite dramatically change the logic of the original SCM. In mathematical terms, the original bilevel optimization problem would then become a multi-objective optimization problem where the weights V govern the relative importance assigned to the empirical fit to the pre-treatment outcomes and the fit to the additional predictors, respectively.

5.2. Panel regression approach to determine predictor weights

There are several possibilities to set weights V based on empirical data. Both Abadie & Gardeazabal (2003) and Abadie et al. (2010) discuss the possibility to use subjectively determined weights V. The default option of the Stata implementation of the *Synth* package is to use regression-based weights V, which are also used as starting values in the R and Matlab implementation of *Synth* (see Abadie et al. (2011)). In this sub-section we similarly resort to a regression-based approach, but propose some modifications to the *Synth* approach.

If panel data of predictors X are available, we propose to first estimate the equation

$$y_{jt}^{\text{pre}} = \mu + X_{jt}'\boldsymbol{\beta} + \gamma_j + \varepsilon_{jt} \qquad j = 1, 2, \ldots, J+1;\ t = 1, 2, \ldots, T^{\text{pre}}. \tag{13}$$

Model (13) can be estimated by standard fixed effects (FE) or random effects (RE) panel data regression. Note that the FE estimator cannot be used when there are time-invariant predictors. The original SCM application to Basque terrorism, to be revisited below, does include some time-invariant predictors. Therefore, we will resort to the RE estimator below, assuming that the random effects γ_j are uncorrelated with the predictors.

Given estimated coefficients $\hat{\boldsymbol{\beta}}$, we propose to assign weights V based the absolute values of the parameter estimates, that is

$$v_k = |\hat{\beta}_k| \Big/ \sum_{j=1}^{K} |\hat{\beta}_j|. \tag{14}$$

We note that the *Synth* algorithm uses the squared values of the parameter estimates to assign weights V. By using the absolute values rather than squared values, one achieves a more equal balance between different predictors.

Having optimized the predictor weights, we apply the two-step procedure proposed in Section 5.1 to optimize the donor weights. Given the optimal donor weights $W^\star$, we estimate the counterfactual as

$$Y_1^{\text{N}} = Y_0 W^\star + (\hat{\gamma}_1 - \hat{\boldsymbol{\gamma}}_0^\top W^\star). \tag{15}$$

Note that the random effects γ_j were not taken into account in the optimization of the donor weights. Therefore, we utilize the estimated random effects to implement the standard bias correction, following Ben-Michael et al. (2021) and Ferman et al. (2020).

We next illustrate the regression-based approach outlined above by reexamining the original SCM application to Basque terrorism. Imputing the missing values by suitable methods (see Appendix D for details), we obtain panel data for most of the predictors during the pre-treatment period. In the RE panel regression to set weights V, we excluded the real GDP per capita, the percentage of the illiterate working-age population, and the sectoral share of non-marketable services to avoid perfect collinearity. Table 3 reports the RE estimates of predictor coefficients and the empirical V weights determined by equation (14) for the Basque example. The percentage of the working age population with some high school and the sectoral share of marketable services are found to be statistically significant predictors. Together with the percentage of the working age population

with high school or higher education, those two significant predictors are the three most influential predictors that receive more than 70% weight. On the other hand, the empirical V weights are relatively balanced among the other predictors, except for population density, which is attributed less than 1% weight. In addition, the overall empirical fit of the RE panel regression is 0.8808, with the between and within effects being 0.8734 and 0.9277, respectively. Note that 78% of the unexplained variation of the outcome is attributed to the random effects and that the random effects are statistically significant.

Given the empirically set V weights, we next determine the optimal W weights to construct the synthetic Basque by using the two-step procedure described in Section 5.1. The donor weight is assigned to Cantabria (79.9%), Catalonia (12.4%), and Madrid (7.7%). Interestingly, Cantabria enters the synthetic control with a large weight. Cantabria is a neighboring region to the Basque Country, but it was not included in any of the the three synthetic controls considered in Section 4.

Figure 2 illustrates the impact of the alternative strategy to set V on the evolution of the synthetic Basque. The time series start from 1960, which is the first year in the panel model. Note that the absolute RE weights approach with bias-correction yields notably better fit to the pre-treatment outcomes than the SCM that does not use any predictors, which is exactly the same as the "global optimum" considered in Section 4 obtained by assigning all weight to a single predictor. The synthetic Basque based on the absolute RE weights still identifies the treatment effect of Basque terrorism on real GDP per capita. However, the treatment effect is considerably smaller than the synthetic control that does not use any predictors. The treatment effect disappears by the mid-1990s. This example illustrates that appropriate use of the predictors does influence the results, and can potentially affect the qualitative conclusions.

One of the key assumptions of any treatment effect model is that the control group is not exposed to the treatment. This assumption does not, strictly speaking, hold in the present application because a significant proportion of Euskadi Ta Askatasuna (ETA)'s terrorism activity took place in other regions, including Madrid and Catalonia, which have large weight in the synthetic control. Abadie & Gardeazabal (2003) indicate that 69% of deaths attributed to terrorism occurred in the Basque Country, which directly implies that almost one third of deaths occurred in the regions that form the donor pool. Further, the specification of the pre-treatment and post-treatment periods (before and after 1970, respectively) could be debated. ETA was founded in 1968 and there were three victims during the pre-treatment period, but only one victim during the first three years of the post-treatment period. The difference between the actual outcome and the counterfactual synthetic control becomes evident from the year 1975 onwards, which matches perfectly with the death of Dictator Franco and the transition towards democracy. While we do not intend to deny the economic cost of ETA's terrorism, perhaps at least some part of the observed treatment effect may be attributed to the economic transition from Franco's dictatorship to democracy, which had varying effects across different regions of Spain. Of course, ETA's terrorism is also closely related to this historical context, but ETA's terrorism did not cause the major political regime shift in Spain.

Table 3: Predictor coefficients and empirical predictor weights for the Basque example.

Predictors	Coefficients	Robust standard errors	Empirical V
Schooling of working age population (%)			
Up to primary school	0.0397	0.0264	0.0532
With some high school	0.2567***	0.0527	0.3439
With high school or above	0.2126	0.2275	0.2848
Investment ratio	-0.0085	0.0068	0.0114
Sectoral shares (%)			
Agriculture, forestry, and fishing	0.0150	0.0335	0.0201
Energy and water	0.0196	0.0389	0.0262
Industry	0.0446	0.0368	0.0598
Construction and engineering	-0.0477	0.0715	0.0639
Marketable services	0.1007**	0.0397	0.1349
Population density	-0.0014	0.0016	0.0019
Intercept	-5.7426**	2.9123	
R^2: within = 0.9277, between = 0.8734, overall = 0.8808			
$\sigma_{\hat{\gamma}} = 0.2062$, $\sigma_{\hat{\varepsilon}} = 0.1099$, $\rho = 0.7789$ (fraction of variance due to γ_i)			

NOTE: * $p \leq 0.10$; ** $p \leq 0.05$; *** $p \leq 0.01$.

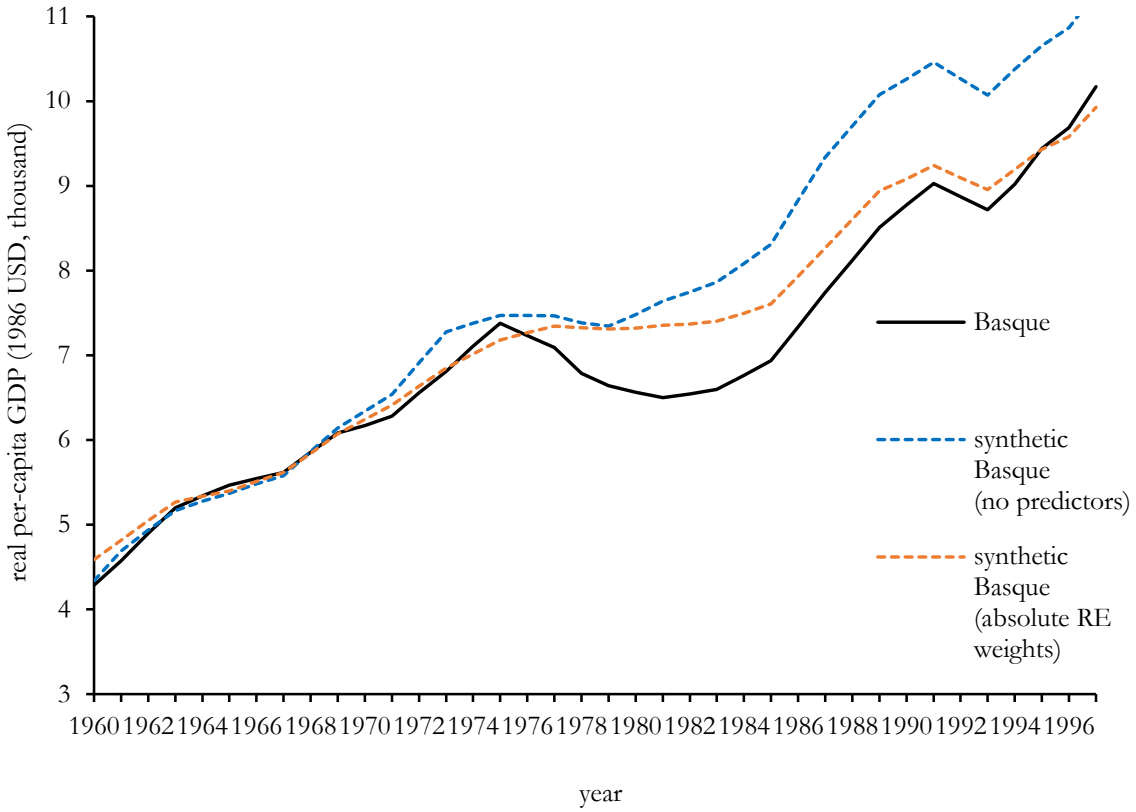


Figure 2: The impact of alternative approaches on the evolution of synthetic Basque.

5.3. Uniform weights to standardized predictors

Suitable panel data are not always available for the purposes of SCM. The original application to California's tobacco control program is one example of such application. Another possibility would be to apply uniform V weights when panel data for the predictors are simply unavailable. In this approach, we propose to first standardize the predictors as

$$z_{ik} = (x_{ik} - \bar{X}_k)) \ / \ \text{std}(X_k).$$

and subsequently apply equal weights $v_k = 1/K$ to the standardized predictors. By doing so, all predictors will count, and the weights are invariant to rescaling or changing the units of measurement.

We next illustrate the application of uniform V weights by revisiting the California tobacco control application. The donor weights are obtained by applying the two-stage procedure proposed in Section 5.1. This yields the following optimal donor weights: Colorado (62.6%), Connecticut (27.8%), Texas (6.5%), and Utah (3.2%). Colorado was included in the synthetic control in the results of Section 4, but the use of standardized uniform predictor weights notably increases its weight. In contrast, Utah was previously assigned the largest weight, but in the present analysis it gets only 3.2% weight.

Figure 3 illustrates the impact of the uniform V on the evolution of the synthetic California. Note that in this example the uniform V approach leads to worse empirical fit to the pre-treatment outcomes than the SCM

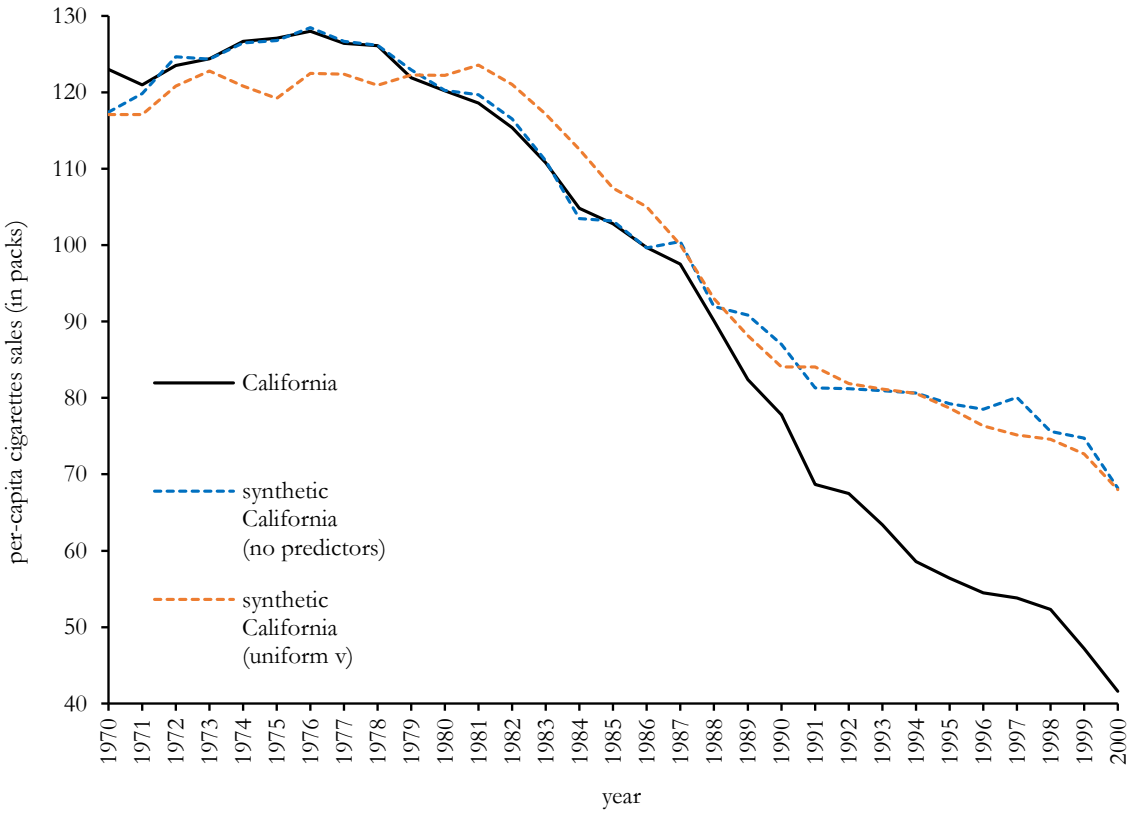


Figure 3: The impact of alternative approaches on the evolution of synthetic California.

that does not use any predictors. There is a trade-off: when we put more emphasis on optimizing the empirical fit with respect to predictors X, then the fit with respect to pre-treatment outcomes is likely to deteriorate, and vice versa. In our interpretation, Figure 3 is a useful illustration of why focusing solely on optimizing the fit with respect to predictors, ignoring the pre-treatment outcomes, is not necessarily a viable solution. In many applications, the good pre-treatment fit of *Synth* is to some extent illusion because it tends to put negligibly small weight to many predictors.

However, it is reassuring to find that the post-treatment outcomes of the synthetic California based on uniform V are very similar to those of the synthetic California in the absence of predictors. Therefore, the use of predictors mainly affects the pre-treatment fit, but not so much the post-treatment. One would be mainly interested in the post-treatment effect, so this would help to support the empirical finding that there was indeed impact. In fact, we suggest that one could examine a range of alternative V weights for testing robustness of the treatment effect (as an additional tool, in addition to the placebo trials and statistical tests that are already known in the literature).

In summary, the main point of Section 5 is to demonstrate that alternative data-driven approaches to determine the weights V are available. The empirical comparisons above demonstrate that the introduction of empirically determined V weights presents a viable remedy to the ill-designed *Synth* algorithm. While the relative merits of the alternative approaches clearly warrant further research, in light of the problems discussed

in Sections 2 and 4, we strongly recommend that the suboptimal weights produced by *Synth* should not be used.

6. Conclusions

SCM has proved a highly appealing approach to estimate causal treatment effects, as a large number of published applications demonstrates. Unfortunately, the computational difficulties caused by joint optimization of the donor weights and the predictor weights not only result as inaccuracy and numerical instability, but in our view, cast serious doubts on the reliability of the original SCM and the *Synth* package. Referring to the secondary title of this article, we would classify the synthetic control methods currently available into three groups.

First, we would argue that decoupling the nested optimization problems of predictor weights and donor weights is the *good* approach and the most promising way forward, in agreement with Albalate et al. (2021). In this paper we developed a simple two-step algorithm to optimize the donor weights when the predictor weights are given *a priori*, and also briefly explored two alternative data-driven approaches to determine the predictor weights using regression analysis or applying uniform weights to standardized predictors, thereby complementing the SHAP approach proposed by Albalate et al. (2021).

Second, we developed an iterative computational algorithm for solving the original SCM problem, which turned out to be a NP-hard bi-level optimization problem. We were the first ones to prove that our SCM algorithm converges to the optimal solution. However, empirical application of the new algorithm strongly suggests that the true optimal solution is typically a corner solution where all predictor weight is assigned to a single predictor. We show that this is the case in the two classic SCM applications by Abadie & Gardeazabal (2003) and Abadie et al. (2010). We stress that development of a better computational algorithm is not the solution that we advocate because it does not help to address the root cause of the problem. Due to the computational complexity of the original SCM formulation, we consider this a *bad* method.

Third, the *ugly* solution is to continue the use of the *Synth* package in applications, despite the accumulating evidence on its numerical instability and demonstrably suboptimal solutions, which may distort the qualitative conclusions. We sincerely hope that the results of this paper would not only contribute to the better understanding of the synthetic control methods, but also help to facilitate the good practices more broadly to the empirical practice in economics, operational research, statistics, and other related areas. The synthetic control methods are already highly influential for decision-making involving important societal problems, and have potential to become even more influential in the future.

We hope that our study could open important avenues for future research, both empirical and methodological studies. From the empirical point of view, the findings of our paper call for systematic replication of the published SCM studies to examine whether and to what extent the use of suboptimal weights produced by *Synth* has affected the qualitative conclusions. Becker & Klößner (2017) is an excellent example of such a replication

study. We hope that the qualitative results of the influential SCM studies prove robust to the optimization errors that are evidently present, but this remains to be tested empirically. Our replication of the two original applications of SCM showed that the suboptimal weights yield somewhat different results than the optimal ones, but fortunately the qualitative conclusions of these two studies remain.

From the methodological point of view, it would be urgent to further examine alternative decoupled data-driven approaches to determine the predictor weights, including regression-based methods, to gain a better understanding of which method is most suitable for the purposes of SCM. While we strongly recommend the users of the classic SCM to determine the predictor weights *a priori*, we do not consider the joint optimization of the predictor weights and the donor weights entirely hopeless. However, the loss function to be minimized requires careful reconsideration to ensure that the optimal solution is meaningful for the intended purposes of using the predictors, and that the problem remains computationally tractable. It would also be helpful to establish more detailed practical guidelines regarding what kind of variables are suitable predictors for SCM. At present, many SCM studies include a mixed set of predictors expressed in levels, logs, differences, and percentage growth rates, which may leave too much room for a user to manipulate the results by creative data transformations.

References

Abadie, A. (2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, *59*, 391–425.

Abadie, A., Diamond, A., & 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*, 493–505.

Abadie, A., Diamond, A., & Hainmueller, J. (2011). Synth: An R package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, *42*, 1–17.

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, *59*, 495–510.

Abadie, A., & Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, *93*, 113–132.

Abadie, A., & L'Hour, J. (2020). A Penalized Synthetic Control Estimator for Disaggregated Data. *Work. Pap., Mass. Inst. Technol., Cambridge, MA*.

Acemoglu, D., Johnson, S., Kermani, A., Kwak, J., & Mitton, T. (2016). The value of connections in turbulent times: Evidence from the United States. *Journal of Financial Economics*, *121*, 368–391.

Albalate, D., Bel, G., & Mazaira-Font, F. A. (2021). Decoupling synthetic control methods to ensure stability, accuracy and meaningfulness. *SERIEs*.

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

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

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

Becker, M., Klößner, S., & Pfeifer, G. (2018). Cross-validating synthetic controls. *Economics Bulletin*, *38*, 603–609.

Ben-Michael, E., Feller, A., & Rothstein, J. (2021). The Augmented Synthetic Control Method. *Journal of the American Statistical Association*.

Bloom, N., & Van Reenen, J. (2007). Measuring and Explaining Management Practices Across Firms and Countries. *The Quarterly Journal of Economics*, *122*, 1351–1408.

Bohn, S., Lofstrom, M., & Raphael, S. (2014). Did the 2007 legal Arizona workers act reduce the state's unauthorized immigrant population? *Review of Economics and Statistics*, *96*, 258–269.

Cavallo, E., Galiani, S., Noy, I., & Pantano, J. (2013). Catastrophic natural disasters and economic growth. *Review of Economics and Statistics*, *95*, 1549–1561.

Charnes, A. (1952). Optimality and Degeneracy in Linear Programming. *Econometrica*, *20*, 160–170.

Cole, M. A., Elliott, R. J., & Liu, B. (2020). The Impact of the Wuhan Covid-19 Lockdown on Air Pollution and Health: A Machine Learning and Augmented Synthetic Control Approach. *Environmental and Resource Economics*, *76*, 553–580.

Dempe, S. (2010). *Foundations of Bilevel Programming*. Kluwer Academic Publishers.

Dempe, S., & Franke, S. (2019). Solution of bilevel optimization problems using the KKT approach. *Optimization*, *68*, 1471–1489.

Doudchenko, N., & Imbens, G. W. (2017). Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. *arXiv preprint arXiv:1610.07748*.

Ferman, B., Pinto, C., & Possebom, V. (2020). Cherry Picking with Synthetic Controls. *Journal of Policy Analysis and Management*, *39*, 510–532.

Gobillon, L., & Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, *98*, 535–551.

Hansen, P., Jaumard, B., & Savard, G. (1992). New Branch-and-Bound Rules for Linear Bilevel Programming. *SIAM Journal on Scientific and Statistical Computing*, *13*, 1194–1217.

Kaul, A., Klößner, S., Pfeifer, G., & Schieler, M. (2021). Standard Synthetic Control Methods: The Case Of Using All Preintervention Outcomes Together With Covariates. *Journal of Business and Economic Statistics*.

Kleven, H. J., Landais, C., & Saez, E. (2013). Taxation and international migration of superstars: Evidence from the European football market. *American Economic Review*, *103*, 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*, 11.

Klößner, S., & Pfeifer, G. (2015). Synthesizing Cash for Clunkers: Stabilizing the Car Market, Hurting the Environment. In *Beiträge zur Jahrestagung des Vereins für Socialpolitik 2015: Ökonomische Entwicklung - Theorie und Politik - Session: Automobiles and the Environment*. ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum WirtschaftKiel und Hamburg volume F13-V1.

Kuosmanen, T., Zhou, X., Eskelinen, J., & Malo, P. (2021). Design Flaw of the Synthetic Control Method. *MPRA Paper 106390*.

Malo, P., Eskelinen, J., Zhou, X., & Kuosmanen, T. (2020). Computing Synthetic Controls Using Bilevel Optimization. *MPRA Paper 104085*.

Sinha, A., Malo, P., & Deb, K. (2013). Efficient Evolutionary Algorithm for Single-Objective Bilevel Optimization. *arXiv preprint arXiv:1303.3901*.

Vicente, L., Savard, G., & Júdice, J. (1994). Descent approaches for quadratic bilevel programming. *Journal of Optimization Theory and Applications*, *81*, 379–399.

Online Supplement for "Synthetic Control Methods: The Good, the Bad, and the Ugly"

Appendix A. Proofs of theorems

Appendix A.1. Regularity Conditions for Parametric Optimization

In this section, we will briefly review a few central concepts from parametric optimization literature that we will later need while discussing the notions of optimality for the synthetic control problem. Without loss of generality, the lower level problem can be stated as a parametric optimization problem

$$\min_y \{f(x,y) : g(x,y) \leq 0, h(x,y) = 0\}, \tag{A.1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$, $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^q$. The constraints

$$\begin{aligned} g(x,y) &= (g_1(x,y), \ldots, g_p(x,y))^\top, \\ h(x,y) &= h_1(x,y), \ldots, h_q(x,y))^\top, \end{aligned}$$

are assumed to be smooth vector-valued functions. The problem is a convex parametric optimization problem, when all functions $f(x,\cdot)$, $g_i(x,\cdot)$, $i = 1, \ldots, p$, are convex and the functions $h_j(x,\cdot)$, $j = 1, \ldots, q$, are affine-linear on $\mathbb{R}^n$ for each fixed $x \in \mathbb{R}^n$. The solution set mapping $\Psi : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is defined by

$$\Psi(x) = \operatorname*{argmin}_y \{f(x,y) : g(x,y) \leq 0, h(x,y) = 0\},$$

which is a point-to-set mapping from the upper level decisions to the set of global optimal solutions of the parametric problem. For convex problems, the solution sets $\Psi(x)$ are closed and convex subsets of $\mathbb{R}^m$.

When it comes to regularity conditions in bilevel programming, the following two conditions have often been utilized. The first condition is concerned with compactness of the feasible set of the lower level problem:

Definition 3 (C). The set $\{(x,y) : \mathbb{R}^n \times \mathbb{R}^m : g(x,y) \leq 0, h(x,y) = 0\}$ is non-empty and compact.

This is enough to guarantee that the set of optimal solutions for the parametric problem

$$\Psi(x) := \operatorname*{argmin}_y \{f(x,y) : g(x,y) \leq 0, h(x,y) = 0\}$$

is non-empty and compact for each $x \in \{z : \Omega(z) \neq \emptyset\}$, where

$$\Omega(x) = \{y \in \mathbb{R}^m : g(x,y) \leq 0, h(x,y) = 0\}$$

is the feasible set mapping for the lower level problem.

The second regularity condition is the commonly applied Mangasarian-Fromowitz constraint qualifications:

Definition 4 (MFCQ). We say that Mangasarian-Fromowitz constraint qualification is satisfied at point (x^0, y^0) if there exists a direction $d \in \mathbb{R}^m$ such that

$$\nabla_y g_i(x^0, y^0)d < 0, \text{ for each } i \in I(x^0, y^0) = \{j : g_j(x^0, y^0) = 0\},$$

$$\nabla_y h_j(x^0, y^0)d = 0, \text{ for each } j = 1, \ldots, q$$

and the gradients of the equality constraints $\{\nabla_y h_j(x^0, y^0) : j = 1, \ldots, q\}$ are linearly independent.

These regularity conditions play an important role in ensuring existence of optimal solutions for optimistic bilevel problems such as the synthetic control problem discussed in this paper. Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ denote the upper level objective function that is minimized with respect to upper-level constraints $X := \{x : G(x) \leq 0\}$, $G : \mathbb{R}^n \to \mathbb{R}^l$. An optimistic solution to a bilevel problem can then be defined as a point solving the following minimization problem:

$$\min_x \{\varphi_0(x) : x \in X\}, \tag{A.2}$$

where $\varphi_0(x) = \min_y \{F(x, y) : \ y \in \Psi(x)\}$.

Theorem 1 (Dempe, 2010)**.** *Let the assumptions (C) and (MFCQ) be satisfied at all points $(x, y) \in X \times \mathbb{R}^m$ with $y \in \Omega(x)$. Then, a global solution of the bilevel problem (A.2) exists provided there is a feasible solution.*

In addition to the existence of optimal solutions, the regularity conditions imply upper-semicontinuity of the optimal solution set mapping.

Definition 5 (Upper semicontinuity). A set-valued mapping $\Psi : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is said to be upper semicontinuous at a point $x \in \mathbb{R}^n$ if, for each open set V with $\Psi(x) \subset V$, there exists an open neighborhood $U_\delta(x)$ of x such that $\Psi(x') \subset V$ for each $x' \in U_\delta(x)$.

In the special case, where Ψ is a single-valued mapping, the notion of upper semicontinuity corresponds to the usual continuity of a function.

Theorem 2 (Bank et al., 1982; Dempe, 2010)**.** *Consider the parametric optimization problem (A.1) at $x = x^0 \in \mathbb{R}^n$ and let the assumptions (C) and (MFCQ) be satisfied for all feasible points (x, y) with $x = x^0$ and $y \in \Omega(x^0)$. Then, the solution set mapping Ψ is upper semicontinuous and the optimal value function φ is continuous at x^0.*

While the solution set mapping is upper semicontinuous under these relatively weak regularity conditions, it is generally not continuous. The continuity of a solution set mapping is possible only under considerably stronger assumptions such as the strong sufficient optimality condition of second order (SSOC) and constant rank constraint qualification (CRCQ).

Definition 6 (SSOC). The strong sufficient optimality condition of second order holds at (x^0, y^0) if for each pair of Lagrange multipliers $(\lambda, \mu) \in \Lambda(x^0, y^0)$ and for each direction $d \neq 0$ with

$$\nabla_y g_i(x^0, y^0)d = 0, \ \forall i \in J(\lambda) := \{j : \lambda_j > 0\},$$

$$\nabla_y h_j(x^0.y^0)d = 0, \ j = 1, \ldots, q$$

we have that

$$d^\top \nabla_{yy} L(x^0, y^0, \lambda, \mu)d > 0.$$

Definition 7 (CRCQ). The constant rank constraint qualification holds at point $(x^0, y^0))$ if there exists an open neighborhood $U_\varepsilon(x^0, y^0)$ of (x^0, y^0) such that for each subset

$$I \subset I(x^0, y^0) := \{i : g_i(x^0, y^0) = 0\}, \ J \subset \{1, \ldots, q\},$$

the family of gradient vectors

$$\{\nabla_y g_i(x, y) : i \in I\} \cup \{\nabla_y h_j(x, y) : j \in J\}$$

has the same rank for all $(x, y) \in U_\varepsilon(x^0, y^0)$.

Let $L(x, y, \lambda, \mu) = f(x, y) + \lambda^\top g(x, y) + \mu^\top h(x, y)$ denote the Lagrangian function of problem (A.1) and let

$$\Lambda(x, y) = \{(\lambda, \mu) \in \mathbb{R}^p \times \mathbb{R}^q : \lambda \geq 0, \lambda^\top g(x, y) = 0, \nabla_y L(x, y, \lambda, \mu) = 0\}$$

be the set of Lagrange multipliers at (x, y).

Theorem 3 (Dempe, 2010)**.** *Consider the problem (A.1) at $x = x^0 \in \mathbb{R}^n$ and let the assumptions (MFCQ), (SSOC), and (CRCQ) be satisfied at (x^0, y^0) with y^0 being a unique local optimal solution. Then, there exists a unique local optimal solution function $y(\cdot)$ that is locally Lipschitz continuous and directionally differentiable at $x = x^0$. The directional derivative in direction r coincides with the unique optimal solution of the following quadratic programming problem*

$$\begin{aligned}
\min_d \quad & 0.5d^\top \nabla^2_{yy} L(x^0, y^0, \lambda^0, \mu^0)d + d^\top \nabla^2_{xy} L(x^0, y^0, \lambda^0, \mu^0)r, \\
s.t. \quad & \nabla_y g_i(x^0, y^0)d + \nabla_x g_i(x^0, y^0)r \begin{cases} = 0, & \text{if } i \in J(\lambda^0), \\ \leq 0, & \text{if } i \in I(x^0, y^0) \setminus J(\lambda^0), \end{cases} \\
& \nabla_y h_j(x^0, y^0)d + \nabla_x h_j(x^0, y^0)r = 0 \text{ for all } j = 1, \ldots, q,
\end{aligned}$$

for any $(\lambda^0, \mu^0) \in \Lambda(x^0, y^0)$ that solve

$$\max_{(\lambda,\mu)\in\Lambda(x^0,y^0)} \nabla_x L(x^0, y^0, \lambda, \mu).$$

Appendix A.2. Proof of Proposition 1

To show existence of a global optimal solution, it is enough to verify that assumptions (C) and (MFCQ) are satisfied.

Let $g(V,W) = -W$ and $h(V,W) = \sum_{j=1}^{J} W_j - 1$ denote the constraints in the lower-level problem. Clearly, the set $\{(V,W) \in \mathbb{R}^{K\times K} \times \mathbb{R}^J : \ g(V,W) \leq 0, \ h(V,W) = 0\}$ is non-empty and compact. Therefore, condition (C) holds.

To check (MFCQ), let $(V_0, W_0) \in \mathcal{V} \times \mathcal{W}$ and define

$$I_0 = \{j : g_j(V_0, W_0) = -W_j = 0\}.$$

If $W_0 > 0$, we have $I_0 = \emptyset$ and (MFCQ) holds trivially. If there exists at least some index j such that $W_{0,j} = 0$, we need to check the gradient conditions. Let $d \in \mathbb{R}^J$ be a candidate direction. From the inequality constraints we have that $\nabla_w g(V,W)d = -d < 0$, which means that for every $j \in I_0$, we require $d_j > 0$. When combined with the equality constraint we have that

$$\nabla_w h(V_0, W_0)d = \sum_{j\in I_0} d_j + \sum_{j\in I_0^c} d_j = 0,$$

where $I_0^c = \{j : g_j(V_0, W_0) \neq 0\}$. Since $h(V_0, W_0) = 0$, all coefficients cannot be zero, the set I_0^c is non-empty. Therefore, we can find d such that (MFCQ) holds. Now the existence of the optimal solution follows from Theorem 1, which concludes the proof.

Appendix A.3. Proof of Proposition 3

Note that the convex combination $X_0\tilde{W}$ is a K-dimensional vector, where each scalar element $X_{0k}W^\star$ is a convex combination of predictor $k = 1, \cdots, K$. Suppose $X_{0k}\tilde{W} = X_{1k}$ for some arbitrary k, but not necessarily for other predictors. In this case, it is easy to verify that $\tilde{W}$ remains an optimal solution to the reduced single-dimensional problem using V_k such that the loss-function of the lower-level problem goes to zero. Since the lower-level loss function cannot be improved, we have $\tilde{W} \in \Psi(V_k)$ and the solution is considered feasible for the bilevel problem (6)-(7). Furthermore, if the original solution was bilevel optimal, then also the other solution $(V_k, \tilde{W})$ remains optimal, since the upper-level objective value depends only on $\tilde{W}$. This concludes the proof.

Appendix A.4. Proof of Proposition 4

Given that the assumptions of Theorem 2 are satisfied, the solution set mapping Ψ_{ε_k} of the regularized lower-level problem (8) is upper semi-continuous. That is, for each sequence $\{(V^k, W^k, \varepsilon_k)\}_{k=1}^{\infty}$ with $\lim_{k\to\infty} V^k = \bar{V}$, $\lim_{k\to\infty} \varepsilon_k = 0+$ and $W^k \in \Psi_{\varepsilon_k}(V^k)$ for all k, each accumulation point of the sequence $\{W^k\}_{k=1}^{\infty}$ is an optimal solution to the lower level problem, i.e. the accumulation points belong to $\Psi_0(\bar{V}) = \Psi(\bar{V})$. Then, by continuity of L_V the first assertion follows.

To show the second assertion it is enough to verify that the regularized lower-level problem meets the assumptions of Theorem 3. This is easy to check because the requirement that $Y_0^\top Y_0$ is positive definite means

that $\nabla_{ww} L_V(V,W)$ is positive definite at each $(V,W) \in \mathcal{V} \times \mathcal{W}$, which means that (SSOC) is satisfied at all feasible points. As a result, Theorem 3 implies that the set $\Psi_{\varepsilon_k}(V^k) = \{W^k(V^k)\}$ is a singleton and the optimal solution function $W^k(V^k)$ is uniquely defined and directionally differentiable at each $\varepsilon_k > 0$. The remaining part of the claim follows from the inequality

$$L_W(V^k, W^k(V^k)) \geq \min_{W \in \mathcal{W}} L_W(V^k, W)$$

that holds due to feasibility. As a result, we have that

$$L_V(V^k, W^k(V^k)) \leq \min_W \{L_V(V^k, W) : \ W \in \Psi(V^k)\},$$

which then implies the last assertion for every fixed $V^k \in \mathcal{V}$. This concludes the proof.

Appendix B. Implementation of SCM algorithm

Appendix B.1. Descent algorithm based on Tykhonov regularization

Based on Proposition 4, the original synthetic control problem can be solved by considering a sequence of single-level problems

$$\min_V \{L_{\varepsilon_k}(V) : \ V \in \mathcal{V}\} \text{ for } \varepsilon_k \to 0+, \tag{B.1}$$

where the implicitly defined objective function $L_{\varepsilon_k}(V) = L_V(V, W^\star_{\varepsilon_k}(V))$ is directionally differentiable with respect to V. In the literature on bilevel programming such approach is commonly referred as Tykhonov regularization (Dempe, 2010). This approach is not often available because of the strictness of (SSOC) and (CRCQ) conditions. However, when these criteria are satisfied, they enable the use of algorithms that are essentially similar to gradient descent.

Let $E\Lambda(V,W)$ be the vertex set of lower-level Lagrange multipliers corresponding to point (V,W),

$$\Lambda(V,W) = \{(\lambda,\mu) : \ \lambda \geq 0, \lambda^\top g(V,W) = 0, \nabla_w \mathcal{L}(V,W,\lambda,\mu) = 0\},$$

where $\mathcal{L}(V,W,\lambda,\mu) = L^\varepsilon_W(V,W) + \lambda^\top g(V,W) + \mu^\top h(V,W)$ denotes the Lagrangian function for the regularized lower level problem. Under (MFCQ) condition, the set $\Lambda(V,W)$ is known to be a non-empty, convex and compact polyhedron. Here functions $g(V,W)$ and $h(V,W)$ denote the vector of lower level inequality constraints and the equality constraint, respectively.

For a fixed vertex $(\lambda^0,\mu^0) \in \Lambda(V^0,W^0)$ at a point $(V,W) = (V^0,W^0)$, we write $\mathcal{I}(\lambda^0)$ to denote the family of all index sets

$$I \subset I(V^0,W^0) := \{i : g_i(V^0,W^0) = 0\}$$

that satisfy the following two conditions:

(C1) There is $(\lambda,\mu) \in E\Lambda(V^0,W^0)$ such that $J(\lambda) := \{i : \lambda_i > 0\} \subset I \subset I(V^0,W^0)$.

(C2) The gradients $\{\nabla_w g_i(V^0, W^0) : i \in I\} \cup \{\nabla_w h(V^0, W^0)\}$ are linearly independent.

Following Dempe (2010), the solution algorithm, which is essentially an adaptation of gradient descent, can be outlined as follows: Tykhonov-Descent: **Input:** Synthetic control problem (6)-(7). **Output:** A Bouligand stationary solution. *Step 1:* Select $V^0 \in \mathcal{V}$, set $k = 0$, choose $\epsilon, \delta \in (0, 1)$, a small $\epsilon' > 0$, a sufficiently small $\kappa > 0$, and a $w < 0$. *Step 2a:* Choose (K^k, λ^k, μ^k) with

$$(\lambda^k, \mu^k) \in E\Lambda(W^\star_{\varepsilon_k}(V^k), V^k) \text{ and } K^k \in \mathcal{I}(\lambda^k)$$

Compute an optimal solution $(d^k, r^k, \gamma^k, \eta^k, s^k)$ for problem (B.2). If $s^k < w$ then go to Step 3. If $s^k \geq w$ and not all possible samples (λ^k, μ^k, K^k) are tried, then continue with Step 2a. If all (λ^k, μ^k, K^k) have been tried, set $w = w/2$. If $|w| < \epsilon'$, go to Step 2b, otherwise continue with Step 2a. *Step 2b:* Choose (K^k, λ^k, μ^k) satisfying

$$K^k \subset I_\kappa(W^\star_{\varepsilon_k}(V^k), V^k) \text{ and (C2) as well as}$$

$$(\lambda^k, \mu^k) \in \underset{(\lambda,\mu)}{\operatorname{argmin}} \{\|\nabla_w \mathcal{L}(W^\star_{\varepsilon_k}(V^k), V^k, \lambda, \mu\|^2 : \ \lambda_j = 0, \ j \notin K^k\}.$$

Here $I_\kappa = \{j : \ -\kappa \leq g_j(V, W) \leq 0\}$ denotes the set of κ-active lower-level inequalities. Compute an optimal solution $(d^k, r^k, \gamma^k, \eta^k, s^k)$ for problem (B.2). If $s^k < w$, go to Step 3. If $s^k \geq w$ and not all (λ^k, μ^k, K^k) have been tried, continue with Step 2b. If all K^k have been tried, then set $w = w/2$. If $|w| < \epsilon'$, then stop. *Step 3:* Choose a largest step-size $t^k \in \{\delta, \delta^2, \delta^3, \delta^4, \dots\}$ such that

$$L_{\varepsilon_k}(V^k + t^k r^k) \leq L_{\varepsilon_k}(V^k) + \epsilon t^k s^k, \ G(V^k + t^k r^k) \leq 0.$$

If $t^k < \epsilon'$, then drop the actual set K^k and continue searching for a new set K^k in Step 2a or 2b. *Step 4:* Set $V^{k+1} = V^k + t^k r^k$, $k = k + 1$. *Step 5:* If a stopping criterion is satisfied, i.e. ε_k is sufficiently small, then stop. Otherwise, set $\varepsilon_{k+1} = \delta \varepsilon_k$ and compute $W^\star_{\varepsilon_{k+1}}(V^{k+1})$ and go to step 2. The directional derivative in Step 2 can be computed using quadratic programming based on Theorem 3 by Dempe (2010). Let $K^k \in \mathcal{I}(\lambda^k)$ be some index set and $\nu^k = (\lambda^k, \mu^k) \in E\Lambda(z^k)$ be a vertex, where $z^k = (V^k, W^k)$. Then the descent direction r^k is obtained as part of a solution to the following problem:

$$\min_{d,r,\gamma,\eta,s} \ s \tag{B.2}$$

$$\begin{aligned}
\text{s.t.} \quad & L'_{\varepsilon_k}(V^k; r^k) := \nabla_w L_V(z^k) d + \nabla_v F(z^k) r \leq s \\
& \nabla_v G_i(V^k) r \leq -G_i(V^k) + s, \ i = 1, \dots, K + 2 \\
& \nabla^2_{ww} \mathcal{L}(z^k, \nu^k) r + \nabla^\top_w g(z^k) \gamma + \nabla^\top_w h(z^k) \eta = 0 \\
& \nabla_w g_i(z^k) d + \nabla_v g_i(z^k) r \begin{cases} = 0, & i \in K^k \\ \leq -g_i(z^k) + s, & i \notin K^k \end{cases} \\
& \nabla_w h(z^k) d + \nabla_y h(z^k) r = 0 \\
& \lambda_i + \gamma_i + s \geq 0, \ i \in K^k, \ \gamma_i = 0, \ i \notin K^k, \ \|r\| \leq 1.
\end{aligned}$$

When the problem has a feasible solution $(d^k, r^k, \gamma^k, \eta^k, s^k)$ such that the objective value is negative, $s^k < 0$, for some index set K^k and vertex ν^k, then the point (V^k, W^k) is not locally optimal. This means that there exists a direction r^k for which the directional derivative of L_{ε_k} is negative at V^k.

When parametrizing the algorithm, it is useful to choose the value for ϵ' to be small enough to ensure that Step 3 terminates only if a set K^k is selected in Step 2b such that the problem (B.2) has a negative optimal value. It is also noteworthy that the Step 2b should be considered only when the value of $L_{\varepsilon_k}(V^k; r^k)$ is sufficiently small and even then only for small κ. Otherwise there is a risk of increasing numerical effort substantially. For discussion on the convergence of this kind of algorithm to a Bouligand stationary point, we refer to Dempe & Schmidt (1996).

Appendix B.2. Algorithm based on KKT approximations

The use of KKT reformulations has been a common practice when solving bilevel problems. Unfortunately, this has turned out to be far more difficult than anticipated. Quite commonly, the local optimal solutions obtained by solving KKT reformulated problems do not correspond to the local optimal solutions of the original bilevel problem. While the KKT reformulations are equivalent to the original problem in terms of global optimal solutions, the equivalence is lost when numerical algorithms need to be used. Since KKT reformulations typically lead to a nonconvex optimization problem, the solution algorithms tend to find only stationary or local optimal solutions, which may not correspond to the solutions of the original problem.

Fortunately, there are still some good news left when it comes to the use of KKT conditions in practice. In their recent paper, Dempe & Franke (2019) suggest a numerically stable approach for handling optimistic bilevel problems with convex lower level problem. The idea is based on a clever approximation of the KKT transformation which enables us to use general solution algorithms for non-convex optimization problems to approximate the local optimal solution of the original bilevel optimization problem.

Now instead of considering the classical KKT reformulation of the problem, the idea developed in the paper by Dempe & Franke (2019) is to construct perturbed problems that approximate the original formulation. Let $\mathcal{L}$ denote the Lagrangian corresponding to the lower level problem,

$$\mathcal{L}_\varepsilon(V, W, \lambda) = L^\varepsilon_W(V, W) + \lambda^\top g(V, W).$$

We then solve a sequence of perturbed problems

$$\begin{aligned} \min_{V,W,\lambda} \; & L_V(V, W) \\ G(V) &\leq 0 \\ ||\nabla_w \mathcal{L}_\varepsilon(V, W, \lambda)|| &\leq e_1 \\ g(V, W) &\leq 0 \\ \lambda &\geq 0, \\ -\lambda_i g_i(V, W) &\leq e_2, \; i = 1, \ldots, J + 2, \end{aligned} \tag{B.3}$$

for $(e_1, e_2) \to 0+$ and $\varepsilon \to 0+$. Here, the norm $||\cdot||$ can be chosen to be for instance the Chebyshev norm $||a||_\infty = \max_{i=1,\ldots,n} |a_i|$ or the usual Euclidean norm $||a||_2 = \sqrt{\sum_{i=1}^n a_i^2}$. The function G is defined such that it matches the definition of set $\mathcal{V} = \{V : G(V) \leq 0\}$ in (5). Similarly, g represents the lower level constraints such that $\mathcal{W} = \{W : \ g(V, W) \leq 0\}$ corresponds to (4).

Earlier, a similar approach of using sequence of perturbed problems to solve bilevel problems has also been considered by Mersha & Dempe (2011), who suggested a specifically tailored algorithm to solve the problem. Later, however, Dempe & Franke (2019) have shown that the assumptions made earlier have been too restrictive and the sequence of perturbed problems can actually be solved by an arbitrary algorithm.

Appendix C. R code

In this appendix we provide the essential R code to help the reader to reproduce our empirical results or adapt the code for their own applications. The latest updates to the code and the technical documentation are available at GitHub: https://github.com/Xun90/SCM-Debug.git. We assume the reader is familiar with the *Synth* R package (see Abadie et al., 2011 for an introduction), and suggest the use of the dataprep() function provided in *Synth* to pre-process the data.

Step 1: Load necessary R packages.

```
library("Synth")      #Synth package
#The two QP solvers used by "Synth" are employed here for a direct comparison with "Synth"
library(kernlab)      #QP solver 1: ipop
library(LowRankQP)    #QP solver 2: LowRankQP, whose results are reported in this study
library(lpSolve)      #LP solver
library(matrixcalc)   #for matrix calculations
```

Step 2: Re-examine the *Synth* results for the California tobacco control application with 1,000 random reorderings of predictors.

```
##loop on 1000 random orders
lossV <- matrix(0, 1000, 1)
lossW <- matrix(0, 1000, 1)
W <- matrix(0, 38, 1000)
V <- matrix(0, 7, 1000)
C <- matrix(0, 7, 1000)
set.seed(42)
for (i in 1:1000){
  row <- sample(nrow(X0))
  C[,i] <- row
  dataprep.out$X0 <- X0[row,]
  dataprep.out$X1 <- as.matrix(X1[row,])
  synth.out <- synth(data.prep.obj = dataprep.out, method = "BFGS")
  lossV[i,] <- synth.out$loss.v
  lossW[i,] <- synth.out$loss.w
  W[,i] <- synth.out$solution.w
```

```
  sorted <- cbind(row, t(synth.out$solution.v))
  sorted <- sorted[order(sorted[,"row"]),]
  V[,i] <- sorted[,2]}
```

Step 3: Re-examine the *Synth* results for the California tobacco control application with 1,000 random reorderings of donors.

```
##loop on 1000 random orders
lossV <- matrix(0, 1000, 1)
lossW <- matrix(0, 1000, 1)
W <- matrix(0, 38, 1000)
V <- matrix(0, 7, 1000)
C <- matrix(0, 38, 1000)
set.seed(42)
for (i in 1:1000){
  column <- sample(ncol(X0))
  C[,i] <- column
  dataprep.out$X0 <- X0[,column]
  dataprep.out$Z0 <- Y0pre[,column]
  synth.out <- synth(data.prep.obj = dataprep.out, method = "BFGS")
  lossV[i,] <- synth.out$loss.v
  lossW[i,] <- synth.out$loss.w
  V[,i] <- t(synth.out$solution.v)
  sorted <- cbind(column, synth.out$solution.w)
  sorted <- sorted[order(sorted[,"column"]),]
  W[,i] <- sorted[,2]}
```

Step 4: Implement the iterative algorithm proposed by Malo et al. (2020) to check for the feasibility of the unconstrained optimum and the possibility of corner solutions.

```
scm.corner <- function(Y1pre,Y0pre,X1,X0){
  ##step1
  Tpre <- dim(Y0pre)[1]
  nDonors <- dim(Y0pre)[2]
  #QP setup
  c1 <- -t(Y0pre) %*% Y1pre
  H1 <- t(Y0pre) %*% Y0pre
  A1 <- matrix(rep(1,nDonors), ncol = nDonors)
  b1 <- 1
  r1 <- 0
  l1 <- matrix(rep(0,nDonors), nrow = nDonors)
  u1 <- matrix(rep(1,nDonors), nrow = nDonors)
  #run QP
  step1_ipop <- ipop(c = c1, H = H1, A = A1, b = b1, l = l1, u = u1, r = r1,
                     margin = 0.0005, maxiter = 1000, sigf = 7, bound = 10) #QP_Solver1
  step1_lowr <- LowRankQP(Vmat = H1, dvec = c1, Amat = A1, bvec = b1, uvec = u1,
                          method = "LU") #QP_Solver2
  W_ipop <- matrix(step1_ipop@primal, nrow = nDonors)
  W_lowr <- step1_lowr$alpha
  L1_ipop <- (t(Y1pre) %*% Y1pre)/Tpre + 2/Tpre * (t(c1) %*% W_ipop
```

```
                                                    + 0.5 * t(W_ipop) %*% H1 %*% W_ipop)
  L1_lowr <- (t(Y1pre) %*% Y1pre)/Tpre + 2/Tpre * (t(c1) %*% W_lowr
                                                    + 0.5 * t(W_lowr) %*% H1 %*% W_lowr)
  ##step2
  #normalize X - Synth
  nvarsV <- dim(X0)[1]
  big.dataframe <- cbind(X0, X1)
  divisor <- sqrt(apply(big.dataframe, 1, var))
  scaled.matrix <- t(t(big.dataframe) %*% ( 1/(divisor)
                                            * diag(rep(dim(big.dataframe)[1], 1)) ))
  X0.scaled <- scaled.matrix[,c(1:(dim(X0)[2]))]
  if(is.vector(X0.scaled)==TRUE)
  {X0.scaled <- t(as.matrix(X0.scaled))}
  X1.scaled <- scaled.matrix[,dim(scaled.matrix)[2]]
  #LP setup
  f.obj_ipop <- (X1.scaled - X0.scaled %*% W_ipop)^2
  f.obj_lowr <- (X1.scaled - X0.scaled %*% W_lowr)^2
  f.con <- rbind(rep(1,nvarsV), diag(x = 1, nrow = nvarsV))
  f.dir <- c("=", rep(">=",nvarsV))
  f.rhs <- c(1, rep(0,nvarsV))
  #run LP
  step2_ipop <- lp ("min", f.obj_ipop, f.con, f.dir, f.rhs)
  step2_lowr <- lp ("min", f.obj_lowr, f.con, f.dir, f.rhs)
  V_ipop <- step2_ipop$solution
  V_lowr <- step2_lowr$solution
  L2_ipop <- step2_ipop$objval
  L2_lowr <- step2_lowr$objval

  scm.corner.out <- list(W = cbind(W_ipop,W_lowr), V = cbind(V_ipop,V_lowr),
                         Lv = c(L1_ipop,L1_lowr), Lw = c(L2_ipop,L2_lowr))
  return(scm.corner.out)}
```

Step 5: Implement the two-step procedure described in Section 5.1. This implementation is currently a hybrid of Section 5.1 and Malo et al. (2020). Since there are currently no reliable solvers in R for the second-stage convex programming problem, we solve the non-Archimedean problem (8) iteratively, decreasing ε towards zero until the objective function reaches the optimal solution of the first-stage QP problem.

```
two.step.iterative <- function(Y1pre,Y0pre,X1.scaled,X0.scaled,SV){
  #SV - predictor weights defined by the user
  ##Solve non-Archimedean problem (8)
  Tpre <- dim(Y0pre)[1]
  nDonors <- dim(Y0pre)[2]
  #QP setup
  A <- matrix(rep(1,nDonors), ncol = nDonors)
  b <- 1
  r <- 0
  l <- matrix(rep(0,nDonors), nrow = nDonors)
  u <- matrix(rep(1,nDonors), nrow = nDonors)
  #Loop on 10 epsilon values (0.1^1 ... 0.1^10) to find the best performer
```

```
L_upper = matrix(0, 10, 2)
L_lower = matrix(0, 10, 2)
W_ipop = matrix(0, nDonors, 10)
W_lowr = matrix(0, nDonors, 10)
for (i in 1:10){
  eps <- 0.1^(i) #epsilon - penalty term
  c <- (-t(X0.scaled) %*% diag(SV1) %*% X1.scaled) - eps * t(Y0pre) %*% Y1pre
  H <- t(X0.scaled) %*% diag(SV1) %*% X0.scaled + eps * t(Y0pre) %*% Y0pre
  #run QP
  QP_ipop <- ipop(c = c, H = H, A = A, b = b, l = l, u = u, r = r,
                  margin = 0.0005, maxiter = 1000, sigf = 7, bound = 10) #QP_Solver1
  QP_lowr <- LowRankQP(Vmat = H, dvec = c, Amat = A, bvec = b, uvec = u,
                       method = "LU") #QP_Solver2
  W_ipop[,i] <- matrix(QP_ipop@primal, nrow = nDonors)
  W_lowr[,i] <- QP_lowr$alpha
  L_upper[i,1] <- 1/Tpre * t(Y1pre - Y0pre %*% W_ipop[,i]) %*%
    (Y1pre - Y0pre %*% W_ipop[,i])
  L_upper[i,2] <- 1/Tpre * t(Y1pre - Y0pre %*% W_lowr[,i]) %*%
    (Y1pre - Y0pre %*% W_lowr[,i])
  L_lower[i,1] <- t(X1.scaled - X0.scaled %*% W_ipop[,i]) %*% diag(SV1) %*%
    (X1.scaled - X0.scaled %*% W_ipop[,i])
  L_lower[i,2] <- t(X1.scaled - X0.scaled %*% W_lowr[,i]) %*% diag(SV1) %*%
    (X1.scaled - X0.scaled %*% W_lowr[,i])}
##Use the first step of the two-step procedure in Section 4.1 to determine epsilon
c1 <- (-t(X0.scaled) %*% diag(SV) %*% X1.scaled)
H1 <- t(X0.scaled) %*% diag(SV) %*% X0.scaled
#run QP
QP_ipop1 <- ipop(c = c1, H = H1, A = A, b = b, l = l, u = u, r = r,
                 margin = 0.0005, maxiter = 1000, sigf = 7, bound = 10) #QP_Solver1
QP_lowr1 <- LowRankQP(Vmat = H1, dvec = c1, Amat = A, bvec = b, uvec = u,
                      method = "LU") #QP_Solver2
W_ipop1 <- matrix(QP_ipop1@primal, nrow = nDonors)
W_lowr1 <- QP_lowr1$alpha
W1 <- cbind(W_ipop1, W_lowr1)
obj_left <- t(X1.scaled) %*% diag(SV) %*% X1.scaled
Lw_ipop1 <- obj_left + 2 * (t(c1) %*% W_ipop1 + 1/2 * t(W_ipop1) %*% H1 %*% W_ipop1)
Lw_lowr1 <- obj_left + 2 * (t(c1) %*% W_lowr1 + 1/2 * t(W_lowr1) %*% H1 %*% W_lowr1)
Lw1 <- c(Lw_ipop1, Lw_lowr1)

two.step.iterative.out <- list(W = cbind(W_ipop,W_lowr), W1 = W1, V = SV,
                       L_upper = L_upper, L_lower = L_lower, Lw1 = Lw1)
return(two.step.iterative.out)}
```

Appendix D. Imputation of missing values

The *Synth* R package contains the original data for the Basque terrorism application. This data set contains incomplete panel data for the predictors across different regions in the pre-treatment period (1960–1969) (see Abadie et al., 2011 for more details). To implement the panel regression approach described in Section 5.2 to

determine predictor weights in the Basque terrorism application, it is necessary to impute the missing values by suitable methods.

For the six sectoral share predictors (i.e., "sec.agriculture", "sec.energy", "sec.industry", "sec.construction", "sec.services.venta", and "sec.services.nonventa"), panel data are available for odd years only (1961, 1963, ..., 1969). We replaced the missing values in even years from 1962 through 1968 with the mean of the data of two adjacent years. We then estimated a linear time trend by regressing the values of 1961–1969, and used the predicted value for the year 1960.

For the four schooling predictors (i.e., "school.illit", "school.prim", "school.med", and "school.high") and the predictor "investment ratio", panel data are available only for the years 1964–1969. Again, we estimated a linear time trend by regressing the values of 1964–1969, and used the predicted values for the years 1960–1963.

Finally, the predictor "population density" was observed only in the year 1969. Since the population density usually changes very slowly, in the absence of better data, we used the observed value of population density in the year 1969 throughout the pre-treatment period 1960–1969.

References

Abadie, A., Diamond, A., Hainmueller, J. (2011). Synth: An R package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, *42*, 1–17.

Bank, B., Guddat, J., Klatte, D., Kummer, B., Tammer, K. (1982). *Non-Linear Parametric Optimization.* Akademie-Verlag, Berlin.

Dempe, S. (2010). *Foundations of Bilevel Programming.* Kluwer Academic Publishers.

Dempe, S., Franke, S. (2019). Solution of bilevel optimization problems using the KKT approach. *Optimization*, *68*, 1471–1489.

Dempe, S., Schmidt, H. (1996). On an algorithm solving two-level programming problems with nonunique lower level solutions. *Computational Optimization and Applications*, *6*, 227–249.

Malo, P., Eskelinen, J., Zhou, X., Kuosmanen, T. (2020). Computing Synthetic Controls Using Bilevel Optimization. *MPRA Paper 104085.*

Mersha, A. G., Dempe, S. (2011). Direct search algorithm for bilevel programming problems. *Computational Optimization and Applications*, *49*, 1–15.