Computing Synthetic Controls Using Bilevel Optimization

Malo, Pekka and Eskelinen, Juha and Zhou, Xun and Kuosmanen, Timo

Aalto University School of Business, Technical University of Munich

November 2020

Computing Synthetic Controls Using Bilevel Optimization

Pekka Malo[1], Juha Eskelinen[1], Xun Zhou[1,2], Timo Kuosmanen[1]

[1] *Aalto University School of Business, Finland*
[2] *Technical University of Munich, Germany*

November 2020

Abstract

The synthetic control method (SCM) is a major innovation in the estimation of causal effects of policy interventions and programs in a comparative case study setting. In this paper, we demonstrate that the data-driven approach to SCM requires solving a bilevel optimization problem. We show how the SCM problem can be solved using iterative algorithms based on Tykhonov descent or KKT approximations.

Keywords: Causal effects; Comparative case studies; Policy impact assessment

JEL Codes: C54; C61; C71

1 Introduction

The synthetic control method (SCM) has emerged as a widely used tool for estimating causal effects of policy interventions and programs in a comparative case study setting. Abadie and Gardeazabal (2003) originally introduced this innovative approach to examine the economic impacts of terrorism in the Basque Country. Abadie et al. (2010) further developed 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, for example, Abadie et al. (2015), Acemoglu et al. (2016), Cavallo et al. (2013), Gobillon and Magnac (2016), and Kleven et al. (2013). Athey and Imbens (2017) refer to SCM as "*arguably the most important innovation in the policy evaluation literature in the last 15 years.*"

Despite the popularity of SCM, rather surprisingly, no explicit mathematical formulation of how the predictor weights and the donor weights are jointly optimized has been presented in the literature. Several recent studies note that the synthetic controls produced by standard algorithms may be numerically unstable and suboptimal (e.g., Becker and Klößner, 2017, 2018; Becker et al., 2018; Klößner et al., 2018).

The purpose of the present paper is to provide a detailed examination of the optimization problem that needs to be solved to compute the synthetic control weights. Unfortunately, the explicit formulation of the SCM problem reveals that computing the synthetic controls requires solving a NP-hard problem, referred to as a bilevel optimization problem (e.g., Hansen et al., 1992; Vicente et al., 1994). That is, computing synthetic controls is a much harder task than any previous SCM studies recognize. This observation helps to explain the numerical instability reported by Klößner et al. (2018), among others. To address this problem, we develop an iterative computational procedure based on the method known as Tykhonov regularization, which is guaranteed to converge to the optimal solution.

The rest of the paper is organized as follows. Section 2 introduces the 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 presents our concluding remarks and discusses some avenues for future research. Proofs of Theorems are presented in the Appendix.

2 Synthetic control method

2.1 Preliminaries

Following the usual notation (e.g., Abadie, 2019), suppose we observe units $j = 1, \dots, J+1$, 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 elements of 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 and 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, 2019). 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.[1] 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 (2019) 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*

[1] 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. (2015) demonstrate that including all pre-treatment outcomes as predictors is not a good idea because the predictors become completely redundant in that case.

particular data-set there are not ex-ante guarantees on the size of the difference $X_1 - X_0W$. 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 and 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,[2] and are usually normalized to sum to unity.[3] That is

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

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

Both Abadie and Gardeazabal (2003) and Abadie et al. (2010) suggest that weights V could be subjectively determined. However, virtually all 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 and 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. By comparing with the original SCM articles, it is easy to verify that the optimal weights $V^\star$, $W^\star$ must be obtained as an optimal solution to the following optimistic bilevel optimization problem

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

[2] 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 and Klößner (2018) set a lower bound $V_{kk} \geq 0.00000001$, which is so low that it has no practical meaning.

[3] 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.

$$\text{s.t.} \quad W(V) \in \Psi(V) := \underset{W \in \mathcal{W}}{\operatorname{argmin}}\, L_W(V, W) = \|X_1 - X_0 W\|_V^2, \qquad (7)$$
$$V \in \mathcal{V},$$

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 and 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) referred 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 and Klößner (2017) and (Klößner et al., 2018), among others. The general-purpose algorithms are simply ill-equipped for the task at hand. If the weights W, V are arbitrarily determined by an algorithm that fails to converge to a feasible and unique global optimum, then all attractive theoretical properties of the estimator fly out of the window.

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. It is noteworthy that surprisingly many of the SCM problems encountered in practice admit either unconstrained solution or a corner solution. 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

$$\min_{W \in \mathcal{W}} L_W^{\varepsilon}(V, W) = \frac{1}{K} \|X_1 - X_0 W\|_{V^\star}^2 + \varepsilon \|Y_1^{\text{pre}} - Y_0^{\text{pre}} W(V)\|^2 \tag{8}$$

where $\varepsilon > 0$ denotes an infinitesimally small non-Archimedean scalar.[4] 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.

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

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*

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

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 $\|X_1 - X_0W\|^2_{V^\star}$ need not be unique. This is particularly relevant when there exist W that satisfy $\|X_1 - X_0W\|^2_{V^\star} = 0$. 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 and 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. (2015) 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, $X_1 = X_0W$, 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 $X_0W = X_1$ 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., Ben-Michael et al., 2018; Doudchenko and Imbens, 2017; Kaul et al., 2015). 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 $W_{\varepsilon_k}^{\star}(V)$, 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 and 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

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

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

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

2. *for each* ε_k*, the unique optimal solution to the regularized lower-level problem* (11)*, denoted by* $W^{\star}_{\varepsilon_k}(V) \in \Psi_{\varepsilon_k}(V)$*, 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 $L_{\varepsilon_k}(V) = L_V(V, W^{\star}_{\varepsilon_k}(V))$ 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 and 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. Indeed, the rationale of using convex combinations is rather similar to the benefit of the doubt weighting (e.g., Cherchye et al., 2007) where zero weights are similarly an issue. 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)).

While we appreciate the data-driven approach to determine the predictor weights, due to the inherent computational difficulties of the joint optimization of weights V and W and the fact that the optimum is likely a corner solution, our proposal is to resort to some alternative data-driven approaches to determine the predictor weights V, and subsequently optimize the donor weights W conditional on V.

4 Conclusions

Synthetic control method has proved a useful approach for estimating causal treatment effects in a comparative case study setting, as a large number of published applications clearly demonstrate. Unfortunately, the standard computational algorithms for jointly solving the donor weights and the predictor weights have proved numerically unstable. In this paper we present the first explicit formulation of the SCM problem as an optimistic bilevel optimization problem. A rigorous mathematical formulation demonstrates that the SCM problem is far from trivial from the computational perspective: the SCM problem is generally NP-hard, and clearly far beyond the scope of the algorithms currently in use.

To address the lack of convergence, the second contribution of our paper was to propose an iterative procedure, which is guaranteed to converge to the optimum under relatively mild assumptions. While this shows that there exists a theoretically valid approach for solving the SCM problem, the optimal solutions to the original SCM formulation are still typically obtained as corner solutions, where most of the predictors have a zero weight. Thus, in practice, it is rarely necessary to apply Tykhonov descent or KKT approximations to find the optimal solution. Instead, an optimal solution is found already during the early stages of the iterative procedure.

The computational difficulties of the original SCM formulation do not change the fact that synthetic controls are very appealing as a concept. We do appreciate the data-driven approach to determine the weights, however, it is important to ensure that the synthetic controls are optimal, and not merely artifacts of a suboptimal algorithm.

Our findings open several avenues for future research, both empirical and methodological studies. From the empirical point of view, it would be interesting to apply the proposed algorithm to replicate published SCM studies in order to examine whether and to what extent the use of suboptimal weights affects the qualitative conclusions. Becker and 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.

From the methodological point of view, the joint optimization of the predictor weights and the donor weights would deserve further attention. In particular, 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. One possibility would be to resort to stepwise optimization of the predictor weights and donor weights, such that the predictor weights are first determined based on some other criteria (e.g., regression analysis) and subsequently the donor weights are optimized taking the predictor weights as given. We leave this as an interesting avenue for future research.

Finally, we hope that the insights of our paper might contribute to further integration of SCM with other estimation approaches such as the difference-in-difference, panel data regression and machine learning; several recent studies such as Abadie and L'Hour (2020), Amjad et al. (2018), Arkhangelsky et al. (2018), Ben-Michael et al. (2018), Doudchenko and Imbens (2017), and Xu (2017) have made impressive progress in this direction.

Appendix A Proofs of theorems

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

$$\min_y \{f(x,y) : g(x,y) \leq 0, h(x,y) = 0\}, \tag{13}$$

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 $h_j(x,\cdot)$, $j=1,\dots,q$, 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

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

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

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

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*

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

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

and the gradients of the equality constraints $\{\nabla_y h_j(x^0, y^0) : j = 1, \dots, q\}$ *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:

$$\min_x \{\varphi_0(x) : x \in X\}, \tag{14}$$

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 (14) 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 (13) 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 (13) 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 (13) 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*

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

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

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

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.

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.

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

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

$$\min_{V} \{L_{\varepsilon_k}(V) : \; V \in \mathcal{V}\} \text{ for } \varepsilon_k \to 0+, \tag{15}$$

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:

procedure 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 (16).

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 (16).

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.

end procedure

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:

$$\min_{d,r,\gamma,\eta,s} \; s \tag{16}$$

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

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 (16) 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 and Schmidt (1996).

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 and 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 and Franke (2019) is to construct perturbed problems that approximate the original formulation. Let $\mathcal{L}$ denote the Lagrangian corresponding to the lower level problem,

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

We then solve a sequence of perturbed problems

$$\begin{aligned} \min_{V,W,\lambda} \; L_V(V, W) & & \\ G(V) &\leq 0 & \\ ||\nabla_w \mathcal{L}_\varepsilon(V, W, \lambda)|| &\leq e_1 & \\ g(V, W) &\leq 0 & \\ \lambda &\geq 0, & \\ -\lambda_i g_i(V, W) &\leq e_2, \; i = 1, \ldots, J+2, & \end{aligned} \tag{17}$$

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 and Dempe (2011), who suggested a specifically tailored algorithm to solve the problem. Later, however, Dempe and 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.

References

Abadie, A. (2019). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*.

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*(490), 493–505.

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, *59*(2), 495–510.

Abadie, A., & Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, *93*(1), 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*(2), 368–391.

Amjad, M., Shah, D., & Shen, D. (2018). Robust synthetic control. *Journal of Machine Learning Research*, *19*(1), 802–852.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2018). Synthetic Difference in Differences. *arXiv preprint arXiv:1812.09970*.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

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

Becker, M., & Klößner, S. (2017). Estimating the economic costs of organized crime by synthetic control methods. *Journal of Applied Econometrics*, *32*(7), 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*(1), 603–609.

Ben-Michael, E., Feller, A., & Rothstein, J. (2018). The Augmented Synthetic Control Method. *arXiv preprint arXiv:1811.04170*.

Cavallo, E., Galiani, S., Noy, I., & Pantano, J. (2013). Catastrophic natural disasters and economic growth. *Review of Economics and Statistics*, *95*(5), 1549–1561.

Charnes, A. (1952). Optimality and Degeneracy in Linear Programming. *Econometrica*, *20*(2), 160–170.

Cherchye, L., Knox Lovell, C. A., Moesen, W., & Van Puyenbroeck, T. (2007). One market, one number? A composite indicator assessment of EU internal market dynamics. *European Economic Review*, *51*(3), 749–779.

Dempe, S., & Franke, S. (2019). Solution of bilevel optimization problems using the KKT approach. *Optimization*, *68*(8), 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*(3), 227–249.

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

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

Gobillon, L., & Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, *98*(3), 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*(5), 1194–1217.

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

Kleven, H. J., Landais, C., & Saez, E. (2013). Taxation and international migration of superstars: Evidence from the European football market. *American Economic Review*, *103*(5), 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*(1), 11.

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

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

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*(2), 379–399.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, *25*(1), 57–76.