Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Combining Matching and Synthetic Control to Tradeoff Biases From Extrapolation and Interpolation

Maxwell Kellogg, Magne Mogstad, Guillaume A. Pouliot & Alexander Torgovitsky

To cite this article: Maxwell Kellogg, Magne Mogstad, Guillaume A. Pouliot & Alexander Torgovitsky (2021) Combining Matching and Synthetic Control to Tradeoff Biases From Extrapolation and Interpolation, Journal of the American Statistical Association, 116:536, 1804-1816, DOI: 10.1080/01621459.2021.1979562

To link to this article: https://doi.org/10.1080/01621459.2021.1979562

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 536, 1804–1816: Theory and Methods Special Section on Synthetic Control Methods
https://doi.org/10.1080/01621459.2021.1979562

Combining Matching and Synthetic Control to Tradeoff Biases From Extrapolation and Interpolation

Maxwell Kellogg[a], Magne Mogstad[b], Guillaume A. Pouliot[c], and Alexander Torgovitsky[d]

[a]National Bureau of Economic Research (NBER); [b]Department of Economics, University of Chicago, Statistics Norway, NBER; [c]Harris School of Public Policy, University of Chicago; [d]Department of Economics, University of Chicago

ABSTRACT

The synthetic control (SC) method is widely used in comparative case studies to adjust for differences in pretreatment characteristics. SC limits extrapolation bias at the potential expense of interpolation bias, whereas traditional matching estimators have the opposite properties. This complementarity motives us to propose a matching and synthetic control (or MASC) estimator as a model averaging estimator that combines the standard SC and matching estimators. We show how to use a rolling-origin cross-validation procedure to train the MASC to resolve tradeoffs between interpolation and extrapolation bias. We use a series of empirically based placebo and Monte Carlo simulations to shed light on when the SC, matching, MASC and penalized SC estimators do (and do not) perform well. Then, we apply these estimators to examine the economic costs of conflicts in the context of Spain.

ARTICLE HISTORY
Received December 2019
Accepted August 2021

KEYWORDS
Causal inference; Comparative case studies; Cross-validation; Forecasting; Model averaging; Program evaluation; Synthetic control

1. Introduction

Estimating the causal effect of an intervention (treatment) is a common task across the social sciences. Longitudinal approaches based on difference-in-differences have long been used for this task. However, the credibility of these methods can be strained when the pretreatment trends or characteristics of the untreated units differ significantly from those of the treated units. This concern can be particularly salient in comparative case studies with units that are large aggregates, such as countries or states. For these applications, the synthetic control (SC) method of Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010, 2015) provides an alluring alternative.

The motivation of the SC method is to limit the extrapolation bias that can occur when units with different pre-treatment characteristics are combined using a traditional adjustment, such as a linear regression. However, the SC estimator *interpolates* by using a convex weighted average of the untreated units to create a synthetic untreated unit with pre-treatment characteristics similar to those of the treated unit. As observed by Abadie, Diamond, and Hainmueller (2010, pp. 495–496), this makes the SC estimator susceptible to interpolation bias. In Section 2.2, we formalize this observation by showing that SC will avoid interpolation bias only if the conditional mean of the outcome is linear in pretreatment characteristics.

As Abadie and L'Hour (2020) observed, the SC estimator belongs to a class of weighting estimators with weights based on pre-treatment characteristics. Many other commonly used estimators in this class, such as nearest neighbor matching, suffer from the opposite drawback of potentially extrapolating too much when suitable untreated units are unavailable. That is, SC controls extrapolation bias while being susceptible to interpolation bias, whereas matching has the opposite properties. This complementarity suggests that an estimator that adaptively combines the SC and matching estimators may be particularly attractive.

In Section 2.5, we propose the matching and synthetic control (or MASC) estimator as a model averaging estimator that combines the standard SC and matching estimators. We show how averaging these two purposefully chosen estimators can defend against the weaknesses of both while preserving their strengths. In Section 3, we show how to choose the weight assigned to each estimator in the MASC through cross-validation, as in Wolpert (1992), Breiman (1996), and Hansen and Racine (2012). Our cross-validation criterion uses an evaluation concept referred to as rolling-origin recalibration in the forecasting literature (e.g., Tashman 2000). One attractive feature of the MASC estimator is that its cross-validated weight can be solved for in closed-form, making it only marginally more difficult to implement than the usual SC estimator. An R package for implementing the MASC is available at https://github.com/maxkllgg/masc.

In Sections 4 and 5, we provide evidence that the MASC estimator performs well in practice. In Section 4, we conduct a placebo study using the data on Spanish terrorism analyzed by Abadie and Gardeazabal (2003). This allows us to evaluate the performance of the matching, SC, penalized SC (Abadie and L'Hour 2020), and MASC estimators by how well they predict a zero treatment effect for untreated units. We find evidence that MASC has lower mean-squared prediction error than the other three alternatives in this application because it is able to

CONTACT Alexander Torgovitsky atorgovitsky@gmail.com Department of Economics, University of Chicago
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

adapt to cases where either SC or matching would do well. Then, in Section 5, we use the same data to re-estimate the effect of terrorism on the GDP of the Basque Country.

Our article is related to a growing literature on SC (see Abadie 2020, for a recent survey). The closest work to ours is the article by Abadie and L'Hour (2020), who propose the penalized SC estimator. The penalized SC and MASC estimators are different, but related in that both assign weights to untreated units while taking into consideration their distance from the treated unit in terms of pretreatment characteristics. In Section 2.6, we show that the penalized SC estimator is the solution to a constrained version of the problem implicitly solved by MASC. Thus, the MASC is a more flexible estimator than the penalized SC. While this does not mean it will have lower prediction error in practice, our empirical results in Sections 4 and 5 show that, at least for the Spanish data, the penalized SC estimator usually coincides with the standard SC estimator, suggesting that the extra flexibility of MASC can be useful.

Also closely related to our work are the articles by Athey et al. (2019) and Viviano and Bradic (2019), who also consider the benefits of model averaging in the context of comparative case studies. The former authors combine several of the regularized SC and matrix completion estimators developed in Doudchenko and Imbens (2016) and Athey et al. (2018), while the latter authors combine a large number of estimators from the machine learning literature. MASC differs from the estimators in these articles both in details and intent. The purpose of MASC is to directly guard against the types of interpolation biases that can occur with SC, and the extrapolation bias that can occur with matching, by adaptively blending them together. Like Athey et al. (2019), we also find that model averaging tends to work quite well, in concordance with a recurring finding of the economic forecasting literature (see, e.g., Stock and Watson 2004, 2006). A contrast with Athey et al. (2019), Viviano and Bradic (2019), and much of the forecasting literature, is that the estimators we average are purposefully chosen to be complementary. This is exactly the case when data-driven model averaging should be especially beneficial; see, for example, Breiman (1996) or Elliot (2011).

2. Synthetic Control and Matching

2.1. Setup

Let Y_{it} denote a scalar outcome for cross-sectional units i at times $t = 1, \ldots, T$, and let $D_i \in \{0, 1\}$ denote a time-invariant binary treatment group indicator. Units in the treated group become treated at an event date, $t^\star$, so that treatment status in time t is given by $D_{it} \equiv D_i \mathbb{1}[t \geq t^\star]$. Associated with the outcome and treatment are potential outcomes $Y_{it}(0)$ and $Y_{it}(1)$, which are related to the observed outcome via $Y_{it} = D_{it} Y_{it}(1) + (1 - D_{it}) Y_{it}(0)$. We also observe a k-dimensional vector of pretreatment covariates, $\boldsymbol{X}_i$, the components of which will typically include some or all of the pretreatment outcomes (Y_{it} for $t < t^\star$), as well as potentially other predetermined characteristics.

Our goal is to estimate the average treatment on the treated (ATT),

$$\begin{aligned} \text{ATT}_t &\equiv \mathbb{E}[Y_{it}(1) - Y_{it}(0) | D_i = 1] \\ &= \mathbb{E}[Y_{it} | D_i = 1] - \mathbb{E}[Y_{it}(0) | D_i = 1] \end{aligned} \tag{1}$$

where $t \geq t^\star$ is some period after the event date, and $\mathbb{E}$ expectation taken with respect to the underlying joint distribution of $(\{Y_{it}(0), Y_{it}(1), Y_{it}\}_{t=1}^T, D_i, \boldsymbol{X}_i)$, which we view as ex-ante identically distributed across cross-sectional units i. Identifying the ATT is a matter of identifying the mean untreated outcomes for the treated group in the post-period, that is, $\beta_t \equiv \mathbb{E}[Y_{it}(0) | D_i = 1]$. A common approach for this is to assume that all differences between the treated and untreated units can be eliminated by conditioning on $\boldsymbol{X}_i$. The formal assumption consists of the following two parts.

Assumption 1. (Selection on observables). If $\boldsymbol{x}$ is in the supports of both $\boldsymbol{X}_i | D_i = 0$ and $\boldsymbol{X}_i | D_i = 1$, then $\mathbb{E}[Y_{it}(0) | D_i = 1, \boldsymbol{X}_i = \boldsymbol{x}] = \mathbb{E}[Y_{it}(0) | D_i = 0, \boldsymbol{X}_i = \boldsymbol{x}]$ for all $t \geq t^\star$.

Assumption 2. (Overlap). The support of $\boldsymbol{X}_i | D_i = 1$ is contained in the support of $\boldsymbol{X}_i | D_i = 0$.

Assumption 1 is a mean-independence version of what is variously described in the literature as ignorable treatment assignment (Rosenbaum and Rubin 1983), unconfoundedness (Imbens and Rubin 2015), or selection on observables (Barnow, Cain, and Goldberger 1980; Heckman and Robb 1985). Together with Assumption 2, it implies that for posttreatment periods $t \geq t^\star$

$$\begin{aligned} \beta_t = \mathbb{E}\Big[\mathbb{E}[Y_{it} | D_i = 0, \boldsymbol{X}_i] \Big| D_i = 1\Big] \equiv \mathbb{E}\left[\gamma_t(\boldsymbol{X}_i) | D_i = 1\right], \\ \text{where} \quad \gamma_t(\boldsymbol{x}) \equiv \mathbb{E}[Y_{it} | D_i = 0, \boldsymbol{X}_i = \boldsymbol{x}]. \end{aligned} \tag{2}$$

That is, β_t is point identified by the outcomes for the untreated group, conditional on covariates, after re-weighting by the distribution of these covariates in the treated group. For further discussion, see, for example, Heckman, Ichimura, and Todd (1997, 1998), Imbens (2004, 2015), or Imbens and Rubin (2015).

Suppose now that we observe a sample of $n + 1$ realizations $\{(y_{i1}, \ldots, y_{iT}, d_i, \boldsymbol{x}_i)\}_{i=1}^{n+1}$ from the distribution of $(Y_{i1}, \ldots, Y_{iT}, D_i, \boldsymbol{X}_i)$. How exactly the sample is drawn is not relevant to the points we will make. For example, it could be that each unit is drawn once with stochastic variation coming from transitory time series innovations. This possibility is allowed for in the factor and autoregressive models considered by Abadie, Diamond, and Hainmueller (2010, Section 2.2), Abadie (2020, sec. 3.3), and Chernozhukov, Wuthrich, and Zhu (2020), among others.

Our focus in this article is the comparative case study setting considered by Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010, 2015), in which the sample drawn contains only a single treated unit. We label this treated unit as $i = 1$, so that $d_1 = 1$, while $d_i = 0$ for all n remaining units $i \geq 2$. Since we only have a single treated unit, we estimate $\mathbb{E}[Y_{it} | D_i = 1]$ by the realization of Y_{1t} in the post-period. Similarly, since the empirical distribution of $\boldsymbol{X}_i$ given $D_i = 1$ is simply a point mass at $\boldsymbol{x}_1$, we estimate β_t with an estimator of $\gamma_t(\boldsymbol{x}_1)$. Thus, we focus on a class of estimators for the ATT of the form

$$\widehat{\text{ATT}}_t \equiv y_{1t} - \hat{\gamma}_t(\boldsymbol{x}_1), \tag{3}$$

where $\hat{\gamma}_t(\boldsymbol{x}_1)$ is an estimator of $\gamma_t(\boldsymbol{x}_1)$. Note that while Assumptions 1 and 2 justify an interest in estimating $\gamma_t(\boldsymbol{x}_1)$, there is no

presumption that any of the untreated units in the sample have pretreatment covariates $\boldsymbol{x}_1$.

The problem we focus on is how to construct $\hat{\gamma}_t(\boldsymbol{x}_1)$. The estimators we consider are all of the form

$$\hat{\gamma}_t = \sum_{i\geq 2} w_i y_{it} \equiv \boldsymbol{y}_{0t}'\boldsymbol{w} \tag{4}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ are weights applied to the observed outcomes $\boldsymbol{y}_{0t} \in \mathbb{R}^n$ for the untreated units at time t. The weights will always be assumed to live in the $(n-1)$-dimensional simplex

$$\mathcal{S} \equiv \left\{ \boldsymbol{w} \in \mathbb{R}^n : \sum_j w_j = 1 \quad \text{and} \quad w_j \geq 0 \text{ for all } j \right\}, \tag{5}$$

so that $\hat{\gamma}_t$ is a convex weighted average of the outcomes for the untreated units at time t. The question is how to choose the weights, $\boldsymbol{w}$.

2.2. Extrapolation Bias and Interpolation Bias

Consider an estimator of form (4), and write it as

$$\hat{\gamma}_t = \sum_{i\geq 2} w_i \left(\gamma_t(\boldsymbol{x}_i) + u_{it}\right) = \overbrace{\sum_{i\geq 2} w_i \gamma_t(\boldsymbol{x}_i)}^{\text{signal}} + \overbrace{\sum_{i\geq 2} w_i u_{it}}^{\text{noise}}, \tag{6}$$

where $u_{it} \equiv y_{it} - \gamma_t(\boldsymbol{x}_i)$ denotes the deviation between y_{it} and its conditional mean. Our focus in this article is on the signal term in Equation (6) and under what conditions it can replicate $\gamma_t(\boldsymbol{x}_1)$. That is, we are concerned with the bias of estimators of form (4), as captured by the behavior of Equation (6) when $u_{it} = 0$ for all i. We can decompose this bias into two components:

$$\underbrace{\gamma_t(\boldsymbol{x}_1) - \sum_{i\geq 2} w_i \gamma_t(\boldsymbol{x}_i)}_{\equiv \text{Bias}(\boldsymbol{w})} = \underbrace{\left[\gamma_t(\boldsymbol{x}_1) - \gamma_t\left(\sum_{i\geq 2} w_i \boldsymbol{x}_i\right)\right]}_{\equiv \text{ExtBias}(\boldsymbol{w})} + \underbrace{\left[\gamma_t\left(\sum_{i\geq 2} w_i \boldsymbol{x}_i\right) - \sum_{i\geq 2} w_i \gamma_t(\boldsymbol{x}_i)\right]}_{\equiv \text{IntBias}(\boldsymbol{w})},$$

where $\text{ExtBias}(\boldsymbol{w})$ is the *extrapolation bias* and $\text{IntBias}(\boldsymbol{w})$ is the *interpolation bias*.

To see the justification of these terms, consider a simple case in which there are two untreated units ($n = 3$), and $\boldsymbol{x}_i \equiv x_i$ is scalar ($k = 1$). Figure 1 plots $(x_i, \gamma_t(x_i))$ for $i = 1, 2, 3$, as well as $\gamma_t(x)$ as a function of x. Notice that x_1 lies between x_2 and x_3, so that it is an element of their convex hull.

One way to use the conditional means of the untreated units ($\gamma_t(x_2)$ and $\gamma_t(x_3)$) to approximate that of the treated unit ($\gamma_t(x_1)$) is to linearly interpolate between x_2 and x_3 to obtain

$$\gamma_t^{\text{li}} \equiv \gamma_t(x_2) + (\gamma_t(x_3) - \gamma_t(x_2))\left(\frac{x_1 - x_2}{x_3 - x_2}\right).$$

This is equivalent to setting w_2 and w_3 to be

$$w_2^{\text{li}} \equiv 1 - \left(\frac{x_1 - x_2}{x_3 - x_2}\right) \quad \text{and} \quad w_3^{\text{li}} \equiv \left(\frac{x_1 - x_2}{x_3 - x_2}\right).$$

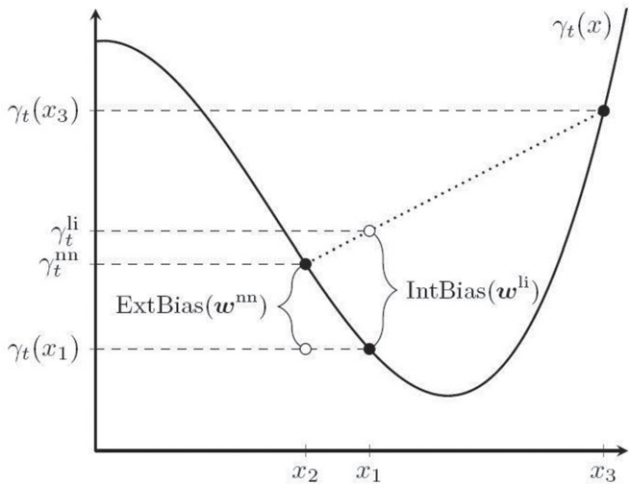


Figure 1. Extrapolation vs. interpolation bias.

Since $x_1 = w_2^{\text{li}} x_2 + w_3^{\text{li}} x_3$, the extrapolation bias associated with the weights $\boldsymbol{w}^{\text{li}} \equiv (w_2^{\text{li}}, w_3^{\text{li}})$ is zero. However, as shown in Figure 1, there is still bias due to interpolation, because $\gamma_t(x)$ is not a linear function of x, and thus

$$\text{IntBias}(\boldsymbol{w}^{\text{li}}) = \gamma_t\left(w_2^{\text{li}} x_2 + w_3^{\text{li}} x_3\right) - \left[w_2^{\text{li}} \gamma_t(x_2) + w_3^{\text{li}} \gamma_t(x_3)\right] \neq 0.$$

Another way to use the untreated units is to simply use the conditional mean for the unit whose value of x_i is closest to x_1. In Figure 1, this is the second untreated unit, $i = 2$. The weights for this approximation strategy are the nearest neighbor weights of $w_2^{\text{nn}} = 1$ and $w_3^{\text{nn}} = 0$, which produce $\gamma_t(x_2)$ as an approximation to $\gamma_t(x_1)$. This approach does not interpolate, so

$$\text{IntBias}(\boldsymbol{w}^{\text{nn}}) = \gamma_t(1 \cdot x_2 + 0 \cdot x_3) - (1 \cdot \gamma_2(x_2) + 0 \cdot \gamma_3(x_3)) = 0.$$

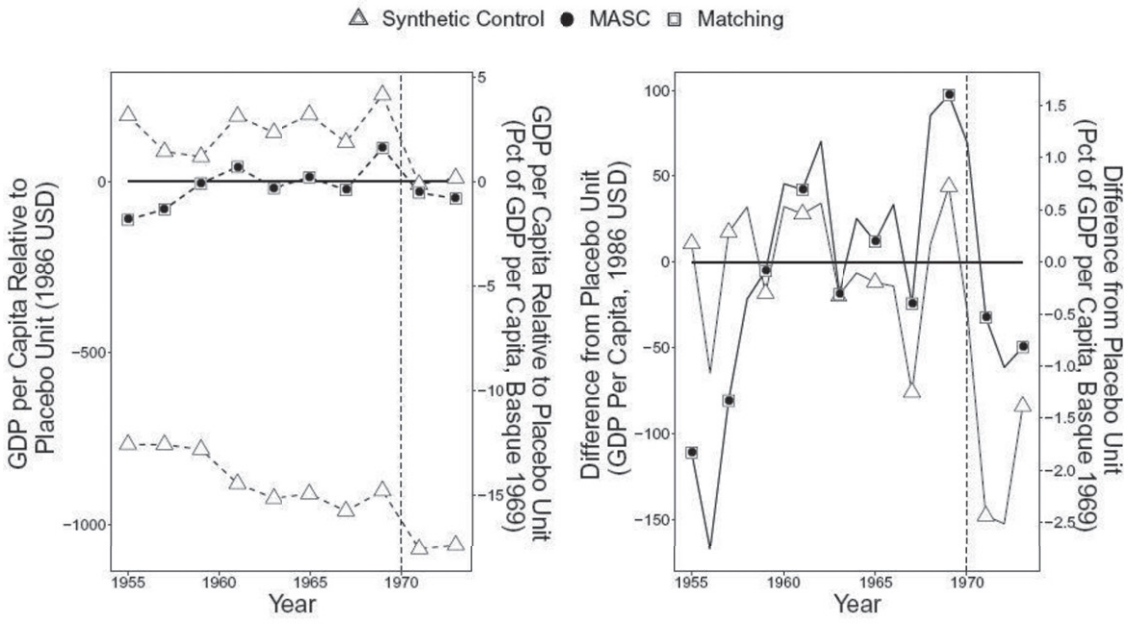
However, it does extrapolate, creating bias to the extent that $\gamma_t(x_1) \neq \gamma_t(x_2)$.

The estimators we consider in this article aim to control interpolation bias, extrapolation bias, or a combination of both, by minimizing bounds on these quantities. Assuming that γ_t is Lipschitz, the magnitude of extrapolation bias can be bounded by

$$|\text{ExtBias}(\boldsymbol{w})| \leq c \left\| \boldsymbol{x}_1 - \sum_{i\geq 2} w_i \boldsymbol{x}_i \right\| \equiv c \times \text{Ext}(\boldsymbol{w}),$$

where $c > 0$ is the Lipschitz constant. Under the same Lipschitz assumption, and assuming that interpolation is actually possible, so that the weights can be chosen to satisfy $\boldsymbol{x}_1 = \sum_{i\geq 2} w_i \boldsymbol{x}_i$ (that is, $\boldsymbol{x}_1$ is in the convex hull of $\{\boldsymbol{x}_i\}_{i\geq 2}$), the magnitude of interpolation bias can be bounded by

$$\begin{aligned} |\text{IntBias}(\boldsymbol{w})| &= \left| \sum_{i\geq 2} w_i \left(\gamma_t(\boldsymbol{x}_1) - \gamma_t(\boldsymbol{x}_i)\right) \right| \\ &\leq \sum_{i\geq 2} w_i |\gamma_t(\boldsymbol{x}_1) - \gamma_t(\boldsymbol{x}_i)| \leq c \sum_{i\geq 2} w_i \|\boldsymbol{x}_1 - \boldsymbol{x}_i\| \\ &\equiv c \times \text{Int}(\boldsymbol{w}). \end{aligned} \tag{7}$$



(a) Untreated outcome paths for selected regions.

(b) Estimated outcome paths for the placebo treated region.

Figure 2. The potential for interpolation bias with the SC estimator.
Notes: This figure depicts a simulation draw from the empirical Monte Carlo described in Section 4.6, using Rioja as the placebo and setting $\boldsymbol{X}_i$ to include pre-treatment outcomes (and no other covariates). The vertical dashed line indicates the beginning of the treatment period. Panel (a) depicts selected control regions which are assigned a weight of at least 0.15 by one of the estimators. Markers indicate which of these controls are assigned weight in the different estimators. Panel (a) plots each estimator using the same markers as in panel (b). In panel (a), paths of GDP per capita over time are plotted relative to the placebo treated region, which is the path lying on the y-intercept. Panel (b) plots the fit and treatment effects (prediction error) for each estimator.

Thus, by choosing $\boldsymbol{w}$ to minimize $\text{Ext}(\boldsymbol{w})$ and/or $\text{Int}(\boldsymbol{w})$, one can control extrapolation and/or interpolation bias.

2.3. The Synthetic Control Estimator

The SC estimator of $\gamma_t(\boldsymbol{x}_1)$ proposed by Abadie and Gardeazabal (2003) and later elaborated by Abadie, Diamond, and Hainmueller (2010, 2015) is defined as follows:

$$\begin{aligned}\hat{\gamma}_t^{\text{sc}} &\equiv \boldsymbol{y}_{0t}'\hat{\boldsymbol{w}}^{\text{sc}} \quad \text{where} \\ \hat{\boldsymbol{w}}^{\text{sc}} &\equiv \arg\min_{\boldsymbol{w}\in\mathcal{S}} \left\|\boldsymbol{x}_1 - \boldsymbol{x}_0'\boldsymbol{w}\right\|^2 \\ &\equiv \arg\min_{\boldsymbol{w}\in\mathcal{S}} \text{Ext}(\boldsymbol{w})^2, \end{aligned} \tag{8}$$

and where we have organized the untreated unit covariates into an $n \times k$ matrix $\boldsymbol{x}_0$. The Euclidean norm in the definition of $\text{Ext}(\boldsymbol{w})$ might be weighted by some symmetric, positive semidefinite matrix, but we omit this from the notation for simplicity. The SC weights, $\hat{\boldsymbol{w}}^{\text{sc}}$, are chosen so that the weighted average of covariates among the untreated units comes as close as possible to matching the covariate vector of the treated unit, subject to the convexity constraint that they are nonnegative and sum to unity.

The SC estimator has a number of attractive properties. By construction, it minimizes the quantity $\text{Ext}(\boldsymbol{w})$ that bounds extrapolation bias. If this quantity can be made zero, then the SC estimator will have no extrapolation bias. This stands in contrast to linear regression, which is known to be subject to potentially large extrapolation biases depending on how it is specified (see Imbens, 2004, pg. 13, or Abadie, 2020). Another benefit of the SC estimator is that the weights $\hat{\boldsymbol{w}}^{\text{sc}}$ are generally sparse, in the sense that they are only non-zero for a few untreated units (Abadie and L'Hour 2020). This aids in transparency by providing a way for experts to use contextual knowledge to evaluate the plausibility of the resulting estimates (Abadie 2020, sec. 4). Also, solving for $\hat{\boldsymbol{w}}^{\text{sc}}$ only requires solving the quadratic program in Equation (8), which is a convex optimization problem.

One concern with the SC estimator is that it is susceptible to interpolation bias. This was noted by Abadie, Diamond, and Hainmueller (2010, pp. 495 and 496), and has been discussed more recently by Abadie and L'Hour (2020), although those authors emphasize non-uniqueness issues that occur with many treated or untreated units. In Figure 2, we illustrate how interpolation biases can arise with the SC estimator. This figure plots several untreated units in a draw from the empirical Monte Carlo simulation introduced in Section 4.6, which uses the Spanish terrorism data of Abadie and Gardeazabal (2003). The simulation considers a placebo exercise in which the "treated unit" is not in fact treated, so that the ground-truth treatment effect is known to be zero. The left-hand side of Figure 2 plots the outcome paths of the untreated units relative to that of a placebo treated unit, while the right-hand side depicts different estimators, which should be zero in the post-period if they are performing well.

Suppose that we follow the recent tradition in the SC literature of taking $\boldsymbol{X}_i$ to include all pretreatment outcomes and no other covariates (Doudchenko and Imbens 2016; Ferman 2020). We temporarily focus on this case because it allows us to examine the methods graphically. In the simulated data

of Figure 2, this produces an SC estimator that is comprised primarily of two regions that have GDP per capita quite different from the placebo region. The SC estimator puts zero weight on the region whose pre-period outcome path oscillates closely to the treated region. Instead, it obtains the best pre-period fit by weighting the two more distant regions. This choice of weights minimizes extrapolation, but creates interpolation.

Whether such interpolation leads $\hat{\gamma}_t^{\text{sc}}$ to be biased depends on the structure of the function $\gamma_t(\boldsymbol{x})$. Assuming that $\boldsymbol{x}_1$ lies in the convex hull of $\boldsymbol{x}_0$, the SC estimator will have no interpolation bias (IntBias($\hat{\boldsymbol{w}}^{\text{sc}}$) = 0) if and only if

$$\sum_{i\geq 2} \hat{w}_i^{\text{sc}} \gamma_t(\boldsymbol{x}_i) = \gamma_t(\boldsymbol{x}_1) = \gamma_t\left(\sum_{i\geq 2} \hat{w}_i^{\text{sc}} \boldsymbol{x}_i\right). \tag{9}$$

In order for Equation (9) to hold, the function γ_t needs to be linear in $\boldsymbol{x}$ on the empirical support of $\boldsymbol{X}_i$. This is a restrictive functional form assumption about an unknown function.

When Equation (9) is not satisfied, interpolation bias will arise. This is illustrated in Figure 2, where the SC estimator fits the pre-period path of the simulated treated unit by interpolating between two regions with very different pre-period paths. However, condition (9) fails, and that the resulting interpolation bias leads $\hat{\gamma}_t^{\text{sc}}$ to be a poor estimate of the post-treatment outcomes of the treated unit. We emphasize that this bias stems from the failure of Equation (9) even assuming perfect fit ($\boldsymbol{x}_1 = \boldsymbol{x}_0'\hat{\boldsymbol{w}}^{\text{sc}}$); it is distinct from the bias due to imperfect fit considered by Ferman and Pinto (2019) and Ben-Michael, Feller, and Rothstein (2019).

Of course, Figure 2 is just one simulation draw, and one which we have selected to show how interpolation bias can arise. The frequency with which it actually arises in applications is an empirical question. We address this empirical question for the Spanish data in Section 4, where we report the results from our full placebo analysis and Monte Carlo simulations.

2.4. The Matching Estimator

Local nonparametric smoothing estimators are a classical way to estimate $\gamma_t(\boldsymbol{x}_1)$. In general, these estimators can be written as follows:

$$\hat{\gamma}_t^{\text{lo}} \equiv \sum_{i\geq 2} \kappa\left(\|\boldsymbol{x}_i - \boldsymbol{x}_1\|\right) y_{it} \equiv \boldsymbol{y}_{0t}'\hat{\boldsymbol{w}}^{\text{lo}}, \tag{10}$$

where κ is a kernel function that determines the weight applied to each untreated observation. For such an estimator to be local, the function κ should be decreasing, so that untreated units with predetermined characteristics more distant from the treated unit are given less weight. Local smoothing estimators do not require the linearity condition (9) that was required for the SC estimator. Instead, they rely only on γ_t being sufficiently smooth in its continuous components (e.g. Fan and Gijbels 1992).

Unlike the SC estimator, local smoothing estimators do not necessarily have sparse, convex weights. However, the specific class of k–nearest neighbors estimators (e.g., Cover 1968) does have weights with these properties. Estimators based on the nearest neighbors idea are widely used for causal inference problems under Assumptions 1 and 2, in which case they are commonly described as matching estimators (e.g. Dehejia and Wahba 1999; Abadie and Imbens 2006).

The matching estimator we consider is defined by choosing a positive integer and equally weighting the m untreated units with pre-period characteristics closest to those of the treated unit. That is,

$$\hat{\gamma}_t^{\text{ma}}(m) \equiv \boldsymbol{y}_{0t}'\hat{\boldsymbol{w}}^{\text{ma}}(m), \tag{11}$$

where the weights $\hat{w}_i^{\text{ma}}(m)$ are $1/m$ for the m units with smallest $\|x_i - x_1\|$ and 0 for all other units. For simplicity, we assume there are no ties. Note that this vector of weights can be written as the solution to the optimization problem

$$\hat{\boldsymbol{w}}^{\text{ma}}(m) = \arg\min_{\boldsymbol{w}\in\mathcal{S}} \underbrace{\sum_{i\geq 2} w_i \|\boldsymbol{x}_1 - \boldsymbol{x}_i\|}_{\equiv \text{Int}(\boldsymbol{w})}$$
$$\text{s.t.} \quad w_i \leq \frac{1}{m} \quad \text{for all } i \geq 2, \tag{12}$$

since the w_i corresponding to the smallest $\|\boldsymbol{x}_1 - \boldsymbol{x}_i\|$ will be pushed up against $1/m$ until the m smallest observations have reached this bound, while all other w_i will be set to 0.

Like the SC estimator, the matching estimator is a sparse, convex weighted average of the post-period outcomes of the untreated units. However, in contrast to the SC estimator, the matching estimator aims to minimize Int($\boldsymbol{w}$), and thus to control interpolation bias. For example, with $m = 1$, the matching estimator selects a single region in Figure 2(a) that oscillates around the placebo region, since its pre-period characteristics are most similar to that of the placebo region. As a consequence, the matching estimator is less susceptible to interpolation bias than the SC estimator, and in this example provides a better estimate of the post-treatment outcomes for the placebo region.

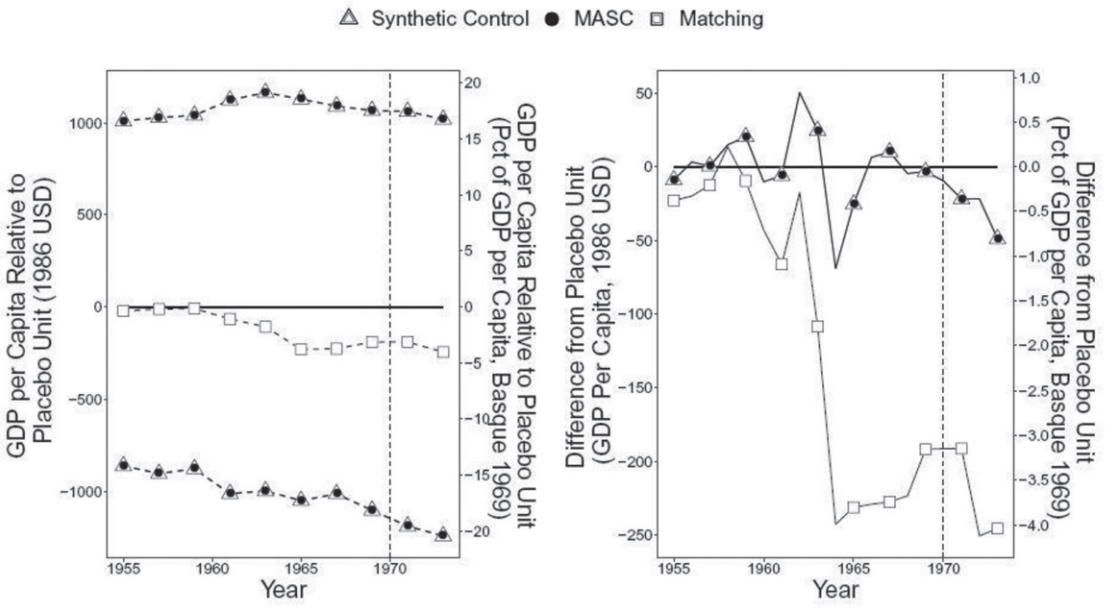
However, the matching estimator is more vulnerable to extrapolation bias than the SC estimator. To see this, consider Figure 3, which reports a different simulation draw. In this draw, the matching estimator uses the single untreated unit that has pre-period path closest to the treated unit, even though their pre-period paths are not actually that close, resulting in considerable bias. In contrast, the SC estimator weights two distant regions in a way that provides an excellent fit to the outcome path of the treated unit throughout both the pre- and post-periods. This is consistent with a case in which γ_t is close to linear, so that Equation (9) is close to satisfied, and the SC estimator has little interpolation bias.

2.5. Model Averaging With the MASC Estimator

Both the SC and matching estimators share a number of appealing properties in common. As illustrated in Figures 2 and 3, however, their drawbacks are different and diametrically opposed: the SC estimator controls extrapolation bias but not interpolation bias, while the matching estimator does the opposite. This complementarity suggests that a model averaging estimator will be able to harness the best properties of both the matching and SC estimators. See, for example, the discussion surrounding Breiman (1996, theor. 1).

With this motivation, we define the MASC estimator as

$$\hat{\gamma}_t^{\text{masc}} \equiv \phi\hat{\gamma}_t^{\text{ma}}(m) + (1-\phi)\hat{\gamma}_t^{\text{sc}} \equiv \boldsymbol{y}_{0t}'\hat{\boldsymbol{w}}^{\text{masc}}$$



(a) Untreated outcome paths for selected regions.

(b) Estimated outcome paths for the placebo treated region.

Figure 3. The potential for extrapolation bias with matching.
Notes: This figure depicts a simulation draw from the empirical Monte Carlo described in Section 4.6, using Navarre as the placebo and setting $\boldsymbol{X}_j$ to include pre-treatment outcomes (and no other covariates). The vertical dashed line indicates the beginning of the treatment period. Panel (a) depicts selected control regions which are assigned a weight of at least 0.15 by one of the estimators. Markers indicate which of these controls are assigned weight in the different estimators. Panel (a) plots each estimator using the same markers as in panel (b). In panel (a), paths of GDP per capita over time are plotted relative to the placebo treated region, which is the path lying on the y-intercept. Panel (b) plots the fit and treatment effects (prediction error) for each estimator.

where $\phi \in [0, 1]$ is a tuning parameter, and $\hat{\boldsymbol{w}}^{\text{masc}} \equiv \phi \hat{\boldsymbol{w}}^{\text{ma}}(m) + (1-\phi)\hat{\boldsymbol{w}}^{\text{sc}}$. In Section 3, we provide a cross-validation procedure for choosing ϕ and m. This allows the MASC to control both interpolation and extrapolation biases in a data-driven way. When interpolation bias is the chief concern, the procedure makes the MASC estimator assign more weight to the matching estimator. In Figure 2, it sets $\phi = 1$ and $m = 1$, so that the MASC exactly coincides with the 1-nearest-neighbor matching estimator. On the other hand, when extrapolation bias is the concern, the procedure assigns more weight to the SC estimator. For example, in Figure 3, it sets $\phi = 0$, so that the MASC exactly coincides with the SC estimator.

Intermediate cases can also arise, as in the simulation draw depicted in Figure 4. In this case, the linearity condition (9) fails, so that the SC estimator suffers from interpolation bias. At the same time, there are no untreated units that closely match the pre-period path of the placebo unit, so the matching estimator suffers from extrapolation bias. The cross-validation procedure chooses $\phi \approx .5$, which allows the MASC estimator to mix the SC estimator with the matching estimator, mitigating both sources of bias.

2.6. The Penalized Synthetic Control Estimator

A related, but different estimator has recently been proposed by Abadie and L'Hour (2020). Those authors start with the SC estimator and add a penalty that discourages choosing units far from the treated unit. Their penalized SC estimator is defined as

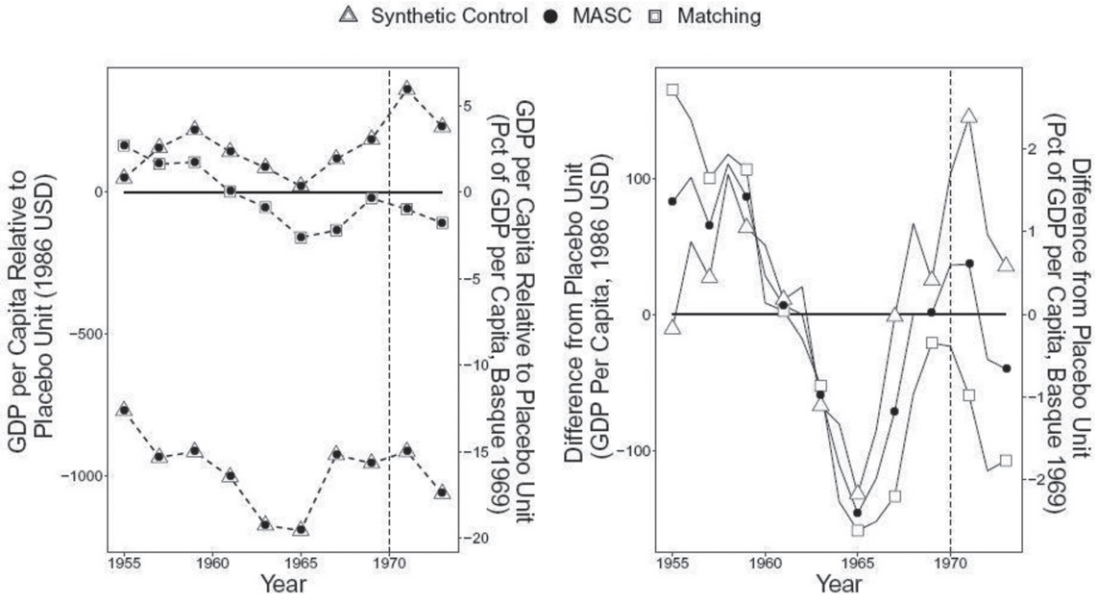
$$\begin{aligned}\hat{\gamma}_t^{\text{pen}} &\equiv \boldsymbol{y}_{0t}'\hat{\boldsymbol{w}}^{\text{pen}} \quad \text{with} \\ \hat{\boldsymbol{w}}^{\text{pen}} &\equiv \arg\min_{\boldsymbol{w}\in\mathcal{S}} (1-\pi)\|\boldsymbol{x}_1 - \boldsymbol{x}_0'\boldsymbol{w}\|^2 \\ &\quad + \pi\left(\sum_{i\geq 2} w_i \|\boldsymbol{x}_i - \boldsymbol{x}_1\|^2\right), \end{aligned} \tag{13}$$

where $\pi \in (0, 1]$ is a tuning parameter chosen through cross-validation that controls the penalty incurred by weighting untreated units with pre-treatment characteristics different from the treated unit. With $\pi = 0$, (13) would reduced to the usual (unpenalized) SC estimator, $\hat{\gamma}_t^{\text{sc}}$. Abadie and L'Hour (2020) exclude $\pi = 0$, but consider the limiting case $\pi \to 0$, which they refer to as "pure synthetic control." With $\pi = 1$, the penalized SC estimator is equal to $\hat{\gamma}_t^{\text{ma}}(m)$ with $m = 1$. (Note that Abadie and L'Hour (2020) parameterize their criterion function slightly differently by normalizing $(1 - \pi)$ to 1 and allowing any $\pi > 0$. The two formulations are equivalent.)

The optimization problem solved by the penalized SC estimator is a constrained version of the one implicitly solved by the MASC estimator. This is because Equation (13) can also be written as follows:

$$\begin{aligned}\hat{\boldsymbol{w}}^{\text{pen}} &= \arg\min_{\boldsymbol{w}^a,\boldsymbol{w}^b\in\mathcal{S}} (1-\pi)\|\boldsymbol{x}_1 - \boldsymbol{x}_0'\boldsymbol{w}^a\|^2 \\ &\quad + \pi\left(\sum_{i\geq 2} w_i^b \|\boldsymbol{x}_i - \boldsymbol{x}_1\|^2\right) \quad \text{s.t.} \quad \boldsymbol{w}^a = \boldsymbol{w}^b, \end{aligned}$$

whereas $\hat{\boldsymbol{w}}^{\text{masc}}$ is the solution to this program (with π replaced by ϕ) when m is fixed at 1 and the constraint $\boldsymbol{w}^a = \boldsymbol{w}^b$ is



(a) Untreated outcome paths for selected regions.

(b) Estimated outcome paths for the placebo treated region.

Figure 4. MASC adapts to control both extrapolation and interpolation bias.
Notes: This figure depicts a simulation draw from the empirical Monte Carlo described in Section 4.6, using Rioja as the placebo and setting $\boldsymbol{X}_i$ to include pre-treatment outcomes (and no other covariates). The vertical dashed line indicates the beginning of the treatment period. Panel (a) depicts selected control regions which are assigned a weight of at least 0.15 by one of the estimators. Markers indicate which of these controls are assigned weight in the different estimators. Panel (a) plots each estimator using the same markers as in panel (b). In panel (a), paths of GDP per capita over time are plotted relative to the placebo treated region, which is the path lying on the y-intercept. Panel (b) plots the fit and treatment effects (prediction error) for each estimator.

dropped. Note that when this constraint is dropped the problem becomes separable in $\boldsymbol{w}^a$ and $\boldsymbol{w}^b$, and the squares on the norms can be removed without changing the optimal solutions. While the MASC estimator takes a convex combination of the SC and matching estimators—which respectively minimize bounds on extrapolation and interpolation bias—the penalized SC estimator minimizes a convex combination of the (squared) SC and matching objective functions. This can lead it to choose an entirely different set of weights.

It is important to mention that we are ignoring a primary motivation provided by Abadie and L'Hour (2020) for the penalized SC estimator, which is its ability to solve the non-uniqueness problem that can arise when solving the SC problem (8). As Abadie and L'Hour (2020) discussed, this problem is usually not an issue when there is a single treated unit, which is the case we consider here. It becomes more likely to be problematic with multiple treated units. In such settings, one could modify the MASC so that it averages between the matching and *penalized* SC estimators. We expect that the resulting estimator would behave similar to the way the MASC behaves when there is a single treated unit.

2.7. Sparsity

A key motivation for the SC method is sparsity in the weights. In comparative case studies, sparsity facilitates the interpretation of the counterfactual estimate and the recognition and assessment of potential biases (see, e.g., Abadie 2020, sec. 4).

The sparsity properties of the MASC estimator are inherited by those of the SC and MA estimators. Suppose that n^{sc} of the components of $\hat{\boldsymbol{w}}^{\text{sc}}$ are nonzero. Then at most $n^{\text{sc}} + m$ of the components of $\hat{\boldsymbol{w}}^{\text{masc}}$ are nonzero. In practice, we often find that the untreated units given nonzero weights by the SC and MA weighting schemes are partially overlapping, so that the actual number of nonzero elements is smaller than this. The smallest number of nonzero components that $\hat{\boldsymbol{w}}^{\text{masc}}$ can have is the minimum of m and n^{sc}, which happens if $\phi = 1$ or $\phi = 0$, respectively. Since m can be as small as 1, the MASC weights could have only one nonzero weight. Thus, the MASC is generically neither more nor less sparse than the SC, MA or penalized SC estimators.

3. Cross-Validation

3.1. Definitions

In this section, we propose a cross-validation procedure for choosing the tuning parameters for the estimators discussed in the previous section. As in Abadie, Diamond, and Hainmueller (2015), our procedure is based on optimizing the fit of the treated unit's outcome series in the pretreatment period. Whereas those authors used a single training-validation split, our procedure uses a series of one-step ahead forecasts, each of which is estimated using data only from periods prior to the forecast date. This is called rolling-origin recalibration in the

forecasting literature (e.g., Tashman 2000; Bergmeir and Benítez 2012), which is related to the rolling-window considered by Swanson and White (1997). The procedure is attractive, because it preserves the temporal structure of the forecasting problem.

We define our folds, $f = 1, \ldots, F$, as consisting of data running between two dates $\underline{t}_f$ and $\bar{t}_f$ in the pre-treatment period. Let $\hat{\gamma}_f(\boldsymbol{\tau})$ denote a generic estimator of the outcome in period $\bar{t}_f+1$ based on data in fold f, where $\boldsymbol{\tau}$ is a vector of tuning parameters. Our cross-validation procedure chooses $\boldsymbol{\tau}$ to minimize the average one-step ahead forecast error, which we denote as

$$\text{cv}(\boldsymbol{\tau}) \equiv \frac{1}{F}\sum_{f=1}^{F}\left(y_{1(\bar{t}_f+1)} - \hat{\gamma}_f(\boldsymbol{\tau})\right)^2. \tag{14}$$

We consider the one-step ahead forecast primarily for concreteness; one can modify the criterion (14) to combine multiple forecast periods under different weights chosen by the researcher. Our R package (https://github.com/maxkllgg/masc) implements such a modification to allow for more general criteria.

The largest that F can be is $t^\star - 2$ if we set $\underline{t}_f = 1$ and $\bar{t}_f = f$ for each $f = 1, \ldots, (t^\star - 2)$. In practice, we use fewer folds than this, and prefer folds that are longer. The bias-variance trade-offs that drive this choice are natural. Folds that end closer to the treatment date ($\bar{t}_f$ closer to $t^\star$) are likely to be more relevant to the posttreatment period. They can also be made longer ($\bar{t}_f - \underline{t}_f$), so that the estimators use more data. On the other hand, we expect that having more folds will decrease the variance of $\text{cv}(\boldsymbol{\tau})$ in repeated samples. Similar trade-offs are also present in cross-validation with independent and identically distributed data (e.g. Hastie, Tibshirani, and Friedman 2009, pp. 242 and 243). The added complication here is that not all folds are equally valuable, so we prefer ones that use data closer to the actual treatment date.

The parameters $\boldsymbol{\tau}$ differ by estimator. The SC estimator has no tuning parameters. (As mentioned earlier, the Euclidean norm defining the SC or matching estimators could be weighted. Abadie, Diamond, and Hainmueller (2015) treated the weights as tuning parameters. We could do this as well, but we have elected not to in the current article because optimizing over the weights introduces computational issues that, while solvable, are not the main focus of our article (Becker and Klößner 2017, 2018).) For the matching estimator, $\boldsymbol{\tau}$ is the number of matches, m. The MASC estimator has both m and the model average parameter, ϕ. The penalized SC estimator has the penalty parameter, π.

3.2. Computation

For the MASC estimator, it is straightforward to find the unconstrained minimum of $\text{cv}(\boldsymbol{\tau}) \equiv \text{cv}(\phi, m)$ in ϕ for any fixed m. Using least-squares algebra, the solution can be shown to be

$$\hat{\phi}(m) \equiv \frac{\sum_{f=1}^{F}(\hat{\gamma}_f^{\text{ma}}(m) - \hat{\gamma}_f^{\text{sc}})(y_{1,\bar{t}_f+1} - \hat{\gamma}_f^{\text{sc}})}{\sum_{f=1}^{F}(\hat{\gamma}_f^{\text{ma}}(m) - \hat{\gamma}_f^{\text{sc}})^2}. \tag{15}$$

This means that cross-validating the MASC is extremely easy computationally. First, compute $\hat{\phi}(m)$ for a set of potential matches, $m \in \mathcal{M}$. Then for each $m \in \mathcal{M}$, set

$$\hat{\phi}^\star(m) \equiv \begin{cases} 0, & \text{if } \hat{\phi}(m) \leq 0 \\ 1, & \text{if } \hat{\phi}(m) \geq 1 \\ \hat{\phi}(m) & \text{otherwise} \end{cases}$$

Finally set $\hat{m}^\star \equiv \arg\min_{m\in\mathcal{M}} \text{cv}(\hat{\phi}^\star(m), m)$, and set $\hat{\phi}^\star \equiv \hat{\phi}^\star(\hat{m}^\star)$. The cross-validated MASC estimator is a weighted average of $\hat{\gamma}^{\text{sc}}$ and $\hat{\gamma}^{\text{ma}}(\hat{m}^\star)$ with weights $(1 - \hat{\phi}^\star)$ and $\hat{\phi}^\star$, respectively.

For the penalized SC estimator, $\text{cv}(\boldsymbol{\tau}) \equiv \text{cv}(\pi)$ is not necessarily convex in π, which makes it harder to find the global minimum. In the results ahead, we use a grid search to cross-validate the penalized SC estimator.

4. Placebo Analyses

4.1. Design

We use a series of empirical placebo analyses to examine the behavior of the estimators described in Section 2. These exercises use the same data as in Abadie and Gardeazabal's (2003) application of the SC method to study the effect of terrorism on per capita GDP in Spain. The data consist of time series on per capita GDP running from 1955 to 1997 for 17 regions in Spain. Additionally, the data includes 12 other covariates observed intermittently for each region over the same interval, representing educational attainment of the labor force (share with primary, high school, or more than high school education), investment (as a share of GDP), sectoral shares (in agriculture/forestry/fishing, energy/water, industry, construction/engineering, marketable services, or nonmarketable services), and population density in each region. The treated unit is the Basque Country, and the treatment is the onset of separatist terrorism, which begins in 1970.

Following Abadie and Gardeazabal (2003), we take $\boldsymbol{X}_i$ to include averages for the 13 characteristics from 1960 to 1969. For the SC estimator, we use the same weighted norm selected by Abadie and Gardeazabal (2003) when solving for the weights in (8). For the penalized SC estimator, we present results both with this norm ("PSC-AG") and with a norm that weights each component by its inverse standard deviation ("PSC-S"), as in Abadie and L'Hour (2020, sec. 6). The matching estimator uses this inverse standard deviation norm.

Abadie and Gardeazabal (2003) performed a placebo analysis using Catalonia as the placebo region. Their stated rationale was that Catalonia is similar to the Basque Country, but with lower exposure to terrorism, and particularly salient for their results, since it received the most weight in their application of the SC method. They found that the SC estimator reproduced the actual per capita GDP for Catalonia quite well, at least until the 1980s. They interpreted this as evidence in support of their estimates for the Basque Country.

Using the same logic, we extend this placebo exercise to all of the untreated regions of Spain, with the exception of the Balearic Islands, Extremadura, and Madrid. The reason for excluding these three regions is that the SC estimator provides a particularly poor fit to their pre-period paths. Given the poor fit, one might argue that it is inappropriate to apply SC to these

Table 1. Performance of alternative estimators in the main Spanish placebo analyses.

	RMSPE					Pre-PeriodFit				
Placebo	SC	MASC	PSC-S	PSC-AG	Matching	SC	MASC	PSC-S	PSC-AG	Matching
Andalusia	501	512	356	507	551	3	15	95	35	58
	(501)	(264)	(252)	(293)	(264)					
Aragon	133	145	227	133	252	24	25	67	24	103
	(133)	(95)	(124)	(133)	(154)					
Asturias	903	787	1,001	903	709	34	75	48	34	120
	(903)	(532)	(979)	(903)	(532)					
Canary Isl	271	199	248	271	227	36	67	336	36	69
	(271)	(144)	(169)	(271)	(154)					
Cantabria	116	131	1,279	109	556	91	86	35	89	115
	(116)	(116)	(1,279)	(109)	(547)					
Cast-Leon	81	82	624	81	274	28	32	442	28	316
	(81)	(80)	(488)	(76)	(274)					
Cast-Mancha	378	271	99	378	328	60	100	145	60	72
	(378)	(87)	(93)	(378)	(100)					
Catalonia	252	268	1,468	252	1,351	18	44	1057	18	1295
	(252)	(252)	(1,468)	(239)	(810)					
Valencia	259	171	374	373	174	37	85	166	66	145
	(259)	(81)	(248)	(125)	(115)					
Galicia	105	99	411	105	422	21	22	316	21	296
	(105)	(64)	(259)	(82)	(285)					
Murcia	329	307	160	362	300	44	65	110	59	92
	(329)	(119)	(160)	(329)	(155)					
Navarre	230	230	397	230	1,034	23	23	76	23	425
	(230)	(230)	(267)	(230)	(267)					
Rioja	202	187	189	192	267	26	39	38	32	188
Average	**289**	**261**	**526**	**300**	**496**	**28**	**47**	**272**	**40**	**253**
	(289)	**(173)**	**(459)**	**(258)**	**(302)**					

Note: PSC-AG uses the weighted norm selected by Abadie and Gardeazabal (2003) when solving for the weights in Equation (8). PSC-S uses a norm weighted by the inverse standard deviation of each component, as in Abadie and L'Hour (2020, sec. 6). Root mean square prediction errors in this table are calculated from 1970 to 1997. Pre-period fit is calculated based on annual per capita GDP from 1960 to 1969. For reference, the pre-period fit of SC for the Basque Country is \$94. The values in parentheses represent the best possible (infeasible) RMSPE that each estimator could obtain for each region if tuning parameters were chosen directly to minimize it. Note that the SC estimator has no such tuning parameters (hence its actual RMSPE is equivalent to its infeasible RMSPE). GDP per capita is measured in 1986 U.S. dollars.

regions. (We thank the associate editors for pointing this out and suggesting that we exclude these three regions. Empirically, we find that including or excluding these regions does not materially change our findings about the relative performance of the various estimators. Results including these regions are provided in Supplemental Appendix C.)

The placebo analyses are conducted separately for each of the remaining 13 untreated regions. We use the same methodology described in Sections 2 and 3, except that now the "treated" unit is a placebo region in which no intervention took place at $t^\star = 1970$. For each estimator, we use data from 1960 to 1969 and cross-validate with $F = 5$ folds, each starting at $\underline{t}_f = 1960$ and ending at $\bar{t}_f \in \{1964, 1965, \ldots, 1968\}$. The number of matches for the matching and MASC estimators is chosen from $\mathcal{M} = \{1, 2, \ldots, 10\}$.

We calculate the mean squared prediction error (MSPE) for each region by taking the differences between its actual and forecasted outcome paths in each of the post-period years (1970–1997), squaring these differences, and then averaging the 28 years. (Our findings about the relative performance of the various estimators do not materially change if we use a shorter post-period, a point we demonstrate in Figure A.1 of the Supplemental Appendix). This procedure produces a MSPE for each placebo region and every estimator. As in Abadie and Gardeazabal (2003), we interpret a low MSPE as evidence that an estimator is performing well. Throughout our discussion, we focus on the square root of the MSPE (the RMSPE) so that errors are interpretable in units of GDP per capita.

4.2. Performance Across Estimators

Results for each placebo region are reported as rows in Table 1. The final row averages values across the 13 placebo regions. The first panel of columns in Table 1 compares the performance of five alternative estimators across each placebo region in terms of RMSPE. Qualitatively, the MASC tends to have lower RMSPE than the other estimators, including the PSC-AG estimator, which frequently coincides with the standard SC estimator. The MASC adapts to regions such as Valencia where matching has low RMSPE but SC has high RMSPE. It also adapts to regions such as Catalonia, Galicia, and Navarre, where matching has high RMSPE, but SC has low RMSPE. In regions like the Canary Islands, Castile-La Mancha, and Rioja, the MASC combines SC and matching in order to have lower RMSPE than both.

The five estimators exhibit noticeable quantitative differences in performance. On average across the placebo regions, the yearly RMSPE of MASC is \$261 per person, equivalent to 4.3% of GDP per capita in the Basque Country in 1969. By comparison, the SC estimator has an average yearly RMSPE of \$289 per person, which amounts to 4.7% of the Basque Country's GDP per capita. The PSC-AG estimator tends to perform worse, with an average yearly RMSPE of \$300 per person, amounting to 4.9% of the Basque Country's GDP per capita. Matching and PSC-S, on the other hand, tend to have even larger prediction errors, with average yearly RMSPEs of \$496 and \$526, respectively. To put these estimates into perspective, the yearly GDP growth of the Basque Country averaged \$159 per person across the years 1955–1969. This means that the average yearly prediction error

of the estimators ranges from 164% (MASC), to 182% (SC), to 189% (PSC-AG), and to over 300% (matching and PSC-S) of the annual GDP growth in the treated region prior to the onset of terrorism.

4.3. Pre-Period Fit and Post-Intervention Prediction Error

The performance of the SC estimator varies across placebo regions. In some regions, its performance is similar to MASC, while in others it performs considerably worse. There are two possible explanations for why SC produces relatively large prediction errors in certain placebo regions. One possible explanation is that the synthetic unit does not fit the pre-period paths in these regions. The other possible explanation is that the pre-period fit is good, and the prediction error results from the susceptibility of the SC estimator to interpolation bias. To evaluate these explanations, we report the pre-period fit of the SC estimator, in the second panel of columns in Table 1.

Overall, the SC estimator is able to fit the pre-period paths of the placebo regions quite well. Indeed, for every one of the placebo regions, we obtain a pre-period fit (RMSE) that is *lower* than that obtained for the Basque Country, which is the actual treated region. (We discuss these results in Section 5, where we find the RMSE of SC when applied to the Basque Country to be \$94 per person.) However, the placebo regions with good (or poor) pre-period fits do not necessarily have small (or large) prediction errors. For example, the SC estimator fits the pre-period data best in the regions of Andalusia and Cantabria. Yet, the prediction error is high in one of these regions, Andalusia, and low in the other region, Cantabria. Pre-period fit is not necessarily a reliable indicator of small prediction error for the other estimators either. For example, MASC has the worst pre-period fit in Andalusia, where the prediction error is the second lowest.

4.4. Cross-Validation and Prediction Error of MASC

The results in the first panel of columns in Table 1 suggest that MASC performs relatively well compared to the other estimators, at least in the current setting. However, MASC also has non-negligible prediction errors in some placebo regions. One potential reason for poor performance is that a suitable control group simply does not exist; that is, no combination of ϕ and m would lead to low prediction error. Another possibility is that a suitable control group exists, but the cross-validation procedure does a poor job locating it.

To distinguish between these two scenarios, we compare the actual RMSPE against the best possible (infeasible) RMSPE that MASC could obtain for each region if ϕ and m were chosen directly to minimize it in the post-treatment period. These results are reported in parentheses in the first panel of columns in Table 1. Results are also reported for matching, PSC-AG, and PSC-S estimators, which (like MASC) have tuning parameters selected by cross-validation. The best possible infeasible RMSPE that the SC estimator can achieve, on the other hand, is equivalent to its actual RMSPE (and hence, the values reported in parentheses are the same as the main values).

Averaging across regions for MASC, the minimum infeasible RMSPE is 34% lower than the actual RMSPE. However, there is a great deal of heterogeneity across regions. For example, the MASC has a relatively high RMSPE in Navarre not because the cross-validation procedure is failing, but because there is no suitable control group (choice of ϕ or m) for that region. On the other hand, MASC has relatively high prediction errors for Castile-La Mancha and Murcia because the cross-validation procedure selects a control group that does considerably worse than the infeasible optimal one.

In Appendix B, we explore alternative cross-validation procedures based on multi-step ahead criteria. Our ability to assess criteria that fully reflect the 28-year length of the post-period is limited by the 15-year length of the pre-period. Our results, though, suggest that multi-step ahead criteria return RMSPEs that tend to be slightly higher than RMSPEs of the one-step ahead rolling-origin cross-validation procedure used in our main results.

4.5. Biases Due to Interpolation Versus Extrapolation

The results in the first panel of columns of Table 1 show that MASC tends to have lower RMSPE than the other estimators in the placebo analyses. In Figure 5, we investigate the reason for this by plotting the pre-period fit (RMSE) and post-period prediction error (RMSPE) as functions of ϕ and π for the MASC and PSC-AG estimators. We report the average across all placebo regions in dotted lines. We also report Castile-La Mancha separately as a solid line, as an example placebo region where MASC deviates from SC.

When $\phi = 0$ the MASC estimator correspond to the standard SC estimator, which controls extrapolation bias by maximizing fit on pre-period outcomes. As $\pi \to 0$, the PSC-AG estimator also corresponds to the standard SC estimator because, with only a small number of control regions, no region falls in the convex span of the characteristics of other regions (Abadie and L'Hour 2020). Panels (a) and (c) of Figure 5 show how pre-period fit tends to deteriorate as ϕ and π increase.

At $\phi = \pi = 1$, both the MASC and PSC-AG estimators correspond to matching estimators which control interpolation bias. Intermediate values of ϕ and π represent a trade-off between controlling extrapolation and interpolation bias. MASC captures this trade-off by assigning weight to both the SC and matching estimators. PSC-AG captures it by changing the relative penalty for giving weight to distant units.

Panels (b) and (d) of Figure 5 show the prediction errors of MASC and PSC-AG. Average prediction error for the MASC is minimized at around $\phi = 0.05$, reflecting observation in Section 4.2 that MASC is able to adapt to the regions where SC performs well (where combining it with matching would substantially increase prediction error), while also blending SC and matching in the regions like Castile-La Mancha where doing so can reduce prediction error. The prediction error of PSC-AG is lowest at $\pi \approx 0$ on average and individually for most regions. One exception where PSC-AG deviates from SC and performs well is Rioja, where it obtains its lowest RMSPE by setting $\pi = 0.10$.

4.6. Monte Carlo Simulations

The placebo analyses in this section have been based on a given dataset, that is, on one particular realization of the underlying

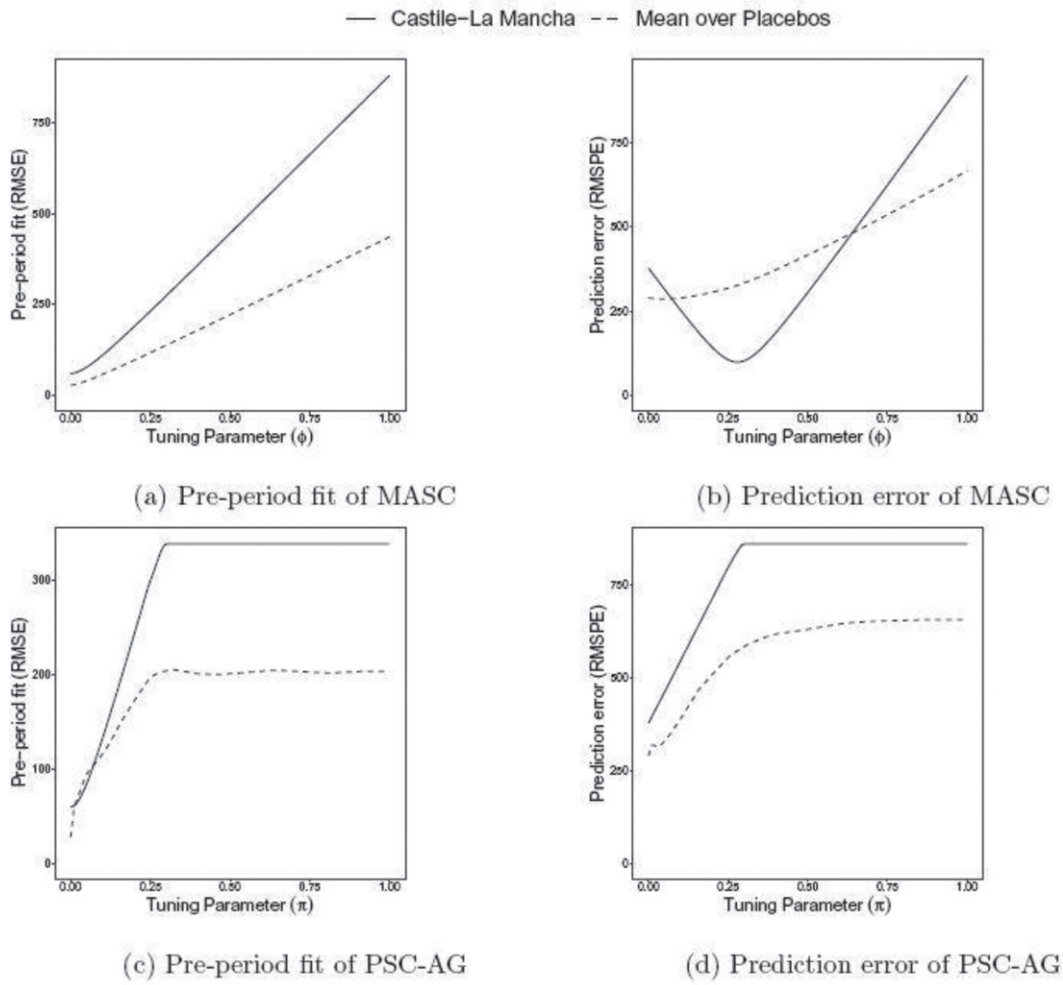


Figure 5. Trading-off between fit and prediction error for MASC and penalized SC.
Note: Each graph depicts how error (prediction error or pre-period fit) evolves as we move from SCs toward matching for the given estimator. Pre-period fit is based on annual per capita GDP from 1960 to 1969 with an unweighted norm. Outcomes are measured in 1986 U.S. dollars. For each placebo, the matching estimator (*m*) is fixed at the value selected by cross-validation. For Castile-La Mancha, $m = 9$. π has been re-scaled so that $\pi = 1$ denotes the *smallest* value at which PSC-AG exactly corresponds to a 1-nearest-neighbor estimator.

data generating process. This raises two questions. The first is whether the relative performance of the alternative estimators would change when looking across multiple realizations of the same data-generating process. The second is how the estimators would perform under alternative data-generating processes. To answer both questions, we conducted a Monte Carlo study, which we discuss in detail in Appendix D.

The results from this simulation support three key insights from the placebo analyses: (i) MASC tends to have lower RMSPE than the other estimators; (ii) the pre-period fit of the SC estimator is not necessarily a strong indicator of its prediction error; and (iii) the cross-validation procedure tends to select suitable control groups for MASC when they exist. However, one may prefer the SC estimator if data is generated from a more restricted process in which regional characteristics are driven only by a small number of latent factors, as a suitable SC is more likely to exist in that setting.

5. Re-Examining the Economic Costs of Conflict

In this section, we re-analyze the Spanish terrorism application of Abadie and Gardeazabal (2003). The goal is to assess if the alternative estimators yield substantively different estimates of the economic costs of conflict for the Basque Country. We continue to use the same estimation procedures as in Section 4, except that now the Basque Country is the treated unit. The economic costs of conflict are calculated by taking the difference between the Basque Country's actual and forecasted outcome path over the post-period.

Figure 6 plots the actual path of per capita GDP for the Basque Country against the counterfactual paths constructed by the MASC, SC, PSC-AG, PSC-S, and matching estimators. It happens to be the case that, through cross-validation, the MASC estimator selects $\phi = 0$ and corresponds exactly to the original SC estimator of Abadie and Gardeazabal (2003). Both of these estimators average together Catalonia with a weight of 0.85 and Madrid with a weight of 0.15. The PSC-AG estimator selects a small but positive penalty π, so that it averages Catalonia with a weight of 0.88 and Madrid with a weight of 0.12, leading to a counterfactual that is very close to the original SC estimator of Abadie and Gardeazabal (2003). The PSC-S estimator, on the other hand, selects a larger penalty $\pi > 0$ and places virtually all weight on Catalonia (the nearest neighbor of the Basque

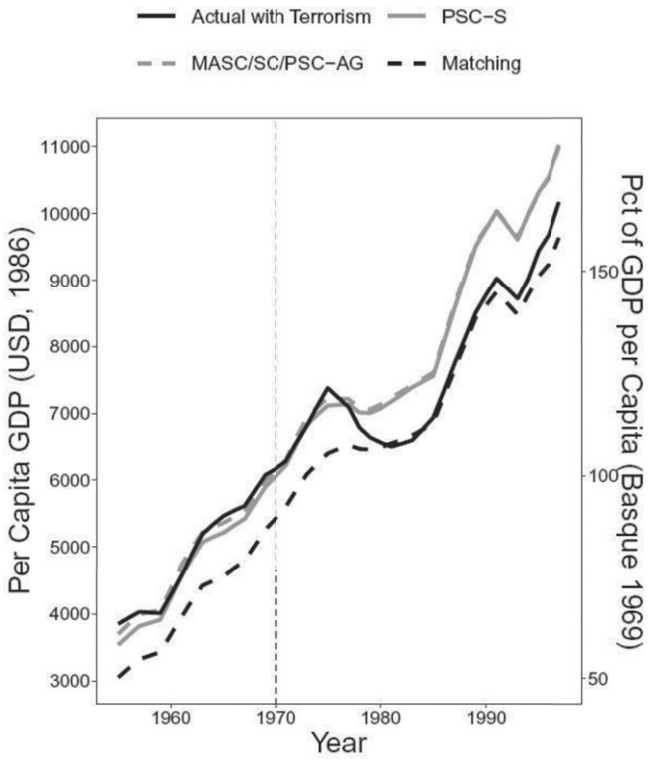


Figure 6. Actual and counterfactual per capita GDP of the Basque Country.
Note: The MASC selects $\phi = 0$ for the Basque Country, so MASC and SC imply the same counterfactual or the Basque Country. The PSC-AG selects a very small value $\pi > 0$, so that its counterfactual differs to a visually indistinguishable extent from SC. GDP per capita is measured in 1986 U.S. dollars.

Country) as a consequence. Its counterfactual therefore differs slightly from the SC, MASC, and PSC-AG. The matching estimator selects $m = 2$ by cross-validation, differing substantially from the other estimators by placing equal weight on Catalonia and Cantabria.

The estimator fits the actual path of per capita GDP in the pre-period fairly closely, with a RMSE of \$94. In the post-period, it experiences a much smaller dip in GDP per capita than the Basque Country actually experienced from the mid-1970's onward. Consequently, the MASC and SC imply a cost of conflict of \$580 per person per year. The PSC-AG has a similar pre-period fit (an RMSE of \$98) and implied cost of conflict (\$572 per person per year) to MASC and SC. PSC-S has comparatively worse fit in the pre-period (an RMSE of \$166) and a smaller implied cost of conflict (\$532 per person per year). The matching estimator has much worse fit (an RMSE of \$782) and implies a *positive* effect of terrorism on per capita GDP of \$331 per person per year.

Appendix Table A.2 reports the results of the placebo test of Abadie, Diamond, and Hainmueller (2010), comparing the SC and MASC estimators against their behaviors in the placebo regions studied in Section 4.

6. Conclusion

One of the major impacts of the SC method has been to recast longitudinal comparative case studies as prediction problems. In this article, we made use of two tools from the machine learning and economic forecasting literature: model averaging and rolling-origin forecast evaluation. By examining the weakness of the SC estimator to interpolation bias, and the weakness of the matching estimator to extrapolation bias, we showed how to use these tools to build a third estimator, the MASC, that is able to effectively avoid both sources of bias. Using both simulated and empirical placebo studies, we found evidence that MASC has lower MSPE than either the matching, SC, or penalized SC estimators. We applied all these estimators to re–examine Abadie and Gardeazabal's (2003) application to the economic costs of conflict in the Basque Country.

We have not discussed in detail the delicate issue of statistical inference in comparative case studies. A variety of inferential methods have been recently proposed for SC and related methods. Li (2019) and Chernozhukov, Wuthrich, and Zhu (2020) develop asymptotic methods that depend on having many pre– and/or post– periods, while Arkhangelsky et al. (2019) propose a jackknife under additional asymptotics in the number of untreated units. Abadie, Diamond, and Hainmueller (2010, 2015), Ferman and Pinto (2017), Firpo and Possebom (2018), Chernozhukov, Wuthrich, and Zhu (2019) and Shaikh and Toulis (2019) develop different types of non-asymptotic randomization tests, while Cattaneo, Feng, and Titiunik (2019) showed how to construct prediction intervals using non-asymptotic bounds.

Supplementary Materials

The supplementary material contains additional results and discussion for the placebo analyses and Monte Carlo simulations.

Funding

Maxwell Kellogg research supported by funding from the National Institute on Aging (T32AG000243). Alexander Torgovitsky research supported by National Science Foundation grant SES-1846832.

References

Abadie, A. (2020), "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," *Journal of Economic Literature*. [1805,1807,1810]

Abadie, A., Diamond, A., and Hainmueller, J. (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of Californias Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505. [1804,1805,1807,1815]

——— (2015), "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59, 495–510. [1804,1805,1807,1810,1811,1815]

Abadie, A., and Gardeazabal, J. (2003), "The Economic Costs of Conflict: A Case Study of the Basque Country," *The American Economic Review*, 93, 113–132. [1804,1805,1807,1811,1812,1814,1815]

Abadie, A., and Imbens, G. W. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267. [1808]

Abadie, A., and L'Hour, J. (2020), "A Penalized Synthetic Control Estimator for Disaggregated Data," Working paper. [1804,1805,1807,1809,1810,1811,1812,1813]

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2019), "Synthetic Difference in Differences," arXiv:1812.09970 [stat]. [1815]

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2018), "Matrix Completion Methods for Causal Panel Data Models," Working Paper No. 25132, National Bureau of Economic Research. [1805]

Athey, S., Bayati, M., Imbens, G., and Qu, Z. (2019), "Ensemble Methods for Causal Effects in Panel Data Settings," *AEA Papers and Proceedings*, 109, pp. 65–70. [1805]

Barnow, B. S., Cain, G. G., Goldberger, A. S. (1980), *Issues in the Analysis of Selectivity Bias*, University of Wisconsin, Inst. for Research on Poverty. [1805]

Becker, M., and Klößner, S. (2017), "Estimating the Economic Costs of Organized Crime by Synthetic Control Methods," *Journal of Applied Econometrics*, 32, 1367–1369. [1811]

——— (2018), "Fast and Reliable Computation of Generalized Synthetic Controls," *Econometrics and Statistics*, 5, 1–19. [1811]

Ben-Michael, E., Feller, A., and Rothstein, J. (2019), "The Augmented Synthetic Control Method," arXiv:1811.04170 [econ, stat]. [1808]

Bergmeir, C., and Benítez, J. M. (2012), "On the Use of Cross-Validation for Time Series Predictor Evaluation," *Information Sciences*, 191, 192–213. [1811]

Breiman, L. (1996), "Stacked Regressions," *Machine Learning*, 24, 49–64. [1804,1805,1808]

Cattaneo, M. D., Feng, Y., and Titiunik, R. (2019), "Prediction Intervals for Synthetic Control Methods," arXiv:1912.07120 [econ, stat]. [1815]

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2019), "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," arXiv:1712.09089 [econ, stat]. [1815]

——— (2020), "Practical and Robust t-Test Based Inference for Synthetic Control and Related Methods," arXiv:1812.10820 [econ]. [1805,1815]

Cover, T. (1968), "Estimation by the Nearest Neighbor Rule," *IEEE Transactions on Information Theory*, 14, 50–55. [1808]

Dehejia, R. H., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062. [1808]

Doudchenko, N., and Imbens, G. W. (2016), "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis," Working Paper 22791, National Bureau of Economic Research. [1805,1807]

Elliot, G. (2011), "Averaging and the Optimal Combination of Forecasts," Working Paper. [1805]

Fan, J., and Gijbels, I. (1992), "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics*, 20, 2008–2036. [1808]

Ferman, B. (2020), "On the Properties of the Synthetic Control Estimator With Many Periods and Many Controls," arXiv:1906.06665 [econ]. [1807]

Ferman, B., and Pinto, C. (2017), "Revisiting the Synthetic Control Estimator," Working paper. [1815]

——— (2019), "Synthetic Controls with Imperfect Pre-Treatment Fit," arXiv:1911.08521 [econ]. [1808]

Firpo, S., and Possebom, V. (2018), "Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets," *Journal of Causal Inference*, 6. [1815]

Hansen, B. E., and Racine, J. S. (2012), "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46. [1804]

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), Springer Series in Statistics, New York: Springer. [1811]

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997), "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, 64, 605–654. [1805]

——— (1998), "Matching as an Econometric Evaluation Estimator," *The Review of Economic Studies*, 65, 261–294. [1805]

Heckman, J. J., and Robb, R. (1985), "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, eds. J. J. Heckman and B. Singer, Cambridge: Cambridge University Press. [1805]

Imbens, G. W. (2004), "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29. [1805,1807]

——— (2015), "Matching Methods in Practice: Three Examples," *Journal of Human Resources*, 50, 373–419. [1805]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge: Cambridge University Press. [1805]

Li, K. T. (2019), "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods," *Journal of the American Statistical Association*, 1–16. [1815]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [1805]

Shaikh, A., and Toulis, P. (2019), "Randomization Tests in Observational Studies with Staggered Adoption of Treatment," arXiv:1912.10610 [stat]. [1815]

Stock, J. H., and Watson, M. W. (2004), "Combination Forecasts of Output Growth in a Seven-Country Data Set," *Journal of Forecasting*, 23, 405–430. [1805]

——— (2006), "Chapter 10 Forecasting With Many Predictors," in *Handbook of Economic Forecasting* (Vol. 1), eds. by G. Elliott, C. W. J. Granger, and A. Timmermann, The Netherlands: Elsevier, 515–554. [1805]

Swanson, N. R., and White, H. (1997), "Forecasting Economic Time Series Using Flexible Versus Fixed Specification and Linear versus Nonlinear Econometric Models," *International Journal of Forecasting*, 13, 439–461. [1811]

Tashman, L. J. (2000), "Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review," *International Journal of Forecasting*, 16, 437–450. [1804,1811]

Viviano, D., and Bradic, J. (2019), "Synthetic Learner: Model-Free Inference on Treatments Over Time," arXiv:1904.01490 [cs, econ, stat]. [1805]

Wolpert, D. H. (1992), "Stacked Generalization," *Neural Networks*, 5, 241–259. [1804]