Journal of the American Statistical Association

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

The Augmented Synthetic Control Method

Eli Ben-Michael, Avi Feller & Jesse Rothstein

To cite this article: Eli Ben-Michael, Avi Feller & Jesse Rothstein (2021) The Augmented Synthetic Control Method, Journal of the American Statistical Association, 116:536, 1789-1803, DOI: 10.1080/01621459.2021.1929245

To link to this article: https://doi.org/10.1080/01621459.2021.1929245

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 536, 1789–1803: Theory and Methods Special Section on Synthetic Control Methods
https://doi.org/10.1080/01621459.2021.1929245

The Augmented Synthetic Control Method

Eli Ben-Michael[a], Avi Feller[b], and Jesse Rothstein[c]

[a]Department of Statistics, Institute for Quantitative Social Sciences, Harvard University, Cambridge, MA; [b]Department of Statistics, Goldman School of Public Policy, University of California, Berkeley, CA; [c]Department of Economics, Goldman School of Public Policy, University of California, Berkeley, CA

ABSTRACT

The synthetic control method (SCM) is a popular approach for estimating the impact of a treatment on a single unit in panel data settings. The "synthetic control" is a weighted average of control units that balances the treated unit's pretreatment outcomes and other covariates as closely as possible. A critical feature of the original proposal is to use SCM only when the fit on pretreatment outcomes is excellent. We propose Augmented SCM as an extension of SCM to settings where such pretreatment fit is infeasible. Analogous to bias correction for inexact matching, augmented SCM uses an outcome model to estimate the bias due to imperfect pretreatment fit and then de-biases the original SCM estimate. Our main proposal, which uses ridge regression as the outcome model, directly controls pretreatment fit while minimizing extrapolation from the convex hull. This estimator can also be expressed as a solution to a modified synthetic controls problem that allows negative weights on some donor units. We bound the estimation error of this approach under different data-generating processes, including a linear factor model, and show how regularization helps to avoid over-fitting to noise. We demonstrate gains from Augmented SCM with extensive simulation studies and apply this framework to estimate the impact of the 2012 Kansas tax cuts on economic growth. We implement the proposed method in the new `augsynth` R package.

ARTICLE HISTORY
Received November 2019
Accepted May 2021

KEYWORDS
Bias correction; Causal inference; Panel data; Synthetic control

1. Introduction

The *synthetic control method* (SCM) is a popular approach for estimating the impact of a treatment on a single unit in panel data settings with a modest number of control units and with many pretreatment periods (Abadie and Gardeazabal 2003; Abadie, Diamond, and Hainmueller 2010, 2015). The idea is to construct a weighted average of control units, known as a synthetic control, that matches the treated unit's pretreatment outcomes. The estimated impact is then the difference in posttreatment outcomes between the treated unit and the synthetic control. SCM has been widely applied—the main SCM papers have over 4000 citations — and has been called "arguably the most important innovation in the policy evaluation literature in the last 15 years" (Athey and Imbens 2017).

A critical feature of the original proposal, not always followed in practice, is to use SCM only when the synthetic control's pretreatment outcomes closely match the pretreatment outcomes for the treated unit (Abadie, Diamond, and Hainmueller 2015). When it is not possible to construct a synthetic control that fits pretreatment outcomes well, the original articles advise against using SCM. At that point, researchers often fall back to linear regression. This allows better (often perfect) pretreatment fit, but does so by applying negative weights to some control units, extrapolating outside the support of the data.

We propose the *augmented synthetic control method* (ASCM) as a middle ground in settings where excellent pretreatment fit using SCM alone is not feasible. Analogous to bias correction for inexact matching (Abadie and Imbens 2011), ASCM begins with the original SCM estimate, uses an outcome model to estimate the bias due to imperfect pretreatment fit, and then uses this to de-bias the SCM estimate. If pretreatment fit is good, then the estimated bias will be small, and the SCM and ASCM estimates will be similar. Otherwise, the estimates will diverge, and ASCM will rely more heavily on extrapolation.

Our primary proposal is to augment SCM with a ridge regression model, which we call *Ridge ASCM*. We show that, like SCM, the Ridge ASCM estimator can be written as a weighted average of the control unit outcomes. We also show that Ridge ASCM weights can be written as the solution to a modified synthetic controls problem, targeting the same imbalance metric as traditional SCM. However, where SCM weights are always nonnegative, Ridge ASCM admits negative weights, using extrapolation to improve pretreatment fit. The regularization parameter in Ridge ASCM directly parameterizes the level of extrapolation by penalizing the distance from SCM weights. By contrast, (ridge) regression alone, which can also be written as a modified synthetic controls problem with possibly negative weights, allows for arbitrary extrapolation and possibly unchecked extrapolation bias.

We relate Ridge ASCMs, improved pretreatment fit to a finite sample bound on estimation error under several data-generating processes (DGPs), including an autoregressive model and the linear factor model often invoked in this setting

CONTACT Avi Feller afeller@berkeley.edu Department of Statistics, Goldman School of Public Policy, University of California, Berkeley, 309 GSPP Main, 2607 Hearst Avenue, Berkeley, CA 94720.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

(Abadie, Diamond, and Hainmueller 2010). Under an autoregressive model, improving pretreatment fit directly reduces bias, and the Ridge ASCM penalty term negotiates a bias-variance trade-off. Under a latent factor model, improving pretreatment fit again reduces bias, though there is now a risk of over-fitting, and the penalty term again directly parameterizes this trade-off. Thus, choosing the hyperparameter will be important for practice; we propose a cross-validation procedure in Section 5.3.

Finally, we describe how the Augmented SCM approach can be extended to incorporate auxiliary covariates other than pretreatment outcomes. We first propose to include the auxiliary covariates in parallel to the lagged outcomes in both the SCM and outcome models. We also propose an alternative when there are relatively few covariates, extending a suggestion from Doudchenko and Imbens (2017): first residualize pre- and posttreatment outcomes against the auxiliary covariates, then fit Ridge ASCM on the residualized outcome series. We show that this controls the estimation error under a linear factor model with auxiliary covariates.

An important question in practice is when to prefer Augmented SCM to SCM alone. We recommend making this decision based on the estimated bias, the computation of which is the first step of implementing the ASCM estimator. If the estimated bias—the difference between the outcome model's fitted values for the treated unit and the synthetic control—is large, then it is worth trading off bias reduction from ASCM for some extrapolation, which the researcher can also assess directly. Since the estimated bias is in the same units as the estimand of interest, researchers can assess what constitutes "large" bias based on context.

We demonstrate the properties of Augmented SCM via both calibrated simulation studies and by using it to examine the effect of an aggressive tax cut in Kansas in 2012 on economic output, finding a substantial negative effect. Overall, we see large gains from ASCM relative to alternative estimators, especially under model mis-specification, in terms of both bias and root mean squared error (RMSE). We implement the proposed methodology in the `augsynth` package for R, available at *https://github.com/ebenmichael/augsynth*.

The article proceeds as follows. Section 1.1 briefly reviews related work. Section 2 introduces notation, the underlying models and assumptions, and the SCM estimator. Section 3 gives an overview of Augmented SCM. Section 4 gives key algorithmic results for Ridge ASCM. Section 5 bounds the Ridge ASCM estimation error under a linear model and under a linear factor model, the standard setting for SCM, and also addresses inference. Section 6 extends the ASCM framework to incorporate auxiliary covariates. Section 7 reports on extensive simulation studies as well as the application to the Kansas tax cuts. Finally, Section 8 discusses some possible directions for further research. The supplementary material includes all of the proofs, as well as additional derivations and technical discussion.

1.1. Related Work

SCM was introduced by Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010, 2015) and is the subject of an extensive methodological literature; see Abadie (2021) and Samartsidis et al. (2019) for recent reviews. We briefly highlight some relevant aspects of this literature.

A group of articles adapts the original SCM proposal to allow for more robust estimation while retaining SCM's simplex constraint on the weights. Robbins, Saunders, and Kilmer (2017), Doudchenko and Imbens (2017), and Abadie and L'Hour (2018) incorporated a penalty on the weights into the SCM optimization problem, building on a suggestion in Abadie, Diamond, and Hainmueller (2015). Gobillon and Magnac (2016) explored dimension reduction strategies and other data transformations that can improve the performance of the subsequent estimator.

A second set of articles relaxes constraints imposed in the original SCM problem, in particular the restriction that control unit weights be nonnegative. Doudchenko and Imbens (2017) argued that there are many settings in which negative weights would be desirable. Amjad, Shah, and Shen (2018) proposed an interesting variant that combines negative weights with a preprocessing step. Powell (2018) instead allowed for extrapolation via a Frisch-Waugh-Lovell-style projection, which similarly generalizes the typical SCM setting. Doudchenko and Imbens (2017) and Ferman and Pinto (2018) both proposed to incorporate an intercept into the SCM problem, which we discuss in Section 3.2.

There have also been several other proposals to reduce bias in SCM, developed independently and contemporaneously with ours. Abadie and L'Hour (2018) also proposed bias correcting SCM using regression. Kellogg et al. (2020) proposed using a weighted average of SCM and matching, trade-off interpolation and extrapolation bias. Arkhangelsky et al. (2019) proposed the *synthetic difference-in-differences* estimator, which is similar to a version of our proposal with a constrained outcome regression.

Finally, there have also been recent proposals to use outcome modeling rather than SCM-style weighting in this setting. These include the matrix completion method in Athey et al. (2017), the generalized synthetic control method in Xu (2017), and the combined approaches in Hsiao et al. (2018). We explore the performance of select methods, both in isolation and within our ASCM framework, in Section 7.1.

2. Overview of the SCM

2.1. Notation and Setup

We consider the canonical SCM panel data setting with $i = 1, \dots, N$ units observed for $t = 1, \dots, T$ time periods; for the theoretical discussion below, we will consider both N and T to be fixed. Let W_i be an indicator that unit i is treated at time $T_0 < T$ where units with $W_i = 0$ never receive the treatment. We restrict our attention to the case where a single unit receives treatment, and follow the convention that this is the first one, $W_1 = 1$; see Ben-Michael, Feller, and Rothstein (2019) for an extension to multiple treated units. The remaining $N_0 = N - 1$ units are possible controls, often referred to as *donor units* in the SCM context. To simplify notation, we limit to one posttreatment observation, $T = T_0 + 1$, though our results are easily extended to larger T.

We adopt the potential outcomes framework (Neyman 1923) and invoke SUTVA, which assumes a well-defined treatment and excludes interference between units; the potential outcomes for unit i in period t under control and treatment are $Y_{it}(0)$ and $Y_{it}(1)$, respectively. We define the treated potential outcome as $Y_{it}(1) = Y_{it}(0) + \tau_{it}$, where the treatment effects τ_{it} are fixed parameters. Since the first unit is treated, the key estimand of interest is $\tau = \tau_{1T} = Y_{1T}(1) - Y_{1T}(0)$. Finally, the observed outcomes are

$$Y_{it} = \begin{cases} Y_{it}(0) & \text{if } W_i = 0 \text{ or } t \leq T_0, \\ Y_{it}(1) & \text{if } W_i = 1 \text{ and } t > T_0. \end{cases} \tag{1}$$

To emphasize that pretreatment outcomes serve as covariates in SCM, we use X_{it}, for $t \leq T_0$, to represent pretreatment outcomes; we use the terms *pretreatment fit* and *covariate balance* interchangeably. With some abuse of notation, we use $\boldsymbol{X}_{0\cdot}$ to represent the N_0-by-T_0 matrix of control unit pretreatment outcomes and $\boldsymbol{Y}_{0T}$ for the N_0-vector of control unit outcomes in period T. With only one treated unit, Y_{1T} is a scalar, and $\boldsymbol{X}_{1\cdot}$ is a T_0-row vector of treated unit pretreatment outcomes. The data structure is then

$$\begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1T_0} & Y_{1T} \\ Y_{21} & Y_{22} & \dots & Y_{2T_0} & Y_{2T} \\ \vdots & & & & \vdots \\ Y_{N1} & Y_{N2} & \dots & Y_{NT_0} & Y_{NT} \end{pmatrix} \equiv \left(\begin{array}{c|c} \underbrace{\begin{matrix} X_{11} & X_{12} & \dots & X_{1T_0} \\ X_{21} & X_{22} & \dots & X_{2T_0} \\ & \vdots & \vdots & \\ X_{N1} & X_{N2} & \dots & X_{NT_0} \end{matrix}}_{\text{pretreatment outcomes}} & \begin{matrix} Y_{1T} \\ Y_{2T} \\ \\ Y_{NT} \end{matrix} \end{array} \right) \equiv \left(\begin{array}{c|c} \boldsymbol{X}_{1\cdot} & Y_{1T} \\ \boldsymbol{X}_{0\cdot} & \boldsymbol{Y}_{0T} \end{array} \right). \tag{2}$$

2.2. Assumptions on the DGP

We now give assumptions on the underlying DGPs for the control potential outcomes. We separate control potential outcomes (before and after T_0) into a model component m_{it} plus an additive noise term $\varepsilon_{it} \sim P(\cdot)$:

$$Y_{it}(0) = m_{it} + \varepsilon_{it}, \tag{3}$$

where epsilon is sub-Gaussian with scale parameter σ. This setup encompasses many common panel data models; see Chernozhukov, Wüthrich, and Zhu (2019) for an extended discussion. Here, we consider two specific versions of Equation (3): (a) for posttreatment time T, $Y_{iT}(0)$ is linear in its lagged values; and (b) for all $t = 1, \dots, T$, $Y_{it}(0)$ is linear in a set of latent factors. In the supplementary material, we also consider the case where m_{it} is a linear model with Lipshitz deviations.

Assumption 1 (Model component). The control potential outcomes are generated according to the following model and error components:

(a) For time period T, the model components m_{iT} are generated as $\sum_{\ell=1}^{T_0} \beta_\ell Y_{i(t-\ell)}(0)$, so the control potential outcomes $Y_{it}(0)$ are

$$Y_{it}(0) = \sum_{\ell=1}^{T_0} \beta_\ell Y_{i(t-\ell)}(0) + \varepsilon_{it}, \tag{4}$$

where $\{\varepsilon_{iT}\}$ have zero mean for each unit:

$$\mathbb{E}\left[\varepsilon_{iT}\right] = 0 \quad \forall i = 1, \dots, N. \tag{5}$$

(b) There are J unknown, latent time-varying factors at time $t = 1, \dots, T$, $\boldsymbol{\mu}_t = \{\mu_{jt}\} \in \mathbb{R}^J$, with $\max_{jt} |\mu_{jt}| \leq M$, and each unit has a vector of unknown factor loadings $\boldsymbol{\phi}_i \in \mathbb{R}^J$. We collect the pre-intervention factors into a matrix $\boldsymbol{\mu} \in \mathbb{R}^{T_0 \times J}$, where the t^{th} row of $\boldsymbol{\mu}$ contains the factor values at time t, $\boldsymbol{\mu}_t'$ and assume that $\frac{1}{T_0}\boldsymbol{\mu}'\boldsymbol{\mu} = \boldsymbol{I}_J$. The model components m_{it} are generated as $m_{it} = \boldsymbol{\phi}_i \cdot \boldsymbol{\mu}_t$, so the control potential outcomes $Y_{it}(0)$ are generated as follows:

$$Y_{it}(0) = \boldsymbol{\phi}_i \cdot \boldsymbol{\mu}_t + \varepsilon_{it} = \sum_{j=1}^{J} \phi_{ij}\mu_{jt} + \varepsilon_{it}, \tag{6}$$

where the noise terms for all units and all periods have zero mean:

$$\mathbb{E}\left[\varepsilon_{it}\right] = 0 \quad \forall i = 1, \dots, N \text{ and } \forall t = 1, \dots, T. \tag{7}$$

We consider both the time-varying factors $\boldsymbol{\mu}_t$ and the unit-varying factor loadings $\boldsymbol{\phi}_i$ to be nonrandom quantities, so the randomness in $Y_{it}(0)$ is only due to the noise term ε_{it}.

Assumptions 1(a) and (b) enable estimation of the missing counterfactual outcome. In Assumption 1(a), the mean-zero noise restrictions hold for the treated unit ($i = 1$), and rule out any unmeasured variables that are correlated with the outcomes and that have different distributions for the treated unit and comparison units. Treatment assignment can depend on the past outcomes, but cannot depend on posttreatment outcomes; furthermore, there cannot be serial correlation between the posttreatment and pretreatment noise. This DGP includes the special case of an autoregressive process of order $K < T_0$. Assumption 1(b) allows for the existence of unmeasured confounders, the factor loadings, that enter into the DGP in a structured way. Treatment assignment can depend on the factor loadings, but cannot depend on the *realized* pretreatment outcomes. We discuss this in more detail in the context of our application in Section 7.

2.3. Synthetic Control Method

The SCM imputes the missing potential outcome for the treated unit, $Y_{1T}(0)$, as a weighted average of the control outcomes, $\boldsymbol{Y}_{0T}'\boldsymbol{\gamma}$ (Abadie and Gardeazabal 2003; Abadie, Diamond, and Hainmueller 2010, 2015). Weights are chosen to balance pretreatment outcomes and possibly other covariates. We consider a version of SCM that chooses weights $\boldsymbol{\gamma}$ as a solution to the constrained optimization problem:

$$\begin{aligned} &\min_{\gamma} \; ||\boldsymbol{V_x}^{1/2}(\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma})||_2^2 \; + \; \zeta \sum_{W_i=0} f(\gamma_i) \\ &\text{subject to} \quad \sum_{W_i=0} \gamma_i = 1 \\ &\qquad\qquad\quad \gamma_i \geq 0 \quad i : W_i = 0 \end{aligned} \tag{8}$$

where the constraints limit $\boldsymbol{\gamma}$ to the simplex $\Delta^{N_0} = \{\boldsymbol{\gamma} \in \mathbb{R}^{N_0} | \gamma_i \geq 0 \; \forall i, \; \sum_i \gamma_i = 1\}$, where $\boldsymbol{V_x} \in \mathbb{R}^{T_0 \times T_0}$ is a symmetric importance matrix and $||\boldsymbol{V_x}^{1/2}(\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\gamma)||_2^2 \equiv$

$(\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\gamma)'\boldsymbol{V_x}(\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma})$ is the 2-norm on $\mathbb{R}^{T_0}$ after applying $\boldsymbol{V_x}^{1/2}$ as a linear transformation, and where $f(\gamma_i)$ is a dispersion penalty on the weights that we discuss below. To simplify the exposition and notation below, we will generally take $\boldsymbol{V_x}$ to be the identity matrix. The simplex constraint in Equation (8) ensures that the weights will be sparse and nonnegative; Abadie, Diamond, and Hainmueller (2010, 2015) argued that enforcing this constraint is important for preserving interpretability.

Equation (8) modifies the original SCM proposal in two ways. First, Equation (8) penalizes the dispersion of the weights with hyperparameter $\zeta \geq 0$, following a suggestion in Abadie, Diamond, and Hainmueller (2015). The choice of penalty is less central when weights is constrained to be on the simplex, but becomes more important below when we relax this constraint (Doudchenko and Imbens 2017). Second, Equation (8) excludes auxiliary covariates; we re-introduce them in Section 6.

When the treated unit's vector of lagged outcomes, $\boldsymbol{X}_{1\cdot}$, is inside the convex hull of the control units' lagged outcomes, $\boldsymbol{X}_{0\cdot}$, the SCM weights in Equation (8) achieve perfect pretreatment fit, and the resulting estimator has many attractive properties. In this setting, Abadie, Diamond, and Hainmueller (2010) showed that SCM will be unbiased under the autoregressive model in Assumption 1(a) and bound the bias under the linear factor model in Assumption 1(b).

Due to the curse of dimensionality, however, achieving perfect (or nearly perfect) pretreatment fit is not always feasible with weights constrained to be on the simplex (see Ferman and Pinto 2018). When "the pretreatment fit is poor or the number of pretreatment periods is small," Abadie, Diamond, and Hainmueller (2015) recommended against using SCM. And even if the pretreatment fit is excellent, Abadie, Diamond, and Hainmueller (2010, 2015) proposed extensive placebo checks to ensure that SCM weights do not overfit to noise. Thus, the conditional nature of the analysis is critical to deploying SCM, excluding many practical settings. Our proposal enables the use of (a modified) SCM approach in many of the cases where SCM alone is infeasible.

3. Augmented SCM

3.1. Overview

We now show how to modify the SCM approach to adjust for poor pretreatment fit. Let $\hat{m}_{iT}$ be an estimator for m_{iT}, the model component of the posttreatment control potential outcome. The *Augmented SCM* (ASCM) estimator for $Y_{1T}(0)$ is

$$\hat{Y}_{1T}^{\text{aug}}(0) = \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} Y_{iT} + \left(\hat{m}_{1T} - \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} \hat{m}_{iT}\right) \tag{9}$$

$$= \hat{m}_{1T} + \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} (Y_{iT} - \hat{m}_{iT}), \tag{10}$$

where weights $\hat{\gamma}_i^{\text{scm}}$ are the SCM weights defined above. Standard SCM is a special case, where $\hat{m}_{iT}$ is a constant. We will largely focus on estimators that are functions of pretreatment outcomes, $\hat{m}_{iT} \equiv \hat{m}(\boldsymbol{X}_i)$, where $\hat{m} : \mathbb{R}^{T_0} \to \mathbb{R}$.

Equations (9) and (10), while equivalent, highlight two distinct motivations for ASCM. Equation (9) directly corrects the SCM estimate, $\sum \hat{\gamma}_i^{\text{scm}} Y_{iT}$, by the imbalance in a particular function of the pretreatment outcomes $\hat{m}(\cdot)$. Intuitively, since $\hat{m}$ estimates the posttreatment outcome, we can view this as an estimate of the bias due to imbalance, analogous to bias correction for inexact matching (Abadie and Imbens 2011). In this form, we can see that SCM and ASCM estimates will be similar if the estimated bias is small, as measured by imbalance in $\hat{m}(\cdot)$. If the estimated bias is large, the two estimators will diverge, and the conditions for appropriate use of SCM will not apply. In independent work, Abadie and L'Hour (2018) also considered a bias-corrected estimator of this form.

Equation (10), by contrast, is analogous to standard doubly robust estimation (Robins, Rotnitzky, and Zhao 1994), which begins with the outcome model but then re-weights to balance residuals. We discuss connections to inverse propensity score weighting and survey calibration in Appendix E in the supplementary material.

3.2. Choice of Estimator

While this setup is general, the choice of estimator $\hat{m}$ is important both for understanding the procedure's properties and for practical performance. We give a brief overview of two special cases: (i) when $\hat{m}$ is linear in pretreatment outcomes; and (ii) when $\hat{m}$ is linear in the comparison units' posttreatment outcomes. Ridge regression is an important example that is linear in both; we explore this estimator further in Sections 4 and 5.

First, consider an estimator that is linear in pretreatment outcomes, $\hat{m}(\boldsymbol{X}) = \hat{\eta}_0 + \hat{\boldsymbol{\eta}} \cdot \boldsymbol{X}$. The augmented estimator (9) is then

$$\hat{Y}_{1T}^{\text{aug}}(0) = \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} Y_{iT} + \sum_{t=1}^{T_0} \hat{\eta}_t \left(X_{1t} - \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} X_{it}\right). \tag{11}$$

Pretreatment periods that are more predictive of the posttreatment outcome will have larger (absolute) regression coefficients and so imbalance in these periods will lead to a larger adjustment. Thus, even if we do not a priori prioritize balance in any particular pretreatment time periods (via the choice of $\boldsymbol{V_x}$), the linear model augmentation will adjust for the time periods that are empirically more predictive of the posttreatment outcome. As we show in Section 4, the ridge-regularized linear model is an important special case in which the resulting augmented estimator is itself a penalized synthetic control estimator. This allows for a more direct analysis of the role of bias correction.

Second, consider an estimator that is a linear combination of comparison units' posttreatment outcomes, $\hat{m}(\boldsymbol{X}) = \sum_{W_j=0} \hat{\alpha}_i(\boldsymbol{X}) Y_{iT}$, for some weighting function $\hat{\alpha} : \mathbb{R}^{T_0} \to \mathbb{R}^{N_0}$. Examples include k-nearest neighbor matching and kernel weighting as well as other "vertical" regression approaches (Athey et al. 2017). The augmented estimator (9) is itself a weighting estimator that adjusts the SCM weights

$$\hat{Y}_{1T}^{\text{aug}}(0) = \sum_{W_i=0} \left(\hat{\gamma}_i^{\text{scm}} + \hat{\gamma}_i^{\text{adj}}\right) Y_{iT}, \quad \text{where}$$

$$\hat{\gamma}_i^{\text{adj}} \equiv \hat{\alpha}_i(\boldsymbol{X}_1) - \sum_{W_j=0} \hat{\gamma}_j^{\text{scm}} \hat{\alpha}_i(\boldsymbol{X}_j). \tag{12}$$

Here, the adjustment term for unit i, $\hat{\gamma}_i^{\text{adj}}$, is the imbalance in a unit i-specific transformation of the lagged outcomes that depends on the weighting function $\alpha(\cdot)$. While $\hat{\gamma}^{\text{scm}}$ are constrained to be on the simplex, the form of $\hat{\gamma}^{\text{adj}}$ makes clear that the overall weights can be negative.

There are many special cases to consider. One is the linear-in-lagged-outcomes model with equal coefficients, $\hat{\eta}_t = \frac{1}{T_0}$, which estimates a fixed-effects outcome model as $\hat{m}(\boldsymbol{X}_i) = \bar{X}_i$. The corresponding treatment effect estimate adjusts for imbalance in all pretreatment time periods equally, and yields a weighted difference-in-differences estimator

$$\hat{\tau}^{\text{de}} = \left(Y_{1T} - \bar{X}_1\right) - \left(\sum_{W_i=0} \hat{\gamma}_i (Y_{iT} - \bar{X}_i)\right)$$
$$= \frac{1}{T_0} \sum_{t=1}^{T_0} \left[(Y_{1T} - X_{1t}) - \left(\sum_{W_i=0} \hat{\gamma}_i (Y_{iT} - X_{it})\right)\right]. \quad (13)$$

An augmented estimator of this form has appeared as the *de-meaned* or *intercept shift SCM* (Doudchenko and Imbens 2017; Ferman and Pinto 2018). As we discuss in Section 6, these proposals balance the residual outcomes $X_{it} - \bar{X}_i$ rather than the raw outcomes X_{it}. See also Arkhangelsky et al. (2019), who extended this to weight across both units and time.

In Section 7.1 we conduct a simulation study to inspect the performance of a range of estimators including other penalized linear models, such as the LASSO; flexible machine learning models, such as random forests; and panel data methods, such as fixed-effects models and low-rank matrix completion methods (Xu 2017; Athey et al. 2017).

4. Ridge ASCM

We now inspect the algorithmic properties for the special case where $\hat{m}(\boldsymbol{X}_i)$ is estimated via a ridge-regularized linear model, which we refer to as *Ridge Augmented SCM* (Ridge ASCM). With Ridge ASCM, the estimator for the posttreatment outcome is $\hat{m}(\boldsymbol{X}_i) = \hat{\eta}_0^{\text{ridge}} + \boldsymbol{X}_i'\hat{\boldsymbol{\eta}}^{\text{ridge}}$, where $\hat{\eta}_0^{\text{ridge}}$ and $\hat{\boldsymbol{\eta}}^{\text{ridge}}$ are the coefficients of a ridge regression of control posttreatment outcomes $\boldsymbol{Y}_{0T}$ on centered pretreatment outcomes $\boldsymbol{X}_{0\cdot}$ with penalty hyperparameter λ^{ridge}

$$\left\{\hat{\eta}_0^{\text{ridge}}, \hat{\boldsymbol{\eta}}^{\text{ridge}}\right\} = \arg\min_{\eta_0, \boldsymbol{\eta}} \; \frac{1}{2} \sum_{W_i=0} (Y_i - (\eta_0 + X_i'\boldsymbol{\eta}))^2$$
$$+ \lambda^{\text{ridge}} ||\boldsymbol{\eta}||_2^2. \quad (14)$$

The Ridge Augmented SCM estimator is then

$$\hat{Y}_{1T}^{\text{aug}}(0) = \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} Y_{iT} + \left(\boldsymbol{X}_1 - \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} \boldsymbol{X}_{i\cdot}\right) \cdot \hat{\boldsymbol{\eta}}^{\text{ridge}}. \quad (15)$$

We first show that Ridge ASCM is a linear weighting estimator as in Equation (12). Unlike augmenting with other linear weighting estimators, when augmenting with ridge regression the implied weights are themselves the solution to a penalized synthetic control problem, as in Equation (8). Using this representation, we show that when the treated unit lies outside the convex hull of the control units, Ridge ASCM improves the pretreatment fit relative to SCM alone by allowing for negative weights and extrapolating away from the convex hull. We also show that ridge regression alone has a representation as a weighting estimator that allows for negative weights.

Allowing for negative weights is an important departure from the original SCM proposal, which constrains weights to be on the simplex. In particular, ridge regression alone allows for arbitrarily negative weights and may have negative weights even when the treated unit is inside of the convex hull. By contrast, Ridge ASCM directly penalizes distance from the sparse, non-negative SCM weights, controlling the amount of extrapolation by the choice of λ^{ridge}, and only resorts to negative weights if the treated unit is outside of the convex hull.

4.1. Ridge ASCM as a Penalized SCM Estimator

We now express both Ridge ASCM and ridge regression alone as special cases of the penalized SCM problem in Equation (8). The Ridge ASCM estimate of the counterfactual is the solution to Equation (8), replacing the simplex constraint with a penalty $f(\gamma_i) = \left(\gamma_i - \hat{\gamma}_i^{\text{scm}}\right)^2$ that penalizes *deviations from the SCM weights.*

Lemma 1. The ridge-augmented SCM estimator (11) is

$$\hat{Y}_{1T}^{\text{aug}}(0) = \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} Y_{iT}, \quad (16)$$

where

$$\hat{\gamma}_i^{\text{aug}} = \hat{\gamma}_i^{\text{scm}} + (\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}})'(\boldsymbol{X}_{0\cdot}'\boldsymbol{X}_{0\cdot} + \lambda^{\text{ridge}}\boldsymbol{I}_{\boldsymbol{T_0}})^{-1}\boldsymbol{X}_{i\cdot}. \quad (17)$$

Moreover, the Ridge ASCM weights $\hat{\boldsymbol{\gamma}}^{\text{aug}}$ are the solution to

$$\min_{\boldsymbol{\gamma} \text{ s.t. } \sum_i \gamma_i = 1} \frac{1}{2\lambda^{\text{ridge}}} \left\| \boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma} \right\|_2^2 + \frac{1}{2} \left\| \boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}^{\text{scm}} \right\|_2^2. \quad (18)$$

When the treated unit is in the convex hull of the control units—so the SCM weights exactly balance the lagged outcomes—the Ridge ASCM and SCM weights are identical. When SCM weights do not achieve exact balance, the Ridge ASCM solution will use negative weights to extrapolate from the convex hull of the control units. The amount of extrapolation is determined both by the amount of imbalance and by the hyperparameter λ^{ridge}. When SCM yields good pretreatment fit or when λ^{ridge} is large, the adjustment term will be small and $\hat{\boldsymbol{\gamma}}^{\text{aug}}$ will remain close to the SCM weights.

We can similarly characterize ridge regression alone as a solution to a penalized SCM problem where the penalty term, $f(\gamma_i) = \left(\gamma_i - \frac{1}{N_0}\right)^2$, penalizes the variance of the weights. Other penalized linear models, such as the LASSO or elastic net, do not have this same representation as a penalized SCM estimator.

Lemma 2. The ridge regression estimator $\hat{Y}_{1T}^{\text{ridge}}(0) \equiv \hat{\eta}_0^{\text{ridge}} + \boldsymbol{X}_1 \cdot \hat{\boldsymbol{\eta}}^{\text{ridge}}$ can be written as $\hat{Y}_{1T}^{\text{ridge}}(0) = \sum_{W_i=0} \hat{\gamma}_i^{\text{ridge}} Y_{iT}$, where the ridge weights $\hat{\boldsymbol{\gamma}}^{\text{ridge}}$ are the solution to

$$\min_{\boldsymbol{\gamma} \mid \sum_i \gamma_i = 1} \frac{1}{2\lambda^{\text{ridge}}} \left\| \boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma} \right\|_2^2 + \frac{1}{2} \left\| \boldsymbol{\gamma} - \frac{1}{N_0} \right\|_2^2. \quad (19)$$

For ridge regression alone, the hyperparameter λ^{ridge} controls the variance of the weights rather than the degree of extrapolation from the simplex. Thus, in order to reduce variance, ridge regression weights might still be negative even if the treated unit is inside of the convex hull and SCM achieves perfect fit.

Figure 1 visualizes this behavior in two dimensions. Figure 1(a) shows the treated unit outside the convex hull of the control units, along with the weighted average of control units using ridge regression and Ridge ASCM weights. For large λ^{ridge}, ridge regression alone begins at the center of the control units (i.e., uniform weights), while Ridge ASCM begins at the SCM solution; both move smoothly toward an exact fit solution as λ^{ridge} is reduced. Figure 1(b) shows the distance from the simplex of these ridge regression and Ridge ASCM weights. Together these figures highlight that ridge regression weights can leave the simplex (i.e., have some negative weights) before the corresponding weighted average is outside of the convex hull, marked in red in both figures. That is, ridge regression weights use negative weights to minimize the variance although it is possible to achieve the same level of balance with nonnegative weights. By contrast, Ridge ASCM weights begin at the SCM solution, which is on the boundary of the simplex, then extrapolate outside the convex hull. Eventually, as $\lambda^{\text{ridge}} \to 0$, both ridge and Ridge ASCM use negative weights to achieve perfect balance, improving the fit relative to SCM alone. The weight vectors differ, however, with the Ridge ASCM weights closer to the simplex.

When achieving excellent pretreatment fit with SCM is possible, Abadie, Diamond, and Hainmueller (2015) argued that we should prefer SCM weights over possibly negative weights: a slight balance improvement is not worth the extrapolation and the loss of interpretability. In this case, the Ridge ASCM weights will be close to the simplex, while the ridge regression weights may be quite far away. When this is not possible, however, and SCM has poor fit, some degree of extrapolation is critical; Ridge ASCM allows the researcher to directly penalize the amount of extrapolation in these cases. See King and Zeng (2006) for a discussion of extrapolation in constructing counterfactuals.

4.2. Ridge ASCM Improves Pretreatment Fit Relative to SCM Alone

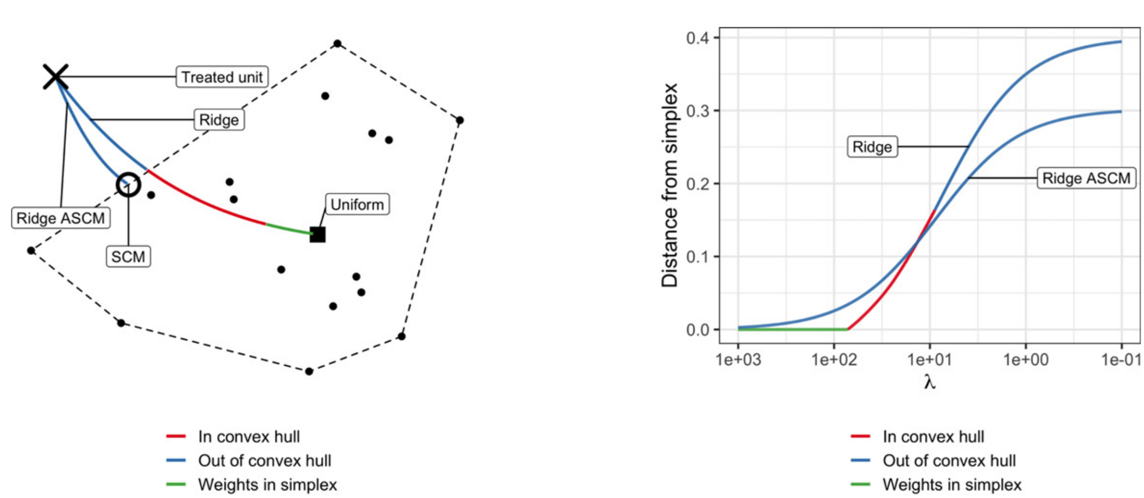
Just as the hyperparameter λ^{ridge} parameterizes the level of extrapolation, it also parameterizes the level of improvement in pretreatment fit over the SCM solution. Because we are removing the nonnegativity constraint and allowing for extrapolation outside of the convex hull, the pretreatment fit from Ridge ASCM will be at least as good as the pretreatment fit from SCM alone, that is, $||\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{aug}}||_2 \leq ||\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}}||_2$. We can exactly characterize the pretreatment fit of Ridge ASCM using the singular value decomposition of the matrix of control outcomes, which will be an important building block in the statistical results below.

Lemma 3. Let $\frac{1}{\sqrt{N_0}}\boldsymbol{X}_{0\cdot} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$ be the singular value decomposition of the matrix of control pre-intervention outcomes, where m is the rank of $\boldsymbol{X}_{0\cdot}$, $\boldsymbol{U} \in \mathbb{R}^{N_0 \times m}$, $\boldsymbol{V} \in \mathbb{R}^{T_0 \times m}$, and $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_m) \in \mathbb{R}^{m \times m}$ is the diagonal matrix of singular values, where d_1 and d_m are the largest and smallest singular values, respectively. Furthermore, let $\widetilde{\boldsymbol{X}}_i = \boldsymbol{V}'\boldsymbol{X}_i$ be the rotation of $\boldsymbol{X}_i$ along the singular vectors of $\boldsymbol{X}_{0\cdot}$. Then $\hat{\boldsymbol{\gamma}}^{\text{aug}}$, the Ridge ASCM weights with hyperparameter $\lambda^{\text{ridge}} = \lambda N_0$ satisfy

$$\left\|\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{aug}}\right\|_2 = \lambda \left\|(\boldsymbol{D} + \lambda \boldsymbol{I})^{-1}\left(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}'\hat{\gamma}^{\text{scm}}\right)\right\|_2 \leq \frac{\lambda}{d_m^2 + \lambda}||\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\gamma}^{\text{scm}}||_2, \tag{20}$$

and the weights from ridge regression alone $\hat{\boldsymbol{\gamma}}^{\text{ridge}}$ satisfy

$$\left\|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{ridge}}\right\|_2 = \lambda\left\|(\boldsymbol{D} + \lambda \boldsymbol{I})^{-1}\widetilde{\boldsymbol{X}}_1\right\|_2 \leq \frac{\lambda}{d_m^2 + \lambda}\left\|\boldsymbol{X}_1\right\|_2. \tag{21}$$



(a) (b)

Figure 1. Ridge ASCM vs. ridge regression alone for a two-dimensional example with the treated unit outside of the convex hull of the control units. Results shown varying λ^{ridge} from 10^3 to 10^{-1}. Green denotes that the weights are inside the simplex, red that the weights are outside the simplex but the weighted average is inside the convex hull, and blue that the weighted average is outside the convex hull. (a) Treated and control units with the convex hull marked as a dashed line. Ridge and Ridge ASCM estimates marked as solid lines. (b) Distance of Ridge and Ridge ASCM weights from the simplex.

From Equation (20), we see that the pretreatment imbalance for Ridge ASCM weights is smaller than that of SCM weights by at least a factor of $\frac{\lambda}{d_m^2+\lambda}$. Thus, Ridge ASCM will achieve strictly better pretreatment fit than SCM alone, except in corner cases where pretreatment fit will be equal, such as when the pretreatment SCM residual $\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}}$ is orthogonal to the lagged outcomes of the control units $\boldsymbol{X}_{0\cdot}$. Since ridge regression penalizes deviations from uniformity, rather than deviations from SCM weights, the relationship for pretreatment imbalance and fit between SCM and ridge regression alone is less clear.

5. Estimation Error for Ridge ASCM

We now relate Ridge ASCM's improved pretreatment fit to improved estimation error under the DGPs in Section 2.2. Under a linear model, improving pretreatment fit directly reduces bias, and the Ridge ASCM penalty term negotiates a bias-variance trade-off. Under a latent factor model, improving pretreatment fit again reduces bias, though there is now a risk of over-fitting. The penalty term also directly parameterizes this trade-off. Thus, choosing the hyperparameter λ^{ridge} is important in practice. We describe a cross-validation hyperparameter selection procedure in Section 5.3. Finally, we discuss inference in Section 5.4.

5.1. Error Under Linearity in Pretreatment Outcomes

We first illustrate the key balancing idea in the simple case in our first DGP, where the posttreatment outcome is a linear combination of lagged outcomes plus additive noise, as in Assumption 1(a). We consider a generic weighting estimator with weights $\hat{\boldsymbol{\gamma}}$ that are independent of the posttreatment outcomes $Y_{1T}, \ldots, Y_{NT}$; both SCM and Ridge ASCM take this form. The difference between the counterfactual outcome $Y_{1T}(0)$ and the weighting estimator $\hat{Y}_{1T}(0)$ decomposes into (i) systemic error due to imbalance in the lagged outcomes $\boldsymbol{X}$, and (ii) idiosyncratic error due to the noise in the posttreatment period:

$$Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i Y_{iT} = \underbrace{\boldsymbol{\beta} \cdot (\boldsymbol{X}_1 - \hat{\gamma}_i \boldsymbol{X}_i)}_{\text{imbalance in } \boldsymbol{X}} + \underbrace{\varepsilon_{1T} - \sum_{W_i=0} \hat{\gamma}_i \varepsilon_{iT}}_{\text{posttreatment noise}}. \quad (22)$$

With this setup, a weighting estimator that exactly balances the lagged outcomes $\boldsymbol{X}$ will eliminate all systematic error. Furthermore, if the vector of autoregression coefficients $\boldsymbol{\beta}$ is sparse, then it suffices to balance only the lagged outcomes with non-zero coefficients; for example, under an AR(K) process, $(\beta_1, \ldots, \beta_{T_0-K-1}) = 0$, it is sufficient to balance only the first K lags.

If the weighting estimator does not perfectly balance the pretreatment outcomes $\boldsymbol{X}$, then there will be a systematic component of the error, with the magnitude depending on the imbalance. Below we construct a finite sample error bound for Ridge ASCM (and for SCM, the special case with $\lambda^{\text{ridge}} = \infty$), building on Lemma 3. This bound on the estimation error holds with high probability over the noise in the posttreatment period $\boldsymbol{\varepsilon}_T$.

Proposition 1. Under the autoregressive model in Assumption 1(a), for any $\delta > 0$ the Ridge ASCM weights with hyperparameter $\lambda^{\text{ridge}} = \lambda N_0$ satisfy the bound

$$\left| Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} Y_{iT} \right| \leq \|\boldsymbol{\beta}\|_2 \underbrace{\left\| \text{diag}\left(\frac{\lambda}{d_j^2 + \lambda} \right) (\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}' \hat{\boldsymbol{\gamma}}^{\text{scm}}) \right\|_2}_{\text{imbalance in } \boldsymbol{X}} + \underbrace{\delta\sigma \left(1 + \|\hat{\boldsymbol{\gamma}}^{\text{aug}}\|_2\right)}_{\text{posttreatment noise}}, \quad (23)$$

with probability at least $1 - 2e^{-\frac{\delta^2}{2}}$, where $\widetilde{\boldsymbol{X}}_i = \boldsymbol{V}'\boldsymbol{X}_i$ is the rotation of $\boldsymbol{X}_i$ along the singular vectors of $\boldsymbol{X}_{0\cdot}$, as above, and σ is the sub-Gaussian scale parameter.

Proposition 1 shows the finite sample error of Ridge ASCM weights is controlled by the imbalance in the lagged outcomes and the L^2 norm of the weights; Lemma A.3 in the supplementary material gives a deterministic bound for $||\hat{\boldsymbol{\gamma}}^{\text{aug}}||_2$. See Athey, Imbens, and Wager (2018) for analogous results on balancing weights in high dimensional cross-sectional settings.

In the special case that SCM weights have perfect pretreatment fit, ASCM and SCM weights will be equivalent, and the estimation error will only be due to the variance of the weights and posttreatment noise. When SCM weights do not achieve perfect pretreatment fit, Ridge ASCM with finite λ extrapolates outside the convex hull, improving pretreatment fit and thus reducing bias. This is subject to the usual bias-variance trade-off: The second term in (23) is increasing in the L^2 norm of the weights, which will generally be larger for ASCM than for SCM. The hyperparameter λ directly negotiates this trade-off.

5.2. Error Under a Latent Factor Model

Following Abadie, Diamond, and Hainmueller (2010), we now consider the case where control potential outcomes are generated according to a linear factor model, as in Assumption 1(b): $Y_{it}(0) = \boldsymbol{\phi}_i \cdot \boldsymbol{\mu}_t + \varepsilon_{it}$. Under this model, the finite-sample error of a weighting estimator depends on the imbalance in the latent factors $\boldsymbol{\phi}$ and a noise term due to the noise at time T:

$$Y_{1T}(0) - \hat{Y}_{1T}(0) = Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i Y_{iT} = \underbrace{\left(\boldsymbol{\phi}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{\phi}_i \right)}_{\text{imbalance in } \boldsymbol{\phi}} \cdot \boldsymbol{\mu}_T + \underbrace{\varepsilon_{1T} - \sum_{W_i=0} \hat{\gamma}_i \varepsilon_{it}}_{\text{noise}}. \quad (24)$$

Balancing the observed pretreatment outcomes $\boldsymbol{X}$ will not necessarily balance the latent factor loadings $\boldsymbol{\phi}$. Following Abadie, Diamond, and Hainmueller (2010), we show in the supplementary material that, under Equation (6), we can decompose the imbalance term as follows:

$$\left(\boldsymbol{\phi}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{\phi}_i\right) \cdot \boldsymbol{\mu}_T = \frac{1}{T_0}\boldsymbol{\mu}' \underbrace{\left(\boldsymbol{X}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{X}_i\right)}_{\text{imbalance in } \boldsymbol{X}} \cdot \boldsymbol{\mu}_T$$
$$- \frac{1}{T_0}\boldsymbol{\mu}' \underbrace{\left(\boldsymbol{\varepsilon}_{1(1:T_0)} - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{\varepsilon}_{i(1:T_0)}\right)}_{\text{approximation error}} \cdot \boldsymbol{\mu}_T, \qquad (25)$$

where $\boldsymbol{\varepsilon}_{i(1:T_0)} = (\varepsilon_{i1}, \ldots, \varepsilon_{iT_0})$ is the vector of pretreatment noise terms for unit i. The first term is the imbalance of observed lagged outcomes and the second term is an approximation error arising from the latent factor structure. In the noiseless case where $\sigma = 0$ and all $\varepsilon_{it} = 0$ deterministically, the approximation error is zero, and it is possible to express $Y_{iT}(0)$ as a linear combination of the pretreatment outcomes, recovering the linear-in-lagged-outcomes case above. However, with $\sigma > 0$ we cannot write the period-T outcome as a linear combination of earlier outcomes plus independent, additive error.

With this setup, we can bound the finite-sample error in Equation (24) for Ridge ASCM weights (and for SCM weights as a special case). This bound is with high probability over the noise in all time periods ε_{it}, and accounts for the noise in the pre- and posttreatment outcomes separately.

Theorem 1. Under the linear factor model in Assumption 1(b), for any $\delta > 0$ the Ridge ASCM weights with hyperparameter $\lambda^{\text{ridge}} = \lambda N_0$ satisfy the bound

$$\left| Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} Y_{1T}(0) \right|$$
$$\leq \frac{JM^2}{\sqrt{T_0}} \left(\underbrace{\left\| \text{diag}\left(\frac{\lambda}{d_j^2 + \lambda}\right) (\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}' \hat{\boldsymbol{\gamma}}^{\text{scm}}) \right\|_2}_{\text{imbalance in } \boldsymbol{X}} \right.$$
$$+ \underbrace{4(1+\delta) \left\| \text{diag}\left(\frac{d_j \sigma}{d_j^2 + \lambda}\right) (\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}' \hat{\boldsymbol{\gamma}}^{\text{scm}}) \right\|_2}_{\text{excess approximation error}}$$
$$\left. + \underbrace{2\sigma \left(\sqrt{\log 2N_0} + \frac{\delta}{2} \right)}_{\text{SCM approximation error}} \right) + \underbrace{\delta\sigma \left(1 + \|\hat{\boldsymbol{\gamma}}^{\text{aug}}\|_2\right)}_{\text{posttreatment noise}} \qquad (26)$$

with probability at least $1 - 6e^{-\frac{\delta^2}{2}} - e^{-2(\log 2 + N_0 \log 5)\delta^2}$, where σ is the sub-Gaussian scale parameter.

Theorem 1 shows that, relative to the linear case in Proposition 1, there is an additional source of error under a latent factor model: approximation error due to balancing lagged outcomes rather than balancing underlying factors. In particular, it is now possible that a control unit only receives a large weight because of idiosyncratic noise, rather than because of similarity in the underlying factors. See Arkhangelsky et al. (2019) and Ferman (2019) for asymptotic analogues of this finite sample bound. As we discuss below, each of the first three terms of the bound in Theorem 1 are directly computable from the observed data, save for the unknown σ parameter.

In the special case where SCM achieves perfect pretreatment fit, considered by Abadie, Diamond, and Hainmueller (2010), the ASCM and SCM weights are equivalent and the error is only due to posttreatment noise and the approximation error. The bound in Theorem 1 accounts for the worst-case scenario where the control unit with the largest weight is only similar to the treated unit due to idiosyncratic noise. The approximation error, and thus the bias, converges to zero in probability as $T_0 \to \infty$ under suitable conditions on the factor loadings $\boldsymbol{\mu}_t$ (see also Ferman and Pinto 2018). Intuitively, as we observe more X_{it}—and can exactly balance each one—we are better able to match on the index $\boldsymbol{\phi}_i \cdot \boldsymbol{\mu}_t$ and, as a result, on the underlying factor loadings. Although we assume independent errors here, in the supplementary material we show that with dependent errors the worst-case error additionally scales with covariance of the error terms.

Without exact balance, Theorem 1 shows that a long preperiod may not be enough to control the error due to imbalance. In this case, Ridge ASCM with $\lambda < \infty$ will extrapolate outside the convex hull, reducing error due to imbalance in the lagged outcomes but possibly over-fitting to noise. Thus, the optimal level of extrapolation will depend on the synthetic control fit and the amount of noise.

Figure 2 illustrates this using SCM weights from the empirical example we discuss in Section 7, where pretreatment fit is good but not perfect. For each value of σ, the figure plots the sum of the imbalance, SCM approximation error, and excess approximation error terms in the bound in Theorem 1, all directly computable from the data for a given σ. At each noise level, a small amount of extrapolation leads to a smaller error bound, but as λ shrinks there is a point where further extrapolation leads to over-fitting and eventually to a worse error bound than without extrapolation. The risk of overfitting is greater when the noise is large (e.g., $\sigma = 0.5$), though even here a sufficiently regularized ASCM estimate has a lower error bound than SCM alone (represented as the $\lambda \to \infty$ bound at the left boundary). When noise is less extreme, the benefits of augmentation are larger and the optimal amount of regularization shrinks.

It is worth noting that Theorem 1 gives a worst-case bound. In Section 7.1 we inspect the typical performance of the Ridge ASCM estimator via extensive simulation studies and find that gains to pretreatment fit through augmentation outweigh increased approximation error in a range of practical settings, including when noise is very large.

Theorem 1 suggests two diagnostics to supplement the estimated bias from Equation (9), based on the first two terms in the bound. For the first term, we can directly assess imbalance in $\boldsymbol{X}$ via the pretreatment RMSE, $\frac{1}{\sqrt{T_0}}\|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}' \hat{\gamma}^{\text{aug}}\|_2$. For the second term, the excess approximation error depends on the unknown noise level, σ. However, as we show in the supplementary material, the excess approximation error is a scaled version of the root mean square distance between the Ridge ASCM weights and the SCM weights, $\frac{1}{\sqrt{N_0}}\|\hat{\gamma}^{\text{aug}} - \hat{\gamma}^{\text{scm}}\|_2$, which is a measure of extrapolation. We report these diagnostics for the empirical application in Section 7. As Figure 2 previews, they

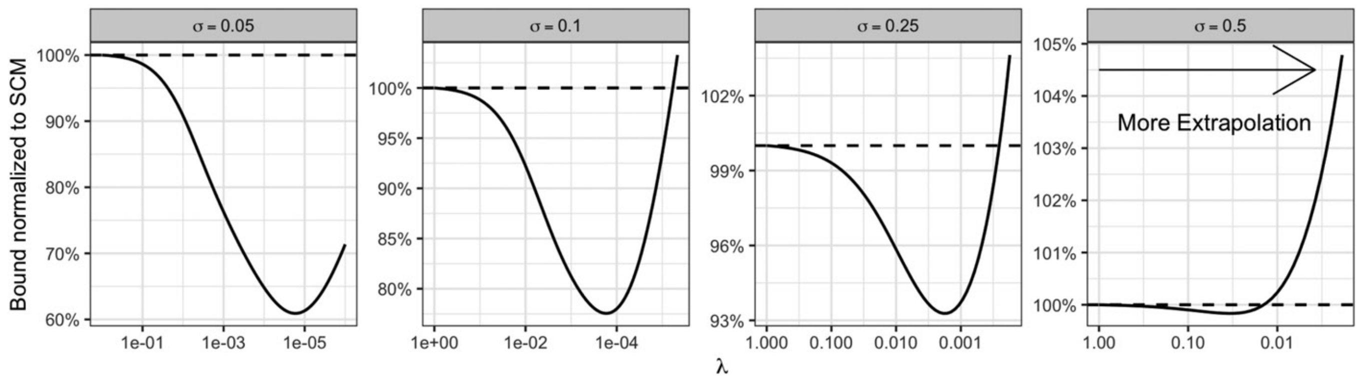


Figure 2. Sketch of the error due to imbalance and approximation error (26) for the linear factor model; the standard deviation of the treated unit's pretreatment outcomes is normalized to one. We fit SCM weights on the empirical example in Section 7 and compute the vector of pretreatment fit. Each line shows the sum of the error due to imbalance in $\boldsymbol{X}$, excess approximation error, and SCM approximation error in Theorem 1 (with $\delta = 0$) for different values of σ. These are normalized so that the SCM solution (with λ large) equals 100%; values below 100% show improvement over the unadjusted weights for a given λ.

support the use of ASCM in this instance, despite what visually appears to be good pretreatment fit for SCM.

5.3. Hyperparameter Selection

We propose a cross-validation approach for selecting λ inspired by the in-time placebo check of Abadie, Diamond, and Hainmueller (2015). Let $\hat{Y}_{1t}^{(-k)} = \sum_{W_i=0} \hat{\gamma}_{i(-k)}^{\text{aug}} Y_{it}$ be the estimate of Y_{1t} where time period k is excluded from fitting the estimator in Equation (17). Abadie, Diamond, and Hainmueller (2015) proposed to compare the difference $Y_{1t} - \hat{Y}_{1t}^{(-t)}$ for some $t \leq T_0$ as a placebo check. We can extend this idea to compute the leave-one-out cross validation MSE over time periods:

$$\text{CV}(\lambda) = \sum_{t=1}^{T_0} \left(Y_{1t} - \hat{Y}_{1t}^{(-t)} \right)^2. \tag{27}$$

We can then choose λ to minimize $\text{CV}(\lambda)$ or follow a more conservative approach such as the "one-standard-error" rule (Hastie, Friedman, and Tibshirani 2009). This proposal is similar to the leave-one-out cross validation proposed by Doudchenko and Imbens (2017), who select hyperparameters by holding out control units and minimizing the MSE of the control units in the posttreatment time T. Finally, only excluding time period t might be inappropriate for some outcome models, for example, the linear model in Section 5.1. In these settings, we can extend the procedure to exclude all time periods $\geq t$ when estimating $\hat{\gamma}_{(-t)}^{\text{aug}}$, as in Kellogg et al. (2020).

5.4. Inference

There is a growing literature on inference for the SCM and variants, going beyond the original proposal in Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010, 2015); see, for example, Kathleen (2020), Toulis and Shaikh (2018), Cattaneo, Feng, and Titiunik (2019), and Chernozhukov, Wuthrich, and Zhu (2018).

We focus here on the conformal inference approach of Chernozhukov, Wüthrich, and Zhu (2019), which has three key steps. First, for a given sharp null hypothesis, $H_0 : \tau = \tau_0$, we create an adjusted posttreatment outcome for the treated unit $\tilde{Y}_{1T} = Y_{1T} - \tau_0$ and extend the original dataset to include the adjusted outcome $\tilde{Y}_{1T}$. Second, we apply the estimator (17) to the extended dataset to obtain adjusted weights $\hat{\boldsymbol{\gamma}}(\tau_0)$. Finally, we compute a p-value by assessing whether the adjusted residual $Y_{1T} - \tau_0 - \sum_{W_i=0} \hat{\gamma}_i(\tau_0) Y_{iT}$ "conforms" with the pretreatment residuals

$$p(\tau_0) = \frac{1}{T} \sum_{t=1}^{T_0} \mathbb{1} \left\{ \left| Y_{1T} - \tau_0 - \sum_{W_i=0} \hat{\gamma}_i(\tau_0) Y_{iT} \right| \leq \left| Y_{1t} - \sum_{W_i=0} \hat{\gamma}_i(\tau_0) Y_{it} \right| \right\} + \frac{1}{T}. \tag{28}$$

Since the counterfactual outcome $Y_{1T}(0)$ is random, inverting this test to construct a confidence interval for τ is equivalent to constructing a conformal *prediction* set (Vovk, Gammerman, and Shafer 2005) for $Y_{1T}(0)$ by using the quantiles of pretreatment residuals:

$$\widehat{C}_Y^{\text{conf}} = \left\{ y \in \mathbb{R} \,\middle|\, \left| y - \sum_{W_i=0} \hat{\gamma}_i(Y_{1T} - y) Y_{iT} \right| \leq q_{T,\alpha}^{+} \left(\left\{ \left| Y_{1t} - \sum_{W_i=0} \hat{\gamma}_i(Y_{1T} - y) Y_{it} \right| \right\} \right) \right\}, \tag{29}$$

where $q_{T,\alpha}^{+}(\{x_t\})$ is the $\lceil (1-\alpha)T \rceil^{th}$ order statistic of $x_1, \ldots, x_T$.

Chernozhukov, Wüthrich, and Zhu (2019) provided several conditions for approximate or exact finite-sample validity of the p-values, and hence coverage of the prediction interval $\widehat{C}_Y^{\text{conf}}$. We briefly discuss two of these conditions here, with a more complete technical treatment in Appendix A in the supplementary material. First, Chernozhukov, Wüthrich, and Zhu (2019) showed exact validity when the residuals $Y_{1t} - \sum_{W_i=0} \hat{\gamma}_i(\tau_0) Y_{it}$ are exchangeable for all $t = 1, \ldots, T$. One sufficient condition for this is that the outcome vectors $(Y_{1t}, \ldots, Y_{Nt})$ are themselves exchangeable for $t = 1, \ldots, T$.

When the residuals are not exchangeable, Chernozhukov, Wüthrich, and Zhu (2019) provided a finite sample bound that relates in-sample prediction error to the validity of $p(\tau_0)$. In Appendix A in the supplementary material, we adapt their SCM bounds to Ridge ASCM by showing that the ridge penalty controls the difference between SCM and Ridge ASCM weights. Under a variant of the basic model (3), the resulting p-value will be valid as the number of pretreatment periods $T_0 \to \infty$. Finally, in Section 7.1, we explore the finite sample coverage probabilities of $\widehat{C}_Y^{\text{conf}}$ under various DGPs and find that they are near their nominal levels.

6. Auxiliary Covariates

Thus far, we have focused exclusively on lagged outcomes as predictors. We now consider the case where there are also a small number of auxiliary covariates $\boldsymbol{Z}_i \in \mathbb{R}^K$ for unit i. These auxiliary covariates may include summaries of lagged outcomes or time-varying covariates such as the pretreatment mean $\bar{X}_i$. Let $\boldsymbol{Z}_{0\cdot} \in \mathbb{R}^{N_0 \times K}$ denote the matrix of donor units' covariates, which we assume are centered, $\bar{\boldsymbol{Z}}_{0\cdot} = \boldsymbol{0}$.

These auxiliary covariates can be incorporated into both the balance objective for SCM and the outcome model used for augmentation in ASCM. For the former, we can extend SCM to choose weights to solve

$$\min_{\gamma \in \Delta^{N_0}} \quad \theta_x \|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma}\|_2^2 + \theta_z \|\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}\gamma\|_2^2 \; + \; \zeta \sum_{W_i=0} f(\gamma_i), \tag{30}$$

where Δ^{N_0} is the N_0-simplex. For the latter, we can augment the SCM weights with an outcome model $\hat{m}(\boldsymbol{X}_i, \boldsymbol{Z}_i)$ that is a function of both the lagged outcomes and auxiliary covariates. For example, we can extend Ridge ASCM to choose $\hat{m}(\boldsymbol{X}, \boldsymbol{Z}) = \hat{\eta}_0 + \boldsymbol{X}'\hat{\boldsymbol{\eta}}_x + \boldsymbol{Z}'\hat{\boldsymbol{\eta}}_z$ and fit via ridge regression:

$$\min_{\eta_0, \boldsymbol{\eta_x}, \boldsymbol{\eta_z}} \; \frac{1}{2} \sum_{W_i=0} (Y_i - (\eta_0 + \boldsymbol{X}_i'\boldsymbol{\eta_x} + \boldsymbol{Z}_i'\boldsymbol{\eta_z}))^2 + \lambda_x \|\boldsymbol{\eta_x}\|_2^2 + \lambda_z \|\boldsymbol{\eta_z}\|_2^2. \tag{31}$$

Both this SCM criterion and augmentation estimator incorporate user-specified weights that determine the importance of balancing each set of covariates (Equation (30)) or the amount of regularization for each set of coefficients (Equation (31)). There are many potential choices for these weights. We discuss two, appropriate to different settings depending on the number of auxiliary covariates.

A sensible default when the dimension of the auxiliary covariates is moderate is to incorporate the lagged outcomes $\boldsymbol{X}$ and the auxiliary covariates $\boldsymbol{Z}$ equally in Equations (30) and (31), setting $\theta_x = \theta_z = 1$ and $\lambda_x = \lambda_z = \lambda^{\text{ridge}}$ (after standardizing auxiliary covariates and lagged outcomes to have equal variance). With this setup the algorithmic results in Section 4 apply for the combined vector of lagged outcomes and auxiliary covariates, $(\boldsymbol{X}_i, \boldsymbol{Z}_i) \in \mathbb{R}^{T_0+K}$. In particular, Ridge ASCM is again a penalized SCM estimator that adjusts the synthetic control weights that solve optimization problem (30) to achieve better balance by extrapolating outside of the convex hull.

An alternative approach when the dimension of the auxiliary covariates is small relative to N (i.e., $K \ll N$) is to fit a regression model that regularizes the lagged outcome coefficients $\boldsymbol{\eta_x}$ but does *not* regularize the auxiliary covariate coefficients $\boldsymbol{\eta_z}$ (i.e., set $\lambda_z = 0$). Lemma 4 below writes the resulting augmented estimator as its corresponding penalized SCM optimization problem, with weights that perfectly balance the auxiliary covariates. This has two key implications. First, since the auxiliary covariates $\boldsymbol{Z}$ are exactly balanced regardless of the balance that the SCM weights achieve alone, we can exclude them from the optimization problem (30). Second, as we show below, the pretreatment fit on the lagged outcomes depends on how well the SCM weights balance the residualized lagged outcomes $\check{\boldsymbol{X}}$, defined in Lemma 4. This suggests modifying Equation (30) to balance $\check{\boldsymbol{X}}$ rather than the lagged outcomes $\boldsymbol{X}$, which leads to a two-step procedure: (i) residualize the pre- and posttreatment outcomes on the auxiliary covariates $\boldsymbol{Z}$; and (ii) estimate Ridge ASCM on the residualized outcomes. This two-step procedure follows from a related proposal in Doudchenko and Imbens (2017).

Lemma 4. Let $\hat{\boldsymbol{\eta}}_x$ and $\hat{\boldsymbol{\eta}}_z$ be the solutions to Equation (31) with $\lambda_x = \lambda^{\text{ridge}}$ and $\lambda_z = 0$. For any weight vector $\hat{\boldsymbol{\gamma}}$ that sums to one, the ASCM estimator from Equation (10) with $\hat{m}(\boldsymbol{X}_i, \boldsymbol{Z}_i) = \boldsymbol{X}_i'\hat{\boldsymbol{\eta}}_{\boldsymbol{x}} + \boldsymbol{Z}_i'\hat{\boldsymbol{\eta}}_{\boldsymbol{z}}$ is

$$\sum_{W_i=0} \hat{\gamma}_i Y_{iT} + \left(\boldsymbol{X}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{X}_i\right)' \hat{\boldsymbol{\eta}}_{\boldsymbol{x}} + \left(\boldsymbol{Z}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{Z}_i\right)' \hat{\boldsymbol{\eta}}_{\boldsymbol{z}} = \sum_{W_i=0} \hat{\gamma}_i^{\text{cov}} Y_{iT}, \tag{32}$$

where the weights $\hat{\boldsymbol{\gamma}}^{\text{cov}}$ are

$$\hat{\gamma}_i^{\text{cov}} = \hat{\gamma}_i + (\check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot})(\check{\boldsymbol{X}}_{0\cdot}'\check{\boldsymbol{X}}_{0\cdot} + \lambda^{\text{ridge}}\boldsymbol{I}_{T_0})^{-1}\check{\boldsymbol{X}}_i + (\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}'\gamma)'(\boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot})^{-1}\boldsymbol{Z}_i, \tag{33}$$

and $\check{\boldsymbol{X}}_i$ is the residual components of a regression of pretreatment outcomes on the control auxiliary covariates

$$\check{\boldsymbol{X}}_i = \boldsymbol{X}_i - \boldsymbol{Z}_i'(\boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot})^{-1}\boldsymbol{Z}_{0\cdot}'\boldsymbol{X}_{0\cdot}. \tag{34}$$

These weights exactly balance the auxiliary covariates, $\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{cov}} = 0$; the imbalance in the lagged outcomes is

$$\left\|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{cov}}\right\|_2 \leq \left(\frac{\lambda^{\text{ridge}}}{\lambda^{\text{ridge}} + N_0\check{d}_r^2}\right) \left\|\check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot}'\hat{\boldsymbol{\gamma}}\right\|_2, \tag{35}$$

where $\check{d}_r$ is the minimal singular value of $\check{\boldsymbol{X}}_0$.

Comparing to the results in Section 4, Lemma 4 shows that the two-step approach penalizes extrapolation from the convex hull *in the residualized space* $\check{\boldsymbol{X}}$, rather than in the lagged outcomes themselves. In essence, by residualizing out the auxiliary covariates $\boldsymbol{Z}$, the two-step approach allows for a possibly large amount of extrapolation in the auxiliary covariates, while carefully penalizing extrapolation in the part of the lagged outcomes that is orthogonal to the covariates.

In Appendix B.3 in the supplementary material, we consider the performance of this estimator when the outcomes follow a linear factor model with either a linear or a nonlinear dependence on auxiliary covariates, focusing on the special case where $\lambda^{\text{ridge}} \to \infty$ and the weights $\hat{\boldsymbol{\gamma}}^{\text{cov}}$ do not extrapolate from the convex hull after residualization. When covariates enter linearly and when K is small relative to N_0, we show that exactly balancing a small number of auxiliary covariates and targeting imbalance in the residuals $\check{\boldsymbol{X}}$ decreases error due to pretreatment fit. When covariates enter nonlinearly, however, there is additional approximation error due to the linear regression specification. Thus, it is important to appropriately transform the covariates in practice. Furthermore with larger numbers of covariates, the approach that incorporates them in parallel to lagged outcomes will be more appropriate.

7. Simulations and Empirical Illustrations

We first conduct extensive simulation studies to assess the performance of different methods, finding substantial gains from ASCM. We then use our approach to examine the effect of an aggressive tax cut on economic output in Kansas in 2012.

7.1. Calibrated Simulation Studies

We now present simulation studies calibrated to our empirical illustration in Section 7.2. Specifically, we use the generalized synthetic control method (Xu 2017) to estimate a factor model with three latent factors based on the series of log gross state product (GSP) per capita, $N = 50$, $T_0 = 89$. We then simulate outcomes using the distribution of estimated parameters and model selection into treatment as a function of the latent factors; see Appendix C in the supplementary material for additional details. We also present results from three additional DGPs, each calibrated to estimates from the same data: (i) the factor model with quadruple the standard deviation of the noise term, (ii) a unit and time fixed effects model, and (iii) an autoregressive model with 3 lags.

We explore the role of augmentation using different outcome estimators. For each DGP, we consider five estimators: (i) SCM alone, (ii) ridge regression alone, (iii) Ridge ASCM, (iv) fixed-effects alone, and (v) De-meaned SCM (i.e., SCM augmented with fixed effects) from Doudchenko and Imbens (2017) and Ferman and Pinto (2018), as shown in Equation (13). See Appendix F in the supplementary material for simulations with additional outcome models for ASCM. Figure 3 shows the Monte Carlo estimate of the absolute bias as a percentage of the absolute bias for SCM, with one panel for each simulation DGP. Figure F.1 in the supplementary material shows the corresponding estimator RMSE.

There are several takeaways. First, augmenting SCM with a ridge outcome regression reduces bias relative to SCM alone—*without* conditioning on excellent pretreatment fit—in all four simulations. This underscores the importance of the recommendation in Abadie, Diamond, and Hainmueller (2010, 2015) to use SCM only in settings with excellent pretreatment fit. Under the baseline factor model and the fixed effect model, the ridge augmentation greatly reduces bias, by more than 75% in the factor model simulation and over 90% in the fixed effects simulation. In the AR(3) model and in the factor model with greater noise, the gains to augmentation relative to SCM are more limited. Second, Ridge ASCM has lower bias than ridge regression alone across all of the simulation settings. Third, when the fixed effects estimator is incorrectly specified, combining it with SCM has much lower bias than either method alone. And even when the fixed-effects estimator is correctly specified, de-meaned SCM has similar bias to the (correctly specified) fixed-effects approach. Finally, Figure F.1 (supplementary material) shows that in all simulations ASCM has lower RMSE than SCM, as the large decrease in bias more than makes up for the slight increase in variance.

Complementing the worst-case analysis in Section 5, we now consider how the typical performance of augmentation relates to the amount of extrapolation and the quality of the original SCM fit. Figure 4 shows the bias and RMSE as a function of λ for the primary factor model simulation, conditional on the quartile of SCM fit. Larger values of λ (and hence smaller adjustments) are to the left, with the left-most points in the plots representing SCM. First, as expected, Augmented SCM substantially reduces bias regardless of SCM pretreatment fit. However, the gains are more modest when the SCM fit is in the best quartile: in this case the bias is non-monotonic in λ and there is some optimal choice of λ that minimizes the bias. Second, it is possible to

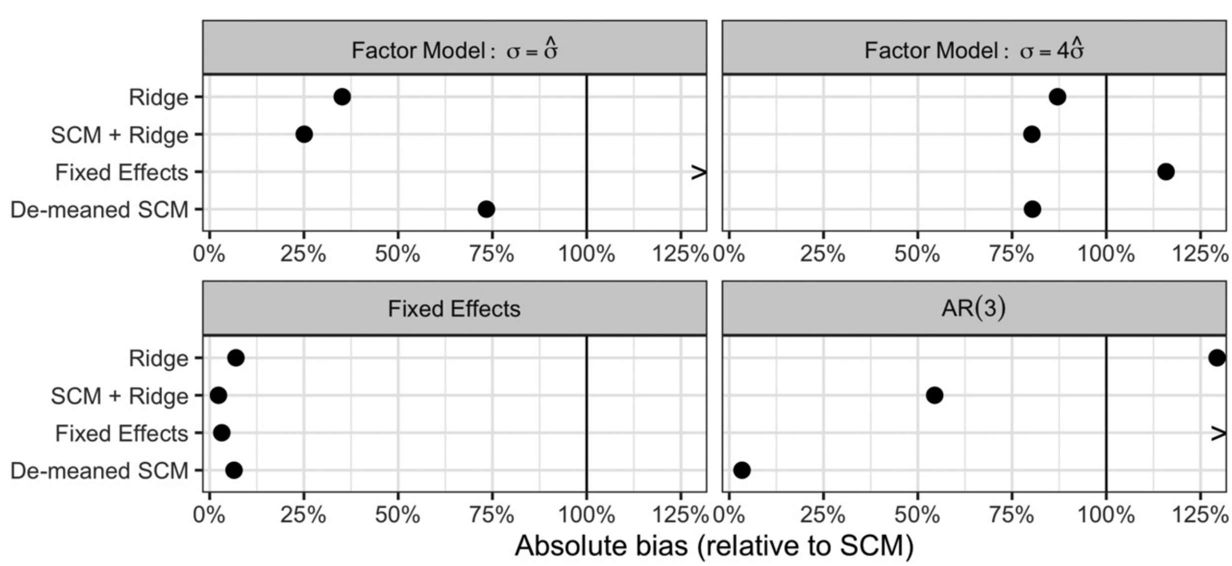


Figure 3. Overall absolute bias, normalized to SCM bias for (a) the factor model simulation, (b) the factor model simulation with quadruple the standard deviation, (c) the fixed effects simulation, and (d) the AR simulation. Arrows indicate values larger than 130%. The SCM estimates reported here are *not* restricted to simulation draws with excellent pretreatment fit; Abadie, Diamond, and Hainmueller (2015) advise against using SCM in such settings.

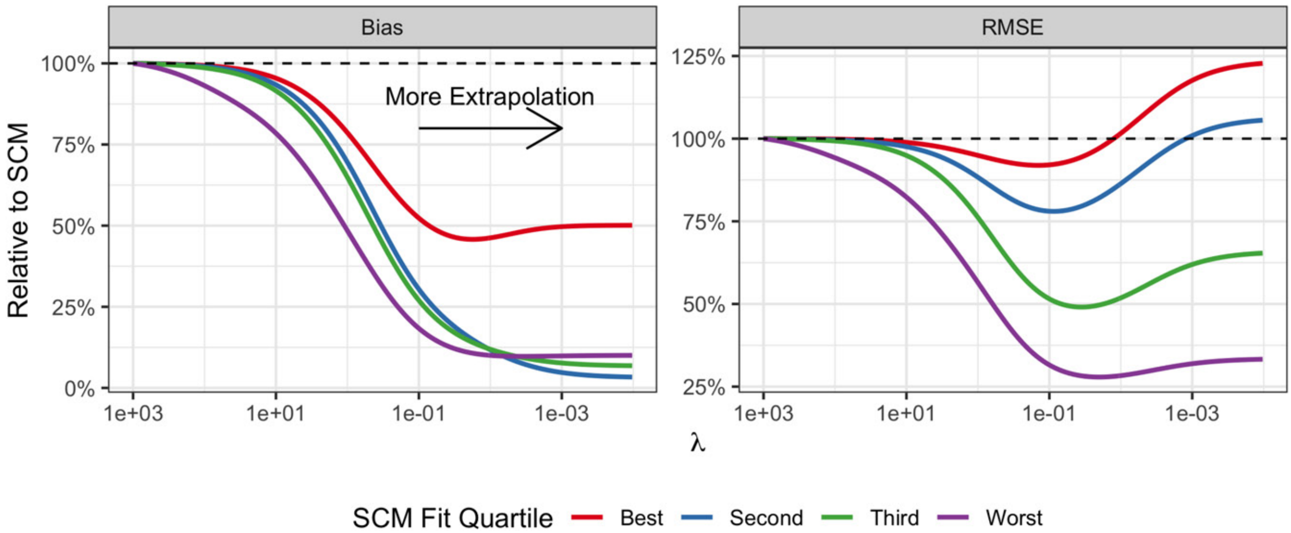


Figure 4. Bias and RMSE of Ridge ASCM, as a percentage of SCM bias and RMSE, versus λ under a linear factor model. Results are divided by the quartile of the SCM fit across all simulations.

Table 1. Coverage for 95% conformal prediction intervals (29) based on 1000 repetitions.

Method	AR(3)	Factor $\sigma =$ Model: $\hat{\sigma}$	Factor $\sigma =$ Model: $4\hat{\sigma}$	Fixed Effects
SCM	0.934	0.926	0.930	0.889
SCM + Ridge	0.932	0.950	0.936	0.939

under-regularize with ASCM, as evident in the RMSE achieving a minimum for an intermediate value of λ. When pretreatment fit is good, augmentation with too-small λ leads to higher RMSE than SCM alone. However, when SCM fit is relatively poor, even minimally regularized ASCM achieves much better bias and RMSE than does SCM.

Finally, Table 1 shows the finite sample coverage of the conformal prediction intervals for $Y_{1T}(0)$. For the four simulation settings, we compute 95% prediction intervals for the posttreatment counterfactual outcome $Y_{1T}(0)$ using the both the SCM and ridge ASCM estimators. We see that the intervals for SCM alone can slightly undercover, due to finite sample bias from poor treatment fit. In contrast, the intervals for ridge ASCM have close to nominal coverage for $Y_{1T}(0)$.

Overall we find that SCM augmented with a penalized regression model has consistently good performance across DGPs. Due to this performance and the method's relative simplicity, we therefore recommend augmenting SCM with penalized regression as a reasonable default in settings where SCM alone has poor pretreatment fit. In particular, we suggest using ridge regression; among the other benefits, Ridge ASCM allows the practitioner to diagnose the level of extrapolation due to the outcome model.

7.2. Illustration: 2012 Kansas Tax Cuts

In 2010, Sam Brownback was elected governor of Kansas, having run on a platform emphasizing tax cuts and deficit reduction (see Rickman and Wang 2018, for further discussion and analysis). Upon taking office, he implemented a substantial personal income tax cut, both lowering rates and reducing credits and deductions. This is a valuable test of "supply side" models: Brownback argued that the tax cuts would increase business activity in Kansas, generating economic growth and additional tax revenues that would make up for the static revenue losses. Kansas' subsequent economic performance has not been impressive relative to its neighbors; however, potentially confounding factors include a drought and declines in the locally important aerospace industry. Finding a credible control for Kansas is thus challenging, and SCM-type approaches offer a potential solution.

We estimate the effect of the tax cuts on log GSP per capita using the second quarter of 2012—when Brownback signed the tax cut bill into law—as the intervention time. We use four primary estimators: (i) SCM alone fit on the entire vector of lagged outcomes, (ii) Ridge ASCM, (iii) Ridge ASCM including auxiliary covariates in parallel to lagged outcomes and (iv) Ridge ASCM on residualized outcomes, as proposed in Section 6. We select the hyperparameter λ via the cross-validation procedure in Section 5.3, following the "one-standard-error" rule with only lagged outcomes, and selecting the minimal λ when including auxiliary covariates. See Figure F.6 (supplementary material). The covariates we include are the pretreatment averages of (i) log state and local revenue per capita, (ii) log average weekly wages, (iii) number of establishments per capita, (iv) the employment level, and (v) log GSP per capita.

These estimators assume that noise is mean zero (Assumption 1). Substantively, under the autoregressive model in Assumption 1(a) this assumes that posttreatment shocks for Kansas will be the same as for other states in expectation; under the linear factor model in Assumption 1(b) this rules out selection on pretreatment shocks. This also rules out unobserved confounders that affect both posttreatment shocks and the decision to enact the Brownback tax cut bill.

Figure 5, known as a "gap plot," shows the difference between Kansas and its synthetic control for all four estimators, along with 95% point-wise confidence intervals intervals computed via the conformal inference procedure from Chernozhukov, Wüthrich, and Zhu (2019). Figure 6 shows log GSP per capita for both Kansas and its synthetic control using SCM and Ridge

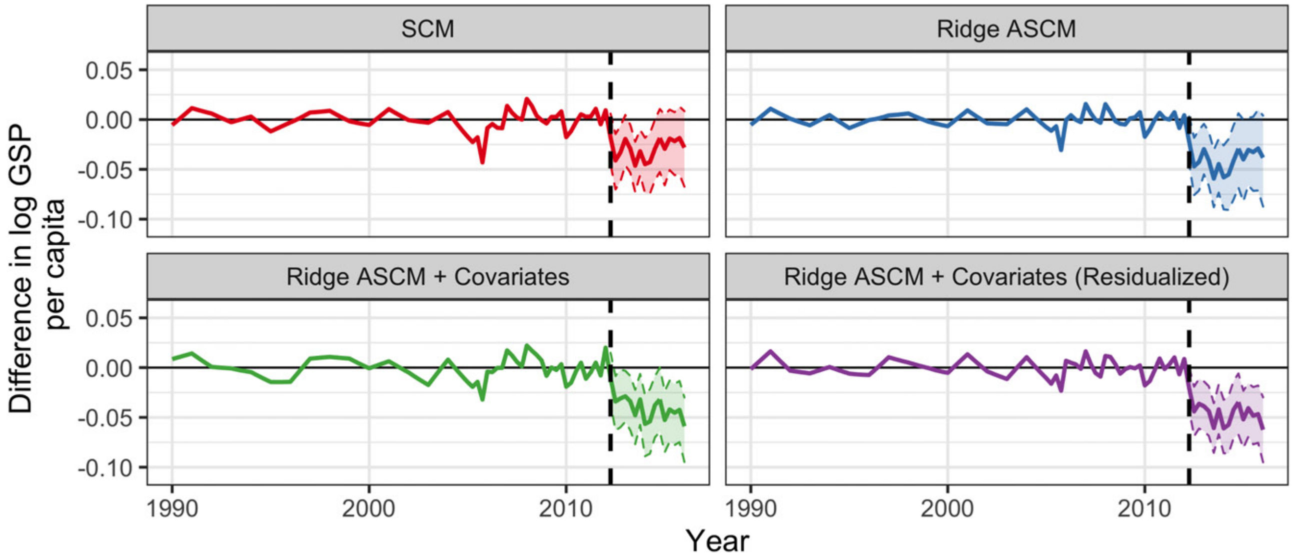


Figure 5. Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on log GSP per capita using SCM, Ridge ASCM, and Ridge ASCM with covariates.

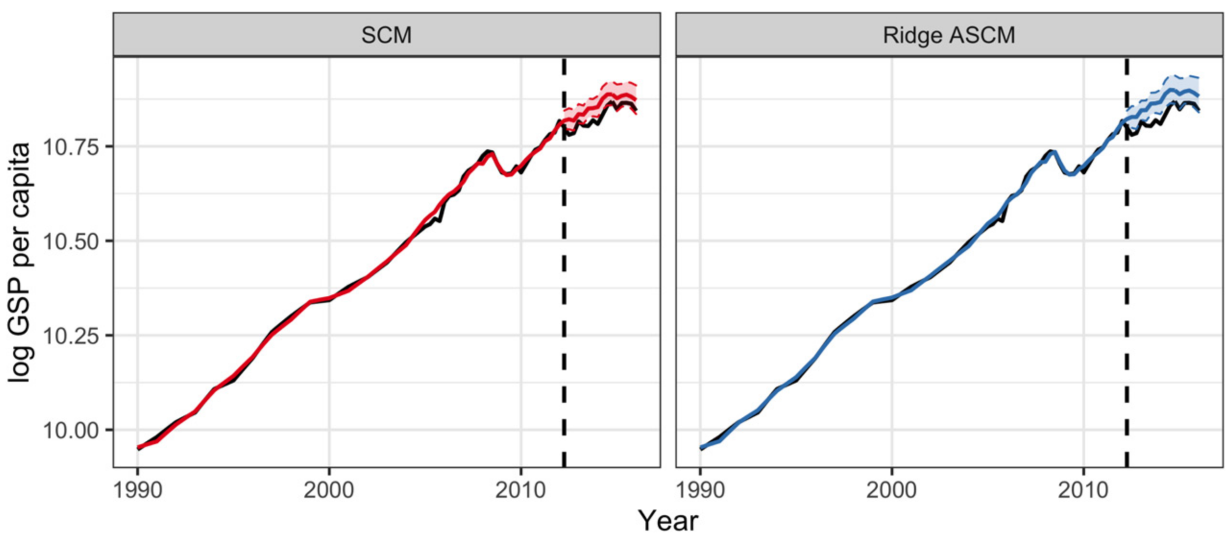


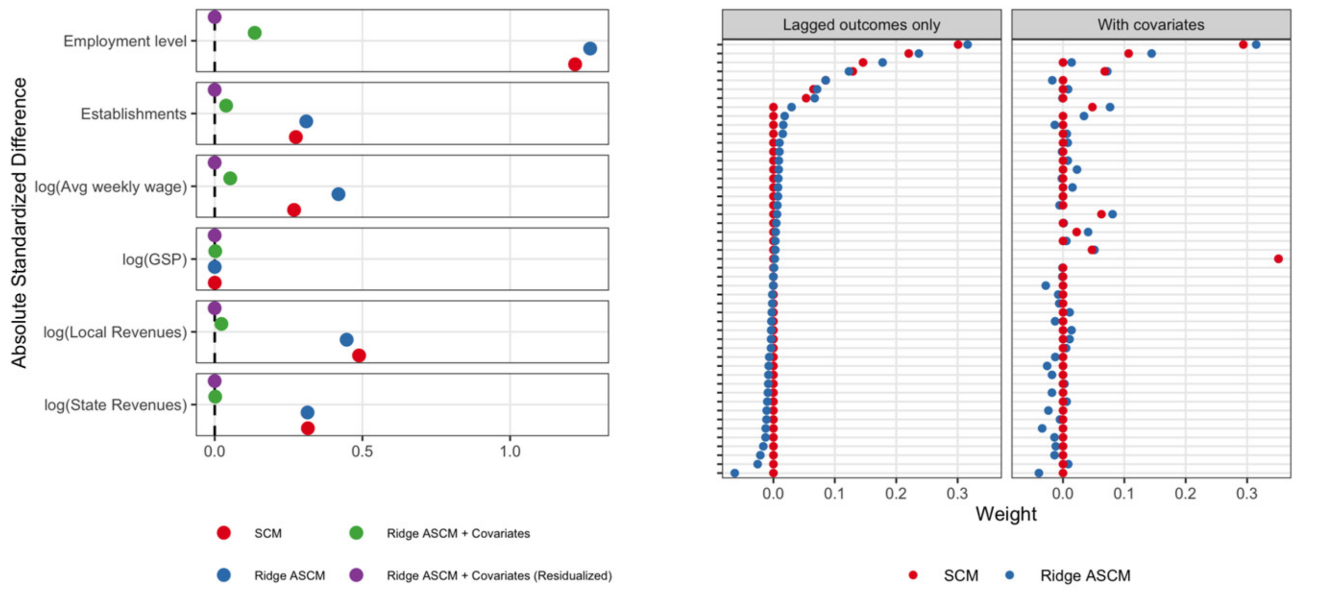
Figure 6. Point estimates along with point-wise 95% conformal prediction intervals for counterfactual log GSP per capita without the tax cuts using SCM, ridge ASCM, and ridge ASCM with covariates, plotting with the observed log GSP per capita in black.

ASCM. Appendix F in the supplementary material shows additional results.

First, the pretreatment fit for SCM alone is relatively good for most of the pre-period, with an overall pretreatment RMSE of about 0.9 log points. However, the fit for SCM alone worsens in 2004–2005, with imbalances of over 4 log points—a pretreatment imbalance as large as the estimated impact. Using ridge regression to assess the possible implications of this pretreatment imbalance, we estimate bias due to pretreatment imbalance of around 1 log point, or roughly a third of the magnitude of the estimated effect. To better understand the estimated bias, we can inspect the ridge regression coefficients for lagged outcomes; see Figure F.9 (supplementary material). While the regression puts the most weight on the two most recent years, the estimated bias due to imbalance in the mid-2000s is just as large as for 2010 and 2011. This suggests that there may be gains to augmentation.

As anticipated, augmenting SCM with ridge regression indeed improves pretreatment fit, with a pretreatment RMSE of 0.65 log points, 25% smaller than the RMSE for SCM alone. This improvement is especially pronounced in the mid 2000s, where SCM imbalance is larger. In the end, despite a large reduction in the pretreatment RMSE, the change in the weights is quite small: the root mean square difference between SCM and Ridge ASCM weights is only 0.01.

Next, we consider including the auxiliary covariates. Adding these auxiliary covariates and augmenting further improves both pretreatment fit and balance on the covariates; see Figure 7(a). Finally, balancing the auxiliary covariates via residualization also improves pretreatment fit. Overall, the estimated impact is consistently negative for all four approaches, with weaker evidence that the effect persists to the end of the observation period.



(a) (b)

Figure 7. (a) Covariate balance for SCM, Ridge ASCM, and ASCM with covariates. Each covariate is standardized to have mean zero and standard deviation one; we plot the absolute difference between the treated unit's covariate and the weighted control units' covariates $\left|Z_{1k} - \sum_{W_i=0} \hat{\gamma} Z_{ik}\right|$. (b) Donor unit weights for (1) SCM alone and (2) Ridge ASCM; left facet uses lagged outcomes only; right facet includes auxiliary covariates.

To check against over-fitting, Figures F.10–F.12 (supplementary material) show in-time placebo estimates for SCM alone, Ridge ASCM, and Ridge ASCM with covariates, with placebo treatment times in the second quarter of 2009, 2010, and 2011. We estimate placebo effects that are near zero with all three placebo treatment times with all three estimators.

Figure 7(a) shows the covariate balance for the four estimators. While SCM and Ridge ASCM achieve excellent fit for the pretreatment average log GSP per capita, neither estimator achieves good balance on the other covariates, most notably the average employment level across the quarters of the preperiod. In contrast, including the auxiliary covariates into both the SCM and ridge optimization problems greatly improves the covariate balance, and—by design—residualizing on the auxiliary covariates perfectly balances them. Moreover, Ridge ASCM on residualized outcomes achieves very good pretreatment fit on the lagged outcomes as shown in Figure 5.

Finally, Figure 7(b) shows the weights on donor units for SCM and Ridge ASCM as well as SCM and Ridge ASCM weights when including covariates jointly with the lagged outcomes (see also, Figure F.14, supplementary material). Here we see the minimal extrapolation property of the ASCM weights. The SCM weights are zero for all but six donor states. The Ridge ASCM weights are similar but deviate slightly from the simplex. As a result, the Ridge ASCM weights retain some of the interpretability of the SCM weights. For the donor units with positive SCM weight, Ridge ASCM places close to the same weight. For the majority of those with zero SCM weight, Ridge ASCM also places a close to zero weight. Only Louisiana receives a meaningful negative weight, with nonnegligible negative weights for only a few other donor units. By contrast, Figure F.13 (supplementary material) shows the weights from ridge regression alone: many of the weights are negative and the weights are far from sparse. Including auxiliary covariates changes the relative importance of different states by adding new information, but the minimal extrapolation property remains.

8. Discussion

SCM is a popular approach for estimating policy impacts at the jurisdiction level, such as the city or state. By design, however, the method is limited to settings where excellent pretreatment fit is possible. For settings when this is infeasible, we introduce Augmented SCM, which controls pretreatment fit while minimizing extrapolation. We show that this approach controls error under a linear factor model and propose several extensions, including to incorporate auxiliary covariates.

There are several directions for future work. First, we can incorporate a sensitivity analysis that directly parameterizes departures from, say, the linear factor model, as in recent approaches for sensitivity analysis for balancing weights (Soriano et al. 2020). Second, we can adapt the ASCM framework to settings with multiple treated units. For instance, there are different approaches in settings when all treated units are treated at the same time: some articles propose to fit SCM separately for each treated unit (e.g., Abadie and L'Hour 2018), while others simply average the units together (e.g., Robbins, Saunders, and Kilmer 2017). The situation is more complicated with staggered adoption, when units take up the treatment at different times; we explore this extension in Ben-Michael, Feller, and Rothstein (2019). Finally, we can consider more complex data structures, such as applications with multiple outcomes series for the same units (e.g., measures of both earnings and total employment in minimum wage studies); hierarchical data structures with outcome information at both the individual and aggregate level (e.g., students within schools); or discrete or count outcomes.

Supplementary Materials

The supplementary materials include additional results on estimation and inference, a discussion of connections to balancing weights and inverse propensity score weighting, and details of the proofs.

Acknowledgments

We thank to Alberto Abadie, Josh Angrist, Matias Cattaneo, Alex D'Amour, Peng Ding, Erin Hartman, Chad Hazlett, Steve Howard, Guido Imbens, Brian Jacob, Pat Kline, Caleb Miles, Luke Miratrix, Sam Pimentel, Fredrik Sävje, Jas Sekhon, Jake Soloff, Panos Toulis, Stefan Wager, Kaspar Wütrich, Yiqing Xu, Alan Zaslavsky, and Xiang Zhou for thoughtful comments and discussion, as well as seminar participants at Stanford, UC Berkeley, UNC, the 2018 Atlantic Causal Inference Conference, COMPIE 2018, and the 2018 Polmeth Conference. We also thank to editors and referees for constructive feedback.

Funding

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Abadie, A. (2021), "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," *Journal of Economic Literature*, 59, 391–425. [1790]

Abadie, A., Diamond, A., and Hainmueller, J. (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505. [1789,1790,1791,1792,1795,1796,1797,1799]

——— (2015), "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59, 495–510. [1789,1790,1791,1792,1794,1797,1799]

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. [1789,1790,1791,1797]

Abadie, A., and Imbens, G. W. (2011), "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business & Economic Statistics*, 29, 1–11. [1789,1792]

Abadie, A., and L'Hour, J. (2018), "A Penalized Synthetic Control Estimator for Disaggregated Data," *Journal of the Acoustical Society of America*. *https://economics.mit.edu/files/18642* [1790,1792,1802]

Amjad, M., Shah, D., and Shen, D. (2018), "Robust Synthetic Control," *The Journal of Machine Learning Research*, 19, 802–852. [1790]

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2019), "Synthetic Difference in Differences," arXiv:1812.09970. [1790,1793,1796]

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2017), "Matrix Completion Methods for Causal Panel Data Models," arxiv: 1710.10251. [1790,1792,1793]

Athey, S., and Imbens, G. W. (2017), "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, 31, 3–32. [1789]

Athey, S., Imbens, G. W., and Wager, S. (2018), "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions," *Journal of the Royal Statistical Society*, Series B, 80, 597–623. [1795]

Ben-Michael, E., Feller, A., and Rothstein, J. (2019), "Synthetic Controls With Staggered Adoption," *Journal of the Royal Statistical Society*, Series B, forthcoming. arXiv: 1912.03290. [1790,1802]

Cattaneo, M. D., Feng, Y., and Titiunik, R. (2019), "Prediction Intervals for Synthetic Control Methods," arXiv: 1912.07120. [1797]

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2018), "Inference on Average Treatment Effects in Aggregate Panel Data Settings," arXiv: 1812.10820. [1797]

——— (2019), "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," *Journal of the American Statistical Association*, *https://doi.org/10.1080/01621459.2021.1920957*. [1791,1797,1800]

Doudchenko, N., and Imbens, G. W. (2017), "Difference-In-Differences and Synthetic Control Methods: A Synthesis," arxiv 1610.07748. [1790,1792,1793,1797,1798,1799]

Ferman, B. (2019), "On the Properties of the Synthetic Control Estimator with Many Periods and Many Controls," *Journal of the American Statistical Association*. arXiv:1906.06665v5 [1796]

Ferman, B., and Pinto, C. (2018), "Synthetic Controls With Imperfect Pre-Treatment Fit," *Quantitative Economics*, Forthcoming. arXiv:1911.08521 [1790,1792,1793,1796,1799]

Gobillon, L., and Magnac, T. (2016), "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls," *Review of Economics and Statistics*, 98, 535–551. [1790]

Hastie, T., Friedman, J., and Tibshirani, R. (2009), *The Elements of Statistical Learning: : Data Mining, Inference, and Prediction*, Springer Series in Statistics (2nd ed.), New York: Springer. [1797]

Hsiao, C., and Zhou, Q. (2018), "Panel Parametric, Semi-Parametric and Nonparametric Construction of Counterfactuals," *Journal of Applied Econometrics*. *https://doi.org/10.1002/jae.2702* [1790]

Kathleen T. L. (2020), "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods," *Journal of the American Statistical Association*, 115, 2068–2083. [1797]

Kellogg, M., Mogstad, M., Pouliot, G., and Torgovitsky, A. (2020), "Combining Matching and Synthetic Controls to Trade Off Biases From Extrapolation and Interpolation," Technical report, National Bureau of Economic Research. [1790,1797]

King, G., and Zeng, L. (2006), "The Dangers of Extreme Counterfactuals," *Political Analysis*, 14, 131–159. [1794]

Neyman, J. (1990 [1923]), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Statistical Science*, 5, 465–472. [1790]

Powell, D. (2018), "Imperfect Synthetic Controls: Did the Massachusetts Health Care Reform Save Lives?" available at SSRN: *https://ssrn.com/abstract=3192710* or *http://dx.doi.org/10.2139/ssrn.3192710* [1790]

Rickman, D. S., and Wang, H. (2018), "Two Tales of Two Us States: Regional Fiscal Austerity and Economic Performance," *Regional Science and Urban Economics*, 68, 46–55. [1800]

Robbins, M., Saunders, J., and Kilmer, B. (2017), "A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention," *Journal of the American Statistical Association*, 112, 109–126. [1790,1802]

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866. [1792]

Samartsidis, P., Seaman, S. R., Presanis, A. M., Hickman, M., De Angelis, D. (2019), "Assessing the Causal Effect of Binary Interventions From Observational Panel Data With Few Treated Units," *Statistical Science*, 34, 486–503. [1790]

Soriano, D., Ben-Michael, E., Bickel, P., Feller, A., and Pimentel, S. (2020), "Interpretable Sensitivity Analysis for Balancing Weights," Technical Report. arxiv: 2102.13218. [1802]

Toulis, P., and Shaikh, A. (2018), "Randomization Tests in Observational Studies With Time-Varying Adoption of Treatment," arXiv:1912.10610. [1797]

Vovk, V., Gammerman, A., and Shafer, G. (2005), *Algorithmic Learning in a Random World*, Boston, MA: Springer. [1797]

Xu, Y. (2017), "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, 25, 57–76. [1790,1793,1799]