NBER WORKING PAPER SERIES

THE AUGMENTED SYNTHETIC CONTROL METHOD

Eli Ben-Michael
Avi Feller
Jesse Rothstein

Working Paper 28885
http://www.nber.org/papers/w28885

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2021

We thank 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 Wutrich, 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 editors and referees for constructive feedback. 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, the U.S. Department of Education, or the National Bureau of Economic Research.

The Augmented Synthetic Control Method
Eli Ben-Michael, Avi Feller, and Jesse Rothstein
NBER Working Paper No. 28885
June 2021
JEL No. C21,C23,E62,H71

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 pre-treatment outcomes and other covariates as closely as possible. A critical feature of the original proposal is to use SCM only when the fit on pre-treatment outcomes is excellent. We propose Augmented SCM as an extension of SCM to settings where such pre-treatment fit is infeasible. Analogous to bias correction for inexact matching, Augmented SCM uses an outcome model to estimate the bias due to imperfect pre-treatment fit and then de-biases the original SCM estimate. Our main proposal, which uses ridge regression as the outcome model, directly controls pre-treatment 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.

Eli Ben-Michael
Institute for Quantitative Social Science
Harvard University
1737 Cambridge Street
Cambridge, MA 02138
ebenmichael@fas.harvard.edu

Avi Feller
Goldman School of Public Policy
University of California, Berkeley
2607 Hearst Avenue
Berkeley, CA 94720
afeller@berkeley.edu

Jesse Rothstein
Goldman School of Public Policy and
Department of Economics
University of California, Berkeley
2607 Hearst Avenue #7320
Berkeley, CA 94720-7320
and NBER
rothstein@berkeley.edu

An R package, augsynth, is available at https://github.com/ebenmichael/augsynth

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 pre-treatment periods (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015). The idea is to construct a weighted average of control units, known as a synthetic control, that matches the treated unit's pre-treatment outcomes. The estimated impact is then the difference in post-treatment outcomes between the treated unit and the synthetic control. SCM has been widely applied — the main SCM papers have over 4,000 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 pre-treatment outcomes closely match the pre-treatment outcomes for the treated unit (Abadie et al., 2015). When it is not possible to construct a synthetic control that fits pre-treatment outcomes well, the original papers advise against using SCM. At that point, researchers often fall back to linear regression. This allows better (often perfect) pre-treatment 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 pre-treatment 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 pre-treatment fit, and then uses this to de-bias the SCM estimate. If pre-treatment fit is good, 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 non-negative, Ridge ASCM admits negative weights, using extrapolation to improve pre-treatment 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 ASCM's improved pre-treatment fit to a finite sample bound on estimation error under several data generating processes, including an autoregressive model and the linear factor model often invoked in this setting (Abadie et al., 2010). Under an autoregressive model, improving pre-treatment fit directly reduces bias, and the Ridge ASCM penalty term negotiates a bias-variance trade-off. Under a latent factor model, improving pre-treatment 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 pre-treatment 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 post-treatment 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 both via 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. We implement the proposed methodology in the `augsynth` package for `R`, available at https://github.com/ebenmichael/augsynth.

The paper 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 appendix 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 et al. (2010, 2015) and is the subject of an extensive methodological literature; see Abadie (2019) and Samartsidis et al. (2019) for recent reviews. We briefly highlight some relevant aspects of this literature.

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

A second set of papers relaxes constraints imposed in the original SCM problem, in particular the restriction that control unit weights be non-negative. Doudchenko and Imbens (2017) argue that there are many settings in which negative weights would be desirable. Amjad et al. (2018) propose an interesting variant that combines negative weights with a pre-processing step. Powell (2018) instead allows 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 propose 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 propose bias correcting SCM using regression. Kellogg et al. (2020) propose using a weighted average of SCM and matching, trading off interpolation and extrapolation bias. Arkhangelsky et al. (2019) propose the *Synthetic Difference-in-Differences* estimator, which can be seen as a special case 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 Synthetic Control Method

2.1 Notation and setup

We consider the canonical SCM panel data setting with $i = 1, \ldots, N$ units observed for $t = 1, \ldots, T$ 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 et al. (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 post-treatment 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 pre-treatment outcomes serve as covariates in SCM, we use X_{it}, for $t \leq T_0$, to represent pre-treatment outcomes; we use the terms *pre-treatment 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 pre-treatment 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 pre-treatment outcomes. The data structure is then:

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

2.2 Assumptions on the data generating process

We now give assumptions on the underlying data generating processes (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 $\varepsilon_{it} \sim P(\cdot)$:

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

This setup encompasses many common panel data models; see Chernozhukov et al. (2019) for an extended discussion. Here we consider two specific versions of Equation (3): (a) for post-treatment time T, $Y_{iT}(0)$ is linear in its lagged values; and (b) for all $t = 1, \ldots, T$, $Y_{it}(0)$ is linear in a set of latent factors. In the Appendix, 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:

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

(b) There are J unknown, latent time-varying factors at time $t = 1, \ldots, T$, $\boldsymbol{\mu}_t = \{\mu_{jt}\} \in \mathbb{R}^J$, 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:

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

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

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

We consider both the time-varying factors $\boldsymbol{\mu}_t$ and the unit-varying factor loadings $\boldsymbol{\phi}_i$ to be non-random 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 post-treatment outcomes; furthermore, there cannot be serial correlation between the post-treatment and pre-treatment noise. This DGP includes the special case of an auto-regressive 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* pre-treatment outcomes. We discuss this in more detail in the context of our application in Section 7.

2.3 Synthetic Control Method

The Synthetic Control Method 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 et al., 2010, 2015). Weights are chosen to balance pre-treatment outcomes and possibly other covariates. We consider a version of SCM that chooses weights $\boldsymbol{\gamma}$ as a solution to the constrained optimization problem:

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

where the constraints limit $\boldsymbol{\gamma}$ to the simplex $\Delta^{N_0} = \{\boldsymbol{\gamma} \in \mathbb{R}^{N_0} \mid \gamma_i \geq 0 \; \forall i, \; \sum_i \gamma_i = 1\}$, where $\boldsymbol{V}_{\boldsymbol{x}} \in \mathbb{R}^{T_0 \times T_0}$ is a symmetric importance matrix and $\|\boldsymbol{V}_{\boldsymbol{x}}^{1/2}(\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\gamma)\|_2^2 \equiv (\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\gamma)'\boldsymbol{V}_{\boldsymbol{x}}(\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma})$ is the 2-norm on $\mathbb{R}^{T_0}$ after applying $\boldsymbol{V}_{\boldsymbol{x}}^{1/2}$ 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 $\boldsymbol{V}_{\boldsymbol{x}}$ to be the identity matrix. The simplex constraint in Equation (8) ensures that the weights will be sparse and non-negative; Abadie et al. (2010, 2015) argue 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 et al. (2015). The choice of penalty is less central when weights are 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 pre-treatment fit, and the resulting estimator has many attractive properties. In this setting, Abadie et al. (2010) show that SCM will be unbiased under the auto-regressive 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) pre-treatment fit is not always feasible with weights constrained to be on the simplex (see Ferman and Pinto, 2018). When "the pre-treatment fit is poor or the number of pre-treatment periods is small," Abadie et al. (2015) recommend against using SCM. And even if the pre-treatment fit is excellent, Abadie et al. (2010, 2015) propose 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 pre-treatment fit. Let $\hat{m}_{iT}$ be an estimator for m_{iT}, the model component of the post-treatment control potential outcome. The *Augmented SCM* (ASCM) estimator for $Y_{1T}(0)$ is:

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

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

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 pre-treatment 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 pre-treatment outcomes $\hat{m}(\cdot)$. Intuitively, since $\hat{m}$ estimates the post-treatment 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 consider a bias-corrected estimator of this form.

Equation (10), by contrast, is analogous to standard doubly robust estimation (Robins et al., 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.

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: (1) when $\hat{m}$ is linear in pre-treatment outcomes; and (2) when $\hat{m}$ is linear in the comparison units' post-treatment 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 pre-treatment outcomes, $\hat{m}(\boldsymbol{X}) = \hat{\eta}_0 + \hat{\boldsymbol{\eta}} \cdot \boldsymbol{X}$. The augmented estimator (9) is then:

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

Pre-treatment periods that are more predictive of the post-treatment 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 pre-treatment 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 post-treatment 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' post-treatment outcomes, $\hat{m}(\boldsymbol{X}) = \sum_{W_i=0} \hat{\alpha}_i(\boldsymbol{X}) Y_{iT}$, 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:

$$\hat{Y}_{1T}^{\text{aug}}(0) = \sum_{W_i=0} \left(\hat{\gamma}_i^{\text{scm}} + \hat{\gamma}_i^{\text{adj}}\right) Y_{iT}, \quad \text{where} \quad \hat{\gamma}_i^{\text{adj}} \equiv \hat{\alpha}_i(\boldsymbol{X}_1) - \sum_{W_j=0} \hat{\gamma}_j^{\text{scm}} \hat{\alpha}_i(\boldsymbol{X}_j). \tag{12}$$

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 pre-treatment time periods equally, and yields a weighted difference-in-differences estimator:

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

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 extend 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 post-treatment 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 post-treatment outcomes $\boldsymbol{Y}_{0T}$ on centered pre-treatment outcomes $\boldsymbol{X}_{0\cdot}$ with penalty hyper-parameter λ^{ridge}:

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

The Ridge Augmented SCM estimator is then:

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

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 pre-treatment 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 $f(\gamma_i) = (\gamma_i - \hat{\gamma}_i^{\text{scm}})^2$ that penalizes *deviations from the SCM weights.*

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

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

where

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

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

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

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 pre-treatment 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:

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

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 1a 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 towards an exact fit solution as λ^{ridge} is reduced. Figure 1b 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 non-negative 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 pre-treatment fit with SCM is possible, Abadie et al. (2015) argue 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 pre-treatment fit relative to SCM alone

Just as the hyper-parameter λ^{ridge} parameterizes the level of extrapolation, it also parameterizes the level of improvement in pre-treatment fit over the SCM solution. Because we are removing the non-negativity constraint and allowing for extrapolation outside of the convex hull, the pre-treatment fit from Ridge ASCM will be at least as good as the pre-treatment fit from SCM alone, i.e., $\|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{aug}}\|_2 \leq \|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}}\|_2$. We can exactly characterize the pre-treatment 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 $\tilde{\boldsymbol{X}}_i = \boldsymbol{V}'\boldsymbol{X}_i$ 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 hyper-parameter $\lambda^{\text{ridge}} = \lambda N_0$ satisfy

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

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

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

From Equation (20), we see that the pre-treatment 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 pre-treatment fit than SCM alone, except in corner cases where pre-treatment fit will be equal, such as when the pre-treatment 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 pre-treatment 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 pre-treatment fit to improved estimation error under the data generating processes in Section 2.2. Under a linear model, improving pre-treatment fit directly reduces bias, and the Ridge ASCM penalty term negotiates a bias-variance trade-off. Under a latent factor model, improving pre-treatment 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 hyper-parameter λ^{ridge} is important in practice. We describe a cross-validation hyper-parameter selection procedure in Section 5.3. Finally, we discuss inference in Section 5.4.

5.1 Error under linearity in pre-treatment outcomes

We first illustrate the key balancing idea in the simple case in our first DGP, where the post-treatment 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 post-treatment 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: (1) systemic error due to imbalance in the lagged outcomes $\boldsymbol{X}$, and (2) idiosyncratic error due to the noise in the post-treatment period:

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

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, $(\beta_1, \dots, \beta_{T_0-K-1}) = 0$, it is sufficient to balance only the first K lags.

If the weighting estimator does not perfectly balance the pre-treatment outcomes $\boldsymbol{X}$, 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 post-treatment period $\boldsymbol{\varepsilon}_T$.

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

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

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 Appendix gives a deterministic bound for $\|\hat{\boldsymbol{\gamma}}^{\text{aug}}\|_2$. See Athey et al. (2018) for analogous results on balancing weights in high dimensional cross-sectional settings.

In the special case that SCM weights have perfect pre-treatment fit, ASCM and SCM weights will be equivalent, and the estimation error will only be due to the variance of the weights and post-treatment noise. When SCM weights do not achieve perfect pre-treatment fit, Ridge ASCM with finite λ extrapolates outside the convex hull, improving pre-treatment 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 et al. (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:

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

Balancing the observed pre-treatment outcomes $\boldsymbol{X}$ will not necessarily balance the latent factor loadings $\boldsymbol{\phi}$. Following Abadie et al. (2010), we show in the appendix that, under Equation (6), we can decompose the imbalance term as:

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

where $\boldsymbol{\varepsilon}_{i(1:T_0)} = (\varepsilon_{i1}, \ldots, \varepsilon_{iT_0})$ is the vector of pre-treatment 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 pre-treatment 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 post-treatment 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

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

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 pre-treatment fit, considered by Abadie et al. (2010), the ASCM and SCM weights are equivalent and the error is only due to post-treatment 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 pre-period 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 pre-treatment 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 pre-treatment 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 pre-treatment 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 Appendix, 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 support the use of ASCM in this instance, despite what visually appears to be good pre-treatment fit for SCM.

5.3 Hyper-parameter selection

We propose a cross-validation approach for selecting λ inspired by the in-time placebo check of Abadie et al. (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 (17). Abadie et al. (2015) propose 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:

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

We can then choose λ to minimize $CV(\lambda)$ or follow a more conservative approach such as the "one-standard-error" rule (Hastie et al., 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 post-treatment time T. Finally, only excluding time period t might be inappropriate for some outcome models, e.g. 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 synthetic control method and variants, going beyond the original proposal in Abadie and Gardeazabal (2003) and Abadie et al. (2010, 2015); see, for example, Li (2019), Toulis and Shaikh (2018), Cattaneo et al. (2019), and Chernozhukov et al. (2018).

We focus here on the conformal inference approach of Chernozhukov et al. (2019), which has three key steps. First, for a given sharp null hypothesis, $H_0 : \tau = \tau_0$, we create an adjusted post-treatment outcome for the treated unit $\tilde{Y}_{1T} = Y_{1T} - \tau_0$ and extend the original data set 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 pre-treatment 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 et al., 2005) for $Y_{1T}(0)$ by using the quantiles of pre-treatment residuals:

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

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

Chernozhukov et al. (2019) provide 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. First, Chernozhukov et al. (2019) show 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 et al. (2019) provide a finite sample bound that relates in-sample prediction error to the validity of $p(\tau_0)$. In Appendix A, 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 pre-treatment 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 data generating processes 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 pre-treatment 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:

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

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 pre-treatment 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 the two-step procedure: (1) residualize the pre- and post-treatment outcomes on the auxiliary covariates $\boldsymbol{Z}$; and (2) 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 (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

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

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 pre-treatment 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

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

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, we consider the performance of this estimator when the outcomes follow a linear factor model with either a linear or a non-linear 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 pre-treatment fit. When covariates enter non-linearly, however, there is additional approximation error due to the linear regression specification. Thus, it is important to appropriately transforming 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 for additional details. We also present results from three additional DGPs, each calibrated to estimates from the same data: (1) the factor model with quadruple the standard deviation of the noise term, (2) a unit and time fixed effects model, and (3) an autoregressive model with 3 lags.

We explore the role of augmentation using different outcome estimators. For each DGP, we consider five estimators: (1) SCM alone, (2) ridge regression alone, (3) Ridge ASCM, (4) fixed effects alone, and (5) 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 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. Appendix Figure F.1 shows the corresponding estimator root mean squared error (RMSE).

There are several takeaways. First, augmenting SCM with a ridge outcome regression reduces bias relative to SCM alone — *without* conditioning on excellent pre-treatment fit —

in all four simulations. This underscores the importance of the recommendation in Abadie et al. (2010, 2015) to use SCM only in settings with excellent pre-treatment 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, Appendix Figure F.1 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 pre-treatment 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 under-regularize with ASCM, as evident in the RMSE achieving a minimum for an intermediate value of λ. When pre-treatment 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 post-treatment 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 data generating processes. 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 pre-treatment 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: (1) SCM alone fit on the entire vector of lagged outcomes, (2) Ridge ASCM, (3) Ridge ASCM including auxiliary covariates in parallel to lagged outcomes and (4) 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 Appendix Figure F.6. The covariates we include are the pre-treatment averages of (1) log state and local revenue per capita, (2) log average weekly wages, (3) number of establishments per capita, (4) the employment level, and (5) log GSP per capita.

These estimators assume that noise is mean zero (Assumption 1). Substantively, under the auto-regressive model in Assumption 1(a) this assumes that post-treatment 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 pre-treatment shocks. This also rules out unobserved confounders that affect both post-treatment 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 et al. (2019). Figure 6 shows the log GSP per capita for both Kansas and its synthetic control using SCM and Ridge ASCM. Appendix F shows additional results.

First, the pre-treatment fit for SCM alone is relatively good for most of the pre-period, with an overall pre-treatment RMSE of about 0.9 log points. However, the fit for SCM alone worsens for in 2004–2005, with imbalances of over 4 log points — a pre-treatment imbalance as large as the estimated impact. Using ridge regression to assess the possible implications of this pre-treatment imbalance, we estimate bias due to pre-treatment 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 Appendix Figure F.9. 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 pre-treatment fit, with a pre-treatment 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 pre-treatment 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 pre-treatment fit and balance on the covariates; see Figure 7a. Finally, balancing the auxiliary covariates via residualization also improves pre-treatment 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.

To check against over-fitting, Appendix Figures F.10, F.11, and F.12 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 7a shows the covariate balance for the four estimators. While SCM and Ridge ASCM achieve excellent fit for the pre-treatment 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 pre-period. 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 pre-treatment fit on the lagged outcomes as shown in Figure 5.

Finally, Figure 7b 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 Appendix Figure F.14). 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 non-negligible negative weights for only a few other donor units. By contrast, Appendix Figure F.13 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 pre-treatment fit is possible. For settings when this is infeasible, we introduce Augmented SCM, which controls pre-treatment 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 could 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 papers 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 et al., 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 et al. (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.

References

Abadie, A. (2019). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*.

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association 105*(490), 493–505.

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

Abadie, A. and J. Gardeazabal (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *The American Economic Review 93*(1), 113–132.

Abadie, A. and G. W. Imbens (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics 29*(1), 1–11.

Abadie, A. and J. L'Hour (2018). A penalized synthetic control estimator for disaggregated data.

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

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2019). Synthetic difference in differences. *arXiv preprint arXiv:1812.09970*.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2017). Matrix Completion Methods for Causal Panel Data Models. *arxiv 1710.10251*.

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

Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(4), 597–623.

Ben-Michael, E., A. Feller, and J. Rothstein (2019). Synthetic controls and weighted event studies with staggered adoption. *arXiv preprint arXiv:1912.03290*.

Cattaneo, M. D., Y. Feng, and R. Titiunik (2019). Prediction intervals for synthetic control methods. *arXiv preprint arXiv:1912.07120*.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2018). Inference on average treatment effects in aggregate panel data settings. *arXiv preprint arXiv:1812.10820*.

Chernozhukov, V., K. Wüthrich, and Y. Zhu (2019). An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls. Technical report.

Doudchenko, N. and G. W. Imbens (2017). Difference-In-Differences and Synthetic Control Methods: A Synthesis. *arxiv 1610.07748*.

Ferman, B. (2019). On the Properties of the Synthetic Control Estimator with Many Periods and Many Controls.

Ferman, B. and C. Pinto (2018). Synthetic controls with imperfect pre-treatment fit.

Gobillon, L. and T. Magnac (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics 98*(3), 535–551.

Hastie, T., J. Friedman, and R. Tibshirani (2009). *The elements of statistical learning*. Springer series in statistics New York.

Hsiao, C., Q. Zhou, et al. (2018). Panel parametric, semi-parametric and nonparametric construction of counterfactuals-california tobacco control revisited. Technical report.

Kellogg, M., M. Mogstad, G. Pouliot, and A. Torgovitsky (2020). Combining matching and synthetic controls to trade off biases from extrapolation and interpolation. Technical report, National Bureau of Economic Research.

King, G. and L. Zeng (2006). The dangers of extreme counterfactuals. *Political Analysis 14*(2), 131–159.

Li, K. T. (2019). Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, 1–16.

Neyman, J. (1990 [1923]). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472.

Powell, D. (2018). Imperfect synthetic controls: Did the massachusetts health care reform save lives?

Rickman, D. S. and H. Wang (2018). Two tales of two us states: Regional fiscal austerity and economic performance. *Regional Science and Urban Economics 68*, 46–55.

Robbins, M., J. Saunders, and B. Kilmer (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*(517), 109–126.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*(427), 846–866.

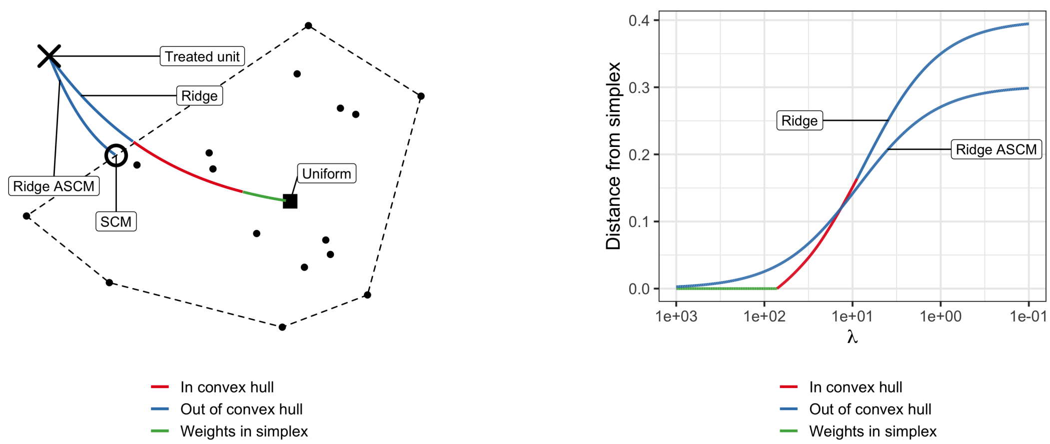
Samartsidis, P., S. R. Seaman, A. M. Presanis, M. Hickman, D. De Angelis, et al. (2019). Assessing the causal effect of binary interventions from observational panel data with few treated units. *Statistical Science 34*(3), 486–503.

Soriano, D., E. Ben-Michael, P. Bickel, A. Feller, and S. Pimentel (2020). Sensitivity analysis for balancing weights. Technical report. working paper.

Toulis, P. and A. Shaikh (2018). Randomization tests in observational studies with time-varying adoption of treatment.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic learning in a random world.* Springer.

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



(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.

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.

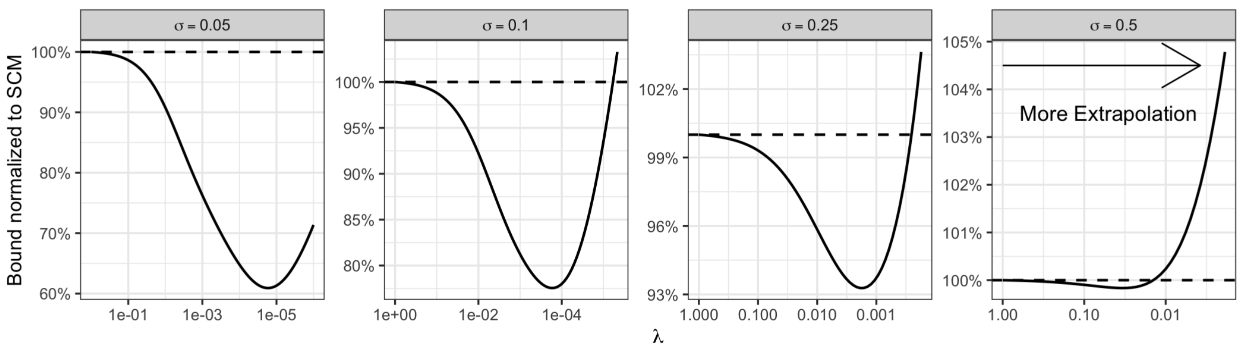


Figure 2: Sketch of the error due to imbalance and approximation error (26) for the linear factor model; the standard deviation of the treated unit's pre-treatment outcomes is normalized to one. We fit SCM weights on the empirical example in Section 7 and compute the vector of pre-treatment fit. Each 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.

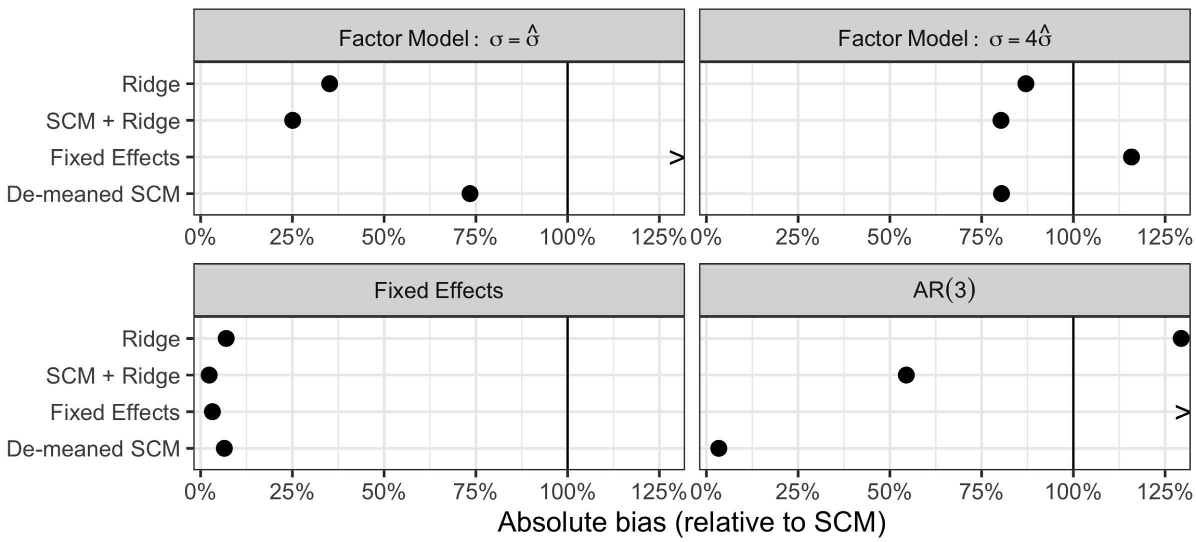


Figure 3: Overall absolute bias, normalized to SCM bias for (a) the factor model simulation, (b) the factor model simulation with quadruple the standard deviation, (c) the fixed effects simulation, and (d) the AR simulation. The SCM estimates reported here are *not* restricted to simulation draws with excellent pre-treatment fit; Abadie et al. (2015) advise against using SCM in such settings.

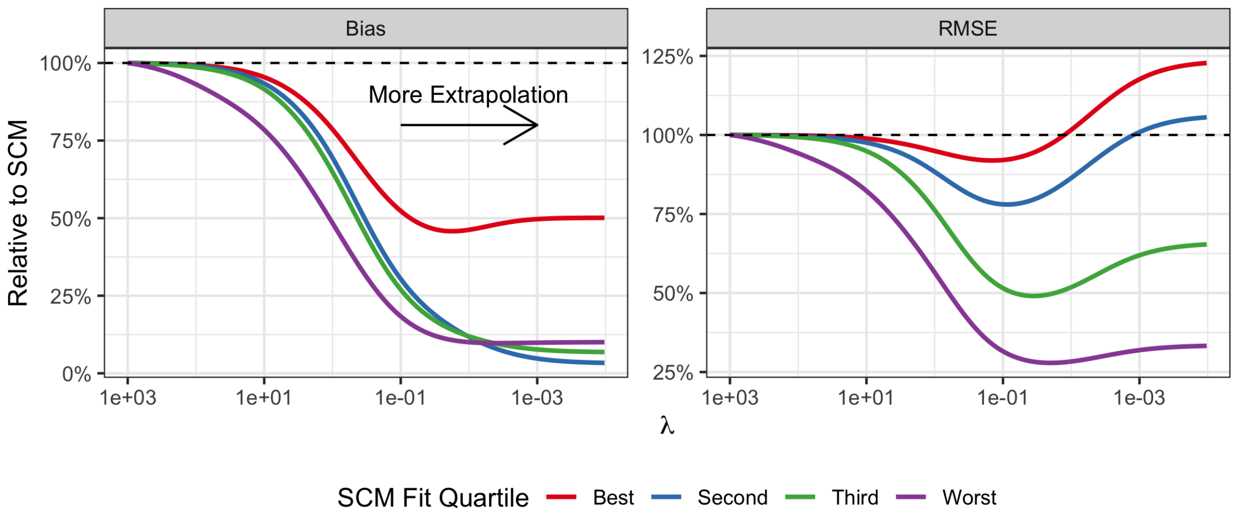


Figure 4: Bias and RMSE of Ridge ASCM, as a percentage of SCM bias and RMSE, versus λ under a linear factor model. Results are divided by the quartile of the SCM fit across all simulations.

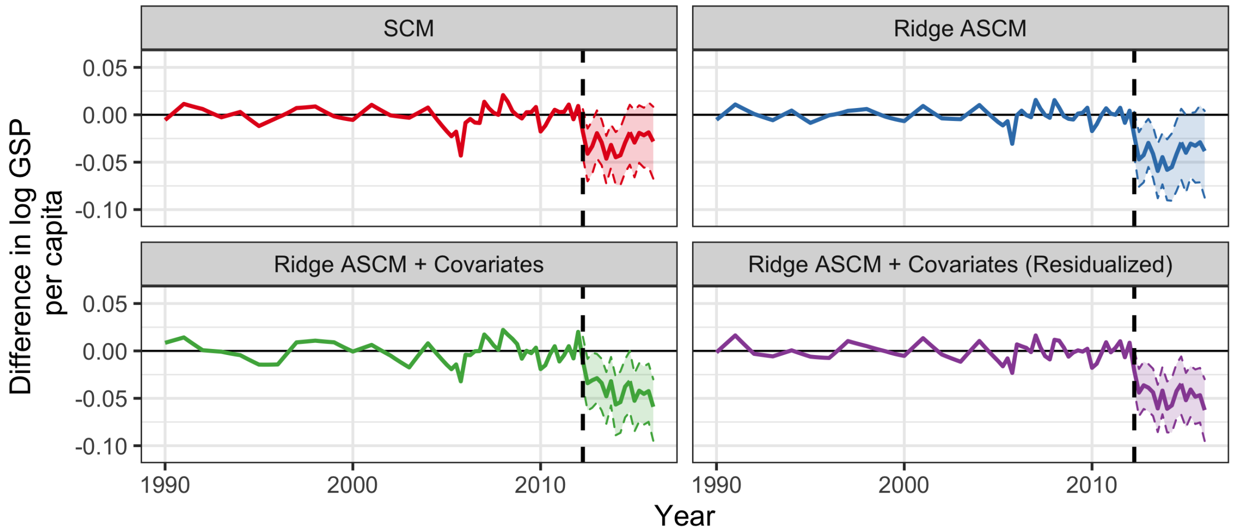


Figure 5: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on log GSP per capita using SCM, Ridge ASCM, and Ridge ASCM with covariates.

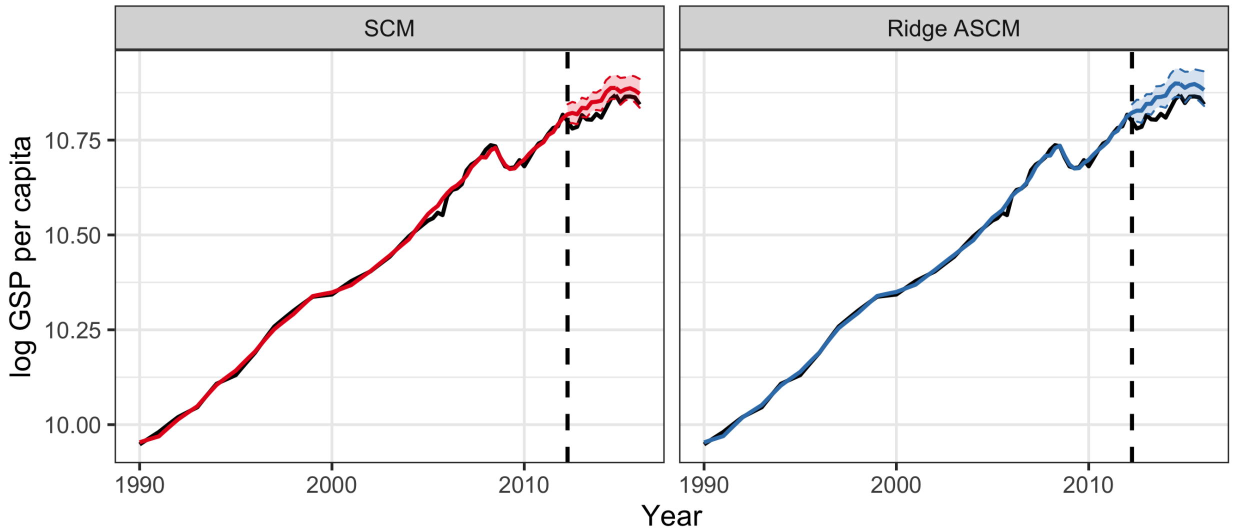
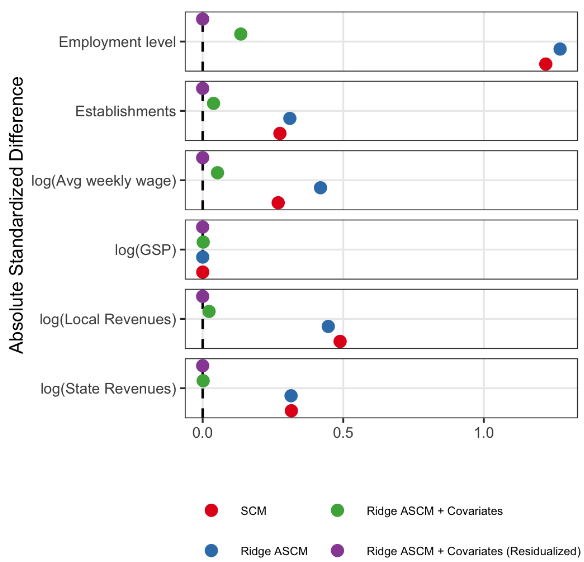
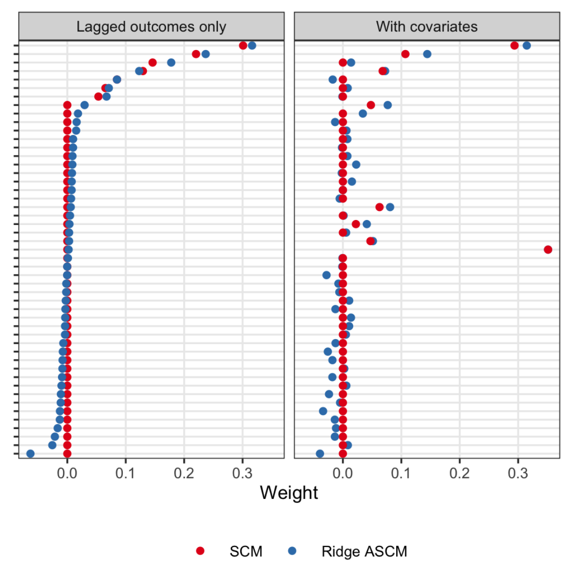


Figure 6: Point estimates along with point-wise 95% conformal prediction intervals for counterfactual log GSP per capita without the tax cuts using SCM, ridge ASCM, and ridge ASCM with covariates, plotting with the observed log GSP per capita in black.



(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 fact includes auxiliary covariates.

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

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

Supplementary Materials for
"The Augmented Synthetic Control Method"

A Inference

We now give additional technical details for the validity of the conformal inference approach of Chernozhukov et al. (2019) with Ridge ASCM, showing approximate validity (as $T_0 \to \infty$) under a set of assumptions.

The approximate validity of the conformal inference procedure in Section 5.4 depends on the predictive accuracy of $\hat{Y}_{it}^{\text{aug}}(0)$ when fit using all periods $t = 1, \ldots, T$, including the post-treatment period T. Denoting $\mathbf{Y}_{\mathbf{1}\cdot} \equiv (\mathbf{X}_{1\cdot}, Y_1) \in \mathbb{R}^T$ to be the full vector of treated unit outcomes and $\mathbf{Y}_{0\cdot} \equiv [\mathbf{X}_{0\cdot}, \mathbf{Y}_{0T}] \in \mathbb{R}^{N_0 \times T}$ be the matrix of comparison unit outcomes, the Ridge ASCM optimization problem in this setting is

$$\min_{\boldsymbol{\gamma} \text{ s.t. } \sum_i \gamma_i = 1} \frac{1}{2\lambda^{\text{ridge}}} \|\boldsymbol{Y}_{1\cdot} - \mathbf{Y}_{0\cdot}' \boldsymbol{\gamma}\|_2^2 + \frac{1}{2} \|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}^{\text{scm}}\|_2^2 . \tag{A.1}$$

We will also consider the constrained form:

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \quad & \|\boldsymbol{Y}_{1\cdot} - \mathbf{Y}_{0\cdot}' \boldsymbol{\gamma}\|_2^2 \\ \text{subject to} \quad & \frac{1}{2} \|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}^{\text{scm}}\|_2 \leq \frac{C}{\sqrt{N_0}} \\ & \sum_i \gamma_i = 1 \end{aligned} \tag{A.2}$$

With these definitions we can characterize the in-sample prediction error of the counterfactual model described by Chernozhukov et al. (2019), which is a version of Equation (3) in an asymptotic framework where T_0 is growing while T is fixed. We state the model and assumptions for asymptotically (in T_0) valid inference below.

Assumption A.1. There exist weights $\gamma^* \in \Delta^{N_0}$ such that the potential outcomes under control for the treated unit ($i = 1$) satisfy

$$Y_{1t}(0) = \sum_{W_i = 1} \gamma_i^* Y_{it} + \varepsilon_{1t},$$

where ε_{1t} are independent of the comparison unit outcomes, $\mathbb{E}[\varepsilon_{1t} Y_{it}] = 0$ for all $W_i = 0$ and $t = 1, \ldots, T$. Furthemore,

1. The data is β-mixing with exponential speed

2. There exist constants $c_1, c_2 > 0$ such that $\mathbb{E}[(Y_{it}\varepsilon_{1t})^2] \geq c_1$ and $\mathbb{E}[|Y_{it}\varepsilon_{1t}|^3] \leq c_2$ for all i such that $W_i = 0$ and $t = 1, \ldots, T$

3. For all i such that $W_i = 0$, $X_{i1}\varepsilon_{11}, \ldots, X_{iT}\varepsilon_{1T}$ is β-mixing with β-mixing coefficient satisfying $\beta(t) \leq a_1 e^{-a^2 t^k}$ for constants $a_1, a_2, k > 0$

4. There exists a constant $c_3 > 0$ such that $\max_{W_i=0} \sum_{t=1}^T X_{it}^2 \varepsilon_{1t}^2 \leq c_3^2 T$ with probability $1 - o(1)$

5. $\log N_0 = o\left(T^{\frac{4k}{3k+4}}\right)$

6. There exists a sequence $\ell_T > 0$ such that $\boldsymbol{Y}_{0t}'(w - \gamma^*) \leq \ell_T \frac{1}{T} \left\|\boldsymbol{Y}_{0\cdot}'(w - \gamma^*)\right\|_2^2$ for all $w \in \Delta^{N_0} + B_2(\frac{C}{\sqrt{N_0}})$, for some constant C where $B_p(a) = \{x \in \mathbb{R} \mid \|x\|_p \leq a\}$, with probability $1 - o(1)$ for $T_0 + 1 \leq t \leq T$

7. The sequence ℓ_T satisfies $\ell_T \left(\log \min\{T, N_0\}\right)^{\frac{1+k}{2k}} \sqrt{T} \to 0$

This setup is nearly identical to the assumptions in Lemma 1 in Chernozhukov et al. (2018); the only key change is for assumption 6 where the bound on the point-wise prediction error is assumed to hold for all weights that are the sum of weights on the simplex Δ^{N_0} and a vector in the L2 ball $B_2\left(\frac{C}{\sqrt{N_0}}\right)$.

Under the model in Assumption A.1, we can characterize the prediction error of the constrained form of Ridge ASCM (A.2) by directly following the development in Chernozhukov et al. (2019), who show asymptotic validity for the conformal procedure with the SCM estimator when it is correctly specified and $\gamma^* \in \Delta^{N_0}$. Lemma A.1 below is equivalent to Lemma 1 in Chernozhukov et al. (2019), and shows that under Assumption A.1 the in-sample prediction error for the constrained form of Ridge ASCM (A.2) is the same as SCM, up to the level of extrapolation C allowed through the constraint $\|\hat{\gamma}^{\text{aug}} - \hat{\gamma}^{\text{scm}}\|_2 \leq \frac{C}{\sqrt{N_0}}$. Then, by Theorem 1 in Chernozhukov et al. (2019) we see that the inference procedure will be valid asymptotically in T_0.

Lemma A.1. Under Assumption A.1, the ridge ASCM weights solving the constrained problem (A.2), $\hat{\gamma}^{\text{aug}}$ satisfy

$$\frac{1}{T}\sum_{t=1}^T \left(\sum_{W_i=0} \hat{\gamma}_i^* Y_{it} - \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} Y_{it}\right)^2 \leq \frac{K_0(2+C)}{\sqrt{T}} \left(\log \min\{T, N_0\}\right)^{\frac{1+k}{2k}} \tag{A.3}$$

and

$$\left|\mu_T \cdot \phi_1 - \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} Y_{iT}\right| \leq \frac{K_0(2+C)}{\sqrt{T}} \ell_T \left(\log \min\{T, N_0\}\right)^{\frac{1+k}{2k}} \tag{A.4}$$

with probability $1 - o(1)$, for some constant K_0 depending on the constants in Assumption A.1.

Proof of Lemma A.1. This proof directly follows Lemma 1 in Chernozhukov et al. (2019). First, notice that

$$\left\|\boldsymbol{Y}_{1\cdot} - \boldsymbol{Y}_{0\cdot}'\hat{\gamma}^{\text{aug}}\right\|_2^2 \leq \left\|\boldsymbol{Y}_{1\cdot} - \boldsymbol{Y}_{0\cdot}'\hat{\gamma}^{\text{scm}}\right\|_2^2 \leq \left\|\boldsymbol{Y}_{1\cdot} - \boldsymbol{Y}_{0\cdot}'\gamma^*\right\|_2^2 = \|\boldsymbol{\varepsilon}_1\|_2^2,$$

where $\boldsymbol{\varepsilon}_1 = (\varepsilon_{11}, \ldots, \varepsilon_{1T}) \in \mathbb{R}^T$ is the vector of noise terms for the treated unit. Next,

$$\boldsymbol{Y}_{1\cdot} - \boldsymbol{Y}_{0\cdot}'\hat{\gamma}^{\text{aug}} = \boldsymbol{Y}_{1\cdot} - \boldsymbol{Y}_{0\cdot}'(\hat{\gamma}^{\text{aug}} - \gamma^* + \gamma^*) = \boldsymbol{\varepsilon}_1 - \boldsymbol{Y}_{0\cdot}'(\hat{\gamma}^{\text{aug}} - \gamma^*)$$

Together, this implies that $\|\boldsymbol{\varepsilon}_1 - \boldsymbol{Y}_{0\cdot}'(\hat{\gamma}^{\text{aug}} - \gamma^*)\|_2^2 \leq \|\boldsymbol{\varepsilon}_1\|_2^2$ and so by expanding the left-hand side we see that by Hölder's inequality

$$\begin{aligned}
\left\|\boldsymbol{Y}_{0\cdot}'(\hat{\gamma}^{\text{aug}} - \gamma^*)\right\|_2^2 &\leq 2\boldsymbol{\varepsilon}_1'\boldsymbol{Y}_{0\cdot}'(\hat{\gamma}^{\text{aug}} - \gamma^*) \\
&\leq 2\left\|\boldsymbol{Y}_{0\cdot}\boldsymbol{\varepsilon}_1\right\|_\infty \|\hat{\gamma}^{\text{aug}} - \gamma^*\|_1 \\
&\leq 2\left\|\boldsymbol{Y}_{0\cdot}\boldsymbol{\varepsilon}_1\right\|_\infty (\|\hat{\gamma}^{\text{scm}} - \gamma^*\|_1 + \|\hat{\gamma}^{\text{aug}} - \hat{\gamma}^{\text{scm}}\|_1)
\end{aligned}$$

Now, since both $\hat{\gamma}^{\text{scm}} \in \Delta^{N_0}$ and $\gamma^* \in \Delta$, $\|\hat{\gamma}^{\text{scm}} - \gamma^*\|_1 \leq 2$. From the constraint in Equation (A.2), $\|\hat{\gamma}^{\text{aug}} - \hat{\gamma}^{\text{scm}}\|_1 \leq \sqrt{N_0}\,\|\hat{\gamma}^{\text{aug}} - \hat{\gamma}^{\text{scm}}\|_2 \leq C$. This implies that

$$\left\|\boldsymbol{Y}_{0\cdot}'(\hat{\gamma}^{\text{aug}} - \gamma^*)\right\|_2^2 \leq 2(2 + C)\left\|\boldsymbol{Y}_{0\cdot}\boldsymbol{\varepsilon}_1\right\|_\infty$$

Lemma 17 in Chernozhukov et al. (2019) shows that

$$P\left(\left\|\boldsymbol{Y}_{0\cdot}\boldsymbol{\varepsilon}_1\right\|_\infty > K_0\left(\log\min\{T, N_0\}\right)^{\frac{1+k}{2k}}\sqrt{T}\right) = o(1).$$

Combining the pieces gives Equation (A.3). Next, combining Equation (A.3) with Assumption A.1(6) gives Equation (A.4). $\square$

B Additional results

B.1 Specialization of Ridge ASCM results to SCM

This appendix section specializes select results from the main text for Ridge ASCM for the special case of SCM, with $\lambda \to \infty$.

First we specialize Proposition 1 to SCM weights by taking $\lambda \to \infty$.

Corollary A.1. Under the linear model (4) with independent sub-Gaussian noise with scale parameter σ, for any $\delta > 0$, for weights $\boldsymbol{\gamma} \in \Delta^{N_0}$ independent of the post-treatment outcomes $(Y_{1T}, \ldots, Y_{NT})$ and for any $\delta > 0$,

$$Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i Y_{iT} \leq \underbrace{\|\boldsymbol{\beta}\|_2 \left\| \boldsymbol{X}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{X}_i \right\|_2}_{\text{imbalance in} \boldsymbol{X}} + \underbrace{\delta\sigma\left(1 + \|\hat{\boldsymbol{\gamma}}\|_2\right)}_{\text{post-treatment noise}}, \tag{A.5}$$

with probability at least $1 - 2e^{-\frac{\delta^2}{2}}$.

We can similarly specialize Theorem 1.

Corollary A.2. Under the linear factor model (6) with independent sub-Gaussian noise with scale parameter σ, for weights $\boldsymbol{\gamma} \in \Delta^{N_0}$ independent of the post-treatment outcomes $(Y_{1T}, \ldots, Y_{NT})$ and for any $\delta > 0$,

$$Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i Y_{iT} \leq \underbrace{\frac{JM^2}{\sqrt{T_0}} \left\| \boldsymbol{X}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{X}_i \right\|_2}_{\text{imbalance in} \boldsymbol{X}} + \underbrace{\frac{2JM^2\sigma}{\sqrt{T_0}} \left(\sqrt{\log 2N_0} + \delta\right)}_{\text{approximation error}} + \underbrace{\delta\sigma\left(1 + \|\hat{\boldsymbol{\gamma}}\|_2\right)}_{\text{post-treatment noise}}, \tag{A.6}$$

with probability at least $1 - 6e^{-\frac{\delta^2}{2}}$.

B.2 Error under a partially linear model with Lipshitz deviations from linearity

We now bound the estimation error for SCM and Ridge ASCM under the basic model (3) when the outcome is only partially linear, with Lipshitz deviations from linearity.

Assumption A.2. For the post-treatment outcome, m_{iT} are generated as $\boldsymbol{\beta} \cdot \boldsymbol{X}_i + f(\boldsymbol{X}_i)$, so the post-treatment control potential outcome is

$$Y_{iT}(0) = \boldsymbol{\beta} \cdot \boldsymbol{X}_i + f(\boldsymbol{X}_i) + \varepsilon_{iT}, \tag{A.7}$$

where $f : \mathbb{R}^{T_0} \to \mathbb{R}$ is L-Lipshitz and where $\{\varepsilon_{iT}\}$ are defined in Assumption 1(a).

Under this model, the L-Lipshitz function $f(\cdot)$ will induce an approximation error from deviating away from the nearest neighbor match.

Theorem A.1. Let $C = \max_{W_i=0} \|\boldsymbol{X}_i\|_2$. Under Assumption A.2, for any $\delta > 0$, the estimation error for the ridge ASCM weights $\hat{\boldsymbol{\gamma}}^{\text{aug}}$ (17) with hyperparameter $\lambda^{\text{ridge}} = N_0\lambda$ is

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

$$\underbrace{CL\left\|\text{diag}\left(\frac{d_j}{d_j^2+\lambda}\right)(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}' \hat{\gamma}^{\text{scm}})\right\|_2}_{\text{excess approximation error}} + \tag{A.8}$$

$$\underbrace{L\sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} \|\boldsymbol{X}_1 - \boldsymbol{X}_i\|_2}_{\text{SCM approximation error}} \quad + \quad \underbrace{\delta\sigma\left(1 + \|\hat{\gamma}^{\text{aug}}\|_2\right)}_{\text{post-treatment noise}}$$

with probability at least $1 - 2e^{-\frac{\delta^2}{2}}$.

We can again specialize this to the SCM weights alone by taking $\lambda \to \infty$.

Corollary A.3. Under Assumption A.2, for any $\delta > 0$, the estimation error for weights on the simplex $\hat{\gamma} \in \Delta^{N_0}$ independent of the post-treatment outcomes $(Y_{1T}, \ldots, Y_{NT})$ is

$$Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i Y_i \leq \|\boldsymbol{\beta}\|_2 \underbrace{\left\|\boldsymbol{X}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{X}_i\right\|_2}_{\text{imbalance in } X} + \underbrace{L\sum_{W_i=0} \hat{\gamma}_i \|\boldsymbol{X}_1 - \boldsymbol{X}_i\|_2}_{\text{approximation error}} + \underbrace{\delta\sigma(1 + \|\hat{\gamma}\|_2)}_{\text{post-treatment noise}} \tag{A.9}$$

with probability at least $1 - 2e^{-\frac{\delta^2}{2}}$.

Inspecting Corollary A.3, we see that in order to control the estimation error, the weights must ensure good pre-treatment fit while only weighting control units that are near to the treated unit. The ratio $L/\|\boldsymbol{\beta}\|_2$ controlling the relative importance of both terms. Abadie and L'Hour (2018) propose finding weights by solving the penalized SCM problem,

$$\min_{\gamma \in \Delta^{N_0}} \left\|\boldsymbol{X}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{X}_i\right\|_2^2 + \lambda \sum_{W_i=0} \hat{\gamma}_i \|\boldsymbol{X}_1 - \boldsymbol{X}_i\|_2^2. \tag{A.10}$$

Comparing this to Corollary A.3, we see that under the partially linear model (A.7) where $f(\cdot)$ is L-Lipshitz, finding weights that limit interpolation error by controling both the overall imbalance in the lagged outcomes as well as the weighted sum of the distances is sufficient to control the error. In the above optimization problem, the hyperparameter λ takes the role of $L/\|\boldsymbol{\beta}\|_2$.

B.3 Error under a linear factor model with covariates

We can quantify the behavior of the two-step procedure from Lemma 4 in controlling the error under a more general form of the linear factor model (6) with covariates (see Abadie et al., 2010; Botosaru and Ferman, 2019, for additional discussion). We can also consider the error under a linear model with auxiliary covariates, as a direct consequence of Lemma 4.

Assumption A.3. The m_{it} are generated as $m_{it} = \sum_{j=1}^{J} \phi_{ij}\mu_{jt} + f_t(\boldsymbol{Z}_i)$ for a time-varying function $f_t : \mathbb{R}^K \to \mathbb{R}$, so the potential outcomes under control are

$$Y_{it}(0) = \sum_{j=1}^{J} \phi_{ij}\mu_{jt} + f_t(\boldsymbol{Z}_i) + \varepsilon_{it}, \tag{A.11}$$

where $\{\varepsilon_{it}\}$ are defined in Assumption 1(b).

To characterize how well the covariates approximate the true function $f(\boldsymbol{Z}_i)$, we will consider the best linear approximation in our data, and define the residual for unit i and time t as $e_{it} = f_t(\boldsymbol{Z}_i) - \boldsymbol{Z}_i'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' f_t(Z)$, where $\boldsymbol{Z} \in \mathbb{R}^{N \times K}$ is the matrix of all auxiliary covariates for all units. For each time period we will characterize the additional approximation error incurred by only balancing the covariates linearly with the *residual sum of squares* $RSS_t = \sum_{i=1}^{n} e_{it}^2$. For ease of exposition, we assume that the control covariates are standardized and rotated, which can always be true after pre-processing, and present results for the simpler case in which we fit SCM on the residualized pre-treatment outcomes rather than ridge ASCM (i.e., we let $\lambda^{\text{ridge}} \to \infty$); the more general version follows immediately by applying Theorem 1.

Theorem A.2. Under the linear factor model with covariates in Assumption A.3, with $\frac{1}{N_0}\boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot} = \boldsymbol{I}_K$, for any $\delta > 0$, $\hat{\gamma}^{\text{cov}}$ in Equation (33) with $\lambda^{\text{ridge}} \to \infty$ satisfies the bound

$$\begin{aligned}\left|Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}^{\text{cov}} Y_{iT}\right| \leq \frac{JM^2}{\sqrt{T_0}} \Bigg(\underbrace{\left\|\check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot}'\hat{\gamma}\right\|_2}_{\text{imbalance in } \check{\boldsymbol{X}}} + \underbrace{4\sigma\sqrt{\frac{K}{N_0}}\|\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}'\hat{\gamma}\|_2}_{\text{excess approximation error}} \Bigg) + \\ \underbrace{\frac{2JM^2\sigma}{\sqrt{T_0}}\left(\sqrt{\log N_0} + \frac{\delta}{2}\right)}_{\text{SCM approximation error}} + \underbrace{(JM^2+1)e_{1\max} + (JM^2+1)\sqrt{RSS_{\max}}\|\hat{\gamma}^{\text{cov}}\|_2}_{\text{covariate approximation error}} \\ + \underbrace{\delta\sigma(1 + \|\hat{\gamma}^{\text{cov}}\|_2)}_{\text{post-treatment noise}}\end{aligned} \tag{A.12}$$

with probability at least $1 - 6e^{-\frac{\delta^2}{2}} - 2e^{-\frac{KN_0(2-\sqrt{\log 5})^2}{2}}$, where $e_{1\max} = \max_t |e_{1t}|$ is the maximal residual for the treated unit and $RSS_{\max} = \max_t RSS_t$ is the maximal residual sum of squares

We can also consider the special case of Theorem A.2 when $f_t(\boldsymbol{Z}_i) = \sum_{k=1}^{K} B_{tk} Z_{ik}$ is a linear function of the covariates, and so

$$Y_{it}(0) = \sum_{j=1}^{J} \phi_{ij}\mu_{jt} + \sum_{k=1}^{K} B_{tk} Z_{ik} + \varepsilon_{it} = \boldsymbol{\phi}_i'\boldsymbol{\mu}_T + \boldsymbol{B}_t'\boldsymbol{Z}_i + \varepsilon_{it}. \tag{A.13}$$

In this case the residuals $e_{it} = 0 \quad \forall i, t$.

Corollary A.4. Under the linear factor model with covariates in Assumption A.3 with $f_t(\boldsymbol{Z}_i) = \sum_{k=1}^{K} B_{tk} Z_{ik}$ as in Equation (A.13), for any $\delta > 0$, $\hat{\gamma}^{\text{cov}}$ in Equation (33) with $\lambda^{\text{ridge}} \to \infty$ satisfies the bound

$$\left| Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}^{\text{cov}} Y_{iT} \right| \leq \frac{JM^2}{\sqrt{T_0}} \left(\underbrace{\left\| \check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot}' \hat{\gamma} \right\|_2}_{\text{imbalance in } \check{\boldsymbol{X}}} + \underbrace{4\sigma \sqrt{\frac{K}{N_0}} \|\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}' \hat{\gamma}\|_2}_{\text{excess approximation error}} \right) + \underbrace{\frac{2JM^2\sigma}{\sqrt{T_0}} \left(\sqrt{\log N_0} + \frac{\delta}{2} \right)}_{\text{SCM approximation error}} + \underbrace{\delta\sigma(1 + \|\hat{\gamma}^{\text{cov}}\|_2)}_{\text{post-treatment noise}} \tag{A.14}$$

with probability at least $1 - 6e^{-\frac{\delta^2}{2}} - 2e^{-\frac{KN_0(2-\sqrt{\log 5})^2}{2}}$.

Building on Lemma 4, Theorem A.2 and Corollary A.4 show that due to the additive, separable structure of the auxiliary covariates in Equation (A.13), controlling the pre-treatment fit in the *residualized* lagged outcomes $\check{\boldsymbol{X}}$ partially controls the error. This justifies directly targeting fit in the residualized lagged outcomes $\check{\boldsymbol{X}}$ rather than targeting raw lagged outcomes $\boldsymbol{X}$. Moreover, the excess approximation error will be small since since the number of covariates K is small relative to N_0 and the auxiliary covariates are measured without noise. As in Section 4.2, we can achieve better balance when we apply ridge ASCM to $\check{\boldsymbol{X}}$ than when we apply SCM alone. Because $\check{\boldsymbol{X}}$ are orthogonal to $\boldsymbol{Z}$ by construction, this comes at no cost in terms of imbalance in $\boldsymbol{Z}$. However, the fundamental challenge of over-fitting to noise still remains, and, as in the case without auxiliary covariates, selecting the tuning parameter remains important. We again propose to follow the cross validation approach in Section 5.3, here using the residualized lagged outcomes $\check{\boldsymbol{X}}$ rather than the raw lagged outcomes $\boldsymbol{X}$.

C Simulation data generating process

We now describe the simulations in detail. We use the Generalized Synthetic Control Method (Xu, 2017) to fit the following linear factor model to the observed series of log GSP per capita ($N = 50, T_0 = 89, T = 105$), setting $J = 3$:

$$Y_{it} = \alpha_i + \nu_t + \sum_{j=1}^{J} \phi_{ij}\mu_{jt} + \varepsilon_{it}. \tag{A.15}$$

We then use these estimates as the basis for simulating data. Appendix Figure F.5 shows the estimated factors $\hat{\boldsymbol{\mu}}$. We use the estimated time fixed effects $\hat{\boldsymbol{\nu}}$ and factors $\hat{\boldsymbol{\mu}}$ and then simulate data using Equation (A.15), drawing:

$$\begin{aligned}
\alpha_i &\sim N(\hat{\bar{\alpha}},\ \hat{\sigma}_\alpha) \\
\phi &\sim N(0,\ \hat{\boldsymbol{\Sigma}}_\phi) \\
\varepsilon_{it} &\sim N(0,\ \hat{\sigma}_\varepsilon),
\end{aligned}$$

where $\hat{\bar{\alpha}}$ and $\hat{\sigma}_\alpha$ are the estimated mean and standard deviation of the unit-fixed effects, $\hat{\boldsymbol{\Sigma}}_\phi$ is the sample covariance of the estimated factor loadings, and $\hat{\sigma}_\varepsilon$ is the estimated residual standard deviation. We also simulate outcomes with quadruple the standard deviation, $\text{sd}(\varepsilon_{it}) = 4\hat{\sigma}_\varepsilon$. We assume a sharp null of zero treatment effect in all DGPs and estimate the ATT at the final time point.

To model selection, we compute the (marginal) propensity scores as

$$\text{logit}^{-1}\{\pi_i\} = \text{logit}^{-1}\{\mathbb{P}(T = 1 \mid \alpha_i, \boldsymbol{\phi}_i)\} = \theta\left(\alpha_i + \sum_j \phi_{ij}\right),$$

where we set $\theta = 1/2$ and re-scale the factors and fixed effects to have unit variance. Finally, we restrict each simulation to have a single treated unit and therefore normalize the selection probabilities as $\frac{\pi_i}{\sum_j \pi_j}$.

We also consider an alternative data generating process that specializes the linear factor model to only include unit- and time-fixed effects:

$$Y_{it}(0) = \alpha_i + \nu_t + \varepsilon_{it}.$$

We calibrate this data generating process by fitting the fixed effects with `gsynth` and drawing new unit-fixed effects from $\alpha_i \sim N(\hat{\bar{\alpha}}, \hat{\sigma}_\alpha)$. We then model selection proportional to the fixed effect as above with $\theta = \frac{3}{2}$. Second, we generate data from an AR(3) model:

$$Y_{it}(0) = \beta_0 + \sum_{j=1}^{3} \beta_j Y_{i(t-j)} + \varepsilon_{it},$$

where we fit $\beta_0, \boldsymbol{\beta}$ to the observed series of log GSP per capita. We model selection as proportional to the last 3 outcomes $\text{logit}^{-1}\pi_i = \theta\left(\sum_{j=1}^{4} Y_{i(T_0-j+1)}\right)$ and set $\theta = \frac{5}{2}$. For this simulation we estimate the ATT at time $T_0 + 1$.

D Proofs

D.1 Proofs for Section 4

Lemma A.2. With $\hat{\eta}_0^{\text{ridge}}$ and $\hat{\boldsymbol{\eta}}^{\text{ridge}}$, the solutions to (14), the ridge estimate can be written as a weighting estimator:

$$\hat{Y}_{1T}^{\text{ridge}}(0) = \hat{\eta}_0^{\text{ridge}} + \hat{\boldsymbol{\eta}}^{\text{ridge}\prime}\boldsymbol{X}_1 = \sum_{W_i=0} \hat{\gamma}_i^{\text{ridge}} Y_{iT}, \tag{A.16}$$

where

$$\hat{\gamma}_i^{\text{ridge}} = \frac{1}{N_0} + (\boldsymbol{X}_1 - \bar{X}_0)'(\boldsymbol{X}_{0\cdot}'\boldsymbol{X}_{0\cdot} + \lambda^{\text{ridge}}\boldsymbol{I}_{T_0})^{-1}\boldsymbol{X}_i. \tag{A.17}$$

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

$$\min_{\boldsymbol{\gamma} \,\mid\, \sum_i \gamma_i = 1} \quad \frac{1}{2\lambda^{\text{ridge}}}\|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma}\|_2^2 + \frac{1}{2}\left\|\boldsymbol{\gamma} - \frac{1}{N_0}\right\|_2^2. \tag{A.18}$$

Proof of Lemmas 1 and A.2. Recall that the lagged outcomes are centered by the control averages. Notice that

$$\begin{aligned}
\hat{Y}_{1T}^{\text{aug}}(0) &= \hat{m}(\boldsymbol{X}_1) + \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}}(Y_{iT} - \hat{m}(\boldsymbol{X}_i)) \\
&= \hat{\eta}_0 + \hat{\eta}'\boldsymbol{X}_1 + \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}}(Y_{iT} - \hat{\eta}_0 - \boldsymbol{X}_i'\hat{\eta}) \\
&= \sum_{W_i=0} (\hat{\gamma}_i^{\text{scm}} + (\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}})(\boldsymbol{X}_{0\cdot}'\boldsymbol{X}_{0\cdot} + \lambda\boldsymbol{I}_{T_0})^{-1}\boldsymbol{X}_i)Y_{iT} \\
&= \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} Y_{iT}
\end{aligned} \tag{A.19}$$

The expression for $\hat{Y}_{1T}^{\text{ridge}}(0)$ follows.

We now prove that $\hat{\boldsymbol{\gamma}}^{\text{ridge}}$ and $\hat{\boldsymbol{\gamma}}^{\text{scm}}$ solve the weighting optimization problems (A.18) and (18). First, the Lagrangian dual to (A.18) is

$$\min_{\alpha,\boldsymbol{\beta}} \frac{1}{2}\sum_{W_i=0}\left(\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i + \frac{1}{N_0}\right)^2 - (\alpha + \boldsymbol{\beta}'\boldsymbol{X}_1) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2, \tag{A.20}$$

where we have used that the convex conjugate of $\frac{1}{2}\left(x - \frac{1}{N_0}\right)^2$ is $\frac{1}{2}\left(y + \frac{1}{N_0}\right)^2 - \frac{1}{2N_0^2}$. Solving for α we see that

$$\sum_{W_i=0} \hat{\alpha} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}_i + 1 = 1$$

Since the lagged outcomes are centered, this implies that

$$\hat{\alpha} = 0$$

Now solving for $\boldsymbol{\beta}$ we see that

$$\boldsymbol{X}_{0\cdot}'\left(\mathbf{1}\frac{1}{N_0} + \boldsymbol{X}_{0\cdot}\hat{\boldsymbol{\beta}}\right) + \lambda\hat{\boldsymbol{\beta}} = \boldsymbol{X}_1$$

This implies that

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_{0\cdot}'\boldsymbol{X}_{0\cdot} + \lambda I)^{-1}\boldsymbol{X}_1$$

Finally, the weights are the ridge weights

$$\hat{\gamma}_i = \frac{1}{N_0} + \boldsymbol{X}_1'(\boldsymbol{X}_{0\cdot}'\boldsymbol{X}_{0\cdot} + \lambda I)^{-1}\boldsymbol{X}_i = \hat{\boldsymbol{\gamma}}_i^{\text{ridge}}$$

Similarly, the Lagrangian dual to (18) is

$$\min_{\alpha,\boldsymbol{\beta}} \frac{1}{2}\sum_{W_i=0}\left(\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i + \hat{\gamma}_i^{\text{scm}}\right)^2 - (\alpha + \boldsymbol{\beta}'\boldsymbol{X}_1) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2, \tag{A.21}$$

where we have used that the convex conjugate of $\frac{1}{2}\left(x - \hat{\gamma}_i^{\text{scm}}\right)^2$ is $\frac{1}{2}\left(y + \hat{\gamma}_i^{\text{scm}}\right)^2 - \frac{1}{2}\hat{\gamma}_i^{\text{scm}2}$. Solving for α we see that $\hat{\alpha} = 0$. Now solving for $\boldsymbol{\beta}$ we see that

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_{0\cdot}'\boldsymbol{X}_{0\cdot} + \lambda I)^{-1}(\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}})$$

Finally, the weights are the ridge ASCM weights

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

□

Proof of Lemma 3. Notice that

$$\begin{aligned}
\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{aug}} &= (I - \boldsymbol{X}_{0\cdot}'\boldsymbol{X}_{0\cdot}(\boldsymbol{X}_{0\cdot}'\boldsymbol{X}_{0\cdot} + N_0\lambda I)^{-1})(\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}}) \\
&= N_0\lambda(\boldsymbol{X}_{0\cdot}'\boldsymbol{X}_{0\cdot} + N_0\lambda I)^{-1}(\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}}) \\
&= \boldsymbol{V}\text{diag}\left(\frac{\lambda}{d_j^2 + \lambda}\right)\boldsymbol{V}'(\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}})
\end{aligned}$$

So since $\boldsymbol{V}$ is orthogonal,

$$\|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{aug}}\|_2 = \left\|\text{diag}\left(\frac{\lambda}{d_j^2 + \lambda}\right)(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}})\right\|_2$$

□

Lemma A.3. The ridge augmented SCM weights with hyperparameter λN_0, $\hat{\boldsymbol{\gamma}}^{\text{aug}}$, satisfy

$$\|\hat{\boldsymbol{\gamma}}^{\text{aug}}\|_2 \leq \|\hat{\boldsymbol{\gamma}}^{\text{scm}}\|_2 + \frac{1}{\sqrt{N_0}}\left\|\text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right)(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}})\right\|_2, \tag{A.22}$$

with $\widetilde{\boldsymbol{X}}_i = V'\boldsymbol{X}_i$ as defined in Lemma 3.

Proof of Lemma A.3. Notice that using the singular value decomposition and by the triangle inequality,

$$\begin{aligned}
\|\hat{\boldsymbol{\gamma}}^{\text{aug}}\|_2 &= \left\|\hat{\boldsymbol{\gamma}}^{\text{scm}} + \boldsymbol{X}_{0\cdot}(\boldsymbol{X}_{0\cdot}'\boldsymbol{X}_{0\cdot} + \lambda I)^{-1}(\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}})\right\|_2 \\
&= \left\|\hat{\boldsymbol{\gamma}}^{\text{scm}} + \boldsymbol{U}\text{diag}\left(\frac{\sqrt{N_0}d_j}{N_0 d_j^2 + \lambda N_0}\right)\boldsymbol{V}'(\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}})\right\|_2 \\
&\leq \|\hat{\boldsymbol{\gamma}}^{\text{scm}}\|_2 + \left\|\text{diag}\left(\frac{d_j}{(d_j^2+\lambda)\sqrt{N_0}}\right)(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}})\right\|_2.
\end{aligned}$$

□

D.2 Proofs for Sections 5, B.1, and B.2

For these proofs we will begin by considering a model where the post-treatment control potential outcomes at time T are linear in the lagged outcomes and include a unit specific term ξ_i.

Assumption A.4. The post-treatment potential outcomes are generated as

$$Y_{iT}(0) = \boldsymbol{\beta} \cdot \boldsymbol{X}_i + \xi_i + \varepsilon_{iT}, \tag{A.23}$$

where $\{\varepsilon_{iT}\}$ are defined as in Assumption 1(a).

Below we will put structure on the unit-specific terms ξ_i, first we write a general finite-sample bound.

Proposition A.1. Under model (A.23) with independent sub-Gaussian noise, for weights $\hat{\boldsymbol{\gamma}}$ independent of the post-treatment residuals $(\varepsilon_{1T}, \ldots, \varepsilon_{NT})$ and for any $\delta > 0$,

$$Y_{1T}(0) - \sum_{W_i=0}\hat{\gamma}_i Y_{iT} \leq \|\boldsymbol{\beta}\|_2 \underbrace{\left\|\boldsymbol{X}_1 - \sum_{W_i=0}\hat{\gamma}_i\boldsymbol{X}_i\right\|_2}_{\text{imbalance in} X} + \underbrace{\left|\xi_1 - \sum_{W_i=0}\hat{\gamma}_i\xi_i\right|}_{\text{approximation error}} + \underbrace{\delta\sigma(1+\|\hat{\boldsymbol{\gamma}}\|_2)}_{\text{post-treatment noise}}, \tag{A.24}$$

with probability at least $1 - 2e^{-\frac{\delta^2}{2}}$.

Proof. First, note that the estimation error is

$$Y_{1T}(0) - \sum_{W_i=0}\hat{\gamma}_i Y_{iT} = \boldsymbol{\beta}\cdot\left(X_1 - \sum_{W_i=0}\hat{\gamma}_i\boldsymbol{X}_i\right) + \left(\rho_1 - \sum_{W_i=0}\hat{\gamma}_i\xi_i\right) + \left(\varepsilon_{1T} - \sum_{W_i=0}\hat{\gamma}_i\varepsilon_{iT}\right) \tag{A.25}$$

Now since the weights are independent of ε_{iT}, by the mean-zero noise assumption in Assumption 1(a) we see that $\varepsilon_{1T} - \sum_{W_i=0}\hat{\gamma}_i\varepsilon_{iT}$ is sub-Gaussian with scale parameter $\sigma\sqrt{1+\|\hat{\boldsymbol{\gamma}}\|_2^2} \leq \sigma\left(1+\|\hat{\boldsymbol{\gamma}}\|_2\right)$. Therefore we can bound the second term:

$$P\left(\left|\varepsilon_{1T} - \sum_{W_i=0}\hat{\gamma}_i\varepsilon_{iT}\right| \geq \delta\sigma\left(1+\|\hat{\boldsymbol{\gamma}}\|_2\right)\right) \leq 2\exp\left(-\frac{\delta^2}{2}\right)$$

The result follows from the triangle inequality and the Cauchy-Schwartz inequality. □

Proof of Proposition 1. Note that under the linear model (4), $\xi_i = 0$ for all i. Now from Lemma 3 we have that

$$\|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\gamma}^{\text{aug}}\|_2 = \left\|\text{diag}\left(\frac{\lambda}{d_j^2 + \lambda}\right)(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}'\hat{\gamma}^{\text{scm}})\right\|_2.$$

Plugging this in to Equation (A.24) completes the proof. □

Proof of Corollary A.1. This is a direct consequence of Proposition A.1 noting that under the linear model (4), $\xi_i = 0$ for all i. □

Random approximation error We now consider the case where ξ_i are random. We can use Proposition A.1 to further bound the approximation error. In particular, we make the following assumption:

Assumption A.5. ξ_i are sub-Gaussian random variables with scale parameter ϖ and are mean-zero, $\mathbb{E}[\xi_i] = 0$ for all $i = 1, \dots, N$.

Lemma A.4. Under Assumption A.5, for weights $\hat{\boldsymbol{\gamma}}$ and any $\delta > 0$ the approximation error satisfies

$$\left|\xi_1 - \sum_{W_i=0}\hat{\gamma}_i\xi_i\right| \leq \delta\varpi + 2\|\hat{\boldsymbol{\gamma}}\|_1\varpi\left(\sqrt{\log 2N_0} + \frac{\delta}{2}\right), \tag{A.26}$$

with probability at least $1 - 4e^{-\frac{\delta^2}{2}}$.

Proof of Lemma A.4. From the triangle inequality and Hölder's inequality we see that

$$\left|\xi_1 - \sum_{W_i=0}\hat{\gamma}_i\xi_i\right| \leq |\xi_1| + \|\hat{\boldsymbol{\gamma}}\|_1 \max_{W_i=0}|\xi_i|.$$

Now since the ξ_i are mean-zero sub-Gaussian with scale parameter ϖ, we have that

$$P\left(|\xi_1| \geq \delta\varpi\right) \leq 2e^{-\frac{\delta^2}{2}}$$

Next, from the union bound, the maximum of the N_0 sub-Gaussian variables $\rho_2, \dots, \rho_N$ satisfies

$$P\left(\max_{W_i=0}|\xi_i| \geq 2\varpi\sqrt{\log 2N_0} + \delta\right) \leq 2e^{-\frac{\delta^2}{2\varpi^2}}.$$

Setting $\delta = \delta\varpi$ and combining the two probabilities with the union bound gives the result. □

Lemma A.5. Under Assumption A.5, for the ridge ASCM weights $\hat{\boldsymbol{\gamma}}^{\text{aug}}$ with hyper-parameter $\lambda^{\text{ridge}} = \lambda N_0$ and for any $\delta > 0$ the approximation error satisfies

$$\left|\xi_1 - \sum_{W_i=0}\hat{\gamma}_i\xi_i\right| \leq 2\varpi\left(\sqrt{\log 2N_0} + \frac{\delta}{2}\right) + \underbrace{(1+\delta)4\varpi\left\|\text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right)\left(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}'\hat{\gamma}^{\text{scm}}\right)\right\|_2}_{\text{excess approximation error}}, \tag{A.27}$$

with probability at least $1 - 4e^{-\frac{\delta^2}{2}} - e^{-2(\log 2 + N_0 \log 5)\delta^2}$.

Proof of Lemma A.5. Again from Hölder's inequality we see that

$$\begin{aligned}\left|\xi_1 - \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} \xi_i\right| &= |\xi_1| + \left|\sum_{W_i=0} (\hat{\gamma}_i^{\text{scm}} + \hat{\gamma}_i^{\text{aug}} - \hat{\gamma}_i^{\text{scm}})\xi_i\right| \\ &\leq |\xi_1| + \|\hat{\boldsymbol{\gamma}}^{\text{scm}}\|_1 \max_{W_i=0} |\xi_i| + \|\hat{\boldsymbol{\gamma}}^{\text{aug}} - \hat{\boldsymbol{\gamma}}^{\text{scm}}\|_2 \sqrt{\sum_{W_i=0} \xi_i^2}.\end{aligned}$$

We have bounded the first two terms in Lemma A.4, now it sufficies to bound the third term. First, from Lemma A.3 we see that

$$\|\hat{\boldsymbol{\gamma}}^{\text{aug}} - \hat{\boldsymbol{\gamma}}^{\text{scm}}\|_2 = \frac{1}{\sqrt{N_0}} \left\|\text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right) \left(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}' \hat{\boldsymbol{\gamma}}^{\text{scm}}\right)\right\|_2.$$

Second, via a standard discretization argument (Wainwright, 2018), we can bound the L^2 norm of the vector $(\xi_2, \ldots, \xi_N)$ as

$$P\left(\sqrt{\sum_{W_i=0} \xi_i^2} \geq 2\varpi\sqrt{\log 2 + N_0 \log 5} + \delta\right) \leq 2\exp\left(-\frac{\delta^2}{2\varpi^2}\right).$$

Setting $\delta = 2\delta\varpi\sqrt{\log 2 + N_0 \log 5}$, noting that $\log 2 + N_0 \log 5 < 4N_0$, we have that

$$\|\hat{\boldsymbol{\gamma}}^{\text{aug}} - \hat{\boldsymbol{\gamma}}^{\text{scm}}\|_2 \sqrt{\sum_{W_i=0} \xi_i^2} \leq (1+\delta)\varpi 4 \left\|\text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right) \left(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}' \hat{\boldsymbol{\gamma}}^{\text{scm}}\right)\right\|_2$$

with probability at least $1 - 2e^{-2(\log 2 + N_0 \log 5)\delta^2}$. Since $\|\hat{\boldsymbol{\gamma}}^{\text{scm}}\|_1 = 1$, combining with Lemma A.4 via the union bound gives the result.

□

Theorem A.3. Under Assumptions A.4 and A.5 model (A.23), for $\hat{\gamma}$ independent of the post-treatment outcomes $(Y_{1T}, \ldots, Y_{NT})$ and for any $\delta > 0$,

$$Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i Y_{iT} \leq \|\boldsymbol{\beta}\|_2 \underbrace{\left\|\boldsymbol{X}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{X}_i\right\|_2}_{\text{imbalance in } X} + \underbrace{\delta\varpi + 2\|\hat{\gamma}\|_1 \varpi\left(\sqrt{\log 2N_0} + \frac{\delta}{2}\right)}_{\text{approximation error}} + \underbrace{\delta\sigma\left(1 + \|\hat{\gamma}\|_2\right)}_{\text{post-treatment noise}}, \tag{A.28}$$

with probability at least $1 - 6e^{-\frac{\delta^2}{2}}$.

Proof of Theorem A.3. The Theorem directly follows from Proposition A.1 and Lemma A.4, combining the two probabilistic bounds via the union bound. □

Theorem A.4. Under Assumptions A.4 and A.5 model (A.23), for any $\delta > 0$, the ridge ASCM weights with hyperparameter $\lambda^{\text{ridge}} = \lambda N_0$ satisfy the bound

$$Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i Y_{iT} \leq \|\boldsymbol{\beta}\|_2 \underbrace{\left\| \text{diag}\left(\frac{\lambda}{d_j^2+\lambda}\right)\left(\widetilde{\boldsymbol{X}}_1 - \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} \widetilde{\boldsymbol{X}}_i\right) \right\|_2}_{\text{imbalance in } X} + \underbrace{2\varpi\left(\sqrt{\log 2N_0} + \frac{\delta}{2}\right)}_{\text{approximation error}}$$
$$\underbrace{(1+\delta)4\varpi \left\| \text{diag}\left(\frac{d_j}{d_j^2+\lambda}\right)\left(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}' \hat{\gamma}^{\text{scm}}\right) \right\|_2}_{\text{excess approximation error}} + \underbrace{\delta\sigma\left(1 + \|\hat{\gamma}\|_2\right)}_{\text{post-treatment noise}}, \tag{A.29}$$

with probability at least $1 - 6e^{-\frac{\delta^2}{2}} - e^{-2(\log 2 + N_0 \log 5)\delta^2}$.

Proof of Theorem A.4. First note that from Lemma 3 we have that

$$\|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}' \hat{\gamma}^{\text{aug}}\|_2 = \left\| \text{diag}\left(\frac{\lambda}{d_j^2+\lambda}\right)(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}' \hat{\gamma}^{\text{scm}}) \right\|_2.$$

The Theorem directly follows from Proposition A.1 and Lemma A.5, combining the two probabilistic bounds via the union bound. ☐

Theorems A.3 and A.4 have several implications when the outcomes follow a linear factor model (6). To see this, we need one more lemma:

Lemma A.6. The linear factor model is a special case of model (A.23) with $\boldsymbol{\beta} = \frac{1}{T_0}\boldsymbol{\mu}\boldsymbol{\mu}_T$ and $\xi_i = \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}\boldsymbol{\varepsilon}_{i(1:T_0)}$. $\|\boldsymbol{\beta}\|_2 \leq \frac{MJ^2}{\sqrt{T_0}}$, and if $\boldsymbol{\varepsilon}_{i(1:T_0)}$ are independent sub-Gaussian vectors with scale parameter σ_{T_0}, then $\frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{i(1:T_0)}$ is sub-Gaussian with scale parameter $\frac{JM^2\sigma_{T_0}}{\sqrt{T_0}}$.

Proof of Lemma A.6. Notice that under the linear factor model, the pre-treatment covariates for unit i satisfy:

$$\boldsymbol{X}_i = \boldsymbol{\mu}\boldsymbol{\phi}_i + \boldsymbol{\varepsilon}_{i(1:T_0)}.$$

Multiplying both sides by $(\boldsymbol{\mu}'\boldsymbol{\mu})^{-1}\boldsymbol{\mu}' = \frac{1}{T_0}\boldsymbol{\mu}'$ and rearranging gives

$$\frac{1}{T_0}\boldsymbol{\mu}'\boldsymbol{X}_i - \frac{1}{T_0}\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{i(1:T_0)} = \boldsymbol{\phi}_i.$$

Then we see that the post treatment outcomes are

$$Y_{iT}(0) = \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{X}_i + \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{i(1:T_0)}.$$

Since $\boldsymbol{\varepsilon}_{i(1:T_0)}$ is a sub-Gaussian vector $v'\boldsymbol{\varepsilon}_{i(1:T_0)}$ is sub-Gaussian with scale parameter σ_{T_0} for any $v \in \mathbb{R}^{T_0}$ such that $\|v\|_2 = 1$. Now notice that $\|\boldsymbol{\mu}_T'\boldsymbol{\mu}'\|_2 \leq \|\boldsymbol{\mu}_T\|_2\|\boldsymbol{\mu}\|_2 \leq MJ^2\sqrt{T_0}$. This completes the proof. ☐

Proof of Corollary A.2. From Lemma A.6 we can apply Theorem A.3 with $\boldsymbol{\beta} = \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'$ and $\xi_i = \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{i(1:T_0)}$. Since ε_{it} are independent sub-Gaussian random variables, $\boldsymbol{\varepsilon}_{i(1:T_0)}$ is a sub-Gaussian vector with scale parameter $\sigma_{T_0} = \sigma$. Noting that $\|\hat{\boldsymbol{\gamma}}\|_1 = \sum_{W_i=0} |\hat{\gamma}_i| = 1$ and applying Lemma A.6 gives the result. $\square$

Proof of Theorem 1. Again from Lemma A.6 we can apply Theorem A.4 with $\boldsymbol{\beta} = \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'$ and $\xi_i = \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{i(1:T_0)}$, so $\varpi = \frac{JM^2\sigma}{\sqrt{T_0}}$. Plugging these values into Theorem A.3 gives the result. $\square$

Corollary A.5 (Approximation error for ridge ASCM with dependent errors)**.** Under the linear factor model (6) with time-dependent errors satisfying $\boldsymbol{\varepsilon}_{i(1:T_0)} \overset{iid}{\sim} N(0, \sigma^2\Omega)$ the approximation error satisfies

$$\left|\xi_1 - \sum_{W_i=0} \hat{\gamma}_i \xi_i\right| = \left|\frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'\left(\boldsymbol{\varepsilon}_{1(1:T_0)} - \sum_{W_i=0}\hat{\gamma}_i\boldsymbol{\varepsilon}_{i(1:T_0)}\right)\right|$$
$$\leq 2\sqrt{\frac{\|\Omega\|_2}{T_0}}JM^2\sigma\left(\sqrt{\log 2N_0} + \delta + (1+\delta)2\left\|\text{diag}\left(\frac{d_j}{d_j^2+\lambda}\right)\left(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}}\right)\right\|_2\right), \tag{A.30}$$

Proof of Corollary A.5. From Lemma A.6, we see that $\xi_i = \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{i(1:T_0)}$ is sub-Guassian with scale parameter $JM^2\sqrt{\frac{\|\Omega\|_2}{T_0}}$. Plugging in to Lemma A.5 gives the result. $\square$

Lipshitz approximation error If ξ_i are Lipshitz functions, we can can also bound the approximation error.

Assumption A.6. $\xi_i = f(\boldsymbol{X}_i)$ where $f : \mathbb{R}^{T_0} \to \mathbb{R}$ is an L-Lipshitz function.

Lemma A.7. Under Assumption A.6, for weights on the simplex $\hat{\boldsymbol{\gamma}} \in \Delta^{N_0}$, the approximation error satisfies

$$\left|\xi_1 - \sum_{W_i=0}\hat{\gamma}_i\xi_i\right| \leq L\sum_{W_i=0}\hat{\gamma}_i\|\boldsymbol{X}_1 - \boldsymbol{X}_i\|_2 \tag{A.31}$$

Proof of Lemma A.7. Since the weights sum to one, we have that

$$\left|\xi_1 - \sum_{W_i=0}\hat{\gamma}_i\xi_i\right| \leq \left|\sum_{W_i=0}\hat{\gamma}_i(f(\boldsymbol{X}_1) - f(\boldsymbol{X}_i))\right|.$$

Now from the Lipshitz property, $|f(\boldsymbol{X}_1) - f(\boldsymbol{X}_i)| \leq L\|\boldsymbol{X}_1 - \boldsymbol{X}_i\|_2$, and so by Jensen's inequalty:

$$\left|\sum_{W_i=0}\hat{\gamma}_i(f(\boldsymbol{X}_1) - f(\boldsymbol{X}_i))\right| \leq L\sum_{W_i=0}\hat{\gamma}_i\|\boldsymbol{X}_1 - \boldsymbol{X}_i\|_2$$

$\square$

Proof of Theorem A.3. The proof follows directly from Proposition A.1 and Lemma A.7. $\square$

Lemma A.8. Let $C = \max_{W_i=0} \|\boldsymbol{X}_i\|_2$. Under Assumption A.6, the ridge ASCM weights $\hat{\boldsymbol{\gamma}}^{\text{aug}}$ (17) with hyperparameter $\lambda^{\text{ridge}} = N_0 \lambda$ satisfy

$$\left|\xi_1 - \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} \xi_i\right| \leq L \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} \|\boldsymbol{X}_1 - \boldsymbol{X}_i\|_2 + CL \left\|\text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right)\left(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}' \hat{\gamma}^{\text{scm}}\right)\right\|_2 \tag{A.32}$$

Proof of Lemma A.8. From the triangle inequality we have that

$$\left|\xi_1 - \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} \xi_i\right| \leq \left|\sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} (f(\boldsymbol{X}_1) - f(\boldsymbol{X}_i))\right| + \left|\sum_{W_i=0} \boldsymbol{X}_i \left(\boldsymbol{X}_{0\cdot}' \boldsymbol{X}_{0\cdot} + \lambda I\right)^{-1} (\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}' \hat{\gamma}^{\text{scm}}) f(\boldsymbol{X}_i)\right|.$$

We have already bounded the first term in Lemma A.7, now we bound the second term. From the Cauchy-Schwartz inequality and since $\|x\|_2 \leq \sqrt{N_0}\|x\|_\infty$ for all $x \in \mathbb{R}^{N_0}$ we see that

$$\begin{aligned}
\left|\sum_{W_i=0} \boldsymbol{X}_i \left(\boldsymbol{X}_{0\cdot}' \boldsymbol{X}_{0\cdot} + \lambda I\right)^{-1} (\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}' \hat{\gamma}^{\text{scm}}) f(\boldsymbol{X}_i)\right| &\leq \sqrt{N_0} \left\|\boldsymbol{X}_{0\cdot} \left(\boldsymbol{X}_{0\cdot}' \boldsymbol{X}_{0\cdot} + \lambda I\right)^{-1} (\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}' \hat{\gamma}^{\text{scm}})\right\|_2 |f(\boldsymbol{X}_i)| \\
&= \left\|\text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right)\left(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}' \hat{\gamma}^{\text{scm}}\right)\right\|_2 |f(\boldsymbol{X}_i)| \\
&\leq CL \left\|\text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right)\left(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}' \hat{\gamma}^{\text{scm}}\right)\right\|_2,
\end{aligned}$$

where the second line comes from Lemma A.3 and the third line from the Lipshitz property. ☐

Proof of Theorem A.1. The proof follows directly from Proposition A.1 and Lemma A.8. ☐

D.3 Proofs for Sections 6 and B.3

Proof of Lemma 4. The regression parameters $\hat{\boldsymbol{\eta}}_x$ and $\hat{\boldsymbol{\eta}}_z$ in Equation (31) are:

$$\hat{\boldsymbol{\eta}}_x = (\check{\boldsymbol{X}}_{0\cdot}' \check{\boldsymbol{X}}_{0\cdot} + \lambda^{\text{ridge}} I)^{-1} \check{\boldsymbol{X}}_{0\cdot}' Y_{0T} \quad \text{and} \quad \hat{\boldsymbol{\eta}}_z = (\boldsymbol{Z}_{0\cdot}' \boldsymbol{Z}_{0\cdot})^{-1} \boldsymbol{Z}_{0\cdot}' Y_{0T} \tag{A.33}$$

Now notice that

$$\begin{aligned}
\hat{Y}_{0T}^{\text{cov}} &= \hat{\boldsymbol{\eta}}_x' \boldsymbol{X}_1 + \hat{\boldsymbol{\eta}}_z' \boldsymbol{Z}_1 + \sum_{W_i=0} (Y_{iT} - \hat{\boldsymbol{\eta}}_x' \boldsymbol{X}_i - \hat{\boldsymbol{\eta}}_z \boldsymbol{Z}_i) \hat{\gamma}_i \\
&= \hat{\boldsymbol{\eta}}_x' (\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}' \hat{\gamma}) + \hat{\boldsymbol{\eta}}_z (\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}' \hat{\gamma}) + \boldsymbol{Y}_{0T}' \hat{\gamma} \\
&= \hat{\boldsymbol{\eta}}_x' (\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}' \hat{\gamma}) - \hat{\boldsymbol{\eta}}_x' \boldsymbol{X}_{0\cdot} (\boldsymbol{Z}_{0\cdot}' \boldsymbol{Z}_{0\cdot})^{-1} (\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}' \hat{\gamma}) + Y_{0T}' \boldsymbol{Z}_{0\cdot} (\boldsymbol{Z}_{0\cdot}' \boldsymbol{Z}_{0\cdot})^{-1} (\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}' \hat{\gamma}) + Y_{0T}' \hat{\gamma} \\
&= \hat{\boldsymbol{\eta}}_x' (\check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot}' \hat{\gamma}) + Y_{0T}' \boldsymbol{Z}_{0\cdot} (\boldsymbol{Z}_{0\cdot}' \boldsymbol{Z}_{0\cdot})^{-1} (\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}' \hat{\gamma}) + Y_{0T}' \hat{\gamma} \\
&= Y_{0T}' \left(\hat{\gamma} + \check{\boldsymbol{X}}_{0\cdot} (\check{\boldsymbol{X}}_{0\cdot}' \check{\boldsymbol{X}}_{0\cdot} + \lambda^{\text{ridge}} \boldsymbol{I}_{T_0})^{-1} (\check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot}' \hat{\gamma}) + \boldsymbol{Z}_{0\cdot} (\boldsymbol{Z}_{0\cdot}' \boldsymbol{Z}_{0\cdot})^{-1} (\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}' \hat{\gamma})\right).
\end{aligned} \tag{A.34}$$

This gives the form of $\hat{\boldsymbol{\gamma}}^{\text{cov}}$. The imbalance in Z is

$$\begin{aligned}\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{cov}} &= \left(\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot}(\boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot})^{-1}\boldsymbol{Z}_1\right) + \left(\boldsymbol{Z}_{0\cdot} - \boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot}(\boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot})^{-1}\boldsymbol{Z}_{0\cdot}\right)'\hat{\boldsymbol{\gamma}} \\ &\quad - \boldsymbol{Z}_{0\cdot}'\check{\boldsymbol{X}}_{0\cdot}(\check{\boldsymbol{X}}_{0\cdot}'\check{\boldsymbol{X}}_{0\cdot} + \lambda^{\text{ridge}} I)^{-1}(\check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot}'\hat{\boldsymbol{\gamma}}) \\ &= 0.\end{aligned} \tag{A.35}$$

The pre-treatment fit is

$$\begin{aligned}\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{cov}} &= \left(\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\boldsymbol{Z}_{0\cdot}(\boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot})^{-1}\boldsymbol{Z}_1\right) + \left(\boldsymbol{X}_{0\cdot} - \boldsymbol{X}_{0\cdot}'\boldsymbol{Z}_{0\cdot}(\boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot})^{-1}\boldsymbol{Z}_{0\cdot}\right)'\hat{\boldsymbol{\gamma}} \\ &\quad - \boldsymbol{X}_{0\cdot}'\check{\boldsymbol{X}}_{0\cdot}(\check{\boldsymbol{X}}_{0\cdot}'\check{\boldsymbol{X}}_{0\cdot} + \lambda^{\text{ridge}}\boldsymbol{I}_{T_0})^{-1}(\check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot}'\hat{\boldsymbol{\gamma}}) \\ &= \left(\boldsymbol{I}_{T_0} - \boldsymbol{X}_{0\cdot}'\check{\boldsymbol{X}}_{0\cdot}(\check{\boldsymbol{X}}_{0\cdot}'\check{\boldsymbol{X}}_{0\cdot} + \lambda^{\text{ridge}}\boldsymbol{I}_{T_0})^{-1}\right)\left(\check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot}'\hat{\boldsymbol{\gamma}}\right) \\ &= \left(\boldsymbol{I}_{T_0} - \check{\boldsymbol{X}}_{0\cdot}'\check{\boldsymbol{X}}_{0\cdot}(\check{\boldsymbol{X}}_{0\cdot}'\check{\boldsymbol{X}}_{0\cdot} + \lambda^{\text{ridge}}\boldsymbol{I}_{T_0})^{-1}\right)\left(\check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot}'\hat{\boldsymbol{\gamma}}\right).\end{aligned} \tag{A.36}$$

This gives the bound on the pre-treatment fit.

□

Proof of Theorem A.2. First, we will separate $f(\boldsymbol{Z})$ into the projection onto $\boldsymbol{Z}$ and a residual. Defining $\boldsymbol{B}_t = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'f_t(\boldsymbol{Z}) \in \mathbb{R}^K$ as the regression coefficient, the projection of $f_t(\boldsymbol{Z}_i)$ is $\boldsymbol{Z}_i'\boldsymbol{B}_t$ and the residual is $e_{it} = f_t(\boldsymbol{Z}_i) - \boldsymbol{Z}_i'\boldsymbol{B}_t$. We will denote the matrix of regression coefficients over $t = 1, \ldots, T_0$ as $\boldsymbol{B} = [\boldsymbol{B}_1, \ldots, \boldsymbol{B}_{T_0}] \in \mathbb{R}^{K \times T_0}$ and denote the matrix of residuals as $\boldsymbol{E} \in \mathbb{R}^{n \times T_0}$, with $\boldsymbol{E}_{1\cdot} = (e_{11}, \ldots, e_{1T_0})$ as the vector of residuals for the treated unit and $\boldsymbol{E}_{0\cdot}$ as the matrix of residuals for the control units.

Then the error is

$$\begin{aligned}\left|Y_{1T}(0) - \sum_{W_i=0}\hat{\gamma}_i^{\text{cov}}Y_{iT}\right| &\leq \left|\boldsymbol{\mu}_T \cdot \left(\boldsymbol{\phi}_1 - \sum_{W_i=0}\hat{\gamma}_i^{\text{cov}}\boldsymbol{\phi}_i\right)\right| + \left|\boldsymbol{B}_t \cdot \left(\boldsymbol{Z}_1 - \sum_{W_i=0}\hat{\gamma}_i^{\text{cov}}\boldsymbol{Z}_i\right)\right| \\ &\quad + \left|e_{1T} - \sum_{W_i=0}\hat{\gamma}^{\text{cov}}e_{iT}\right| + \left|\varepsilon_{1T} - \sum_{W_i=0}\hat{\gamma}_i^{\text{cov}}\varepsilon_{iT}\right|\end{aligned}$$

Since $\hat{\gamma}_i^{\text{cov}}$ exactly balances the covariates, the second term is equal to zero. We can bound the third term with Hölder's inequality:

$$\left|e_{1T} - \sum_{W_i=0}\hat{\gamma}^{\text{cov}}e_{iT}\right| \leq |e_{1T}| + \sqrt{RSS}_T\|\hat{\gamma}^{\text{cov}}\|_2$$

In previous theorems we have bounded the last term with high probability. Only the error due to imbalance remains.

Denote $\boldsymbol{\varepsilon}_{0(1:T_0)}$ as the matrix of pre-treatment noise for the control units, where the rows correspond to $\boldsymbol{\varepsilon}_{2(1:T_0)}, \ldots, \boldsymbol{\varepsilon}_{N_0(1:T_0)}$. Building on Lemma A.6, we can see that the error due to imbalance in ϕ is equal to

$$\boldsymbol{\mu}_T \cdot \left(\boldsymbol{\phi}_1 - \sum_{W_i=0} \hat{\gamma}_i^{\text{cov}} \boldsymbol{\phi}_i\right) = \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'(\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{cov}}) - \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'(\boldsymbol{\varepsilon}_{1(1:T_0)} - \boldsymbol{\varepsilon}_{0(1:T_0)}'\hat{\boldsymbol{\gamma}}^{\text{cov}}) \\ - \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'B'(\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{cov}}) - \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'(\boldsymbol{E}_{1\cdot} - \boldsymbol{E}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{cov}}). \tag{A.37}$$

By construction, $\hat{\boldsymbol{\gamma}}^{\text{cov}}$ perfectly balances the covariates, and combined with Lemma 4, the error due to imbalance in ϕ simplifies to

$$\boldsymbol{\mu}_T \cdot \left(\boldsymbol{\phi}_1 - \sum_{W_i=0} \gamma_i \boldsymbol{\phi}_i\right) = \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'(\check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot}'\hat{\gamma}) - \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'(\boldsymbol{\varepsilon}_{1(1:T_0)} - \boldsymbol{\varepsilon}_{0(1:T_0)}'\hat{\boldsymbol{\gamma}}^{\text{cov}}) - \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'(\boldsymbol{E}_{1\cdot} - \boldsymbol{E}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{cov}}).$$

We now turn to bounding the noise term and the error due to the projection of $f(Z)$ on to Z. First, notice that

$$\frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{0(1:T_0)}'\hat{\boldsymbol{\gamma}}^{\text{cov}} = \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{0(1:T_0)}'\hat{\boldsymbol{\gamma}}^{\text{scm}} + \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{0(1:T_0)}'\boldsymbol{Z}_{0\cdot}(\boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot})^{-1}(\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}}).$$

We have bounded the first term on the right hand side in Lemma A.4. To bound the second term, notice that $\sum_{W_i=0}\sum_{t=1}^{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}_{t\cdot}Z_{ik}\varepsilon_{it}$ is sub-Gaussian with scale parameter $\sigma MJ^2\sqrt{T_0}\|Z_{\cdot k}\|_2 = MJ^2\sigma\sqrt{T_0N_0}$. We can now bound the L^2 norm of $\frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{0(1:T_0)}'\boldsymbol{Z}_{0\cdot} \in \mathbb{R}^K$:

$$P\left(\frac{1}{T_0}\|\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{0(1:T_0)}'\boldsymbol{Z}_{0\cdot}\|_2 \geq 2JM^2\sigma\left(\sqrt{\frac{N_0K\log 5}{T_0}} + \delta\right)\right) \leq 2\exp\left(-\frac{T_0\delta^2}{2}\right)$$

Replacing δ with $\sqrt{\frac{KN_0}{T_0}}(2 - \sqrt{\log 5})$ and with the Cauchy-Schwarz inequality we see that

$$\frac{1}{T_0}\left|\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{0(1:T_0)}'\boldsymbol{Z}_{0\cdot}(\boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot})^{-1}(\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}'\hat{\boldsymbol{\gamma}})\right| \leq 4JM^2\sigma\sqrt{\frac{K}{T_0N_0}}\|\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}}\|_2$$

with probability at least $1 - 2\exp\left(-\frac{KN_0(2-\sqrt{\log 5})^2}{2}\right)$.

Next we turn to the residual term. By Hölder's inequality and using that for a matrix $\boldsymbol{A}$, the operator norm is bounded by $\|\boldsymbol{A}\|_2 \leq \sqrt{\text{trace}(\boldsymbol{A}'\boldsymbol{A})}$ we see that

$$\begin{aligned}\left|\frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'(\boldsymbol{E}_{1\cdot} - \boldsymbol{E}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{cov}})\right| &\leq \frac{JM^2}{\sqrt{T_0}}\left(\|\boldsymbol{E}_{1\cdot}\|_2 + \|\hat{\boldsymbol{\gamma}}^{\text{cov}}\|_2\|\boldsymbol{E}_{0\cdot}\|_2\right) \\ &\leq JM^2\left(\max_{t=1,\ldots,T_0}|e_{1t}| + \|\hat{\boldsymbol{\gamma}}^{\text{cov}}\|_2\sqrt{\frac{1}{T_0}\sum_{t=1}^{T_0}RSS_t}\right) \\ &\leq JM^2\left(\max_{t=1,\ldots,T_0}|e_{1t}| + \|\hat{\boldsymbol{\gamma}}^{\text{cov}}\|_2\sqrt{\max_t RSS_t}\right),\end{aligned}$$

where we have used that $\frac{1}{\sqrt{T_0}}\|\boldsymbol{E}_{1\cdot}\|_2 \leq \max_{t=1,\ldots,T_0}|e_{1t}|$ and $\text{trace}(\boldsymbol{E}_{0\cdot}'\boldsymbol{E}_{0\cdot}) = \sum_{t=1}^{T_0}RSS_t$.

Combining with Lemma 4 and putting together the pieces with the union bound gives the result. $\square$

E Connection to balancing weights and IPW

We have motivated Augmented SCM via bias correction. An alternative motivation comes from the connection between SCM and inverse propensity score weighting (IPW). This is also comparable in form to the generalized regression estimator in survey sampling (Cassel et al., 1976; Breidt and Opsomer, 2017), which has been adapted to the causal inference setting by, among others, Athey et al. (2018) and Hirshberg and Wager (2018).

First, notice that the SCM weights from the constrained optimization problem in Equation (8) are a form of *approximate balancing weights*; see, for example, Zubizarreta (2015); Athey et al. (2018); Tan (2017); Wang and Zubizarreta (2018); Zhao (2018). Unlike traditional inverse propensity score weights, which indirectly minimize covariate imbalance by estimating a propensity score model, balancing weights seek to *directly* minimize covariate imbalance, in this case L^2 imbalance. Balancing weights have a Lagrangian dual formulation as inverse propensity score weights (see, for example Zhao and Percival, 2017; Zhao, 2018; Chattopadhyay et al., 2020). Extending these results to the SCM setting, the Lagrangian dual of the SCM optimization problem in Equation (8) has the form of a propensity score model. Importantly, as we discuss below, it is not always appropriate to interpret this model as a propensity score.

We first derive the Lagrangian dual for a general class of balancing weights problems, then specialize to the penalized SCM estimator (8).

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \quad & \underbrace{h_\zeta(\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma})}_{\text{balance criterion}} + \sum_{W_i=0} \underbrace{f(\gamma_i)}_{\text{dispersion}} \\ \text{subject to} \quad & \sum_{W_i=0} \gamma_i = 1. \end{aligned} \tag{A.38}$$

This formulation generalizes Equation (8) in two ways: first, we remove the non-negativity constraint and note that this can be included by restricting the domain of the strongly convex dispersion penalty f. Examples include the re-centered L^2 dispersion penalties for ridge regression and ridge ASCM, an entropy penalty (Robbins et al., 2017), and an elastic net penalty (Doudchenko and Imbens, 2017). Second, we generalize from the squared L^2 norm to a general balance criterion h_ζ; another promiment example is an L^∞ constraint (see e.g. Zubizarreta, 2015; Athey et al., 2018).

Proposition A.2. The Lagrangian dual to Equation (A.38) is

$$\min_{\alpha,\boldsymbol{\beta}} \quad \underbrace{\sum_{W_i=0} f^*(\alpha + \boldsymbol{\beta}' X_{i\cdot}) - (\alpha + \boldsymbol{\beta}'\boldsymbol{X}_1) +}_{\text{loss function}} \quad \underbrace{h_\zeta^*(\boldsymbol{\beta})}_{\text{regularization}} \quad , \tag{A.39}$$

where a convex, differentiable function g has convex conjugate $g^*(\boldsymbol{y}) \equiv \sup_{\boldsymbol{x}\in \mathrm{dom}(g)}\{\boldsymbol{y}'\boldsymbol{x} - g(\boldsymbol{x})\}$. The solutions to the primal problem (A.38) are $\hat{\gamma}_i = f^{*\prime}(\hat{\alpha} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}_i)$, where $f^{*\prime}(\cdot)$ is the first derivative of the convex conjugate, $f^*(\cdot)$.

There is a large literature relating balancing weights to propensity score weights. This literature shows that the loss function in Equation (A.39) is an M-estimator for the propensity score and thus will be consistent for the propensity score parameters under large N asymptotics. The dispersion measure $f(\cdot)$ determines the link function of the propensity score model, where the odds of treatment are $\frac{\pi(x)}{1-\pi(x)} = f^{*\prime}(\alpha + \boldsymbol{\beta}' x)$. Note that un-penalized SCM, which can yield multiple solutions, does not have a well-defined link function. We extend the duality to a general set of balance criteria so that Equation (A.39) is a regularized M-estimator of the propensity score parameters where the balance criterion $h_\zeta(\cdot)$ determines the type of regularization through its conjugate $h^*_\zeta(\cdot)$. This formulation recovers the duality between entropy balancing and a logistic link (Zhao and Percival, 2017), Oaxaca-Blinder weights and a log-logistic link (Kline, 2011), and L^∞ balance and L^1 regularization (Wang and Zubizarreta, 2018). This more general formulation also suggests natural extensions of both SCM and ASCM beyond the L^2 setting to other forms, especially L^1 regularization.

Specializing proposition A.2 to a squared L^2 balance criterion $h_\zeta(x) = \frac{1}{2\zeta}\|x\|_2^2$ as in the penalized SCM problems yields that the dual propensity score coefficients $\boldsymbol{\beta}$ are regularized by a ridge penalty. In the case of an entropy dispersion penalty as Robbins et al. (2017) consider, the donor weights $\hat{\boldsymbol{\gamma}}$ have the form of IPW weights with a logistic link function, where the propensity score is $\pi(\boldsymbol{X}_i) = \text{logit}^{-1}(\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i)$, the odds of treatment are $\frac{\pi(\boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)} = \exp(\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i) = \gamma_i$.

We emphasize that while Proposition A.2 shows that the the estimated weights have the IPW form, in SCM settings it may not always be appropriate to interpret the dual problem as a propensity score reflecting stochastic selection into treatment. For example, this interpretation would not be appropriate in some canonical SCM examples, such as the analysis of German reunification in Abadie et al. (2015).

Proof of Proposition A.2. We can augment the optimization problem (A.38) with auxiliary variables ϵ, yielding:

$$\begin{aligned} &\min_{\boldsymbol{\gamma},\boldsymbol{\epsilon}} \; h_\zeta(\boldsymbol{\epsilon}) + \sum_{W_i=0} f(\gamma_i). \\ &\text{subject to } \boldsymbol{\epsilon} = \boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma} \\ &\sum_{W_i=0} \gamma_i = 1 \end{aligned} \tag{A.40}$$

The Lagrangian is

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\epsilon}, \alpha, \boldsymbol{\beta}) = \sum_{i|W_i=0} f(\gamma_i) + \alpha(1-\gamma_i) + h_\zeta(\boldsymbol{\epsilon}) + \boldsymbol{\beta}'(\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma} - \boldsymbol{\epsilon}). \tag{A.41}$$

The dual maximizes the objective

$$\begin{aligned} q(\alpha, \boldsymbol{\beta}) &= \min_{\boldsymbol{\gamma},\boldsymbol{\epsilon}} \mathcal{L}(\boldsymbol{\gamma}, \epsilon, \alpha, \boldsymbol{\beta}) \\ &= \sum_{W_i=0} \min_{\gamma_i}\{f(\gamma_i) - (\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i)\gamma_i\} + \min_{\boldsymbol{\epsilon}}\{h_\zeta(\boldsymbol{\epsilon}) - \boldsymbol{\beta}'\boldsymbol{\epsilon}\} + \alpha + \boldsymbol{\beta}'\boldsymbol{X}_1 \\ &= -\sum_{W_i=0} f^*(\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i) + \alpha + \boldsymbol{\beta}'\boldsymbol{X}_1' - h^*_\zeta(\boldsymbol{\beta}), \end{aligned} \tag{A.42}$$

By strong duality the general dual problem (A.39), which minimizes $-q(\alpha, \boldsymbol{\beta})$, is equivalent to the primal balancing weights problem. Given the $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ that minimize the Lagrangian dual objective,

$-q(\alpha, \boldsymbol{\beta})$, we recover the donor weights solution to (A.38) as

$$\hat{\gamma}_i = f^{*\prime}(\hat{\alpha} + \hat{\boldsymbol{\beta}}' \boldsymbol{X}_i). \tag{A.43}$$

$\square$

F Additional figures

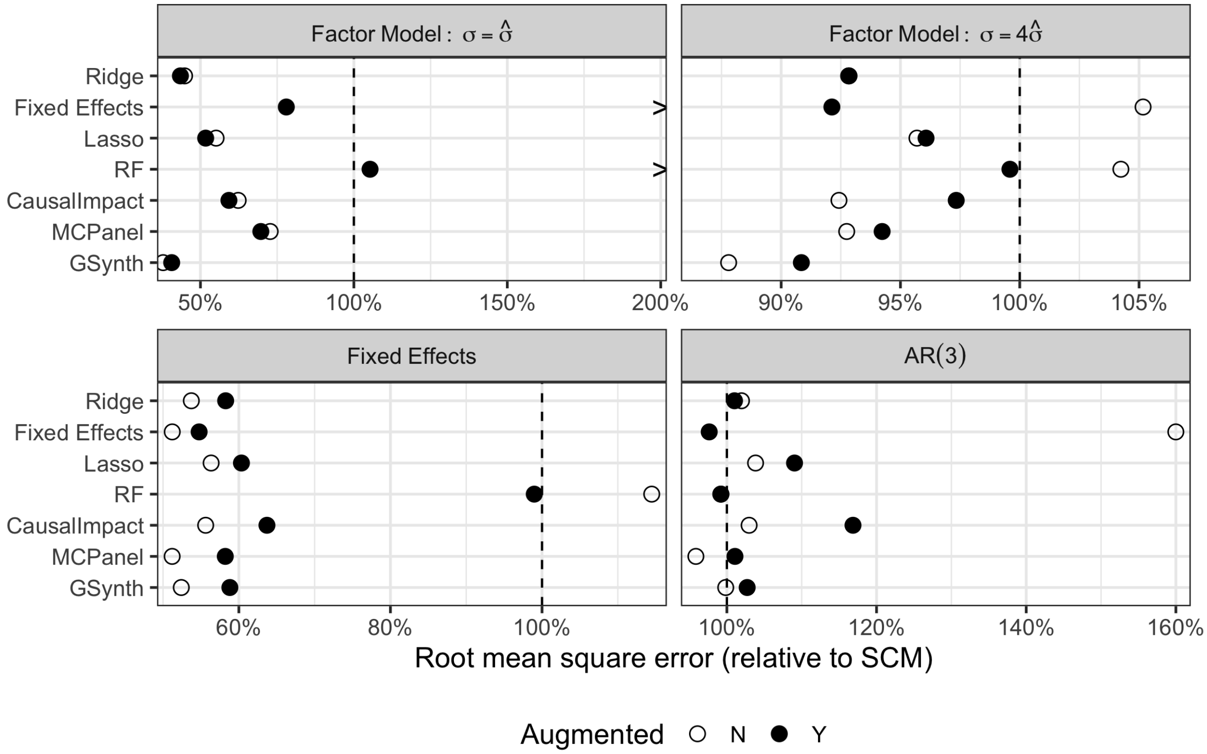


Figure F.1: RMSE for different augmented and non-augmented estimators across outcome models.

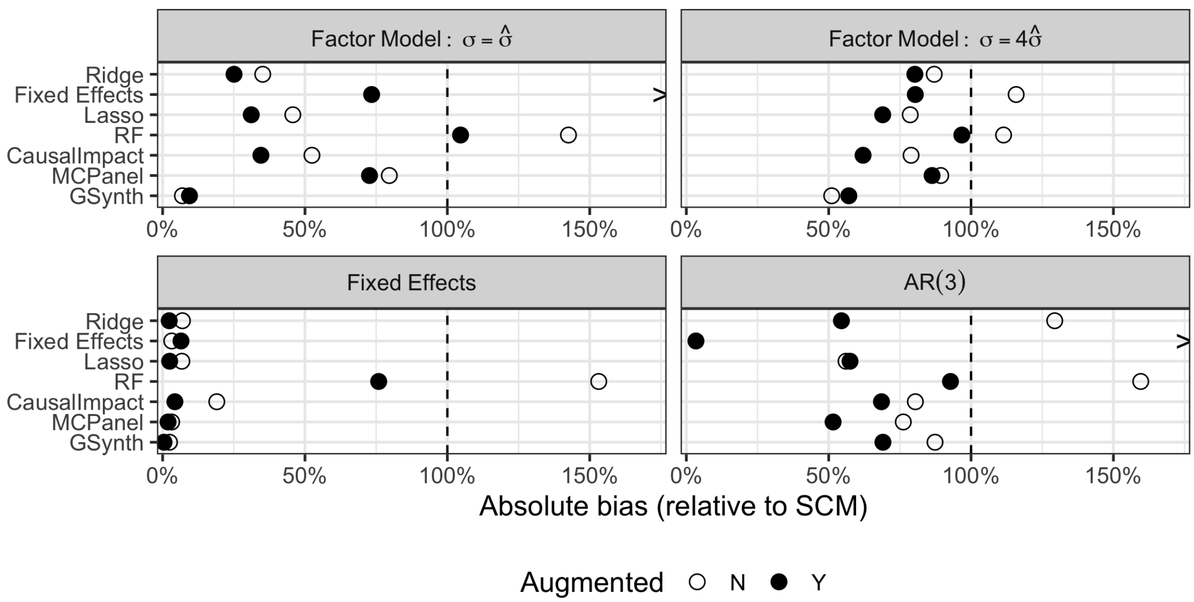


Figure F.2: Absolute bias (as a percentage of SCM bias) for ridge, fixed effects, and several machine learning and panel data outcome models, and their augmented versions using the same data generating processes as Figure 3.

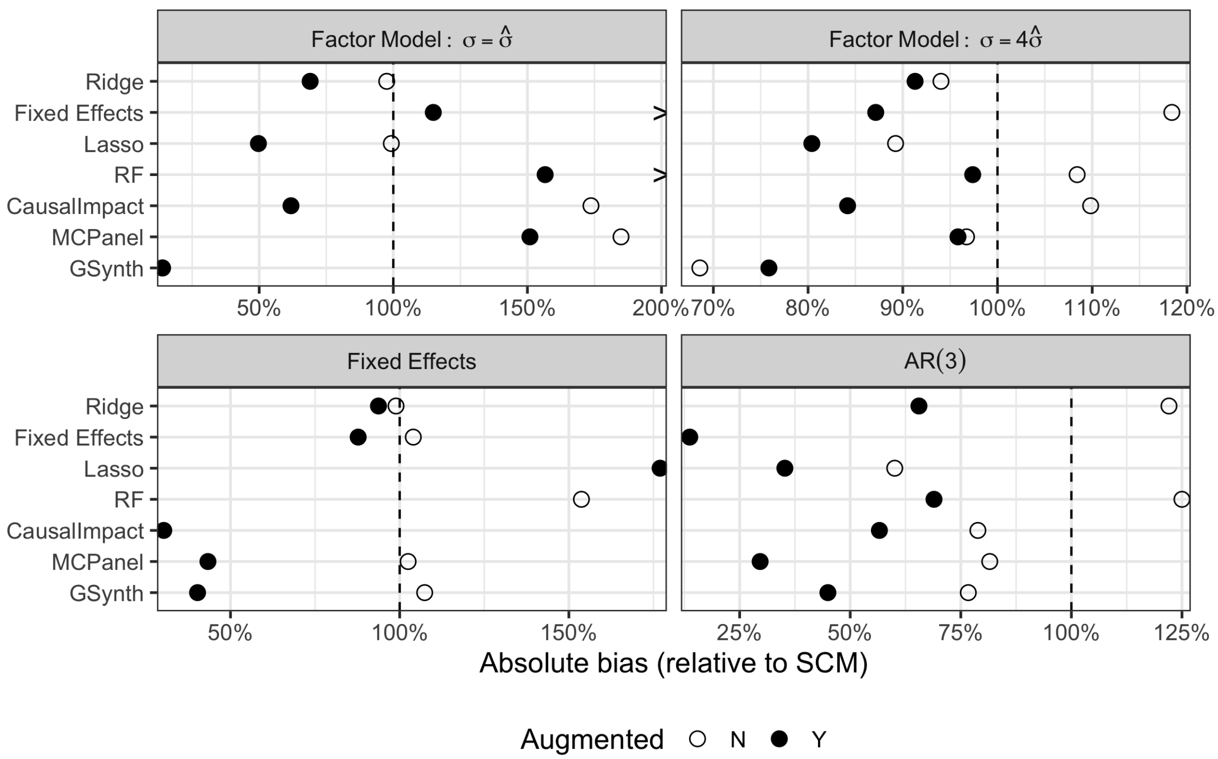


Figure F.3: Bias for different augmented and non-augmented estimators across outcome models conditioned on SCM fit in the top quintile.

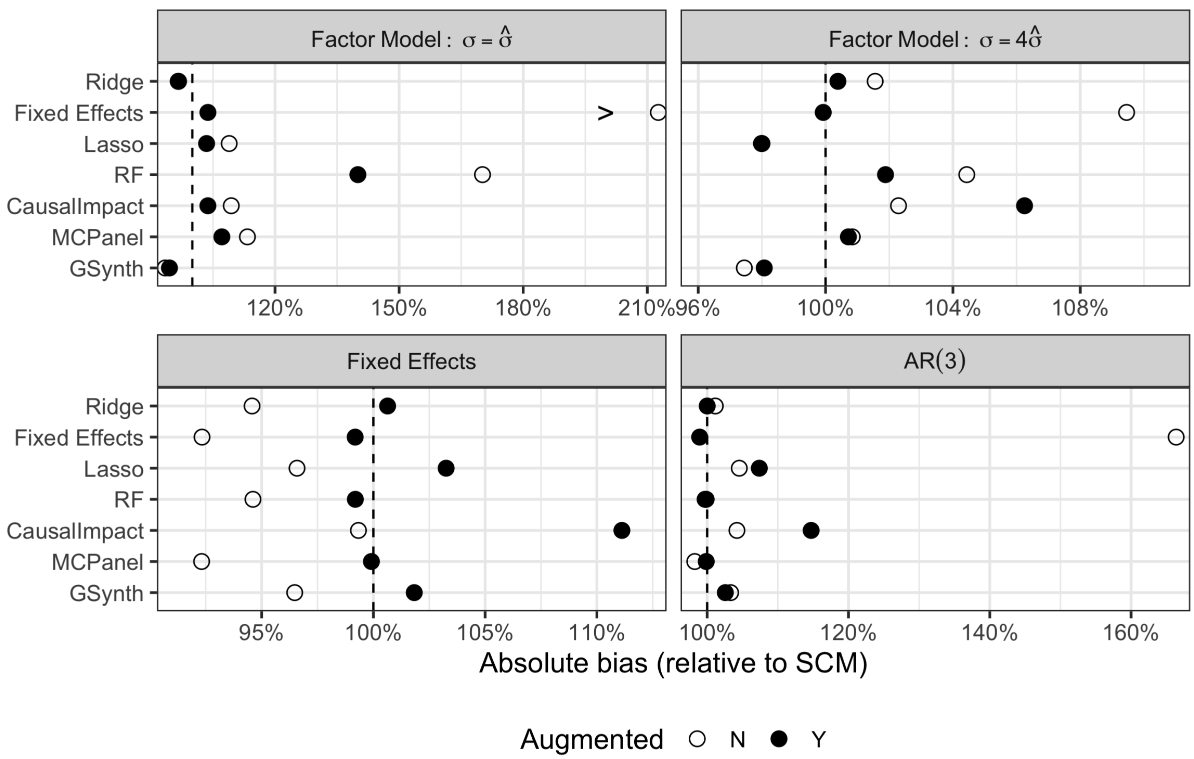


Figure F.4: RMSE for different augmented and non-augmented estimators across outcome models conditioned on SCM fit in the top quntile.

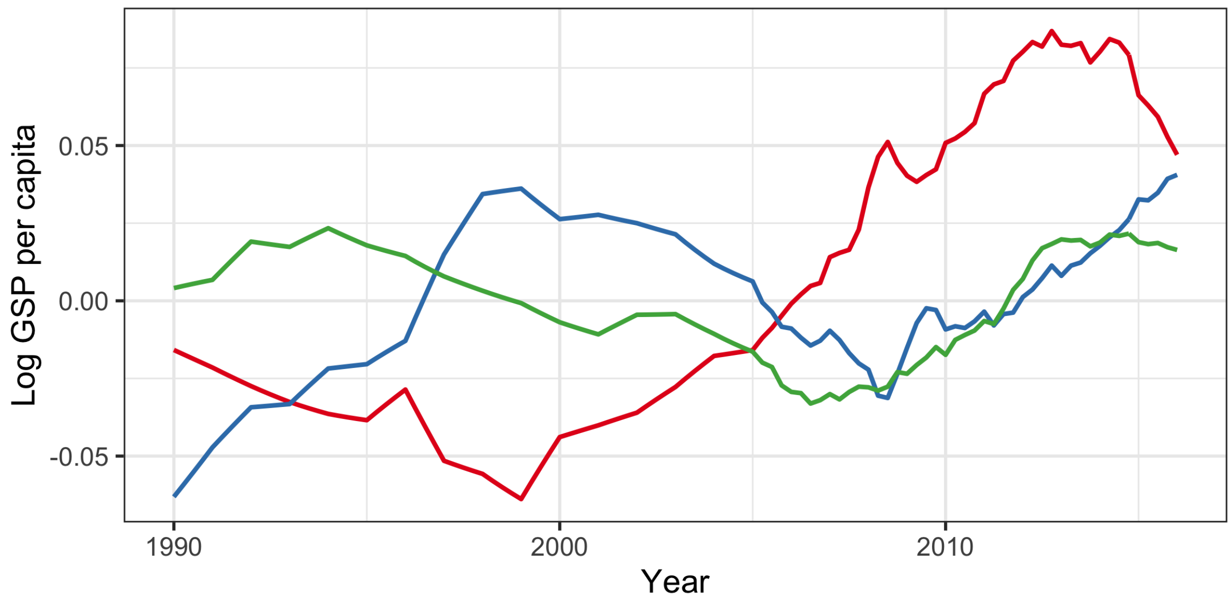


Figure F.5: Latent factors for calibrated simulation studies.

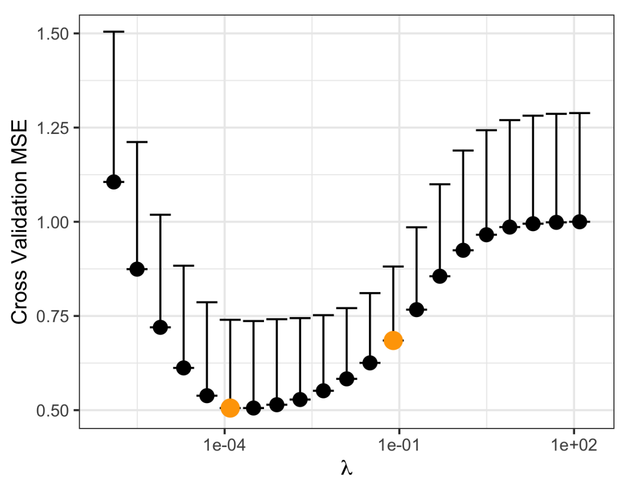


Figure F.6: Cross validation MSE and one standard error computed according to Equation (27). The minimal point, and the maximum λ within one standard error of the minimum are highlighted.

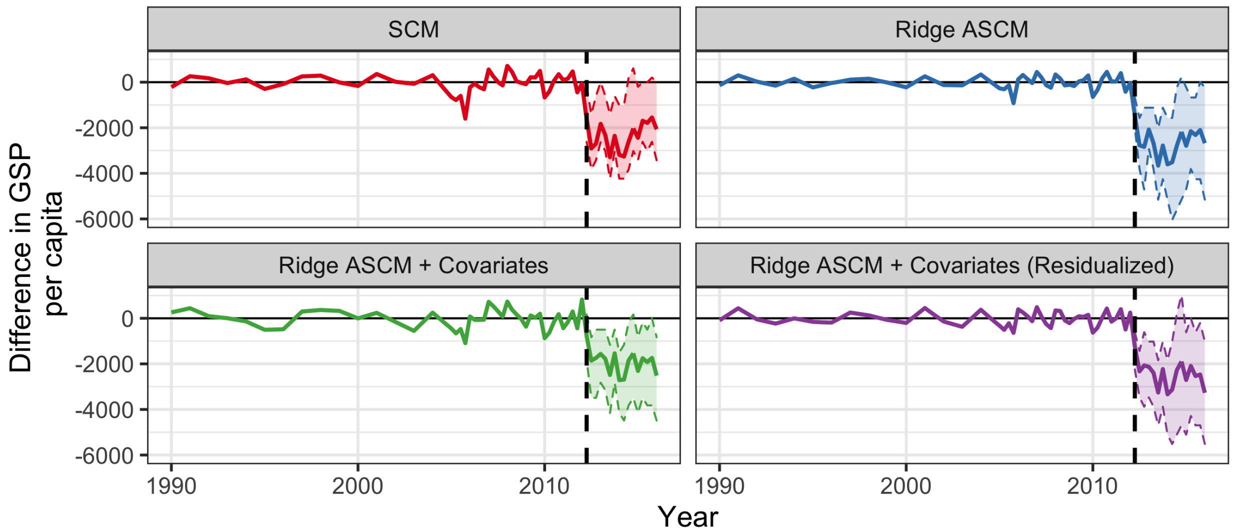


Figure F.7: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on GSP per capita using SCM, ridge ASCM, and ridge ASCM with covariates.

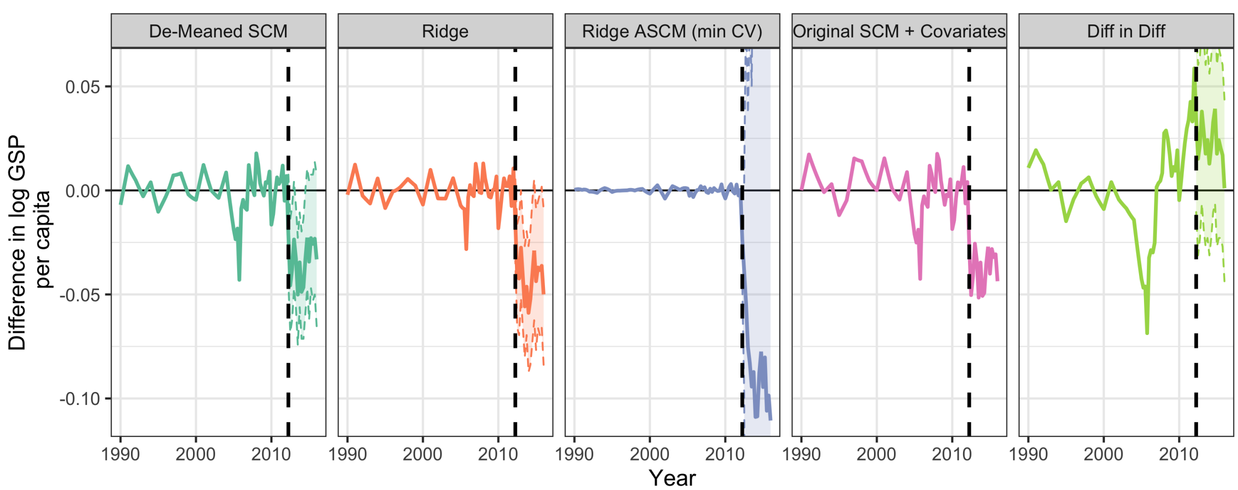


Figure F.8: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on log GSP per capita using de-meaned SCM, ridge regression alone, ridge ASCM with λ chosen to minimize the cross validated MSE, the original SCM proposal with covariates (Abadie et al., 2010), and a two-way fixed effects differences in differences estimate.

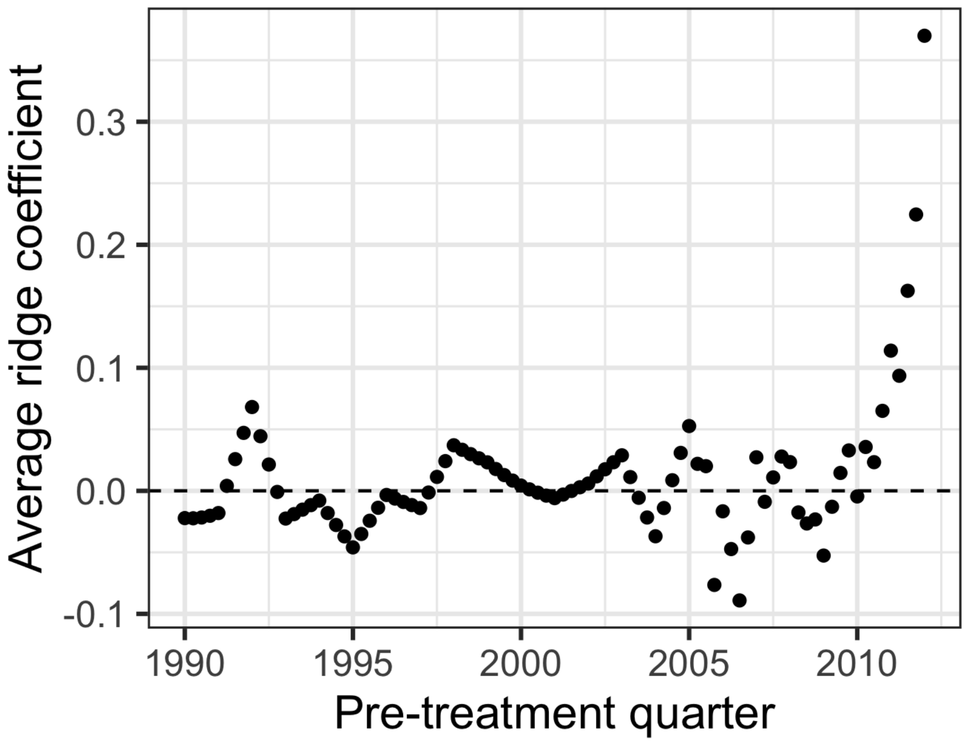


Figure F.9: Ridge regression coefficients for each pre-treatment quarter, averaged across post-treatment quarters.

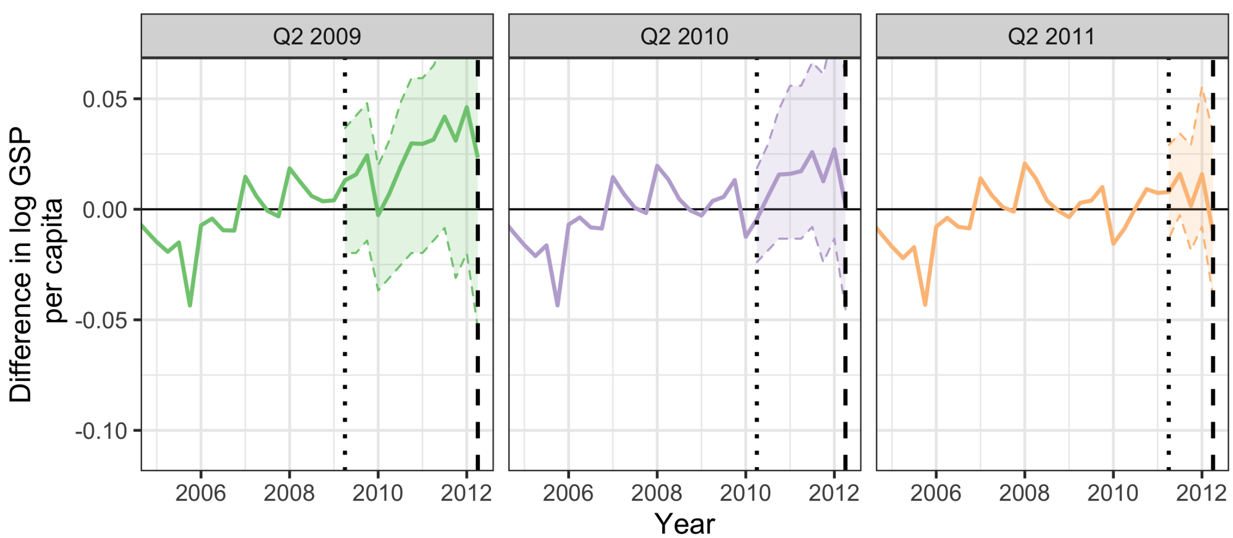


Figure F.10: Placebo point estimates along with 95% conformal confidence intervals for SCM with placebo treatment times in Q2 2009, 2010, and 2011. Scale begins in 2005 to highlight placebo estimates.

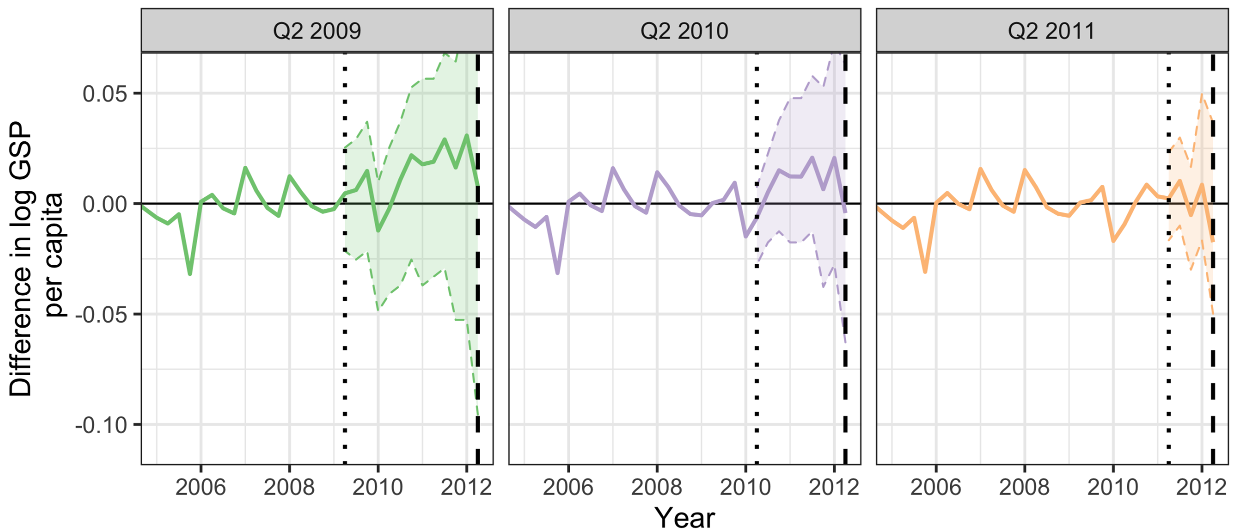


Figure F.11: Placebo point estimates along with 95% conformal confidence intervals for ridge ASCM with placebo treatment times in Q2 2009, 2010, and 2011. Scale begins in 2005 to highlight placebo estimates.

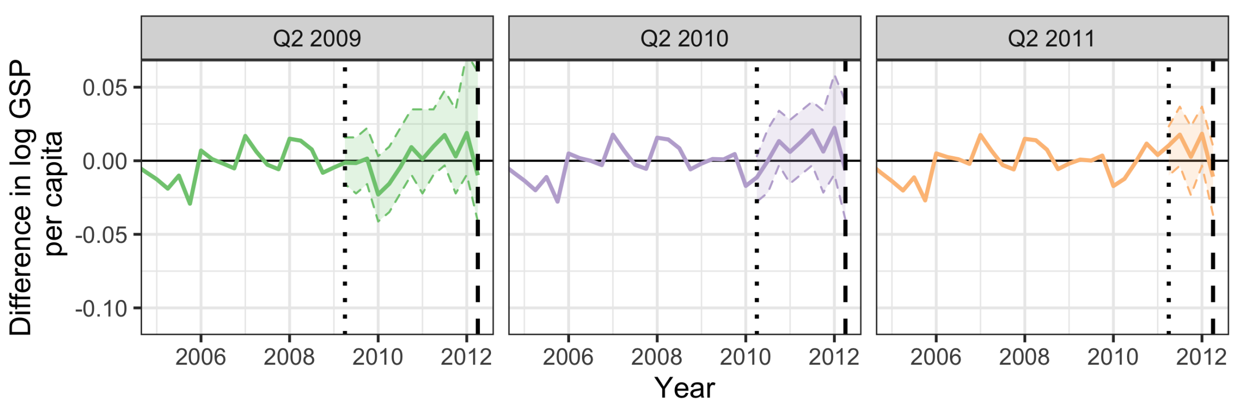


Figure F.12: Placebo point estimates along with 95% conformal confidence intervals for Ridge ASCM with covariates with placebo treatment times in Q2 2009, 2010, and 2011. The time period begins in 2005 and ends in Q1 2012 to highlight placebo estimates.

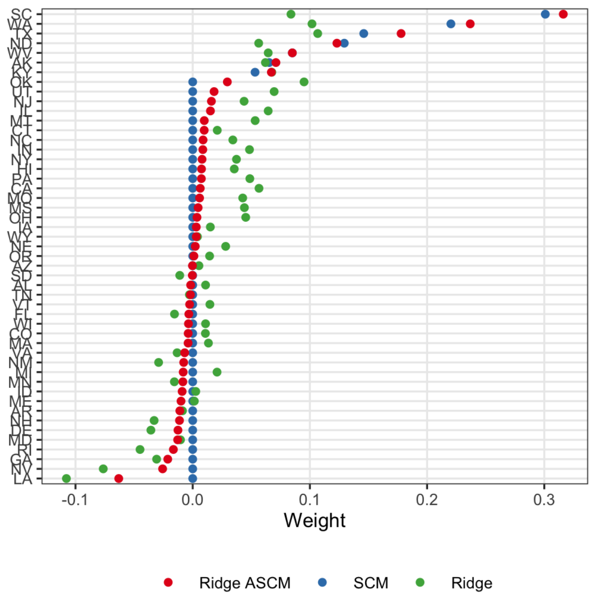


Figure F.13: Donor unit weights for SCM, ridge regression, and ridge ASCM balancing lagged outcomes.

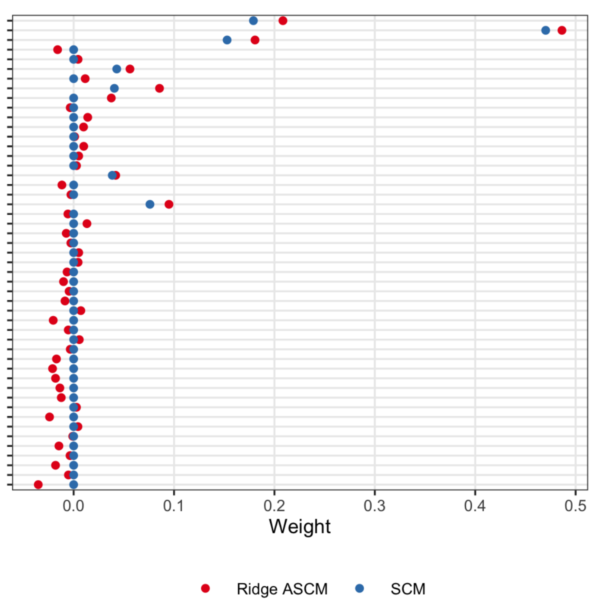


Figure F.14: Donor unit weights for SCM and ridge ASCM fit on lagged outcomes after residualizing out auxiliary covariates.

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association 105*(490), 493–505.

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

Abadie, A. and J. L'Hour (2018). A penalized synthetic control estimator for disaggregated data.

Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(4), 597–623.

Botosaru, I. and B. Ferman (2019). On the role of covariates in the synthetic control method. *The Econometrics Journal 22*(2), 117–130.

Breidt, F. J. and J. D. Opsomer (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques. *Statistical Science 32*(2), 190–205.

Cassel, C. M., C.-E. Sarndal, and J. H. Wretman (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika 63*(3), 615–620.

Chattopadhyay, A., Christopher H. Hase, and J. R. Zubizarreta (2020). Balancing Versus Modeling Approaches to Weighting in Practice. *Statistics in Medicine in press*.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2018). Inference on average treatment effects in aggregate panel data settings. *arXiv preprint arXiv:1812.10820*.

Chernozhukov, V., K. Wüthrich, and Y. Zhu (2019). An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls. Technical report.

Doudchenko, N. and G. W. Imbens (2017). Difference-In-Differences and Synthetic Control Methods: A Synthesis. *arxiv 1610.07748*.

Hirshberg, D. A. and S. Wager (2018). Augmented Minimax Linear Estimation.

Kline, P. (2011). Oaxaca-Blinder as a reweighting estimator. In *American Economic Review*, Volume 101, pp. 532–537.

Robbins, M., J. Saunders, and B. Kilmer (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*(517), 109–126.

Tan, Z. (2017). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data.

Wainwright, M. (2018). *High dimensional statistics: a non-asymptomatic viewpoint.*

Wang, Y. and J. R. Zubizarreta (2018). Minimal Approximately Balancing Weights: Asymptotic Properties and Practical Considerations.

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

Zhao, Q. (2018). Covariate Balancing Propensity Score by Tailored Loss Functions. *Annals of Statistics*, forthcoming.

Zhao, Q. and D. Percival (2017). Entropy balancing is doubly robust. *Journal of Causal Inference 5*(1).

Zubizarreta, J. R. (2015). Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data. *Journal of the American Statistical Association 110*(511), 910–922.