The Augmented Synthetic Control Method*

Eli Ben-Michael, Avi Feller, and Jesse Rothstein

UC Berkeley

November 2018

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 as closely as possible. The curse of dimensionality, however, means that SCM does not generally achieve exact balance, which can bias the SCM estimate. We propose an extension, Augmented SCM, which uses an outcome model to estimate the bias due to covariate imbalance and then de-biases the original SCM estimate, analogous to bias correction for inexact matching. We motivate this approach by showing that SCM is a (regularized) inverse propensity score weighting estimator, with pre-treatment outcomes as covariates and a ridge penalty on the propensity score coefficients. We give theoretical guarantees for specific cases and propose a new inference procedure. We demonstrate gains from Augmented SCM with extensive simulation studies and apply this framework to canonical SCM examples. We implement the proposed method in the new `augsynth` `R` package.

*email: `afeller@berkeley.edu`. We thank Josh Angrist, Alex D'Amour, Peng Ding, Erin Hartman, Chad Hazlett, Steve Howard, Guido Imbens, Pat Kline, Caleb Miles, Luke Miratrix, Sam Pimentel, Fredrik Sävje, Jas Sekhon, Jake Soloff, Panos Toulis, Stefan Wager, 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.

1 Introduction

The *synthetic control method* (SCM) is a popular approach for estimating the impact of a treatment on a single unit in settings with a modest number of control units and with many pre-treatment outcomes for all units (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).

An important limitation of this approach is that, while SCM minimizes imbalance in pre-treatment outcomes, it generally fails to achieve exact balance due to the curse of dimensionality (Ferman and Pinto, 2018). The resulting imbalance can lead to bias.

To address this, we propose the *augmented synthetic control method* (ASCM). Analogous to bias correction for inexact matching (Abadie and Imbens, 2011), ASCM uses an outcome model to estimate the bias due to covariate imbalance and then de-biases the original SCM estimate. If the estimated bias is small, then the SCM and ASCM estimates will be similar.

We relate our proposal to similar estimators by demonstrating that SCM is a regularized inverse propensity score weighting (IPW) estimator, using pre-treatment outcomes as covariates and penalizing the propensity score coefficients via a ridge penalty. This equivalence, which has not previously been noted, relies on a dual representation of the SCM constrained optimization problem. From this perspective, augmented SCM is analogous to augmented IPW in cross-sectional settings (Robins et al., 1994). The IPW perspective also allows us to draw upon the large propensity score literature to better understand SCM and especially to clarify ambiguity about inference and testing.

We make several additional contributions. First, we characterize the finite sample bias for SCM and highlight the role of covariate imbalance. In particular, Abadie et al. (2010) show that SCM is (asymptotically) unbiased under an assumption that SCM weights yield exact balance on the lagged outcomes. This is a strong assumption, however; such weights exist if and only if the treated unit's pre-treatment time series is inside the convex hull of the control units' time series. Similar to other matching and weighting estimators (Robins and Ritov, 1997; Abadie and Imbens, 2011), SCM is therefore subject to the curse of dimensionality — the probability that exact balancing weights exist vanishes as the number of time periods grows (Ferman and Pinto, 2018). Recognizing this, we bound the bias due to covariate imbalance for general weighting estimators under a linear factor model; the results in Abadie et al. (2010) are a special case of this bound. We also propose estimating this bias directly using an outcome model. While we advocate using this to de-bias standard SCM, the estimated bias itself is a useful diagnostic for researchers using SCM without augmentation.

Second, we show that, unlike SCM, ASCM extrapolates outside the convex hull of the control units, possibly leading to negative weights. This ensures much closer balance, reducing bias, but rests more heavily on modeling assumptions, such as linearity. We give theoretical results when the outcome model is ridge regression, and show that covariate imbalance and the corresponding bias will generally be lower for ridge-augmented SCM than for either SCM or ridge regression alone. We also show that the ASCM framework can incorporate flexible outcome models, including panel data methods like matrix completion (Athey et al., 2017) and the generalized synthetic control method (Xu, 2017), and off-the-shelf machine learning models like random forests and neural networks. In addition, ASCM can accommodate auxiliary, time-invariant covariates to further reduce bias. Despite these advantages, we recommend that, as with any model-based estimator, users devote extra effort to checking model specification, especially in settings where the ASCM and SCM weights yield different estimates.

Third, we draw on the IPW connection to clarify inference and testing for SCM. Abadie et al. (2010, 2015) propose a widely adopted testing procedure for SCM based on a uniform permutation approach. Firpo and Possebom (2017) interpret this test as a Fisher randomization test, though Abadie et al. (2015) interpret it as a placebo test that does not require randomization for validity. The connection between SCM and IPW suggests that the randomization-based interpretation is a natural one, though other interpretations are possible. From this perspective, however, a uniform permutation test will be invalid — a valid randomization test should weight permutations by the propensity score (Branson and Bind, 2018). Unfortunately, a weighted permutation test is not practical in most SCM settings. We instead propose a model-based inference procedure that is generally conservative, and which performs well in simulations.

Fourth, we contribute to the growing literature on approximate balancing weights in high dimensions (Athey et al., 2018; Tan, 2018; Wang and Zubizarreta, 2018). This literature has focused on settings with sparsity. We extend results to settings with a latent factor model, which motivates using L^2 penalties and ridge regression, rather than L^∞ and Lasso as in existing approaches. Our development will therefore be useful beyond SCM. Finally, we implement the proposed methodology in the `augsynth` package for `R`, available at `github.com/ebenmichael/augsynth`.

The paper proceeds as follows. Section 2 introduces notation and the SCM estimator. Section 3 introduces Augmented SCM and characterizes covariate balance for the special case of ridge ASCM. Section 4 bounds the bias under a linear factor model, the standard setting for SCM, showing that ASCM will generally lead to lower bias than SCM alone. Section 5 demonstrates the equivalence of SCM and IPW. Section 6 discusses testing and inference. Section 7 extends the ASCM framework to incorporate auxiliary covariates. Section 8 reports on numerical illustrations as well as extensive simulation studies. Finally, Section 9 discusses some outstanding issues and 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. We briefly highlight three relevant strands of research.

The first strand assesses the performance of the original SCM estimator under different settings. Kreif et al. (2016), Gobillon and Magnac (2016), Wan et al. (2018), among others, assess the general performance of SCM methods. Botosaru and Ferman (2017) and Kaul et al. (2018) explore the role of auxiliary covariates in SCM and note several pathologies. Ferman and Pinto (2018) consider the behavior of SCM when the weights fail to exactly balance the lagged outcomes, showing that the resulting SCM weights do not converge to oracle weights when the number of time periods grows. Powell (2018) similarly explores variants of SCM without exact balance.

The second strand explores inference and testing for SCM. Abadie et al. (2010, 2015) propose a placebo-based testing approach with a uniform permutation. Firpo and Possebom (2017) interpret this approach as a randomization-based test of the sharp null hypothesis of no impact and derive conditions under which this is valid from a finite sample perspective. Toulis and Shaikh (2018) also take a randomization-based perspective, concluding that the proposed permutation procedure cannot be interpreted as a valid randomization test. Alternatively, Hahn and Shi (2017) assess the same procedure from the perspective of a placebo test, arguing that the approach fails to control Type I error. Ando and Sävje (2013) similarly evaluate this approach under an exchangeability assumption, which they argue is unlikely to be satisfied in practice. Chernozhukov et al. (2017) consider alternative permutations that instead exploit the time series structure of the problem. Finally, several papers consider a sampling approach to inference (Doudchenko and Imbens, 2017; Robbins et al., 2017; Imai et al., 2018).

The third strand extends SCM to allow for more robust estimation and for new data structures. Building on a suggestion in Abadie et al. (2015), several papers have connected SCM to penalized regression (Doudchenko and Imbens, 2017; Abadie and L'Hour, 2018; Minard and Waddell, 2018). Hazlett and Xu (2018) instead outline a promising approach for improving SCM estimation by first using a kernel approach to transform the raw lagged outcomes. There have also been several proposals to use outcome modeling rather than SCM-style weighting. 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). Relatedly, Doudchenko and Imbens (2017) relax the SCM restriction that control unit weights be non-negative, arguing that there are many settings in which negative weights would be desirable.

Finally, our work also builds on the recent literature on balancing weights, also known as calibrated propensity scores. This literature modifies the traditional propensity score estimator and instead estimates weights that directly balance covariate means between treated and control units, rather than balancing them indirectly by first estimating the propensity score coefficients.

Examples include Hainmueller (2011), Graham et al. (2012), Imai and Ratkovic (2013), Zubizarreta (2015), Tan (2017), and Wang and Zubizarreta (2018). Several also combine weighting and outcome modeling, including Athey et al. (2018), Tan (2018), and Hirshberg and Wager (2018). We similarly build on recent results that highlight the connections between covariate balancing approaches and the implied propensity score model. Examples include Robins et al. (2007), Imai and Ratkovic (2013), Zhao and Percival (2017), Zhao (2018), Tan (2017), and Wang and Zubizarreta (2018).

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. Let W_i be an indicator that unit i is treated. We assume that all treated units receive the treatment at a common time $T_0 < T$; units with $W_i = 0$ never receive the treatment. There are a total of $N_1 = \sum_i W_i$ treated units and $N_0 = N - N_1$ control units, often referred to as *donor units* in the SCM context. The outcome is Y and is typically continuous.[1]

We adopt the potential outcomes framework (Neyman, 1923; Rubin, 1974) and invoke SUTVA, which assumes a well-defined treatment and excludes interference between units (Rubin, 1980). The potential outcomes for unit i in period t under control and treatment are $Y_{it}(0)$ and $Y_{it}(1)$, respectively.[2] 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 keep notation simple, we assume that there is only one post-treatment observation, $T = T_0 + 1$, though our results are easily extended to larger $T - T_0$. We therefore use Y_i to represent the single post-treatment observation for unit i, dropping t from the subscript. We use X_{it}, for $t \leq T_0$, to represent pre-treatment outcomes, which emphasizes that pre-treatment outcomes serve as covariates in SCM. With some abuse of notation, we use $X_{0\cdot}$ to represent the N_0-by-T_0 matrix of control unit pre-treatment outcomes and Y_0 for the N_0-vector of control unit post-treatment outcomes. For most of our discussion, we will restrict attention to the case where only a single unit, $i = 1$, is treated; that is, $N_1 = 1$ and $N_0 = N - 1$. Thus, Y_1 is a scalar, and $X_{1\cdot}$ is a T_0 row vector of treated unit pre-treatment outcomes.[3] The data structure is then:

[1] The source of randomness varies across the SCM literature. As we discuss in Section 4, we focus on the setting where the units are fixed and uncertainty comes from noisy realizations of the latent factor model. This is in contrast to randomization inference in which treatment assignment is the only source of randomness.

[2] To simplify the discussion, we assume that potential outcomes under both treatment and control exist for all units in all time periods $t > T_0$. With some additional technical caveats, we could relax this assumption to only require that $Y_{it}(0)$ be well-defined for all units, without also requiring control units to have well-defined potential outcomes under treatment.

[3] With multiple treated units treated at the same time, we can overload Y_1 and $X_{1\cdot}$ to denote *averages* across

$$\begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1T_0} & Y_{1T} \\ Y_{21} & Y_{22} & \dots & Y_{2T_0} & Y_{2T} \\ \vdots & & & & \vdots \\ Y_{N1} & Y_{N2} & \dots & Y_{NT_0} & Y_{NT} \end{pmatrix} \equiv \left(\begin{array}{cccc|c} X_{11} & X_{12} & \dots & X_{1T_0} & Y_1 \\ \hline X_{21} & X_{22} & \dots & X_{2T_0} & Y_2 \\ \vdots & & & & \vdots \\ \underbrace{X_{N1} \quad X_{N2} \quad \dots \quad X_{NT_0}}_{\text{pre-treatment outcomes}} & & & & Y_N \end{array}\right) \equiv \left(\begin{array}{c|c} X_{1\cdot} & Y_1 \\ \hline X_{0\cdot} & Y_0 \end{array}\right) \tag{2}$$

Finally, the treatment effect of interest is $\tau = \tau_{1T} = Y_1(1) - Y_1(0) = Y_1 - Y_1(0)$.

2.2 Synthetic Control Method

The Synthetic Control Method imputes the missing potential outcome for the treated unit, $Y_1(0)$, as a weighted average of the control outcomes, $Y_0'\gamma$. Abadie and Gardeazabal (2003) and Abadie et al. (2010, 2015) propose choosing weights γ as a solution to the constrained optimization problem:

$$\begin{aligned} \min_{\gamma} \quad & \|X_{1\cdot} - X_{0\cdot}'\gamma\|_2^2 \\ \text{subject to} \quad & \sum_{i=2}^{N} \gamma_i = 1 \\ & \gamma_i \geq 0 \quad i = 2, \dots, N \end{aligned} \tag{3}$$

where $\|\cdot\|_2^2$ is the squared 2-norm on $\mathbb{R}^{T_0}$, $\sum_{t=1}^{T_0}(X_{1t} - X_{0t}'\gamma)^2$, and the constraints limit γ to the N_0 simplex, $\gamma \in \Delta^{N_0}$.

Equation (3), known as the *constrained regression* formulation of SCM, follows the recent methodological literature and focuses solely on balancing the lagged outcomes X (see Doudchenko and Imbens, 2017; Ferman and Pinto, 2018; Powell, 2018). By contrast, the original SCM formulation also includes auxiliary covariates Z and a weighted L^2 norm. As Kaul et al. (2018) and others have shown, (3) is identical to the original SCM proposal in a range of practical settings. We discuss additional complications that arise in the original SCM formulation in Section 7.

2.3 SCM generally fails to achieve exact balance

The SCM weights in Equation (3) minimize the imbalance of pre-treatment outcomes between the treated unit and synthetic control. Abadie et al. (2010) show that the resulting estimator is unbiased (asymptotically in T_0) under an assumption that exact balance can be achieved, that is, $X_{1t} - X_{0t}'\hat{\gamma}^{\text{scm}} = 0$ for all t. This is a strong assumption, however. Weights that yield exact balance exist if and only if the treated unit is inside the convex hull of the control units.

treated units. The implicit estimand is the (Sample) Average Treatment Effect on the Treated. We return to this point in the discussion in Section 9.

Achieving exact balance is therefore subject to the curse of dimensionality (Robins and Ritov, 1997; Abadie and Imbens, 2011; D'Amour et al., 2017). Informally, the probability that exact balancing weights exist vanishes as the dimension of the covariates, T_0, grows large. For example, under a population model, Zhao and Percival (2017) show that the probability that the treated unit is in the convex hull of the control units decreases exponentially in the number of covariates. For exact balancing weights to exist (with high probability), the number of control units must therefore be exponentially larger than T_0, a far cry from the typical SCM setting with $N \sim T_0$. Ferman and Pinto (2018) investigate this question in the context of SCM under a linear factor model, arguing that the SCM weights fail to exactly balance the underlying factors with infinite T_0.[4]

As with other matching or weighting estimators, failing to balance covariates can introduce bias.[5] Heuristically, let $Y_i(0) = m(X_i) + \varepsilon_{iT}$, where $m(X_i)$ is some function of lagged outcomes X_i, and ε_{it} is mean-zero noise conditionally independent of treatment assignment. Then the bias for a weighting estimator is $m(X_1) - \sum_{W_i=0} \hat{\gamma}_i m(X_i)$. This is zero if $m(X)$ is linear with respect to X and the weights achieve exact balance, but may be non-zero if either condition is not satisfied. For instance, as we show in Section 4, under a linear factor model, $m(X)$ is approximately linear for large T_0. At the same time, achieving exact balance is particularly challenging for large T_0, suggesting that SCM will be biased. In the next section, we propose to estimate this bias directly and then use this estimate to de-bias the original SCM estimate.

3 Augmented SCM

3.1 Overview

Our main contribution is to propose an Augmented SCM (ASCM) estimator that combines SCM with outcome modeling. Specifically, let $\hat{m}(X_i)$ be an estimated outcome model under control. Then $\hat{m}(X_1) - \sum_{W_i=0} \hat{\gamma}_i \hat{m}(X_i)$ is an estimate of the SCM bias. We propose the following bias-

[4] We confirm this intuition with simulation. In one simple simulation, presented in Figure 1, the probability of achieving exact balance drops from 47 percent with $T_0 = 4$ to 1 percent with $T_0 = 9$. In "calibrated" simulations, presented in Section 8, there is not a single Monte Carlo draw in which SCM achieves exact balance.

[5] One approach to mitigating bias is to proceed with the analysis only if covariate imbalance is small. Specifically, Abadie et al. (2015) recommend against using SCM when "the pre-treatment fit is poor or the number of pre-treatment periods is small." There is little guidance about what constitutes poor fit, however, and common practice is fairly ad hoc. An estimate of the bias based on a prognostic score $\hat{m}(\cdot)$ is a natural summary for assessing quality of fit. However, while this pre-screening might be an effective approach for a specific application, it could lead to a file drawer problem. In our simulations in Section 8, we find that a range of methods give excellent performance when we condition on cases with good covariate balance, but that this conditioning substantially limits the utility of the methods. While SCM does poorly when covariate balance is poor, many alternatives (including ASCM) are more robust to imbalance.

reducing estimator for $Y_1(0)$:

$$\hat{Y}_1^{\text{aug}}(0) = \sum_{W_i=0} \hat{\gamma}_i Y_i \quad + \quad \left(\hat{m}(X_1) - \sum_{W_i=0} \hat{\gamma}_i \hat{m}(X_i) \right), \tag{4}$$

$$= \hat{m}(X_1) \quad + \quad \sum_{W_i=0} \hat{\gamma}_i (Y_i - \hat{m}(X_i)), \tag{5}$$

where $\hat{\gamma}$ are SCM weights. This specializes to standard SCM when we set $\hat{m}(\cdot)$ to be a constant.

Equations (4) and (5), while equivalent, highlight two distinct motivations for ASCM. Equation (4) directly corrects the SCM estimate, $\sum_{W_i=0} \hat{\gamma}_i Y_i$, by the estimated bias, $\hat{m}(X_1) - \sum_{W_i=0} \hat{\gamma}_i \hat{m}(X_i)$. This is analogous to bias correction for inexact matching (Abadie and Imbens, 2011). If the estimated bias is small, then the SCM and ASCM estimates will be similar. Equation (5) is analogous to Augmented IPW (Robins et al., 1994), which begins with the outcome model but uses SCM to re-weight the residuals. This is 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 Tan (2018). We develop the IPW analogy further in Section 5.

The ASCM framework can incorporate any choice of outcome model. For example, we can use recent proposals for outcome modeling in the SCM setting, including matrix completion (`MCP`; Athey et al., 2017), the generalized synthetic control method (`gsynth`; Xu, 2017), and Bayesian structural time series modeling (`causalImpact`; Brodersen et al., 2015). Alternatively, we can use generic supervised learning methods, such as random forests and neural networks. The ASCM framework also nests recent proposals from Doudchenko and Imbens (2017) and Ferman and Pinto (2018) for "de-meaned SCM" or SCM with an intercept shift, which correspond to a simple unit fixed effects outcome model, $m(X_i) = \alpha_i$. Ferman and Pinto (2018) show that this estimator dominates both standard difference-in-differences and SCM, asymptotically in T_0. We explore these options via simulation in Section 8. Finally, in Section 7, we generalize ASCM to include additional covariates beyond lagged outcomes.

3.2 Ridge-augmented SCM

We now explore the special case of ASCM where the outcome model $m(X)$ is fit with (linear) ridge regression. Using results from survey sampling, we first show that we can write this estimator as a single weighting estimator (possibly with negative weights) that adjusts the SCM weights to allow for better balance. Using this equivalence, we then show that ridge-augmented SCM generally has better covariate balance than either SCM or ridge regression alone, though it may have higher variance. While we restrict our theoretical results to ridge-augmented SCM, we anticipate that these can be extended to more general, non-linear outcome models via the model calibration weights

framework of Wu and Sitter (2001).

Let $\hat{m}(X_i) = X_{i\cdot}'\hat{\eta}$, where $\hat{\eta}$ are the coefficients of a ridge regression of control post-treatment outcomes Y_0 on pre-treatment outcomes $X_{0\cdot}$ with penalty hyper-parameter λ. Ridge regression, like OLS, has a closed-form expression as a weighting estimator (see Appendix A.1). Thus, the augmented SCM estimator merely combines two different weighting estimators, SCM and ridge regression.

Lemma 1. The ridge-augmented SCM estimator specialization of (5) is:

$$\hat{Y}_1^{\text{aug}}(0) = \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} Y_i \quad + \left(X_{1\cdot} - \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} X_{i\cdot} \right) \cdot \hat{\eta} = \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} Y_i, \tag{6}$$

where

$$\hat{\gamma}_i^{\text{aug}} = \hat{\gamma}_i^{\text{scm}} + (X_1 - X_{0\cdot}'\hat{\gamma}^{\text{scm}})'(X_{0\cdot}'X_{0\cdot} + \lambda I_{T_0})^{-1} X_{i\cdot}. \tag{7}$$

Lemma 1 shows that ASCM weights — which, unlike SCM weights, can be negative — adjust the raw SCM weights to achieve better covariate balance. When SCM weights exactly balance the lagged outcomes, ridge ASCM and SCM weights are equivalent, and when SCM yields good balance or the tuning parameter λ is large, the estimated bias is small and the two weights are close to each other. Conversely, when SCM has poor balance and λ is small, the adjustment will be large and the weights will be far apart.[6] As a result, ridge-augmented SCM weights will generally achieve better pre-treatment fit than weights from SCM alone, although at the cost of higher variance.

Lemma 2. Let $\sigma_{\min}^2$ be the minimum eigenvalue of the sample covariance matrix $\frac{1}{N_0} X_{0\cdot}' X_{0\cdot}$. Then with $\kappa = \frac{\lambda}{N_0 \sigma_{\min}^2} + 1 > 1$, the ridge-augmented SCM weights $\hat{\gamma}^{\text{aug}}$ satisfy:

$$\|X_1 - X_{0\cdot}'\hat{\gamma}^{\text{aug}}\|_2 \leq \frac{\kappa - 1}{\kappa} \quad \left\| X_1 - X_{0\cdot}'\hat{\gamma}^{\text{scm}} \right\|_2 \tag{8}$$

and

$$\|\hat{\gamma}^{\text{aug}}\|_2 \leq \|\hat{\gamma}^{\text{scm}}\|_2 + \frac{1}{\kappa \sigma_{\min} \sqrt{N_0}} \quad \left\| X_1 - X_{0\cdot}'\hat{\gamma}^{\text{scm}} \right\|_2 \tag{9}$$

Since $\frac{\kappa-1}{\kappa} < 1$, ridge-augmented SCM will have strictly better covariate balance than SCM alone, except in the special case of exact balance. At the same time, ridge-augmented SCM weights will generally have larger variance than pure SCM weights, with larger discrepancies for worse SCM imbalance. Intuitively, this larger variance arises because, unlike SCM weights, ASCM weights can be negative, which can increase the spread of the weights.

Allowing for negative weights is an important departure from standard SCM: Abadie et al. (2010, 2015) argue that negative weights are undesirable because they are difficult to interpret and

[6] In practice we choose λ using K-fold cross-validation for the outcome ridge regression.

allow for extrapolation.[7] The SCM non-negativity constraint implies choosing a synthetic control within the convex hull created by the control units, even when the treated unit lies outside this hull, and the distance from the treated unit to the convex hull creates the potential for bias. Thus, even with moderate dimensional covariates, an estimator that constrains weights to be non-negative will be biased in practice. ASCM, by contrast, uses negative weights to extrapolate outside of the convex hull, ensuring much closer balance but resting more heavily on the assumption that the expected value of $Y_1(0)$ is (approximately) linear in the control outcomes. We confirm this intuition with simulations in Section 8. In cases where covariate balance is excellent, there is little penalty to restricting weights to be non-negative; otherwise, this constraint can lead to severe bias, at least relative to ASCM. Doudchenko and Imbens (2017) make a similar point with an analogy to bias correction for matching, arguing that negative weights play an important role in reducing bias when exact matches are infeasible.

Finally, in the appendix we compare ridge-augmented SCM to ridge regression alone, again relying on a representation of ridge regression as a weighting estimator. As we show, the ridge regression weights correspond to a special case of the elastic-net synthetic controls estimator proposed by Doudchenko and Imbens (2017), with the elastic-net parameter set to zero, and to a penalized version of the Oaxaca-Blinder weights considered by Kline (2011). Ridge regression achieves worse balance than ridge ASCM, but yields lower sampling variance. In the special case with no regularization, ridge regression weights reduce to the standard regression weights discussed in Abadie et al. (2015), and both standard regression and ridge ASCM will yield perfect balance.

4 Bias under a linear factor model

We now characterize the bias under the linear factor model considered in Abadie et al. (2010), with treatment assignment that is ignorable given unobserved factor loadings. We begin with bias for a general weighting estimator and then turn to SCM, showing that the results in Abadie et al. (2010) are a special case of our bound. Finally, we show that ridge ASCM will generally have a tighter bias bound than SCM alone. We confirm these results with simulations in Section 8.

4.1 Bias for a general weighting estimator

Following the setup in Abadie et al. (2010), we assume that there are J latent time-varying factors $\mu_t = \{\mu_{jt}\}$, $j = 1, \ldots, J$, with $\max_{jt} |\mu_{jt}| \leq M$, where J will typically be small relative to N. Each unit has a vector of factor loadings $\phi_i \in \mathbb{R}^J$. Control potential outcomes are weighted averages of

[7] The question of interpreting negative weights has been heavily debated in the surveys literature, where negative weights can arise with (generalized) regression estimates (Fuller, 2002). The main drawback is that, from a design-based perspective, negative survey weights can no longer be interpreted as sampling weights (see Lohr, 2007). From a model-based perspective, however, requiring non-negative weights is technically arbitrary.

these factors plus additive noise ε_{it}:

$$Y_{it}(0) = \sum_{j=1}^{J} \phi_{ij}\mu_{jt} + \varepsilon_{it} = \phi_i'\mu_t + \varepsilon_{it}, \tag{10}$$

where the only random quantities are the noise terms ε_{it}. Slightly abusing notation, we collect the pre-intervention factors into a matrix $\mu \in \mathbb{R}^{T_0 \times J}$, where the tth row of μ contains the factor values at time t, μ_t'.

Following Bai (2009) and Xu (2017), we assume that the factors are orthogonal and normalized, i.e. that $\frac{1}{T_0}\mu'\mu = I_J$, and that treatment assignment W_i is ignorable given a unit's factor loadings,

$$\mathbb{E}_{\varepsilon_T}\left[Y_i(0) \mid \phi_i, W_i\right] = \mathbb{E}_{\varepsilon_T}\left[Y_i(0) \mid \phi_i\right],$$

where the expectation is taken with respect to $\varepsilon_T = (\varepsilon_{1T}, \ldots, \varepsilon_{NT})$. We further assume that the error terms ε_{it} are independent (across units and over time) sub-Gaussian random variables with scale parameter σ.

Under the linear factor model and ignorability given ϕ_i, an estimator that balances ϕ_i will yield an unbiased estimate of $Y_1(0)$, and if exact balance is not achieved the bias will be proportional to the level of imbalance in ϕ_i. However, ignorability given the latent factors ϕ_i does *not* generally imply ignorability given the observable X_i, and ensuring balance in X_i will not necessarily ensure balance in ϕ_i. We show in the appendix that the bias of any weighting estimator with weights γ can be expressed as

$$\begin{aligned}\mathbb{E}_{\varepsilon_T}\left[Y_1(0) - \sum_{W_i=0}\gamma_i Y_i\right] &= \left(\phi_1 - \sum_{W_i=0}\gamma_i\phi_i\right)\cdot\mu_T \\ &= \frac{1}{T_0}\underbrace{\left(\mu' X_1 - \sum_{W_i=0}\gamma_i\mu' X_i\right)}_{\text{imbalance in } X}\cdot\mu_T - \frac{1}{T_0}\underbrace{\left(\mu'\varepsilon_1 - \sum_{W_i=0}\gamma_i\mu'\varepsilon_i\right)}_{\text{approximation error}}\cdot\mu_T,\end{aligned} \tag{11}$$

The first term is the imbalance of observed lagged outcomes and the second term is an approximation error arising from the latent factor structure. Using our assumption that the noise is sub-Gaussian, we can bound this bias using the triangle inequality.

Theorem 1. Under the linear factor model (10), for any weights $\gamma \in \mathbb{R}^{N_0}$ such that $\sum_i \gamma_i = 1$,

$$\left|\mathbb{E}_{\varepsilon_T}\left[Y_1(0) - \sum_{W_i=0}\gamma_i Y_i\right]\right| \leq M\sqrt{J}\underbrace{\frac{1}{\sqrt{T_0}}\|X_{1\cdot} - X_{0\cdot}'\gamma\|_2}_{\text{imbalance in } X} + \underbrace{M^2 J\frac{C}{\sqrt{T_0}}\left(\frac{1}{\sqrt{N_1}} + \|\gamma\|_2\right)}_{\text{approximation error}} \tag{12}$$

with probability at least $1 - 2e^{-8J \log 5} > 0.99$, where $\max_{jt} |\mu_{jt}| \leq M$ and $C = 8\sigma\sqrt{\log 5}$.

This result holds for any weights that sum to one, and does not require non-negativity. The approximation error is generally non-zero since the outcome is linear in ϕ rather than linear in X; however, with large T_0 the approximation error will be small. Interestingly, this term depends on the L^2 norm of the weights, $\|\gamma\|_2$, which does not generally enter into bias bounds in simpler settings. Similar results exist for bias in other panel data settings. For instance, the approximation error in Equation (12) is analogous to so-called "Nickell (1981) bias" that arises in short panels.

4.2 Bias for SCM

To apply Theorem 1 to SCM, we consider two cases: the special case with exact balance and the more general case with approximate balance. In the exact balance case, the first term of (12) is zero and the second term goes to zero, with high probability, as $T_0 \to \infty$. This is the basis of the Abadie et al. (2010) claim that SCM is asymptotically (nearly) unbiased in a factor model with exact balance. Intuitively, the lagged outcome for unit i at time t, X_{it}, is a noisy proxy for the index $\phi_i'\mu_t$. Thus, as we observe more X_{it} — and can exactly balance each one — we are better able to match on this index and, as a result, on the underlying factor loadings.

If we do not assume exact balance, the bias bound contains two competing terms: the approximation error, which is decreasing in T_0, and the imbalance in X, which is non-decreasing in T_0. Figure 1 illustrates this tension via simulation.[8] As T_0 increases, the average covariate imbalance (in terms of RMSE in X) also increases, eventually leveling off. The bias decreases initially as the approximation error falls, but levels off to a positive value as imbalance comes to dominate.

While Theorem 1 only provides an upper bound on the bias, rather than an expression for the bias itself, we nonetheless argue that SCM will still be biased with large T_0. Ferman and Pinto (2018) show that — even if weights exist that exactly balance the latent factor loadings — as $T_0 \to \infty$ the SCM weights will converge to a solution that *does not* exactly balance the latent factors, leading to bias. Theorem 1 complements their asymptotic analysis with a finite sample bound that holds for all weighting estimators, and explicitly includes the level of imbalance in the lagged outcomes. In Appendix A.3 we discuss further connections to their results and the duality between SCM and IPW presented in the following section. Our simulation evidence, both in Figure 1 and in Section 8, is consistent with conclusion that SCM is biased in a range of scenarios.

[8] We perform 1000 simulations of a simple fixed effects model (which is a special case of the linear factor model) where fixed effects are drawn from a mixture of two Gaussians with $N = 100$, $N_1 = 1$, and varying T_0. The assumed means are 0 and 0.2, with common variance 0.1, mixing proportion 0.8, and additive noise $\varepsilon_{it} \sim N(0, 0.5^2)$. We use a logistic selection model with the (unobserved) fixed effects, and normalize the probabilities so that a single unit is treated. With this setup the treated unit's fixed effect is typically in the convex hull of the control fixed effects.

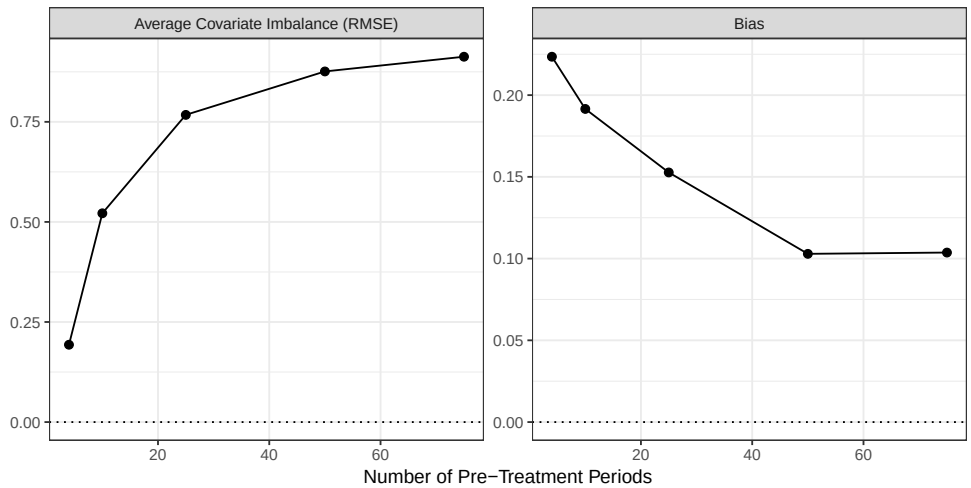


Figure 1: Covariate imbalance (in terms of RMSE) and bias for a simple fixed effects model with $N = 100$ and increasing T_0.

4.3 Bias for Augmented SCM

Lemma 2 shows that ridge ASCM will generally have better covariate balance than SCM. We now combine this with Theorem 1 to show that ridge ASCM will also have a tighter bias bound than SCM alone. While our theoretical discussion is again limited to bounds on the bias, simulations in Section 8 confirm that the realized bias is consistently smaller for ridge ASCM than SCM alone, at least in the settings we consider.

Corollary 1. Let $\sigma^2_{\min}$ be the minimum eigenvalue of the sample covariance matrix $\frac{1}{N_0} X_{0\cdot}' X_{0\cdot}$. Then with $\kappa = \frac{\lambda}{N_0 \sigma^2_{\min}} + 1 > 1$, the ridge-augmented SCM weights (6) exist and under the linear factor model (10) the bias satisfies the bound

$$\left| \mathbb{E}_{\varepsilon_T} \left[Y_1(0) - \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} Y_i \right] \right| \leq \underbrace{\frac{M\sqrt{J}}{\sqrt{T_0}} \left(\frac{\kappa - 1}{\kappa} + \frac{8\sigma M \sqrt{J}}{\sqrt{N_c}\sigma_{\min}\kappa} \right) \|X_{1\cdot} - X_{0\cdot}' \hat{\gamma}^{\text{scm}}\|_2}_{\text{imbalance in } X} + \underbrace{M^2 J \frac{C}{\sqrt{T_0}} \left(\frac{1}{\sqrt{N_1}} + \|\hat{\gamma}^{\text{scm}}\|_2 \right)}_{\text{approximation error}} \tag{13}$$

with probability at least $1 - 2e^{-8J \log 5} > 0.99$, where $\max_{jt} |\mu_{jt}| \leq M$ and $C = 8\sigma\sqrt{\log 5}$.

As in the general case in Theorem 1, the level of imbalance and the complexity of the weights play possibly competing roles. Specifically, the regularization parameter λ (or the transformation κ) controls how much to prioritize balance. In the extreme case with $\lambda = \infty$, ridge ASCM is equivalent to SCM. As we reduce λ, the imbalance in X decreases but $\|\hat{\gamma}^{\text{aug}}\|_2$ increases, which increases the

approximation error. Note that if $X_{0\cdot}$ is full rank (requiring $N > T_0$), ridge ASCM weights exists with $\lambda = 0$ and ridge ASCM *exactly balances* the lagged outcomes, even where non-augmented SCM does not.

Finally, while the constant terms in Theorem 1 and Corollary 1 are relatively loose, it is instructive to consider the bias bounds for SCM and ridge ASCM when the number of control units is large relative to the number of latent factors. In this case, the approximation error is small and it is bias-optimal to minimize imbalance in the lagged outcomes, setting λ to be small. Thus, the bias bound for ridge ASCM will be lower than the corresponding bound for SCM. As we show through simulation in Section 8, this intuition holds even when the number of control units is only slightly larger than the number of latent factors, and even in this case the reduction in bias outweighs any increase in variance.

5 Connection to other balancing estimators

We now connect SCM to other balancing estimators, which builds intuition, helps to motivate the Augmented SCM estimator proposed in Section 3, and gives additional clarity for inference and testing. First, we notice that a form of SCM, which we call penalized SCM, is a special case of an approximate covariate balancing weights estimator. Second, we extend existing results on the duality between balancing weights and inverse propensity score weights to show that SCM is indeed a form of inverse propensity score weighting.

5.1 Penalized SCM

There may be no unique solution to the original SCM problem in Equation (3), if multiple sets of weights achieve exact balance on the lagged outcomes. Following Abadie et al. (2015) and Doudchenko and Imbens (2017), we modify the original SCM procedure to penalize the dispersion of the weights with a strongly convex *dispersion function*, $f(\gamma_i)$. For a sufficiently small penalty, penalized SCM will be nearly identical to standard SCM in cases where the latter has a unique solution. However, the penalized SCM problem is guaranteed to have a unique solution for any positive penalization, which is analytically convenient.

To fix ideas, we consider the entropy penalty, $f(\gamma_i) = \gamma_i \log \gamma_i$, used in Robbins et al. (2017).[9]

[9] Many other dispersion functions are possible, including an elastic net penalty (Doudchenko and Imbens, 2017), a measure of pairwise distance (Abadie and L'Hour, 2018), and a measure of outcome variance (Minard and Waddell, 2018); we discuss alternative dispersion penalties in Section 5.3. An additional motivation for penalized SCM is that the standard SCM weights can often be unstable, with very few (often just three or four) donor units receiving positive weights. While this may minimize bias, it has high variance; in practice, researchers might want to accept some bias for lower variance.

Specifically, the *entropy penalized SCM* weights solve:

$$\begin{aligned} \min_{\gamma} \quad & \frac{1}{2\zeta}\|X_{1\cdot} - X_{0\cdot}'\gamma\|_2^2 + \sum_{W_i=0} \gamma_i \log \gamma_i \\ \text{subject to} \quad & \sum_{W_i=0} \gamma_i = 1 \\ & \gamma_i \geq 0 \quad i = 2, \ldots, N \end{aligned} \tag{14}$$

The hyperparameter $\zeta > 0$ sets the relative priority of minimizing the entropy term, which penalizes very large and very small weights, versus covariate imbalance. A larger ζ means greater covariate imbalance, and thus more bias, but lower variance due to the more dispersed weight vector. As with other regularized estimators, if a unique solution exists to the un-regularized (3), we can find an arbitrarily close solution to (14) by setting ζ sufficiently small.

With this setup, we can immediately see that *entropy balancing* (Hainmueller, 2011) is a special case of Equation (14) in which the weights yield exact balance. Thus, entropy penalized SCM can also be motivated as entropy balancing with approximate—rather than exact—balance. More broadly, SCM is one example of a broad class of *covariate balancing weights* that have recently become more prominent. In addition to entropy balancing, examples include stable balancing weights (Zubizarreta, 2015), approximate residual balancing (Athey et al., 2018), and several others (Tan, 2017; Wang and Zubizarreta, 2018; Zhao, 2018). See Graham et al. (2012), Chan et al. (2016), and Li et al. (2017) for related examples with slightly different forms. We return to the general case in Section 5.3.

5.2 SCM is regularized IPW

Once we recognize that SCM yields covariate balancing weights, we can leverage recent results connecting balancing and propensity score weights to show that SCM is a form of inverse propensity score weighting. First, we show the duality between entropy-penalized SCM with exact balance and unregularized IPW. This result follows Zhao and Percival (2017), though our argument is more direct. Second, we relax the restriction that the weights yield exact balance. In this case, SCM is equivalent to regularized IPW. We extend results from Wang and Zubizarreta (2018) to the case of an L^2 norm and show that entropy-penalized SCM maps to a ridge penalty on the propensity score coefficients. In the appendix we derive a general duality between balancing weights and propensity score estimation that encompasses many estimators in the literature, including extensions to other link functions.

5.2.1 Penalized SCM with exact balance is unregularized IPW

We consider the penalized SCM estimator in Equation (14). Enforcing exact balance, this estimator can be rewritten as:

$$\begin{aligned} \min_{\gamma} \quad & \sum_{W_i=0} \gamma_i \log \gamma_i \\ \text{subject to} \quad & X_{1\cdot} = X_{0\cdot}'\gamma \\ \text{and} \quad & \sum_{W_i=0} \gamma_i = 1. \end{aligned} \tag{15}$$

Note that the entropy dispersion penalty includes the SCM non-negativity constraint. In the appendix, we show that this problem has an equivalent Lagrangian dual:

$$\min_{\alpha,\beta} \sum_{W_i=0} \exp(\alpha + \beta' X_{i\cdot}) - (\alpha + \beta' X_{1\cdot}), \tag{16}$$

where β and α are the Lagrange multipliers on the two constraints in (15). The solution to (16), $\{\hat{\alpha}, \hat{\beta}\}$, corresponds to weights on the control units:

$$\hat{\gamma}_i = \exp(\hat{\alpha} + \hat{\beta}' X_{i\cdot}) = \frac{\text{logit}^{-1}(\hat{\alpha} + \hat{\beta}' X_{i\cdot})}{1 - \text{logit}^{-1}(\hat{\alpha} + \hat{\beta}' X_{i\cdot})}.$$

With this link between the primal problem (15) and the Lagrangian dual (16), we can see that the exact balance constraints in the primal correspond to the first order conditions of the dual:

$$\begin{aligned} \sum_{W_i=0} \underbrace{\exp(\alpha + \beta' X_{i\cdot})}_{=\gamma_i} &= 1 \qquad && \text{and} \\ \sum_{W_i=0} \underbrace{\exp(\alpha + \beta' X_{i\cdot})}_{=\gamma_i} X_{it} &= X_{1t} \qquad && t = 1, \ldots, T_0. \end{aligned} \tag{17}$$

Thus, the donor weights $\hat{\gamma}$ have the form of IPW weights with a logistic link function, where the propensity score is $\pi(x_i) = \text{logit}^{-1}(\alpha + \beta' x_i)$ and the odds of treatment are $\frac{\pi(x_i)}{1-\pi(x_i)} = \exp(\alpha + \beta' x_i) = \gamma_i$. Importantly, the parameters in Equation (17) are fit via *calibrated* estimation rather than the two-step, maximum likelihood-based approach that is standard for IPW but is impractical in typical SCM settings.[10] Specifically, the Lagrangian dual (17) fits the propensity score coefficients so that the implied odds of treatment, $\gamma_i = \exp(\alpha + \beta' X_{i\cdot})$, lie on the simplex and the weighted mean of the control units' covariates is exactly equal to the value of the treated unit, $\sum_{W_i=0} X_i' \gamma_i = X_1$.

[10] In traditional IPW, the propensity score model is first estimated via maximum likelihood, then the estimated propensity scores $\widehat{e}(X_i)$ are used to form IPW weights. McCullagh and Nelder (1989) and King and Zeng (2001) show that the MLE for logistic regression can be badly biased when the response is a rare event. Since $1 - \widehat{e}(X_i)$ is in the denominator, the re-weighting step can amplify this bias. Calibrated propensity score estimation also yields consistent estimates of the true propensity score under appropriate conditions; see, for example, Zhao and Percival (2017) and Tan (2017).

See Tan (2017) and Wang and Zubizarreta (2018) for additional discussion.

5.2.2 Penalized SCM with approximate balance is regularized IPW

We now relax the unrealistic constraint that SCM yield exact balance. In the dual perspective of balancing weights as IPW, allowing for approximate balance is equivalent to regularizing the propensity score model. Specifically, analogous to the argument under exact balance, the Lagrangian dual to the entropy-penalized SCM problem (14) includes a propensity score model that now includes a ridge penalty on the propensity score coefficients:

$$\min_{\alpha,\beta} \sum_{W_i=0} \exp(\alpha + \beta' X_{i\cdot}) - (\alpha + \beta' X_{1\cdot}) + \frac{\zeta}{2}\|\beta\|_2^2. \tag{18}$$

In this form it is clear that ζ controls the level of regularization of the propensity score parameters, which maps back to the weights γ_i. When ζ is large, the parameter estimates $\hat{\beta}$ will be near zero, implying that the weights will be near uniform. Conversely, when ζ is small, $\hat{\beta}$ may be large in magnitude, allowing for extreme weights that prioritize lower bias at the price of higher variance. In practice, SCM implicitly chooses ζ to be as small as possible, so weights are extreme; it is common for only three or four units to receive positive weights. See Doudchenko and Imbens (2017) for a discussion of choosing ζ by cross validation.

5.3 General primal-dual connection

Finally, we briefly characterize the entropy-penalized SCM problem as a special case of a broader class of problems. Specifically, we can re-write the penalized SCM problem in Equation (14) in a more general form (see Ben-Michael et al., 2018, for additional discussion):

$$\begin{aligned} \min_{\gamma} \quad & \underbrace{h_\zeta(X_{1\cdot} - X_{0\cdot}'\gamma)}_{\text{balance criterion}} + \sum_{W_i=0} \underbrace{f(\gamma_i)}_{\text{dispersion}} \\ \text{subject to} \quad & \sum_{W_i=0} \gamma_i = 1. \end{aligned} \tag{19}$$

This general form has two key components: f is a strongly convex measure of dispersion (e.g. entropy), which guarantees uniqueness, and h_ζ is a measure of distance (e.g. the L^2 norm). This formulation covers several estimators, including penalized SCM, entropy balancing (Hainmueller, 2011), Oaxaca-Blinder weights (Kline, 2011), and minimal approximately balancing weights (Wang and Zubizarreta, 2018). In the appendix we extend the above arguments to derive the Lagrangian dual of this general balancing weights problem (19):

$$\min_{\alpha,\beta} \sum_{W_i=0} f^*(\alpha + \beta' X_{i\cdot}) - (\alpha + \beta' X_{1\cdot}) + h_\zeta^*(\beta), \tag{20}$$

where a convex, differentiable function g has convex conjugate $g^*(y) \equiv \sup_{x \in \text{dom}(g)}\{y'x - g(x)\}$. The solutions to the primal problem (19) are $\hat{\gamma}_i = f^{*\prime}(\hat{\alpha} + \hat{\beta}' X_i)$, where $f^{*\prime}(\cdot)$ is the first derivative of the convex conjugate, $f^*(\cdot)$.

The two components of the primal problem (19) control the propensity score model and how it is regularized. The dispersion measure $f(\cdot)$ determines the link function of the propensity score model, where the odds of treatment are $\frac{\pi(x)}{1-\pi(x)} = f^{*\prime}(\alpha + \beta' x)$. Note that un-penalized SCM, which can yield multiple solutions, does not have a well-defined link function. 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.

6 Testing and inference

6.1 Permutation-based testing

Abadie et al. (2010, 2015) propose a widely adopted testing procedure for SCM based on a uniform permutation test. Interpretation of this test, however, is unsettled. Abadie et al. (2015) justify it as a placebo test that does not require randomization for validity.[11] Firpo and Possebom (2017) instead justify it as a Fisher randomization test with uniform treatment assignment and then assess sensitivity to this assumption.[12]

The IPW equivalence suggests that the randomization-based perspective is a natural interpretation. A valid randomization-based test for an IPW estimator, however, requires weighting the permutation distribution by the propensity score. In the appendix, we describe valid testing with known, varying propensity scores (see Branson and Bind, 2018), and show that the uniform permutation test differs from this ideal case. Thus, we argue that uniform permutation testing will generally be invalid from the randomization-based perspective, though other perspectives are possible.

A weighted permutation test using the estimated propensity score is promising in principle. Toulis and Shaikh (2018) give conditions under which the corresponding p-values will be valid asymptotically. In practice, however, there is little reason to expect those conditions to hold in

[11] Hahn and Shi (2017) argue that the placebo test interpretation only holds under strong assumptions on both the model and outcome distribution. Ando and Sävje (2013) show that a similar motivation, which requires exchangeability of the test statistic, is unlikely to hold since units on the border of the convex hull of the factor loadings are systematically different from the treated unit, which is typically assumed to be inside the convex hull.

[12] Chernozhukov et al. (2017) propose an alternative permutation approach that permutes the time periods rather than the treatment assignment. The key assumption for the validity of this approach is unbiasedness of the estimator, which, as we argue, is likely violated in practice for SCM. Simulation results in the appendix confirm that this approach is invalid in the settings we consider.

typical, finite sample SCM settings, with a heavily regularized propensity score based on a single, treated unit. Simulation evidence presented in the appendix confirms this pessimistic outlook. Moreover, even if this approach were valid, it would be impractical in most SCM settings: since SCM weights are generally sparse, few permutations will have non-zero weight, and the p-value distribution under both the null and alternative will be pathological. For instance, it is quite common for SCM weights to be positive for only three or four donor units. If four units each get equal weight, the lowest possible p-value is $p_{\text{scm}} = 0.25$.

6.2 Model-based inference

As an alternative to permutation testing, we propose a model-based approach to inference based on the placebo distribution and show that the resulting inference is conservative under some assumptions on the error distribution. Specifically, we consider a generic outcome model with independent, sub-Gaussian noise ε_{it}, $Y_{it}(0) = m_{it} + \varepsilon_{it}$. To simplify exposition and notation, we initially assume that the noise terms ε_{it} are homoskedastic at each time t, with common variance $\nu_t = \text{Var}(\epsilon_{it})$. Under this model, the variance of the SCM or ASCM treatment effect estimate is:

$$\text{Var}_{\varepsilon_T}(\hat{\tau}) = \text{Var}_{\varepsilon_T}\left(Y_{1T} - \sum_{W_i=0} \hat{\gamma}_i Y_{iT}\right) = \left(\frac{1}{N_1} + \|\hat{\gamma}\|_2^2\right)\nu_T, \tag{21}$$

where ν_T is the residual variance at post-treatment time T and where the variance conditions on the observed lagged outcomes and treatment assignment. With uniform weights, Equation (21) reduces to the usual variance of the difference in means under homoscedasticity.

We propose to estimate the noise variance ν_T via the average squared *placebo gap*:

$$\hat{\nu}_T = \frac{1}{N_0} \sum_{W_i=0} (Y_i - \tilde{Y}_i)^2, \tag{22}$$

where $\widetilde{Y}_i$ is the leave-one-out SCM estimate of Y_i, $\sum_{j \neq i} \widetilde{\gamma}_j^{(i)} Y_j^{\text{obs}}$. Importantly, $\hat{\nu}_T$ is a *conservative* estimator for ν_T, in the sense of having a positive bias. To see this, note that the variance of placebo gaps is strictly larger than the variance of the noise: $\text{Var}_{\varepsilon_T}(Y_i - \tilde{Y}_i) = \left(1 + \|\tilde{\gamma}^{(i)}\|_2^2\right)\nu_T > \nu_T$. We then estimate the sampling variance of $\hat{\tau}$ as

$$\widehat{\text{Var}}(\hat{\tau}) = \left(\frac{1}{N_1} + \|\hat{\gamma}\|_2^2\right)\hat{\nu}_T. \tag{23}$$

Proposition 1 shows that this is a conservative estimate for $\text{Var}(\hat{\tau})$.

Proposition 1. Let $\text{Bias}_i \equiv \mathbb{E}_{\varepsilon_T}[Y_i - \tilde{Y}_i]$. Under the linear factor model (10), $\widehat{\text{Var}}(\hat{\tau})$ is conservative:

$$\mathbb{E}_{\varepsilon_T}\left[\widehat{\text{Var}}(\hat{\tau}) - \text{Var}(\hat{\tau})\right] = \left(\frac{1}{N_1} + \|\hat{\gamma}\|_2^2\right)\left[\frac{1}{N_0}\sum_{W_i=0}\text{Bias}_i^2 + \frac{\nu_T}{N_0}\sum_{W_i=0}\|\tilde{\gamma}^{(i)}\|_2^2\right] > 0. \tag{24}$$

Proposition 1 shows that the upward bias of $\widehat{\text{Var}}(\hat{\tau})$ depends on the bias of the placebo estimates and the L^2-norm of the placebo weights.[13] Building on results from Section 4, these quantities will be small when the control units are similar to each other and tightly packed together, which will limit both the bias and the spread of the weights. Conversely, the upward bias of $\widehat{\text{Var}}(\hat{\tau})$ will be large if there are extreme outliers in the control group.

Many modifications to the simple variance estimate in (23) are possible. For example, we can divide the ith squared placebo gap by its contribution to the bias, $1 + \|\tilde{\gamma}^{(i)}\|_2^2$. We can similarly extend to the fully heteroskedastic case by weighting the mean squared placebo gap. In addition, we can extend this variance estimate to the setting with multiple post-treatment time periods by separately estimating $\hat{\nu}_t$ for each post-treatment time period and plugging in to (23).

While we advocate the model-based inference approach, there are many possible alternatives. First, the bootstrap is an attractive approach for inference for IPW estimators in cross-sectional settings (Funk et al., 2011). The SCM setting, however, falls between traditional IPW, where standard case-resampling bootstrap works well, and matching-style estimators, where it can fail (Abadie and Imbens, 2008). The simulations in Section 8 include a simple non-parametric bootstrap and show that it has poor coverage. Second, Robbins et al. (2017) propose estimating standard errors following a survey sampling approach. This method, however, is restricted to estimators with non-negative weights. Finally, we could pursue a design-based approach to inference, which follows naturally from both the IPW perspective and the randomization-based testing approach for SCM (e.g., Doudchenko and Imbens, 2017). Our initial simulations, however, found that both the design-based and survey sampling approaches performed poorly in settings with a single treated unit. We leave a thorough investigation to future work.

7 Incorporating auxiliary covariates

7.1 Extending Augmented SCM

So far, we have focused on a simplified version of SCM that uses only pre-treatment outcomes as covariates. The original SCM formulation, however, also includes auxiliary covariates $Z_i \in \mathbb{R}^d$, which are typically time invariant and which can also include functions of the lagged outcomes, such as the pre-treatment average. We assume these covariates are centered by the control unit

[13] Doudchenko and Imbens (2017) use a different justification for using $\hat{\nu}_T$ as the estimate for $\text{Var}_{\varepsilon_T}(\hat{\tau})$. Proposition 1 shows that in some cases $\hat{\nu}_T$ may be conservative enough to give valid confidence intervals for $\hat{\tau}$; these intervals may undercover in other situations, however, especially with more than one treated unit.

averages, and for now assume that the matrix of control unit covariates, Z_0, is full rank. Abadie et al. (2010) propose a variant of Equation (3) that minimizes a weighted imbalance measure in Z, with weights chosen to minimize imbalance of the lagged outcomes X. Using our IPW perspective, we show in Appendix A.4.1 that this approach fits a (calibrated) propensity score model using Z as covariates and a weighted ridge penalty, with weights chosen to minimize the imbalance in X.[14]

Our results suggest three transparent alternatives for incorporating auxiliary covariates. First, we can expand ASCM to include Z alongside X, both in the SCM balance criterion and in the outcome model, $m(X_i, Z_i)$. This is natural from the (augmented) IPW perspective; including Z in the balance criterion is equivalent to including it in the propensity score equation. See Appendix A.4.2 for additional details, including the possibility of giving different weights to X and Z in the SCM balance criterion and the outcome model regularization.

Second, we can partition the variables in the ASCM framework, using SCM to balance the lagged outcomes and the outcome model to adjust for possible bias due to imbalance in the auxiliary covariates. Specifically, this approach uses our ASCM proposal, Equation (5), but expresses the outcome model solely in terms of Z. Because we assume that Z is full rank, we can fit this model by OLS, yielding $\hat{m}(Z_i) = Z_i'\hat{\eta}^{\text{ols}}$, where $\hat{\eta}^{\text{ols}} = (Z_0'Z_0)^{-1}Z_0'Y_0$.

To understand this approach, consider linear projections of X and Y onto Z, using coefficients estimated from control observations. We can write the projections as $\hat{X} = HX_{0\cdot}$ and $\hat{Y} = HY_0$, where $H = Z(Z_0'Z_0)^{-1}Z_0'$, and write the residuals from these projections as $\check{X} \equiv X - \hat{X}$ and $\check{Y} \equiv Y - \hat{Y}$, respectively. Lemma 3 indicates that ASCM with OLS on Z as the outcome model is a weighting estimator that exactly balances the covariates Z and thus the projections $\hat{X}$. The remaining imbalance comes solely from the residual component $\check{X}$.

Lemma 3. For any weight vector $\hat{\gamma}$ that sums to one, the ASCM estimator from Equation (5) that uses $\hat{m}(Z_i) = Z_i'\hat{\eta}^{\text{ols}}$ as the outcome model is a weighting estimator,

$$\tilde{Y}_1 - \sum_{W_i=0} \hat{\gamma}_i \tilde{Y}_i = Y_1 - \sum_{W_i=0} \hat{\gamma}_i Y_i \;\; - \left(Z_{1\cdot} - \sum_{W_i=0} \hat{\gamma}_i Z_{i\cdot} \right)' \hat{\eta}^{\text{ols}} = Y_1 - \sum_{W_i=0} \hat{\gamma}_i^{\text{cov}} Y_i, \tag{25}$$

where

$$\hat{\gamma}_i^{\text{cov}} = \hat{\gamma}_i + (Z_1 - Z_0'\hat{\gamma})'(Z_0'Z_0)^{-1}Z_i, \tag{26}$$

and the imbalance in the lagged outcomes and the auxiliary covariates are

$$X_1 - X_{0\cdot}'\hat{\gamma}^{\text{cov}} = \check{X}_1 - \check{X}_{0\cdot}'\hat{\gamma} \quad \text{and} \quad Z_1 - Z_{0\cdot}'\hat{\gamma}^{\text{cov}} = 0. \tag{27}$$

This has several implications. First, the approach outlined above, using $\hat{m}(Z_i) = Z_i'\hat{\eta}^{\text{ols}}$ as

[14] This two-step method induces certain pathologies. For instance, Kaul et al. (2018) point out that SCM applications commonly include *all* lagged outcomes in Z. In this case, covariates in Z other than the lagged outcomes receive zero weight, and the two-step estimator reduces to the simpler version in Equation (3).

the outcome model in ASCM, is equivalent to first residualizing Y against Z, then estimating the treatment effect as $\check{Y}_1 - \check{Y}_0'\hat{\gamma}^{\text{scm}}$, where $\hat{\gamma}^{\text{scm}}$ are the SCM weights that minimize imbalance in X.

Second, this suggests an improvement on this approach that replaces $\hat{\gamma}^{\text{scm}}$ with SCM- or ASCM-style weights that minimize imbalance in $\check{X}$ instead of X. This can be implemented in the form of a partitioned regression: first residualize X and Y against Z, then apply SCM or ASCM to the residuals $\check{X}$ and $\check{Y}$. By Lemma 3, this perfectly balances Z. Furthermore, by achieving better balance on the residuals $\check{X}$, this approach achieves better balance on X than is obtained by separately balancing the raw lagged outcomes and fitting OLS on the auxiliary covariates.[15] In addition, because we can represent this approach as a weighting estimator, we can compute standard errors with the procedure in Section 6.2.

Finally, the partitioned regression procedure as outlined requires Z_0 to be full rank. If it is not (e.g., if $d > N_0$), we can still follow the partitioned regression procedure, though the residuals will only be approximately orthogonal to the auxiliary covariates, and Equation (27) will not hold. In practice we suggest following the dimension reduction approaches in the following section, and do this in the numerical illustration in Section 8.

7.2 Dimension reduction

While Abadie et al. (2010) suggest including (functions of) the lagged outcomes in Z, there is little guidance on how to do so in practice. Consistent with the literature on balancing covariates in high dimensions (e.g., Ning et al., 2017), we argue that a principled approach is to choose functions that best approximate the conditional expectation of $Y_i(0)$, $m(X_i)$.

This can take several forms. First, we can manually select summary statistics based on an assumed functional form for $m(X_i)$. Abadie et al. (2010) argue for this under an autoregressive model; if Y_{it} follows an $AR(p)$ model, then exactly balancing the most recent p entries is sufficient for unbiasedness. Doudchenko and Imbens (2017) consider the case where Y_{it} includes unit-specific intercepts; here, balancing the pre-treatment average outcome $\bar{X}_i \equiv \frac{1}{T_0}\sum_{t=1}^{T_0} X_{it}$ is sufficient to capture the available information about the underlying factors.[16] Richer models would point to additional summary statistics, such as the pre-treatment trend or, in a general factor model, estimates of the factor loadings from a singular value decomposition $X = \hat{\Phi}\hat{\mu}'$; see Gobillon and Magnac (2016) for similar suggestions.

Alternatively, we can directly model the prognostic score (Hansen, 2008). For example, we could use the ridge regression model in Section 3.2 to estimate $\hat{Y}_i(0)$ for each unit, then fit SCM using

[15] In addition, under the linear factor model with covariates considered by Abadie et al. (2010) and Xu (2017), the bias bound in Theorem 1 will still hold because $\hat{\gamma}^{\text{cov}}$ perfectly balances the covariates. Note that while this partitioned ASCM approach always exactly balances the auxiliary covariates, whether it achieves better balance on the original X than SCM alone, without auxiliary covariates, is data dependent. We thus propose this approach as a sensible default, and as with SCM and ASCM encourage careful balance checking.

[16] Doudchenko and Imbens (2017) and Ferman and Pinto (2018) propose applying SCM to de-meaned data $X_{it} - \bar{X}_i$. This can be seen as a special case of our partitioned regression proposal.

this as the only covariate. Or, analogous to the V matrix in Abadie et al. (2010), we could weight the covariates by their relative predictive power for $Y_i(0)$.[17] Hazlett and Xu (2018) also suggest a promising kernel-based approach in the context of SCM, first proposing a kernel representation of the lagged outcomes and then balancing a low-rank approximation to the kernel.

We evaluate specific implementations of dimension reduction in our simulations in Section 8. Simulation evidence indicates that SCM using a well-chosen dimension reduction can improve over SCM using the full vector of pre-treatment outcomes, but that ASCM using the full covariate vector and a flexible outcome model generally improves on either.

8 Simulations and numerical illustrations

We now turn to a numerical illustration and simulations. We focus on two prominent SCM examples: the impact of Proposition 99 on cigarette consumption in California (Abadie et al., 2010) and the impact of terrorism on the Basque economy (Abadie and Gardeazabal, 2003). We describe the Prop. 99 setting below and describe the Basque example in the appendix. We generate simulation studies designed to mimic each application. Overall, we find that SCM is badly biased across these simulations and that there can be substantial gains from ASCM and related approaches. We also use these simulations to show that standard uniform permutation tests are generally invalid and to evaluate coverage of the variance estimator in Section 6.2.

8.1 Illustration: California Prop. 99

In their seminal paper, Abadie et al. (2010) apply SCM to the impact of California Proposition 99 on per-capita cigarette sales. Enacted in 1989, Prop. 99 increased cigarette taxes in addition to launching several anti-tobacco projects in the state. The authors find that, for the 1989–2000 period, Prop. 99 reduced annual cigarette sales by roughly 25 percent. A small cottage industry has emerged around re-analyzing this example (e.g., Doudchenko and Imbens, 2017; Hazlett and Xu, 2018; Hsiao et al., 2018).

Figures 2 and 3 show the results of our re-analysis of the California data. Figure 2 shows gap plots of the ATT estimates, plus or minus two standard errors calculated using our model-based proposal in Section 6.2. We consider three estimators: (1) SCM alone,[18] (2) ridge-augmented SCM, and (3) the partitioned strategy described in Section 7, using an OLS regression on four auxiliary covariates and the pre-treatment average as the outcome model and balancing the residuals with

[17] This is the L^2 analog of the common "covariate screening" approach for high-dimensional covariates. For example, Ning et al. (2017) use a Lasso regression of Y on X as a pre-processing step to select important covariates and then balance that selected subset. See also Belloni et al. (2014).

[18] Note that the SCM estimator used here balances all the lagged outcomes. The original Abadie et al. (2010) study used four auxiliary covariates (cigarette prices, per capita income, the share of the population aged 15-24, and per capita beer consumption) and three lagged outcomes (1975, 1980, and 1988) in Z (see Section 7). Perhaps surprisingly, we obtain nearly identical estimates using all the lagged outcomes and no auxiliary covariates.

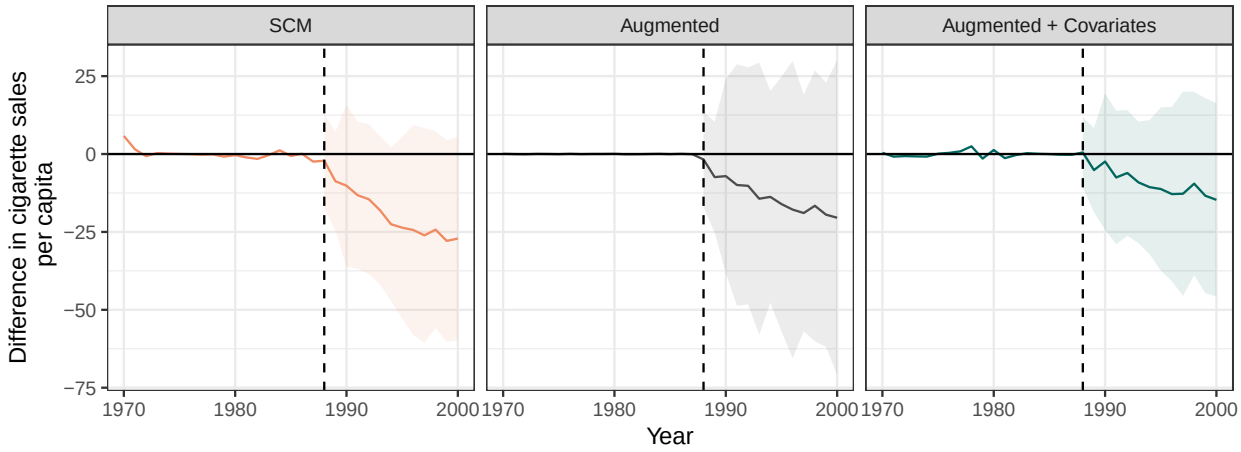


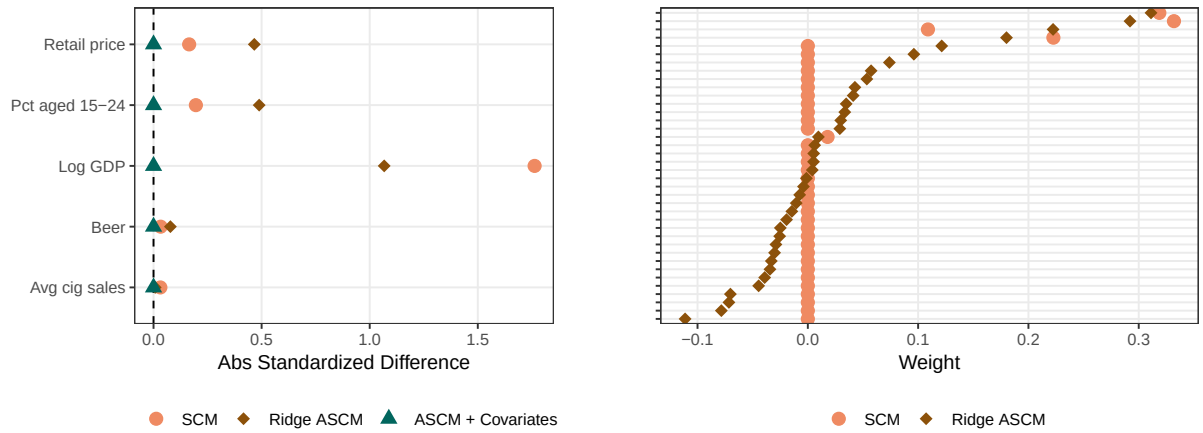
Figure 2: Point estimates ± two standard errors of the ATT for the California Prop. 99 example (Abadie et al., 2010) using SCM, ridge ASCM, and ASCM with covariates. Note that ridge ASCM achieves nearly perfect balance for the pre-period outcomes.

SCM. Figure 3a shows the balance across auxiliary covariates and, in the last row, across pre-treatment average outcomes. Recall that by Lemma 3, the third estimator achieves exact balance on the auxiliary covariates. Figure 3b shows the donor unit weights for SCM and ridge ASCM.

The different estimators lead to somewhat similar stories. Our SCM estimate shows relatively weak evidence for a large, negative effect of Prop. 99 on cigarette sales, of around 26 packs per capita in 1997, close to the estimate in Abadie et al. (2010) of 24 packs per capita. By contrast, the ridge ASCM estimate shows a weaker effect of 20 packs per capita, with wider standard errors. The covariate-adjusted ASCM has similar-sized standard errors to SCM but an estimated effect of around 13 packs per capita, roughly half the size of the SCM estimate but close to the estimated 12 packs per capita in the original Prop. 99 study (Glantz and Balbach, 2000).

We can use the tools we have developed to understand these differences. The balance in the lagged outcomes and auxiliary covariates, while good, is not exact for SCM, which suggests that the estimates are still likely biased. The augmented estimators substantially improve balance: the baseline ridge augmented estimator achieves nearly perfect balance of the lagged outcomes, but leaves some imbalance of other covariates, and the covariate-adjusted ASCM estimator achieves perfect balance on the auxiliary covariates and roughly half the imbalance on pre-treatment outcomes as SCM. This suggests that bias will be smaller for the augmented estimators, especially the covariate-adjusted estimator, than for SCM alone.

Finally, as anticipated, the weights for ridge ASCM are more variable than the sparse, non-negative standard SCM weights—the L^2 norm of the ridge ASCM weights is roughly 12 percent larger than that of the SCM weights—and many of the weights are negative. Thus, ridge ASCM is extrapolating outside of the convex hull in order to achieve nearly perfect balance on the pre-



(a) (b)

Figure 3: (a) Covariate balance for SCM, ridge ASCM, and ASCM with covariates. (b) Donor unit weights for SCM and ridge ASCM.

treatment outcomes. As a consequence, the estimate relies more heavily on global linearity than SCM alone. In addition, these more variable weights increase sampling variability. Finally, we note that the standard error estimates assume homoskedasticity across units, but not across time periods, and so the standard error widen and tighten at different time points.

8.2 Calibrated simulation studies

We now turn to simulation studies in which the true data generating process is known. To make our simulation studies as realistic as possible, we conduct "calibrated" simulation studies (Kern et al., 2016), based on estimates of linear factor models fit to the California ($N = 38$, $T_0 = 18$) and Basque ($N = 17$, $T_0 = 20$) examples. We use the Generalized Synthetic Control Method (Xu, 2017) to estimate factor models with four latent factors for each application. We then simulate outcomes using the distribution of estimated parameters. We model selection into treatment as a function of the latent factors. Additional details are in Appendix A.6. We also present results from two additional DGPs, each calibrated to estimates from the Basque data: a unit and time fixed effects model, and an auto-regressive model with 4 lags.

8.2.1 Simulation results: estimation

SCM and ridge ASCM First, we explore the role of augmentation using simple outcome models. For each DGP, we consider five estimators: (1) SCM alone, (2) ridge regression alone, (3) ridge ASCM, (4) fixed effects alone, and (5) SCM augmented with fixed effects. Figure 4 shows the main results. Each panel displays the average absolute bias for all five estimators across all simulations. In addition, all panels show the absolute bias conditional on the simulation yielding good balance,

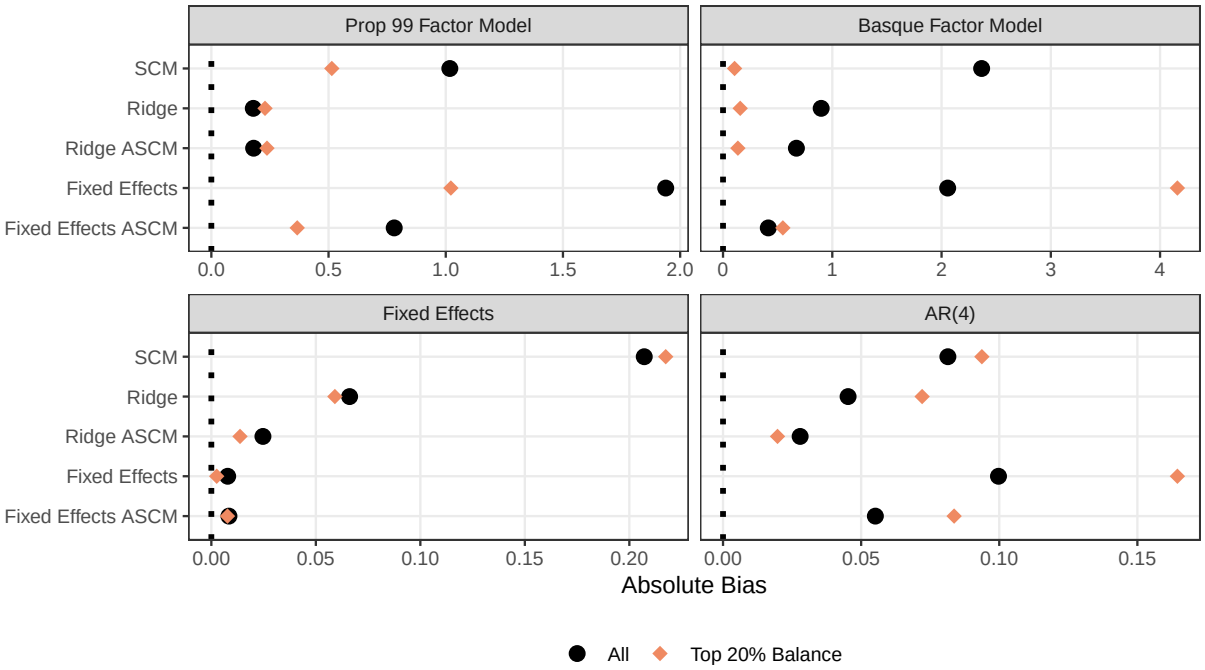


Figure 4: Overall bias and bias conditional on the simulation being in the highest quintile by covariate balance for (a) the Prop. 99 simulation, (b) the Basque simulation, (c) the fixed effects simulation, and (d) the AR simulation. In panels (a) and (b), the outcome is scaled such that the residual standard deviation is one.

defined as being in the bottom quintile of the distance from X_1 to the convex hull of X_0 (i.e., of the SCM imbalance). Appendix Figure C.4 shows the estimator RMSE for the same simulations.

There are several takeaways. First, SCM is badly biased for all four DGPs when we consider all simulation draws. Restricting to simulation draws with good balance in the calibrated factor model simulations, however, dramatically reduces the bias in the factor model simulations, especially that based on the Basque data. For the fixed effects and AR(4) simulations, unbiased estimation requires balancing the pre-treatment average and last 4 pre-treatment time periods, respectively. Thus, conditioning on low RMSE across all periods does not improve the bias of SCM under these DGPs.

Second, the simple ridge outcome regression substantially reduces bias relative to SCM in all four simulations — by more than 75 percent in the fixed effects and California simulations. The fixed effects outcome model does quite well when it matches the DGP, but poorly in the other three simulations. Third, augmented SCM always reduces bias relative to SCM alone and generally reduces bias relative to outcome modeling alone. When the outcome model alone performs well, as ridge does in the Prop. 99 simulation, there is little room for gains from ASCM. However, when the outcome model has larger bias, as with ridge regression for the fixed effects and AR simulations

and the fixed effects estimator for all non-fixed effects simulations, ASCM significantly reduces the bias. Even when the outcome model is *more biased* than SCM, augmenting SCM with it reduces bias relative to SCM, suggesting an interesting robustness property. Fourth, ridge ASCM is less sensitive to baseline imbalance in the simulation than is SCM. When SCM has poor balance, ridge ASCM can extrapolate outside of the convex hull of the control units, and when SCM has good balance so does ridge ASCM. Finally, Appendix Figure C.4 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.

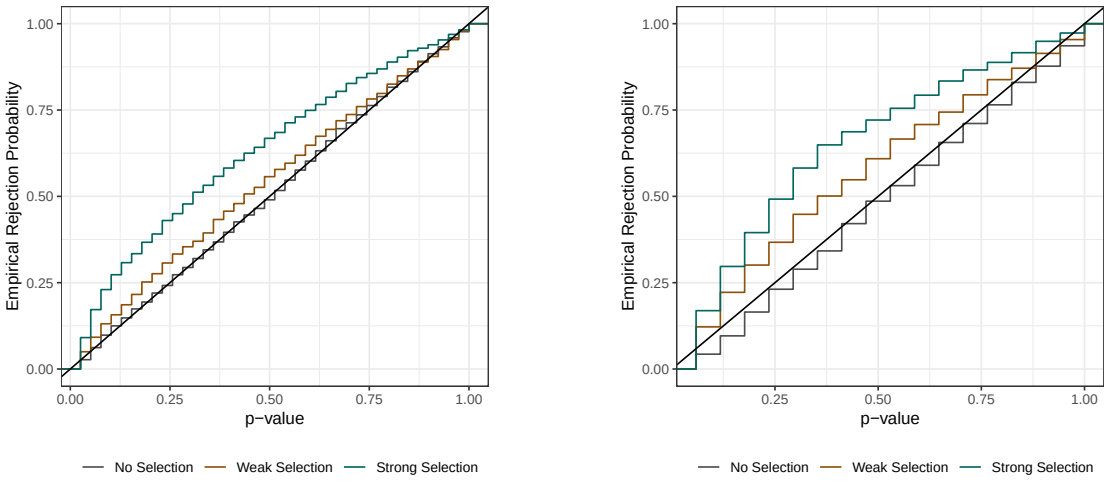
Machine learning models In Appendix C we evaluate alternative outcome models for use in ASCM. For each DGP we consider SCM augmented with (1) a random forest, (2) a two layer feed-forward neural network, (3) matrix completion using `MCPanel` (Athey et al., 2017), (4) Lasso, and (5) `CausalImpact` (Brodersen et al., 2015). We compare ASCM to the pure outcome models as well as pure SCM. We find wide variation across models, with a random forest model and a two layer feed-forward neural network performing relatively poorly and Lasso and `CausalImpact` (Brodersen et al., 2015) doing better. In all cases, augmenting SCM with the outcome model reduces the bias in SCM and generally improves on the pure outcome model as well. In particular, even augmenting SCM with a flexible but poorly performing machine learning method results in bias reduction relative to SCM alone.

Dimension reduction Appendix Figure C.7 shows the performance of combining dimension reduction with standard SCM. In general, fitting SCM using either (1) unit long-run averages or (2) the prognostic score fit via ridge regression has much lower bias than fitting SCM on all lagged outcomes, though much more bias than the outcome modeling and ASCM estimates shown in Figure 4.

8.2.2 Simulation results: testing and inference

Testing. Using the Basque and California calibrated simulations, Figure 5 shows the empirical distribution for the p-values from uniform permutation tests, p_{unif} as defined in Appendix A.5, with a placebo gap test statistic, $\left|W_i(Y_i^{\text{obs}} - \widetilde{Y}_i)\right|$, where $\widetilde{Y}_i$ is the leave-one-out synthetic control for Y_i, $\sum_{j \neq i} \widetilde{\gamma}_j^{(i)} Y_j^{\text{obs}}$. Each plot shows three simulations with varying levels of selection: zero, weak, and strong selection. The pattern is striking. While the uniform permutation p-values are valid under zero selection — when the propensity scores are constant across units — they are clearly invalid when this does not hold. This pattern holds across a large number of simulation settings.[19]

[19] An interesting feature of our simulations is that the empirical Type I error rate differed substantially by test statistic. We obtained better performance with the RMSPE test statistic proposed by Abadie et al. (2010). We leave a thorough investigation to future work.



(a) California

(b) Basque

Figure 5: Uniform permutation test is invalid with selection.

Inference. Finally, we evaluate the model-based inference method described in Section 6.2 using the Prop. 99 simulation. Appendix Table 1 shows coverage results for the leave-one-out estimate of the standard error as well as for the bootstrap. As anticipated, we find that our proposed variance estimator is conservative, at least in settings with small bias, and that over-coverage is more dramatic when the placebo weights have worse balance. By contrast, the bootstrap badly under-covers across simulation settings.

9 Discussion

SCM remains a widely popular approach for estimating causal effects in policy settings. Despite this popularity, however, SCM remains poorly understood. It also has a major limitation: in typical applications, SCM fails to exactly balance pre-treatment outcomes, creating bias. We address this by proposing the Augmented SCM estimator, which combines SCM and outcome modeling. By extrapolating outside the convex hull of the control units, ASCM achieves better balance and reduces bias, albeit at the price of additional model dependence. ASCM naturally accommodates additional covariates when available to further reduce bias.

We also offer a new perspective on SCM — that SCM is a form of inverse propensity score weighting — which clarifies some open issues. Many of the ideas we present build off suggestions in the original SCM papers; our new perspective on SCM will hopefully help make these ideas more accessible. We also use this perspective to argue that, at least from a randomization-based perspective, the commonly used uniform permutation test is invalid. We instead propose a model-based inference approach for both SCM and ASCM that works well in simulations.

While Augmented SCM is promising, the SCM problem is fundamentally challenging and there is no silver bullet. As with any analysis, researchers should assess sensitivity to modeling choices and should be restrained in interpreting results. In particular, ASCM is vulnerable to departures from the outcome model, so sensitivity to this should be carefully explored.

There are several directions for future work. Perhaps most important is to develop comprehensive diagnostics and other model checks for Augmented SCM and, related, to find coherent and valid design-based methods for testing and inference in typical SCM settings. This is difficult since any valid approach must successfully handle both small sample sizes and very noisy propensity score estimates. While our model-based procedure is straightforward, valid design-based inference would be particularly attractive.

A second area for future inquiry is the application of the ASCM framework to settings with multiple treated units. When all treated units are treated at the same time, the IPW perspective suggests that averaging the treated units is a natural strategy. By contrast, recent proposals to fit SCM separately for each treated unit (e.g., Dube and Zipperer, 2015) correspond to fitting a separate propensity score for each treated unit, which can be unnatural. The situation is more complicated with staggered adoption, when units take up the treatment at different times (e.g., Toulis and Shaikh, 2018; Imai et al., 2018). This is an important direction for future work.

A third potential extension is to 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) or hierarchical data structures with outcome information at both the individual and aggregate level (e.g., students within schools).

Finally, many of the ideas we develop for the panel data setting can be extended to more traditional cross-sectional observational studies. As we note in Section 5.3, in some settings, minimizing an L^2 imbalance will be more attractive than minimizing the standard L^∞ imbalance (Athey et al., 2018; Tan, 2017; Ning et al., 2017).

In the end, SCM remains a powerful tool for researchers studying interventions at the aggregate level, and we believe that Augmented SCM helps to mitigate some of SCM's shortcomings. This remains an inherently challenging problem — in many cases, we are trying to impute the missing potential outcome for a single unit — and developing valid statistical methods for this setting is an important open question.

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. 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 (2008). On the failure of the bootstrap for matching estimators. *Econometrica 76*(6), 1537–1557.

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.

Ando, M. and F. Sävje (2013). Hypothesis testing with the synthetic control method.

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: De-Biased Inference of Average Treatment Effects in High Dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Ben-Michael, E., A. Feller, and J. Rothstein (2018). The equivalence between balancing and propensity score weights.

Botosaru, I. and B. Ferman (2017). On the role of covariates in the synthetic control method.

Branson, Z. and M.-A. Bind (2018). Randomization-based inference for bernoulli trial experiments and implications for observational studies. *Statistical methods in medical research*, 0962280218756689.

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

Brodersen, K. H., F. Gallusser, J. Koehler, N. Remy, and S. L. Scott (2015). Inferring Causal Impact using Bayesian Structural Time-Series Models. *The Annals of Applied Statistics 9*(1), 247–274.

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.

Chan, K. C. G., S. C. P. Yam, and Z. Zhang (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(3), 673–700.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2017). An exact and robust conformal inference method for counterfactual and synthetic controls. *arXiv preprint arXiv:1712.09089*.

D'Amour, A., P. Ding, A. Feller, L. Lei, and J. Sekhon (2017). Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*.

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

Dube, A. and B. Zipperer (2015). Pooling multiple case studies using synthetic controls: An application to minimum wage policies.

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

Firpo, S. and V. Possebom (2017). Synthetic Control Method: Inference , Sensitivity Analysis and Confidence Sets.

Fuller, W. A. (2002). Regression estimation for survey samples (with discussion). *Survey Methodology 28*(1), 5–23.

Funk, M. J., D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian (2011). Doubly robust estimation of causal effects. *American journal of epidemiology 173*(7), 761–767.

Glantz, S. A. and E. D. Balbach (2000). *Tobacco War: Inside the California Battles*. Berkeley: University of California Press.

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

Graham, B. S., C. C. de Xavier Pinto, and D. Egel (2012). Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies 79*(3), 1053–1079.

Hahn, J. and R. Shi (2017). Synthetic control and inference. *Econometrics 5*(4), 52.

Hainmueller, J. (2011). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis 20*, 25–46.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika 95*(2), 481–488.

Hazlett, C. and Y. Xu (2018). Trajectory balancing: A general reweighting approach to causal inference with time-series cross-sectional data.

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

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

Imai, K., I. S. Kim, and E. Wang (2018). Matching Methods for Causal Inference with Time-Series Cross-Section Data.

Imai, K. and M. Ratkovic (2013). Covariate balancing propensity score. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 76*(1), 243–263.

Kaul, A., S. Klößner, and G. Pfeifer (2018). Synthetic Control Methods: Never Use All Pre-Intervention Outcomes Together With Covariates.

Kern, H. L., E. A. Stuart, J. Hill, and D. P. Green (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness 9*(1), 103–127.

King, G. and L. Zeng (2001). Logistic Regression in Rare Events Data. *Political Analysis 9*(02), 137–163.

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

Kreif, N., R. Grieve, D. Hangartner, A. J. Turner, S. Nikolova, and M. Sutton (2016). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health economics 25*(12), 1514–1528.

Li, F., K. L. Morgan, and A. M. Zaslavsky (2017). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 1–11.

Lohr, S. L. (2007). Struggles with survey weighting and regression modeling. comment. *Statistical science* (2), 175–178.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models*, Volume 37. CRC press.

Minard, S. and G. R. Waddell (2018). Dispersion-weighted synthetic controls.

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

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society*, 1417–1426.

Ning, Y., S. Peng, and K. Imai (2017). High dimensional propensity score estimation via covariate balancing.

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

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. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science 22*(4), 544–559.

Robins, J. M. and Y. Ritov (1997). Toward a curse of dimensionality appropriate(coda) asymptotic theory for semi-parametric models. *Statistics in medicine 16*(3), 285–319.

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.

Rosenbaum, P. R. (1987). Model-Based Direct Adjustment. *Journal of the American Statistical Association 82*(398), 387–394.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688.

Rubin, D. B. (1980). Comment on "randomization analysis of experimental data: The fisher randomization test". *Journal of the American Statistical Association 75*(371), 591–593.

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

Tan, Z. (2018). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *arXiv preprint arXiv:1801.09817*.

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

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

Wan, S.-K., Y. Xie, and C. Hsiao (2018). Panel data approach vs synthetic control method. *Economics Letters 164*, 121–123.

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

Wu, C. and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association 96*(453), 185–193.

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.

A Appendix

A.1 Ridge regression as a weighting estimator

Just as we can write ridge ASCM as a weighting estimator, we can write ridge regression alone as a weighting estimator.

Lemma 4. With $\hat{\eta}_0^{\text{ridge}}$ and $\hat{\eta}^{\text{ridge}}$, the solution to the ridge regression of post-treatment outcomes on pre-treatment outcomes for control units, the ridge estimate can be written as a weighting estimator:

$$\hat{Y}_1^{\text{ridge}}(0) = \hat{\eta}_0^{\text{ridge}} + \hat{\eta}^{\text{ridge}\prime} X_1 = \sum_{W_i=0} \hat{\gamma}_i^{\text{ridge}} Y_i, \tag{28}$$

where

$$\hat{\gamma}_i^{\text{ridge}} = \frac{1}{N_0} + X_1'(X_{0\cdot}'X_{0\cdot} + \lambda I_{T_0})^{-1} X_i. \tag{29}$$

We can also characterize the imbalance and L^2 norm of the ridge regression weights:

Lemma 5. Let $\sigma_{\min}^2$ be the minimum eigenvalue of the sample covariance matrix $\frac{1}{N_0} X_{0\cdot}' X_{0\cdot}$. Then with $\kappa = \frac{\lambda}{N_0 \sigma_{\min}^2} + 1 > 1$, the ridge weights $\hat{\gamma}^{\text{ridge}}$ satisfy:

$$\|X_1 - X_{0\cdot}'\hat{\gamma}^{\text{ridge}}\|_2 \leq \frac{\kappa - 1}{\kappa} \quad \left\| X_1 - \frac{1}{N_0} X_{0\cdot}' \mathbf{1} \right\|_2 \tag{30}$$

and

$$\|\hat{\gamma}^{\text{ridge}}\|_2 \leq \frac{1}{\sqrt{N_0}} + \frac{1}{\kappa \sigma_{\min} \sqrt{N_0}} \left\| X_1 - \frac{1}{N_0} X_{0\cdot}' \mathbf{1} \right\|_2 \tag{31}$$

Next, we can use Lemmas 4 and 5 with Theorem 1 to characterize the bias of ridge regression under a linear factor model. To the best of our knowledge this is a novel result.

Corollary 2. Let $\sigma_{\min}^2$ be the minimum eigenvalue of the sample covariance matrix $\frac{1}{N_0} X_{0\cdot}' X_{0\cdot}$. Then with $\kappa = \frac{\lambda}{N_0 \sigma_{\min}^2} + 1 > 1$, the ridge regression weights (29) exist and under the linear factor model (10) the bias satisfies the bound

$$\begin{aligned} \left| \mu_T \cdot \left(\phi_1 - \sum_{W_i=0} \phi_i \hat{\gamma}_i^{ridge} \right) \right| \leq \frac{M\sqrt{J}}{\sqrt{T_0}} \left(\frac{\kappa - 1}{\kappa} + \frac{8\sigma M \sqrt{J}}{\sqrt{N_c} \sigma_{\min} \kappa} \right) \left\| X_{1\cdot} - \frac{1}{N_0} X_{0\cdot}' \mathbf{1} \right\|_2 \\ + 8\sigma M^2 J \sqrt{\frac{\log 5}{T_0}} \left(\frac{1}{\sqrt{N_1}} + \frac{1}{\sqrt{N_0}} \right) \end{aligned} \tag{32}$$

with probability at least $1 - 2e^{-8J \log 5} > 0.99$.

Finally, we note that ridge regression weights are a special case of the elastic-net synthetic control estimator proposed by Doudchenko and Imbens (2017) and correspond to an L^2 penalized version of the Oaxaca-Blinder weights considered by Kline (2011). Furthermore, these weights fit into the general balance weights formulation (19) with an L^2 dispersion penalty $f(\gamma_i) = \frac{1}{2}\gamma_i^2$ and an L^2 balance criterion $h_\lambda(x) = \frac{1}{2\lambda}\|x\|_2^2$.

Lemma 6. The ridge weights (29) are the solution to the general balancing weights problem (19) with an L^2 balance criterion and dispersion function, i.e. $\hat{\gamma}^{ridge}$ is the solution to

$$\min_{\gamma|\mathbf{1}'\gamma_i=1} \frac{1}{2\lambda}\|X_{1\cdot} - \gamma' X_{0\cdot}\|_2^2 + \frac{1}{2}\|\gamma\|_2^2 \tag{33}$$

A.2 Derivation of the general dual problem

We now derive the Lagrangian dual for the general balancing weights problem (19). We can augment the optimization problem with auxiliary variables ϵ, yielding:

$$\begin{aligned} \min_{\gamma,\epsilon} \ & h_\zeta(\epsilon) + \sum_{W_i=0} f(\gamma_i). \\ \text{subject to } & \epsilon = X_{1\cdot} - X_{0\cdot}'\gamma \\ & \sum_{W_i=0} \gamma_i = 1 \end{aligned} \tag{34}$$

The Lagrangian is

$$\mathcal{L}(\gamma, \epsilon, \alpha, \beta) = \sum_{i|W_i=0} f(\gamma_i) + \alpha(1-\gamma_i) + h_\zeta(\epsilon) + \beta'(X_{1\cdot} - X_{0\cdot}'\gamma - \epsilon). \tag{35}$$

The dual maximizes the objective

$$\begin{aligned} q(\alpha,\beta) &= \min_{\gamma,\epsilon} \mathcal{L}(\gamma,\epsilon,\alpha,\beta) \\ &= \sum_{W_i=0} \min_{\gamma_i}\{f(\gamma_i) - (\alpha + \beta' X_i)\gamma_i\} + \min_{\epsilon}\{h_\zeta(\epsilon) - \beta'\epsilon\} + \alpha + \beta' X_{1\cdot} \\ &= -\sum_{W_i=0} f^*(\alpha + \beta' X_i) + \alpha + \beta' X_{1\cdot}' - h_\zeta^*(\beta), \end{aligned} \tag{36}$$

where a convex, differentiable function g has convex conjugate $g^*(y) \equiv \sup_{x\in\text{dom}(g)}\{y'x - g(x)\}$. By strong duality the general dual problem (20), which minimizes $-q(\alpha,\beta)$, is equivalent to the primal balancing weights problem. Given the $\hat{\alpha}$ and $\hat{\beta}$ that minimize the Lagrangian dual objective, $-q(\alpha,\beta)$, we recover the donor weights solution to (19) as

$$\hat{\gamma}_i = f^{*\prime}(\hat{\alpha} + \hat{\beta}' X_i), \tag{37}$$

The convex congjugate of the entropy function is $f^*(x) = \exp(x-1)$, and the convex conjugate of the exact balancing constraint is uniformly zero. Thus, in the special case of entropy penalized SCM with exact balance (15) the specialization of the general dual problem (20) is

$$\min_{\alpha,\beta} \sum_{W_i=0} \exp(\alpha + \beta' X_{i\cdot} - 1) - (\alpha + \beta' X_{1\cdot} - 1) \tag{38}$$

By redefining α to be $\alpha - 1$, we get the dual form (16).

The convex conjugate of $h_\zeta(x) = \frac{1}{2\zeta}\|x\|_2^2$ is $h_\zeta^*(x) = \frac{\zeta}{2}\|x\|_2^2$ so following the general dual problem

(20) the dual problem to entropy penalized SCM is

$$\min_{\alpha,\beta} \sum_{W_i=0} \exp(\alpha + \beta' X_{i\cdot} - 1) - (\alpha + \beta' X_{1\cdot} - 1) + \frac{\zeta}{2}\|\beta\|_2^2. \tag{39}$$

We again absorb the constant subtraction term into the intercept and redefine α to be $\alpha - 1$, which gives the form of the dual problem presented in the main text (18).

A.3 With large T_0, SCM is IPW on factor loadings with ridge penalty

One limitation of Theorem 1 is that it is an *upper bound* on the bias. However, the perspective of SCM as approximate balancing weights and its dual view as IPW provide insight into existing results in the literature that show that without exact balance, SCM will indeed be biased even with infinite pre-treatment time periods. In particular, Ferman and Pinto (2018) show that SCM will be biased by considering the behavior of the un-penalized SCM optimization problem (3) as $T_0 \to \infty$. Under a more general set of assumptions, Ferman and Pinto (2018) show that the SCM weights converge to the minimizer of the asymptotic MSE. Under the linear factor model (10) and our set of assumptions, this implies that the solution to the unpenalized SCM problem (3) converges to $\tilde{\gamma}$, where

$$\tilde{\gamma} = \operatorname*{argmin}_{\gamma \in \Delta^{n-1}} \frac{1}{2\sigma^2} \left\|\phi_{1\cdot} - \phi_{c\cdot}'\gamma\right\|_2^2 + \frac{1}{2}\|\gamma\|_2^2. \tag{40}$$

Notice that the un-penalized SCM problem (3) that approximately balances the lagged outcomes converges to a *penalized* SCM problem that approximately balances the factor loadings, where the noise level σ^2 controls the level of penalization. The dual formulation of (40) fits a propensity score with linear odds and a ridge penalty:

$$\min_{\alpha,\beta} \frac{1}{2} \sum_{W_i=0} \left(\max\{0, \alpha + \beta'\phi_i'\}\right)^2 - (\alpha + \beta'\phi_1) + \frac{\sigma^2}{2}\|\beta\|_2^2 \tag{41}$$

Even as the number of pre-treatment periods goes to infinity, the estimator is still regularized: with larger noise variance σ^2, the dual propensity score parameters shrink more towards 0 and thus the weights shrink towards uniformity. Notice that this is exactly the same phenomenon as OLS with measurement error in the covariates. Therefore, as Ferman and Pinto (2018) also conclude, even asymptotically SCM will not balance the factor loadings and thus the SCM estimator will be biased asymptotically in T_0, except in degenerate cases.

A.4 Auxiliary Covariates

A.4.1 Auxiliary covariates and a reweighted L^2 norm

The original SCM formulation (Abadie et al., 2010, 2015) includes covariates $Z_i \in \mathbb{R}^d$, possibly including (some function of) the lagged outcomes, and finds weights $\hat{\gamma}(V)$ that minimize the imbalance in the covariates measured by a V-weighted L^2 norm, for a symmetric and positive definite matrix $V \in \mathbb{R}^{d\times d}$:

$$\min_{\gamma(V)\in\Delta^{N_c-1}} (Z_1 - Z_c^\intercal \gamma(V))^\intercal V (Z_1 - Z_0^\intercal \gamma(V)) \tag{42}$$

where the rows of $Z_c \in \mathbb{R}^{N_c \times d}$ are the augmented covariates Z_i for the control units. V is typically restricted to be diagonal.

Thus, rather than balancing lagged outcomes, this version of SCM minimizes imbalance across the auxiliary covariates. Abadie et al. (2010) suggest choosing V to minimize the L^2 imbalance for lagged outcomes X, when the synthetic control is formed by $\hat{\gamma}(V)$:

$$\min_{V \text{ symmetric P.D.}} (X_{1\cdot} - X_{0\cdot}'\hat{\gamma}(V))^\intercal (X_{1\cdot} - X_{0\cdot}^\intercal \hat{\gamma}(V)) \tag{43}$$

This results in a two step procedure that searches over possible values of V, solving (42) each time and choosing the V which minimizes (43). The V matrix prioritizes balance in some variables over others. In our framework the V matrix defines the balance criterion and thus the type of regularization. To see this, consider the dual (20) to the entropy-penalized version of (42)

$$\min_{\alpha,\beta} \sum_{i|W_i=0} \exp(\alpha + \beta' Z_i) - (\alpha + \beta' Z_1) + \frac{\zeta}{2} \|V^{-\frac{1}{2}}\beta\|_2^2. \tag{44}$$

This dual problem is instructive: the original SCM fits a penalized propensity score model using Z as covariates and with a ridge penalty, reweighted by V, where V is chosen to optimize the pre-intervention outcome imbalance obtained by the estimated propensity score. This procedure does not guarantee balance on the auxiliary covariates, only using them as a proxy for the lagged outcomes. See Kaul et al. (2018) for further discussion.

A.4.2 Balancing auxiliary covariates directly

An alternative to incorporating covariates through an outcome model (as we propose in Section 7) is to explicitly balance both the auxiliary covariates and the lagged outcomes. We separate the balance criterion into two pieces, one for the lagged outcomes X_i and one for the auxiliary covariates Z_i. For example, we can use separate L^2 balance criteria:

$$\min_{\gamma} \quad \frac{1}{2\zeta_X} \|X_{1\cdot} - X_{0\cdot}'\gamma\|_2^2 + \frac{1}{2\zeta_Z} \|Z_{1\cdot} - Z_{c\cdot}'\gamma\|_2^2 + \sum_{W_i=0} \gamma_i \log \gamma_i, \tag{45}$$

where the two hyper-parameters ζ_X and ζ_Z control the trade off between prioritizing balance in the lagged outcomes and auxiliary covariates. The Lagrangian dual (20) estimates a propensity score using both the lagged outcomes and the covariates, with separate ridge penalties:

$$\min_{\alpha,\beta_X,\beta_Z} \sum_{W_i=0} \exp(\alpha + \beta_X' X_i + \beta_Z' Z_i') - (\alpha + \beta_X' X_1 + \beta_Z' Z_1) + \frac{\zeta_X}{2} \|\theta_X\|_2^2 + \frac{\zeta_Z}{2} \|\beta_Z\|_2^2 \tag{46}$$

We can follow a similar procedure in the outcome model by also separately regularizing the lagged outcomes and auxiliary covariates.

A.5 Weighted randomization tests with a single treated unit

We now consider the permutation test from a randomization-based perspective. First, we outline the necessary conditions for a valid randomization test with varying, known propensity scores (Rosenbaum, 1987; Branson and Bind, 2018; Toulis and Shaikh, 2018), then we show the uniform

permutation test of Abadie et al. (2010, 2015) differs from this ideal case. We then briefly discuss a weighted randomization with a single treated unit, which we use in the simulation studies.

Following Firpo and Possebom (2017), we consider the randomization-based test for the sharp null hypothesis of no treatment effect whatsoever:

$$H_0 : Y_i(1) = Y_i(0), \quad \forall i, \tag{47}$$

where we again assume that both potential outcomes are defined for all units. Consider a test statistic, $S(\mathbf{W}, \mathbf{Y}(0), \mathbf{Y}(1))$, where $\mathbf{W}$ is the vector of treatment assignment indicators, $W_1, \ldots, W_N$, and $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ are the vectors of potential outcomes. Under the sharp null in Equation (47), $\mathbf{Y}(0) = \mathbf{Y}(1) = \mathbf{Y}^{\text{obs}}$, so we can simplify the test statistic to $S(\mathbf{W}, \mathbf{Y}^{\text{obs}})$. Example test statistics include the simple difference-in-means, $S(\mathbf{W}, \mathbf{Y}^{\text{obs}}) = \frac{1}{N_1} \sum W_i Y_i^{\text{obs}} - \frac{1}{N_0} \sum (1 - W_i) Y_i^{\text{obs}}$, as well as the *placebo gap*, $S(\mathbf{W}, \mathbf{Y}^{\text{obs}}) = \frac{1}{N_1} \sum_{W_i=1} Y_i^{\text{obs}} - \widetilde{Y}_i$, where $\widetilde{Y}_i$ is the leave-one-out SCM estimate of Y_i, $\sum_{j \neq i} \widetilde{\gamma}_j^{(i)} Y_j^{\text{obs}}$. For simplicity, we restrict test statistics to be non-negative, for example by taking the absolute value, which yields one-sided p-values.

For the setting with a known assignment mechanism, we draw the vector of treatment assignments $\mathbf{W}' \sim \Pr(\mathbf{W}|\mathbf{X})$, with observed assignment $\mathbf{W}^{\text{obs}}$ and observed test statistic $S^{\text{obs}} = S(\mathbf{W}^{\text{obs}}, \mathbf{Y}^{\text{obs}})$. Note that this is a departure from type of uncertainty discussed in the main text. The valid (one-sided) p-value is then:

$$p_{\text{wt}} = \mathbb{E}_{\mathbf{W}' \sim \Pr(\mathbf{W}|\mathbf{X})} \left[\mathbb{I} \left\{ S(\mathbf{W}', \mathbf{Y}^{\text{obs}}) \geq S^{\text{obs}} \right\} \right], \tag{48}$$

where $\mathbb{I}$ is the indicator function and where the expectation is taken with respect to the (known) randomization distribution, $\Pr(\mathbf{W}|\mathbf{X})$.

It is well known that the p-value in Equation (48) is valid in any finite sample, in the sense that $\mathbb{P}(p_{\text{wt}} \leq \alpha) \leq \alpha$ for every $\alpha \in [0, 1]$, so long as $\Pr(\mathbf{W}|\mathbf{X})$ is indeed the correct randomization distribution; see, for example Branson and Bind (2018) and Toulis and Shaikh (2018). Conversely, taking the expectation with respect to the *incorrect* randomization distribution may yield an invalid p-value. In particular, the commonly used uniform permutation test calculates:

$$p_{\text{unif}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{S_i \geq S_1\}, \tag{49}$$

where S_i is the test statistic when unit i receives the treatment. This will only be valid when treatment assignment is constant across all units.

To make progress, we impose additional structure on Equation (48). For simplicity, let $\pi_i \equiv \mathbb{P}(W_i = 1)$ be known, independent, unit-level treatment probabilities under a Bernoulli design, $\mathbf{W} \in \{0, 1\}^N$. To reflect standard SCM settings, we further restrict the possible randomizations to draws with a single treated unit, $N_1 = 1$. Following Branson and Bind (2018), the probability that unit i receives treatment is therefore

$$\widetilde{\pi}_i = \mathbb{P}(W_i = 1 | N_1 = 1) = \frac{\pi_i \prod_{j \neq i} (1 - \pi_j)}{\sum_{k=1}^{N} \pi_k \prod_{j \neq k} (1 - \pi_j)}. \tag{50}$$

Continuing the convention that unit 1 is the observed treated unit, $W_1 = 1$, we can therefore

re-write Equation (48) as:

$$
\begin{aligned}
p_{\text{wt}} &= \sum_{i=1}^{N} \mathbb{I}\{S_i \geq S_1\} \; \mathbb{P}(W_i = 1 \mid N_1 = 1) \\
&= \sum_{i=1}^{N} \mathbb{I}\{S_i \geq S_1\} \; \frac{\pi_i \prod_{j \neq i}(1-\pi_j)}{\sum_{k=1}^{N} \pi_k \prod_{j \neq k}(1-\pi_j)} \\
&= \sum_{i=1}^{N} \mathbb{I}\{S_i \geq S_1\} \; \frac{\widetilde{\pi}_i}{\sum \widetilde{\pi}_k}.
\end{aligned}
\tag{51}
$$

where $\widetilde{\pi}_i \equiv \pi_i \prod_{j \neq i}(1-\pi_j)$. We can then use the duality between SCM and IPW to modify p_{wt} to obtain a valid p-value for the SCM setting with one treated unit:

$$
\begin{aligned}
p_{\text{scm}} &= \sum_{i=1}^{N} \mathbb{I}\{S_i \geq S_1\} \; \frac{\frac{\hat{\gamma}_i}{1+\hat{\gamma}_i} \prod_{j \neq i} \frac{1}{1+\hat{\gamma}_i}}{\sum_{k=1}^{N} \frac{\hat{\gamma}_k}{1+\hat{\gamma}_k} \prod_{j \neq k} \frac{1}{1+\hat{\gamma}_j}} \\
&= \sum_{i=1}^{N} \mathbb{I}\{S_i \geq S_1\} \; \frac{\widetilde{\gamma}_i}{\sum \widetilde{\gamma}_k},
\end{aligned}
\tag{52}
$$

where $\hat{\gamma}_i$ are the estimated SCM weights and where $\widetilde{\gamma} = \frac{\hat{\gamma}_i}{1+\hat{\gamma}_i} \prod_{j \neq i} \frac{1}{1+\hat{\gamma}_i}$.If $\hat{\gamma}_i$ are, in fact, the true treatment odds, then p_{scm} will be a valid p-value, analogous to p_{wt}. A practical complication is that unpenalized SCM does not yield a propensity score estimate for the treated unit; we need the implied propensity score link function to have a closed form to estimate $\hat{\gamma}_1$. Thus this procedure is limited to penalized variants of SCM.

A.6 Simulation data generating process

We now describe the simulations in detail. For the Basque example, $N = 17$ and $T_0 = 20$; for the California example, $N = 38$ and $T_0 = 18$. For each, we use the Generalized Synthetic Control Method (Xu, 2017) to fit the following linear factor model, setting $J = 4$:

$$
Y_{it} = \alpha_i + \nu_t + \sum_{j=1}^{J} \phi_{ij} \mu_{jt} + \varepsilon_{it}. \tag{53}
$$

We then use these estimates as the basis for simulating data.Appendix Figure C.3 shows the estimated factors $\hat{\mu}$. We use the estimated time fixed effects $\hat{\nu}$ and factors $\hat{\mu}$ and then simulate data using Equation (53), drawing:

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

where $\hat{\bar{\alpha}}$ and $\hat{\sigma}_\alpha$ are the estimated mean and standard deviation of the unit-fixed effects, $\hat{\Sigma}_\phi$ is the sample covariance of the estimated factor loadings, and $\hat{\sigma}_\varepsilon$ is the estimated residual standard

deviation. We assume a sharp null of zero treatment effect in all DGPs.

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

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

where we set $\theta = 1/2$ unless otherwise noted and re-scale the fixed effects α to have unit variance. Finally, we restrict each simulation to have a single treated unit and therefore normalize the selection probabilities π_i as in Equation (50).

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 to the Basque data 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. Second, we generate data from an AR(4) model:

$$Y_{it}(0) = \rho \sum_{j=1}^{4} Y_{i(t-j)} + \varepsilon_{it},$$

where we set $\rho = \frac{1}{3}$. We model selection as proportional to the last 4 outcomes $\text{logit}^{-1}\pi_i = \theta\left(\sum_{j=1}^{4} Y_{i(T_0-j+1)}\right)$. Following the Basque data example, we set $N = 17$ and $T_0 = 20$.

B Proofs

Lemma 7. Under the linear factor model (10), for weights γ with $\sum_i \gamma_i = 1$, the bias is

$$\mathbb{E}\left[Y_{1T}(0) - \sum_{W_i=0} \gamma_i Y_{iT}\right] = \frac{1}{T_0}\mu_T \cdot \underbrace{\left(\mu' X_1 - \sum_{W_i=0} \gamma_i \mu' X_i\right)}_{\text{imbalance}} - \frac{1}{T_0}\mu_T \cdot \underbrace{\left(\mu'\varepsilon_1 - \sum_{W_i=0} \gamma_i \mu' \varepsilon_i\right)}_{\text{approx. error}}. \quad (54)$$

Proof of Lemma 7. Collecting the factor loadings into a matrix $\Phi \in \mathbb{R}^{N\times J}$, notice that

$$\mu(\Phi_{1\cdot} - \Phi'_{c\cdot}\gamma) = X_{1\cdot} - \varepsilon_1 - (X_{0\cdot} - \varepsilon_c)'\gamma$$

Multiplying both sides by $(\mu'\mu)^{-1}\mu'$ and recalling that $\frac{1}{T_0}\mu'\mu = I$, this gives that

$$\Phi_{1\cdot} - \Phi'_{c\cdot}\gamma = \frac{1}{T_0}\mu'(X_{1\cdot} - X'_{0\cdot}\gamma) - \frac{1}{T_0}\mu'(\varepsilon_1 - \varepsilon'_c\gamma)$$

Therefore the bias is

$$\begin{aligned}\mathbb{E}\left[Y_{1T}(0) - \sum_{W_i=0} \gamma_i Y_{iT}\right] &= \mu_T \cdot \left(\phi_1 - \sum_{W_i=0} \gamma_i \phi_i\right) \\ &= \mu_T \cdot \left(\frac{1}{T_0}\mu'(X_{1\cdot} - X'_{0\cdot}\gamma) - \frac{1}{T_0}\mu'(\varepsilon_1 - \varepsilon'_c\gamma)\right)\end{aligned}$$

□

Lemma 8. For weights γ and ϵ_{it} sub-Gaussian with scale parameter σ and independent,

$$\frac{1}{T_0}\left\|\mu_T \cdot \left(\mu'\varepsilon_1 - \sum_{W_i=0} \gamma_i \mu' \varepsilon_i\right)\right\|_2 \leq 8M^2 J\sigma\sqrt{\frac{\log 5}{T_0}}\,(1 + \|\gamma\|_2) \quad (55)$$

with probability at least $1 - 2e^{-8J\log 5}$.

Proof of Lemma 8. Notice that by Cauchy-Schwarz

$$\frac{1}{T_0}\left\|\mu_T \cdot \left(\mu'\varepsilon_1 - \sum_{W_i=0} \gamma_i \mu' \varepsilon_i\right)\right\|_2 \leq \frac{1}{T_0}\|\mu_T\|_2 \|\mu'(\varepsilon_1 - \varepsilon'_c\gamma)\|_2$$

The jth element of $\mu'(\varepsilon_1 - \varepsilon'_c\gamma)$ can be written as a sum over T_0 sub-Gaussian random variables, $\sum_{t=1}^{T_0} \mu_{jt}(\varepsilon_{1t} - \varepsilon'_{ct}\gamma)$. Using the fact that $|\mu_{jt}| \leq M$, each element of the sum has scale parameter $M\sigma\sqrt{1 + \|\gamma\|_2^2} \leq M\sigma(1 + \|\gamma\|_2)$, and so the sum is sub-Gaussian with scale parameter $M\sigma\sqrt{T_0}(1 + \|\gamma\|_2)$. Using a standard discretization argument (Wainwright, 2018), the L^2 norm satisfies the tail bound

$$P(\|\mu'(\varepsilon_1 - \varepsilon'_c\gamma)\|_2 \geq 4M\sigma\sqrt{T_0 J \log 5}(1 + \|\gamma\|_2) + x) \leq 2e^{-\frac{x^2}{M^2\sigma^2 T_0(1+\|\gamma\|_2)^2}}$$

Plugging in $x = 4M\sigma\sqrt{T_0 J \log 5}(1 + \|\gamma\|_2)$ gives that $\|\mu'(\varepsilon_1 - \varepsilon_c'\gamma)\|_2 \leq 8M\sigma\sqrt{T_0 J \log 5}(1 + \|\gamma\|_2)$ with probability at least $1 - 2e^{-8J\log 5}$.

Finally, note that $\|\mu_T\|_2 \leq \sqrt{J}\|\mu_t\|_\infty \leq M\sqrt{J}$. Combining the pieces gives the result. □

Proof of Theorem 1. From Lemma 7 and using the triangle inequality, the absolute bias satisfies the bound

$$\left|\mathbb{E}\left[\mu_T\phi_1 - \sum_{W_i=0}\gamma_i Y_{iT}\right]\right| \leq \frac{1}{T_0}\left\|\mu_T \cdot \left(\mu' X_1 - \sum_{W_i=0}\gamma_i \mu' X_i\right)\right\|_2 + \frac{1}{T_0}\left\|\mu_T \cdot \left(\mu'\varepsilon_1 - \sum_{W_i=0}\gamma_i\mu'\varepsilon_i\right)\right\|_2$$

Recall that $\mu'\mu = T_0 I_J$, so $\|\mu'\|_2 = \|\mu\|_2 = \sqrt{T_0}$ Again using Cauchy-Schwarz and $\|\mu_T\|_2 \leq M\sqrt{J}$, we have the bound

$$\frac{1}{T_0}\left\|\mu_T \cdot \left(\mu' X_1 - \sum_{W_i=0}\gamma_i\mu' X_i\right)\right\|_2 \leq \frac{M\sqrt{J}}{T_0}\|\mu'\|_2\|X_1 - X_{0\cdot}'\gamma\|_2 \leq M\sqrt{\frac{J}{T_0}}\|X_1 - X_{0\cdot}'\gamma\|_2$$

Combining this with Lemma 8 yields that

$$\left|\mathbb{E}\left[\mu_T\phi_1 - \sum_{W_i=0}\gamma_i Y_{iT}\right]\right| \leq M\sqrt{\frac{J}{T_0}}\|X_1 - X_{0\cdot}'\gamma\|_2 + 8M^2 J\sigma\sqrt{\frac{\log 5}{T_0}}\,(1 + \|\gamma\|_2)$$

□

Proof of Lemmas 1 and 4. Recall that the lagged outcomes are centered by the control averages. Notice that

$$\begin{aligned}
\hat{Y}_1(0)^{\text{aug}} &= \hat{m}(X_1) + \sum_{W_i=0}\hat{\gamma}_i(Y_{iT} - \hat{m}(X_i)) \\
&= \hat{\eta}_0 + \hat{\eta}' X_1 + \sum_{W_i=0}\hat{\gamma}_i(Y_{iT} - \hat{\eta}_0 - X_i'\hat{\eta}) \\
&= \sum_{W_i=0}(\hat{\gamma}_i + (X_1 - X_{0\cdot}'\hat{\gamma})(X_{0\cdot}'X_{0\cdot} + \lambda I_{T_0})^{-1}X_i)Y_i \\
&= \sum_{W_i=0}\hat{\gamma}_i^{\text{aug}} Y_i
\end{aligned} \tag{56}$$

The expression for $\hat{Y}_1(0)^{ridge}$ is clear. □

Proof of Lemmas 2 and 5. Recall that the lagged outcomes are centered by the control averages. The imbalance of the ridge-augmented weights (7) is

$$\begin{aligned}
\|X_{1\cdot} - \hat{\gamma}^{aug\prime}X_{0\cdot}\|_2 &= \|(X_{1\cdot} - \hat{\gamma}'X_{0\cdot})(I - (X_{0\cdot}'X_{0\cdot} + \lambda I)^{-1}X_{0\cdot}'X_{0\cdot}\|_2 \\
&\leq \frac{\lambda}{\lambda + N_c\sigma_{\min}^2}\|X_{1\cdot} - \hat{\gamma}'X_{0\cdot}\|_2.
\end{aligned} \tag{57}$$

The 2-norm of the weights is

$$\begin{aligned}\|\hat{\gamma}^{aug}\|_2 &= \left\|\hat{\gamma} + X_{1\cdot}(X_{0\cdot}'X_{0\cdot} + \lambda I)^{-1}X_{0\cdot}'\right\|_2 \\ &\leq \|\hat{\gamma}\|_2 + \frac{\sqrt{N_c}\sigma_{\min}}{\lambda + N_c\sigma_{\min}^2}\|X_{1\cdot} - \hat{\gamma}'X_{0\cdot}\|_2.\end{aligned} \tag{58}$$

The imbalance of the ridge regression weights is:

$$\begin{aligned}\|X_{1\cdot} - \hat{\gamma}^{ridge\prime}X_{0\cdot}\|_2 &= \|X_{1\cdot}(I - (X_{0\cdot}'X_{0\cdot} + \lambda I)^{-1}X_{0\cdot}'X_{0\cdot}\|_2 \\ &\leq \left(1 - \frac{N_c\sigma_{\min}^2}{N_c\sigma_{\min}^2 + \lambda}\right)\|X_{1\cdot}\|_2 \\ &\leq \frac{\lambda}{\lambda + N_c\sigma_{\min}^2}\|X_{1\cdot}\|_2.\end{aligned} \tag{59}$$

The 2-norm of the weights is

$$\begin{aligned}\|\hat{\gamma}^{ridge}\|_2 &= \left\|\frac{1}{N_c}\mathbf{1} + X_{1\cdot}(X_{0\cdot}'X_{0\cdot} + \lambda I)^{-1}X_{0\cdot}'\right\|_2 \\ &\leq \frac{1}{\sqrt{N_c}} + \left\|X_{1\cdot}(X_{0\cdot}'X_{0\cdot} + \lambda I)^{-1}X_{0\cdot}'\right\|_2 \\ &\leq \frac{1}{\sqrt{N_c}} + \frac{\sqrt{N_c}\sigma_{\min}}{\lambda + N_c\sigma_{\min}^2}\|X_{1\cdot}\|_2.\end{aligned} \tag{60}$$

□

Proof of Corollaries 1 and 2. Plugging in the results of Lemmas 2 and 5 into Theorem 1 and simplifying the expression directly yields the results. □

Proof of Lemma 6. The Lagrangian dual to (33) is

$$\min_{\alpha,\beta} \frac{1}{2}\sum_{W_i=0}(\alpha + \beta'X_i)^2 - (\alpha + \beta'X_1) + \frac{\lambda}{2}\|\beta\|_2^2 \tag{61}$$

Solving for α we see that

$$\sum_{W_i=0}\hat{\alpha} + \hat{\beta}'X_i = 1$$

Since the lagged outcomes are centered, this implies that

$$\hat{\alpha} = \frac{1}{N_c}$$

Now solving for β we see that

$$X_{0\cdot}'\left(\mathbf{1}\frac{1}{N_c} + X_{0\cdot}\hat{\beta}\right) + \lambda\hat{\beta} = X_1$$

This implies that

$$\hat{\beta} = (X_{0\cdot}'X_{0\cdot} + \lambda I)^{-1}X_1$$

Finally, the weights are the ridge weights

$$\hat{\gamma}_i = \frac{1}{N_c} + X_1'(X_{0\cdot}'X_{0\cdot} + \lambda I)^{-1}X_i$$

□

Proof of Lemma 3. By replacing X with Z in Lemma 1 we find the form of the weights. The imbalance is

$$\begin{aligned} X_1 - \hat{\gamma}^{\text{cov}\prime}X_{0\cdot} &= X_1 - (\hat{\gamma}^{\text{scm}\prime} + (Z_1 - Z_0'\hat{\gamma}^{\text{scm}})'(Z_0'Z_0)^{-1}Z_0')X_{0\cdot} \\ &= X_1 - Z_1'(Z_0'Z_0)^{-1}X_{0\cdot} - \hat{\gamma}^{\text{scm}\prime}(X_{0\cdot} - Z_0(Z_0'Z_0)^{-1}Z_0'X_{0\cdot}) \\ &= X_1 - \widehat{X}_1 - (X_{0\cdot} - \widehat{X}_{0\cdot})'\hat{\gamma}^{\text{scm}} \end{aligned} \tag{62}$$

□

Proof of Proposition 1. The bias of $\widehat{\text{Var}}(\hat{\tau})$ is

$$\mathbb{E}_{\varepsilon_T}\left[\widehat{\text{Var}}(\hat{\tau}) - \text{Var}(\hat{\tau})\right] = \left(\frac{1}{N_1} + \|\hat{\gamma}\|_2^2\right)\mathbb{E}_{\varepsilon_T}[\hat{\nu}_T - \nu_T] \tag{63}$$

We now compute $\mathbb{E}[\hat{\nu}_T - \nu_T]$:

$$\begin{aligned} \mathbb{E}_{\varepsilon_T}[\hat{\nu}_T - \nu_T] &= \frac{1}{N_0}\sum_{W_i=0}\mathbb{E}_{\varepsilon_T}[(Y_i - \tilde{Y}_i)^2] - \nu_T \\ &= \frac{1}{N_0}\sum_{W_i=0}\text{Bias}_i^2 + \text{Var}_{\varepsilon_T}(Y_i - \tilde{Y}_i) - \nu_T \\ &= \frac{1}{N_0}\sum_{W_i=0}\text{Bias}_i^2 + (1 + \|\tilde{\gamma}^{(i)}\|_2^2)\nu_T - \nu_T \\ &= \frac{1}{N_0}\sum_{W_i=0}\text{Bias}_i^2 + \|\tilde{\gamma}^{(i)}\|_2^2\nu_T \\ &> 0 \end{aligned} \tag{64}$$

□

C Additional numerical results and simulations

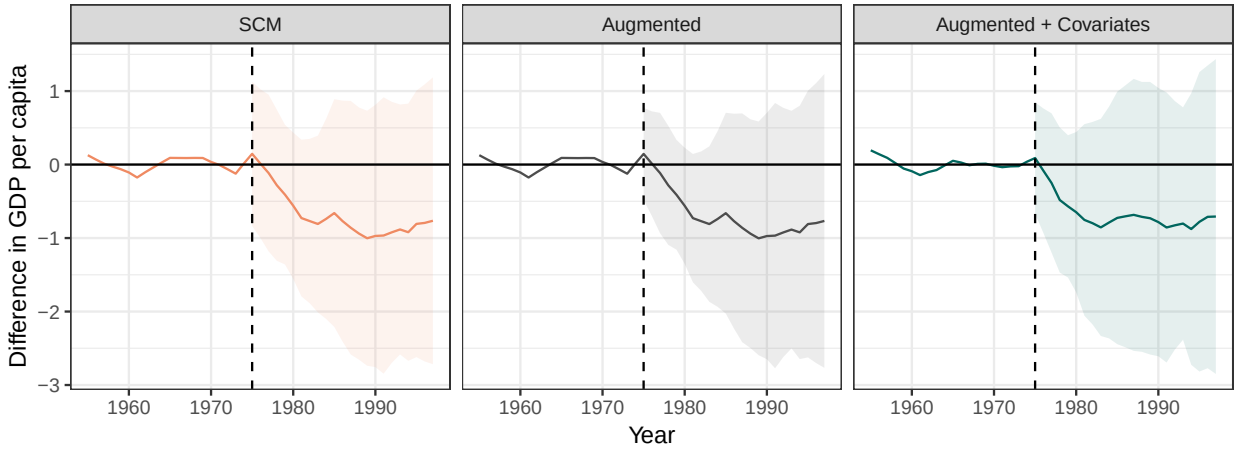
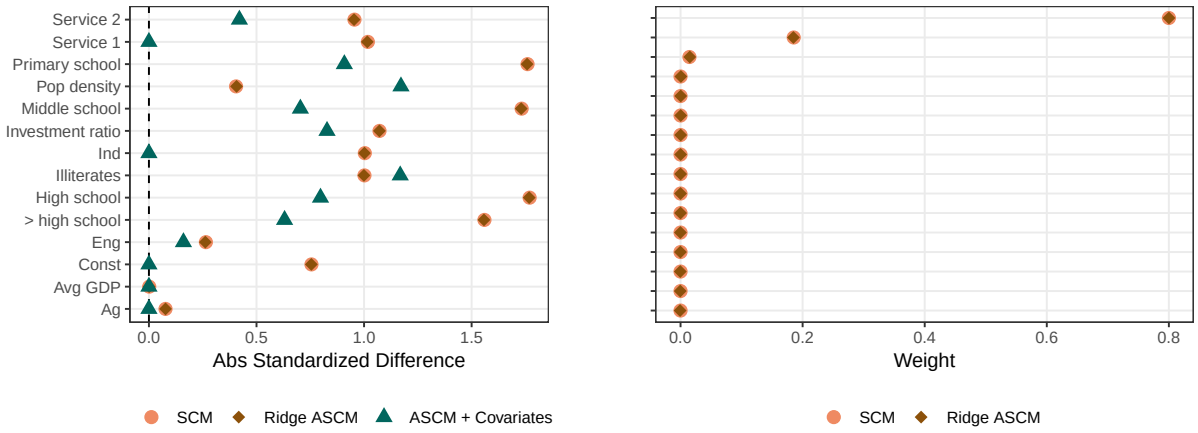


Figure C.1: Point estimates ± two standard errors of the ATT for the Basque example (Abadie and Gardeazabal, 2003) using SCM, ridge ASCM, and ASCM with covariates. Note that ridge ASCM achieves nearly perfect balance for the pre-period outcomes.

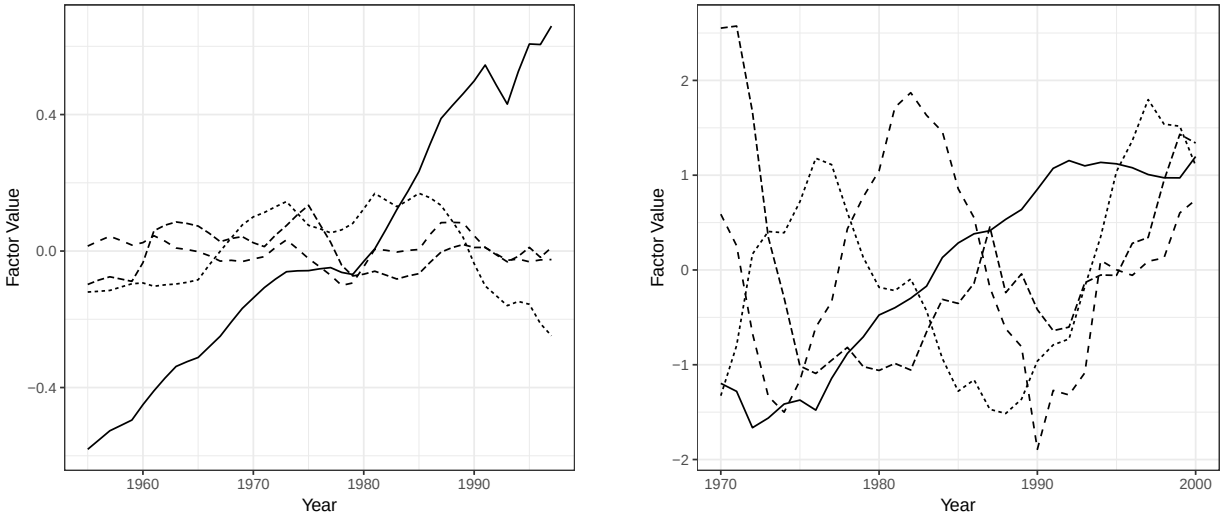


(a) Covariate balance

(b) Weights

Figure C.2: (a) Covariate balance in the Basque example for SCM, ridge ASCM, and ASCM with covariates. (b) Donor unit weights in the Basque example for SCM and ridge ASCM. The ridge and ridge-augmented SCM estimates yield exact balance for the pre-period outcomes and that SCM and ridge-augmented SCM have nearly identical weights.

Basque analysis In addition to replication the Prop. 88 study in Abadie et al. (2010), we replicate the study on the impact of terrorism on the Basque ecnonomy in Abadie and Gardeazabal (2003). As before we estimate the effect using (1) SCM on the lagged outcomes, (2) ridge ASCM



(a) Basque

(b) California

Figure C.3: Estimated factors for the Basque and California examples, using GSynth (Xu, 2017) with $J = 4$.

on the lagged outcomes, and (3) the covariate-adjusted SCM estimator described in Section 7.[20] Figure C.1 shows the effect estimates and their standard errors, Figure C.2a shows the imbalance in the auxiliary covariates and C.2b compares the weights for SCM and ridge ASCM. Note that in covariate balance is quite good for SCM and thus SCM and ridge ASCM are nearly identical; adjusting for covariates improves balance in the lagged outcomes by roughly 7.5%.

Machine learning methods We evaluate alternative outcome models for use in ASCM. For each DGP we consider SCM augmented with (1) a random forest, (2) a two layer feed-forward neural network, (3) matrix completion using `MCPanel` (Athey et al., 2017), (4) Lasso, and (5) `CausalImpact` (Brodersen et al., 2015). We compare ASCM to the pure outcome models as well as pure SCM.

Figure C.5 shows the absolute bias for these methods and Appendix Figure C.6 shows the RMSE. We broadly see the same results as with ridge ASCM. In our simulations, augmenting SCM with flexible outcome models always reduces the bias, with some models improving SCM more than others. Additionally, ASCM can have much better performance than either SCM or outcome modelling alone. For instance, with such small data the neural network is badly biased in almost all simulations, but augmenting SCM with a neural network model greatly reduces the bias relative to both estimators; this is especially stark in the Basque simulation, and we see similar behavior for `CausalImpact` in the fixed effects simulation.

Overall we find consistently good performance across data generating processes from SCM augmented with a penalized regression model. Due to this good performance, the small number of units and high dimensional structure of typical SCM settings, and their relative simplicity, we

[20]Following the discussion in Section 7.2, we use a Lasso regression of post-treatment outcomes on the auxiliary covariates to reduce the dimension of covariates to adjust for. The Lasso selected the average pre-treatment GDP per capita, and measures of industry share for the agricultural, construction, industrial, and service sectors.

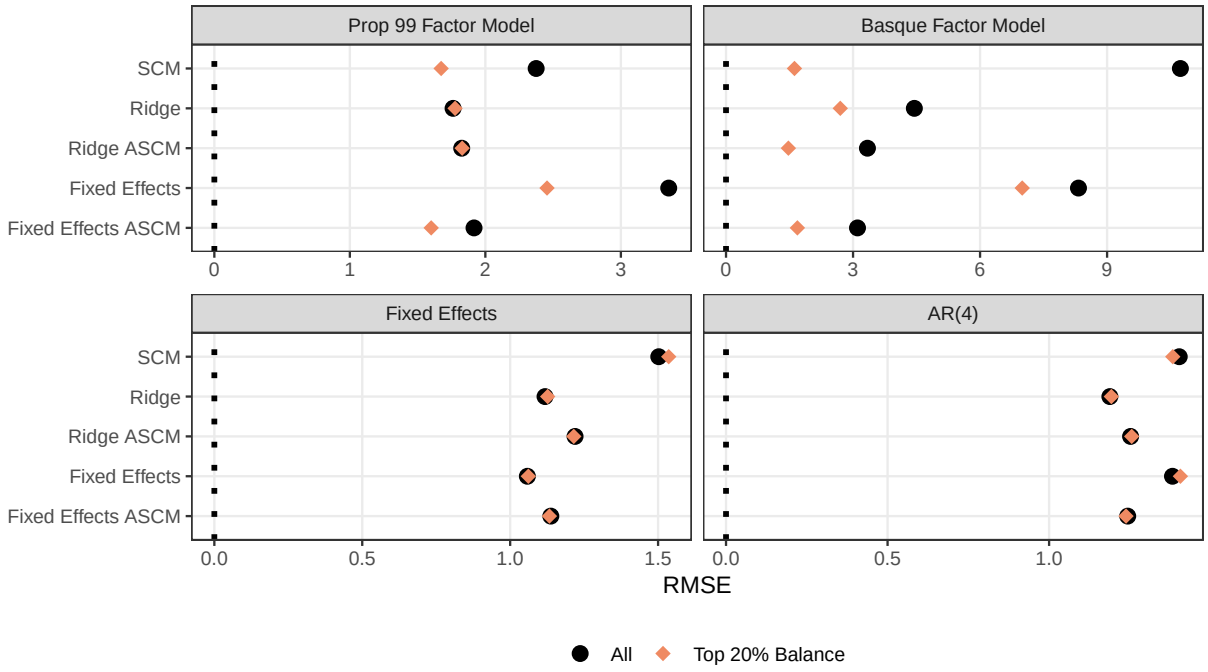


Figure C.4: RMSE for different augmented and non-augmented estimators across outcome models.

recommend augmenting SCM with penalized regression in practice. In particular we suggest using ridge regression; among the other benefits, ridge ASCM allows the practitioner to diagnose the level of extrapolation that the outcome model adds.

Alternatives to the uniform permutation test In Appendix Figure C.8 we show the empirical distribution for the p-value under strong selection using uniform permutation, uniform permutation restricting to units with positive SCM weight, two conformal inference approaches described in Chernozhukov et al. (2017), a weighted randomization test using the true propensity scores, and a weighted randomization test using propensity scores estimated with entropy-penalized SCM (14). In general we find that the conformal inference methods and uniform permutation are invalid. The ad-hoc solution of restricting the uniform permutation to control units with positive SCM appears to be valid but lacks power, and the weighted randomization test with the estimated propensity score appears to lack any power whatsoever. This simulation evidence suggests that valid and powerful permutation-style hypothesis testing for SCM remains an open problem.

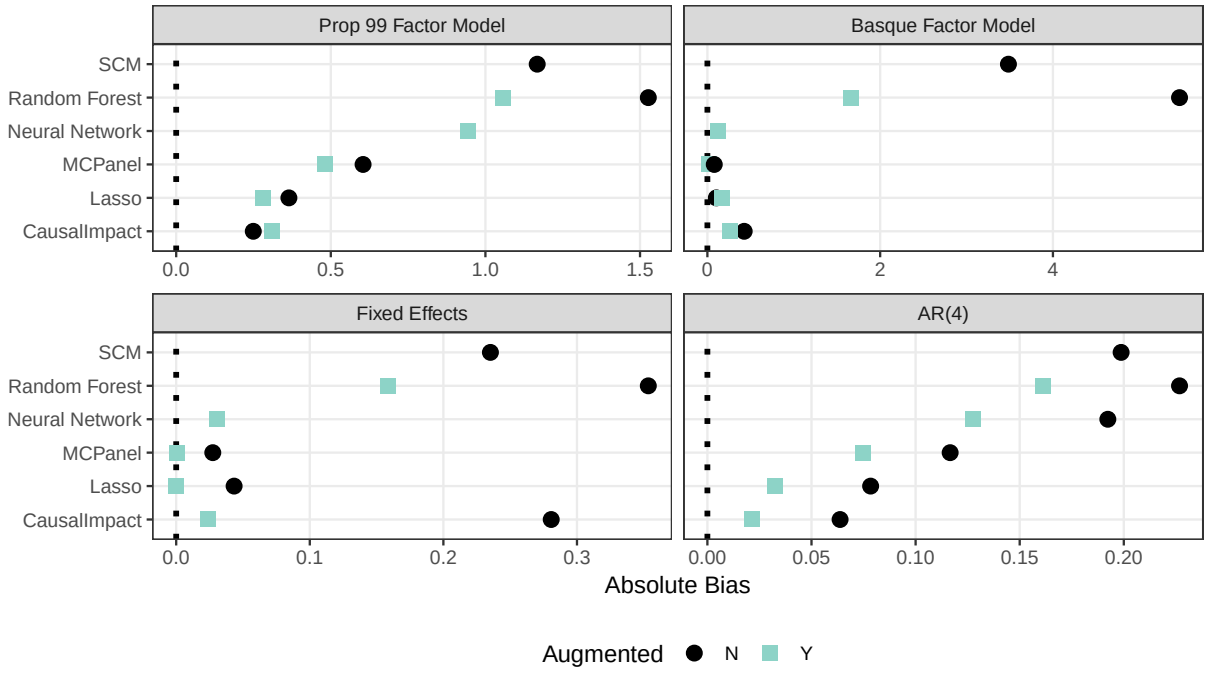


Figure C.5: Absolute bias for SCM, several machine learning and panel data outcome models, and their augmented versions using the same data generating processes as Figure 4. For three panels the Neural Network bias is not shown: for the Prop. 99 simulation it is 7.61, for the Basque simulation it is 28.3, and for the fixed effects simulation it is 3.63.

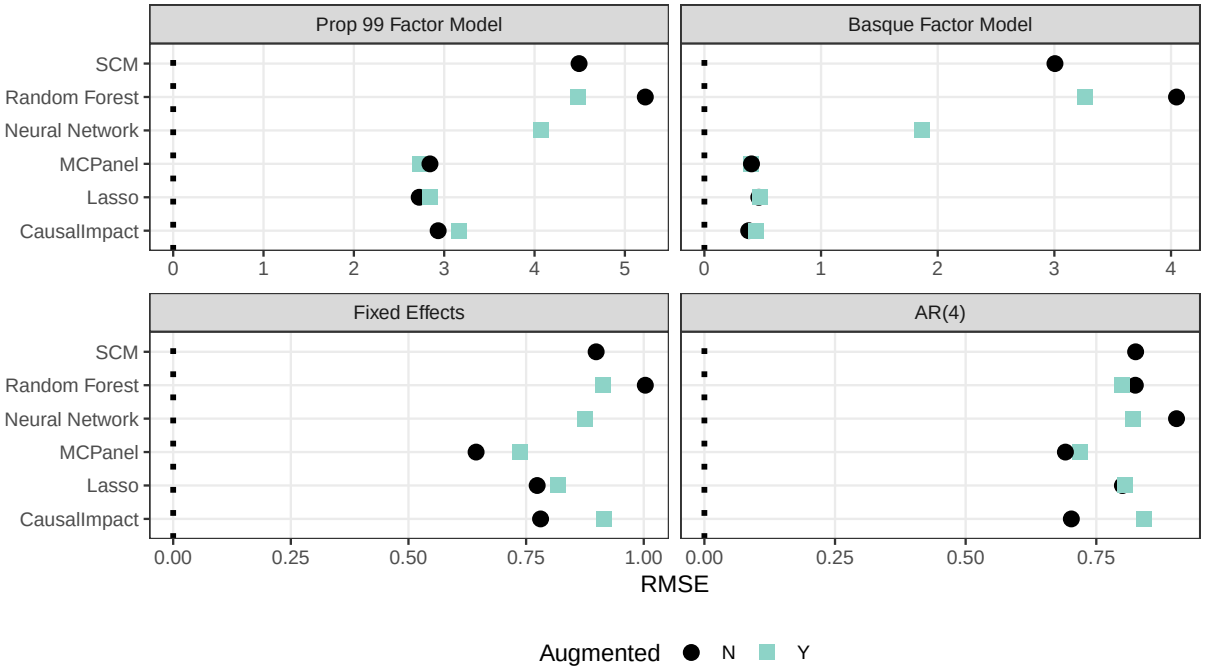
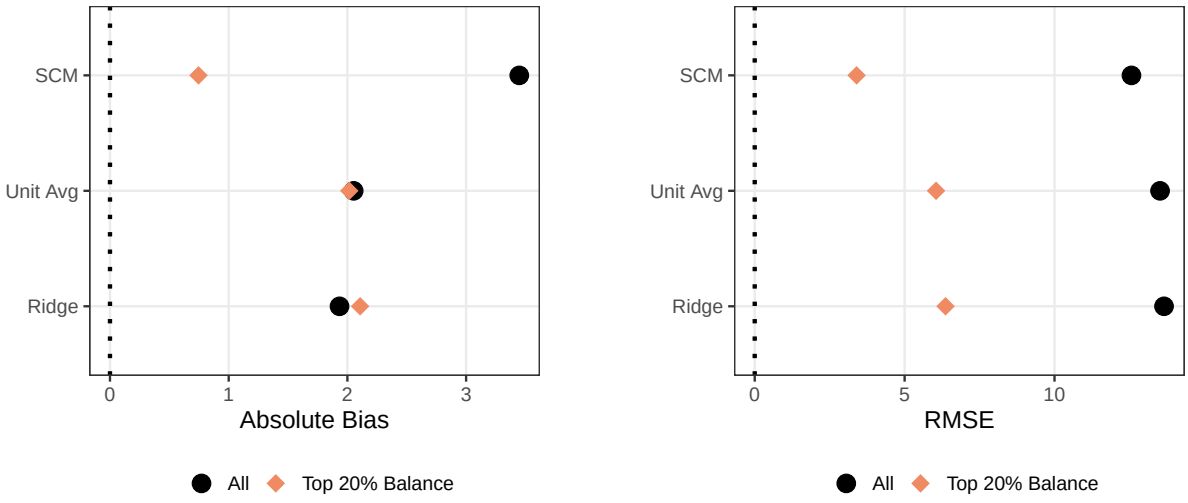


Figure C.6: RMSE for different augmented and non-augmented estimators across outcome models, using flexible machine learning and panel data approaches.



(a) Bias

(b) RMSE

Figure C.7: Strategies for dimension reduction. Basque simulation.

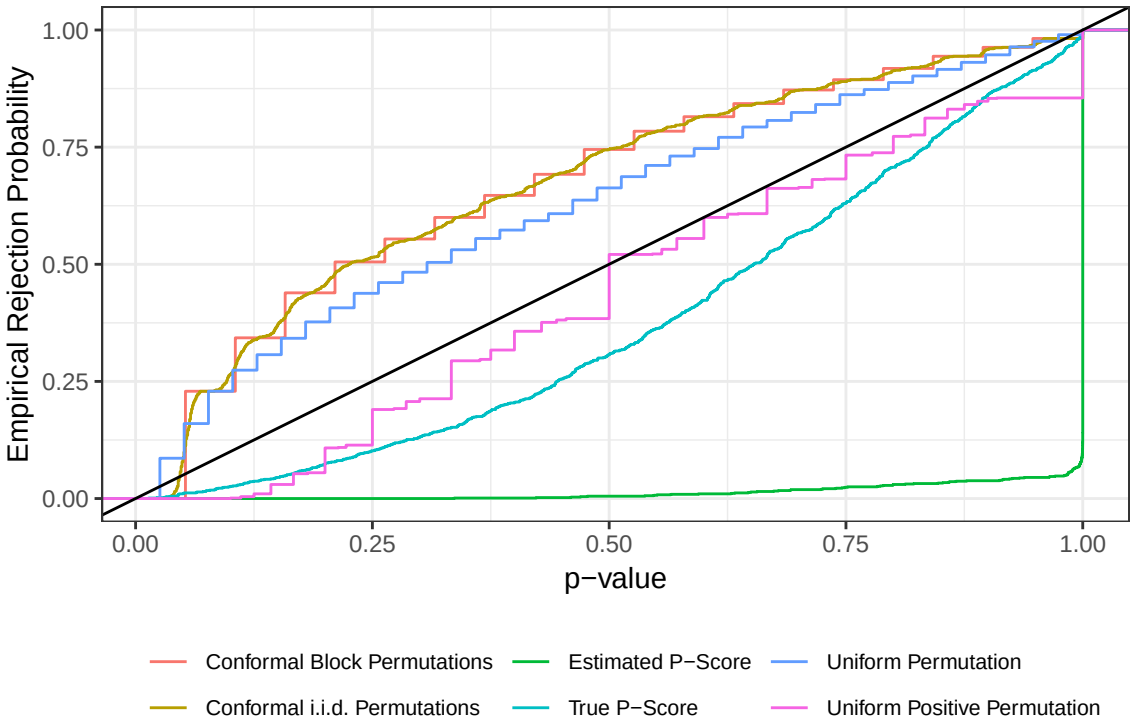


Figure C.8: Cumulative density functions for various methods of constructing a p-value under strong selection ($\theta = 1$) for the Prop. 99 simulation. (1) Conformal block permutations and (2) i.i.d. permutations are as described in Chernozhukov et al. (2017). (3) Estimated P-Score uses entropy penalized SCM (14) to estimate the propensity score, uses the cross validation technique described in Doudchenko and Imbens (2017) to select the hyper parameter, and computes a weighted p-value (52). (4) True P-Score Uses the true propensity score in the weighted p-value (52). (5) Uniform Permutation performs a uniform permutation test over the control units and (6) Uniform Positive Permutation performs a uniform permutation test, restricted to control units with non-zero SCM weight.

Estimation method	Inference method	Bias	RMSE	Coverage for ATT	Average CI Length
Ridge ASCM	Bootstrap	0.521	1.613	0.704	1.751
	Leave one out	0.521	1.613	0.973	4.053
SCM	Bootstrap	1.939	3.421	0.550	2.297
	Leave one out	1.939	3.421	0.917	6.028

Table 1: Coverage for standard errors in the Prop. 99 simulation. Standard errors are estimated using a simple non-parametric blocked bootstrap, and the leave out one method described in Section 6.2. The bias, RMSE, and average confidence interval length are all scaled by the residual standard deviation.