Title
Why weight? Weighting approaches for causal inference with panel and cross-sectional data

Author
Ben-Michael, Elijahu

Publication Date
2020

Peer reviewed|Thesis/dissertation

Why weight? Weighting approaches for causal inference with panel and cross-sectional data

by

Elijahu E Ben-Michael

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Avi Feller, Co-chair
Professor Peng Ding, Co-chair
Professor Bin Yu
Professor Jesse Rothstein

Fall 2020

Why weight? Weighting approaches for causal inference with panel and cross-sectional data

Abstract

Why weight? Weighting approaches for causal inference with panel and cross-sectional data

by

Elijahu E Ben-Michael

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Avi Feller, Co-chair

Professor Peng Ding, Co-chair

In observational studies, researchers wish to study the effect of a treatment without directly controlling treatment assignment. These studies are particularly useful when it is uneconomical, unethical, or infeasible for researchers to manipulate treatment in a controlled setting. They also offer insight into how treatment affects large, naturally occurring populations, and so they are indispensible counterparts to randomized trials, which are typically conducted on smaller, unrepresentative study samples. A key feature of randomized trials is that researchers can use randomized treatment assignment to ensure that the treated and control sample do not differ in observed and unobserved characteristics, on average. Observational studies have no such guarantee, as treatment assignment occurs through some unknown process. In practice, this unknown process often results in substantial differences between the treated and control samples, leading to substantial biases in naive comparisons between the two groups and confounding the relationship between the outcome and treatment.

There are many methods that attempt to overcome this bias by searching for ways in which the treatment and control groups are comparable, e.g. by restricting the sample to the region around a discontinuity in treatment assignment, or matching treated and control units based on baseline characteristics. In this thesis we take a similar approach, using weighting estimators that take a weighted average of treated and control outcomes, with weights that make the two groups directly comparable. If the treated and control groups differ on pre-treatment characteristics that are highly correlated with the outcome, then comparisons between the two groups will be highly biased. However, if we can find weights so that the two groups are balanced on these pre-treatment characteristics after weighting, then the bias will be negligible. Therefore, in this thesis we address the problem of confounding by addressing the problem of imbalance, finding weights that directly optimize for balance between the weighted treated and control samples.

Each of the chapters in this thesis follows a common "recipe". First, we write the estimation error of a weighting estimator explicitly in terms of balance. This informs *what* aspects of the pre-treatment characteristics we should balance. Then we show *how* to achieve balance, constructing a convex optimization problem that directly controls the balance, with a tradeoff between better balance and lower variance. Finally, in some settings we cannot find weights that achieve a sufficient level of balance. In these cases, we can account for any remaining imbalance by combining the weighting estimator with a predictive model of the outcome. Chapter 1 briefly covers the broad strokes of this general recipe in a simplified observational study setting, where the goal is to estimate the average treatment effect for the treated population. The subsequent chapters apply this recipe to answer questions in the social sciences by developing weighting approaches to estimate the treatment effect on the treated in three different settings.

Chapter 2 considers estimating treatment effects in comparative case-studies, where a single unit is treated and there is access to a long series of pre-intervention outcomes. In this setting, variants of weighting estimators that ensure balance on pre-intervention outcomes are known as the synthetic control method (SCM), where the "synthetic control" is a weighted average of comparison units. By inspecting the estimation error we see that an important feature of the original SCM proposal is to use it only when the weights have excellent balance on pre-intervention outcomes. This chapter primarily focuses on the final step, proposing Augmented SCM as an extension of SCM when it is not possible to achieve good-enough pre-treatment fit. The main proposal is to use ridge regression to de-bias the original SCM estimate; we show that this estimator can itself be written as a modified synthetic controls problem, allowing for limited extrapolation in order to improve pre-treatment fit. We then use this framework to inspect the impact of an aggressive tax cut in Kansas in 2012, finding evidence that the tax cuts hindered economic growth. We implement this estimation procedure in a new `R` package, `augsynth`.[1]

Chapter 3 builds on Chapter 2 to adapt the synthetic control method to estimating treatment effects with staggered adoption of treatment by different units at different times. Current practice is to fit SCM separately for each treated unit, averaging the resulting estimates. Following the recipe above, we show that the estimation error depends on both the average imbalance across the synthetic controls and the imbalance of the average of the synthetic controls. We propose finding "partially pooled" SCM weights that minimize both the average and treated-unit specific fits. Finally, we combine these weights with a fixed effects estimate of the outcomes. We then apply this method to measure the impact of teacher collective bargaining laws on school spending, finding minimal impacts. As in Chapter 2, we implement this procedure in the `augsynth` `R` package.

Finally, Chapter 4 focuses on estimating treatment effects for subgroups in observational studies with cross-sectional data, analyzing a pilot study on letters of recommendation in

[1] Available at https://github.com/ebenmichael/augsynth.

UC Berkeley undergraduate admissions. Here, we are interested in understanding how the effect of submitting a letter of recommendation varies for under-represented students and for applicants with different *a priori* probabilities of admission. Again following the general recipe, we build on results in Chapter 3 to see that the estimation error for a subgroup depends on the "local balance" within the subgroup. Using this, we develop balancing weights that solve a convex optimization problem to directly optimize for the local balance within subgroups while maintaining global covariate balance between the overall treated and control samples. We then show that this approach has a dual representation as inverse propensity score weighting with a hierarchical propensity score model and use a random forest to de-bias the weighting estimator. Overall, we find that the impact of letters of recommendation is higher for applicants with a higher predicted probability of admission, and find mixed evidence of differences for under-represented minority applicants.

To my family.

Contents

List of Figures

List of Tables

Acknowledgments

I'd like to thank my advisor Avi Feller, who has been instrumental to both my intellectual and professional development. Through long working sessions, quick messages, and impromptu video calls, Avi always made time for me, even when I didn't know to ask for it. He helped me workshop and develop ideas from vague musings to legitimate projects, and gave me constant feedback on framing, presentation, and writing. Avi encouraged me to cast a wide net, introducing me to many different collaborators and spawning several fruitful projects with a single email. Finally, Avi played an enormous amount of defense for me, protecting my time and dealing with the various administrative issues that arose. In short, I owe him for his excellent guidance. I'm not sure that I can ever pay it back, but I certainly plan to pay it forward.

The other members of my committee also greatly shaped me as a statistician and researcher. Peng Ding taught me what it means to be a careful and precise statistician. His deep well of knowledge and insight was invaluable, and I am grateful to him for agreeing to serve as the co-chair of the committee. Without Jesse Rothstein, the work in this thesis would be a jumbled mess. He helped to refine many of the ideas and gave important perspective and intellectual grounding, serving as a voice of reason and redirecting me the many times I lost the thread. In addition, Jesse helped to bring Chapter 4 from a loose set of notes to a fully-fledged chapter with the UC Berkeley admissions pilot study. Finally, I am thankful to have had Bin Yu serve both as an informal mentor and as the chair of my qualifying exam committee. Bin made me feel at home in her group meetings and is a role model for any aspiring applied statistician. She is also a great ally for the graduate students in the department.

In the past few years I had the chance to work with and learn from many other excellent researchers. Skip Hirshberg and Jose Zubizarreta—whose research was influential to me well before I met them—have helped to clarify my thinking on the work in this dissertation and beyond, and have given me great guidance. Jas Sekhon, through his courses and reading group, exposed me to causal inference and taught me to read social science research with a critical eye. Erin Hartman, Luke Keele, Luke Miratrix, and Elizabeth Stuart have been incredible collaborators and mentors, and I look forward to continuing to work with them. I would also like to thank La Shana Porlaris, as well as the entire Statistics department staff, without whom I surely would have missed every single degree requirement.

I've been incredibly fortunate for my fellow students, who helped me grow personally and intellectually. In particular, I want to thank Jake Soloff and Bryan Liu, who have been with me since those many late nights in Evans; Jamie Murdoch for our camaraderie; Sara Stoudt for her giant heart; Steve Howard for the hours making music; Rebecca Barter, Ryan Giordano, and Kellie Ottoboni for their guidance as senior students. I'd also like to thank past and current SGSA presidents for their work creating a cohesive graduate student community. Of course I'd be remiss not to thank the stats lounge foosball table and everyone who played on it for the much needed diversion. There are several people to thank outside of the world of Berkeley. First, my friends, who reminded me that there even is an outside

world, and in particular Noah Zinsmeister for housing me in two boroughs and Saratoga Springs, and both Luke Dodge and Varshini Narayanan for being there for me anytime and anywhere. Second, Dominika Blach, for righting me when I stumble and lifting me up. Finally, my parents and brother: without them I don't know where I'd be.

Chapter 1

Balancing weights for causal inference

We begin by describing the general approach to weighting in observational studies in a simplified setting. For units $i = 1, \ldots, n$, we observe an outcome $Y_i \in \mathbb{R}$, a binary treatment variable $W_i \in \{0, 1\}$, and a set of d pre-treatment baseline characteristics $X_i \in \mathbb{R}^d$. Let $n_1 = \sum_i W_i$ be the number of treated units and $n_0 = n - n_1$ be the number of control units. Following the potential outcomes framework (Neyman, 1923; Holland, 1986), we will posit two potential outcomes for unit i, $Y_i(1)$ and $Y_i(0)$, that correspond to the outcome under treatment and the outcome under control, respectively. Implicitly, this notation rules out interference between units and different forms of treatment (Rubin, 1980). The observed outcome is thus $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$. We will further assume that $(X_i, W_i, Y_i(0), Y_i(1))$ are sampled i.i.d. from a distribution $\mathcal{P}(\cdot)$. We will begin with two key restrictions on this distribution, that allow for estimation of causal effects in this setting.

Assumption 1.1 (Ignorable treatment assignment)**.** The potential outcomes are independent of treatment assignment given the baseline characteristics:

$$Y(0), Y(1) \perp\!\!\!\perp W \mid X \tag{1.1}$$

Assumption 1.2 (One Sided Overlap)**.** The *propensity score* $e(x) = P(W = 1 \mid X = x) < 1$.

Together, Assumptions 1.1 and 1.2 are usually known as *strong ignorability* (Rosenbaum and Rubin, 1983). In the chapters that follow we will consider variants of these foundational assumptions, assessing their credibility in the three applications. Another important conditional expectation is the *prognostic score*, $m(x, w) \equiv \mathbb{E}[Y(w) \mid X = x]$, where $m(X, 1)$ and $m(X, 0)$ are the conditional expectation of the treated and control potential outcome given X, respectively. The strong ignorability assumptions yield non-parametric identification of the (population) average treatment effect on the treated via

$$\mathbb{E}[Y(1) - Y(0) \mid W = 1] = \mathbb{E}[Y \mid W = 1] - \mathbb{E}\left[\frac{e(X)}{1 - e(X)} Y \mid W = 0\right]. \tag{1.2}$$

With this setup, we follow Athey et al. (2018) and focus on estimating the (conditional) average treatment effect on the treated:

$$\tau = \mu_1 - \mu_0 \quad \text{where} \quad \mu_1 = \frac{1}{n_1}\sum_{W_i=1} m(X_i, 1), \quad \mu_0 = \frac{1}{n_1}\sum_{W_i=1} m(X_i, 0). \tag{1.3}$$

It is straightforward to estimate μ_1 as the average of the outcomes for the treated sample.; however, estimating μ_0 requires more care. As we will see below, the problem is that we only observe control potential outcomes for units that receive the control condition, *not* for the units that receive the treatment condition. In this chapter and those that follow, we will be interested in estimating μ_0 via a weighted average of the control units with weights $\hat{\gamma}$:

$$\hat{\mu}_0 = \sum_{W_i=0} \hat{\gamma}_i Y_i. \tag{1.4}$$

One way to choose the weights $\hat{\gamma}$ is as a plug-in version of the identification result in Equation (1.2). This approach, often called *inverse propensity score weighting*, first finds an estimator of the propensity score $\hat{e}(x)$ and then sets the weights to be an estimate of the odds of treatment $\hat{\gamma}_i = \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}$ (Imbens and Rubin, 2015). The finite sample performance of these weighting approaches usually suffers when there are many covariates: one issue is that these weights involve inverting the estimate $1 - \hat{e}(X_i)$, which can behave poorly if the propensity score is close to one (Athey et al., 2018). Instead, our focus will be on choosing these weights to explicitly balance covariates, and therefore to explicitly reduce the estimation error.

1.1 Choosing what to balance

In order to decide how to find the weights, we will first inspect the estimation error: the difference between the counterfactual mean for the treated group, μ_0, and our estimator $\hat{\mu}_0$. Denoting the residual $\varepsilon_i \equiv Y_i - m(X_i, 0)$, the estimation error decomposes into two terms: the error due to imbalance and the error due to noise,

$$\hat{\mu}_0 - \mu_0 = \underbrace{\frac{1}{n_1}\sum_{W_i=1} m(X_i, 0) - \sum_{W_i=0} \hat{\gamma}_i m(X_i, 0)}_{\text{imbalance in prognostic}} + \underbrace{\sum_{W_i=0} \hat{\gamma}_i \varepsilon_i}_{\text{noise}}. \tag{1.5}$$

Throughout, we will consider *design-based* weights that are independent of the outcomes. Therefore, the noise term will be mean-zero and so any bias is due to imbalance in the prognostic score $m(\cdot, 0)$. On the other hand, the variance of the estimator will be proportional to the sum of the squared weights $\|\hat{\gamma}\|_2^2$.[1]

[1] In the survey sampling context a transformation of the square 2-norm is known as the *effective sample size* $n^{\text{eff}} = \|\hat{\gamma}\|_2^{-2} \left(\sum_i \hat{\gamma}_i\right)^2$.

Ideally, we would attempt to minimize both the bias and the variance; however, we do not know the true prognostic score. In the absence of this knowledge we can instead minimize an *upper bound* on the bias. Specifically, if we posit that the prognostic score is in a class of functions $\mathcal{M}$, then the imbalance term will be less than the worst-case imbalance in the model class:

$$\left|\frac{1}{n_1}\sum_{W_i=1} m_0(X_i) - \sum_{W_i=0}\hat{\gamma}_i m_0(X_i)\right| \leq \max_{m\in\mathcal{M}}\left|\frac{1}{n_1}\sum_{W_i=1} m(X_i) - \sum_{W_i=0}\hat{\gamma}_i m(X_i)\right| \equiv \text{imbalance}_{\mathcal{M}}(\gamma). \tag{1.6}$$

Therefore, to specify what to balance we first choose a model class $\mathcal{M}$. An important special case is where the prognostic score is assumed to be linear in a basis of the covariates, $\phi(x)$, with a bounded coefficient vector, $\Phi^p = \{\beta\cdot\phi(x) \mid \|\beta\|_p \leq C\}$. With this model class, by Hölder's inequality the bias depends on the imbalance in the basis functions, measured via the dual norm,

$$\text{imbalance}_{\Phi^p}(\gamma) = C\left\|\frac{1}{n_1}\sum_{W_i=1}\phi(X_i) - \sum_{W_i=0}\hat{\gamma}_i\phi(X_i)\right\|_q \quad \text{where} \quad \frac{1}{p}+\frac{1}{q}=1. \tag{1.7}$$

There are two common choices for Φ^p. First, if we believe that β is (approximately) sparse, we we can choose $p=1$—so that the one-norm of β is bounded—and then $\text{imbalance}_{\Phi^1}(\gamma)$ measures imbalance by the L^∞ norm of the imbalance vector, i.e. the difference between the treated and re-weighted control samples in the least-balanced transformation of the covariates. Many approaches to approximate balancing weights use some variant of this model class, including Zubizarreta (2015); Athey et al. (2018); Ning et al. (2017); Tan (2018); Wang and Zubizarreta (2019). Second, we may think that the covariates have roughly equal contribution and so we set $p=2$. With this choice of model class, $\text{imbalance}_{\Phi^2}(\gamma)$ measures imbalance in the transformed covariates through the L^2 norm, which penalizes larger imbalances more heavily than smaller imbalances and controls the imbalance in the *average* transformation of the covariates rather than the least-balanced one. As we will see in Section 2.4, ridge regression implicitly find weights for this model class. Related approaches use a reproducing kernel Hilbert space $\mathcal{H}_k$, with associated kernel k, as the model class. In this case the prognostic score is linear in the *reproducing kernel feature map* $\phi(x) = k(\cdot, x)$, and we can compute $\text{imbalance}_{\mathcal{H}_k}(\gamma)$ via the kernel trick; see, e.g., Kallus (2016); Wong and Chan (2017); Hazlett (2020); Hirshberg et al. (2019).

1.2 Finding weights via convex optimization

Having decided what to balance, we will now consider how to construct the weights. We will find weights that minimize a monotonic function $h_\lambda(\cdot)$ of the imbalance – parameterized by a hyperparameter λ – while regularizing the weights with a strictly convex *dispersion function*, $f:\mathbb{R}\to\mathbb{R}$ that penalizes non-uniform weights. This produces a trade-off between

bias and variance where better balance comes at the price of lower precision. To fix ideas, we will consider balancing the model class Φ^p. In this case we solve

$$\min_{\gamma} \quad h_\lambda \left(\left\| \frac{1}{n_1} \sum_{W_i=1} \phi(X_i) - \sum_{W_i=0} \hat{\gamma}_i \phi(X_i) \right\|_q \right) + \sum_{W_i=1} f(\gamma_i). \tag{1.8}$$

This formulation leads to several choices in constructing the optimization problem.

Constrained vs Lagrangian form. There are several monotonic transformations $h_\lambda(\cdot)$ to consider. One choice is a soft penalty with a scaling factor λ, $h_\lambda(x) = \frac{1}{\lambda} x^2$. Here, λ determines how much to prioritize balance—and hence bias reduction—against the dispersion of the weights. When λ is small we place greater emphasis on balance, and when λ is large we place more emphasis on variance.

An alternative choice to the soft penalty is to explicitly constrain the imbalance, $h_\lambda(x) = \mathcal{I}(x \leq \lambda)$.[2] With this choice, the optimization (1.8) finds the most uniform weights subject to a maximal allowed amount of imbalance. Here again λ controls the bias variance trade-off, but the constraint guarantees that the imbalance will be less than λ. Note that from the usual correspondence between the constrained and Lagrangian forms of an optimization problem, there exists some choice of the hyperparameters so that the soft-penalized and constrained approaches have the same solution. However, in many settings is may be beneficial to explicitly constrain the optimization problem.

Exact balance and the dispersion function. The simplest choice of balance criterion is to infinitely penalize any imbalance in any basis function by choosing $h(x) = \mathcal{I}(x = 0)$. Then, the particular choice of norm becomes irrelevant, and we find weights that *exactly balance* the basis functions $\phi(x)$. Deville and Särndal (1992) and Deville et al. (1993) consider this procedure in the survey sampling context, solving (1.8) with various dispersion measures for continuous or categorical covariates. More recently, there has been renewed interest in exact balancing weights in the causal inference literature. Entropy balancing (Hainmueller, 2011) is one such proposal, and Chan et al. (2016) explicitly adopt the framework of Deville and Särndal (1992) to the causal inference setting.

When enforcing exact balance, the only free parameter in the balancing weights optimization problem (1.8) is the choice of *dispersion function* $f(\cdot)$, which penalizes non-uniformity in the weights: different choices of dispersion function will yield different weights that exactly balance the covariates. One choice is an L^2 dispersion penalty, $f(\gamma_i) = \gamma_i^2$, which directly controls the variance of the weighting estimator. In fact, the minimum variance weights that exactly balance $p < n$ basis functions corresponds to linear regression weights (Kline, 2011); see Chapter 2. Chan et al. (2016) and Wang and Zubizarreta (2019) consider a wide variety

[2] $\mathcal{I}(A) = \begin{cases} 0 & x \in A \\ \infty & x \notin A \end{cases}$ is the indicator function.

of dispersion penalties including the entropy dispersion penalty $f(\gamma_i) = \gamma_i \log \gamma_i$ considered by Hainmueller (2011).

Unfortunately, with covariates of even moderate dimension, we cannot expect to be able to achieve exact balance. In many practical and empirical applications we often must suffice with *approximate balance*.[3] In general, if we can find weights that exactly balance the covariates, we should consider protecting against imbalance in a wider function class, e.g. by including more transformations of the covariates or balancing a non-parametric function class. Therefore, the particular functional form of the dispersion function is not crucial, and in the chapters below we default to controlling the variance of the weights with an L^2 dispersion penalty.

Further constraints on the weights and ruling out extrapolation These weights can also include post-processing steps in a principled way inside the optimization problem. For example, trimming weights to prevent any particular unit from receiving too much weight is a popular post-hoc processing step in traditional MLE-based IPW estimation. We can directly include this constraint into the optimization problem, rather than applying a post-hoc transformation to the weights. We can incorporate a lower and upper bound on the weights, $L \leq \gamma_i \leq U$ by including an infinite penalty in the dispersion function: for a dispersion function f, we can create a new dispersion function $\tilde{f}(\gamma_i) = \begin{cases} f(\gamma_i) & L \leq \gamma_i \leq U \\ \infty & \gamma_i > U, \gamma_i < L \end{cases}$. In the survey-sampling context, Deville and Särndal (1992) show how to combine several dispersion functions with weight truncation.

Another frequent post-hoc transformation is normalizing the inverted estimated propensity score weights to sum to one. We can also include this constraint in a principled manner into optimization problem (1.8). To constrain the weights to sum to one we can modify the model class Φ^p to include the set of constant functions $\tilde{\Phi}^p \equiv \Phi^p \cup \{m(x) = c \mid c \in \mathbb{R}\}$ and enforce exact balance on the set of constant functions, so that $h(\text{imbalance}_{\tilde{\Phi}^p}(\gamma)) = \begin{cases} h(\text{imbalance}_{\Phi^p}(\gamma)) & \sum_{W_i=0} \gamma_i = 1 \\ \infty & \sum_{W_i=0} \gamma_i \neq 1 \end{cases}$. A sum–to-one constraint on the weights combined with a non-negativity constraint, forbids *extrapolation* outside of the support of the data. Specifically, we limit the re-weighted covariate distribution to be in the *convex hull* of the treated units, $\{\sum_{W_i=0} \gamma_i \phi(X_i) \mid \gamma_i \geq 0, \sum_i \gamma_i = 1\}$. In so doing, we are ensuring that our estimator is *interpolating* between control units, which is often preferable to extrapolation. We discuss the relative merits of interpolation and extrapolation in Chapter 2.

[3] Zhao and Percival (2017) show that if we restrict the weights to be non-negative and sum to one (see next paragraph), a sufficient condition for exact balance is to have *exponentially* many more units than covariates (or basis functions).

The duality between balancing weights and propensity score modelling

We now connect the weights from Equation (1.8) to inverse propensity weights by inspecting the Lagrangian dual problem. We will see that solving (1.8) in fact fits a regularized propensity score model with a different loss function than the usual MLE approach, where the balance criterion determines the type and level of regularization.

We begin by deriving the Lagrangian dual for optimization problem (1.8).

Proposition 1.1. The Lagrangian dual to (1.8) is

$$\min_{\beta} \underbrace{\frac{1}{n}\sum_{i=1}^{n}(1-W_i)f^*(\phi(X_i)\cdot\beta) - W_i\phi(X_i)\cdot\beta}_{\text{balancing loss function}} + \underbrace{h^*(\|\beta\|_p)}_{\text{regularization}}, \tag{1.9}$$

where for a convex function f, $f^*(y) \equiv \sup_x \langle x, y\rangle - f(x)$ is the *convex conjugate*.
If $\hat{\beta}$ is the solution to the Lagrangian dual, the unit weights $\hat{\gamma}_i$ that solve the primal problem (1.8) are recovered as

$$\hat{\gamma}_i = f^{*\prime}(\hat{\beta}\cdot\phi(X_i)) \tag{1.10}$$

This Lagrangian dual (1.9) will appear in different forms and settings in the chapters that follow.[4] It consists of two key components, the first determined by the choice of dispersion function f and the second determined by the model class Φ^p and choice of either a soft penalty or a hard constraint. First, the dispersion function f controls the form of the *loss function*, the first two terms in Equation (1.9). In particular, through the dual we can write the weights as functions of the covariates, $\hat{\gamma}(X_i) = f^{*\prime}(\beta\cdot\phi(X_i))$. As we discuss below, these can be viewed as estimates of the odds of treatment $\frac{e(x)}{1-e(x)} \approx f^{*\prime}(\beta\cdot\phi(x))$, so $\beta\cdot\phi(x)$ is the inverse propensity score in a "natural parameter" scale, and the derivative of the convex conjugate of the dispersion function, $f^{*\prime}$ determines the link function. In particular, if $f(\gamma_i) = \frac{1}{2}\gamma_i^2$ is the two norm of the weights, the link function is the identity—$f^{*\prime}(x) = x$—and if the weights are constrained to be non-negative, then the link function enforces that constraint: $f^{*\prime}(x) = \max\{0, x\}$. Finally, if $g(\gamma_i) = \gamma_i \log \gamma_i$ is the entropy of the weights, then the link function is exponential $f^{*\prime}(x) = e^x$.

Second, the choice of model class Φ^p determines the type of regularization through the norm $\|\beta\|_p$.[5] For example, for the class of (approximately) sparse linear models Φ^1 with a hard constraint $h_\lambda(x) = \mathcal{I}(x \le \lambda)$, the dual problem is regularized via the L^1 norm, $h^*_\lambda(\|\beta\|_p) = \lambda\|\beta\|_1$; this enforces sparsity in the dual variables (see e.g. Wang and Zubizarreta,

[4] Proposition 1.1 is equivalent to Proposition A.2 in Appendix A. It is stated here, with slightly different notation, for clarity.

[5] The choice of hard or soft penalty controls whether this norm is squared or not. When $h_\lambda(x) = \mathcal{I}(x \le \lambda)$ is a hard constraint, $h^*_\lambda(\|\beta\|_p)) = \lambda\|\beta\|_p$ and when $h_\lambda(x) = \frac{1}{2\lambda}x^2$ is a soft constraint, $h^*_\lambda(\|\beta\|_p) = \frac{\lambda}{2}\|\beta\|_p^2$. When $h_\lambda(x) = |x|$, the dual problem is constrained: $h^*_\lambda(\|\beta\|_p) = \mathcal{I}(\|\beta\|_p \le \lambda)$.

2019). To see this, note that the zero-subgradient condition of the dual problem (1.9) implies that the only components of $\hat{\beta}$ that are non-zero are those that correspond to covariates that are on the constraint boundary:

$$\hat{\beta}_j \neq 0 \Leftrightarrow \left| \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i) - \sum_{W_i=1} \hat{\gamma}_i \phi_j(X_i) \right| = \lambda. \tag{1.11}$$

So the zero sub-gradient condition ensures that the dual solution will be sparse, with the only active coefficients corresponding to the binding constraints (Zubizarreta, 2015). In the following chapters we will often be balancing Φ^2 with a soft penalty, where the dual problem has a ridge regularization $h^*(\|\beta\|_p) = \frac{\lambda}{2}\|\beta\|_2^2$.

M-estimation of the inverse propensity score The zero gradient condition for the balancing loss in the dual problem (1.9) shows that it is an M-estimator for the propensity score, and so the weights are a regularized M-estimator of the inverse propensity score. First, consider the expected value of the loss function:

$$\bar{q}(\beta) \equiv \mathbb{E}\left[(1 - W_i) f^*(\phi(X_i) \cdot \beta) - W_i \phi(X_i) \cdot \beta\right]. \tag{1.12}$$

The gradient is the level of imbalance in $\phi(X)$, and the zero gradient condition is a population balance condition for the odds of treatment:

$$\nabla \bar{q}(\beta) = 0 \quad \Leftrightarrow \quad \mathbb{E}\left[(1 - W_i) f^{*\prime}(\phi(X_i) \cdot \beta)\phi(X_i)\right] = \mathbb{E}\left[W_i \phi(X_i)\right]. \tag{1.13}$$

By the balancing property of the propensity score, we see that the treatment odds $\frac{e(X_i)}{1-e(X_i)} = f^{*\prime}(\phi(X_i)'\beta)$ satisfies the zero gradient condition. Since the population loss $\bar{q}$ is convex, this implies that the inverse propensity score is a minimizer, and if $\bar{q}$ is *strictly* convex then the inverse propensity score is the unique minimizer.[6] Thus, the dual problem (1.9) is a *regularized M estimator* for the inverse propensity score. Zhao and Percival (2017) use a variant of this M-estimation argument for the special case with a logistic link using entropy balancing with covariates entering linearly (Hainmueller, 2011). Tan (2017) and Wang and Zubizarreta (2019) extend this argument to the setting with sparsity, using an L^∞ balance criterion. Zhao (2018) considers a broader class of loss functions under the equivalent characterization as proper scoring rules and considers balancing weights for various causal estimands.

1.3 Augmentation and bias correction

As we note above, and will see in the chapters that follow, in moderate to high dimensional settings we can only expect *approximate* rather than *exact* balance. From Equation (1.5), we

[6] One necessary condition for this is that the propensity score is correctly specified.

can see that the remaining imbalance leaves room for bias in our estimator. To account for imbalance that remains after weighting, we can try to explicitly estimate and adjust for the bias. We begin with an estimate of the prognostic score, e.g. fit via least squares regression on the control sample,

$$\min_{m \in \mathcal{M}} \sum_{W_i=0} (Y_i - m(X_i, 0))^2. \tag{1.14}$$

With this estimator $\hat{m}(\cdot, 0)$ in hand, we can estimate the bias by plugging it in to Equation (1.5). We can directly correct for this via the *augmented* or *bias-corrected* estimator

$$\hat{\mu}^{\text{aug}} = \underbrace{\sum_{W_i=0} \hat{\gamma}_i Y_i}_{\text{weighting estimator}} + \underbrace{\frac{1}{n_1} \sum_{W_i=1} \hat{m}(X_i, 0) - \sum_{W_i=0} \hat{\gamma}_i \hat{m}(X_i, 0)}_{\text{estimate of bias due to imbalance}}. \tag{1.15}$$

This approach is analogous to bias-correction for inexact matching (Abadie and Imbens, 2011), and through the dual relation above we can see that it is similar to the Augmented IPW estimator (Robins et al., 1994).

In the following chapters we will make use of bias-corrected estimators of this form in various contexts. To see the benefit of this approach, we can return to the error decomposition in Equation (1.5). By adjusting for the estimated bias, the estimation error now depends on the imbalance in the *error* in the prognostic score, $\delta m(x) \equiv \hat{m}(x, 1) - m(x, 1)$, rather than the imbalance in the prognostic score itself:

$$\hat{\mu}_0^{\text{aug}} - \mu_0 = \underbrace{\frac{1}{n_1} \sum_{W_i=1} \delta m(X_i, 0) - \sum_{W_i=0} \hat{\gamma}_i \delta m(X_i, 0)}_{\text{imbalance in error}} + \underbrace{\sum_{W_i=0} \hat{\gamma}_i \varepsilon_i}_{\text{noise}}. \tag{1.16}$$

Intuitively, if the error in the estimated prognostic score is small then the bias will also be small. Various different augmented and bias-corrected approaches use this argument, relying on Hölder's inequality, to characterize the error in both high dimensional settings (Ning et al., 2017; Athey et al., 2018; Hirshberg and Wager, 2018) and nonparametric settings (Kallus, 2016; Wong and Chan, 2017; Zhao, 2018). Hirshberg and Wager (2019) propose a variant of this approach that takes Equation (1.16) as a starting point, and finds weights to minimize the worst-case imbalance in the *regression error*, rather than the worst-case imbalance in the prognostic score.

1.4 Dissertation roadmap

This chapter—based on unpublished material coauthored with David Hirshberg, Jose Zubizarreta, and Avi Feller—lays out the foundation for the remaining chapters. Each chapter—based on material coauthored with Avi Feller and Jesse Rothstein—will use a variant of the bias-corrected weighting estimator above, with special attention given to one or all of the

three steps. First, inspect the estimation error of the weighting estimator under a single or multiple model classes. Second, find weights that control (an upper bound of) this estimation error. Finally, use an outcome model to correct for remaining imbalance.

Chapter 2 considers using this procedure to estimate causal effects in comparative case studies, with a single treated unit and outcome measures both before and after treatment. Inspecting the estimation error, under several different outcome models it is sufficient to control the balance in pre-treatment outcomes. Various weighting approaches that control this imbalance are known as the "Synthetic Control Method." From this starting point, Chapter 2 primarily focuses on the bias-correction (or augmentation) step. In particular the chapter uses ridge regression to predict post-treatment outcomes from pre-treatment outcomes, and analyzes the numerical and statistical properties of this bias-corrected weighting estimator.

Chapter 3 continues this thread by generalizing to the setting where multiple units adopt a treatment at different times. Starting with current practice—estimating a separate synthetic control for each treated unit—we show that the estimation error can be controlled by both the average imbalance across the synthetic controls and the imbalance of the average of the synthetic controls. We then propose a "partially pooled" synthetic control estimator that minimizes a convex combination of these two imbalances. Following the procedure laid out in this chapter, we conclude by bias-correcting the synthetic control estimates, focusing on predicting post-treatment outcomes with a simple two-way fixed effects model.

Finally, Chapter 4 uses these ideas to estimate subgroup treatment effects in observational studies, designing an observational study evaluating a UC Berkeley pilot program on letters of recommendation for undergraduate admissions. We first show that the estimation error for subgroup treatment effects can be controlled by the "local imbalance" within each subgroup, while the estimation error for the overall treatment effect additionally requires control over the "global imbalance" across subgroups. Continuing with the procedure, we then find weights which minimize the local imbalance within each subgroup while exactly balancing the treated and control samples across the dataset. We conclude by considering bias correction by predicting outcomes via the LASSO or with a random forest.

Chapter 2

The Augmented Synthetic Control Method

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

2.1 Introduction

The *synthetic control method* (SCM) is a popular approach for estimating the impact of a treatment on a single unit in panel data settings with a modest number of control units and with many pre-treatment periods (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015). The idea is to construct a weighted average of control units, known as a synthetic control, that matches the treated unit's pre-treatment outcomes. The estimated impact is then the difference in post-treatment outcomes between the treated unit and the synthetic control. SCM has been widely applied — the main SCM papers have over 4,000 citations — and has

been called "arguably the most important innovation in the policy evaluation literature in the last 15 years" (Athey and Imbens, 2017).

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

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

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

We relate Ridge ASCM's improved pre-treatment fit to a finite sample bound on estimation error under several data generating processes, including an autoregressive model and the linear factor model often invoked in this setting (Abadie et al., 2010). Under an autoregressive model, improving pre-treatment fit directly reduces bias, and the Ridge ASCM penalty term negotiates a bias-variance trade-off. Under a latent factor model, improving pre-treatment fit again reduces bias, though there is now a risk of over-fitting, and the penalty term again directly parameterizes this trade-off. Thus, choosing the hyperparameter will be important for practice; we propose a cross-validation procedure in Section 2.5.

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

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

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

The chapter proceeds as follows. Section 2.1 briefly reviews related work. Section 2.2 introduces notation, the underlying models and assumptions, and the SCM estimator. Section 2.3 gives an overview of Augmented SCM. Section 2.4 gives key numerical results for Ridge ASCM. Section 2.5 bounds the Ridge ASCM estimation error under a linear model and under a linear factor model, the standard setting for SCM, and also addresses inference. Section 2.6 extends the ASCM framework to incorporate auxiliary covariates. Section 2.7 reports on extensive simulation studies as well as the application to the Kansas tax cuts. Finally, Section 4.6 discusses some possible directions for further research. The supplementary materials in Appendix A includes all of the proofs, as well as additional derivations and technical discussion.

Related work

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

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

A second set of papers relaxes constraints imposed in the original SCM problem, in particular the restriction that control unit weights be non-negative. Doudchenko and Imbens (2017) argue that there are many settings in which negative weights would be desirable. Amjad et al. (2018) propose an interesting variant that combines negative weights with a pre-processing step. Powell (2018) instead allows for extrapolation via a Frisch-Waugh-Lovell-style projection, which similarly generalizes the typical SCM setting. Doudchenko

and Imbens (2017) and Ferman and Pinto (2018) both propose to incorporate an intercept into the SCM problem, which we discuss in Section 2.3.

There have also been several other proposals to reduce bias in SCM, developed independently and contemporaneously with ours. Abadie and L'Hour (2018) also propose bias correcting SCM using regression. Kellogg et al. (2020) propose using a weighted average of SCM and matching, trading off interpolation and extrapolation bias. Arkhangelsky et al. (2019) propose the *Synthetic Difference-in-Differences* estimator, which can be seen as a special case of our proposal with a constrained outcome regression.

Finally, there have also been recent proposals to use outcome modeling rather than SCM-style weighting in this setting. These include the matrix completion method in Athey et al. (2017), the generalized synthetic control method in Xu (2017), and the combined approaches in Hsiao et al. (2018). We explore the performance of select methods, both in isolation and within our ASCM framework, in Section 2.7.

2.2 Overview of the Synthetic Control Method

Notation and setup

We consider the canonical SCM panel data setting with $i = 1, \ldots, N$ units observed for $t = 1, \ldots, T$ time periods; for the theoretical discussion below, we will consider both N and T to be fixed. Let W_i be an indicator that unit i is treated at time $T_0 < T$ where units with $W_i = 0$ never receive the treatment. We restrict our attention to the case where a single unit receives treatment, and follow the convention that this is the first one, $W_1 = 1$; see Chapter 3 for an extension to multiple treated units. The remaining $N_0 = N - 1$ units are possible controls, often referred to as *donor units* in the SCM context. To simplify notation, we limit to one post-treatment observation, $T = T_0 + 1$, though our results are easily extended to larger T.

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

$$Y_{it} = \begin{cases} Y_{it}(0) & \text{if } W_i = 0 \text{ or } t \leq T_0 \\ Y_{it}(1) & \text{if } W_i = 1 \text{ and } t > T_0. \end{cases} \tag{2.1}$$

To emphasize that pre-treatment outcomes serve as covariates in SCM, we use X_{it}, for $t \leq T_0$, to represent pre-treatment outcomes; we use the terms *pre-treatment fit* and *covariate balance* interchangeably. With some abuse of notation, we use $\boldsymbol{X}_{0\cdot}$ to represent the N_0-by-T_0 matrix of control unit pre-treatment outcomes and $\boldsymbol{Y}_{0T}$ for the N_0-vector of control unit

outcomes in period T. With only one treated unit, Y_{1T} is a scalar, and $\boldsymbol{X}_{1\cdot}$ is a T_0-row vector of treated unit pre-treatment outcomes. The data structure is then:

$$\begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1T_0} & Y_{1T} \\ Y_{21} & Y_{22} & \dots & Y_{2T_0} & Y_{2T} \\ \vdots & & & & \vdots \\ Y_{N1} & Y_{N2} & \dots & Y_{NT_0} & Y_{NT} \end{pmatrix} \equiv \left(\begin{array}{cccc|c} X_{11} & X_{12} & \dots & X_{1T_0} & Y_{1T} \\ \hline X_{21} & X_{22} & \dots & X_{2T_0} & Y_{2T} \\ \vdots & & & & \vdots \\ \underbrace{X_{N1} & X_{N2} & \dots & X_{NT_0}}_{\text{pre-treatment outcomes}} & Y_{NT} \end{array}\right) \equiv \left(\begin{array}{c|c} \boldsymbol{X}_{1\cdot} & Y_{1T} \\ \hline \boldsymbol{X}_{0\cdot} & \boldsymbol{Y}_{0T} \end{array}\right) \tag{2.2}$$

Assumptions on the data generating process

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

$$Y_{it}(0) = m_{it} + \varepsilon_{it}. \tag{2.3}$$

This setup encompasses many common panel data models; see Chernozhukov et al. (2019) for an extended discussion. In Section 2.5, we consider two specific versions of (2.3): $Y_{it}(0)$ is linear in its lagged values; and $Y_{it}(0)$ is linear in a set of latent factors. In the supplementary materials, we also consider the case where m_{it} is a linear model with Lipshitz deviations from linearity. The results in Section 2.4 are purely numeric and do not rest on specific assumptions about the underlying model.

We begin with our assumptions on the distribution of the noise terms, followed by assumptions on the model component.

Assumption 2.1 (Noise component)**.** The noise terms ε_{it} for $i = 1, \dots, N$ and $t = 1, \dots, T$ are independent across units and time, and are sub-Gaussian with scale parameter σ.

(a) In our first DGP, we assume that the post-treatment noise terms $\varepsilon_{1T}, \dots, \varepsilon_{NT}$ have zero mean for each unit:

$$\mathbb{E}\left[\varepsilon_{iT}\right] = 0 \quad \forall i = 1, \dots, N. \tag{2.4}$$

(b) In our second DGP, we further assume that the noise terms for all units *and all periods* $t = 1, \dots, T$ have zero mean:

$$\mathbb{E}\left[\varepsilon_{it}\right] = 0 \quad \forall i = 1, \dots, N \text{ and } \forall t = 1, \dots, T. \tag{2.5}$$

Assumption 2.2 (Model component)**.** The control potential outcomes are generated according to the following model and error components:

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

$$Y_{it}(0) = \sum_{\ell=1}^{T_0} \beta_\ell Y_{i(t-\ell)}(0) + \varepsilon_{it}. \tag{2.6}$$

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

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

$$Y_{it}(0) = \boldsymbol{\phi}_i \cdot \boldsymbol{\mu}_t + \varepsilon_{it} = \sum_{j=1}^{J} \phi_{ij}\mu_{jt} + \varepsilon_{it}. \tag{2.7}$$

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

Together, the pair of Assumptions 2.1 and 2.2 enable estimation of the missing counterfactual outcome. In particular, the mean-zero noise restrictions hold for the treated unit ($i = 1$), and rule out any unmeasured variables that are correlated with the outcomes and that have different distributions for the treated unit and comparison units. Under the DGP in Assumption 2.2(a), treatment assignment can depend on the past outcomes, but cannot depend on post-treatment outcomes; furthermore, there cannot be serial correlation between the post-treatment and pre-treatment noise. This DGP includes the special case of an auto-regressive process of order $K < T_0$. Under the DGP in Assumption 2.2(b), treatment assignment can depend on the factor loadings, but cannot depend on the *realized* pre-treatment outcomes. We discuss this in more detail in the context of our application in Section 2.7.

Synthetic Control Method

The Synthetic Control Method imputes the missing potential outcome for the treated unit, $Y_{1T}(0)$, as a weighted average of the control outcomes, $\boldsymbol{Y}_{0T}'\boldsymbol{\gamma}$ (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015). Weights are chosen to balance pre-treatment outcomes and possibly other covariates. We consider a version of SCM that chooses weights $\boldsymbol{\gamma}$ as a solution

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

to the constrained optimization problem:

$$\begin{aligned} \min_{\gamma} \quad & \|\boldsymbol{V}_{\boldsymbol{x}}^{1/2}(\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma})\|_2^2 \; + \; \zeta \sum_{W_i=0} f(\gamma_i) \\ \text{subject to} \quad & \sum_{W_i=0} \gamma_i = 1 \\ & \gamma_i \geq 0 \quad i : W_i = 0 \end{aligned} \tag{2.8}$$

where the constraints limit $\boldsymbol{\gamma}$ to the simplex $\Delta^{N_0} = \{\boldsymbol{\gamma} \in \mathbb{R}^{N_0} \mid \gamma_i \geq 0 \; \forall i, \; \sum_i \gamma_i = 1\}$, and where $\boldsymbol{V}_{\boldsymbol{x}} \in \mathbb{R}^{T_0 \times T_0}$ is a symmetric importance matrix and $\|\boldsymbol{V}_{\boldsymbol{x}}^{1/2}(\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\gamma)\|_2^2 \equiv (\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\gamma)'\boldsymbol{V}_{\boldsymbol{x}}(\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma})$ is the 2-norm on $\mathbb{R}^{T_0}$ after applying $\boldsymbol{V}_{\boldsymbol{x}}^{1/2}$ as a linear transformation. To simplify the exposition and notation below, we will generally take $\boldsymbol{V}_{\boldsymbol{x}}$ to be the identity matrix. The simplex constraint in Equation (2.8) ensures that the weights will be sparse and non-negative; Abadie et al. (2010, 2015) argue that enforcing this constraint is important for preserving interpretability.

Equation (2.8) modifies the original SCM proposal in two ways.[2] First, Equation (2.8) penalizes the dispersion of the weights with hyperparameter $\zeta \geq 0$, following a suggestion in Abadie et al. (2015). The choice of penalty is less central when weights are constrained to be on the simplex, but becomes more important below when we relax this constraint (Doudchenko and Imbens, 2017). Second, Equation (2.8) excludes auxiliary covariates; we re-introduce them in Section 2.6.

When the treated unit's vector of lagged outcomes, $\boldsymbol{X}_{1\cdot}$, is inside the convex hull of the control units' lagged outcomes, $\boldsymbol{X}_{0\cdot}$, the SCM weights in Equation (2.8) achieve perfect pre-treatment fit, and the resulting estimator has many attractive properties. In this setting, Abadie et al. (2010) show that SCM will be unbiased under the auto-regressive model in Assumption 2.2(a) and bound the bias under the linear factor model in Assumption 2.2(b).

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

[2]Equation (2.8) follows the recent methodological literature and directly optimizes for the pre-treatment fit, minimizing the (possibly weighted) imbalance of pre-treatment outcomes between the treated unit and the weighted control mean. In Section 2.5, we argue that this a natural quantity to target under both linearity and a latent factor model. Many choices are possible, however, and we can easily modify Equation (2.8) to balance other summary measures and functions of the lagged outcomes.

2.3 Augmented SCM

Overview

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

$$\hat{Y}_{1T}^{\text{aug}}(0) = \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} Y_{iT} \quad + \left(\hat{m}_{1T} - \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} \hat{m}_{iT} \right) \tag{2.9}$$

$$= \hat{m}_{1T} \quad + \quad \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} (Y_{iT} - \hat{m}_{iT}), \tag{2.10}$$

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

Equations (2.9) and (2.10), while equivalent, highlight two distinct motivations for ASCM. Equation (2.9) directly corrects the SCM estimate, $\sum \hat{\gamma}_i^{\text{scm}} Y_{iT}$, by the imbalance in a particular function of the pre-treatment outcomes $\hat{m}(\cdot)$. Intuitively, since $\hat{m}$ estimates the post-treatment outcome, we can view this as an estimate of the bias due to imbalance, analogous to bias correction for inexact matching (Abadie and Imbens, 2011). In this form, we can see that SCM and ASCM estimates will be similar if the estimated bias is small, as measured by imbalance in $\hat{m}(\cdot)$. If the estimated bias is large, the two estimators will diverge, and the conditions for appropriate use of SCM will not apply. In independent work, Abadie and L'Hour (2018) also consider a bias-corrected estimator of this form.

Equation (2.10), by contrast, is analogous to standard doubly robust estimation (Robins et al., 1994), which begins with the outcome model but then re-weights to balance residuals. We discuss connections to inverse propensity score weighting and survey calibration in Appendix A.5.

Choice of estimator

While this setup is general, the choice of estimator $\hat{m}$ is important both for understanding the procedure's properties and for practical performance. We give a brief overview of two special cases: (1) when $\hat{m}$ is linear in the pre-treatment outcomes; and (2) when $\hat{m}$ is linear in the comparison units. Ridge regression is an important example that is linear in both pre-treatment outcomes and comparison units; we explore this estimator further in Sections 2.4 and 2.5.

First, consider an estimator that is linear in pre-treatment outcomes, $\hat{m}(\boldsymbol{X}) = \hat{\eta}_0 + \hat{\boldsymbol{\eta}} \cdot \boldsymbol{X}$.

The augmented estimator (2.9) is then:

$$\hat{Y}_{1T}^{\text{aug}}(0) = \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} Y_{iT} \;\; + \sum_{t=1}^{T_0} \hat{\eta}_t \left(X_{1t} - \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} X_{it} \right). \tag{2.11}$$

Pre-treatment periods that are more predictive of the post-treatment outcome will have larger (absolute) regression coefficients and so imbalance in these periods will lead to a larger adjustment. Thus, even if we do not *a priori* prioritize balance in any particular pre-treatment time periods (via the choice of $\boldsymbol{V_x}$), the linear model augmentation will adjust for the time periods that are empirically more predictive of the post-treatment outcome. As we show in Section 2.4, the ridge-regularized linear model is an important special case in which the resulting augmented estimator is itself a penalized synthetic control estimator. This allows for a more direct analysis of the role of bias correction.

Second, consider an estimator that is a linear combination of comparison units, $\hat{m}(\boldsymbol{X}) = \sum_{W_i=0} \hat{\alpha}_i(\boldsymbol{X}) Y_{iT}$, for some weighting function $\hat{\alpha} : \mathbb{R}^{T_0} \to \mathbb{R}^{N_0}$. Examples include k-nearest neighbor matching and kernel weighting as well as other "vertical" regression approaches (Athey et al., 2017). The augmented estimator (2.9) is itself a weighting estimator that adjusts the SCM weights:[3]

$$\hat{Y}_{1T}^{\text{aug}}(0) = \sum_{W_i=0} \left(\hat{\gamma}_i^{\text{scm}} + \hat{\gamma}_i^{\text{adj}} \right) Y_{iT}, \quad \text{where} \quad \hat{\gamma}_i^{\text{adj}} \equiv \hat{\alpha}_i(\boldsymbol{X}_1) - \sum_{W_j=0} \hat{\gamma}_j^{\text{scm}} \hat{\alpha}_i(\boldsymbol{X}_j). \tag{2.12}$$

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

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

$$\hat{\tau}^{\text{de}} = \left(Y_{1T} - \bar{X}_1\right) - \left(\sum_{W_i=0} \hat{\gamma}_i (Y_{iT} - \bar{X}_i) \right) = \frac{1}{T_0} \sum_{t=1}^{T_0} \left[(Y_{1T} - X_{1t}) - \left(\sum_{W_i=0} \hat{\gamma}_i (Y_{iT} - X_{it}) \right) \right]. \tag{2.13}$$

An augmented estimator of this form has appeared as the *de-meaned* or *intercept shift SCM* (Doudchenko and Imbens, 2017; Ferman and Pinto, 2018).[4] See also Arkhangelsky et al. (2019), who extend this to weight across both units and time.

In Section 2.7 we conduct a simulation study to inspect the performance of a range of estimators including: other penalized linear models, such as the LASSO; flexible machine

[3] We thank an anonymous reviewer for suggesting this presentation.

[4] In these proposals, the SCM weights balance the *residual* outcomes $X_{it} - \bar{X}_i$ rather than the raw outcomes X_{it}. We further consider balancing residuals in Section 2.6.

learning models, such as random forests; and panel data methods, such as fixed effects models and low-rank matrix completion methods (Xu, 2017; Athey et al., 2017).

2.4 Ridge ASCM: Numerical results

We now inspect the algorithmic and numerical properties for the special case where $\hat{m}(\boldsymbol{X}_i)$ is estimated via a ridge-regularized linear model, which we refer to as *Ridge Augmented SCM* (Ridge ASCM). With Ridge ASCM, the estimator for the post-treatment outcome is $\hat{m}(\boldsymbol{X}_i) = \hat{\eta}_0^{\text{ridge}} + \boldsymbol{X}_i'\hat{\boldsymbol{\eta}}^{\text{ridge}}$, where $\hat{\eta}_0^{\text{ridge}}$ and $\hat{\boldsymbol{\eta}}^{\text{ridge}}$ are the coefficients of a ridge regression of control post-treatment outcomes $\boldsymbol{Y}_{0T}$ on centered pre-treatment outcomes $\boldsymbol{X}_{0\cdot}$ with penalty hyper-parameter λ^{ridge}:[5]

$$\left\{\hat{\eta}_0^{\text{ridge}}, \hat{\boldsymbol{\eta}}^{\text{ridge}}\right\} = \arg\min_{\eta_0,\boldsymbol{\eta}} \ \frac{1}{2}\sum_{W_i=0}(Y_i - (\eta_0 + X_i'\boldsymbol{\eta}))^2 + \lambda^{\text{ridge}}\|\boldsymbol{\eta}\|_2^2. \tag{2.14}$$

The Ridge Augmented SCM estimator is then:

$$\hat{Y}_{1T}^{\text{aug}}(0) = \sum_{W_i=0}\hat{\gamma}_i^{\text{scm}}Y_{iT} + \left(\boldsymbol{X}_1 - \sum_{W_i=0}\hat{\gamma}_i^{\text{scm}}\boldsymbol{X}_{i\cdot}\right)\cdot\hat{\boldsymbol{\eta}}^{\text{ridge}}. \tag{2.15}$$

We first show that Ridge ASCM is a linear weighting estimator as in Equation (2.12). Unlike augmenting with other linear weighting estimators, when augmenting with ridge regression the implied weights are themselves the solution to a penalized synthetic control problem, as in Equation (2.8). Using this representation, we show that when the treated unit lies outside the convex hull of the control units, Ridge ASCM improves the pre-treatment fit relative to SCM alone by allowing for negative weights and extrapolating away from the convex hull. We also show that ridge regression alone has a representation as a weighting estimator that allows for negative weights.

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

Ridge ASCM as a penalized SCM estimator

We now express both Ridge ASCM and ridge regression alone as special cases of the penalized SCM problem in Equation (2.8). The Ridge ASCM estimate of the counterfactual is the so-

[5]Similar to the synthetic controls problem, we can regularize time periods differently with a generalized ridge penalty $\boldsymbol{\eta}'\boldsymbol{\Lambda}\boldsymbol{\eta}$ using an importance matrix $\boldsymbol{\Lambda}$. Following the typical case with diagonal elements, the generalized ridge penalty reduces to separate regularization on each time period.

lution to Equation (2.8), replacing the simplex constraint with a penalty $f(\gamma_i) = (\gamma_i - \hat{\gamma}_i^{\text{scm}})^2$ that penalizes *deviations from the SCM weights*.

Lemma 2.1. The ridge-augmented SCM estimator (2.11) is:

$$\hat{Y}_{1T}^{\text{aug}}(0) = \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} Y_{iT}, \tag{2.16}$$

where

$$\hat{\gamma}_i^{\text{aug}} = \hat{\gamma}_i^{\text{scm}} + (\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}})'(\boldsymbol{X}_{0\cdot}'\boldsymbol{X}_{0\cdot} + \lambda^{\text{ridge}}\boldsymbol{I}_{\boldsymbol{T_0}})^{-1}\boldsymbol{X}_{i\cdot}. \tag{2.17}$$

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

$$\min_{\boldsymbol{\gamma} \text{ s.t. } \sum_i \gamma_i = 1} \frac{1}{2\lambda^{\text{ridge}}} \|\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma}\|_2^2 + \frac{1}{2}\|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}^{\text{scm}}\|_2^2 . \tag{2.18}$$

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

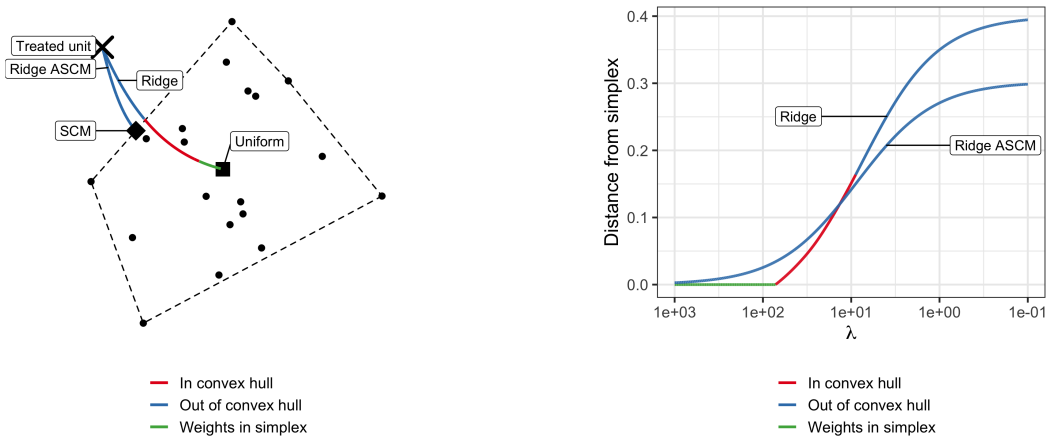
We can similarly characterize ridge regression alone as a solution to a penalized SCM problem where the penalty term, $f(\gamma_i) = \left(\gamma_i - \frac{1}{N_0}\right)^2$, penalizes the variance of the weights. Other penalized linear models, such as the LASSO or elastic net, do not have this same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

Lemma 2.2. The ridge regression estimator $\hat{Y}_{1T}^{\text{ridge}}(0) \equiv \hat{\eta}_0^{\text{ridge}} + \boldsymbol{X}_1 \cdot \hat{\boldsymbol{\eta}}^{\text{ridge}}$ can be written as $\hat{Y}_{1T}^{\text{ridge}}(0) = \sum_{W_i=0} \hat{\gamma}_i^{\text{ridge}} Y_{iT}$, where the ridge weights $\hat{\boldsymbol{\gamma}}^{\text{ridge}}$ are the solution to:

$$\min_{\boldsymbol{\gamma} \mid \sum_i \gamma_i = 1} \frac{1}{2\lambda^{\text{ridge}}} \|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\boldsymbol{\gamma}\|_2^2 + \frac{1}{2}\left\|\boldsymbol{\gamma} - \frac{1}{N_0}\right\|_2^2 . \tag{2.19}$$

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

Figure 2.1 visualizes this behavior in two dimensions. Figure 2.1a shows the treated unit outside the convex hull of the control units, along with the weighted average of control units using ridge regression and Ridge ASCM weights. For large λ^{ridge}, ridge regression alone begins at the center of the control units (i.e., uniform weights), while Ridge ASCM begins at the SCM solution; both move smoothly towards an exact fit solution as λ^{ridge} is reduced. Figure 2.1b shows the distance from the simplex of these ridge regression and Ridge ASCM weights.



(a) Treated and control units with the convex hull marked as a dashed line. Ridge and Ridge ASCM estimates marked as solid lines.

(b) Distance of ridge and Ridge ASCM weights from the simplex.

Figure 2.1: Ridge ASCM vs. ridge regression alone for a two-dimensional example with the treated unit outside of the convex hull of the control units. Results shown varying λ^{ridge} from 10^3 to 10^{-1}. Green denotes that the weights are inside the simplex, red that the weights are outside the simplex but the weighted average is inside the convex hull, and blue that the weighted average is outside the convex hull.

Together these figures highlight that ridge regression weights can leave the simplex (i.e., have some negative weights) before the corresponding weighted average is outside of the convex hull, marked in red in both figures. That is, ridge regression weights use negative weights to minimize the variance although it is possible to achieve the same level of balance with non-negative weights. By contrast, Ridge ASCM weights begin at the SCM solution, which is on the boundary of the simplex, then extrapolate outside the convex hull. Eventually, as $\lambda^{\text{ridge}} \to 0$, both ridge and Ridge ASCM use negative weights to achieve perfect balance, improving the fit relative to SCM alone. The weight vectors differ, however, with the Ridge ASCM weights closer to the simplex.

When achieving excellent pre-treatment fit with SCM is possible, Abadie et al. (2015) argue that we should prefer SCM weights over possibly negative weights: a slight balance improvement is not worth the extrapolation and the loss of interpretability. In this case, the Ridge ASCM weights will be close to the simplex, while the ridge regression weights may be quite far away. When this is not possible, however, and SCM has poor fit, some degree of extrapolation is critical; Ridge ASCM allows the researcher to directly penalize the amount of extrapolation in these cases.[6]

[6]See King and Zeng (2006) for a discussion of extrapolation in constructing counterfactuals. As they note: "If we learn that a counterfactual question involves extrapolation, we still might wish to proceed if the question is sufficiently important, but we would be aware of how much more model dependent our answers

Ridge ASCM improves pre-treatment fit relative to SCM alone

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

Lemma 2.3. Let $\frac{1}{\sqrt{N_0}}\boldsymbol{X}_{0\cdot} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$ be the singular value decomposition of the matrix of control pre-intervention outcomes, where m is the rank of $\boldsymbol{X}_{0\cdot}$, $\boldsymbol{U} \in \mathbb{R}^{N_0 \times m}$, $\boldsymbol{V} \in \mathbb{R}^{T_0 \times m}$, and $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_m) \in \mathbb{R}^{m \times m}$ is the diagonal matrix of singular values, where d_1 and d_m are the largest and smallest singular values, respectively. Furthermore, let $\tilde{\boldsymbol{X}}_i = \boldsymbol{V}'\boldsymbol{X}_i$ be the rotation of $\boldsymbol{X}_i$ along the singular vectors of $\boldsymbol{X}_{0\cdot}$. Then $\hat{\boldsymbol{\gamma}}^{\text{aug}}$, the Ridge ASCM weights with hyper-parameter $\lambda^{\text{ridge}} = \lambda N_0$ satisfy

$$\|\boldsymbol{X}_{1\cdot} - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{aug}}\|_2 = \lambda \left\| (\boldsymbol{D} + \lambda \boldsymbol{I})^{-1} \left(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}'\hat{\gamma}^{\text{scm}}\right) \right\|_2 \leq \frac{\lambda}{d_m^2 + \lambda} \|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\gamma}^{\text{scm}}\|_2, \quad (2.20)$$

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

$$\|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{ridge}}\|_2 = \lambda \left\| (\boldsymbol{D} + \lambda \boldsymbol{I})^{-1} \widetilde{\boldsymbol{X}}_1 \right\|_2 \leq \frac{\lambda}{d_m^2 + \lambda} \|\boldsymbol{X}_1\|_2. \quad (2.21)$$

From Equation (2.20), we see that the pre-treatment imbalance for Ridge ASCM weights is smaller than that of SCM weights by at least a factor of $\frac{\lambda}{d_m^2+\lambda}$. Thus, Ridge ASCM will achieve strictly better pre-treatment fit than SCM alone, except in corner cases where pre-treatment fit will be equal, such as when the pre-treatment SCM residual $\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{scm}}$ is orthogonal to the lagged outcomes of the control units $\boldsymbol{X}_{0\cdot}$. Since ridge regression penalizes deviations from uniformity, rather than deviations from SCM weights, the relationship for pre-treatment imbalance and fit between SCM and ridge regression alone is less clear.

2.5 Ridge ASCM: Estimation error

We now relate Ridge ASCM's improved pre-treatment fit to improved estimation error under the data generating processes in Section 2.2. Under a linear model, improving pre-treatment fit directly reduces bias, and the Ridge ASCM penalty term negotiates a bias-variance trade-off. Under a latent factor model, improving pre-treatment fit again reduces bias, though there is now a risk of over-fitting. The penalty term also directly parameterizes this trade-off. Thus, choosing the hyper-parameter λ^{ridge} is important in practice. We describe a cross-validation hyper-parameter selection procedure in Section 2.5. Finally, we discuss inference in Section 2.5.

would be."

Error under linearity

We first illustrate the key balancing idea in the simple case in our first DGP, where the post-treatment outcome is a linear combination of lagged outcomes plus additive noise, as in Assumption 2.2(a). We consider a generic weighting estimator with weights $\hat{\boldsymbol{\gamma}}$ that are independent of the post-treatment outcomes $Y_{1T}, \ldots, Y_{NT}$; both SCM and Ridge ASCM take this form. Under linearity, the difference between the counterfactual outcome $Y_{1T}(0)$ and the weighting estimator $\hat{Y}_{1T}(0)$ decomposes into: (1) systemic error due to imbalance in the lagged outcomes $\boldsymbol{X}$, and (2) idiosyncratic error due to the noise in the post-treatment period:

$$Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i Y_{iT} = \underbrace{\boldsymbol{\beta} \cdot \left(\boldsymbol{X}_1 - \sum_{W_i=0} \boldsymbol{X}_i \right)}_{\text{imbalance in } \boldsymbol{X}} + \underbrace{\varepsilon_{1T} - \sum_{W_i=0} \hat{\gamma}_i \varepsilon_{iT}}_{\text{post-treatment noise}}. \tag{2.22}$$

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

If the weighting estimator does not perfectly balance the pre-treatment outcomes $\boldsymbol{X}$, there will be a systematic component of the error, with the magnitude depending on the imbalance. Below we construct a finite sample error bound for Ridge ASCM (and for SCM, the special case with $\lambda^{\text{ridge}} = \infty$), building on Lemma 2.3. This bound on the estimation error holds with high probability over the noise in the post-treatment period $\boldsymbol{\varepsilon}_T$.

Proposition 2.1. Under the auto-regressive model in Assumption 2.2(a), for any $\delta > 0$ the Ridge ASCM weights with hyperparameter $\lambda^{\text{ridge}} = \lambda N_0$ satisfy the bound

$$\left| Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} Y_{iT} \right| \leq \|\boldsymbol{\beta}\|_2 \underbrace{\left\| \text{diag}\left(\frac{\lambda}{d_j^2 + \lambda} \right) (\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_0' \hat{\boldsymbol{\gamma}}^{\text{scm}}) \right\|_2}_{\text{imbalance in } \boldsymbol{X}} + \underbrace{\delta\sigma \left(1 + \|\hat{\boldsymbol{\gamma}}^{\text{aug}}\|_2\right)}_{\text{post-treatment noise}}, \tag{2.23}$$

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

Proposition 2.1 shows the finite sample error of Ridge ASCM weights is controlled by the imbalance in the lagged outcomes and the L^2 norm of the weights; Lemma A.3 in the supplementary materials gives a deterministic bound for $\|\hat{\boldsymbol{\gamma}}^{\text{aug}}\|_2$. See Athey et al. (2018) for analogous results on balancing weights in high dimensional cross-sectional settings.

In the special case that SCM weights have perfect pre-treatment fit, ASCM and SCM weights will be equivalent, and the estimation error will only be due to the variance of the weights and post-treatment noise. When SCM weights do not achieve perfect pre-treatment

fit, Ridge ASCM with finite λ extrapolates outside the convex hull, improving pre-treatment fit and thus reducing bias. This is subject to the usual bias-variance trade-off: The second term in (2.23) is increasing in the L^2 norm of the weights, which will generally be larger for ASCM than for SCM. The hyperparameter λ directly negotiates this trade off.

Error under a latent factor model

Following Abadie et al. (2010), we now consider the case where control potential outcomes are generated according to a linear factor model, as in Assumption 2.2(b): $Y_{it}(0) = \boldsymbol{\phi}_i \cdot \boldsymbol{\mu}_t + \varepsilon_{it}$. Under this model, the finite-sample error of a weighting estimator depends on the imbalance in the latent factors $\boldsymbol{\phi}$ and a noise term due to the noise at time T:

$$Y_{1T}(0) - \hat{Y}_{1T}(0) = Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i Y_{iT} = \underbrace{\left(\boldsymbol{\phi}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{\phi}_i\right) \cdot \boldsymbol{\mu}_T}_{\text{imbalance in } \boldsymbol{\phi}} + \underbrace{\varepsilon_{1T} - \sum_{W_i=0} \hat{\gamma}_i \varepsilon_{it}}_{\text{noise}}. \quad (2.24)$$

Balancing the observed pre-treatment outcomes $\boldsymbol{X}$ will not necessarily balance the latent factor loadings $\boldsymbol{\phi}$. Following Abadie et al. (2010), we show in the supplementary materials that, under Equation (2.7), we can decompose the imbalance term as:

$$\left(\boldsymbol{\phi}_1 - \sum_{W_i=0} \gamma_i \boldsymbol{\phi}_i\right) \cdot \boldsymbol{\mu}_T = \frac{1}{T_0} \boldsymbol{\mu}' \underbrace{\left(\boldsymbol{X}_1 - \sum_{W_i=0} \gamma_i \boldsymbol{X}_i\right)}_{\text{imbalance in } \boldsymbol{X}} \cdot \mu_T - \frac{1}{T_0} \boldsymbol{\mu}' \underbrace{\left(\boldsymbol{\varepsilon}_{1(1:T_0)} - \sum_{W_i=0} \gamma_i \boldsymbol{\varepsilon}_{i(1:T_0)}\right)}_{\text{approximation error}} \cdot \boldsymbol{\mu}_T, \quad (2.25)$$

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

With this setup, we can bound the finite-sample error in Equation (2.24) for Ridge ASCM weights (and for SCM weights as a special case). This bound is with high probability over the noise in all time periods ε_{it}, and accounts for the noise in the pre- and post-treatment outcomes separately.

Theorem 2.1. Under the linear factor model in Assumption 2.2(b), for any $\delta > 0$ the Ridge

ASCM weights with hyperparameter $\lambda^{\text{ridge}} = \lambda N_0$ satisfy the bound

$$\left| Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} Y_{1T}(0) \right| \leq \frac{JM^2}{\sqrt{T_0}} \Bigg(\underbrace{\left\| \text{diag}\left(\frac{\lambda}{d_j^2 + \lambda} \right) (\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_0' \hat{\boldsymbol{\gamma}}^{\text{scm}}) \right\|_2}_{\text{imbalance in } \boldsymbol{X}} + \underbrace{4(1+\delta) \left\| \text{diag}\left(\frac{d_j \sigma}{d_j^2 + \lambda} \right) (\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_0' \hat{\gamma}^{\text{scm}}) \right\|_2}_{\text{excess approximation error}} + \underbrace{2\sigma \left(\sqrt{\log 2N_0} + \frac{\delta}{2} \right)}_{\text{SCM approximation error}} \Bigg) + \underbrace{\delta\sigma \left(1 + \|\hat{\boldsymbol{\gamma}}^{\text{aug}}\|_2\right)}_{\text{post-treatment noise}} \tag{2.26}$$

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

Theorem 2.1 shows that, relative to the linear case in Proposition 2.1, there is an additional source of error under a latent factor model: approximation error due to balancing lagged outcomes rather than balancing underlying factors. In particular, it is now possible that a control unit only receives a large weight because of idiosyncratic noise, rather than because of similarity in the underlying factors. See Arkhangelsky et al. (2019) and Ferman (2019) for asymptotic analogues of this finite sample bound. As we discuss below, each of the first three terms of the bound in Theorem 2.1 are directly computable from the observed data, save for the unknown σ parameter.

In the special case where SCM achieves perfect pre-treatment fit, considered by Abadie et al. (2010), the ASCM and SCM weights are equivalent and the error is only due to post-treatment noise and the approximation error. The bound in Theorem 2.1 accounts for the worst case scenario where the control unit with the largest weight is only similar to the treated unit due to idiosyncratic noise. The approximation error, and thus the bias, converges to zero in probability as $T_0 \to \infty$ under suitable conditions on the factor loadings $\boldsymbol{\mu}_t$ (see also Ferman and Pinto, 2018). Intuitively, as we observe more X_{it} — and can exactly balance each one — we are better able to match on the index $\boldsymbol{\phi}_i \cdot \boldsymbol{\mu}_t$ and, as a result, on the underlying factor loadings.[7]

Without exact balance, Theorem 2.1 shows that a long pre-period may not be enough to control the error due to imbalance. In this case, Ridge ASCM with $\lambda < \infty$ will extrapolate outside the convex hull, reducing error due to imbalance in the lagged outcomes but possibly

[7]We show in the supplementary material that with dependent errors the probability of the worst-case error additionally scales with the maximum eigenvalue of the covariance matrix. Dependence leads to a more complicated error structure overall; we leave a thorough analysis of this to future work.

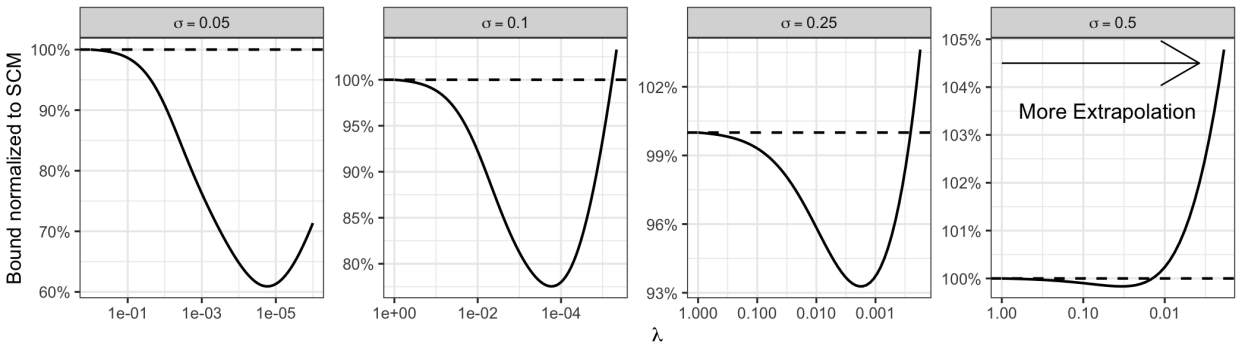


Figure 2.2: Sketch of the error due to imbalance and approximation error (2.26) for the linear factor model; the standard deviation of the treated unit's pre-treatment outcomes is normalized to one. We fit SCM weights on the empirical example in Section 2.7 and compute the vector of pre-treatment fit. Each line shows the sum of the error due to imbalance in $\boldsymbol{X}$, excess approximation error, and SCM approximation error in Theorem 2.1 (with $\delta = 0$) for different values of σ. These are normalized so that the SCM solution (with λ large) equals 100%; values below 100% show improvement over the unadjusted weights for a given λ.

over-fitting to noise. Thus, the optimal level of extrapolation will depend on the synthetic control fit and the amount of noise.

Figure 2.2 illustrates this using SCM weights from the empirical example we discuss in Section 2.7, where pre-treatment fit is good but not perfect. For each value of σ, the figure plots the sum of the imbalance, SCM approximation error, and excess approximation error terms in the bound in Theorem 2.1, all directly computable from the data for a given σ. At each noise level, a small amount of extrapolation leads to a smaller error bound, but as λ shrinks there is a point where further extrapolation leads to over-fitting and eventually to a worse error bound than without extrapolation. The risk of overfitting is greater when the noise is large (e.g., $\sigma = 0.5$), though even here a sufficiently regularized ASCM estimate has a lower error bound than SCM alone (represented as the $\lambda \to \infty$ bound at the left boundary). When noise is less extreme, the benefits of augmentation are larger and the optimal amount of regularization shrinks.

It is worth noting that Theorem 2.1 gives a worst-case bound. In Section 2.7 we inspect the typical performance of the Ridge ASCM estimator via extensive simulation studies and find that gains to pre-treatment fit through augmentation outweigh increased approximation error in a range of practical settings, including when noise is very large.

Theorem 2.1 suggests two diagnostics to supplement the estimated bias from Equation (2.9), based on the first two terms in the bound. For the first term, we can directly assess imbalance in $\boldsymbol{X}$ via the pre-treatment RMSE, $\frac{1}{\sqrt{T_0}}\|\boldsymbol{X}_1 - \boldsymbol{X}_0'\hat{\gamma}^{\text{aug}}\|_2$. For the second term, the excess approximation error depends on the unknown noise level, σ. However, as we show in the supplementary materials, the excess approximation error is a scaled version of the root mean square distance between the Ridge ASCM weights and the SCM weights,

$\frac{1}{\sqrt{N_0}}\|\hat{\gamma}^{\text{aug}} - \hat{\gamma}^{\text{scm}}\|_2$, which is a measure of extrapolation. We report these diagnostics for the empirical application in Section 2.7. As Figure 2.2 previews, they support the use of ASCM in this instance, despite what visually appears to be good pre-treatment fit for SCM.

Hyper-parameter selection

We propose a cross-validation approach for selecting λ inspired by the in-time placebo check of Abadie et al. (2015). Let $\hat{Y}_{1t}^{(-k)} = \sum_{W_i=0} \hat{\gamma}_{i(-k)}^{\text{aug}} Y_{it}$ be the estimate of Y_{1t} where time period k is excluded from fitting the estimator in (2.17). Abadie et al. (2015) propose to compare the difference $Y_{1t} - \hat{Y}_{1t}^{(-t)}$ for some $t \leq T_0$ as a placebo check. We can extend this idea to compute the leave-one-out cross validation MSE over time periods:

$$CV(\lambda) = \sum_{t=1}^{T_0} \left(Y_{1t} - \hat{Y}_{1t}^{(-t)} \right)^2. \tag{2.27}$$

We can then choose λ to minimize $CV(\lambda)$ or follow a more conservative approach such as the "one-standard-error" rule (Hastie et al., 2009). This proposal is similar to the leave-one-out cross validation proposed by Doudchenko and Imbens (2017), who select hyperparameters by holding out control units and minimizing the MSE of the control units in the post-treatment time T. Finally, only excluding time period t might be inappropriate for some outcome models, e.g. the linear model in Section 2.5. In these settings we can extend the procedure to exclude all time periods $\geq t$ when estimating $\hat{\gamma}_{(-t)}^{\text{aug}}$, as in Kellogg et al. (2020).

Inference

There is a growing literature on inference for the synthetic control method and variants, going beyond the original proposal in Abadie and Gardeazabal (2003) and Abadie et al. (2010, 2015); see, for example, Li (2017), Toulis and Shaikh (2018), Cattaneo et al. (2019), and Chernozhukov et al. (2018).

We focus here on the conformal inference approach of Chernozhukov et al. (2019), which has three key steps. First, for a given sharp null hypothesis, $H_0 : \tau = \tau_0$, we create an adjusted post-treatment outcome for the treated unit $\tilde{Y}_{1T} = Y_{1T} - \tau_0$ and extend the original data set to include the adjusted outcome $\tilde{Y}_{1T}$. Second, we apply the estimator (2.17) to the extended dataset to obtain adjusted weights $\hat{\boldsymbol{\gamma}}(\tau_0)$. Finally, we compute a p-value by assessing whether the adjusted residual $Y_{1T} - \tau_0 - \sum_{W_i=0} \hat{\gamma}_i(\tau_0) Y_{iT}$ "conforms" with the pre-treatment residuals:[8]

$$p(\tau_0) = \frac{1}{T} \sum_{t=1}^{T_0} \mathbb{1}\left\{ \left| Y_{1T} - \tau_0 - \sum_{W_i=0} \hat{\gamma}_i(\tau_0) Y_{iT} \right| \leq \left| Y_{1t} - \sum_{W_i=0} \hat{\gamma}_i(\tau_0) Y_{it} \right| \right\} + \frac{1}{T}. \tag{2.28}$$

[8] Chernozhukov et al. (2019) consider several choices of test statistic and permutation distributions across time periods. For a single post treatment time their main proposals reduce to Equation (2.28).

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

$$\widehat{C}_Y^{\text{conf}} = \left\{ y \in \mathbb{R} \;\middle|\; \left| y - \sum_{W_i=0} \hat{\gamma}_i (Y_{1T} - y) Y_{iT} \right| \leq q^+_{T,\alpha} \left(\left| Y_{1t} - \sum_{W_i=0} \hat{\gamma}_i (Y_{1T} - y) Y_{it} \right| \right) \right\}, \quad (2.29)$$

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

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

When the residuals are not exchangeable, Chernozhukov et al. (2019) provide a finite sample bound that relates in-sample prediction error to the validity of $p(\tau_0)$. In Appendix A.1, we adapt their SCM bounds to Ridge ASCM by showing that the ridge penalty controls the difference between SCM and Ridge ASCM weights. Under a variant of the basic model (2.3), the resulting p-value will be valid as the number of pre-treatment periods $T_0 \to \infty$. Finally, in Section 2.7 we explore the finite sample coverage probabilities of $\widehat{C}_Y^{\text{conf}}$ under various data generating processes and find that they are near their nominal levels.

2.6 Auxiliary covariates

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

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

$$\min_{\gamma \in \Delta^{N_0}} \quad \theta_x \|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\gamma\|_2^2 + \theta_z \|\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}\gamma\|_2^2 \; + \; \zeta \sum_{W_i=0} f(\gamma_i), \quad (2.30)$$

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

$$\min_{\eta_0, \boldsymbol{\eta_x}, \boldsymbol{\eta_z}} \quad \frac{1}{2} \sum_{W_i=0} (Y_i - (\eta_0 + \boldsymbol{X}_i'\boldsymbol{\eta_x} + \boldsymbol{Z}_i'\boldsymbol{\eta_z}))^2 + \lambda_x \|\boldsymbol{\eta_x}\|_2^2 + \lambda_z \|\boldsymbol{\eta_z}\|_2^2. \quad (2.31)$$

Both this SCM criterion and augmentation estimator incorporate user-specified weights that determine the importance of balancing each set of covariates (Equation 2.30) or the amount of regularization for each set of coefficients (Equation 2.31). There are many potential choices for these weights. We discuss two, appropriate to different settings depending on the number of auxiliary covariates.

A sensible default when the dimension of the auxiliary covariates is moderate is to incorporate the lagged outcomes $\boldsymbol{X}$ and the auxiliary covariates $\boldsymbol{Z}$ equally in Equations (2.30) and (2.31), setting $\theta_x = \theta_z = 1$ and $\lambda_x = \lambda_z = \lambda^{\text{ridge}}$ (after standardizing auxiliary covariates and lagged outcomes to have equal variance). With this setup the numerical and algorithmic results in Section 2.4 apply for the combined vector of lagged outcomes and auxiliary covariates, $(\boldsymbol{X}_i, \boldsymbol{Z}_i) \in \mathbb{R}^{T_0+K}$. In particular, Ridge ASCM is again a penalized SCM estimator that adjusts the synthetic control weights that solve optimization problem (2.30) to achieve better balance by extrapolating outside of the convex hull.

An alternative approach when the dimension of the auxiliary covariates is small relative to N (i.e., $K \ll N$) is to fit a regression model that regularizes the lagged outcome coefficients $\boldsymbol{\eta_x}$ but does *not* regularize the auxiliary covariate coefficients $\boldsymbol{\eta_z}$ (i.e., set $\lambda_z = 0$). Lemma 2.4 below writes the resulting augmented estimator as its corresponding penalized SCM optimization problem, with weights that perfectly balance the auxiliary covariates. This has two key implications. First, since the auxiliary covariates $\boldsymbol{Z}$ are exactly balanced regardless of the balance that the SCM weights achieve alone, we can exclude them from the optimization problem (2.30). Second, as we show below, the pre-treatment fit on the lagged outcomes depends on how well the SCM weights balance the residualized lagged outcomes $\check{\boldsymbol{X}}$. This suggests modifying Equation (2.30) to balance $\check{\boldsymbol{X}}$ rather than the lagged outcomes $\boldsymbol{X}$, which leads to the two-step procedure: (1) residualize the pre- and post-treatment outcomes on the auxiliary covariates $\boldsymbol{Z}$; and (2) estimate Ridge ASCM on the residualized outcomes. This two-step procedure follows from a related proposal in Doudchenko and Imbens (2017).

Lemma 2.4. Let $\hat{\boldsymbol{\eta}}_x$ and $\hat{\boldsymbol{\eta}}_z$ be the solutions to (2.31) with $\lambda_x = \lambda^{\text{ridge}}$ and $\lambda_z = 0$. For any weight vector $\hat{\boldsymbol{\gamma}}$ that sums to one, the ASCM estimator from Equation (2.10) with $\hat{m}(\boldsymbol{X}_i, \boldsymbol{Z}_i) = \boldsymbol{X}_i'\hat{\boldsymbol{\eta}}_{\boldsymbol{x}} + \boldsymbol{Z}_i'\hat{\boldsymbol{\eta}}_{\boldsymbol{z}}$ is

$$\sum_{W_i=0} \hat{\gamma}_i Y_{iT} \;+ \left(\boldsymbol{X}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{X}_i\right)' \hat{\boldsymbol{\eta}}_{\boldsymbol{x}} \;+ \left(\boldsymbol{Z}_1 - \sum_{W_i=0} \hat{\gamma}_i \boldsymbol{Z}_i\right)' \hat{\boldsymbol{\eta}}_{\boldsymbol{z}} = \sum_{W_i=0} \hat{\gamma}_i^{\text{cov}} Y_{iT}, \tag{2.32}$$

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

$$\hat{\gamma}_i^{\text{cov}} = \hat{\gamma}_i + (\check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot})(\check{\boldsymbol{X}}_{0\cdot}'\check{\boldsymbol{X}}_{0\cdot} + \lambda^{\text{ridge}}\boldsymbol{I}_{T_0})^{-1}\check{\boldsymbol{X}}_i + (\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}'\gamma)'(\boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot})^{-1}\boldsymbol{Z}_i, \tag{2.33}$$

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

$$\check{\boldsymbol{X}}_i = \boldsymbol{X}_i - \boldsymbol{Z}_i'(\boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot})^{-1}\boldsymbol{Z}_{0\cdot}'\boldsymbol{X}_{0\cdot}. \tag{2.34}$$

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

$$\|\boldsymbol{X}_1 - \boldsymbol{X}_{0\cdot}'\hat{\boldsymbol{\gamma}}^{\text{cov}}\|_2 \leq \left(\frac{\lambda^{\text{ridge}}}{\lambda^{\text{ridge}} + N_0\check{d}_r^2}\right) \left\|\check{\boldsymbol{X}}_1 - \check{\boldsymbol{X}}_{0\cdot}'\hat{\boldsymbol{\gamma}}\right\|_2, \tag{2.35}$$

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

Comparing to the numerical results in Section 2.4, Lemma 2.4 shows that the two-step approach penalizes extrapolation from the convex hull *in the residualized space* $\check{\boldsymbol{X}}$, rather than in the lagged outcomes themselves. In essence, by residualizing out the auxiliary covariates $\boldsymbol{Z}$, the two-step approach allows for a possibly large amount of extrapolation in the auxiliary covariates, while carefully penalizing extrapolation in the part of the lagged outcomes that is orthogonal to the covariates.

In Appendix A.2, we consider the performance of this estimator when the outcomes follow a linear factor model with either a linear or a non-linear dependence on auxiliary covariates, focusing on the special case where $\lambda^{\text{ridge}} \to \infty$ and the weights $\hat{\boldsymbol{\gamma}}^{\text{cov}}$ do not extrapolate from the convex hull after residualization. When covariates enter linearly and when K is small relative to N_0, we show that exactly balancing a small number of auxiliary covariates and targeting imbalance in the residuals $\check{\boldsymbol{X}}$ decreases error due to pre-treatment fit. When covariates enter non-linearly, however, there is additional approximation error due to the linear regression specification. Thus, it is important to appropriately transforming the covariates in practice. Furthermore with larger numbers of covariates, the approach that incorporates them in parallel to lagged outcomes will be more appropriate.

2.7 Simulations and empirical illustrations

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

Calibrated simulation studies

We now present simulation studies calibrated to our empirical illustration in Section 2.7. Specifically, we use the Generalized Synthetic Control Method (Xu, 2017) to estimate a factor model with three latent factors based on the series of log GSP per capita ($N = 50$, $T_0 = 89$). We then simulate outcomes using the distribution of estimated parameters and model selection into treatment as a function of the latent factors; see Appendix A.3 for additional details. We also present results from three additional DGPs, each calibrated to estimates from the same data: (1) the factor model with quadruple the standard deviation of the noise term, (2) a unit and time fixed effects model, and (3) an autoregressive model with 3 lags.

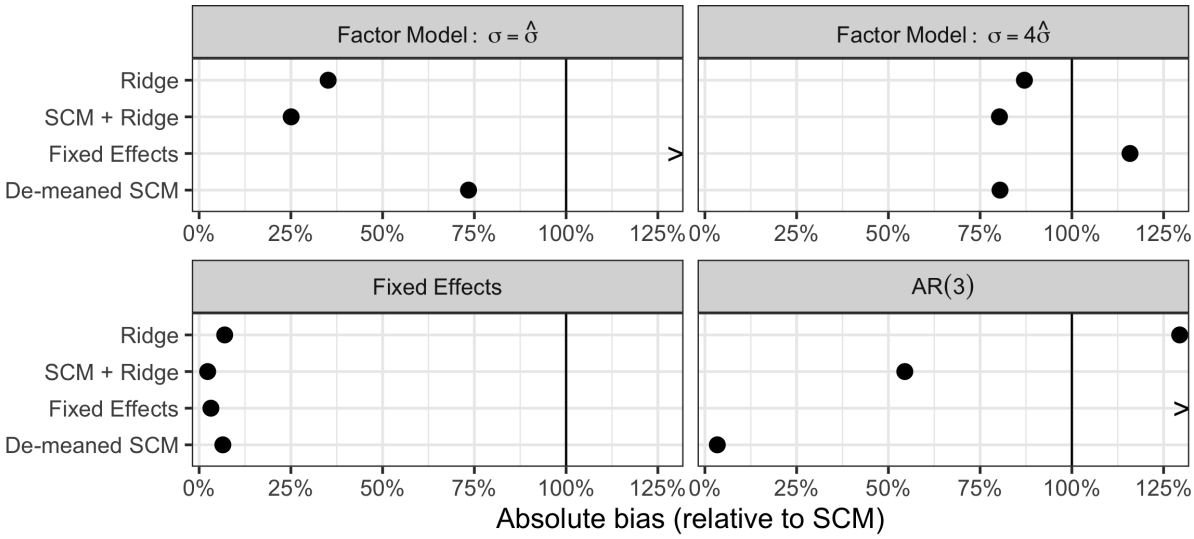


Figure 2.3: Overall absolute bias, normalized to SCM bias for (a) the factor model simulation, (b) the factor model simulation with quadruple the standard deviation, (c) the fixed effects simulation, and (d) the AR simulation. The SCM estimates reported here are *not* restricted to simulation draws with excellent pre-treatment fit; Abadie et al. (2015) advise against using SCM in such settings.

We explore the role of augmentation using different outcome estimators. For each DGP, we consider five estimators: (1) SCM alone, (2) ridge regression alone, (3) Ridge ASCM, (4) fixed effects alone, and (5) De-meaned SCM (i.e., SCM augmented with fixed effects) from Doudchenko and Imbens (2017) and Ferman and Pinto (2018), as shown in Equation (2.13).[9] Figure 2.3 shows the Monte Carlo estimate of the absolute bias as a percentage of the absolute bias for SCM, with one panel for each simulation DGP; Appendix Figure A.1 shows the corresponding estimator root mean squared error (RMSE).

There are several takeaways. First, augmenting SCM with a ridge outcome regression reduces bias relative to SCM alone — *without* conditioning on excellent pre-treatment fit — in all four simulations. This underscores the importance of the recommendation in Abadie et al. (2010, 2015) to use SCM only in settings with excellent pre-treatment fit.[10] Under the baseline factor model and the fixed effect model, the ridge augmentation greatly reduces

[9] Additional simulations shown in Appendix A.6 also consider alternative outcome models for use in ASCM: (1) LASSO, (2) a random forest, (3) `CausalImpact` (Brodersen et al., 2015), (4) matrix completion using `MCPanel` (Athey et al., 2017) and (5) fitting the factor model directly with `gsynth` (Xu, 2017). The results are consistent with those for Ridge ASCM, with meaningful gains from augmentation relative to SCM alone.

[10] Abadie et al. (2010, 2015) also strongly recommend incorporating auxiliary covariates, weighted by their predictive power, into the procedure, noting that this is important for further reducing bias. For simplicity, the simulations do not include auxiliary covariates.

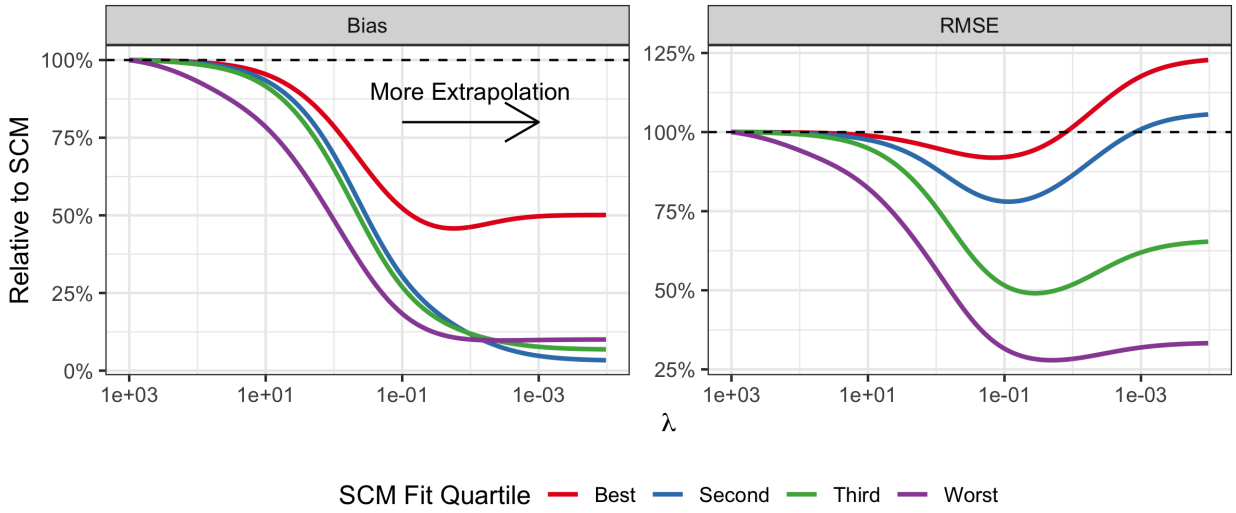


Figure 2.4: Bias and RMSE of Ridge ASCM, as a percentage of SCM bias and RMSE, versus λ under a linear factor model. Results are divided by the quartile of the SCM fit across all simulations.

bias, by more than 75% in the factor model simulation and over 90% in the fixed effects simulation. In the AR(3) model and in the factor model with greater noise, the gains to augmentation relative to SCM are more limited. Second, Ridge ASCM has lower bias than ridge regression alone across all of the simulation settings. Third, when the fixed effects estimator is incorrectly specified, combining it with SCM has much lower bias than either method alone. And even when the fixed effects estimator is correctly specified, de-meaned SCM has similar bias to the (correctly specified) fixed effects approach. Finally, Appendix Figure A.1 shows that in all simulations ASCM has lower RMSE than SCM, as the large decrease in bias more than makes up for the slight increase in variance.

Complementing the worst-case analysis in Section 2.5, we now consider how the typical performance of augmentation relates to the amount of extrapolation and the quality of the original SCM fit. Figure 2.4 shows the bias and RMSE as a function of λ for the primary factor model simulation, conditional on the quartile of SCM fit. Larger values of λ (and hence smaller adjustments) are to the left, with the left-most points in the plots representing SCM. First, as expected, Augmented SCM substantially reduces bias regardless of SCM pre-treatment fit. However, the gains are more modest when the SCM fit is in the best quartile: in this case the bias is non-monotonic in λ and there is some optimal choice of λ that minimizes the bias. Second, it is possible to under-regularize with ASCM, as evident in the RMSE achieving a minimum for an intermediate value of λ. When pre-treatment fit is good, augmentation with too-small λ leads to higher RMSE than SCM alone. However, when SCM fit is relatively poor, even minimally regularized ASCM achieves much better

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

Table 2.1: Coverage for 95% conformal prediction intervals (2.29) based on 1000 repetitions.

bias and RMSE than does SCM.

Finally, Table 2.1 shows the finite sample coverage of the conformal prediction intervals for $Y_{1T}(0)$. For the four simulation settings we compute 95% prediction intervals for the post-treatment counterfactual outcome $Y_{1T}(0)$ using the both the SCM and ridge ASCM estimators. We see that the intervals for SCM alone can slightly undercover, due to finite sample bias from poor treatment fit. In contrast, the intervals for ridge ASCM have close to nominal coverage for $Y_{1T}(0)$.

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

Illustration: 2012 Kansas tax cuts

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

We estimate the effect of the tax cuts on log gross state product (GSP) per capita using the second quarter of 2012 — when Brownback signed the tax cut bill into law — as the intervention time. We use four primary estimators: (1) SCM alone fit on the entire vector of lagged outcomes, (2) Ridge ASCM, (3) Ridge ASCM including auxiliary covariates in parallel to lagged outcomes and (4) Ridge ASCM on residualized outcomes, as proposed in Section 2.6.[11] These estimators rely on the mean-zero noise Assumption 2.1. Substantively, under

[11]The covariates we include are the pre-treatment averages of (1) log state and local revenue per capita, (2) log average weekly wages, (3) number of establishments per capita, (4) the employment level, and (5) log

the auto-regressive model in Assumption 2.2(a) this assumes that post-treatment shocks for Kansas will be the same as for other states in expectation; under the linear factor model in Assumption 2.2(b) this rules out selection on pre-treatment shocks. This also rules out unobserved confounders that affect both post-treatment shocks and the decision to enact the Brownback tax cut bill.

Figure 2.5, known as a "gap plot", shows the difference between Kansas and its synthetic control for all four estimators, along with 95% point-wise confidence intervals intervals computed via the conformal inference procedure from Chernozhukov et al. (2019). Figure 2.6 shows the log GSP per capita for both Kansas and its synthetic control using SCM and Ridge ASCM. Appendix A.6 shows additional results.

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

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

Next we consider including the auxiliary covariates. Adding these auxiliary covariates and augmenting further improves both pre-treatment fit and balance on the covariates; see Figure 2.7a. Finally, balancing the auxiliary covariates via residualization also improves pre-treatment fit. Overall, the estimated impact is consistently negative for all four approaches, with weaker evidence that the effect persists to the end of the observation period.

To check against over-fitting, Appendix Figures A.10, A.11, and A.12 show in-time placebo estimates for SCM alone, Ridge ASCM, and Ridge ASCM with covariates, with placebo treatment times in the second quarter of 2009, 2010, and 2011. We estimate placebo effects that are near zero with all three placebo treatment times with all three estimators.

Figure 2.7a shows the covariate balance for the four estimators. While SCM and Ridge ASCM achieve excellent fit for the pre-treatment average log GSP per capita, neither esti-

GSP per capita. For the augmented estimator on the lagged outcomes we select the hyperparameter λ^{ridge} as the largest λ within one standard error of the λ that minimizes the cross-validation placebo fit $CV(\lambda)$; see Section 2.5. Appendix Figure A.6 plots $CV(\lambda)$. When including the auxiliary covariates we use the minimal λ. Results are consistent using outcomes scales other than the standard normalization of log GSP per capita (see Appendix A.6).

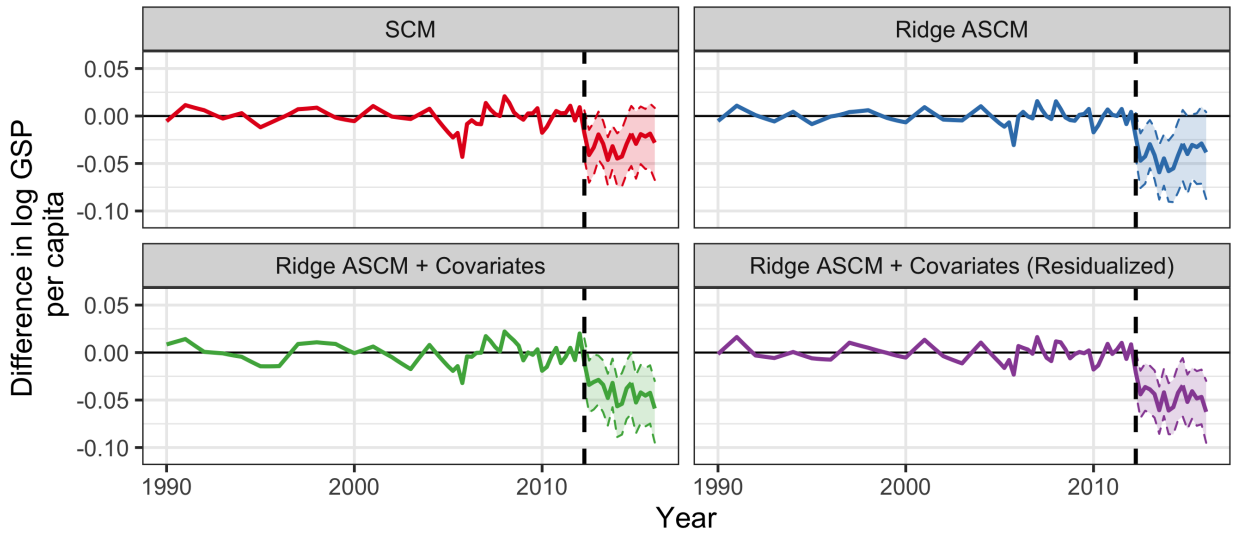


Figure 2.5: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on log GSP per capita using SCM, Ridge ASCM, and Ridge ASCM with covariates.

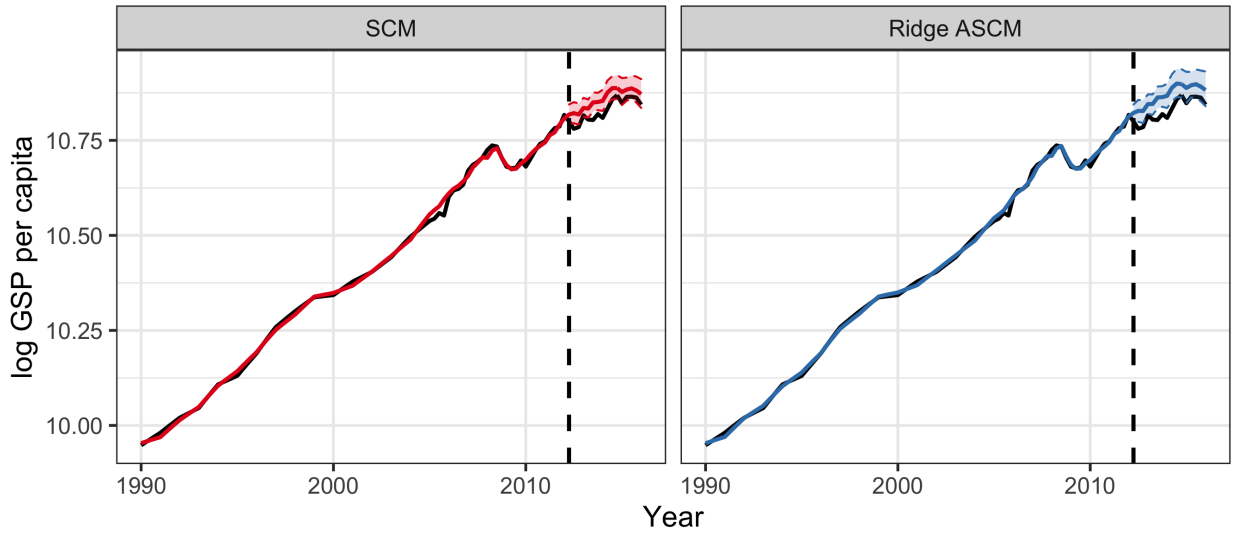
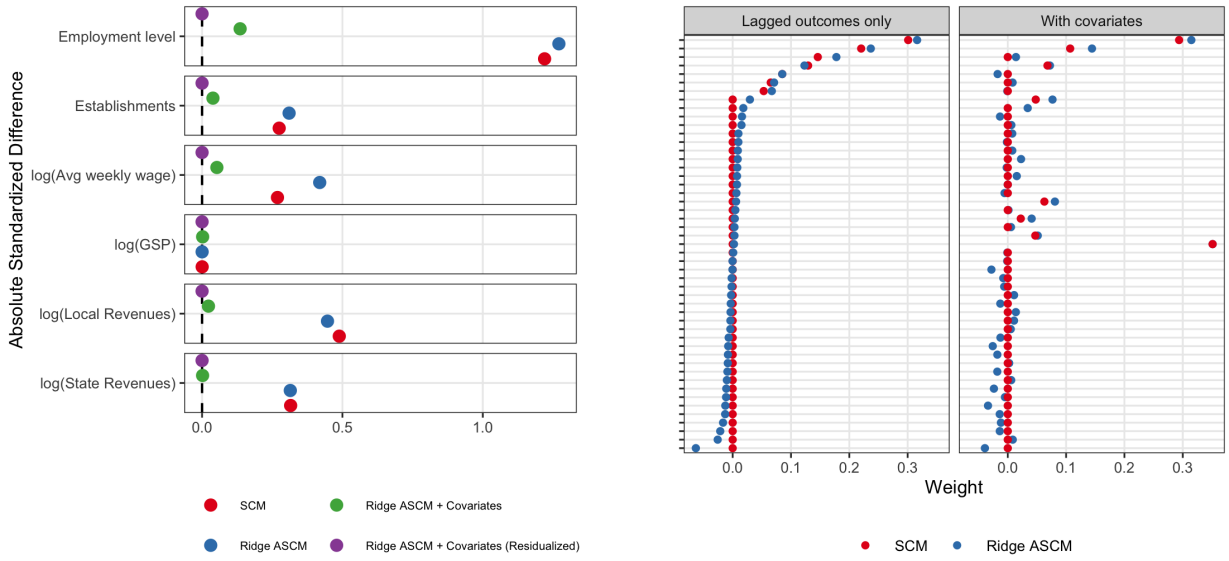


Figure 2.6: Point estimates along with point-wise 95% conformal prediction intervals for counterfactual log GSP per capita without the tax cuts using SCM, ridge ASCM, and ridge ASCM with covariates, plotting with the observed log GSP per capita in black.



(a) (b)

Figure 2.7: (a) Covariate balance for SCM, Ridge ASCM, and ASCM with covariates. Each covariate is standardized to have mean zero and standard deviation one; we plot the absolute difference between the treated unit's covariate and the weighted control units' covariates $\left|Z_{1k} - \sum_{W_i=0} \hat{\gamma} Z_{ik}\right|$. (b) Donor unit weights for (1) SCM alone and (2) Ridge ASCM; left facet uses lagged outcomes only; right fact includes auxiliary covariates.

mator achieves good balance on the other covariates, most notably the average employment level across the quarters of the pre-period. In contrast, including the auxiliary covariates into both the SCM and ridge optimization problems greatly improves the covariate balance, and — by design — residualizing on the auxiliary covariates perfectly balances them. Moreover, Ridge ASCM on residualized outcomes achieves very good pre-treatment fit on the lagged outcomes as shown in Figure 2.5.

Finally, Figure 2.7b shows the weights on donor units for SCM and Ridge ASCM as well as SCM and Ridge ASCM weights when including covariates jointly with the lagged outcomes (see also Appendix Figure A.14). Here we see the minimal extrapolation property of the ASCM weights. The SCM weights are zero for all but six donor states. The Ridge ASCM weights are similar but deviate slightly from the simplex. As a result, the Ridge ASCM weights retain some of the interpretability of the SCM weights. For the donor units with positive SCM weight, Ridge ASCM places close to the same weight. For the majority of those with zero SCM weight, Ridge ASCM also places a close to zero weight. Only Louisiana receives a meaningful negative weight, with non-negligible negative weights for only a few other donor units. By contrast, Appendix Figure A.13 shows the weights from ridge regression alone: many of the weights are negative and the weights are far from sparse. Including auxiliary covariates changes the relative importance of different states by adding

new information, but the minimal extrapolation property remains.

2.8 Discussion

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

There are several directions for future work. First, we could incorporate a sensitivity analysis that directly parameterizes departures from, say, the linear factor model, as in recent approaches for sensitivity analysis for balancing weights (Soriano et al., 2020). Second, we can adapt the ASCM framework to settings with multiple treated units. For instance, there are different approaches in settings when all treated units are treated at the same time: some papers propose to fit SCM separately for each treated unit (e.g., Abadie and L'Hour, 2018), while others simply average the units together (e.g., Robbins et al., 2017). The situation is more complicated with staggered adoption, when units take up the treatment at different times; we explore this extension next in Chapter 3. Finally, we can consider more complex data structures, such as applications with multiple outcomes series for the same units (e.g., measures of both earnings and total employment in minimum wage studies); hierarchical data structures with outcome information at both the individual and aggregate level (e.g., students within schools); or discrete or count outcomes.

Chapter 3

Synthetic Controls with Staggered Adoption

Staggered adoption of policies by different units at different times creates promising opportunities for observational causal inference. The synthetic control method (SCM) is a recent addition to the evaluation toolkit but is designed to study a single treated unit and does not easily accommodate staggered adoption. In this chapter, we generalize SCM to the staggered adoption setting. Current practice involves fitting SCM separately for each treated unit and then averaging. We show that the average of separate SCM fits does not necessarily achieve good balance for the average of the treated units, leading to possible bias in the estimated effect. We propose "partially pooled" SCM weights that instead minimize both average and state-specific imbalance, and show that the resulting estimator controls bias under a linear factor model. We also combine our partially pooled SCM weights with traditional fixed effects methods to obtain an augmented estimator that improves over both SCM weighting and fixed effects estimation alone. We assess the performance of the proposed method via extensive simulations and apply our results to the question of whether teacher collective bargaining leads to higher school spending, finding minimal impacts. We implement the proposed method in the `augsynth` R package.

3.1 Introduction

Jurisdictions often adopt policies at different times, creating promising opportunities for observational causal inference. In our motivating application, 33 states passed laws between 1964 and 1987 mandating that school districts bargain with teachers unions (Hoxby, 1996; Paglayan, 2019); our goal is to estimate the impact of these laws on student expenditures and teacher salaries.

Estimating causal effects under staggered adoption remains challenging, however. One promising approach is to use the *synthetic control method* (SCM; Abadie et al., 2010, 2015). Developed for the case with a single treated unit, SCM estimates the counterfactual untreated

outcome via a weighted average of untreated units, with weights chosen to match the treated unit's pre-treatment outcomes as closely as possible. Some applied researchers have used SCM in staggered adoption settings by estimating SCM weights separately for each treated unit and then averaging the estimates (see, e.g., Dube and Zipperer, 2015; Donohue et al., 2019). This approach is not well understood, however, especially in applications like ours where some treated units have poor SCM fits.

We develop SCM for the staggered adoption setting. We first consider two immediate adaptations: *separate SCM*, which reflects the current practice of estimating weights that separately minimize the pre-treatment imbalance for each treated unit; and *pooled SCM*, which instead estimates weights that minimize the average pre-treatment imbalance across all treated units. Both approaches have drawbacks. Separate SCM can lead to poor fit for the average, leading to possible bias. Pooled SCM, by contrast, can achieve nearly perfect fit for the average but can yield substantially worse unit-specific fits, making the estimator susceptible to interpolation bias from non-linearity and settings where the outcome process varies over time.

Our main proposal is *partially pooled SCM*, which finds intermediate weights between these two extremes. First, we show that, under a linear factor model, the error of the Average Treatment Effect on the Treated (ATT) estimate decomposes into error stemming from the pooled fit and from state-specific fits. By minimizing both quantities, partially pooled SCM thus directly controls the corresponding bias. We also motivate our proposal by examining the Lagrangian dual of the constrained optimization problem, showing that method partially pools parameters in the dual parameter space.

Partially pooled SCM in general does not perfectly balance both the unit-specific and pooled average pre-treatment outcomes. Chapter 2 demonstrates that augmenting SCM with an outcome model can correct for possible bias due to imperfect pre-treatment fit in settings with a single unit. We now extend these augmented panel data methods to the staggered adoption setting. While the framework is general, we focus here on augmentation with an average of pre-treatment outcomes, as would arise from a fixed effects specification. We refer to the augmented estimator as a *weighted event study*; the (unweighted) "event study" estimator is common in econometrics but has a number of pathologies that our weighted approach avoids (e.g., Abraham and Sun, 2018; Callaway and Sant'Anna, 2018). We can also view this as adapting intercept-shifted SCM (Doudchenko and Imbens, 2017; Ferman and Pinto, 2018) to the staggered adoption case.

We apply our methods to estimating the impact of mandatory teacher collective bargaining and show that they achieve better pre-treatment balance than existing approaches. We find no impact of teacher collective bargaining laws on either teacher salaries or student expenditures, consistent with several recent papers (Frandsen, 2016; Paglayan, 2019) but counter to earlier claims (most notably Hoxby, 1996).

Related work. This chapter contributes to several active methodological literatures. First, there is a large and active applied econometrics literature on challenges and remedies for

two-way fixed effects models with multiple treated units, including event study models; see Borusyak and Jaravel (2017); Abraham and Sun (2018); Athey and Imbens (2018); Goodman-Bacon (2018); Callaway and Sant'Anna (2018); see also Xu (2017) and Athey et al. (2017) for recent generalizations of these models. SCM has also attracted a great deal of attention; see Abadie (2019) for a recent review. Several recent papers have explored SCM with multiple treated units. In the case where all units adopt treatment at the same time, some propose to first average the units and then estimate SCM weights for the average, analogous to our fully pooled SCM estimate; for discussion, see Kreif et al. (2016); Robbins et al. (2017). An alternative is Abadie and L'Hour (2018), who instead propose to estimate separate SCM weights for each treated unit. In particular, they propose a penalized SCM approach that aims to reduce interpolation bias, allowing for weights that move continuously between standard SCM and nearest-neighbor matching. Our approach complements these papers by adapting some of these ideas to the staggered adoption setting. For some other examples of SCM under staggered adoption, see also Dube and Zipperer (2015); Toulis and Shaikh (2018); Donohue et al. (2019).

We also build on recent papers that combine outcome modeling and SCM in panel data settings, which themselves adapt "doubly robust" estimators more common in the (cross sectional) observational studies literature. To date, these approaches have been limited to the case with a single treated unit or, if multiple units are treated, to a single adoption time. Along with Chapter 2,Ferman and Pinto (2018); Abadie and L'Hour (2018); Chernozhukov et al. (2018); Arkhangelsky et al. (2019) all propose versions of bias correction. See also Arkhangelsky and Imbens (2019) for a more general discussion of double robustness in panel data settings.

Motivating example: Teacher collective bargaining. The United States, like other developed countries, spends substantial resources on public education. Approximately 80% of education spending goes to teacher salaries and benefits (U.S. Department of Education, National Center for Education Statistics, 2018), and recent research points to teacher quality as a key determinant of student outcomes (Jackson et al., 2014). Over recent decades, the teacher employment relationship has changed dramatically via the introduction of unions and collective bargaining agreements (Goldstein, 2015). Critics identify these as a "harmful anachronism" and "the most daunting impediments" to education reform (Hess and West, 2006). A major 2018 Supreme Court decision, *Janus v AFSCME*, is expected to weaken teachers' unions, bringing renewed attention to this area and raising interest in understanding the effects of teacher collective bargaining.

Since 1964, a number of states have passed laws mandating that school districts bargain with teachers' unions.[1] Given the strong criticism directed at teachers' unions, there is surprisingly little evidence that they, or the mandatory bargaining laws, have any effect at all. In a seminal study, Hoxby (1996) uses state-level changes in collective bargaining laws to

[1]Another 10 states allow but do not require collective bargaining, while 7 prohibit it. We focus on identifying the effects of mandates.

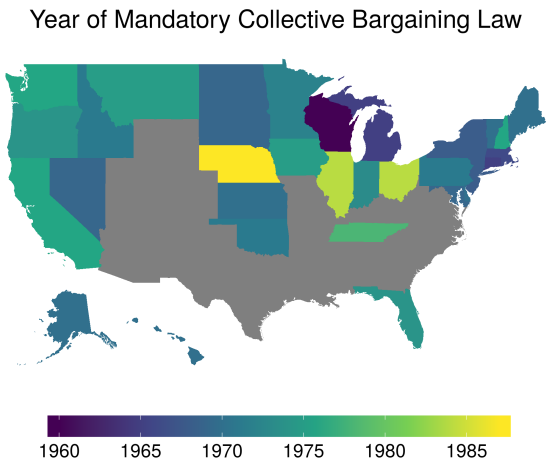


Figure 3.1: Staggered adoption of mandatory collective bargaining laws from 1964 to 1990.

argue that teacher collective bargaining raises teacher salaries and school expenditures but reduces student outcomes. Several more recent papers have disputed Hoxby's conclusions, however. Using a panel of school districts, Lovenheim (2009) finds little effect of unionization on teacher pay or class size. Frandsen (2016) similarly finds little effect of state unionization laws on teacher pay. Finally, Paglayan (2019) extends the historical state-level data set from Hoxby (1996). In a two-way fixed effect event study specification, she finds precisely estimated zero effects of mandatory bargaining laws on school expenditures and teacher salaries. See Appendix B.3 for additional discussion of Paglayan (2019).

Motivated in part by recent criticisms of event study models (Goodman-Bacon, 2018), we revisit the Paglayan (2019) analysis using different methods. Figure 3.1 shows adoption times of state mandatory bargaining laws between 1964 and 1990. Adoptions were spread across 14 separate years, though 16 states adopted laws between 1965 and 1970. Following Paglayan (2019), our main outcomes of interest are per-pupil student expenditures and teacher salaries, both measured in log 2010 dollars. We observe these outcomes back to 1959 for 49 states; we exclude Washington DC and Wisconsin, which adopted a mandatory bargaining law in 1960 and thus has only one year of pre-intervention data. This gives between 6 and 28 years of data before the adoption of mandatory bargaining, with an average of 13 years.

Chapter roadmap. Section 3.2 lays out the technical background. Section 3.3 develops SCM for the staggered adoption setting and introduces partially pooled SCM. Section 3.4 gives theoretical results for the generalized SCM approaches. Section 3.5 introduces augmented panel data methods and the weighted event study estimator. Section 3.6 describes a calibrated simulation study. Section 3.7 gives additional results for the teacher collective bargaining application. Finally, Section 4.6 discusses some directions for future work. The

supplementary materials in Appendix B include further analyses and technical results.

3.2 Preliminaries

Setup and notation

We consider a panel data setting where we observe outcomes Y_{it} for $i = 1, \ldots, N$ units over $t = 1, \ldots, T$ time periods. In the teacher collective bargaining application, $N = 49$ and $T = 39$. Some but not all of the units, indicated by $W_i = 1$, adopt the treatment during the panel; once units adopt treatment, they stay treated for the remainder of the panel. See Imai and Kim (2019) for a more general discussion. Let T_i represent the date that unit i receives treatment, with $T_i = \infty$ denoting never-treated units. Without loss of generality, we order units so that $T_1 \leq T_2 \leq \cdots \leq T_N$. We assume that there are non-zero number of never-treated units, $N_0 \equiv N - \sum_i W_i > 0$, and we let $J = N - N_0 = \sum_i W_i$. To clearly differentiate ever treated units, we index them by $j = 1, \ldots, J$. For treated units, we require a sufficient number of time periods both before and after treatment: we assume that $T_1 \gg 1$ and $T_J \leq T - K$ for some $K > 0$, representing the longest lagged treatment effect we will examine.[2]

We adopt a potential outcomes framework to express causal quantities (Neyman, 1923; Rubin, 1974) and assume stable treatment and no interference between units (SUTVA; Rubin, 1980). In principle, each unit i in each time t might have a distinct potential outcome for each potential treatment time s, $Y_{it}(s)$, for $s = 1, \ldots, T, \infty$. Following Athey and Imbens (2018), we adopt two assumptions that impose mild substantive restrictions but dramatically simplify the notation. First, we assume "no anticipation": prior to treatment, a unit's potential outcomes are equal to the control potential outcome, i.e. $Y_{it}(s) = Y_{it}(0)$ for $t < s$, with treatment time s. Second, we assume "invariance to history": following treatment, a unit's potential outcomes are equal to the treated potential outcome, $Y_{it}(s) = Y_{it}(1)$ for any $0 < s \leq t$, and do not depend on the timing of treatment. These assumptions allow us to use just two potential outcomes, $Y_{it}(0)$ and $Y_{it}(1)$; the observed outcome is $Y_{it} = \mathbb{1}\{t < T_i\}Y_{it}(0) + \mathbb{1}\{t \geq T_i\}Y_{it}(1)$ for units with $W_i = 1$ and is $Y_{it} = Y_{it}(0)$ for all t for units with $W_i = 0$. The first assumption is relatively innocuous, and is a generalization of the SUTVA assumption typically employed in cross-sectional studies (Rubin, 1980). The second is stronger, ruling out treatment effects that phase in over time. However, this is not too restrictive since we do not restrict how $\{Y_{it}(0), Y_{it}(1)\}$ vary across units. See Imai and Kim (2019) for related discussion.

[2] This ensures that we observe both pre-treatment outcomes and the outcome measures of interest for all treated units, and that there are untreated comparisons for even the last treated units; $N_0 = 17$, $T_1 = 6$, and $K = 10$ in our application. See Athey and Imbens (2018) for a discussion of the setting in which all units eventually adopt treatment.

Estimands

As is common in event studies, we focus on effects a specified period after treatment onset. For treated unit j, we index "event time" relative to treatment time T_j by $k = t - T_j$. The unit-level treatment effect for treated unit j at event time $k \geq 0$ is:

$$\tau_{jk} = Y_{j,T_j+k}(1) - Y_{j,T_j+k}(0).$$

Our primary estimand of interest is the Average Treatment Effect on the Treated k periods after treatment onset, sometimes called the "event study" or "dynamic" ATT (Abraham and Sun, 2018):

$$\text{ATT}_k \equiv \frac{1}{J}\sum_{j=1}^{J} \tau_{jk} = \frac{1}{J}\sum_{j=1}^{J} Y_{j,T_j+k}(1) - Y_{j,T_j+k}(0).$$

We are also interested in the average post-treatment effect, averaging across k: $\text{ATT} = \frac{1}{K}\sum_{k=1}^{K} \text{ATT}_k$. Our methods generalize to many other estimands; see Callaway and Sant'Anna (2018) for examples in this setting.

We observe all treated potential outcomes for treated units post-treatment adoption; that is, $Y_{j,T_j+k}(1)$ is observed for $k \leq K \leq T - T_J$ for all $W_j = 1$. The key challenge is therefore to impute the average of the missing un-treated potential outcomes:

$$\mu_k \equiv \frac{1}{J}\sum_{j=1}^{J} Y_{j,T_j+k}(0).$$

It is useful to define the set of *not yet treated* units, which are the potential "donor units" for SCM. For fixed event time k, the possible donor units for treated unit j are those units that are either never treated or are not yet treated by time $T_j + k$. We denote these as $\mathcal{D}_{jk} = \{i : W_i = 0 \text{ or } T_i > T_j + k\}$. For a given treated unit j, $\mathcal{D}_{jk}$ can differ across event time k since units adopt treatment over time; however we use a fixed set of donors for each treated unit. Following Paglayan (2019), we set the maximum value of leads to $K = 10$, and restrict our attention to the set $\mathcal{D}_j \equiv \mathcal{D}_{jK}$. This reduces the number of available donor units but simplifies both estimation and exposition.

Finally, auxiliary covariates play an important role in many panel data settings. Consistent with the analysis in Paglayan (2019), we do not include such covariates here. However, it is straightforward to extend our results to consider auxiliary covariates in parallel to the lagged outcomes.

SCM for a single treated unit

In the synthetic control method, the counterfactual outcome under control is estimated from a weighted average, known as a *synthetic control* of untreated units, where weights are chosen to minimize the squared imbalance between the lagged outcomes for the treated unit and the weighted control ("donor") units.

For a fixed treated unit j, we consider a modified version of the original SCM estimator of Abadie et al. (2010, 2015). In this version, the SCM weights $\hat{\gamma}_j$ for treated unit j are the solution to a constrained optimization problem:

$$\min_{\gamma_j \in \Delta_j^{\text{scm}}} \frac{1}{2(T_j-1)} \sum_{\ell=1}^{T_j-1} \left(Y_{j,T_j-\ell} - \sum_{i=1}^{N} \gamma_{ij} Y_{i,T_j-\ell} \right)^2 + \lambda \sum_{i=1}^{N} f(\gamma_{ij}), \tag{3.1}$$

where $\gamma_j \in \Delta_j^{\text{scm}}$ is an N-vector that represents SCM donor unit weights, with elements $\{\gamma_{ij}\}$ that satisfy $\gamma_{ij} \geq 0$ for all i, $\sum_i \gamma_{ij} = 1$, and $\gamma_{ij} = 0$ whenever i is not a possible donor, $i \notin \mathcal{D}_j$.[3] Equation (3.1) modifies the original SCM proposal in two key ways. First, where Abadie et al. (2010, 2015) balance auxiliary covariates, we focus exclusively on lagged outcomes Second, following a suggestion in Abadie et al. (2015), we include a term that penalizes the weights toward uniformity, with hyperparameter λ; see Doudchenko and Imbens (2017); Abadie and L'Hour (2018). In settings with perfect pre-treatment fit, the choice of penalty can be important as Equation (3.1) may not have a unique solution for $\lambda = 0$. This is not the case in our setting, however, and so we largely view this term as a technical convenience.

The SCM estimate of the missing potential outcome for treated unit j at event time k, $Y_{j,T_j+k}(0)$, is then:

$$\hat{Y}_{j,T_j+k}(0) = \sum_{i=1}^{N} \hat{\gamma}_{ij} \, Y_{i,T_j+k},$$

with estimated treatment effect $\hat{\tau}_{jk}^{\text{scm}} = Y_{j,T_j+k} - \hat{Y}_{j,T_j+k}(0)$. Thus, the optimization problem (3.1) minimizes the L_2 norm of the imbalance between the treated unit and the synthetic control over the pre-treatment period. Alternatively, it can be seen as minimizing the sum of squared placebo treatment effects on pre-treatment outcomes.

A central question for SCM is assessing whether $\hat{Y}_{j,T_j+k}(0)$ is a reasonable estimate for $Y_{j,T_j+k}(0)$. Abadie et al. (2010) argue that, in addition to other model checks, SCM will be a compelling estimator when the placebo estimates are close to zero, i.e. $Y_{j,T_j-\ell} - \hat{Y}_{j,T_j-\ell} \approx 0$ for all lags ℓ. Accordingly, Abadie et al. (2010, 2015) recommend only proceeding with an SCM analysis if the pre-treatment fit is excellent. Figure 3.2a shows SCM "gap plots," of $Y_{j,T_j-\ell} - \hat{Y}_{j,T_j-\ell}$ against ℓ for three illustrative treated states taken one at a time. Ohio shows relatively good pre-treatment fit; however, the SCM estimates for Illinois and New York fail to closely track those states' pre-treatment outcomes, suggesting SCM is likely to give misleading estimates for these states.

[3]Recall that the possible donor units for unit j, $\mathcal{D}_j$, are the *not yet treated* units, defined as either never-treated units or units that have not yet been treated at time $T_j + K$ where we set $K = 10$. Thus, the set of possible donor states for Michigan, which adopted mandatory collective bargaining in 1965, includes the never-treated states as well as Nebraska, which also adopted mandatory collective bargaining in 1987.

3.3 Generalizing to staggered adoption: Partially Pooled SCM

We now extend SCM to the staggered adoption setting. The literature provides little guidance about how to do this, and applied researchers have used a range of ad hoc approaches. We start by outlining two immediate generalizations: separate SCM, which estimates weights that separately minimize the pre-treatment imbalance for each treated unit; and pooled SCM, which estimates weights that minimize the average pre-treatment imbalance across all treated units. Both approaches have drawbacks, however. Separate SCM can lead to poor fit for the average of the treated units, leading to bias in the estimated ATT. Pooled SCM, by contrast, can achieve nearly perfect fit for the average but yields substantially worse state-specific fits, making the estimator more susceptible to interpolation bias from non-linearity and when both treatment adoption and the outcome process vary across time. Recognizing this, we then propose *partially pooled SCM*, which finds intermediate weights between these two extremes. We turn to the theoretical properties of this approach in the next section.

Separate SCM for each treated unit

A number of applied researchers have confronted the problem of using SCM when there are multiple treated units with staggered adoption. These studies have taken each treated unit one at a time, forming a separate synthetic control for each, and then estimating the ATT by averaging the unit-specific SCM estimates (see, for example, Dube and Zipperer, 2015; Donohue et al., 2019). We can re-write this *separate SCM* strategy as solving a single joint optimization problem over a matrix of weights $\Gamma = [\gamma_1, \ldots, \gamma_J] \in \mathbb{R}^{N \times J}$.[4]

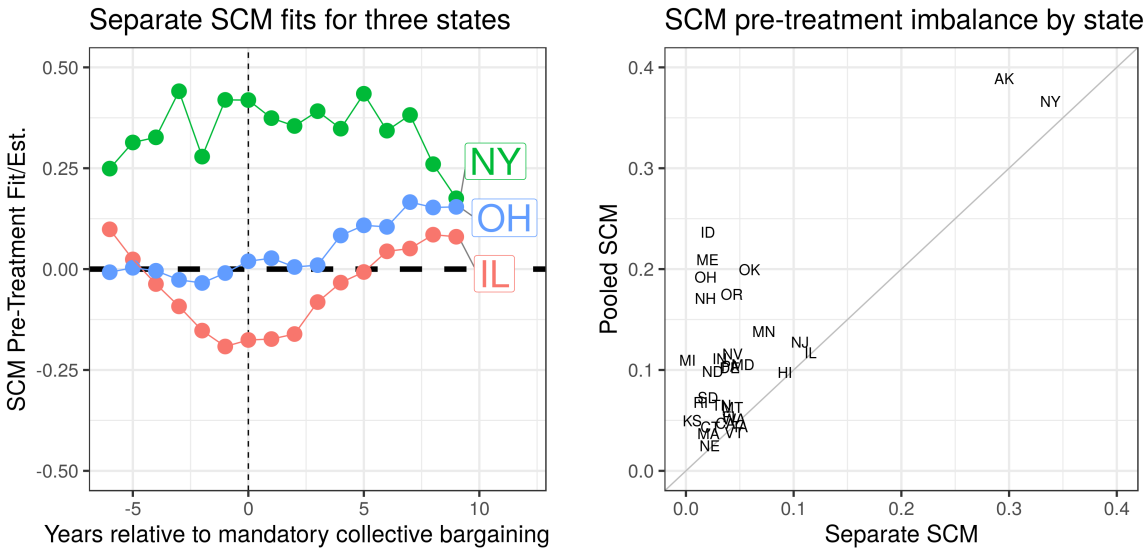
$$\min_{\gamma_1,\ldots,\gamma_J \in \Delta_j^{\text{scm}}} \underbrace{\frac{1}{2J}\sum_{j=1}^{J}\left[\frac{1}{T_j-1}\sum_{\ell=1}^{T_j-1}\left(Y_{j,T_j-\ell} - \sum_{i=1}^{N}\gamma_{ij}Y_{i,T_j-\ell}\right)^2\right]}_{q^{\text{sep}}} + \lambda\sum_{j=1}^{J}\sum_{i=1}^{N} f(\gamma_{ij}), \quad (3.2)$$

where q^{sep} is the average pre-intervention mean square error across the J treated units. The estimated ATT is then:

$$\widehat{\text{ATT}}_k = \frac{1}{J}\sum_{j=1}^{J}\left[Y_{j,T_j+k} - \sum_{i=1}^{N}\hat{\gamma}_{ij}\, Y_{i,T_j+k}\right] = \frac{1}{J}\sum_{j=1}^{J} Y_{j,T_j+k} - \frac{1}{J}\sum_{j=1}^{J}\hat{Y}_{j,T_j+k}(0), \quad (3.3)$$

where the last term imputes the missing (average) potential outcome by averaging over the unit-specific SCM estimates.

[4]Recall that we adopt the convention that each γ_j is an N-vector, with entries corresponding to inadmissible donor units fixed at zero. We fix the same value of the hyper-parameter λ across all problems. It is straightforward to generalize this to a separate λ_j for each treated unit, but complicates the exposition. In principle, we could also give different weights to different units in the ATT, for example, prioritizing larger states over smaller states or weighting units by dose (e.g., Dube and Zipperer, 2015).



(a) SCM "gap plots" for three illustrative states

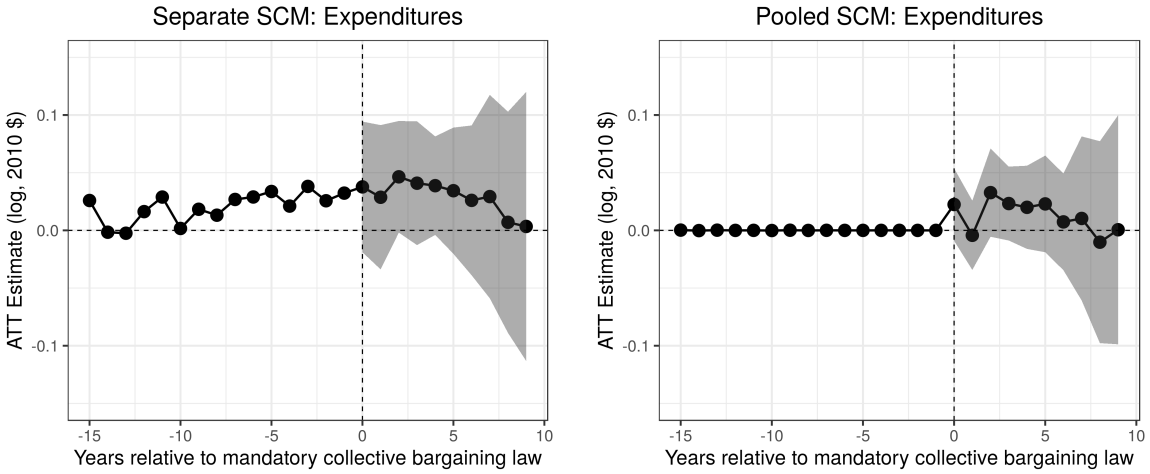
(b) SCM pre-treatment fits by state

Figure 3.2: (a) The SCM pre-treatment fit for Ohio is good overall. The pre-treatment fit for Illinois is poor: the SCM estimate fails to match an important pre-treatment trend. The pre-treatment fit for New York is quite bad: the pre-treatment imbalance for New York is roughly an order of magnitude larger than typical estimates for the impact of teacher mandatory bargaining. (b) SCM fits by state show that Separate SCM gives better pre-treatment fit for all treated states.

As with SCM for a single treated unit, an important question is when this separate SCM strategy will yield a reasonable estimate of ATT_k. One possible justification is that $\widehat{\text{ATT}}_k$ will be an unbiased estimate of ATT_k if the set of the state-specific SCM estimates $\{\hat{Y}_{j,T_j+k}(0)\}_j$ are all unbiased for the corresponding potential outcomes $\{Y_{j,T_j+k}(0)\}_j$. However, Figure 3.2a, which plots the placebo gaps for three example states, shows this is not the case in our application. This suggests that the separate SCM strategy will not yield convincing estimates in our setting. We expect that in many applications there will be several treated units with poor pre-treatment fit (see e.g. Dube and Zipperer, 2015; Donohue et al., 2019). This motivates the search for other approaches.

Pooled SCM

An alternative strategy is to estimate weights that balance the average across treated units directly. We call this *pooled SCM*. Specifically, we modify the separate SCM problem in



(a) Separate SCM (b) Pooled SCM

Figure 3.3: Estimated ATT on per-pupil expenditure (log, 2010 $) using (a) separate SCM, and (b) pooled SCM.

Equation (3.2) to:

$$\min_{\gamma_1,\ldots,\gamma_J \in \Delta^{\text{scm}}} \underbrace{\frac{1}{L}\sum_{\ell=1}^{L}\left[\sum_{T_j>\ell}\left(Y_{j,T_j-\ell} - \sum_{i=1}^{N}\gamma_{ij}Y_{i,T_j-\ell}\right)\right]^2}_{q^{\text{pool}}} + \lambda\sum_{j=1}^{J}\sum_{i=1}^{N} f(\gamma_{ij}), \tag{3.4}$$

where $L = T_J - 1$ is the maximum number of observed lags, q^{pool} is the imbalance between the average synthetic control and the average treated unit at each lag ℓ, summed over the possible ℓ, and the weights are again constrained to be non-negative, to sum to one for each j, and to be zero for any i not in the set of donors $\mathcal{D}_j$. Intuitively, by minimizing q^{pool} the pooled SCM approach finds weights that minimize the placebo estimates for the ATT, rather than weights that minimize the average unit-specific placebo estimates, as in q^{sep}. We can see this in Figure 3.3, which shows the implied placebo estimates for the ATT using the two approaches: The placebo ATT estimates are consistently positive for separate SCM weights and are all nearly identical to zero for pooled SCM weights.

At the same time, the pooled SCM weights generally yield worse state-specific fits, which do not enter the objective function in Equation (3.4). Figure 3.2b plots the state-level pre-treatment imbalances for separate SCM vs pooled SCM, showing that the separate SCM fit is better for all treated states. The resulting estimator is therefore more susceptible to interpolation biases due to non-linearity (see e.g. Abadie and L'Hour, 2018, for a setting with abundant micro data). Furthermore, as we show through simulation in Appendix B.2,

even under linearity the pooled SCM estimator can be poor when both treatment adoption and the outcome process vary over time.

Partially pooled SCM

We can now define our main proposal, *partially pooled SCM*, which finds weights that minimize a convex combination of state-level imbalance, q^{sep}, and pooled imbalance, q^{pool}:

$$\min_{\gamma_1,\ldots,\gamma_J \in \Delta^{\text{scm}}} \quad \frac{\nu}{2}\, q^{\text{pool}}(\Gamma) + \frac{(1-\nu)}{2}\, q^{\text{sep}}(\Gamma) \;+\; \lambda \sum_{j=1}^{J} \sum_{i=1}^{N} f(\gamma_{ij}), \tag{3.5}$$

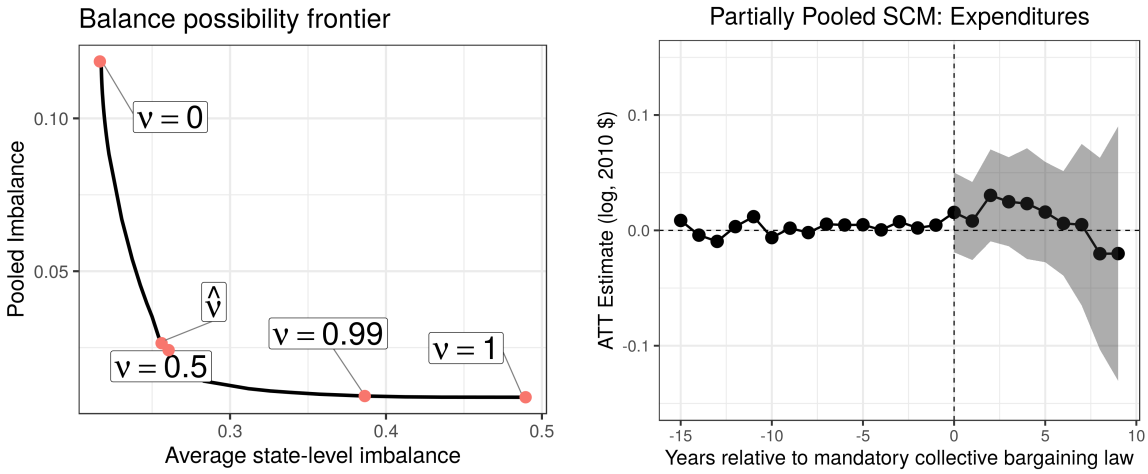
with hyperparameter $\nu \in [0, 1]$. This optimization problem nests both the separate SCM approach (3.2) with $\nu = 0$ and the pooled SCM approach (3.4) with $\nu = 1$. In the next section, we show that intermediate values of ν correspond to a *partial pooling* solution for the γ weights in the dual parameter space, and discuss the specific choice of ν.

Figure 3.4a shows the *balance possibility frontier* for all ν: the impact of changing ν on the pooled imbalance q^{pool} (the y-axis) and on the average state-level imbalance q^{sep} (the x-axis) for the teacher collective bargaining application.[5] The end points, $\nu = 0$ and $\nu = 1$ correspond to separate SCM and pooled SCM, respectively. As ν rises, pooled imbalance falls while state-level imbalance rises, though at different rates. Moving from the separate SCM estimate of $\nu = 0$ to a partially pooled SCM estimate of $\nu = 0.5$ reduces the pooled imbalance by over 90 percent, with more modest further reductions as $\nu \to 1$. This is consistent with Figure 3.2, which shows poor fit for $\nu = 0$ and nearly perfect pre-treatment fit for $\nu = 1$.

Meanwhile, average state-level imbalance increases relatively slowly as ν rises from 0 to 0.5, increasing more rapidly as ν nears its upper limit. Even a very small deviation from the pooled SCM solution, such as from $\nu = 1$ to $\nu = 0.99$, cuts the average state-level imbalance by roughly one-fifth with essentially no change in the overall imbalance. Due to the number of degrees of freedom involved, in many cases the pooled imbalance will be near zero for $\nu = 1$, and the objective function q^{pool} will be relatively flat in the neighborhood of the pooled solution. Therefore we expect that in many cases it will be possible to trade off a small increase in pooled imbalance for a large decrease in the state-level imbalance, adding robustness to the estimator at relatively little cost.

Based on Figure 3.4a, the intermediate estimate with $\nu = 0.5$ has very similar global pre-treatment imbalance to the fully pooled estimator, $\nu = 1$, with only a modest increase in state-level imbalance relative to the separate SCM estimate, $\nu = 0$. This is reflected in Figure 3.4b, which shows the placebo ATT estimates for partially pooled SCM. While the imbalance for the ATT is slightly larger than for pooled SCM, it is substantially better than for separate SCM.

[5] See King et al. (2017) and Pimentel and Kelz (2019) for other examples of balance frontiers in observational settings.



(a) The *balance possibility frontier* (b) Partially pooled SCM

Figure 3.4: (a) The trade-off between pooled imbalance and state-specific imbalance, where $\nu = 0$ is the separate SCM solution and $\nu = 1$ is the pooled SCM solution. The large distance in average state-level imbalance between $\nu = 0.99$ and $\nu = 1$ suggest meaningful gains in balance from deviating from the complete pooling estimate even by a small amount. (b) Partially pooled SCM estimates for per-pupil current expenditures (log, 2010 \$), with the heuristic $\hat{\nu} = \sqrt{q^{\text{pool}}} \Big/ \sqrt{q^{\text{sep}}} \approx 0.44$; see Section 3.4.

We can now turn to the ATT estimates themselves. Figure 3.4b shows $\widehat{\text{ATT}}_k$ for $k \in [0, 10]$ for partially pooled SCM. Following Arkhangelsky et al. (2019), we quantify uncertainty using the (leave-one-unit-out) jackknife to estimate standard errors and (asymptotic) Normality to compute 95% confidence intervals.[6] Similar to the estimates from separate SCM and pooled SCM — and consistent with Paglayan (2019) — we find no effect of mandatory teacher collective bargaining laws on student expenditures. We explore additional analyses in Section 3.7.

3.4 Theoretical results for partially pooled SCM

This section explores some of the theoretical properties of SCM under staggered adoption, especially partially pooled SCM. First, we show that, under a linear factor model, the error of the ATT estimate decomposes into error stemming from imperfection of the pooled fit and from imperfections in the state-specific fits. By minimizing a weighted average of these

[6]The theoretical results in Arkhangelsky et al. (2019) focus on the setting with simultaneous adoption; we leave the formal extensions to this setting for future work. While we do not pursue it here, we anticipate that we could also apply the influence function-based inference method in Callaway and Sant'Anna (2018), especially for the weighted event study estimators we consider in Section 3.5.

quantities, partially pooled SCM, with appropriately chosen ν, can thus directly control the corresponding bias. We also motivate partially pooled SCM by examining the Lagrangian dual of the constrained optimization problem, showing that method partially pools parameters in the dual parameter space.

Pre-treatment fit and bias under a linear factor model

We consider the bias of a generic weighting estimator under staggered adoption. Following the recent literature on panel data methods, we consider data generated by a linear factor model; in Appendix B.4 we show analogous results for a time-varying autoregressive process. For ease of exposition we consider the case where we balance the first $t = 1, \ldots, L$ time periods for each unit, and we focus on the (absolute) error in estimating the ATT at event time k, $\left|\widehat{\text{ATT}}_k - \text{ATT}_k\right|$.

We assume that there are F latent time-varying factors, where F is typically small relative to both N and T. We let $\mu_t \in \mathbb{R}^F$ represent the vector of factor values at time t, and assume it is bounded: $\max_t \|\mu_t\|_\infty \leq M$. Each unit has a vector of time-invariant factor loadings $\phi_i \in \mathbb{R}^F$, and the control potential outcomes are generated as:

$$Y_{it}(0) = \phi_i' \mu_t + \varepsilon_{it}, \tag{3.6}$$

where ε_{it} is independent, sub-Gaussian additive noise with scale parameter σ.

Two summaries of the factor values at the various treatment times will be important to our analysis: $\bar{\mu}_k = \frac{1}{J}\sum_{j=1}^{J} \mu_{T_j+k} \in \mathbb{R}^F$, the *average factor value* across the J treatment times, and $S_k^2 = \frac{1}{J}\sum_{j=1}^{J} \|\mu_{T_j+k} - \bar{\mu}_k\|_2^2$, the corresponding *variance*. Together, the relative values of the magnitude of the mean $\|\bar{\mu}_k\|_2$ and the standard deviation S_k measure the amount of heterogeneity in the factors over treatment times.[7] Our first result is that under the standard linear factor model the error bound for the ATT depends on *both* the pooled pre-treatment fit and the state-specific pre-treatment fits, where $\|\bar{\mu}_k\|_2$ and S_k control their relative importance for the bound.

Theorem 3.1. For $\hat{\gamma}_1, \ldots, \hat{\gamma}_J \in \Delta^{\text{scm}}$ where $\hat{\gamma}_j$ is independent of $\varepsilon_{\cdot T_j+k}$ and $\delta > 0$, if $Y_{it}(0)$

[7] For example, if all treatment times are the same, $T_1 = \ldots = T_J$, then the standard deviation $S_k = 0$. Similarly, in the special case of a unit fixed effects model (e.g. a single, time constant factor) the standard deviation is also zero. Conversely, if the factor values vary widely over time, the standard deviation S_k will be large relative to the magnitude of the average factor value $\|\bar{\mu}_k\|_2$.

follows a linear factor model (3.6) the error for $\widehat{\text{ATT}}_k$ is

$$\left|\widehat{\text{ATT}}_k - \text{ATT}_k\right| \leq \frac{M\sqrt{F}}{\sqrt{L}} \left(\|\bar{\mu}_k\|_2 \underbrace{\sqrt{\sum_{t=1}^{L} \left(\frac{1}{J}\sum_{j=1}^{J} Y_{jt} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} Y_{it} \right)^2}}_{\text{bias due to pooled fit}} \right.$$
$$\left. + S_k \underbrace{\sqrt{\frac{1}{J}\sum_{j=1}^{J}\sum_{t=1}^{L} \left(Y_{jt} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} Y_{it} \right)^2}}_{\text{bias due to individual fit}} \right) \tag{3.7}$$
$$+ \underbrace{\frac{\sigma M^2 F}{\sqrt{L}} \left(3\delta + 2\sqrt{\log NJ}\right)}_{\text{bias due to approximation error}} + \underbrace{\frac{\delta\sigma}{\sqrt{J}} \left(1 + \|\hat{\Gamma}\|_F\right)}_{\text{variance}}$$

with probability at least $1 - 6e^{-\frac{\delta^2}{2}}$, where $\max_t \|\mu_t\|_\infty \leq M$ and $\|\Gamma\|_F^2 = \sum_{j=1}^{J}\sum_{i=1}^{N} \gamma_{ij}^2$ is the Frobenius norm of the weights.

Theorem 3.1 shows that the error for the ATT is bounded by several distinct terms, each of which can be controlled by the partially pooled optimization problem (3.5). First, we can directly control the variance term due to noise by penalizing the dispersion of the weights, e.g., $\|\hat{\Gamma}\|_F^2$. Second, there is an approximation error that arises due to balancing — and possibly over-fitting to — noisy outcomes, rather than to the true underlying factor loadings. In the worst case, the J synthetic controls put maximal weight on the control units with the largest noise. Constraining the weights to lie in the simplex reduces the impact of this worst case, however, and the error decreases as more lagged outcomes are balanced; see Chapter 2 and Abadie et al. (2010); Arkhangelsky et al. (2019) for further discussion.

We are most interested in the bias arising from the pooled and state-specific fits. Theorem 3.1 shows that the bias in a weighting estimator for the ATT is controlled by choosing weights that optimize a weighted sum of these two pre-treatment fits. The relative importance of these fits is governed by the ratio of the average factor value $\|\bar{\mu}_k\|_2$ and the factor standard deviation S_k. When the average factor value is large relative to the standard deviation, then the level of pooled fit is more important than the state-specific fits. Conversely, when the factors vary widely over treatment times then the state-specific fits outweigh the pooled fit. In the special case where $S_k = 0$, such as when all treatment times are the same or under a unit (one-way) fixed effects model, Theorem 3.1 shows that *only* the pooled level of fit is important for bias, at least under a linear factor model. In Appendix B.4 we show that a time-varying auto-regressive model also exhibits this behavior.

Following Theorem 3.1, the (infeasible) optimization problem would choose hyperparameter $\nu = \frac{\|\bar{\mu}_k\|_2}{\|\bar{\mu}_k\|_2 + S_k}$ for some k, minimizing the first two terms in Equation (3.7). However,

in general both $\|\bar{\mu}_k\|_2$ and S_k are unknown, so we propose a heuristic to set ν based on the ratio of q^{pool} and q^{sep}. First we solve the partially pooled SCM problem (3.5) with $\nu = 0$ (i.e. the separate SCM problem), then set ν to be the ratio of q^{pool} and q^{sep}: $\hat{\nu} = \sqrt{q^{\text{pool}}} \Big/ \sqrt{q^{\text{sep}}} \in [0, 1]$.[8] If the separate SCM problem (3.2) achieves good pooled balance on its own, this approach will set a small ν. Conversely, if the pooled balance is poor, ν will be large in order to account for this discrepancy. In the teacher bargaining example, we choose $\hat{\nu} \approx 0.44$ for the per-pupil expenditure outcome, which is close to the $\nu = 0.5$ plotted in Figure 3.4a.

Partially pooled SCM: Dual shrinkage

We now inspect the Lagrangian dual problem to the partially pooled SCM problem (3.5), showing that the optimization problem partially pools a set of state-specific dual variables toward global dual variables. We focus on balancing the first $L = T_1 - 1$ lagged outcomes, which are observed for each treated unit; see Appendix B.4 for the general case.

For each treated unit j, the sum-to-one constraint induces a Lagrange multiplier $\alpha_j \in \mathbb{R}$, and the state-level balance measure induces a set of Lagrange multipliers $\beta_j \in \mathbb{R}^L$, with elements $\beta_{\ell j}$. We combine these dual parameters into a vector $\alpha = [\alpha_1, \ldots, \alpha_J] \in \mathbb{R}^J$ and a matrix $\beta = [\beta_1, \ldots, \beta_J] \in \mathbb{R}^{L \times J}$. In addition to the J sets of Lagrange multipliers — one for each treated unit — the pooled balance measure in the partially pooled SCM problem Equation (3.5) induces a set of global Lagrange multipliers $\mu_\beta \in \mathbb{R}^L$. As we see in the following proposition, in the dual problem the parameters $\beta_1, \ldots, \beta_J$ are regularized toward this set of pooled Lagrange multipliers, μ_β.

Proposition 3.1. The Lagrangian dual to Equation (3.5) is:

$$\min_{\alpha, \mu_\beta, \beta} \mathcal{L}(\alpha, \beta) + \frac{\lambda J L}{2(1-\nu)} \sum_{j=1}^{J} \|\beta_j - \mu_\beta\|_2^2 + \frac{\lambda L}{2\nu} \|\mu_\beta\|_2^2. \tag{3.8}$$

Where the dual objective function is

$$\mathcal{L}(\alpha, \beta) \equiv \sum_{j=1}^{J} \left[\sum_{W_i = 0} f^* \left(\alpha_j + \sum_{\ell=1}^{L} \beta_{\ell j} Y_{i, T_j - \ell} \right) - \left(\alpha_j + \sum_{\ell=1}^{L} \beta_{\ell j} Y_{j, T_1 - \ell} \right) \right], \tag{3.9}$$

and $f^*(y) = \sup_x x'y - f(x)$ is the convex conjugate of f.[9] For treated unit j, the synthetic control weight on unit i is $\hat{\gamma}_{ij} = f^{*\prime} \left(\hat{\alpha}_j + \sum_{\ell=1}^{L} \hat{\beta}_{\ell j} Y_{j, T_1 - \ell} \right)$.

[8] Note that by the triangle inequality, $\sqrt{q^{\text{pool}}} \leq \sqrt{q^{\text{sep}}}$. Thus their ratio is bounded above by 1. If the SCM fits are perfect for each state, $q^{\text{sep}} = 0$, then the overall fit will also be perfect, $q^{\text{pool}} = 0$, and we define $\nu = 0$. This is not a common situation.

[9] For example, if $f(x) = x \log x$ is an entropy penalty, then $f^*(y) = \exp(y - 1)$ is an exponential. If $f(x) = \frac{1}{2} x^2$, then $f^*(y) = \frac{1}{2} y^2$.

The dual objective function (3.9) is an example of a *calibrated loss function* for propensity score parameters (see e.g. Zhao, 2018; Wang and Zubizarreta, 2019); we develop this propensity score connection in Appendix B.1. More relevant for our purposes is the form of the regularization terms in (3.8). Proposition 3.1 highlights that the estimator partially pools the individual synthetic controls to the pooled synthetic control *in the dual parameter space*, with ν controlling the level of pooling. When $\nu = 0$ in the separate SCM problem, the parameters $\beta_1, \dots \beta_J$ are shrunk towards zero rather than a set of global parameters. By contrast, when $\nu = 1$, $\beta_1, \dots, \beta_J$ are constrained to be equal to μ_β, fitting a single pooled synthetic control in the dual parameter space. By choosing $\nu \in (0, 1)$, we move continuously between the two extremes of J separate Lagrangian dual problems and a single dual problem, regularizing the individual β_js toward the pooled μ_β, allowing for some limited differences between the J dual parameters.

3.5 Combining SCM and outcome modeling

We have established that the partially pooled SCM estimator achieves nearly as good overall balance as the fully pooled estimator, while achieving much better balance for each state. Nevertheless, balance will typically be imperfect, especially at the state level. We now follow Chapter 2 and combine partially pooled SCM with outcome modeling, which can correct for imperfect pre-treatment balance in the SCM estimator. We first describe the general framework to combine SCM with an arbitrary panel data imputation method. We then focus on augmentation with a (possibly weighted) average of pre-treatment outcomes, which we refer to as a *weighted event study*.

Augmentation with generic panel data methods

Constructing the augmented estimator proceeds in three steps. First, we consider a working model for the potential outcome under control, k periods after treatment time T_j: $Y_{i,T_j+k}(0) = m_{ijk} + \varepsilon_{i,T_j+k}$; we give specific examples below. We estimate m_{ijk} with a pilot estimate $\hat{m}_{ijk}$, and define the corresponding *residuals*, $\dot{Y}_{i,T_j+k} \equiv Y_{i,T_j+k} - \hat{m}_{ijk}$ for event time k. Second, we estimate SCM weights $\hat{\gamma}^*_{ij}$ using these residuals, i.e., by modifying the balance criteria q^{pool} and q^{sep} in Equation (3.5) to depend on the residuals $\{\dot{Y}_{i,T_j+k}\}$ rather than "raw" $\{Y_{i,T_j+k}\}$. Finally, we impute the counterfactual for treated unit j, k periods after treatment as:

$$\begin{aligned} \hat{Y}^{\text{aug}}_{j,T_j+k} &= \sum_{i=1}^{n} \hat{\gamma}^*_{ij} Y_{i,T_j+k} \quad + \left(\hat{m}_{jjk} - \sum_{i=1}^{n} \hat{\gamma}^*_{ij} \hat{m}_{ijk} \right) \\ &= \hat{m}_{jjk} \qquad\qquad\qquad + \sum_{i=1}^{n} \hat{\gamma}^*_{ij} \left(Y_{i,T_j+k} - \hat{m}_{ijk} \right). \end{aligned} \tag{3.10}$$

Following Chapter 2 we can view this approach as analogous to bias correction for matching (Rubin, 1973; Abadie and Imbens, 2011), where $\hat{m}_{jjk} - \sum_{i=1}^{n} \hat{\gamma}_{ij}^{*} \hat{m}_{ijk}$ is an estimate of the bias. As with partially pooled SCM, we then estimate $\widehat{\text{ATT}}_k^{\text{aug}} = \frac{1}{J} \sum_{j=1}^{J} \hat{\tau}_{jk}^{\text{aug}}$.

This formulation is quite general and can accommodate any panel data imputation method for the pilot estimate $\hat{m}_{ijk}$. We focus next on simple outcome models, especially unit fixed effects. More broadly, however, we can estimate the factor model (3.6) directly, as in the *generalized SCM* approach of Xu (2017) and set the pilot estimate to be the imputed counterfactual $\hat{m}_{ijk} = \hat{\phi}_i' \hat{\mu}_{T_j+k}$. Alternatively, we can estimate $\hat{m}_{ijk}$ using a direct matrix completion approach (Hastie et al., 2015; Athey et al., 2017). We inspect the performance of augmenting SCM with a factor model through simulation in Section 3.6 and apply it to the teacher collective bargaining example in Section 3.7.

Weighted event studies

Our primary recommendation is to augment partially pooled SCM with a (possibly weighted) average of pre-treatment outcomes, which we refer to as a *weighted event study*; see Abraham and Sun (2018); Callaway and Sant'Anna (2018) for futher discussion of event study models. Here, the pilot estimate for unit i, k periods after treatment time T_j is a weighted average of the pre-treatment outcomes:

$$\hat{m}_{ijk} = \hat{\eta}_j' Y_{iT_j}^{\text{pre}} \equiv \sum_{\ell=1}^{T_j-1} \hat{\eta}_{j\ell} Y_{i,T_j-\ell}, \tag{3.11}$$

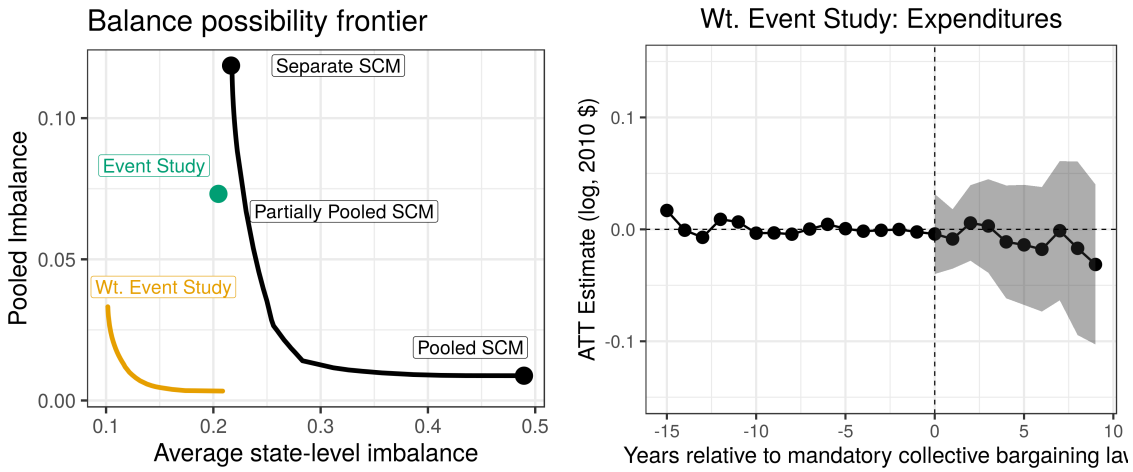
where the weights η_j need not be on the simplex. The treatment effect estimates for τ_{jk}, the impact for treated unit j at event time k, have a particularly useful form:

$$\hat{\tau}_{jk}^{\text{aug}} = \left(Y_{j,T_j+k} - \sum_{\ell=1}^{T_j-1} \hat{\eta}_{j\ell} Y_{j,T_j-\ell} \right) - \sum_{i=1}^{N} \hat{\gamma}_{ij}^{*} \left(Y_{i,T_j+k} - \sum_{\ell=1}^{T_j-1} \hat{\eta}_{j\ell} Y_{i,T_j-\ell} \right), \tag{3.12}$$

where $\widehat{\text{ATT}}_k^{\text{aug}}$ is the simple average of $\hat{\tau}_{jk}^{\text{aug}}$ over treated units j. We can view this approach as a weighted average over all possible two-period, two-group difference-in-differences estimates. Specifically, the base difference-in-differences estimate compares the "single difference" in outcomes for treated unit j at two time points, $Y_{j,T_j+k} - Y_{j,T_j-\ell}$, to the "single difference" in outcomes for donor unit i at the same time points, $Y_{i,T_j+k} - Y_{i,T_j-\ell}$. Equation (3.12) fixes treated unit j and event time k, but then takes a double-weighted average, first over pre-treatment periods to form a "synthetic pre treatment time period", then over donor units to track pre-intervention trends (see Arkhangelsky et al., 2019, for additional discussion).

Our default approach is to set uniform weights over time periods, $\hat{\eta}_{j\ell} = \frac{1}{T_j - 1}$:

$$\hat{\tau}_{jk}^{\text{aug}} = \frac{1}{T_j - 1} \sum_{\ell=1}^{T_j-1} \left[\left(Y_{j,T_j+k} - Y_{j,T_j-\ell} \right) - \sum_{i=1}^{N} \hat{\gamma}_{ij}^{*} \left(Y_{i,T_j+k} - Y_{i,T_j-\ell} \right) \right], \tag{3.13}$$



(a) The balance possibility frontier for the weighted event study, and SCM alone.

(b) Weighted event study, $\hat{\nu} = 0.266$

Figure 3.5: (a) The balance possibility frontier for SCM alone and for the weighted event study model, which combines SCM and fixed effects, as well as the implied imbalance for fixed effects alone. Incorporating unit-level fixed effects leads to substantial improvements in balance. We use Equation (3.12) to estimate the event study estimator and compute the implied balance as $\sqrt{\sum_{\ell=2}^{L} \hat{\delta}^2_{-\ell}}$, the RMSE of the placebo estimates. (b) Weighted event study estimates for per-pupil current expenditure (log, 2010 $).

which is equivalent to augmenting SCM with a unit fixed effects model, $m_{ijk} = \frac{1}{T_j - 1} \sum_{\ell=1}^{T_j - 1} Y_{i,T_j - \ell}$. This approach extends the intercept-shifted or de-meaned SCM estimator, which has attractive robustness properties (Doudchenko and Imbens, 2017; Ferman and Pinto, 2018), to the staggered adoption setting.

A second special case is the *unweighted* event study model that imposes uniform weights over units, $\hat{\gamma}^*_{ij} = 1/\|\mathcal{D}_j\|$, as well as over time periods. In this form, Equation (3.12) is the simple average over all two-period, two-group DID estimates averaged over all pre-treatment lags ℓ and donor units i.[10]

Figure 3.5 shows the weighted event study estimates for the teacher collective bargaining application, with $\hat{\nu}$ chosen by applying the procedure in Section 3.3 to the residuals $\dot{Y}$. Figure 3.5a shows the balance possibility frontier for SCM alone and for the weighted event study estimator, as well as the implied imbalance for the event study estimator alone. The frontier

[10]This parallels recent proposals from, among others, Abraham and Sun (2018) and Callaway and Sant'Anna (2018). We can also consider estimating η_t rather than restricting them to be uniform. Following Chapter 2, we could estimate these weights via ridge regression, which would allow for negative weights; we could also restrict these weights to be on the simplex as in Arkhangelsky et al. (2019). We leave a thorough analysis of these estimators to future work.

for the weighted event study estimator is a clear improvement over either the FE or SCM estimates alone, regardless of the level of tuning parameter ν. We see similar results when examining the state-specific fits; see, for example, Appendix Figure B.4. The left of Figure 3.5b shows the placebo estimates from Equation (3.12), where $k < 0$.[11] By design, the augmentation improves pre-treatment fit relative to either the event study model or SCM alone. As with the estimates for partially pooled SCM alone, the weighted event study estimates show no impact of mandatory teacher collective bargaining on student expenditures.

3.6 Simulation study

We now consider the performance of different approaches in a simulation study calibrated to the collective bargaining dataset; we turn to the impacts of mandatory teacher collective bargaining laws in the actual data in the next section. We use the Generalized Synthetic Control Method, implemented in the `R` package `gsynth` (Xu, 2017) to estimate the parameters of simple data generating processes that best fit these data. Specifically, we estimate an interactive two-way fixed effects model with a 2-dimensional latent time-varying factor $\mu_t \in \mathbb{R}^2$ and unit-specific coefficients $\phi_i \in \mathbb{R}^2$:

$$Y_{it} = \text{int} + \text{unit}_i + \text{time}_t + \phi_i' \mu_t + \varepsilon_{it}. \tag{3.14}$$

We estimate (3.14) using untreated units and time periods, then estimate the variance-covariance matrix of the unit fixed effects and factor loadings, $\hat{\Sigma}$, and the variance of the error term $\hat{\sigma}_\varepsilon^2$. We then generate simulated data sets with the same dimensions as the data, $N = 49$ and $T = 39$, using the estimated $\{\widehat{\text{time}}_t, \hat{\mu}_t\}$, and drawing $\{\text{unit}_i, \phi_i\} \overset{\text{iid}}{\sim} \text{MVN}(0, \hat{\Sigma})$ and $\varepsilon_{it} \overset{\text{iid}}{\sim} N(0, \hat{\sigma}_\varepsilon^2)$. We impose a sharp null of no treatment effect, $Y_{it}(1) = Y_{it}(0) = Y_{it}$.

A key component of the simulation model is selection into treatment. We fix the treatment times to be the same as in the teacher unionization application, and set the probability that unit i is treated at each treatment time to be $\pi_i = \text{logit}(\theta_0 + \theta_1(\text{unit}_i + \phi_{i1} + \phi_{i2}))$. For each treatment time, we assign treatment to those units not already treated with probability π_i, sweeping through the fixed set of treatment times. We set $\theta_0 = -2.7$ and $\theta_1 = -1$ to ensure that around 32 units are eventually treated in each simulation draw, following the distribution of the data. We provide additional simulation results under a two-way fixed effects model and a random-effects autoregressive model in Appendix B.2.

We consider several estimators for the average post-treatment effect ATT. Figure 3.6 shows five: (1) A simple difference-in-differences estimator (i.e., an unweighted event study),

[11] These placebo checks differ from those typically performed in traditional event studies, which test for the parallel trends assumption by comparing pre-treatment outcomes between treated and control units. These tests generally have low power, however; see, e.g., Roth (2018); Bilinski and Hatfield (2018); Kahn-Lang and Lang (2019). In contrast, the weighted event study estimator uses pre-treatment outcomes to select donor units that best balance the treated units, in effect optimizing for the placebo test. It is still possible to inspect pre-treatment fit, as in standard SCM, but this is best seen as an assessment of the quality of the match rather than as a formal placebo test.

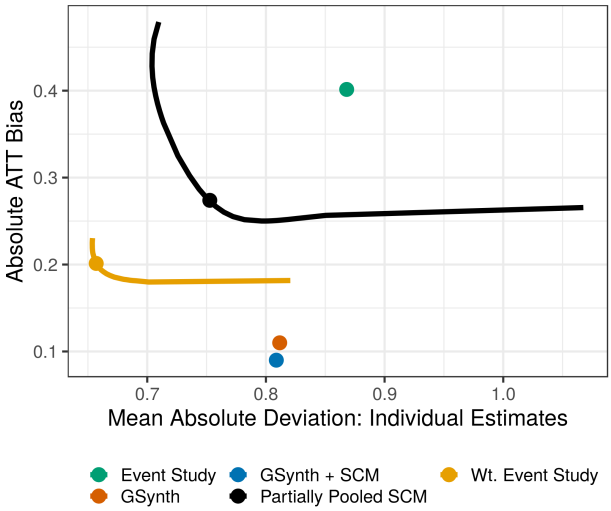


Figure 3.6: Monte Carlo estimates of the bias for the overall ATT vs the MAD of the individual ATT estimates. The lines trace out values for $\nu \in [0, 1]$, the points are the average value using the heuristic $\hat{\nu}$.

(2) the partially pooled SCM estimator, with a "bias frontier" as we vary ν between 0 and 1, (3) the weighted event study estimator that combines fixed effects and partially pooled SCM, again presented as a bias frontier, (4) directly estimating the factor model with `gsynth`, and (5) augmenting partially pooled SCM with `gsynth` using the heuristic value of $\hat{\nu}$. The vertical axis of each panel shows the absolute bias for the ATT, $\left|\mathbb{E}\left[\text{ATT} - \widehat{\text{ATT}}\right]\right|$, while the horizontal axis shows the Mean Absolute Deviation (MAD) of the individual average post-treatment effect estimates, $\mathbb{E}\left[\frac{1}{J}\sum_{W_j=1}|\tau_j - \hat{\tau}_j|\right]$. Appendix Figure B.1 additionally plots the the MAD and RMSE for both estimates.

There are several key takeaways from Figure 3.6. First, the unweighted event study model is misspecified here, and does not do particularly well at controlling either ATT or unit-level bias. Second, partially pooled SCM significantly reduces the bias for the overall ATT relative to separate SCM, and a small amount of pooling also leads to slightly better individual ATT estimates. The gains to pooling, however, diminish for ν close to 1, with the fully pooled SCM yielding poor individual ATT estimates and slightly worse overall ATT estimates than partially pooled SCM. Third, the weighted event study estimator dominates either of the alternatives in terms of both pooled and state-level imbalance. Here again there are gains to partially pooling SCM, although the gains are limited together with the fixed effects augmentation. Finally, as expected, directly estimating the oracle `gsynth` model has lower bias for the overall ATT. However, the MAD for the individual ATT estimates is similar to partially pooled SCM, due to the low number of pre-intervention periods for many of the treated units. Combining `gsynth` and SCM leads to very similar estimates in this case.

Appendix Figure B.2 shows the results for a two-way fixed effects model, where the unweighted event study is the oracle estimator. In this setting we see that the pooled SCM estimate has half the bias of the separate SCM estimate, with no ill effects from pooling. Additionally, all forms of augmentation lead to nearly unbiased estimators. Appendix Figure B.3 shows the results for the random effects AR model. In this setting it is possible to over pool, with both the separate and fully pooled SCM estimators performing worse than partially pooled SCM. In addition, although the fixed effects model is misspecified, the partially pooled weighted event study performs better than either partially pooled SCM or the event study alone.

3.7 Impacts of mandatory teacher collective bargaining laws

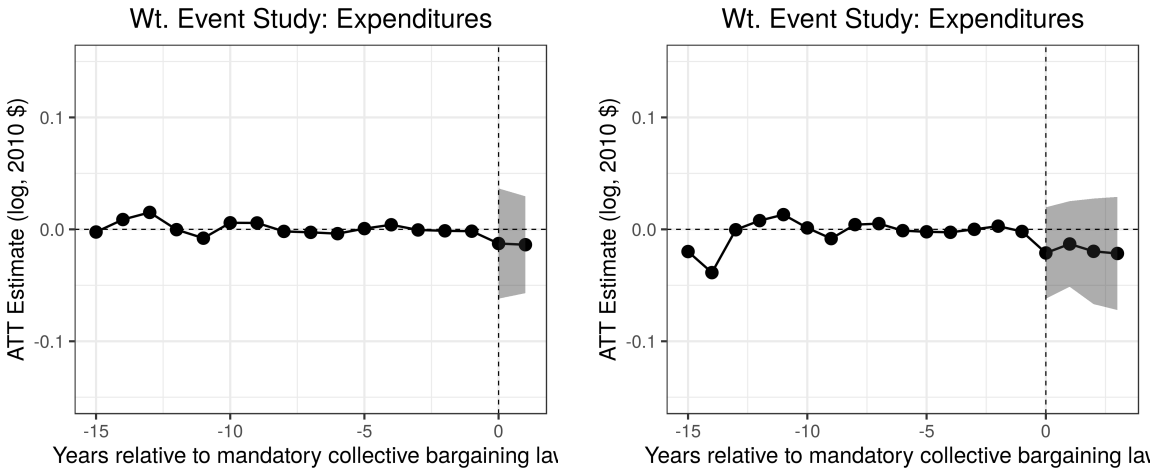
We now return to our primary application of the impact of mandatory teacher collective bargaining. We first consider additional analyses on per-pupil expenditures and then turn to the effects on teacher salary.

Effects on per-pupil expenditures

As we show in Figure 3.5, we find no meaningful effects of mandatory teacher collective bargaining on per-pupil expenditures. Pooled across the ten years after treatment adoption, the overall estimate from the combined event study and SCM model is essentially zero: $\widehat{\text{ATT}} = -0.01$, or a 1 percent reduction in per-pupil expenditures, with an approximate 95% confidence interval of $[-0.049, +0.029]$. Supplementing our main analyses, Appendix Figure B.5 presents corresponding estimates from the generalized synthetic control method (Xu, 2017), both alone and combined with partially pooled SCM, showing similar null results overall. Taken together, these estimates are in stark contrast to the results from Hoxby (1996), who argues for a 12 percent positive effect, although she gives a range of estimates.

We can assess the strength of evidence by conducting robustness and placebo checks. First, following Abadie et al. (2015), we begin by assessing out-of-sample validity via *in time placebo checks*. These checks re-index treatment time to be earlier in order to hold out some pre-treatment time periods (i.e. setting $T_j' = T_j - x$ for some x), then estimate placebo effects for the held-out pre-intervention time periods. Figure 3.7 shows the placebo estimates for the weighted event study estimator with placebo treatment time two and four periods before the true treatment time. Both estimators achieve excellent pre-treatment fit and estimate small negative placebo effects that are indistinguishable from zero.

Next, we consider the result of trimming states with poor pre-treatment fit, following common practice in the matching and SCM literatures. Figure 3.8a shows the state-level fit for both partially pooled SCM and the weighted event study; two states, New York and Alaska, have especially bad pre-treatment fits without augmentation, though interestingly



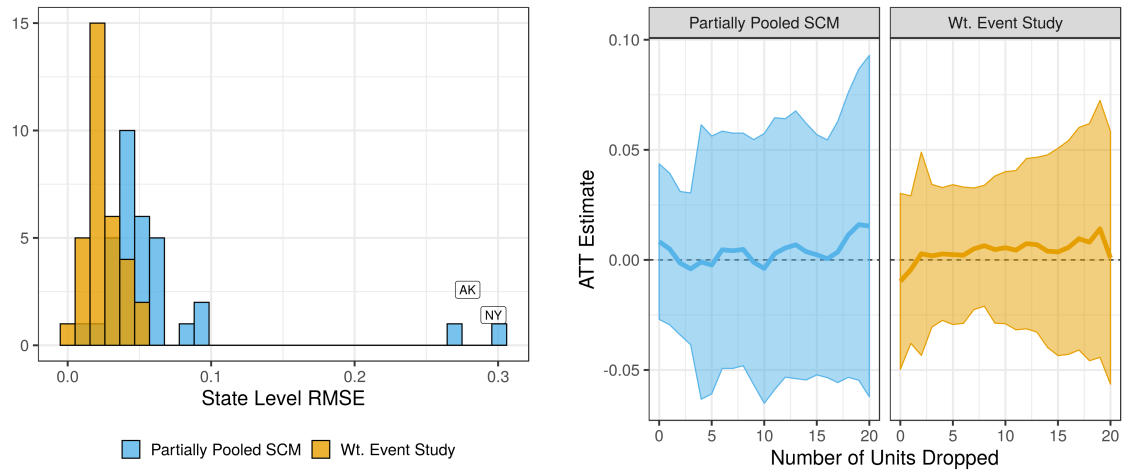
(a) Two year in-time placebo estimates. $\widehat{\text{ATT}} = -0.015$, approximate 95% confidence interval $[-0.076, 0.046]$.

(b) Four year in-time placebo estimates. $\widehat{\text{ATT}} = -0.018$, approximate 95% confidence interval $[-0.067, 0.031]$.

Figure 3.7: Placebo estimates for per-pupil expenditures re-indexing treatment time to (a) two and (b) four years before the true treatment time. The placebo effects are very close to zero and are indistinguishable from zero at this level of precision.

the augmented model fits these states much better. Figure 3.8b shows the overall ATT estimates and 95% confidence intervals when removing an increasing number of treated units with poor fits, in the order of state-level fit shown in Figure 3.8a. We see that the substantive conclusions remain the same.

Appendix B.3 includes several additional analyses of the impact on per-pupil expenditures. First, an important feature of SCM-based methods is that we can directly inspect the weights. Appendix Figures B.6 and B.7 show the state-specific weights over donor states for each treated unit for partially pooled SCM and the weighted event study, respectively. Appendix Figure B.8 shows the number of times each potential donor state is part of a treated state's synthetic control. Taken together, these figures highlight the role of augmentation in constructing more plausible estimators. For the weights from SCM alone, both Illinois and Wyoming are consistently important donor states; after removing the unit fixed effects, the weights are much more evenly distributed across the donor pool, suggesting that estimates are not overly reliant on a single control unit. Finally, we can assess the sensitivity of our estimates to the particular choice of pooling parameter ν. Appendix Figure B.9b shows the overall ATT estimates for partially pooled SCM and the weighted event study estimator varying ν from separate SCM $\nu = 0$ to pooled SCM $\nu = 1$. We see that the partially pooled SCM estimates are more sensitive to the choice of ν, but no choice of ν substantively changes the conclusions for either estimator.



(a) Distribution of state-level fits

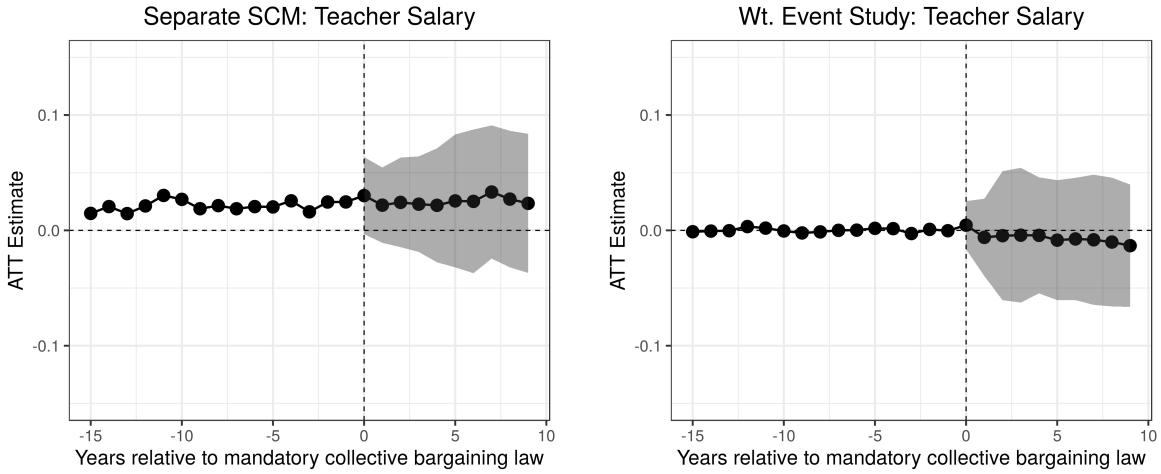
(b) Dropping 1 to 20 treated units according to their worst fit.

Figure 3.8: (a) The distribution of state-level fits (in terms of RMSE) with and without augmentation; Alaska and New York are clear outliers on the original scale, but have similar pre-treatment fits to other states after removing pre-treatment averages. (b) Estimates are not especially sensitivity to dropping an increasing number of units (ranked by pre-treatment imbalance), although the uncertainty intervals are wider with fewer units in the analysis.

Effects on average teacher salary

Thus far, we have focused on the impacts of teacher collective bargaining agreements on expenditures, finding no effect overall. One possible explanation is that school districts are able to divert funds from other purposes to fund higher teacher salaries with no net effect on total expenditures. In Figure 3.9, we therefore repeat the analysis focusing on average teacher salaries (in log, 2010 $) as the outcome of interest. Figure 3.9a shows the separate SCM estimate, which, similar to our discussion in Section 3.3, shows poor balance for average pre-treatment outcomes. By contrast, Figure 3.9b shows the weighted event study estimate, which has excellent pre-treatment balance. Estimates with partially pooled SCM alone are similar.

Consistent with the estimates on expenditures, the estimates from Figure 3.9 do not show any meaningful impact of mandatory teacher collective bargaining on teacher salaries. Specifically, impacts larger than around 0.04 are outside the 95% intervals, even nine years after implementation of the laws. When we average over all post-treatment years, as in Paglayan (2019, Table 2), the estimate is again essentially zero: $\widehat{\text{ATT}} = -0.006$ with an approximate 95% confidence interval of $[-0.051, +0.039]$. While not as severe as for per-pupil expenditures, Hoxby (1996)'s estimate that unions raise teacher salaries by 5 percent is also outside this interval.



(a) Separate SCM

(b) Weighted event study, $\hat{\nu} = 0.22$

Figure 3.9: (a) Separate SCM and (b) weighted event study estimates for the impact of mandatory collective bargaining laws on average teacher salary (log, 2010 $).

3.8 Discussion

In this chapter, we develop a new framework for estimating the impact of a treatment adopted gradually by units over time. In our motivating example, 33 states have enacted laws mandating school districts to bargain with teachers unions (Paglayan, 2019), and we seek to estimate the effects of these laws on educational expenditures and teacher pay. To do so, we adapt SCM to the staggered adoption setting. We argue that current practice of estimating separate SCM weights for each treated unit is unlikely to yield good results, but also that fully pooled SCM may over-correct; our preferred approach, partially pooled SCM, finds weights that balance both state-specific and overall pre-treatment fit. We then augment SCM with a simple average of pre-treatment outcomes, which yields a weighted event study estimator that has advantages over either the event study or SCM estimator alone. We apply this approach to the teacher bargaining example and, consistent with recent analyses, find precisely estimated null effects on teacher salaries and student expenditures.

We briefly note some directions for future work. First, we could extend these ideas to settings with multiple treated units but where treatment can "shut off" for some units, deviating from the staggered adoption structure. This would necessarily require additional assumptions; see, for example, Imai and Kim (2019). We could similarly incorporate other structure from our application. For example, in staggered adoption settings where multiple units adopt treatment at the same time, we could add a layer in the hierarchy and more closely pool units treated at the same time while still partially pooling different treatment cohorts.

Second, many SCM analyses explore multiple outcomes. As in other SCM studies, we

treat each outcome separately, choosing different synthetic control weights for each. In many settings, however, lagged values from one outcome may predict future values of another, suggesting that balancing multiple outcome variables would be useful. This seems especially important in settings like ours with relatively few units.

Finally, we focus on relatively simple outcome models, and in particular a simple pre-treatment average. More complex models are possible and may be desirable. For example, Fesler and Pender (2019) apply the Ridge Augmented SCM proposal in Chapter 2 to a staggered adoption setting, modeling each treated unit separately. Partial pooling may be helpful here. In another direction, we might consider an outcome model that incorporates the time weights used in Arkhangelsky et al. (2019). We anticipate that, unlike in the simple case with unit fixed effects, these augmented approaches likely require more elaborate shrinkage estimation, such as via matrix penalties.

Chapter 4

Varying impacts of letters of recommendation on college admissions

In a pilot study during the 2016-17 admissions cycle, the University of California, Berkeley invited many applicants for freshman admission to submit letters of recommendation. We are interested in estimating how impacts vary for under-represented applicants and applicants with differing *a priori* probability of admission. Assessing treatment effect variation in observational studies is challenging, however, because differences in estimated impacts across subgroups reflect both differences in impacts and differences in covariate balance. To address this, we develop balancing weights that directly optimize for "local balance" within subgroups while maintaining global covariate balance between treated and control populations. We then show that this approach has a dual representation as a form of inverse propensity score weighting with a hierarchical propensity score model. In the UC Berkeley pilot study, our proposed approach yields excellent local and global balance, unlike more traditional weighting methods, which fail to balance covariates within subgroups. We find that the impact of letters of recommendation increases with the predicted probability of admission, with mixed evidence of differences for under-represented minority applicants.

4.1 Introduction and motivation

In a pilot study during the 2016-17 admissions cycle, the University of California, Berkeley invited some applicants for freshman admission to submit letters of recommendation (LORs) as part of their applications. Unlike other highly selective universities, UC Berkeley had never previously asked applicants to submit letters from teachers and guidance counselors. Ideally, these letters would support what the university calls "holistic review": looking beyond reductive summaries (e.g., SAT scores) and examining the whole applicant, taking account of any contextual factors and obstacles overcome (Hout, 2005). However, there was also legitimate concern that applicants from disadvantaged backgrounds might not have access to adults who could write strong letters, and that the use of letters would further disadvantage

these students.

In this chapter, we design an observational study of the impact of submitting a letter of recommendation on subsequent admission using data from this pilot program. Our goal is to understand how these impacts vary for under-represented applicants and for applicants with differing *a priori* probabilities of admission.

Assessing treatment effect variation in observational studies is challenging, even when, as here, subgroups are pre-specified. Variation in estimated impacts reflect both actual treatment effect variation and differences in covariate balance across groups. Traditional Inverse Propensity Score Weighting (IPW) is one standard approach: first estimate a propensity score model via logistic regression, including treatment-by-subgroup interaction terms; construct weights based on the estimated model; and then compare IPW estimates across subgroups (see Green and Stuart, 2014; Lee et al., 2019). Estimated weights from traditional IPW methods, however, are only guaranteed to have good covariate balancing properties asymptotically. Balancing weights estimators, by contrast, instead find weights that directly minimize a measure of covariate imbalance, often yielding better finite sample performance (Zubizarreta, 2015; Athey et al., 2018; Hirshberg and Wager, 2019). Both balancing weights and traditional IPW, however, face a curse of dimensionality when estimating subgroup effects: it is difficult to achieve exact balance on all covariates within each subgroup, or, equivalently, balance all covariate-by-subgroup interactions.

We therefore develop an approximate balancing weights approach tailored to estimating subgroup treatment effects, with a focus on the UC Berkeley LOR pilot study. Specifically, we present a convex optimization problem that finds weights that directly target the level of local imbalance within each subgroup — ensuring *approximate* local covariate balance — while guaranteeing *exact* global covariate balance between the treated and control samples. We show that controlling local imbalance controls the estimation error of subgroup-specific effects, allowing us to better isolate treatment effect variation. We also show that, even when the target estimand is the overall treatment effect, ensuring both exact global balance and approximate local balance reduces the overall estimation error.

Next, we demonstrate that this proposal has a dual representation as inverse propensity weighting with a hierarchical propensity score model, building on recent connections between balancing weights and propensity score estimation (Zhao and Percival, 2017; Tan, 2017; Chattopadhyay et al., 2020). In particular, finding weights that minimize both global and local imbalance corresponds to estimating a propensity score model in which the subgroup-specific parameters are partially pooled toward a global propensity score model. Any remaining imbalance after weighting may lead to bias. To adjust for this, we also combine the weighting approach with an outcome model, analogous to bias correction for matching (Rubin, 1973; Athey et al., 2018).

After assessing its properties, we use this approach to estimate the impacts of letters of recommendation during the 2016 UC Berkeley undergraduate admissions cycle. We focus on variation in the effect on admissions rates based on under-represented minority (URM) status and on the *a priori* predicted probability of admission, estimated using data from the prior year's admissions cycle. First, we show that the proposed weights indeed yield excellent

local and global balance, while traditional propensity score weighting methods yield poor local balance. We then find evidence that the impact of letters increases with the predicted probability of admission. Applicants who are very unlikely to be admitted see little benefit from letters of recommendation while applicants on the cusp of acceptance see a larger, positive impact.

The evidence on the differential effects by URM status is more mixed. Overall, the point estimates for URM and non-URM applicants are close to each other. However, these estimates are noisy and mask important variation by *a priori* probability of admission. For applicants with the highest baseline admission probabilities, we estimate larger impacts for non-URM than URM applicants, though these estimates are sensitive to augmentation with an outcome model. For all other applicants, we estimate the reverse: larger impacts for URM than non-URM applicants. Since URM status is correlated with the predicted probability of admission, this leads to a Simpson's Paradox-type pattern for subgroup effects, with a slightly larger point estimate for non-URM applicants pooled across groups (Bickel et al., 1975; VanderWeele and Knol, 2011).

These results hinge on estimating higher-order interaction terms with the treatment. This suggests caution but also highlights the advantages of a design-based approach (Rubin, 2008). Since we separate the design and analysis phases, we can carefully assess covariate balance and overlap in the subgroups of interest — and can tailor the weights to target these quantities directly. This is a challenge for many recent approaches that use automatic machine learning methods to regularize the complexity of estimated heterogeneous treatment effects (Carvalho et al., 2019). Nonetheless, we view our proposed approach as a complement to — not a substitute for — these approaches and explore an augmented estimator as part of our analysis.

The importance of higher-order interactions also suggests that, as in all observational studies, our results are sensitive to violating the strong assumption of ignorable treatment assignment. Thus, we argue our analysis is a reasonable first look at this question, best understood alongside other approaches that rest on different assumptions (such as those in Rothstein, 2017). In Appendix C.1, we explore one alternative approach that instead leverages unique features of the UC Berkeley pilot study, which included an additional review without the letters of recommendation from a sample of 10,000 applicants. The results from this approach are broadly similar to the estimates from the observational study, differing mainly in regions with relatively poor overlap.

The chapter proceeds as follows. In the next section we introduce the letter of recommendation pilot program at UC Berkeley. Section 4.2 introduces the problem setup and notation, and discusses related work. Section 4.3 proposes and analyzes the approximate balancing weights approach. Section 4.4 presents a simulation study. Section 4.5 presents empirical results on the effect of letters of recommendation. Section 4.6 concludes with a discussion about possible extensions. The appendix includes additional theoretical discussion and analysis.

A pilot program for letters of recommendation in college admissions

As we discuss above, there is considerable debate over the role of letters of recommendation in college admissions. LORs have the potential to offer insight into aspects of the applicant not captured by the available quantitative information or by the essays that applicants submit (Kuncel et al., 2014). At the same time, letters from applicants from under-resourced high school may be less informative or prejudicial against the applicant, due, e.g., to poor writing or grammar, or to lower status of the letter writer; see Schmader et al. (2007) as an example.

The UC Berkeley LOR pilot study is a unique opportunity to assess this question; Rothstein (2017) discusses implementation details. For this analysis, we restrict our sample to non-athlete California residents who applied to either the College of Letters and Science or the College of Engineering at UC Berkeley in the 2016 admissions cycle. This leaves 40,541 applicants, 11,143 of whom submitted LORs. For the purposes of this study, we follow the university in defining a URM applicant as one who is a low-income student, a student in a low-performing high school, a first-generation college student, or from an underrepresented racial or ethnic group. We focus our analysis on the impacts for applicants who both were invited to and subsequently did submit LORs.[1]

Selection into treatment

UC Berkeley uses a two-reader evaluation system. Each reader scores applicants on a three-point scale, as "No," "Possible," or "Yes." Application decisions are based on the combination of these two scores and the major to which a student has applied. In the most selective majors (e.g., mechanical engineering), an applicant typically must receive two "Yes" scores to be admitted, while in others a single "Yes" is sufficient. In the LOR pilot, applicants were invited to submit letters based in part on the first reader score, and the LORs, if submitted, were made available to the second reader.

As in any observational study of causal effects, selection into treatment is central. Decisions to submit letters were a two-step process. Any applicant who received a "Possible" score from the first reader was invited. In addition, due to concerns that first read scores would not be available in time to be useful, an index of student- and school-level characteristics was generated, and applicants with high levels of the index were invited as well.[2]

[1] We could use the methods discussed here to explore a range of different quantities. For this target, the net effect of LORs on admission includes differential rates of submission of a letter given invitation. While non-URM applicants submitted letters at a higher rate than URM applicants, the majority of the discrepancy arises from applicants who were unlikely to be admitted *a priori* (Rothstein, 2017).

[2] The index was generated from a logistic regression fit to data from the prior year's admissions cycle, predicting whether an applicant received a "Possible" score (versus either a "No" or a "Yes"). Applicants with predicted probabilities from this model greater than 50% were invited to submit LORs. Because we observe all of the explanatory variables used in the index, this selection depends only on observable covariates. A small share of applicants with low predicted probabilities received first reads after January 12, 2017, the last date that LOR invitations were sent, and were not invited even if they received "Possible" scores.

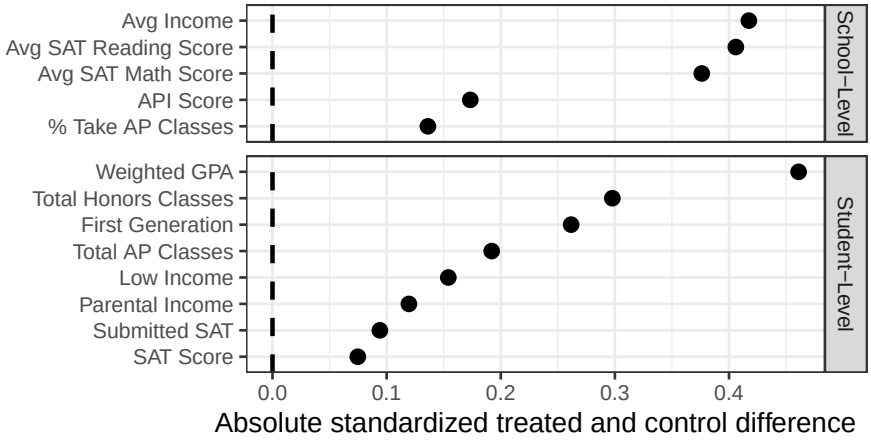


Figure 4.1: Absolute difference in means, standardized by the pooled standard deviation, between applicants submitting and not submitting letters of recommendation for several key covariates. By design, applicants submitting letters of recommendation disproportionately have a "Possible" score from the first reader (70% of treated applicants vs. 4% of untreated applicants).

Of the 40,451 total applicants, 14,596 were invited to submit a letter. Approximately 76% of those invited to submit letters eventually submitted them, and no applicant submitted a letter who was not invited to.

For this analysis, we assume that submission of LORs is effectively random conditional on the first reader score and on both student- and school-level covariates. In particular, the *interaction* between the covariates and the first reader score plays an important role in the overall selection mechanism, as applicants who received a score of "No" or "Yes" from the first reader could still have been asked to submit an LOR based on their individual and school information. Figure 4.1 shows covariate imbalance for several key covariates — measured as the absolute difference in means divided by the pooled standard deviation — for applicants who submitted LORs versus those who did not.[3] We see that there are large imbalances in observable applicant characteristics, most notably average school income, GPA, the number of honors and AP classes taken, and SAT score. There were also large imbalances in first reader scores (not shown in Figure 4.1): 70% of applicants that submitted

[3]The full set of student-level variables we include in our analysis are: weighted and unweighted GPA, GPA percentile within school, parental income and education, SAT composite score and math score, the number of honors courses and percentage out of the total available, number of AP courses, ethnic group, first generation college student status, and fee waiver status. The school level variables we control for are: average SAT reading, writing, and math scores, average ACT score, average parental income, percent of students taking AP classes, and the school Academic Performance Index (API) evaluated through California's accountability tests. For students that did not submit an SAT score but did submit an ACT score, we imputed the SAT score via the College Board's SAT to ACT concordance table. For the 992 applicants with neither an SAT nor an ACT score, we impute the SAT score as the average among applicants from the school.

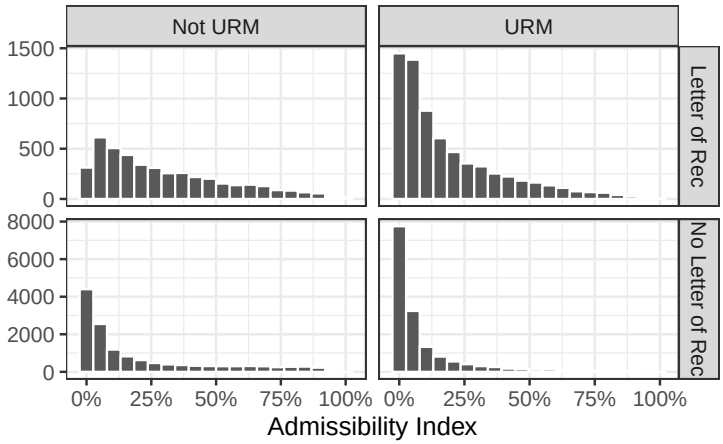


Figure 4.2: Distribution of the "admissibility index" — an estimate of the *a priori* probability of acceptance — for the 2016 UC Berkeley application cohort, separated into URM and non-URM and those that submitted a letter versus those that did not.

LORs had "Possible" scores, compared to only 4% of those who did not.

Heterogeneity across *a priori* probability of admission

To better understand who was invited to submit LORs and any differential impacts between URM and non-URM applicants, we construct a univariate summary of applicant- and school-level characteristics. We use logistic regression to estimate the probability of admission given observable characteristics using the *prior year* (2015) admissions data.[4] We then use this model to predict *a priori* admissions probabilities for the applicants of interest in 2016; we refer to these predicted probabilities as the Admissibility Index (AI). The overall AUC in predicting 2016 admissions is 0.88 and the mean square error is 10% (see Appendix Table C.1). However, the predictive accuracy decreases for higher AI applicants, slightly under-estimating the probability of admissions for middle-tier applicants and over-estimating for the highest admissibility applicants (see Appendix Figure C.1). Additionally, predictive performance is better for URM applicants than non-URM applicants, particularly for applicants to the College of Engineering (see Appendix Figure C.2).

Figure 4.2 shows the AI distribution for the 2016 applicant cohort, broken out by URM status and LOR submission. There are several features of this distribution that have important implications for our analysis. First, although the probability of admission is quite low overall, applicants across nearly the full support of probabilities submitted LORs. This is primarily because applicants who received "Possible" scores from the first readers come from a wide range of admissibility levels. This will allow us to estimate heterogeneous effects across the full distribution, with more precision for applicants with lower AIs. Second,

[4]This is a different model than the logistic regression used by the admissions office, which predicted a reviewer score of "Possible" rather than admission.

AI Range	URM	Number of Applicants	Number Submitting LOR	Proportion Treated
$< 5\%$	URM	11,832	2,157	18%
	Not URM	6,529	607	9%
5% - 10%	URM	3,106	1,099	35%
	Not URM	2,099	536	25%
10% - 20%	URM	2,876	1,212	42%
	Not URM	2,495	828	33%
$> 20\%$	URM	4,645	2,345	50%
	Not URM	6,959	2,359	34%

Table 4.1: Number of applicants and proportion treated by subgroup.

because the admissions model disproportionately predicted that URM students had high chances of receiving "Possible" scores, many more URM applicants were invited to submit letters than non-URM applicants, and so our estimates for URM applicants will be more precise than those for non-URM applicants.

From Figure 4.2 we know that the distribution of AI varies between URM and non-URM applicants, and so apparent differences in estimated effects between the two groups may be due to compositional differences. Therefore, in the subsequent sections we will focus on estimating effects within subgroups defined by both URM status and admissibility. To do this, we define subgroups by creating four (non-equally-sized) strata of the AI: $< 5\%$, $5\% - 10\%$, $10\% - 20\%$ and $> 20\%$. Interacting with URM status, this leads to eight non-overlapping subgroups; we will marginalize over these to estimate the other subgroup effects above. Table 4.1 shows the total number of applicants in each of the eight groups, along with the proportion submitting letters of recommendation. As we discuss in Section 4.5, we will further divide each of these subgroups by first reader score and college, to ensure exact balance on these important covariates.

4.2 Treatment effect variation in observational studies

Setup and estimands

We now describe the letter of recommendation study as an observational study where for each applicant $i = 1, \ldots, n$, we observe applicant and school level-covariates $X_i \in \mathcal{X}$; a group indicator $G_i \in \{1, \ldots, K\}$ denoting e.g., URM status or coarsened AI; a binary indicator for submitting a letter of recommendation $W_i \in \{0, 1\}$; and whether the applicant is admitted, which we denote as $Y_i \in \{0, 1\}$. We assume that for each applicant, (X_i, G_i, W_i, Y_i) are sampled i.i.d. from some distribution $\mathcal{P}(\cdot)$. Additionally, let n_{1g} and n_{0g} be the number of treated and control units in subgroup $G_i = g$, respectively. Following the potential outcomes

framework (Neyman, 1923; Holland, 1986), we assume SUTVA (Rubin, 1980) and posit two potential outcomes $Y_i(0)$ and $Y_i(1)$ for each applicant i, corresponding to i's outcome if that applicant submits a letter of recommendation or not, respectively; the observed outcome is $Y_i = W_i Y_i(1) + (1 - W_i)Y_i(0)$.[5] In this study we are interested in estimating two types of effects. First, we wish to estimate the overall Average Treatment Effect on the Treated (ATT), the treatment effect for applicants who submit a letter,

$$\tau = \mathbb{E}[Y(1) - Y(0) \mid W = 1],$$

where we denote $\mu_1 = \mathbb{E}[Y(1) \mid W = 1]$ and $\mu_0 = \mathbb{E}[Y(0) \mid W = 1]$. Second, for each subgroup $G_i = g$, we would like to estimate the Conditional ATT (CATT),

$$\tau_g = \mathbb{E}[Y(1) - Y(0) \mid G = g, W = 1], \tag{4.1}$$

where similarly we denote $\mu_{1g} = \mathbb{E}[Y(1) \mid G = g, W = 1]$ and $\mu_{0g} = \mathbb{E}[Y(0) \mid G = g, W = 1]$.

Estimating μ_{1g} is relatively straightforward: we can simply use the average outcome for treated units in group g, $\hat{\mu}_{1g} \equiv \frac{1}{n_{1g}} \sum_{G_i=g} W_i Y_i$. However, estimating μ_{0g} is more difficult due to confounding; we focus much of our discussion on imputing this counterfactual mean for the group of applicants who submitted letters of recommendation. To do this, we rely on two key assumptions that together form the usual *strong ignorability* assumption (Rosenbaum and Rubin, 1983).

Assumption 4.1 (Ignorability)**.** The potential outcomes are independent of treatment given the covariates and subgroup:

$$Y(1), Y(0) \perp\!\!\!\perp W \mid X, G. \tag{4.2}$$

Assumption 4.2 (One Sided Overlap)**.** The *propensity score* $e(x, g) \equiv P(W = 1 \mid X = x, G = g)$ is less than 1:

$$e(X, G) < 1. \tag{4.3}$$

In our context, Assumption 4.1 says that conditioned on the first reader score and applicant- and school-level covariates, submission of an LOR is independent of the potential admissions outcomes. Due to the selection mechanism we describe in Section 4.1, we believe that this is a reasonable starting point for estimating these impacts; see Rothstein (2017) and Appendix C.1 for alternatives. Assumption B.2 corresponds to assuming that no applicant would have been guaranteed to submit a letter of recommendation. Although some applicants were guaranteed to be *invited* to submit an LOR, we believe that this is a reasonable assumption for actually submitting a letter. In Section 4.5 we assess overlap empirically.

With this setup, let $m_0(x, g) = \mathbb{E}[Y(0) \mid X = x, G = g]$ be the *prognostic score*, the expected control outcome conditioned on covariates X and group membership G. Under Assumptions 4.1 and B.2, we have the standard identification result:

$$\mu_{0g} = \mathbb{E}[m_0(X, G) \mid W = 1] = \mathbb{E}\left[\frac{e(X, G)}{1 - e(X, G)} Y \mid W = 0\right]. \tag{4.4}$$

[5] There is a possibility of interference induced by the number of admitted applicants being capped. With 6874 admitted students, we consider the potential interference to be negligible

Therefore we can obtain a plug-in estimate for μ_{0g} with an estimate of the prognostic score, $m_0(\cdot,\cdot)$, an estimate of propensity score, $e(\cdot,\cdot)$, or an estimate of the treatment odds themselves, $\frac{e(\cdot,\cdot)}{1-e(\cdot,\cdot)}$. We next review existing methods for such estimation, turning to our proposed weighting approach in the following section.

Related work: methods to estimate subgroup treatment effects

There is an extensive literature on estimating varying treatment effects in in observational studies; see Anoke et al. (2019) and Carvalho et al. (2019) for recent discussions. This is an active area of research, and we narrow our discussion here to methods that assess heterogeneity across pre-defined, discrete subgroups. In particular, we will focus on linear weighting estimators that take a set of weights $\hat{\gamma} \in \mathbb{R}^n$, and estimate μ_{0g} as a weighted average of the control outcomes in the subgroup:

$$\hat{\mu}_{0g} \equiv \frac{1}{n_{1g}} \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) Y_i. \tag{4.5}$$

Many estimators take this form; we focus on design-based approaches that do not use outcome information in constructing the estimators (Rubin, 2008). See Hill (2011); Künzel et al. (2019); Carvalho et al. (2019); Nie and Wager (2019) for discussions of approaches that instead focus on outcome modeling.

Methods based on estimated propensity scores. A canonical approach in this setting is Inverse Propensity Weighting (IPW) estimators for μ_{0g} (see Green and Stuart, 2014). Traditionally, this proceeds in two steps: first estimate the propensity score $\hat{e}(x, g)$, e.g. via logistic regression; second, estimate μ_{0g} as in Equation (4.5), with weights $\hat{\gamma}_i = \frac{\hat{e}(X_i,G_i)}{1-\hat{e}(X_i,G_i)}$:

$$\hat{\mu}_{0g} = \sum_{W_i=0, G_i=g} \frac{\hat{e}(X_i, G_i)}{1 - \hat{e}(X_i, G_i)} Y_i \tag{4.6}$$

where these are "odds of treatment" weights to target the ATT. A natural approach to estimating $\hat{e}(X_i, G_i)$, recognizing that G_i is discrete, is to estimate a logistic model for treatment separately for each group or, equivalently, with full interactions between G_i and (possibly transformed) covariates $\phi(X_i) \in \mathbb{R}^p$:

$$\text{logit}(e(x, g)) = \alpha_g + \beta_g \cdot \phi(x). \tag{4.7}$$

Due to the high-dimensional nature of the problem, it is often infeasible to estimate Equation (4.7) without any regularization: the treated and control units might be completely separated, particularly when some groups are small. Classical propensity score modeling with random effects is one common solution, but can be numerically unstable in settings similar to this (Zubizarreta and Keele, 2017). Other possible solutions in high dimensions include

L^1 penalization (Lee et al., 2019), hierarchical Bayesian modeling (Li et al., 2013), and generalized boosted models (McCaffrey et al., 2004). In addition, Dong et al. (2020) propose a stochastic search algorithm to estimate a similar model when the number of subgroups is large, and Li (2017) and Yang et al. (2020) propose *overlap weights*, which upweight regions of greater overlap. We explore overlap weights further in Section 4.5.

Under suitable assumptions and conditions, methods utilizing the estimated propensity score will converge to the true ATT asymptotically. However, in high dimensional settings with a moderate number of subgroups these methods can often fail to achieve good covariate balance in the sample of interest; as we show in Section 4.5, these methods fail to balance covariates in the UC Berkeley LOR study. The key issue is that traditional IPW methods focus on estimating the propensity score itself (i.e., the conditional probability of treatment) rather than finding weights that achieve good in-sample covariate balance.

Balancing weights. Unlike traditional IPW, balancing weights estimators instead find weights that directly target in-sample balance. One example is the Stable Balancing Weights (SBW) proposal from Zubizarreta (2015), which finds the minimum variance weights that achieve a user-defined level of covariate balance in $\phi(X_i) \in \mathbb{R}^p$:

$$\begin{aligned} \min_{\gamma} \quad & \|\gamma\|_2^2 \\ \text{subject to} \quad & \max_j \left| \frac{1}{n_1} \sum_{W_i=1} \phi_j(X_i) - \frac{1}{n_1} \sum_{W_i=0} \gamma_i \phi_j(X_i) \right| \leq \delta, \end{aligned} \tag{4.8}$$

for weights γ, typically constrained to the simplex, and for allowable covariate imbalance δ. These methods have a long history in calibrated survey weighting (see, e.g. Deming and Stephan, 1940; Deville et al., 1993), and have recently been extensively studied in the observational study context (e.g. Hainmueller, 2011; Zubizarreta, 2015; Athey et al., 2018; Hazlett, 2020; Hirshberg et al., 2019). They have also been shown to estimate the propensity score with a loss function designed to achieve good balance (Zhao and Percival, 2017; Wang and Zubizarreta, 2019; Chattopadhyay et al., 2020).

While balancing weights achieve better balance than the traditional IPW methods above, we must take special care to use them appropriately when estimating subgroup treatment effects. As we will show in Section 4.5, designing balancing weights estimators without explicitly incorporating the subgroup structure also fails to balance covariates within subgroups in the LOR study. We turn to designing such weights in the next section.

4.3 Approximate balancing weights for treatment effect variation

Now we describe a specialization of balancing weights that minimizes the bias for subgroup treatment effect estimates. This approach incorporates the subgroup structure into the

balance measure and optimizes for the "local balance" within each subgroup. First we show that the error for the subgroup treatment effect estimate is bounded by the level of local imbalance within the subgroup. Furthermore, the error for estimating the overall ATT depends on both the global balance and the local balance within each subgroup. We then describe a convex optimization problem to minimize the level of imbalance within each subgroup while ensuring exact global balance in the full sample. Next, we connect the procedure to IPW with a hierarchical propensity score model, using the procedure's Lagrangian dual formulation. We conclude by describing how to augment the weighting estimate with an outcome model.

Local balance, global balance, and estimation error

Subgroup effects

We initially consider the role of local imbalance in estimating subgroup treatment effects. This is the subgroup-specific specialization of standard results in balancing weights. We will compare the estimate $\hat{\mu}_{0g}$ to $\tilde{\mu}_{0g} \equiv \frac{1}{n_{1g}} \sum_{G_i=g} W_i m_0(X_i, g)$, our best approximation to μ_{0g} if we knew the true prognostic score. Defining the residual $\varepsilon_i = Y_i - m_0(X_i, G_i)$, the error is

$$\hat{\mu}_{0g} - \tilde{\mu}_{0g} = \underbrace{\frac{1}{n_{1g}} \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) m_0(X_i, g) - \frac{1}{n_{1g}} \sum_{G_i=g} W_i m_0(X_i, g)}_{\text{bias}_g} + \underbrace{\frac{1}{n_{1g}} \sum_{G_i=g} (1 - W_i) \hat{\gamma}_i \varepsilon_i}_{\text{noise}} . \tag{4.9}$$

Since the weights $\hat{\gamma}$ are *design-based*, they will be independent of the outcomes, and the noise term will be mean-zero and have variance proportional to the sum of the squared weights $\frac{1}{n_{1g}^2} \sum_{G_i=g} (1 - W_i) \hat{\gamma}_i^2$.[6] At the same time, the conditional bias term, bias_g, depends on the imbalance in the true prognostic score $m_0(X_i, G_i)$. The idea is to bound this imbalance by the worst-case imbalance in all functions m in a model class $\mathcal{M}$. While the setup is general,[7] we describe the approach assuming that the prognostic score within each subgroup is a linear function of transformed covariates $\phi(X_i) \in \mathbb{R}^p$ with L^2-bounded coefficients; i.e., $\mathcal{M} = \{m_0(x, g) = \eta_g \cdot \phi(x) \mid \|\eta_g\|_2 \leq C\}$. We can then bound the bias by the level of *local imbalance* within the subgroup via the Cauchy-Schwarz inequality:

$$|\text{bias}_g| \leq C \underbrace{\left\| \frac{1}{n_{1g}} \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i=g} W_i \phi(X_i) \right\|_2}_{\text{local imbalance}} . \tag{4.10}$$

[6] In the general case with heteroskedastic errors, the variance of the noise term is $\frac{1}{n_{1g}^2} \sum_{G_i=g} \hat{\gamma}_i^2 \text{Var}(\varepsilon_i) \leq \max_i\{\text{Var}(\varepsilon_i)\} \frac{1}{n_{1g}^2} \sum_{G_i=g} \hat{\gamma}_i^2$.

[7] See Wang and Zubizarreta (2019) for the case where the prognostic score can only be approximated by a linear function; see Hazlett (2020) for a kernel representation and Hirshberg et al. (2019) for a general nonparametric treatment.

Based on Equation (4.10), we could control local bias solely by controlling local imbalance. This approach would be reasonable if we were solely interested in subgroup impacts. In practice, however, we are also interested in the overall effect, as well as in aggregated subgroup effects, such as the impact for all URM applicants, not just the specific URM $\times$ AI stratum. We can estimate these aggregated effects by taking a weighted average of the subgroup-specific estimates, e.g. we estimate μ_{0g} as $\hat{\mu}_0 = \sum_{g=1}^{K} \frac{n_{1g}}{n_1} \hat{\mu}_{0g} = \frac{1}{n_1} \sum_{W_i=0} n_{1G_i} \hat{\gamma}_i Y_i$. As we show in in both the simulations in Section 4.4 and the analysis of the LOR pilot study in Section 4.5, incorporating the global balance constraint leads to negligible changes in the level of local balance and the performance of the subgroup estimators, but can lead to large improvements in the global balance and the performance of the overall estimate. Thus, there seems to be little downside in terms of subgroup estimates from an approach that controls both local and global imbalance — but large gains for overall estimates, as we discuss next.

Overall treatment effect

The imbalance within each subgroup continues to play a key role in estimating the overall treatment effect, alongside global balance. To see this, we again compare to our best estimate if we knew the prognostic score, $\tilde{\mu}_0 = \frac{1}{n_1} \sum_{g=1}^{K} n_{1g} \tilde{\mu}_{0g}$, and see that the local imbalance plays a part. The error is

$$\begin{aligned} \hat{\mu}_0 - \tilde{\mu}_0 = \bar{\eta} \cdot \left(\frac{1}{n_1} \sum_{i=1}^{n} n_{1G_i} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_1} \sum_{i=1}^{n} W_i \phi(X_i) \right) \; + \\ \frac{1}{n_1} \sum_{g=1}^{k} n_{1g} \left(\eta_g - \bar{\eta} \right) \cdot \left(\sum_{G_i=g} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i=g} W_i \phi(X_i) \right) \; + \\ \frac{1}{n_1} \sum_{i=1}^{n} \hat{\gamma}_i (1 - W_i) \varepsilon_i, \end{aligned} \tag{4.11}$$

where $\bar{\eta} \equiv \frac{1}{K} \sum_{g=1}^{K} \eta_g$ is the average of the model parameters across all subgroups. Again using Cauchy-Schwarz we see that the overall bias is controlled by the *local imbalance* within each subgroup as well as the *global balance* across subgroups:

$$\begin{aligned} |\text{bias}| \leq \|\bar{\eta}\|_2 \underbrace{\left\| \frac{1}{n_1} \sum_{i=1}^{n} n_{1G_i} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_1} \sum_{i=1}^{n} W_i \phi(X_i) \right\|_2}_{\text{global balance}} + \\ \sum_{g=1}^{G} \frac{n_{1g}}{n_1} \|\eta_g - \bar{\eta}\|_2 \underbrace{\left\| \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i=g} W_i \phi(X_i) \right\|_2}_{\text{local balance}}. \end{aligned} \tag{4.12}$$

In general, we will want to achieve *both* good local balance within each subgroup and good global balance across subgroups. Equation (4.12) shows that the relative importance of local

and global balance for estimating the overall ATT is controlled by the level of similarity in the outcome process across groups. In the extreme case where the outcome process does not vary across groups — i.e., $\eta_g = \bar{\eta}$ for all g — then controlling the global balance is sufficient to control the bias. In the other extreme where the outcome model varies significantly across subgroups — e.g., $\|\eta_g - \bar{\eta}\|_2$ is large for all g — we will primarily seek to control the local imbalance within each subgroup in order to control the bias for the ATT. Typically, we expect that interaction terms are weaker than "main effects," i.e., $\|\eta_g - \bar{\eta}\|_2 < \|\bar{\eta}\|_2$ (see Cox, 1984; Feller and Gelman, 2015). As a result, our goal is to find weights that prioritize global balance while still achieving good local balance.

Optimizing for both local and global balance

We now describe a convex optimization procedure to find weights that optimize for local balance while ensuring exact global balance across the sample. The idea is to stratify across subgroups and find approximate balancing weights within each stratum, while still constraining the overall level of balance. In our setting, we stratify on first reader score, URM status, the coarsened AI measure, and the college that the applicant is applying to; see Section 4.5. We then find weights $\hat{\gamma}$ that solve the following optimization problem:

$$\begin{aligned} \min_{\gamma} \quad & \sum_{g=1}^{K} \left\| \sum_{G_i=g, W_i=0} \gamma_i \phi(X_i) - \sum_{G_i=g, W_i=1} \phi(X_i) \right\|_2^2 + \frac{\lambda_g}{2} \sum_{G_i=G, W_i=0} \gamma_i^2 \\ \text{subject to} \quad & \sum_{W_i=0} \gamma_i \phi(X_i) = \sum_{W_i=1} \phi(X_i) \\ & \sum_{G_i=G, W_i=0} \gamma_i = n_{1g} \\ & \gamma_i \geq 0 \quad \forall i = 1, \ldots, n \end{aligned} \tag{4.13}$$

The optimization problem (4.13) has several key components. First, following Equation (4.10) we try to find weights that minimize the local imbalance for each stratum defined by G; this is a proxy for the stratum-specific bias. We also constrain the weights to *exactly balance* the covariates globally over the entire sample. Equivalently, this finds weights that achieve exact balance marginally on the covariates $\phi(X_i)$ and only approximate balance for the interaction terms $\phi(X_i) \times \mathbb{1}_{G_i}$, placing greater priority on main effects than interaction terms. Taken together, this ensures that we are minimizing the overall bias as well as the bias within each stratum. In principle, weights that exactly balance the covariates within each stratum would also yield exact balance globally. Typically, however, the sample sizes are too small to achieve exact balance within each stratum, and so this combined approach

at least guarantees global balance.[8] From Equation (4.12), we can see that if there is a limited amount of heterogeneity in the baseline outcome process across groups, the global exact balance constraint will limit the estimation error when estimating the ATT, even if local balance is relatively poor. While we choose to enforce exact global balance, we could also limit to *approximate* global balance, with the relative importance of local and global balance controlled by an additional hyperparameter set by the analyst.

Second, we include an L^2 regularization term that penalizes the sum of the squared weights in the stratum; from Equation (4.9), we see that this is a proxy for the variance of the weighting estimator. For each stratum, the optimization problem includes a hyperparameter λ_g that negotiates the bias-variance tradeoff within that stratum. When λ_g is small, the optimization prioritizes minimizing the bias through the local imbalance, and when λ is large it prioritizes minimizing the variance through the sum of the squared weights. As a heuristic, we set $\lambda_g = \frac{1}{n_g}$: for larger strata where better balance is possible, this heuristic will prioritize balance — and thus bias — over variance; for smaller strata, by contrast, this will prioritize lower variance.

We also incorporate two additional constraints on the weights. We include a fine balance constraint (Rosenbaum et al., 2007): within each stratum the weights sum up to the number of treated units in that stratum, n_{1g}. Since each stratum maps to only one subgroup, this also guarantees that the weights sum to the number of treated units in each subgroup. We also restrict the weights to be non-negative, which stops the estimates from extrapolating outside of the support of the control units (King and Zeng, 2006). Together, these induce several stability properties, including that the estimates are sample bounded.

In our setting the strata G are part of a hierarchy: each stratum is a unique combination of first reader score, URM status, admissibility group, and college. Thus, we could also extend the optimization problem in Equation (4.13) to balance intermediate levels between global balance and local balance. Incorporating additional balance constraints for each intermediate level, is unwieldy in practice due to the proliferation of hyperparameters. Instead, we expand $\phi(x)$ to include additional interaction terms between covariates and levels of the hierarchy. In our application, we interact the admissibility index with both URM status and the AI group, which means that we exactly balance AI within each URM-AI group.

Finally, we compute the variance of our estimator conditioned on the design $(X_1, Z_1, W_1), \ldots, (X_n, Z_n, W_n)$ or, equivalently, conditioned on the weights. The conditional variance is

$$\mathrm{Var}(\hat{\mu}_{0g} \mid \hat{\gamma}) = \frac{1}{n_{1g}^2} \sum_{G_i=g} (1 - W_i)\hat{\gamma}_i^2 \mathrm{Var}(Y_i). \tag{4.14}$$

Using the i^{th} residual to estimate $\mathrm{Var}(Y_i)$ yields the empirical sandwich estimator for the treatment effect

$$\widehat{\mathrm{Var}}(\hat{\mu}_{1g} - \hat{\mu}_{0g} \mid \hat{\gamma}) = \frac{1}{n_{1g}^2} \sum_{G_i=g} W_i(Y_i - \hat{\mu}_{1g})^2 + \frac{1}{n_{1g}^2} \sum_{G_i=g} (1 - W_i)\hat{\gamma}_i^2 (Y_i - \hat{\mu}_{0g})^2, \tag{4.15}$$

[8] This constraint induces a dependence across the strata, so that the optimization problem does not decompose into J sub-problems.

where, as above, $\hat{\mu}_{1g}$ is the average outcome for applicants in subgroup g who submit an LOR. This is the fixed-design Huber-White heteroskedastic robust standard error for the weighted average. See Hirshberg et al. (2019) for discussion on asymptotic normality and semi-parametric efficiency for estimators of this form.

Dual relation to partially pooled propensity score estimation

Thus far, we have motivated the approximate balancing weights approach by appealing to the connection between local bias and local balance. We now draw on recent connections between approximate balancing weights and (calibrated) propensity score estimation through the Lagrangian dual problem. The weights that solve optimization problem (4.13) correspond to estimating the inverse propensity weights with a (truncated) linear odds function with the stratum G interacted with the covariates $\phi(X)$,[9]

$$\frac{P(W=1 \mid X=x, G=g)}{1-P(W=1 \mid X=x, G=g)} = \left[\alpha_g + \beta_g \cdot \phi(x)\right]_+, \tag{4.16}$$

where the coefficients β_g are *partially pooled* towards a global model.

To show this, we first derive the Lagrangian dual. For each stratum g, the sum-to-n_{1g} constraint induces a dual variable $\alpha_g \in \mathbb{R}$, and the local balance measure induces a dual variable $\beta_g \in \mathbb{R}^p$. These dual variables are part of the *balancing loss function* for stratum z:

$$\mathcal{L}_g(\alpha_g, \beta_g) \equiv \sum_{W_i=0, G_i=g} \left[\alpha_g + \beta_g \cdot \phi(X_i)\right]_+^2 - \sum_{W_i=1, G_i=g} \left(\alpha_g + \beta_g \cdot \phi(X_i)\right), \tag{4.17}$$

where $[x]_+ = \max\{0, x\}$. With this definition we can now state the Lagrangian dual.

Proposition 4.1. With $\lambda_g > 0$, if a feasible solution to (4.13) exists, the Lagrangian dual is

$$\min_{\alpha, \beta_1, \ldots, \beta_J, \mu_\beta} \sum_{g=1}^{K} \underbrace{\mathcal{L}_g(\alpha_g, \beta_g)}_{\text{balancing loss}} + \underbrace{\sum_{z=1}^{J} \frac{\lambda_g}{2} \|\beta_g - \mu_\beta\|_2^2}_{\text{shrinkage to global variable}}. \tag{4.18}$$

If $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_J$ are the solutions to the dual problem, then the solution to the primal problem (4.13) is

$$\hat{\gamma}_i = \left[\hat{\alpha}_{Z_i} + \hat{\beta}_{Z_i} \cdot \phi(X_i)\right]_+. \tag{4.19}$$

The Lagrangian dual formulation sheds additional light on the approximate balancing weights estimator. First, applying results on the connection between approximate balancing weights and propensity score estimation (e.g., Zhao and Percival, 2017; Wang and Zubizarreta, 2019; Hirshberg and Wager, 2019; Chattopadhyay et al., 2020), we see that this

[9] The truncation arises from constraining weights to be non-negative, and the linear odds form arises from penalizing the L^2 norm of the weights. We can consider other penalties that will lead to different forms.

approach estimates propensity scores of the form (4.16). This corresponds to a fully interacted propensity score model where the coefficients on observed covariates vary across strata. Recall that we find *approximate* balancing weights for each stratum because the number of units per stratum might be relatively small; therefore we should not expect to be able to estimate this fully interacted propensity score well.

The dual problem in Equation (4.18) also includes a global dual variable μ_β induced by the global balance constraint in the primal problem (4.13). Because we enforce *exact* global balance, this global model is not regularized. However, by penalizing the deviations between the stratum-specific variables and the global variables via the L^2 norm, $\|\beta_g - \mu_\beta\|_2^2$, the dual problem *partially pools* the stratum-specific parameters towards a global model. Thus, we see that the approximate balancing weights problem in Equation (4.13) corresponds to a hierarchical propensity score model (see, e.g. Li et al., 2013), as in Section 4.2, fit with a loss function designed to provide covariate balance. Excluding the global constraint removes the global dual variable μ_β, and the dual problem shrinks the stratum-specific variables β_g towards zero without any pooling. In contrast, ignoring the local balance measure by setting $\lambda_g \to \infty$ constrains the stratum-specific variables β_g to all be *equal* to the global variable μ_β, resulting in a fully pooled estimator.

Finally, recall that in the primal problem (4.13), the hyperparameter λ_g controlled the bias-variance tradeoff within stratum z between prioritizing local balance or effective sample size. In the dual problem λ_g performs the same role by controlling the level of partial pooling. When λ_g is large the dual parameters are heavily pooled towards the global model, and when λ_g is small the level of pooling is reduced. By setting $\lambda_g = \frac{1}{n_g}$ as above, larger strata will be pooled less than smaller strata.[10]

Augmentation with an outcome estimator

The balancing weights we obtain via the methods above may not achieve perfect balance, leaving the potential for bias. We can augment the balancing weights estimator with an outcome model, following Chapter 2 and other similar proposals in a variety of settings (see, e.g. Athey et al., 2018; Hirshberg and Wager, 2019). Analogous to bias correction for matching (Rubin, 1973) or model-assisted estimation in survey sampling (Särndal et al., 2003), the essential idea is to adjust the weighting estimator using an estimate of the bias. Specifically, we can estimate the prognostic score $m_0(x, g)$ with a working model $\hat{m}_0(x, g)$, e.g., with a flexible regression model. An estimate of the bias in group g is then:

$$\widehat{\text{bias}}_g = \frac{1}{n_{1g}} \sum_{W_i=1, G_i=g} \hat{m}_0(X_i, g) - \frac{1}{n_{1g}} \sum_{W_i=0, G_i=g} \hat{\gamma}_i \hat{m}_0(X_i, g). \tag{4.20}$$

This is the bias due to imbalance in estimated prognostic score in group g *after* weighting. With this estimate of the bias, we can explicitly bias-correct our weighting estimator,

[10] It is also possible to have covariate-specific shrinkage by measuring imbalance in the primal problem (4.13) with a *weighted* L^2 norm, leading to an additional p hyper-parameters. We leave exploring this extension and hyper-parameter selection methods to future work.

estimating μ_{0g} as

$$\begin{aligned}\hat{\mu}_{0g}^{\text{aug}} &\equiv \hat{\mu}_{0g} + \widehat{\text{bias}}_g \\ &= \frac{1}{n_{1g}} \sum_{W_i=0, G_i=g} \hat{\gamma}_i Y_i + \left[\frac{1}{n_{1g}} \sum_{W_i=1, G_i=g} \hat{m}_0(X_i, g) - \frac{1}{n_{1g}} \sum_{W_i=0, G_i=g} \hat{\gamma}_i \hat{m}_0(X_i, g) \right].\end{aligned} \tag{4.21}$$

Thus, if the balancing weights fail to achieve good covariate balance in a given subgroup, the working outcome model, $\hat{m}_0(X_i, g)$, can further adjust for any differences.

4.4 Simulation study

Before estimating the differential impacts of letters of recommendation, we first present simulations assessing the performance of our proposed approach versus traditional inverse propensity score weights fit via logistic regression. For $n = 10,000$ units, we draw $d = 50$ covariates $X_{id} \overset{iid}{\sim} N(0, 1)$ and subgroup indicators $G_i \in \{1, \ldots, G\}$ as Multinomial$(\frac{1}{G}, \ldots, \frac{1}{G})$, where $G \in \{10, 50\}$. We then use a separate logistic propensity score model for each group following Equation (4.7),[11]

$$\text{logit } e(X_i, G_i) = \alpha_{G_i} + (\mu_\beta + U_g^\beta \odot B_g^\beta) \cdot X_i, \tag{4.22}$$

and also use a separate linear outcome model for each group,

$$Y_i(0) = \eta_{0G_i} + (\mu_\eta + U_g^\eta \odot B_g^\eta) \cdot X_i + \varepsilon_i, \tag{4.23}$$

where $\varepsilon_i \sim N(0, 1)$ and $\odot$ denotes element-wise multiplication. We then draw group-specific treatment effects $\tau_g \overset{\text{iid}}{\sim} N(0, 1)$ and set the treated potential outcome as $Y_i(1) = Y_i(0) + \tau_{G_i} W_i$. The true ATT in simulation j is thus $\tau_j = \frac{1}{n_1} \sum_{i=1}^n W_i (Y_i(1) - Y_i(0))$.

We draw the fixed effects and varying slopes for each group according to a hierarchical model with sparsity. We draw the fixed effects as $\alpha_g \overset{iid}{\sim} N(0, 1)$ and $\eta_{0g} \overset{iid}{\sim} N(0, 1)$. For the slopes, we first start with a mean slope vector $\mu_\beta, \mu_\eta \in \{-\frac{3}{\sqrt{d}}, \frac{3}{\sqrt{d}}\}^K$, where each element is chosen independently with uniform probability. Then we draw isotropic multivariate normal random variables $U_g^\beta, U_g^\eta \overset{iid}{\sim} MVN(0, I_d)$. Finally, we draw a set of d binary variables $B_{gj}^\beta, B_{gj}^\eta$ Bernoulli with probability $p = 0.25$. The slope is then constructed as a set of sparse deviations from the mean vector: $\mu_\beta + U_g^\beta \odot B_g^\beta$ for the propensity score and $\mu_\eta + U_g^\eta \odot B_g^\eta$ for the outcome model.

For $j = 1, \ldots, m$ with $m = 500$ Monte Carlo samples, we estimate the treatment effects for group g, $\hat{\tau}_{gj}$, and the overall ATT, $\hat{\tau}_j$, and compute a variety of metrics. Following the

[11] The logistic specification differs from the truncated linear odds in Equation 4.16. If the transformed covariates $\phi(X_i)$ include a flexible basis expansion, the particular form of the link function will be less important.

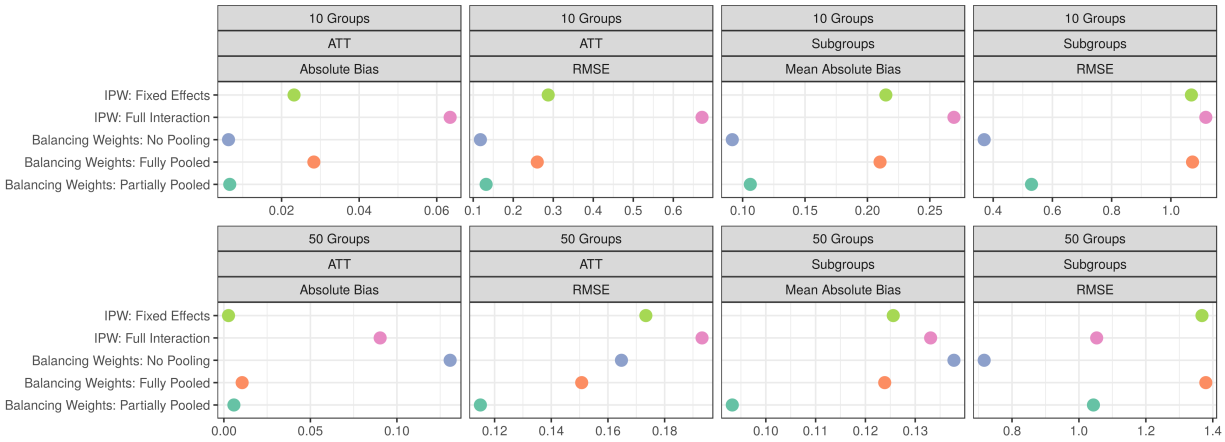


Figure 4.3: Performance of approximate balancing weights and traditional IPW with logistic regression for estimating subgroup treatment effects.

metrics studied by Dong et al. (2020), for subgroup treatment effects we compute (a) the mean absolute bias across the G treatment effects, $\frac{1}{m}\sum_{j=1}^{m}\left|\frac{1}{g}\sum_{g=1}^{G}\hat{\tau}_{gj}-\tau_g\right|$, and (b) the mean root mean square error $\sqrt{\frac{1}{mG}\sum_{j=1}^{m}\sum_{g=1}^{G}(\hat{\tau}_{gj}-\tau_g)^2}$. For the overall ATT we measure (a) the absolute bias $\left|\frac{1}{m}\sum_{j=1}^{m}\hat{\tau}_j-\tau_j\right|$ and (b) the root mean square error $\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\hat{\tau}_j-\tau_j)^2}$.

We compute treatment effects for five weighting estimators:

- *Partially pooled balancing weights:* approximate balancing weights that solve (4.13), using G as the stratifying variable and prioritizing local balance by setting $\lambda_g = \frac{1}{n_{1g}}$.
- *Fully pooled balancing weights:* approximate balancing weights that solve (4.13), but ignore local balance by setting λ to be very large and fully pooling towards the global model. This is equivalent to stable balancing weights in Equation (4.8) with an exact balance constraint $\delta = 0$.
- *No pooled balancing weights:* approximate balancing weights that solve (4.13), but without the exact global balance constraint.
- *Full interaction IPW:* traditional IPW with a fully interacted model that estimates a separate propensity score within each stratum as in Equation (4.7).
- *Fixed effects IPW:* full interaction IPW with stratum-specific coefficients constrained to be equal to a global parameter $\beta_g = \beta$ for all g.

We fit each logistic regression via maximum likelihood with an L^1 penalty to induce sparsity; for the fully interacted specification we also include a set of global parameters μ_β so that the slope for group g is $\mu_\beta + \Delta_g$, with an L^1 penalty for each component. For

both logistic regression specifications, we estimate the models with `glmnet` (Friedman et al., 2010) using an L^1 penalty on the parameters with hyperparameter chosen through 5-fold cross validation.[12]

Figure 4.3 shows the results for the overall ATT and for subgroup effects. We see that with 10 subgroups, prioritizing local balance with either the partially pooled or no-pooled approximate balancing approaches yields lower bias and RMSE than ignoring local balance entirely with the fully pooled approach. These approaches also have better performance than either of the traditional logistic regression approaches. In this setting where there are 1,000 units per group, it is possible to achieve good balance in each group and there is no benefit to partially pooling via the exact global balance constraint. However, with 50 subgroups and 200 units per group, it is difficult to balance within each subgroup and there is a benefit to partial pooling. Partially pooling balancing weights yields much lower bias for the overall ATT than the no-pooled approach, and has lower bias for the subgroup effects as well, although this comes at the cost of higher RMSE for subgroup effects.

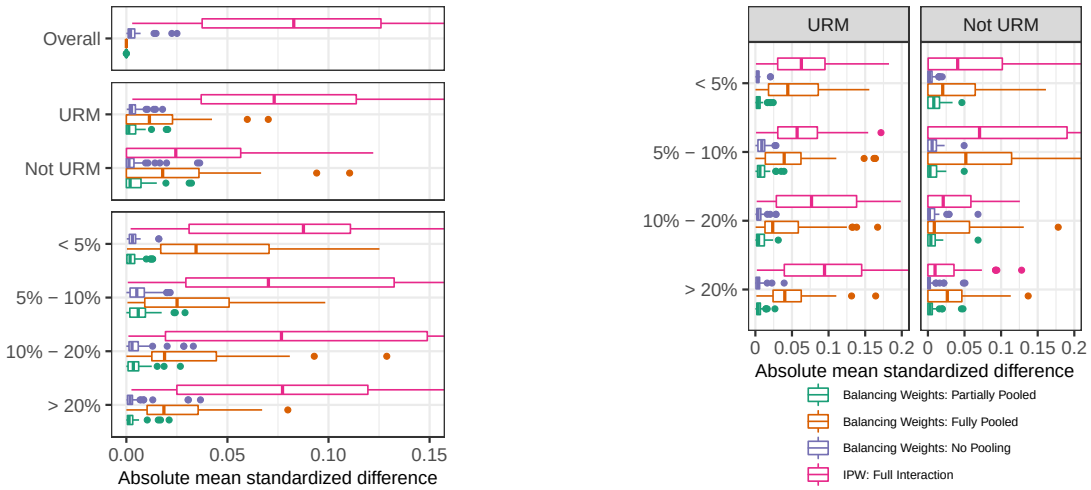
4.5 Differential impacts of letters of recommendation

We now turn to estimating the differential impacts of letters of recommendation on admissions decisions. We focus on the eight subgroups defined in Table 4.1, based on the interaction between URM status (2 levels) and admissibility index (4 levels). Due to the selection mechanism described in Section 4.1, however, it is useful to create even more fine-grained strata and then aggregate to these eight subgroups. Specifically, we define $G = 41$ fine-grained strata based on URM status, AI grouping, first reader score, and college applied to.[13] While we are not necessarily interested in treatment effect heterogeneity across all 41 strata, this allows us to exactly match on key covariates and then aggregate to obtain the primary subgroup effects.

Another key component in the analysis is the choice of transformation of the covariates $\phi(\cdot)$. Because we have divided the applicants into many highly informative strata, we choose $\phi(\cdot)$ to include all of the raw covariates. Additionally, because of the importance of the admissibility index, we also include a natural cubic spline for AI with knots at the sample quantiles. Finally, we include the output of the admissions model and a binary indicator for whether the predicted probability of a "Possible" score is greater than 50%. If desired, we could also consider other transformations such as a higher order polynomial transformation, using a series of basis functions for all covariates, or computing inner products via the kernel trick to allow for an infinite dimensional basis (see, e.g. Hazlett, 2020; Wang and Zubizarreta,

[12]This amounts to partial pooling towards a sparse global model with sparse deviations. We can also consider partially pooling via multilevel modelling.

[13]Of the 48 possible strata, we drop 7 strata where no applicants submitted a letter of recommendation. These are non-URM applicants in both colleges in the two lowest AI strata but where the first reader assigned a "Yes" or "No". This accounts for $\sim 2\%$ of applicants. The remaining 41 strata have a wide range of sizes with a few very large strata. Min: 15, p25: 195, median: 987, p75: 1038, max: 8000



(a) Overall and by URM status and AI.

(b) By URM status interacted with AI.

Figure 4.4: The distribution of imbalance in each component of $\phi(X)$ after weighting with both the partially- and fully-pooled balancing weights estimators, as well as the fully interacted IPW estimator.

2019; Hirshberg and Wager, 2019). We further prioritize local balance in the admissibility index by exactly balancing the AI within each URM $\times$ subgroup. As we discuss above, this ensures local balance in the admissibility index at an intermediate level of the hierarchy between global balance and local balance. Finally, we standardize each component of $\phi(X)$ to have mean zero and variance one.

Diagnostics: local balance checks and assessing overlap

Before estimating effects, we first assess the level of local balance within each subgroup, following the discussion in Section 4.3. We consider the five estimators described in Section 4.4. We also use the estimated fully interacted propensity score model to create subgroup overlap weights as in Yang et al. (2020).

Figure 4.4 shows the distribution of the imbalance in each of the 51 (standardized) components of $\phi(X)$, for the three balancing weights approaches as well as the fully interacted IPW estimator. The fully interacted IPW approach has very poor balance overall, due in part to the difficulty of estimating the high-dimensional propensity score model. As expected, both the fully- and partially-pooled balancing weights achieve perfect balance overall; however, only the partially pooled balancing weights achieve excellent local balance. The partially- and no-pooled approaches have similar global and local balance overall, but the partially-pooled approach sacrifices a small amount of local balance for an improvement in global balance. Appendix Figure C.3 shows these same metrics for the fixed effects IPW and overlap weights, which uses the same propensity score estimates as in the fully interacted

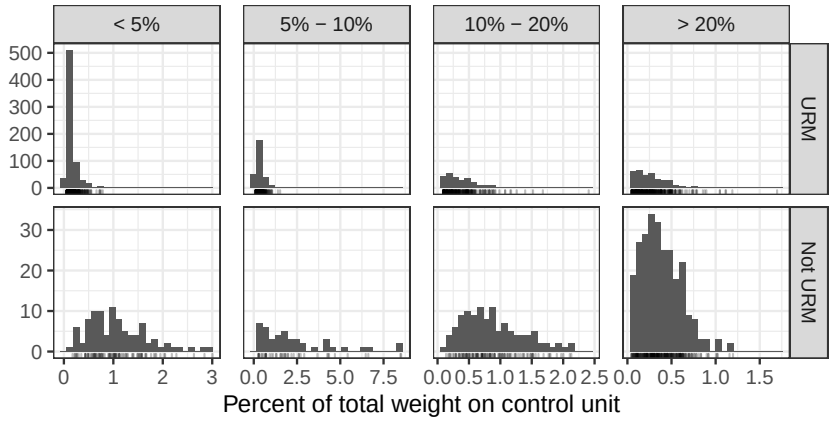


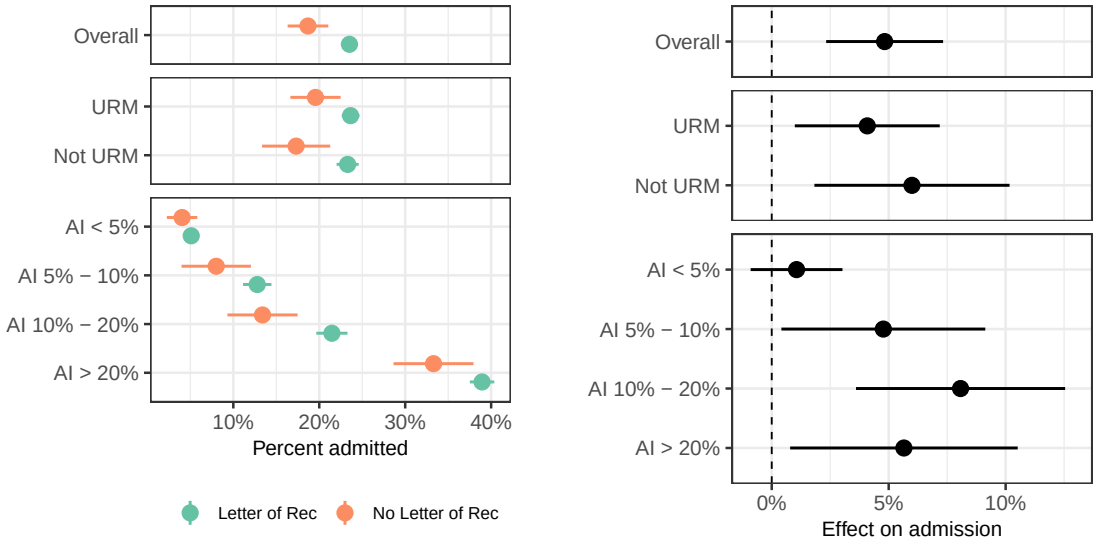
Figure 4.5: Weights on control units from solving the approximate balancing weights problem (4.13). Not pictured: the vast majority of control units that receive zero weight.

IPW approach. Both yield poor local balance.

Appendix Figure C.4 shows imbalance in the one-dimensional summary admissibility index. Our proposed approach, which directly balances this summary index within URM and AI subgroups, again achieves excellent balance overall and within each group. This is not true for other approaches, especially fully interacted IPW weights, which fail to achieve reasonable balance in the admissibility index for most subgroups, with *worse* imbalance relative to the unweighted comparisons for some subgroups. Here we see the effect of partial pooling. The no-pooled approach—only targeting balance within the fine-grained strata, ignoring global balance as well as balance in our primary subgroups—fails to achieve good balance in some subgroups, notably for high admissibility URM applicants, while the partially pooled approach achieves exact balance by design.

Finally, we assess overlap within each subgroup. A key benefit of weighting approaches is that any overlap issues manifest in the distribution of our weights $\hat{\gamma}$. Figure 4.5 plots the distribution of the weights over the comparison applicants by URM status and AI group, normalized by the number of treated applicants in the subgroup. The vast majority of control units receive zero weight and are excluded from the figure. Of the 28,556 applicants who did not submit an LOR, only 5,702 (20%) receive positive weight. This is indicative of a lack of "left-sided" overlap: very many applicants who did not submit a letter of recommendation had nearly zero odds of doing so in the pilot program. This is problematic for estimating the overall average treatment effect, but is less of a concern when we focus on estimating the average treatment effect on the treated.

For each AI subgroup we also see that the distribution of weights is skewed more positively for the non-URM applicants. In particular, for the lower AI, non-URM subgroups we see a non-trivial number of comparison applicants that "count for" over 2% of the re-weighted sample, with a handful of outliers that are equivalent to over 5%. While large weights do not necessarily affect the validity of the estimator — though they suggest caution in



(a) Treated and re-weighted control percent admitted.

(b) Estimated effects on admission.

Figure 4.6: Estimated treated and control means and treatment effect of letters of recommendation on admission ± two standard errors, overall and by URM status and Admissibility Index.

terms of "right-sided" overlap — large weights decrease the effective sample size, reducing the precision of our final estimates. Appendix Figure C.5 shows the effective sample size, $n_{1g}/\sum_{G_i=g}(1-W_i)\hat{\gamma}_i^2$, for each subgroup g. We see that the URM subgroups have larger effective sample sizes than the non-URM subgroups, with particularly stark differences for the lower AI subgroups. Furthermore, for all non-URM subgroups with AI $\leq$ 20%, the effective sample size is $\leq$ 100. From this, we should expect to have far greater precision in the estimates for URM applicants than non-URM applicants.

Treatment effect estimates

After assessing local balance and overlap, we can now turn to estimating the differential impacts of letters of recommendation. Figure 4.6 shows (1) the percent of applicants who submitted an LOR who were accepted, $\hat{\mu}_{1g}$ (2) the imputed counterfactual mean, $\hat{\mu}_{0g}$ and (3) the ATT, $\hat{\mu}_{1g} - \hat{\mu}_{0g}$. The standard errors are computed via the sandwich estimator in Equation (4.15). Overall, we estimate an increase in admission rates of 5 percentage points (pp). While we estimate a larger effect for non-URM applicants (6 pp) than URM applicants (4 pp), there is insufficient evidence to distinguish between the two effects. Second, we see a roughly positive trend between treatment effects and the AI, potentially with a peak for the 10%-20% group. This is driven by the very small estimated effect for applicants with AI

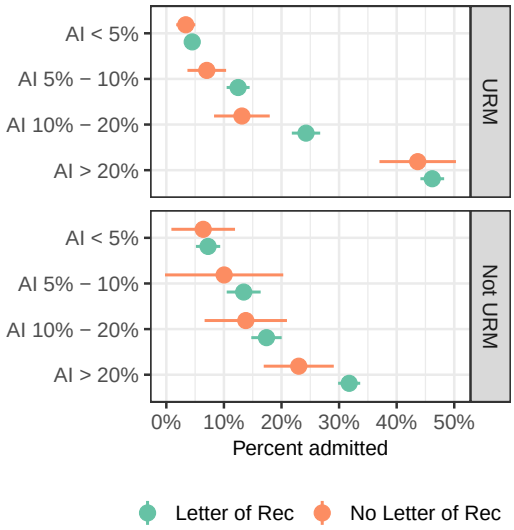
$< 5\%$ and who are thus very unlikely to be accepted *a priori*. Substantively, this corresponds to letters of recommendation having a very low impact for applicants unlikely to be accepted, but a larger impact for applicants that are perhaps on the cusp on acceptance. Appendix Figure C.6 shows an estimate of the log risk ratio, $\log \frac{\mathbb{E}[Y(1)|G=g]}{\mathbb{E}[Y(0)|G=g]}$, for the subgroups. From the estimated risk ratios, we see that this pattern, while noisy, is consistent with impacts that are roughly constant on the multiplicative scale, perhaps with a dip for both the low and high admissibility applicants.

Figure 4.7 further stratifies the subgroups, showing the effects jointly by URM status and AI. While the point estimate for the overall increase in admission rates is slightly larger for non-URM applicants than for URM applicants, this is mainly a composition effect. For applicants very unlikely to be admitted (AI $< 5\%$) the point estimates are nearly identical for URM and non-URM applicants, although the URM subgroup is estimated much more precisely. For the next two levels of the admissibility index (AI between 5% and 20%), URM applicants have a higher estimated impact, with imprecise estimates for non-URM applicants. For the highest admissibility groups (AI $> 20\%$), non-URM applicants have larger positive effects, though again these estimates are noisy. Since URM applicants have lower AI on average, the overall estimate is also lower for URM applicants. Furthermore, the peak in the effect for middle-tier applicants is more pronounced for URM applicants than non-URM applicants. From Figure 4.7a we see that this is primarily because high admissibility URM applicants with a letter of recommendation are admitted at very high rates; the imputed baseline after re-weighting is similarly large.

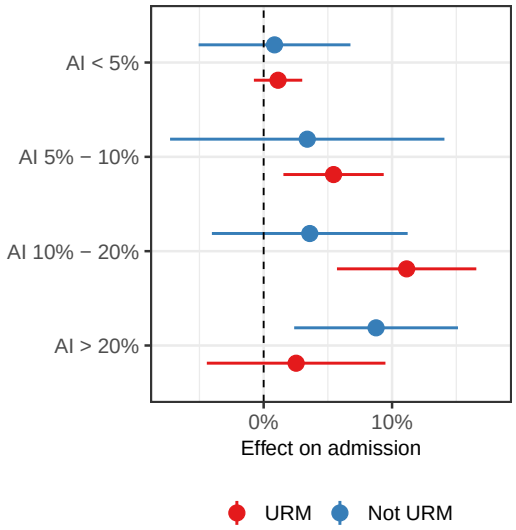
We also consider augmenting the weighting estimator with an estimate of the prognostic score, $\hat{m}(x, g)$. In Appendix Figure C.7 we show estimates after augmenting with ridge regression; we compute standard errors via Equation (4.15), replacing $Y_i - \hat{\mu}_{0g}$ with the empirical residual $Y_i - \hat{m}(X_i, g)$. Because the partially pooled balancing weights achieve excellent local balance for $\phi(X)$, augmenting with a model that is also linear in $\phi(X)$ results in minimal adjustment. We therefore augment with a nonlinear outcome model, random forests. Tree-based estimators are a natural choice for a nonlinear outcome model, creating "data-dependent strata" similar in structure to the strata we define for G. For groups where the weights $\hat{\gamma}$ have good balance across the estimates $\hat{m}(x, g)$, there will be little adjustment due to the outcome model. Conversely, if the raw and bias-corrected estimate disagree for a subgroup, then the weights have poor local balance across important substantive data-defined strata. For these subgroups we should be more cautious of our estimates.

Figure 4.8 shows the random forest-augmented effect estimates relative to the un-augmented estimates; the difference between the two is the estimated bias. Overall, the random forest estimate of the bias is negligible and, as a result, the un-adjusted and adjusted estimators largely coincide. Augmentation, however, does seem to stabilize the higher-order interaction between AI and URM status, with particularly large adjustments for the highest AI group (AI $\geq 20\%$). This suggests that we should be wary of over-interpreting any change in the relative impacts for URM and non-URM applicants as AI increases.

In the Appendix we consider alternative estimates. First, Appendix Figure C.8 shows the estimated effects on admission rates using all five weighting procedures we consider above.

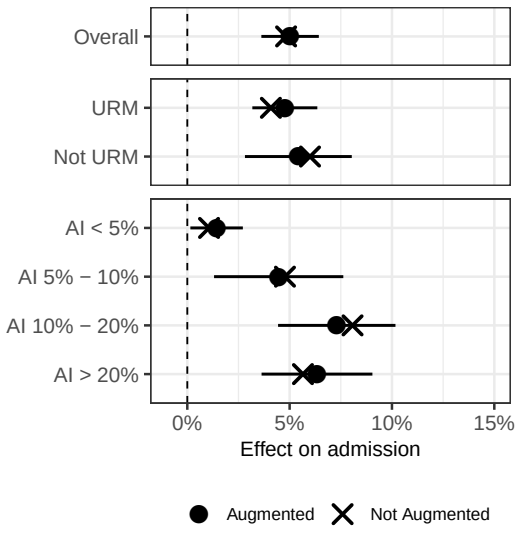


(a) Treated and re-weighted control percent admitted.

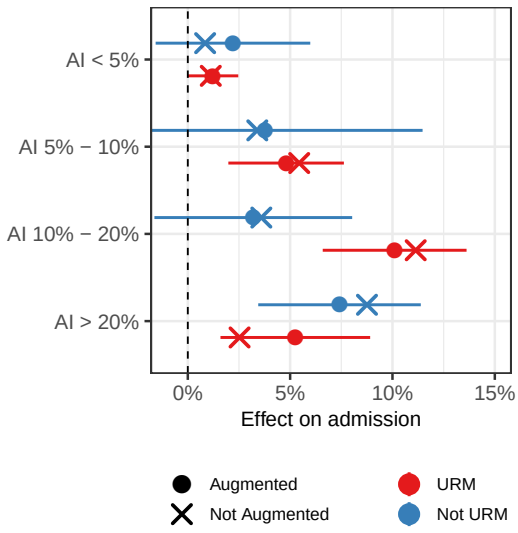


(b) Estimated effects on admission.

Figure 4.7: Estimated treated and control means and treatment effect of letters of recommendation on admission ± two standard errors, further broken down by URM status interacted with the Admissibility Index.



(a) Overall and by URM status and AI.



(b) By URM status interacted with AI.

Figure 4.8: Estimated effect of letters of recommendation on admission rates with and without augmentation via a random forest outcome model.

Despite failing to achieve good local balance, the IPW approaches and fully pooled balancing weights approach yield effect estimates that are similar to our proposed approach. The overlap weighting approach of Li (2017), however, leads to substantively different conclusions, perhaps due to the change in the estimand. These differences appear to be driven by that estimator's *negative* estimated effect of LORs for high admissibility, non-URM applicants, suggesting that there are other substantively important sources of heterogeneity beyond URM status and admissibility.

Second, we consider effects on an intermediate outcome: whether the second reader — who has access to the LOR — gives a "Yes" score. Because these are *design-based* weights, we use the same set of weights to estimate effects on both second reader scores and admissions decisions. With this outcome we can also make use of a within-study design to estimate treatment effects, leveraging scores from additional third readers who did not have access to the letters of recommendation; we describe this design in Appendix C.1. Appendix Figures C.9 and C.10 show the results for both approaches. Overall for second reader scores we see a similar structure of heterogeneity as for admission rates, although there does not appear to be an appreciable decline in the treatment effect for the highest admissibility non-URM applicants. The two distinct approaches yield similar patterns of estimates overall, with the largest discrepancy for applicants with a predicted probability of admission between 5% and 10%, particularly for non-URM applicants. However, this group has a very low effective sample size, and so the weighting estimates are very imprecise.

Taken together, these results paint a relatively clear picture of differential impact of letters of recommendation across applicants' *a priori* probability of admission. Treatment effects are low for applicants who are unlikely to be accepted and high for applicants on the margin for whom letters provide useful context, with some evidence of a dip for the highest admissibility applicants. Our estimates of differential impacts between URM and non-URM students are more muddled, due to large sampling errors, and do not support strong conclusions. Point estimates indicate that LORs benefit URM applicants more than they do non-URM applicants at all but the highest academic indexes. Because non-URM applicants are overrepresented in the high-AI category, the point estimate for the average treatment effect is larger for non-URMs; however, there is insufficient precision to distinguish between the two groups.

4.6 Discussion

Estimating heterogeneous treatment effects and assessing treatment effect variation in observational studies is a challenge, even for pre-specified subgroups. Focusing on weighting estimators that estimate subgroup treatment effects by re-weighting control outcomes, we show that the estimation error depends on the level of *local imbalance* between the treated and control groups after weighting. We then present a convex optimization problem that finds approximate balancing weights that directly target the level of local imbalance within each subgroup, while ensuring exact global balance to also estimate the overall effect. Using

this method to estimate heterogeneous effects in the UC Berkeley letters of recommendation pilot study, we find evidence that letters of recommendation lead to better admissions outcomes for stronger applicants, with mixed evidence of differences between URM and non-URM applicants.

There are several directions for future methodological work. First, we directly estimate the effect of submitting an LOR among those who submit. However, we could instead frame the question in terms of non-compliance and use the *invitation* to submit an LOR as an instrument for submission. Using the approximate balancing weights procedure described above we could adjust for unequal invitation probabilities, and estimate the effect on compliers via weighted two-stage least squares. Second, we could consider deviations from the ignorability assumption via a sensitivity analysis. One potential path is to extend the balancing weights sensitivity procedure from Soriano et al. (2020) to the setting with distinct subgroups. Third, we could adapt our approach to explore treatment effect variation in other types of observational studies, for instance in settings that mimic the structure of multisite trials.

Bibliography

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies 72*(1), 1–19.

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

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

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

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

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

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

Abraham, S. and L. Sun (2018). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects.

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

Anoke, S. C., S.-L. Normand, and C. M. Zigler (2019). Approaches to treatment effect heterogeneity in the presence of confounding. *Statistics in medicine 38*(15), 2797–2815.

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

Arkhangelsky, D. and G. W. Imbens (2019). Double-robust identification for causal panel data models. *arXiv preprint arXiv:1909.09412*.

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. and G. W. Imbens (2018). Design-based analysis in difference-in-differences settings with staggered adoption. Technical report, National Bureau of Economic Research.

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

Bates, D., M. Mächler, B. M. Bolker, and S. C. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1).

Bickel, P. J., E. A. Hammel, and J. W. O'Connell (1975). Sex bias in graduate admissions: Data from Berkeley. *Science 187*(4175), 398–404.

Bilinski, A. and L. A. Hatfield (2018). Seeking evidence of absence: reconsidering tests of model assumptions. *arXiv preprint arXiv:1805.03273*.

Borusyak, K. and X. Jaravel (2017). Revisiting event study designs. *Available at SSRN 2826228*.

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

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

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.

Callaway, B. and P. H. C. Sant'Anna (2018). Difference-in-Differences With Multiple Time Periods and an Application on the Minimum Wage and Employment.

Carvalho, C., A. Feller, J. Murray, S. Woody, and D. Yeager (2019). Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge. *Observational Studies 5*, 21–35.

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.

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

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.

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

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

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

Cox, D. R. (1984). Interaction. *International Statistical Review/Revue Internationale de Statistique*, 1–24.

Deming, W. E. and F. F. Stephan (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics 11*(4), 427–444.

Deville, J. C. and C. E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*(418), 376–382.

Deville, J. C., C. E. Särndal, and O. Sautory (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association 88*(423), 1013–1020.

Dong, J., J. L. Zhang, S. Zeng, and F. Li (2020). Subgroup balancing propensity score. *Statistical Methods in Medical Research 29*(3), 659–676.

Donohue, J. J., A. Aneja, and K. D. Weber (2019). Right-to-carry laws and violent crime: A comprehensive assessment using panel data and a state-level synthetic control analysis. *Journal of Empirical Legal Studies 16*(2), 198–247.

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.

Feller, A. and A. Gelman (2015). Hierarchical models for causal effects. *Emerging Trends in the Social and Behavioral Sciences: An interdisciplinary, searchable, and linkable resource*, 1–16.

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

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

Fesler, L. and M. Pender (2019). Local promise programs: Varying impacts on enrollment, graduation, and financial outcomes.

Frandsen, B. R. (2016). The effects of collective bargaining rights on public employee compensation: Evidence from teachers, firefighters, and police. *ILR Review 69*(1), 84–112.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software 33*(1).

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

Goldstein, D. (2015). *The teacher wars: A history of America's most embattled profession*. Anchor.

Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing. Technical report, National Bureau of Economic Research.

Green, K. M. and E. A. Stuart (2014). Examining moderation analyses in propensity score methods: Application to depression and substance use. *Journal of consulting and clinical psychology 82*(5), 773.

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

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

Hastie, T., R. Mazumder, J. D. Lee, and R. Zadeh (2015). Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *Journal of Machine Learning Research 16*, 3367–3402.

Hazlett, C. (2020). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Statistica Sinica*.

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

Hess, F. M. and M. R. West (2006). A better bargain: Overhauling teacher collective bargaining for the 21st century. *Program on Education Policy and Governance, Harvard University*.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics 20*(1), 217–240.

Hirshberg, D. and S. Wager (2019). Augmented Minimax Linear Estimation.

Hirshberg, D. A., A. Maleki, and J. Zubizarreta (2019). Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296*.

Hirshberg, D. A. and S. Wager (2018). Debiased inference of average partial effects in single-index models. *arXiv preprint arXiv:1811.02547*.

Holland, P. W. (1986). Statistics and Causal Inference: Rejoinder. *Journal of the American Statistical Association 81*(396), 968.

Hout, M. (2005). Berkeley's comprehensive review method for making freshman admissions decisions: An assessment. Technical report, University of California, Berkeley.

Hoxby, C. M. (1996). How teachers' unions affect education production. *The Quarterly Journal of Economics 111*(3), 671–718.

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

Imai, K. and I. S. Kim (2019). On the use of two-way fixed effects regression models for causal inference with panel data.

Imai, K. and D. A. Van Dyk (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association 99*(467), 854–866.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika 87*(3), 706–710.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jackson, C. K., J. E. Rockoff, and D. O. Staiger (2014). Teacher effects and teacher-related policies. *Annu. Rev. Econ. 6*(1), 801–825.

Kahn-Lang, A. and K. Lang (2019). The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business & Economic Statistics*, 1–14.

Kallus, N. (2016). Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*.

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

King, G., C. Lucas, and R. A. Nielsen (2017). The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science 61*(2), 473–489.

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

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.

Kuncel, N. R., R. J. Kochevar, and D. S. Ones (2014). A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *International Journal of Selection and Assessment 22*(1), 101–107.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America 116*(10), 4156–4165.

Lee, Y., T. Q. Nguyen, and E. A. Stuart (2019). Partially Pooled Propensity Score Models for Average Treatment Effect Estimation with Multilevel Data.

Li, F., A. M. Zaslavsky, and M. B. Landrum (2013). Propensity score weighting with multilevel data. *Statistics in Medicine 32*(19), 3373–3387.

Li, K. T. (2017). Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods.

Lovenheim, M. F. (2009). The effect of teachers' unions on education production: Evidence from union election certifications in three midwestern states. *Journal of Labor Economics 27*(4), 525–587.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods 9*(4), 403.

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

Nie, X. and S. Wager (2019). Quasi-Oracle Estimation of Heterogeneous Treatment Effects.

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

Paglayan, A. S. (2019). Public-sector unions and the size of government. *American Journal of Political Science 63*(1), 21–36.

Pimentel, S. D. and R. R. Kelz (2019). Optimal tradeoffs in matched designs for observational studies. Technical report.

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

Pustejovsky, J. E. and E. Tipton (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics 36*(4), 672–683.

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

Robbins, M., J. Saunders, and B. Kilmer (2017). A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention. *Journal of the American Statistical Association 112*(517), 109–126.

Robins, J. M., A. Rotnitzky, L. Ping Zhao, and L. Ping ZHAO (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association 89427*, 846–866.

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., R. N. Ross, and J. H. Silber (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association 102*(477), 75–83.

Rosenbaum, P. R. and D. B. Rubin (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika 70*(1), 41–55.

Roth, J. (2018). Should we condition on the test for pre-trends in difference-in-difference designs? *arXiv preprint arXiv:1804.01208*.

Rothstein, J. (2017). The impact of letters of recommendation on UC Berkeley admissions in the 2016-17 cycle. Technical report, California Policy Lab.

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 185–203.

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.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics 2*(3), 808–840.

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

Särndal, C.-E., B. Swensson, and J. Wretman (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Schmader, T., J. Whitehead, and V. H. Wysocki (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex roles 57*(7-8), 509–514.

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

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.

U.S. Department of Education, National Center for Education Statistics (2018). Fast facts: Expenditures. Technical report.

VanderWeele, T. J. and M. J. Knol (2011). Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions. *Annals of internal medicine 154*(10), 680–683.

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

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

Wang, Y. and J. R. Zubizarreta (2019). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*.

Wong, R. K. and K. C. G. Chan (2017). Kernel-based covariate functional balancing for observational studies. *Biometrika 105*(1), 199–213.

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

Yang, S., E. Lorenzi, G. Papadogeorgou, D. Wojdyla, F. Li, and L. Thomas (2020). Subgroup covariates balancing via the overlap weights.

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.

Zubizarreta, J. R. and L. Keele (2017). Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system. *Journal of the American Statistical Association 112*(518), 547–560.

Appendix A

Supplementary materials for Chapter 2

A.1 Inference

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

The approximate validity of the conformal inference procedure in Section 2.5 depends on the predictive accuracy of $\hat{Y}_{it}^{\text{aug}}(0)$ when fit using all periods $t = 1, \ldots, T$, including the post-treatment period T. Denoting $\mathbf{Y_1}_\cdot \equiv (\mathbf{X}_{1\cdot}, Y_1) \in \mathbb{R}^T$ to be the full vector of treated unit outcomes and $\mathbf{Y}_{0\cdot} \equiv [\mathbf{X}_{0\cdot}, \mathbf{Y}_{0T}] \in \mathbb{R}^{N_0 \times T}$ be the matrix of comparison unit outcomes, the Ridge ASCM optimization problem in this setting is

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

We will also consider the constrained form:

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

With these definitions we can characterize the in-sample prediction error of the counterfactual model described by Chernozhukov et al. (2019), which is a version of Equation (2.3) in an asymptotic framework where T_0 is growing while T is fixed. We state the model and assumptions for asymptotically (in T_0) valid inference below.

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

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

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

1. The data is β-mixing with exponential speed
2. There exist constants $c_1, c_2 > 0$ such that $\mathbb{E}[(Y_{it}\varepsilon_{1t})^2] \geq c_1$ and $\mathbb{E}[|Y_{it}\varepsilon_{1t}|^3] \leq c_2$ for all i such that $W_i = 0$ and $t = 1, \ldots, T$
3. For all i such that $W_i = 0$, $X_{i1}\varepsilon_{11}, \ldots, X_{iT}\varepsilon_{1T}$ is β-mixing with β-mixing coefficient satisfying $\beta(t) \leq a_1 e^{-a^2 t^k}$ for constants $a_1, a_2, k > 0$
4. There exists a constant $c_3 > 0$ such that $\max_{W_i=0} \sum_{t=1}^{T} X_{it}^2 \varepsilon_{1t}^2 \leq c_3^2 T$ with probability $1 - o(1)$
5. $\log N_0 = o\left(T^{\frac{4k}{3k+4}}\right)$
6. There exists a sequence $\ell_T > 0$ such that $\mathbf{Y}_{0t}'(w - \gamma^*) \leq \ell_T \frac{1}{T} \|\mathbf{Y}_{0\cdot}'(w - \gamma^*)\|_2^2$ for all $w \in \Delta^{N_0} + B_2(\frac{C}{\sqrt{N_0}})$, for some constant C where $B_p(a) = \{x \in \mathbb{R} \mid \|x\|_p \leq a\}$, with probability $1 - o(1)$ for $T_0 + 1 \leq t \leq T$
7. The sequence ℓ_T satisfies $\ell_T \left(\log \min\{T, N_0\}\right)^{\frac{1+k}{2k}} \sqrt{T} \to 0$

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

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

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

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

A.2 Additional results

Specialization of Ridge ASCM results to SCM

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

First we specialize Proposition 2.1 to SCM weights by taking $\lambda \to \infty$.

Corollary A.1. Under the linear model (2.6) with independent sub-Gaussian noise with scale parameter σ, for any $\delta > 0$, for weights $\boldsymbol{\gamma} \in \Delta^{N_0}$ independent of the post-treatment outcomes $(Y_{1T}, \ldots, Y_{NT})$ and for any $\delta > 0$,

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

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

We can similarly specialize Theorem 2.1.

Corollary A.2. Under the linear factor model (2.7) with independent sub-Gaussian noise with scale parameter σ, for weights $\boldsymbol{\gamma} \in \Delta^{N_0}$ independent of the post-treatment outcomes $(Y_{1T}, \ldots, Y_{NT})$ and for any $\delta > 0$,

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

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

Error under a partially linear model with Lipshitz deviations from linearity

We now bound the estimation error for SCM and Ridge ASCM under the basic model (2.3) when the outcome is only partially linear, with Lipshitz deviations from linearity.

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

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

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

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

Theorem A.1. Let $C = \max_{W_i=0} \|\boldsymbol{X}_i\|_2$. Under Assumption A.2, for any $\delta > 0$, the estimation error for the ridge ASCM weights $\hat{\boldsymbol{\gamma}}^{\text{aug}}$ (2.17) with hyperparameter $\lambda^{\text{ridge}} = N_0\lambda$ is

$$\begin{aligned}\left|Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i^{\text{aug}} Y_{1T}\right| \leq \|\boldsymbol{\beta}\|_2 \underbrace{\left\|\text{diag}\left(\frac{\lambda}{d_j^2+\lambda}\right)(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_0'\hat{\boldsymbol{\gamma}}^{\text{scm}})\right\|_2}_{\text{imbalance in } X} + \\ \underbrace{CL\left\|\text{diag}\left(\frac{d_j}{d_j^2+\lambda}\right)(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_0'\hat{\boldsymbol{\gamma}}^{\text{scm}})\right\|_2}_{\text{excess approximation error}} + \\ \underbrace{L\sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} \|\boldsymbol{X}_1 - \boldsymbol{X}_i\|_2}_{\text{SCM approximation error}} \quad + \quad \underbrace{\delta\sigma\left(1 + \|\hat{\boldsymbol{\gamma}}^{\text{aug}}\|_2\right)}_{\text{post-treatment noise}}\end{aligned} \tag{A.8}$$

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

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

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

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

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

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

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

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

Error under a linear factor model with covariates

We can quantify the behavior of the two-step procedure from Lemma 2.4 in controlling the error under a more general form of the linear factor model (2.7) with covariates (see Abadie et al., 2010; Botosaru and Ferman, 2019, for additional discussion). We can also consider the error under a linear model with auxiliary covariates, as a direct consequence of Lemma 2.4.

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

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

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

To characterize how well the covariates approximate the true function $f(\boldsymbol{Z}_i)$, we will consider the best linear approximation in our data, and define the residual for unit i and time t as $e_{it} = f_t(\boldsymbol{Z}_i) - \boldsymbol{Z}_i'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' f_t(Z)$, where $\boldsymbol{Z} \in \mathbb{R}^{N\times K}$ is the matrix of all auxiliary covariates for all units. For each time period we will characterize the additional approximation error incurred by only balancing the covariates linearly with the *residual sum of squares* $RSS_t = \sum_{i=1}^{n} e_{it}^2$. For ease of exposition, we assume that the control covariates are standardized and rotated, which can always be true after pre-processing, and present results for the simpler case in which we fit SCM on the residualized pre-treatment outcomes rather than ridge ASCM (i.e., we let $\lambda^{\text{ridge}} \to \infty$); the more general version follows immediately by applying Theorem 2.1.

Theorem A.2. Under the linear factor model with covariates in Assumption A.3, with

$\frac{1}{N_0}\boldsymbol{Z}_{0\cdot}'\boldsymbol{Z}_{0\cdot} = \boldsymbol{I}_K$, for any $\delta > 0$, $\hat{\boldsymbol{\gamma}}^{\text{cov}}$ in Equation (2.33) with $\lambda^{\text{ridge}} \to \infty$ satisfies the bound

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

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

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

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

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

Corollary A.4. Under the linear factor model with covariates in Assumption A.3 with $f_t(\boldsymbol{Z}_i) = \sum_{k=1}^K B_{tk}Z_{ik}$ as in Equation (A.13), for any $\delta > 0$, $\hat{\boldsymbol{\gamma}}^{\text{cov}}$ in Equation (2.33) with $\lambda^{\text{ridge}} \to \infty$ satisfies the bound

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

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

Building on Lemma 2.4, Theorem A.2 and Corollary A.4 show that due to the additive, separable structure of the auxiliary covariates in Equation (A.13), controlling the pre-treatment fit in the *residualized* lagged outcomes $\check{\boldsymbol{X}}$ partially controls the error. This justifies

directly targeting fit in the residualized lagged outcomes $\check{\boldsymbol{X}}$ rather than targeting raw lagged outcomes $\boldsymbol{X}$. Moreover, the excess approximation error will be small since since the number of covariates K is small relative to N_0 and the auxiliary covariates are measured without noise. As in Section 2.4, we can achieve better balance when we apply ridge ASCM to $\check{\boldsymbol{X}}$ than when we apply SCM alone. Because $\check{\boldsymbol{X}}$ are orthogonal to $\boldsymbol{Z}$ by construction, this comes at no cost in terms of imbalance in $\boldsymbol{Z}$. However, the fundamental challenge of over-fitting to noise still remains, and, as in the case without auxiliary covariates, selecting the tuning parameter remains important. We again propose to follow the cross validation approach in Section 2.5, here using the residualized lagged outcomes $\check{\boldsymbol{X}}$ rather than the raw lagged outcomes $\boldsymbol{X}$.

A.3 Simulation data generating process

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

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

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

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

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

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

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

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

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

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

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

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

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

A.4 Proofs

Proofs for Section 2.4

Lemma A.2. With $\hat{\eta}_0^{\text{ridge}}$ and $\hat{\boldsymbol{\eta}}^{\text{ridge}}$, the solutions to (2.14), the ridge estimate can be written as a weighting estimator:

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

where

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

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

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

Proof of Lemmas 2.1 and A.2. Recall that the lagged outcomes are centered by the control averages. Notice that

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

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

We now prove that $\hat{\boldsymbol{\gamma}}^{\text{ridge}}$ and $\hat{\boldsymbol{\gamma}}^{\text{scm}}$ solve the weighting optimization problems (A.18) and (2.18). First, the Lagrangian dual to (A.18) is

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

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

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

Since the lagged outcomes are centered, this implies that

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

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

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

This implies that

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

Finally, the weights are the ridge weights

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

Similarly, the Lagrangian dual to (2.18) is

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

where we have used that the convex conjugate of $\frac{1}{2}(x - \hat{\gamma}_i^{\text{scm}})^2$ is $\frac{1}{2}(y + \hat{\gamma}_i^{\text{scm}})^2 - \frac{1}{2}\hat{\gamma}_i^{\text{scm2}}$. Solving for α we see that $\hat{\alpha} = 0$. Now solving for $\boldsymbol{\beta}$ we see that

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

Finally, the weights are the ridge ASCM weights

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

□

Proof of Lemma 2.3. Notice that

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

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

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

□

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

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

with $\widetilde{\boldsymbol{X}}_i = V'\boldsymbol{X}_i$ as defined in Lemma 2.3.

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

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

□

Proofs for Sections 2.5, A.2, and A.2

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

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

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

where $\{\varepsilon_{iT}\}$ are defined as in Assumption 2.1(a).

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

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

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

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

Proof. First, note that the estimation error is

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

Now since the weights are independent of ε_{iT}, by Assumption 2.1(a) we see that $\varepsilon_{1T} - \sum_{W_i=0} \hat{\gamma}_i \varepsilon_{iT}$ is sub-Gaussian with scale parameter $\sigma\sqrt{1 + \|\hat{\boldsymbol{\gamma}}\|_2^2} \leq \sigma\,(1 + \|\hat{\boldsymbol{\gamma}}\|_2)$. Therefore we can bound the second term:

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

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

Proof of Proposition 2.1. Note that under the linear model (2.6), $\xi_i = 0$ for all i. Now from Lemma 2.3 we have that

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

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

Proof of Corollary A.1. This is a direct consequence of Proposition A.1 noting that under the linear model (2.6), $\xi_i = 0$ for all i. □

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

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

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

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

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

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

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

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

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

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

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

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

□

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

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

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

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

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

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

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

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

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

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

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

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

□

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

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

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

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

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

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

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

Proof of Theorem A.4. First note that from Lemma 2.3 we have that

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

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

Theorems A.3 and A.4 have several implications when the outcomes follow a linear factor model (2.7). To see this, we need one more lemma:

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

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

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

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

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

Then we see that the post treatment outcomes are

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

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

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

Proof of Theorem 2.1. Again from Lemma A.6 we can apply Theorem A.4 with $\boldsymbol{\beta} = \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'$ and $\xi_i = \frac{1}{T_0}\boldsymbol{\mu}_T'\boldsymbol{\mu}'\boldsymbol{\varepsilon}_{i(1:T_0)}$, so $\varpi = \frac{JM^2\sigma}{\sqrt{T_0}}$. Plugging these values into Theorem A.3 gives the result. □

Corollary A.5 (Approximation error for ridge ASCM with dependent errors)**.** Under the linear factor model (2.7) with time-dependent errors satisfying $\boldsymbol{\varepsilon}_{i(1:T_0)} \overset{iid}{\sim} N(0, \sigma^2\Omega)$ the approximation error satisfies

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

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

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

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

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

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

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

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

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

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

□

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

Lemma A.8. Let $C = \max_{W_i=0} \|\boldsymbol{X}_i\|_2$. Under Assumption A.6, the ridge ASCM weights $\hat{\boldsymbol{\gamma}}^{\text{aug}}$ (2.17) with hyperparameter $\lambda^{\text{ridge}} = N_0\lambda$ satisfy

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

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

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

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

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

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

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

Proofs for Sections 2.6 and A.2

Proof of Lemma 2.4. The regression parameters $\hat{\boldsymbol{\eta}}_x$ and $\hat{\boldsymbol{\eta}}_z$ in Equation (2.31) are:

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

Now notice that

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

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

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

The pre-treatment fit is

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

This gives the bound on the pre-treatment fit.

□

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

Then the error is

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

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

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

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

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

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

By construction, $\hat{\boldsymbol{\gamma}}^{\text{cov}}$ perfectly balances the covariates, and combined with Lemma 2.4, the error due to imbalance in ϕ simplifies to

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

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

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

We have bounded the first term on the right hand side in Lemma A.4. To bound the second term, notice that $\sum_{W_i=0} \sum_{t=1}^{T_0} \boldsymbol{\mu}_T' \boldsymbol{\mu}_t . Z_{ik} \varepsilon_{it}$ is sub-Gaussian with scale parameter $\sigma M J^2 \sqrt{T_0} \|Z_{\cdot k}\|_2 = M J^2 \sigma \sqrt{T_0 N_0}$. We can now bound the L^2 norm of $\frac{1}{T_0} \boldsymbol{\mu}_T' \boldsymbol{\mu}' \boldsymbol{\varepsilon}_{0(1:T_0)}' \boldsymbol{Z}_{0\cdot} \in \mathbb{R}^K$:

$$P\left(\frac{1}{T_0} \|\boldsymbol{\mu}_T' \boldsymbol{\mu}' \boldsymbol{\varepsilon}_{0(1:T_0)}' \boldsymbol{Z}_{0\cdot}\|_2 \geq 2 J M^2 \sigma \left(\sqrt{\frac{N_0 K \log 5}{T_0}} + \delta \right) \right) \leq 2 \exp\left(-\frac{T_0 \delta^2}{2} \right)$$

Replacing δ with $\sqrt{\frac{K N_0}{T_0}}(2 - \sqrt{\log 5})$ and with the Cauchy-Schwarz inequality we see that

$$\frac{1}{T_0} \left| \boldsymbol{\mu}_T' \boldsymbol{\mu}' \boldsymbol{\varepsilon}_{0(1:T_0)}' \boldsymbol{Z}_{0\cdot} (\boldsymbol{Z}_{0\cdot}' \boldsymbol{Z}_{0\cdot})^{-1} (\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}' \hat{\boldsymbol{\gamma}}) \right| \leq 4 J M^2 \sigma \sqrt{\frac{K}{T_0 N_0}} \|\boldsymbol{Z}_1 - \boldsymbol{Z}_{0\cdot}' \hat{\boldsymbol{\gamma}}^{\text{scm}}\|_2$$

with probability at least $1 - 2\exp\left(-\frac{K N_0 (2 - \sqrt{\log 5})^2}{2}\right)$.

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

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

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

Combining with Lemma 2.4 and putting together the pieces with the union bound gives the result. □

A.5 Connection to balancing weights and IPW

We have motivated Augmented SCM via bias correction. An alternative motivation comes from the connection between SCM and inverse propensity score weighting (IPW). This is also comparable in form to the generalized regression estimator in survey sampling (Cassel et al., 1976; Breidt and Opsomer, 2017), which has been adapted to the causal inference setting by, among others, Athey et al. (2018) and Hirshberg and Wager (2019).

First, notice that the SCM weights from the constrained optimization problem in Equation (2.8) are a form of *approximate balancing weights*; see, for example, Zubizarreta (2015); Athey et al. (2018); Tan (2017); Wang and Zubizarreta (2019); Zhao (2018). Unlike traditional inverse propensity score weights, which indirectly minimize covariate imbalance by estimating a propensity score model, balancing weights seek to *directly* minimize covariate imbalance, in this case L^2 imbalance. Balancing weights have a Lagrangian dual formulation as inverse propensity score weights (see, for example Zhao and Percival, 2017; Zhao, 2018; Chattopadhyay et al., 2020). Extending these results to the SCM setting, the Lagrangian dual of the SCM optimization problem in Equation (2.8) has the form of a propensity score model. Importantly, as we discuss below, it is not always appropriate to interpret this model as a propensity score.

We first derive the Lagrangian dual for a general class of balancing weights problems, then specialize to the penalized SCM estimator (2.8).

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

This formulation generalizes Equation (2.8) in two ways: first, we remove the non-negativity constraint and note that this can be included by restricting the domain of the strongly convex dispersion penalty f. Examples include the re-centered L^2 dispersion penalties for ridge regression and ridge ASCM, an entropy penalty (Robbins et al., 2017), and an elastic net penalty (Doudchenko and Imbens, 2017). Second, we generalize from the squared L^2 norm to a general balance criterion h_ζ; another promiment example is an L^∞ constraint (see e.g. Zubizarreta, 2015; Athey et al., 2018).

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

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

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

There is a large literature relating balancing weights to propensity score weights. This literature shows that the loss function in Equation (A.39) is an M-estimator for the propensity score and thus will be consistent for the propensity score parameters under large N asymptotics. The dispersion measure $f(\cdot)$ determines the link function of the propensity score model, where the odds of treatment are $\frac{\pi(x)}{1-\pi(x)} = f^{*\prime}(\alpha + \boldsymbol{\beta}'x)$. Note that un-penalized SCM, which can yield multiple solutions, does not have a well-defined link function. We extend the duality to a general set of balance criteria so that Equation (A.39) is a regularized M-estimator of the propensity score parameters where the balance criterion $h_\zeta(\cdot)$ determines the type of regularization through its conjugate $h^*_\zeta(\cdot)$. This formulation recovers the duality between entropy balancing and a logistic link (Zhao and Percival, 2017), Oaxaca-Blinder weights and a log-logistic link (Kline, 2011), and L^∞ balance and L^1 regularization (Wang and Zubizarreta, 2019). This more general formulation also suggests natural extensions of both SCM and ASCM beyond the L^2 setting to other forms, especially L^1 regularization.

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

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

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

$$\begin{aligned} &\min_{\boldsymbol{\gamma},\boldsymbol{\epsilon}} \; h_\zeta(\boldsymbol{\epsilon}) + \sum_{W_i=0} f(\gamma_i). \\ &\text{subject to } \boldsymbol{\epsilon} = \boldsymbol{X}_1 - \boldsymbol{X}_0'.\boldsymbol{\gamma} \\ &\sum_{W_i=0} \gamma_i = 1 \end{aligned} \tag{A.40}$$

The Lagrangian is

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\epsilon}, \alpha, \boldsymbol{\beta}) = \sum_{i|W_i=0} f(\gamma_i) + \alpha(1-\gamma_i) + h_\zeta(\boldsymbol{\epsilon}) + \boldsymbol{\beta}'(\boldsymbol{X}_1 - \boldsymbol{X}_0'.\boldsymbol{\gamma} - \boldsymbol{\epsilon}). \tag{A.41}$$

The dual maximizes the objective

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

By strong duality the general dual problem (A.39), which minimizes $-q(\alpha, \boldsymbol{\beta})$, is equivalent to the primal balancing weights problem. Given the $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ that minimize the Lagrangian dual objective, $-q(\alpha, \boldsymbol{\beta})$, we recover the donor weights solution to (A.38) as

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

□

A.6 Additional figures

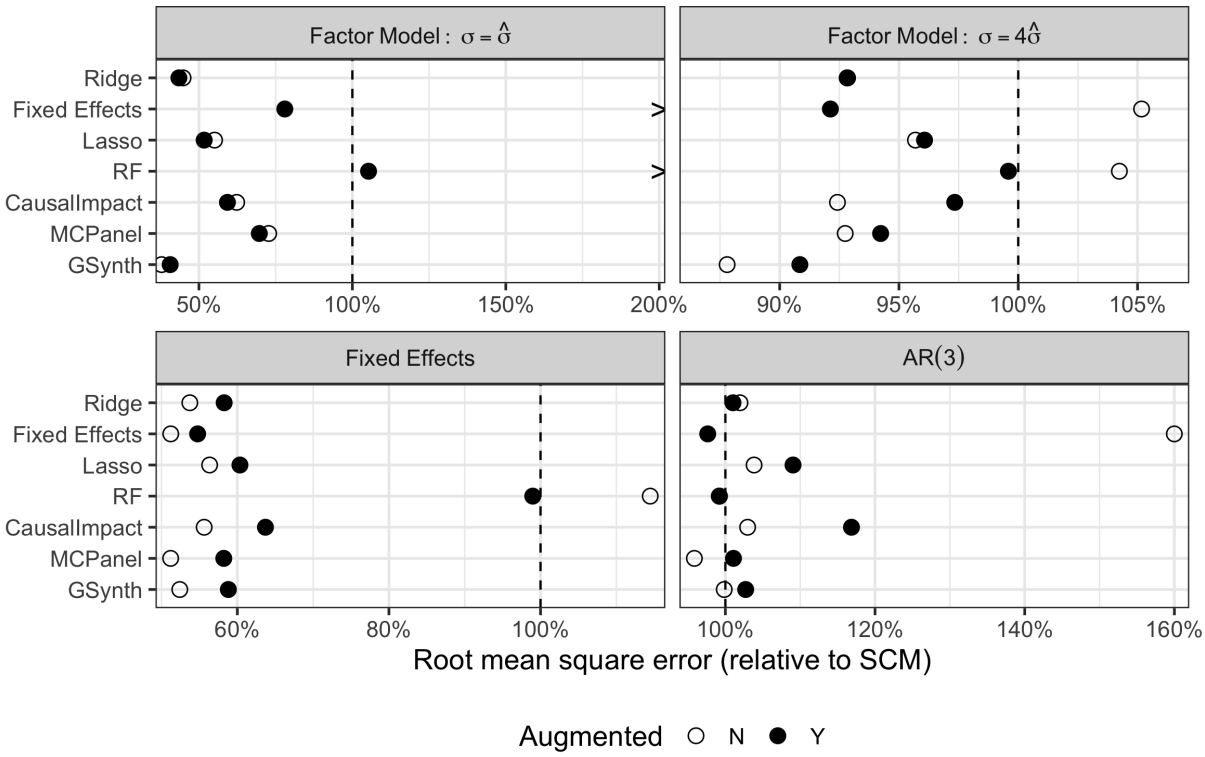


Figure A.1: RMSE for different augmented and non-augmented estimators across outcome models.

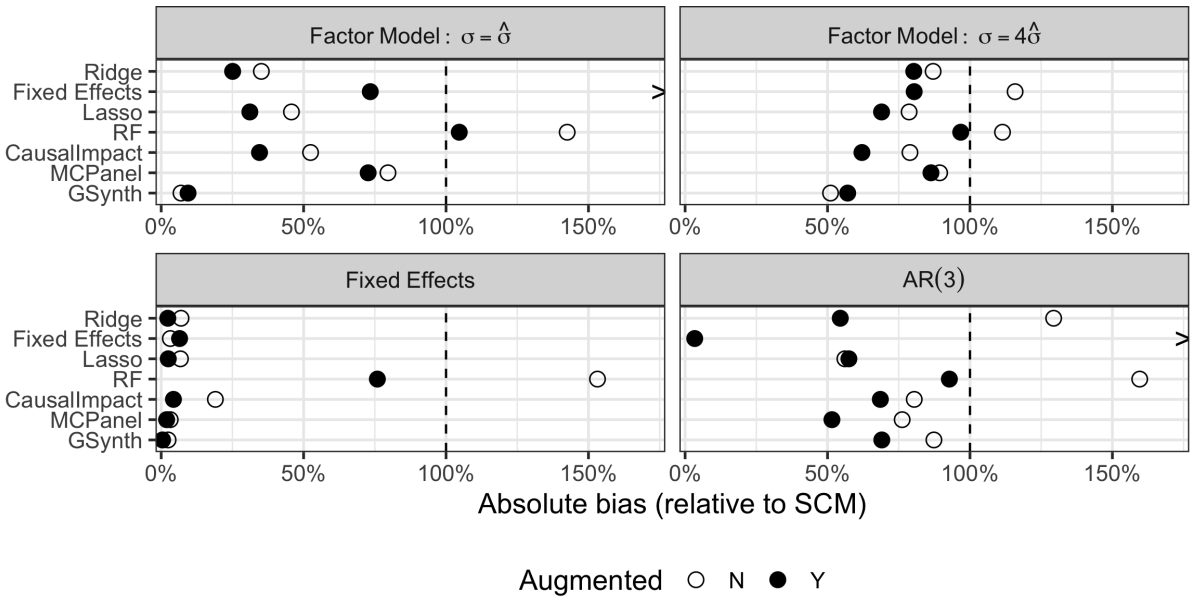


Figure A.2: Absolute bias (as a percentage of SCM bias) for ridge, fixed effects, and several machine learning and panel data outcome models, and their augmented versions using the same data generating processes as Figure 2.3.

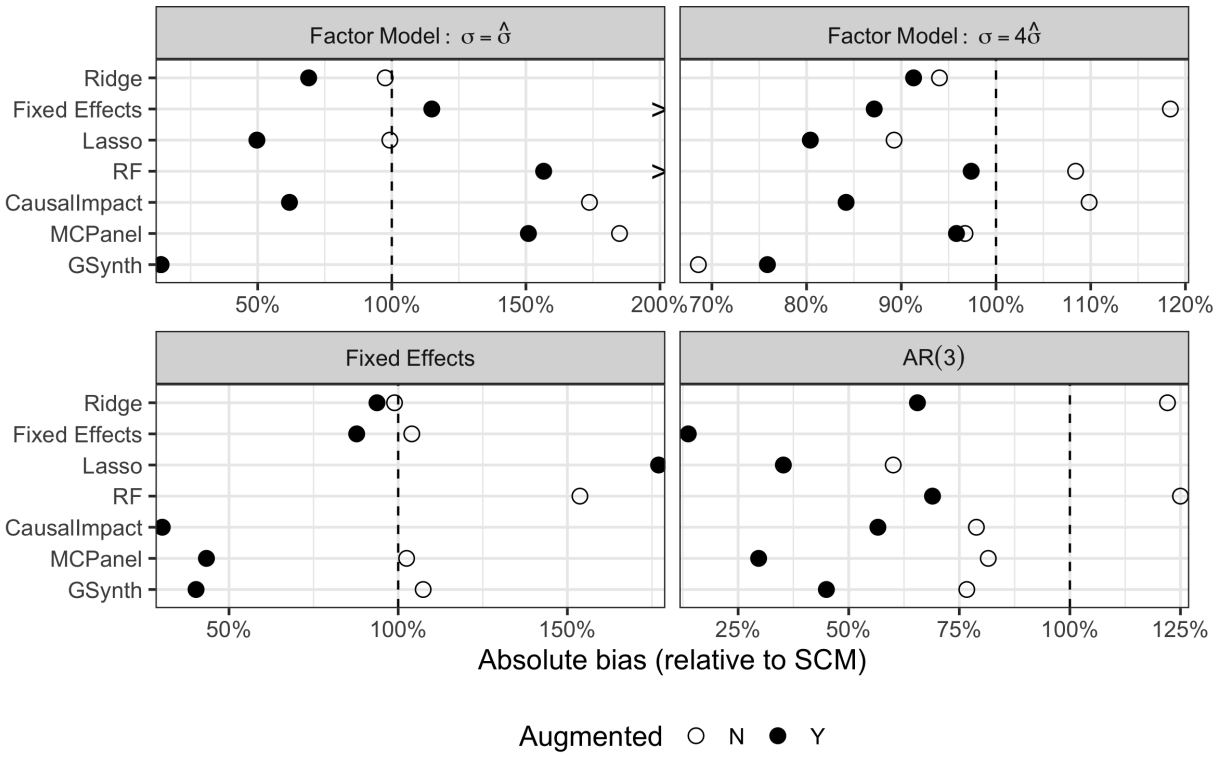


Figure A.3: Bias for different augmented and non-augmented estimators across outcome models conditioned on SCM fit in the top quintile.

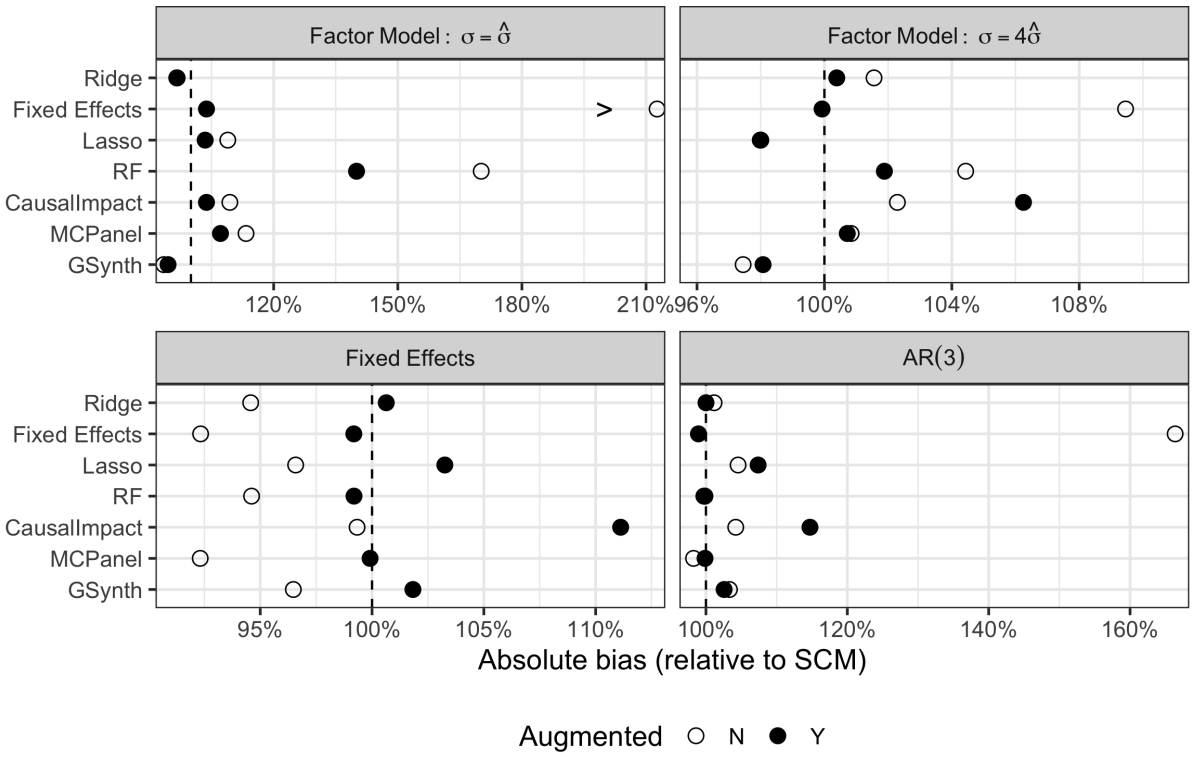


Figure A.4: RMSE for different augmented and non-augmented estimators across outcome models conditioned on SCM fit in the top quntile.

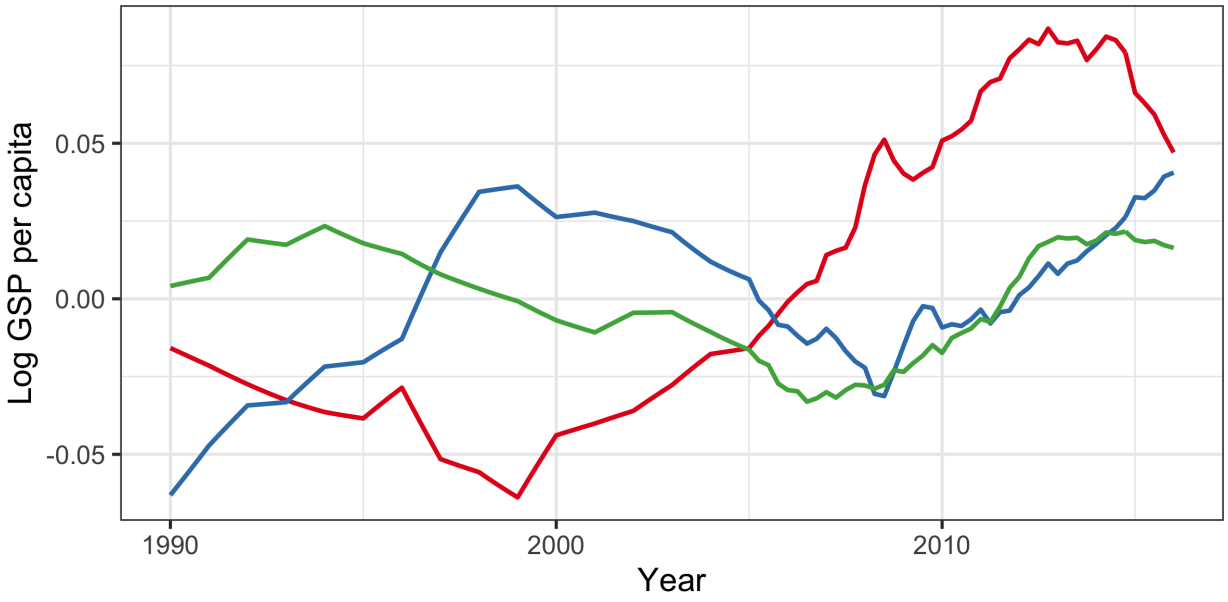


Figure A.5: Latent factors for calibrated simulation studies.

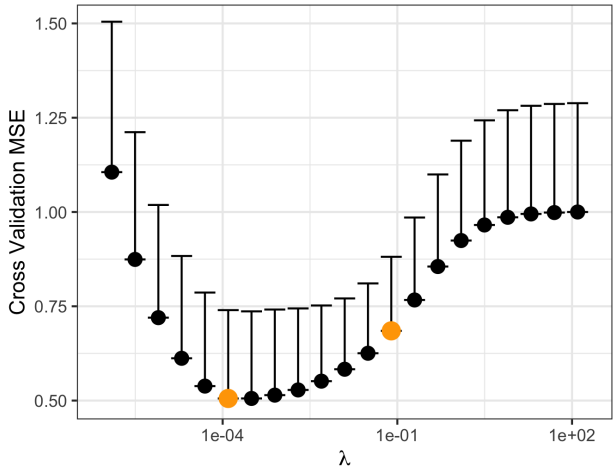


Figure A.6: Cross validation MSE and one standard error computed according to Equation (2.27). The minimal point, and the maximum λ within one standard error of the minimum are highlighted.

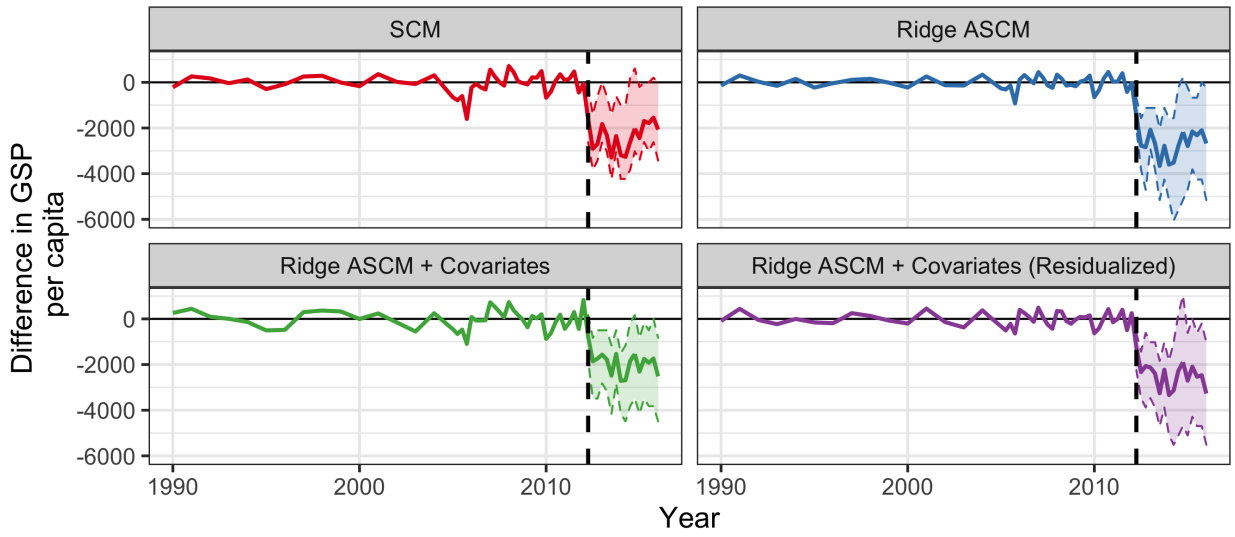


Figure A.7: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on GSP per capita using SCM, ridge ASCM, and ridge ASCM with covariates.

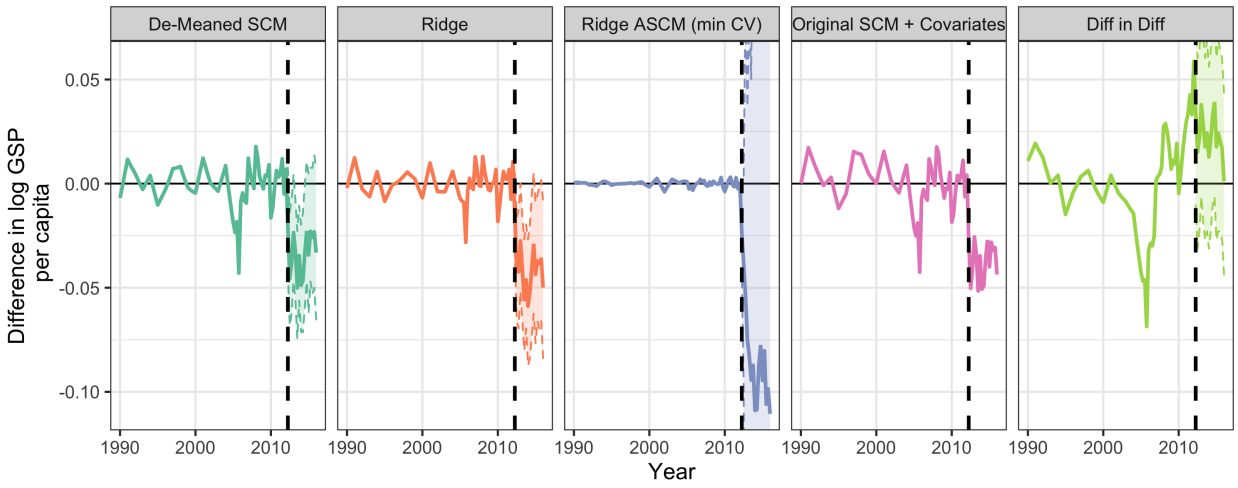


Figure A.8: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on log GSP per capita using de-meaned SCM, ridge regression alone, ridge ASCM with λ chosen to minimize the cross validated MSE, the original SCM proposal with covariates (Abadie et al.,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

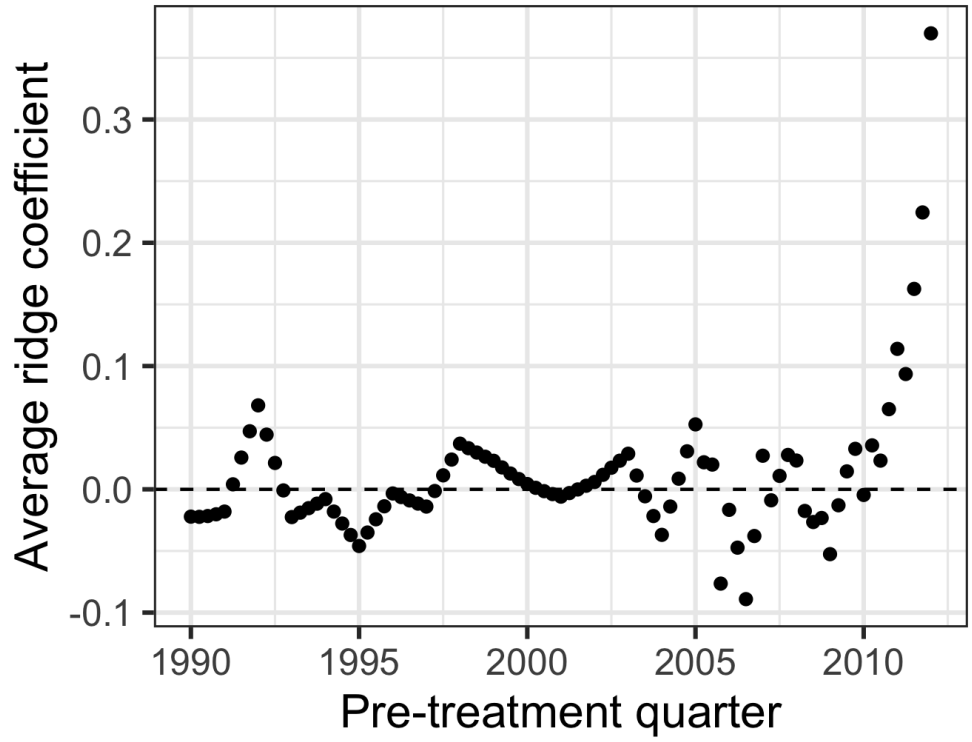


Figure A.9: Ridge regression coefficients for each pre-treatment quarter, averaged across post-treatment quarters.

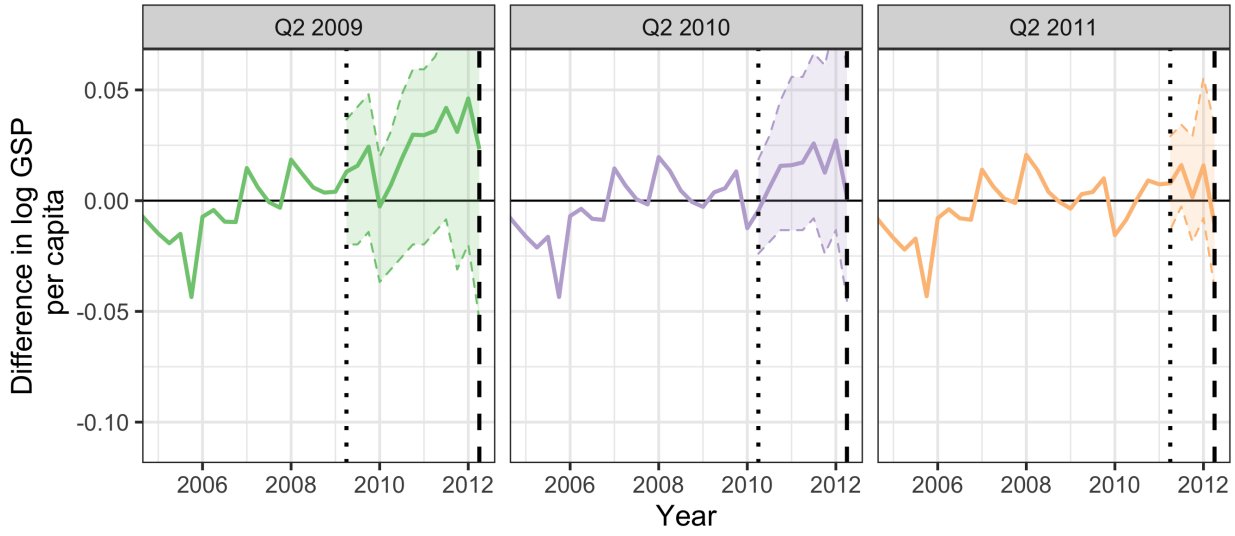


Figure A.10: Placebo point estimates along with 95% conformal confidence intervals for SCM with placebo treatment times in Q2 2009, 2010, and 2011. Scale begins in 2005 to highlight placebo estimates.

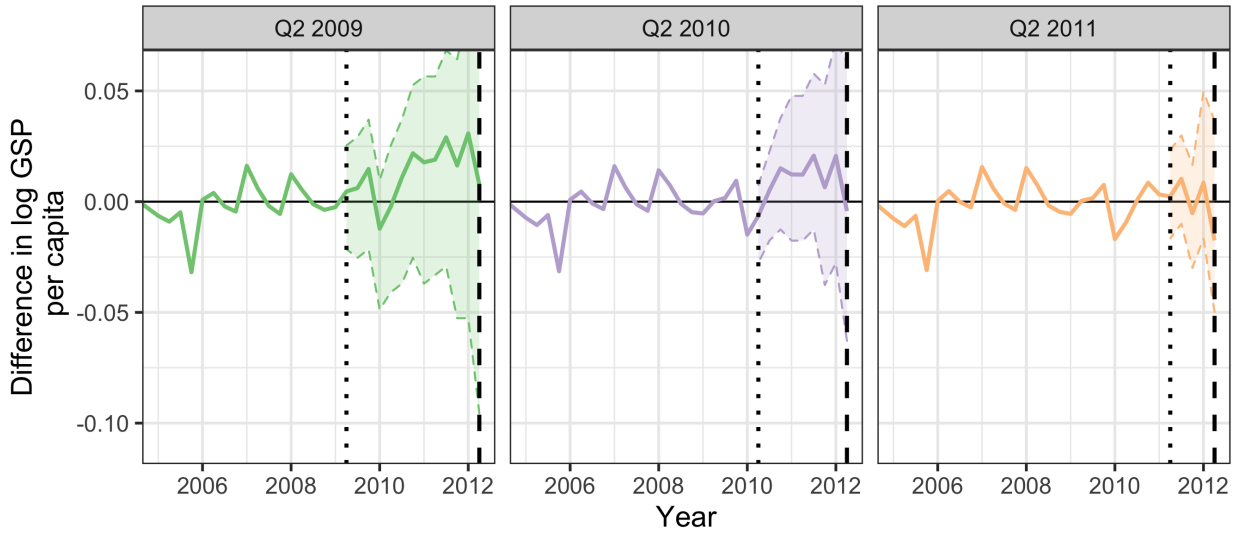


Figure A.11: Placebo point estimates along with 95% conformal confidence intervals for ridge ASCM with placebo treatment times in Q2 2009, 2010, and 2011. Scale begins in 2005 to highlight placebo estimates.

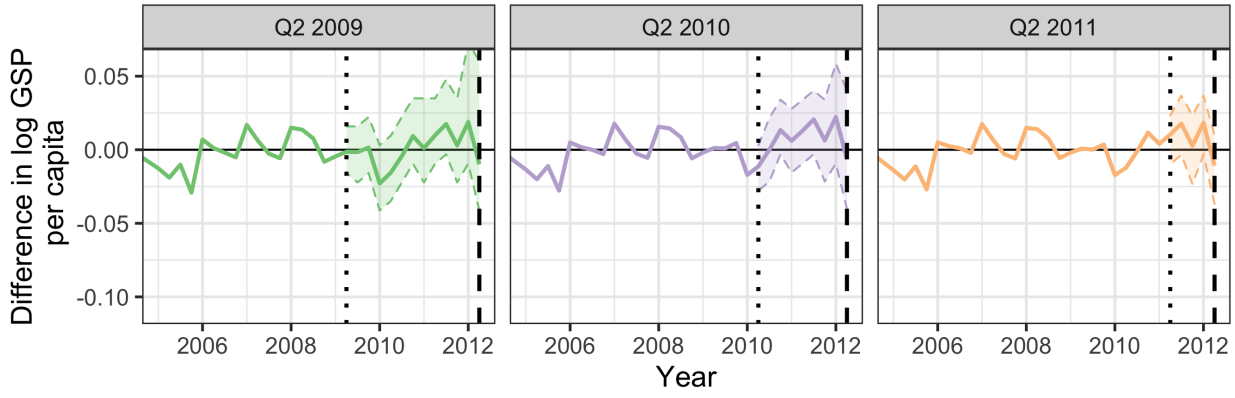


Figure A.12: Placebo point estimates along with 95% conformal confidence intervals for Ridge ASCM with covariates with placebo treatment times in Q2 2009, 2010, and 2011. The time period begins in 2005 and ends in Q1 2012 to highlight placebo estimates.

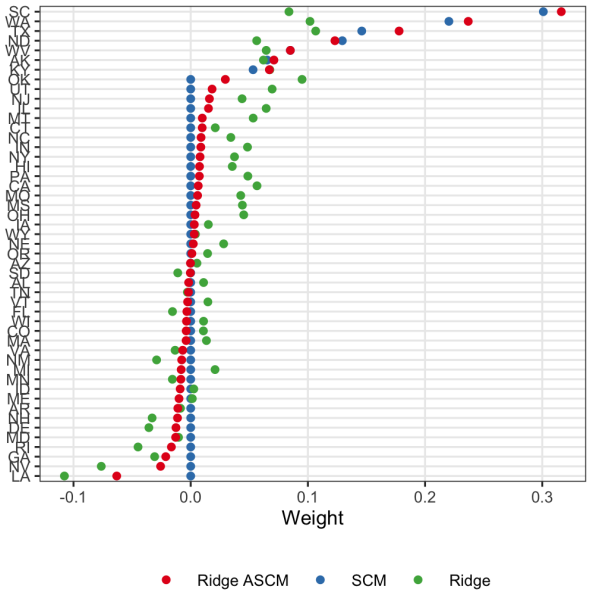


Figure A.13: Donor unit weights for SCM, ridge regression, and ridge ASCM balancing lagged outcomes.

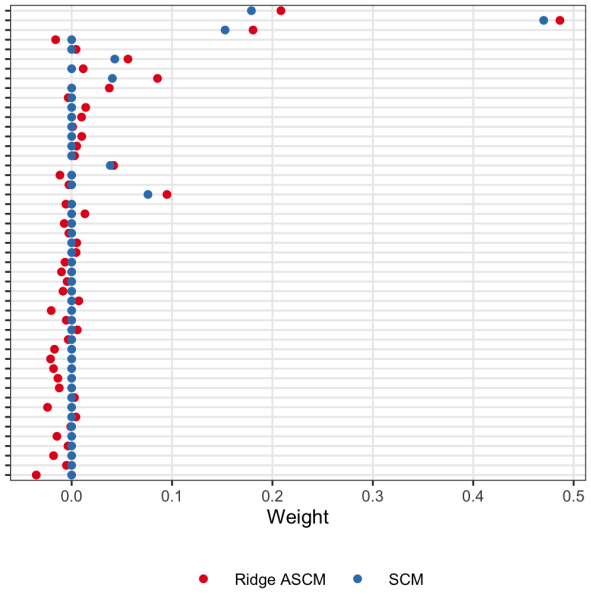


Figure A.14: Donor unit weights for SCM and ridge ASCM fit on lagged outcomes after residualizing out auxiliary covariates.

Appendix B

Supplementary materials for Chapter 3

B.1 The dual perspective: generalized propensity score weighting and conditional parallel trends

As we discuss in the main text, we can view partially pooled SCM as a form of balancing weights. By exploiting the duality between balancing weights and inverse propensity weighting, we can then interpret these weights as a form of (generalized) propensity score weighting (Imbens, 2000; Imai and Van Dyk, 2004). We further use this dual interpretation to show identification under a particular conditional parallel trends assumption (Abadie, 2005). For this section we consider combining units into treatment time cohorts indexed by Z_i where $Z_i = j$ if $T_i = T_j < \infty$ and $Z_i = 0$ if $T_i = \infty$ for control units. This is equivalent to fully pooling the synthetic control for units that share a treatment time.

Interpretation: inverse generalized propensity score weighting

In the case of a single treated unit, Appendix A.5 relates the synthetic control problem to propensity score estimation through the Lagrangian dual; see also Zhao and Percival (2017), Zhao (2018), and Wang and Zubizarreta (2019). A simple extension of that result shows that the loss function $\mathcal{L}(\alpha, \beta)$ in the dual problem (3.8) estimates the parameters of J separate propensity score models that are linear in the lagged outcomes, with link function $f^{*\prime}$; i.e. the propensity score model follows

$$f'\left(\frac{P(Z_i = j \mid Y_{i,T_j-L}, \ldots, Y_{i,T_j-1})}{P(Z_i = 0 \mid Y_{i,T_j-L}, \ldots, Y_{i,T_j-1})}\right) = \alpha_j + \sum_{\ell=1}^{L} \beta_{\ell j} Y_{i,T_j-\ell}, \quad j = 1, \ldots, J. \tag{B.1}$$

This dual problem has the form of a propensity score model for the probability of assignment to treatment level j, $Z_i = j$, relative to control, $Z_i = 0$, conditioned on the previous T_1 outcomes. The loss function $\mathcal{L}(\alpha, \beta)$ is a so-called *calibrated* loss, rather than the more standard

likelihood approach. Using a calibrated loss generally leads to better finite sample properties for the resulting weights than more traditional methods; see Tan (2017). Jointly, this series of separate logistic models has the form of a multinomial regression for the *generalized propensity score* (Imbens, 2000).

Armed with this IPW interpretation of the loss function, we can shed more light on the regularization in Equation (3.8) by comparing it to the corresponding regularized maximum likelihood estimate. This estimate is a *maximum a priori* estimate of the following hierarchical propensity score model (Li et al., 2013):

$$\begin{aligned} f'\left(\frac{P(Z_i = j \mid Y_{i,T_j-L}, \ldots, Y_{i,T_j-1})}{P(Z_i = 0 \mid Y_{i,T_j-L}, \ldots, Y_{i,T_j-1})}\right) &= \alpha_j + \sum_{\ell=1}^{L} \beta_{\ell j} Y_{i,T_j-\ell}, \quad j = 1, \ldots, J \\ \beta_{\ell j} \mid \mu_{\beta\ell} &\sim N\left(\mu_{\beta\ell}, \frac{(1-\nu)}{\lambda}\right) \\ \mu_{\beta\ell} &\sim N\left(0, \frac{\nu}{\lambda}\right). \end{aligned} \tag{B.2}$$

Written in a Bayesian form, we see that the dual problem (3.8) shrinks cohort specific parameters β_j towards a *global model* of treatment μ_β.

Finally, we note that while the *form* of Equation (B.1) holds whenever $\lambda > 0$, interpreting this model as a generalized propensity score requires additional considerations. First, treatment levels must be well defined. This is most natural when we consider cohorts with multiple treated units, and so can imagine another unit adopting treatment at the same time. Second, all control units must have well-defined treated potential outcomes. This is plausible in our collective bargaining example — we can conceive of never treated states adopting such laws — but is not always appropriate, such as in classic SCM examples (Abadie et al., 2015).

Connection to semiparametric DID and identification under conditional parallel trends

With this IPW interpretation, we consider the oracle estimator that uses the true (unpenalized) propensity score weights to estimate the ATT and show that it identifies causal treatment effects under a conditional parallel trends assumption. To do so, we show that this approach is a version of *semiparametric DID* (Abadie, 2005; Callaway and Sant'Anna, 2018) and then apply existing results. Unlike existing methods, however, the weighted event study approach instead conditions on pre-treatment dynamics, specifically the residuals after subtracting off the pre-treatment average.

To formalize these results, we make some additional assumptions that are standard in the event studies literature (see Callaway and Sant'Anna, 2018) but which might not necessarily hold in all settings. First, we assume that the observed units are sampled from an underlying population.

Assumption B.1 (Sampling)**.** $\{Y_{i1}, \ldots, Y_{iT}, T_i\}_{i=1}^{N} \overset{iid}{\sim} \mathcal{P}(\cdot)$ for some joint distribution $\mathcal{P}(\cdot)$

Second, we assume that every unit has a non-zero probability of adopting treatment at any time, both overall and conditional on lagged residuals.

Assumption B.2 (Overlap)**.** $\mathbb{P}(Z = j) > 0$ for all $j = 0, \ldots, J$.

Finally, we relax the parallel trends assumption to only hold conditionally. Following Hazlett and Xu (2018), we assume that parallel trends holds given the vector of residuals, $\dot{Y}_{it} = Y_{it} - \bar{Y}_{i,j}^{\text{pre}}$.

Assumption B.3 (Conditional parallel trends)**.** For $t' < t$ and $j = 1, \ldots, J$,

$$\mathbb{E}[Y_{it}(0) - Y_{it'}(0) \mid \dot{Y}_{i1}, \ldots, \dot{Y}_{iT_j-1}, Z = j] = \mathbb{E}[Y_{it}(0) - Y_{it'}(0) \mid \dot{Y}_{i1}, \ldots, \dot{Y}_{iT_j-1}, Z = 0]. \quad \text{(B.3)}$$

Assumption B.3 loosens the usual parallel trends assumption by allowing trends to differ depending on how the lagged outcomes deviate from their baseline value. Thus, we are essentially conditioning on pre-treatment "dynamics," rather than pre-treatment levels. For instance, even if two states have very different levels of student expenditures, under conditional parallel trends we can compare them so long as they have similar pre-treatment trends and shocks. See Hazlett and Xu (2018) for further discussion.

Given these assumptions, we now return to the SCM-weighted event study estimator in Equation (3.12). Let $\hat{p}_{ij} = \frac{\hat{\gamma}_{ij}}{1-\hat{\gamma}_{ij}}$ be the implied propensity score for unit i for treatment $Z_j = j$ from Equation (B.1). Then we can re-write Equation (3.12) as:

$$\hat{\tau}_{jk}^{\text{aug}} = \frac{1}{T_j - 1} \sum_{t=1}^{T_j-1} \left[\left(Y_{j,T_j+k} - Y_{j,t}\right) - \sum_{i=1}^{N} \frac{\hat{p}_{ij}}{1 - \hat{p}_{ij}} \left(Y_{i,T_j+k} - Y_{i,t}\right) \right]. \quad \text{(B.4)}$$

This is identical in form to the semiparametric DID estimators proposed by Abadie (2005) and Callaway and Sant'Anna (2018).[1] We can then immediately apply Theorem 1 in Callaway and Sant'Anna (2018). Specifically, the population analog of $\hat{\tau}_{jk}^{\text{aug}}$ identifies τ_{jk} under Assumptions B.1-B.3:

$$\begin{aligned} &\frac{1}{T_j - 1} \sum_{t=1}^{T_j-1} \Bigg\{ \mathbb{E}[Y_{T_j+k}(1) - Y_t(0) \mid Z = j] \; - \\ &\qquad \frac{\mathbb{P}(Z = j \mid \dot{Y}_1, \ldots, \dot{Y}_{T_j-1})}{\mathbb{P}(Z = 0 \mid \dot{Y}_1, \ldots, \dot{Y}_{T_j-1})} \mathbb{E}\left[Y_{T_j+k}(0) - Y_t(0) \mid Z = 0\right] \Bigg\} \quad \text{(B.5)} \\ = \;& \frac{1}{T_j - 1} \sum_{t=1}^{T_j-1} \mathbb{E}[Y_{T_j+k}(1) - Y_{T_j+k}(0) \mid Z = j] \\ = \;& \mathbb{E}[Y_{T_j+k}(1) - Y_{T_j+k}(0) \mid Z = j]. \end{aligned}$$

[1] The estimator in Callaway and Sant'Anna (2018) also standardizes by the the sum of $\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}$. This sum equals 1 by construction when $\hat{\Gamma}$ is estimated via the calibrated approach above. The estimator in Callaway and Sant'Anna (2018) also restricts estimation to $t = T_j - 1$. We slightly modify their Theorem 1 under the stronger assumption that conditional parallel trends hold for all pre-treatment times, $t = 1, \ldots, T_j$

This shows that the weighted event study approach estimates causal effects under weaker assumptions than are needed for traditional event studies. This is not the only proposal to generalize DID: existing semiparametric approaches condition on auxiliary or time-invariant covariates, rather than on lagged outcomes. We anticipate that blended strategies that condition both on auxiliary covariates and pre-treatment outcome dynamics may be attractive in many applications.

B.2 Additional simulation results

We now describe the two-way fixed effects model and the autoregressive model that we use in our simulation studies. First, the two-way fixed effects model is:

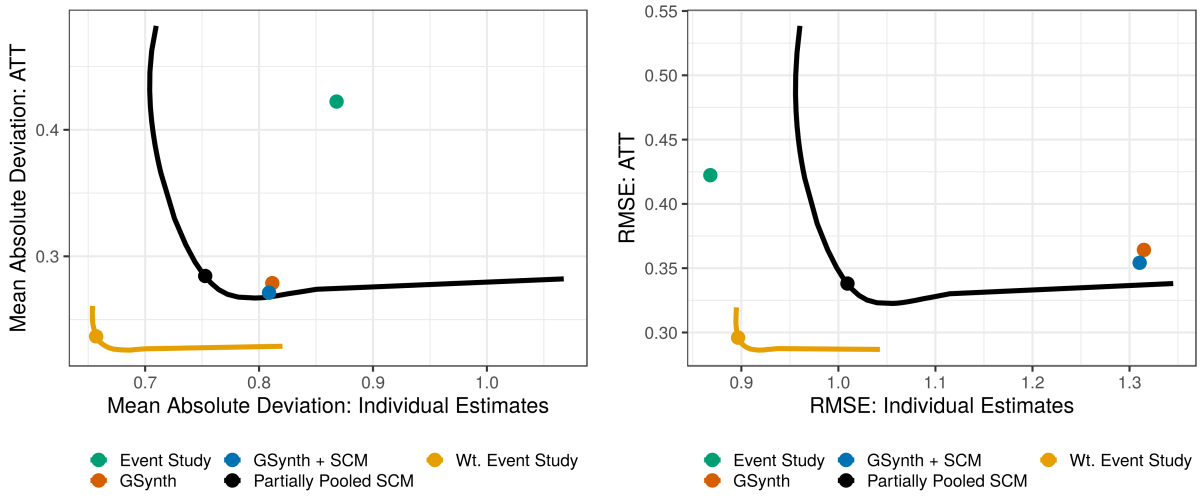
$$Y_{it} = \text{int} + \text{unit}_i + \text{time}_t + \varepsilon_{it}. \tag{B.6}$$

Here, both unit and time effects are normalized to have mean zero. We estimate (B.6) using just the un-treated observations, and extract the estimated variance of the unit effects, $\hat{\Sigma}$, and of the error term, $\hat{\sigma}^2_\varepsilon$. We then generate simulated data sets with the same $N = 49$ and $T = 38$ where $\text{unit}_i \overset{\text{iid}}{\sim} N(0, \hat{\Sigma})$ and $\varepsilon_{it} \overset{\text{iid}}{\sim} N(0, \hat{\sigma}^2_\varepsilon)$. We impose a sharp null of no treatment effect, $Y_{it}(1) = Y_{it}(0) = Y_{it}$. This model satisfies the parallel trends assumption needed for traditional event studies and DID models. We set the probability that unit i is treated at each treatment time to be $\pi_i = \text{logit}(\theta_0 + \theta_1 \cdot \text{unit}_i)$, with $\theta_0 = -2.7$ and $\theta_1 = -1$ to ensure that around 30 units are eventually treated in each simulation draw.

We also consider a random effects autoregressive model:

$$\begin{aligned} Y_{it} &= \sum_{\ell=1}^{3} \rho_\ell Y_{i,t-\ell} + \varepsilon_{it} \\ \rho &\sim N(\mu_\rho, \sigma^2_\rho). \end{aligned} \tag{B.7}$$

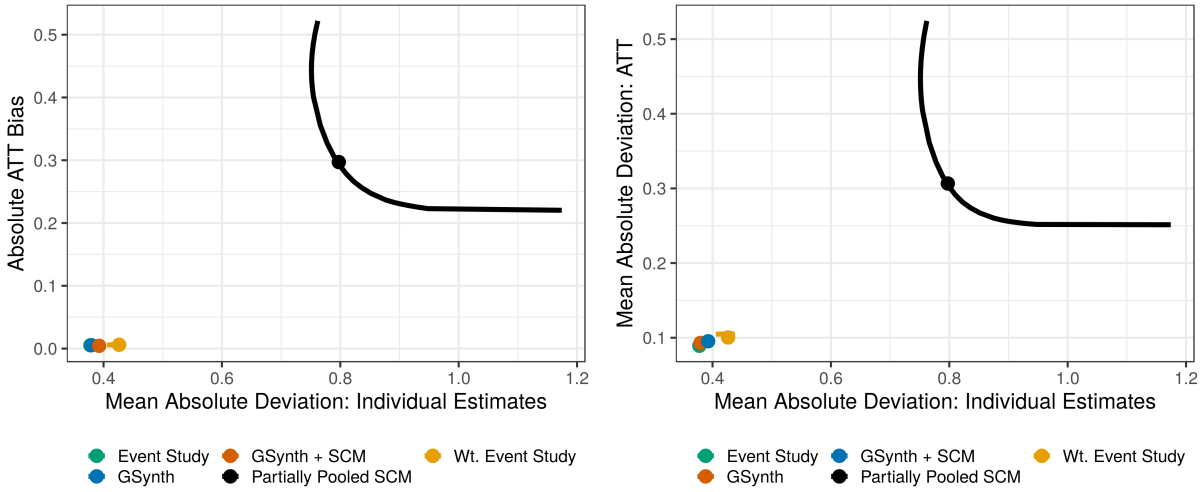
We fit this random effects model using `lme4` (Bates et al., 2015) to get estimates $\hat{\mu}_\rho$ and $\hat{\sigma}_\rho$. In order to increase the level of heterogeneity across time, we simulate from this hierarchical model with 8 times the standard deviation $8\hat{\sigma}_\rho$. We allow selection to depend on the three lagged outcomes $\pi_i = \text{logit}\left(\theta_0 + \theta_1 \sum_{\ell=1}^{L} Y_{i,t-\ell}\right)$, where $\theta_0 = \log 0.04$ and $\theta_1 = -2$.



(a) MAD for overall ATT vs MAD for individual ATT estimates

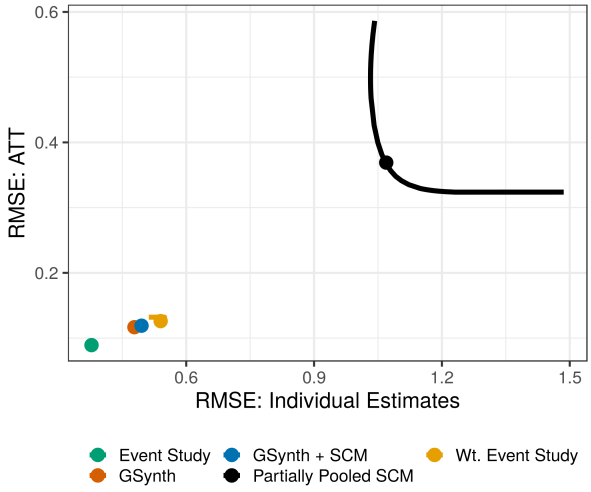
(b) RMSE for overall ATT vs RMSE for individual ATT estimates

Figure B.1: Monte Carlo estimates of the MAD and RMSE for the overall ATT and the individual ATT estimates under a linear factor model (3.14).



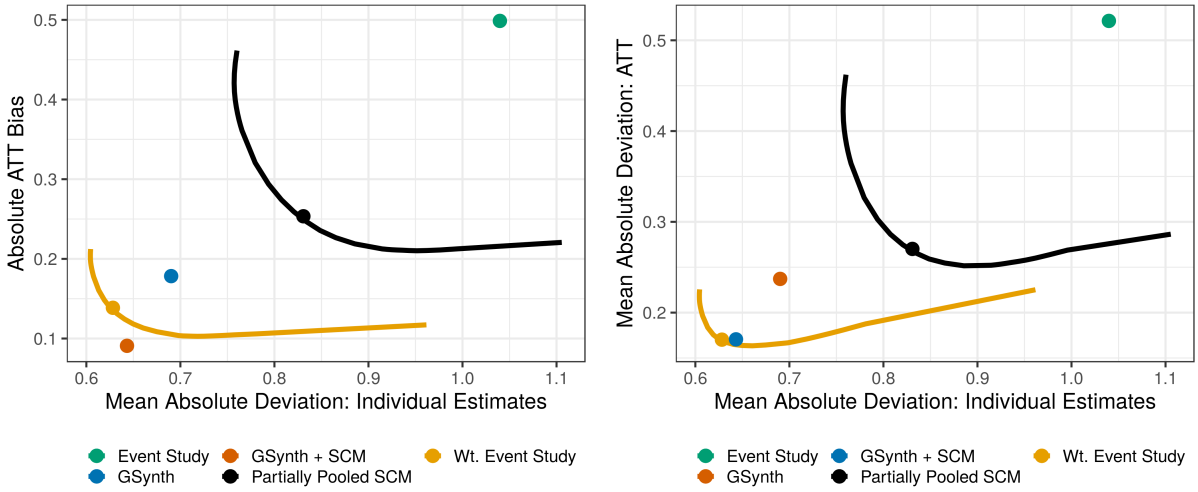
(a) Bias for overall ATT vs MAD for individual ATT estimates

(b) MAD for overall ATT vs MAD for individual ATT estimates



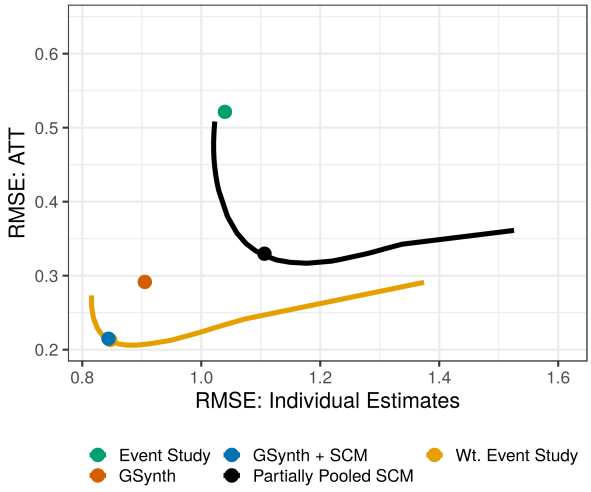
(c) RMSE for overall ATT vs RMSE for individual ATT estimates

Figure B.2: Monte Carlo estimates of the MAD and RMSE for the overall ATT and the individual ATT estimates under a two-way fixed effects model (B.6).



(a) Bias for overall ATT vs MAD for individual ATT estimates

(b) MAD for overall ATT vs MAD for individual ATT estimates



(c) RMSE for overall ATT vs RMSE for individual ATT estimates

Figure B.3: Monte Carlo estimates of the MAD and RMSE for the overall ATT and the individual ATT estimates under a random effects AR model (B.7).

B.3 Additional results and figures for mandatory collective bargaining

Event study estimates for the teacher collective bargaining application

A common approach for estimating causal effects under staggered adoption is by fitting a variant of the two-way fixed effect model, known as the *event study*, or dynamic treatment effect, specification:

$$Y_{it} = \text{unit}_i + \text{time}_t + \sum_{\ell=2}^{L} \delta_\ell \mathbb{1}\{T_i = t - \ell\} + \sum_{k=0}^{K} \tau_k \mathbb{1}\{T_i = t + k\} + \varepsilon_{it}, \tag{B.8}$$

for outcome Y_{it} for state i at time t, where $L = T_J$ and $K = T - T_1$ are the maximum number of pre-treatment outcomes (lags) and post-treatment outcomes (leads) observed for all treated units in the sample. Following standard practice, the coefficient for δ_1 is normalized to zero. This is arbitrary, and researchers sometimes impose a different normalization (e.g., $\delta_L = 0$). If all units are eventually treated, a second normalization is required. Paglayan (2019) uses a slightly different normalization to the equation shown here, though the differences are immaterial. See Abraham and Sun (2018); Callaway and Sant'Anna (2018) for further discussion of this workhorse model.

Figures B.10 and B.11a shows the results from estimating Equation (B.8) on the full data in Paglayan (2019) for per-pupil expenditures and average teacher salary, respectively; we show standard errors clustered by state (Pustejovsky and Tipton, 2018). The placebo estimates to the left of treatment adoption time, the coefficients $\hat{\delta}_\ell$ from Equation (B.8), show that states that pass laws have declining expenditures — and, to a lesser degree, declining salaries — in the several years prior to adoption, relative to other states at the same time. These "pre-trends" suggest that the critical *parallel trends* assumption is likely violated in this setting, raising doubts about the estimates to the right of the treatment adoption time, the coefficients $\hat{\tau}_k$ from Equation (B.8). Paglayan (2019) estimates Equation (B.8) using only the ever-treated states, for which there is no evidence against the parallel trends assumption. Figures B.10 and B.11a are nearly identical to the corresponding figures from the supplementary materials in Paglayan (2019).

Additional figures

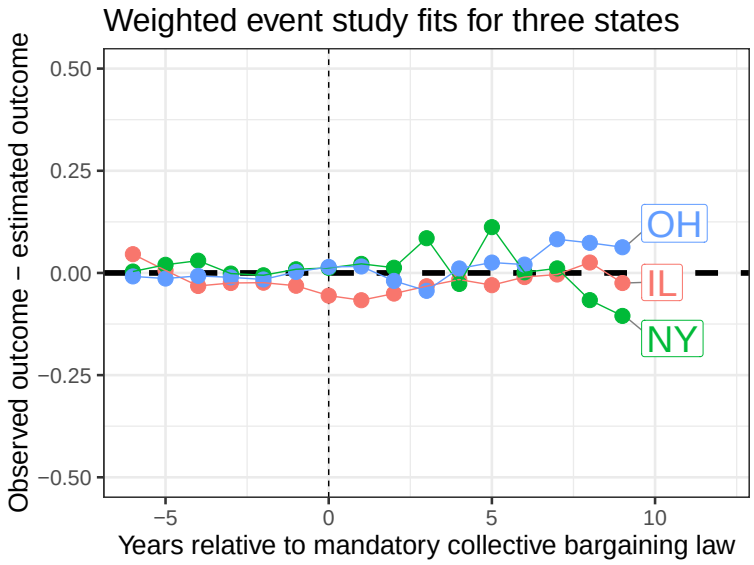
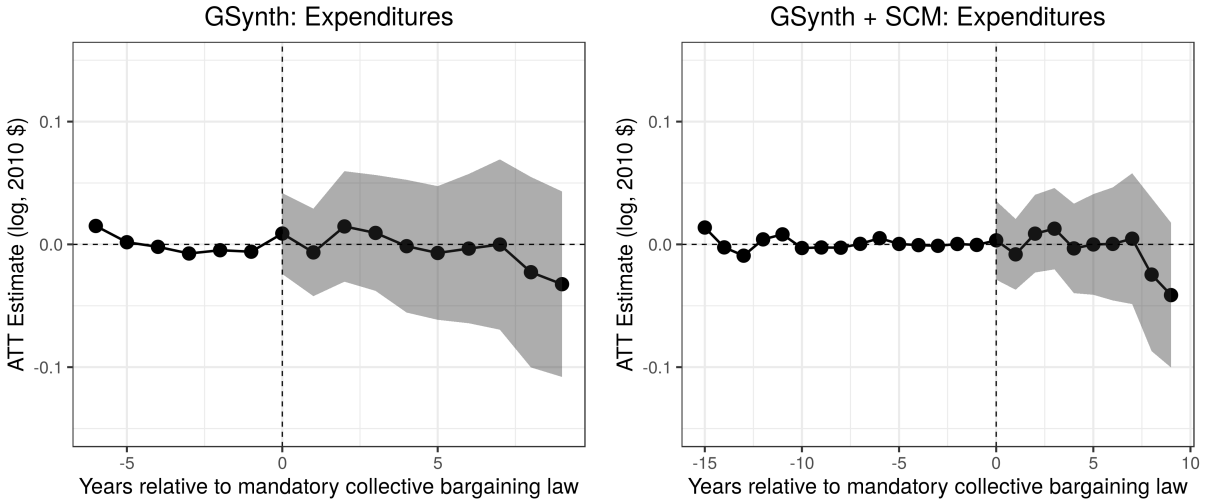


Figure B.4: Illustrative fits for the weighted event study



(a) `gsynth` alone

(b) Partially pooled SCM augmented with `gsynth`

Figure B.5: `gsynth` and augmented estimates for per-pupil student expenditures (log, 2010 $).

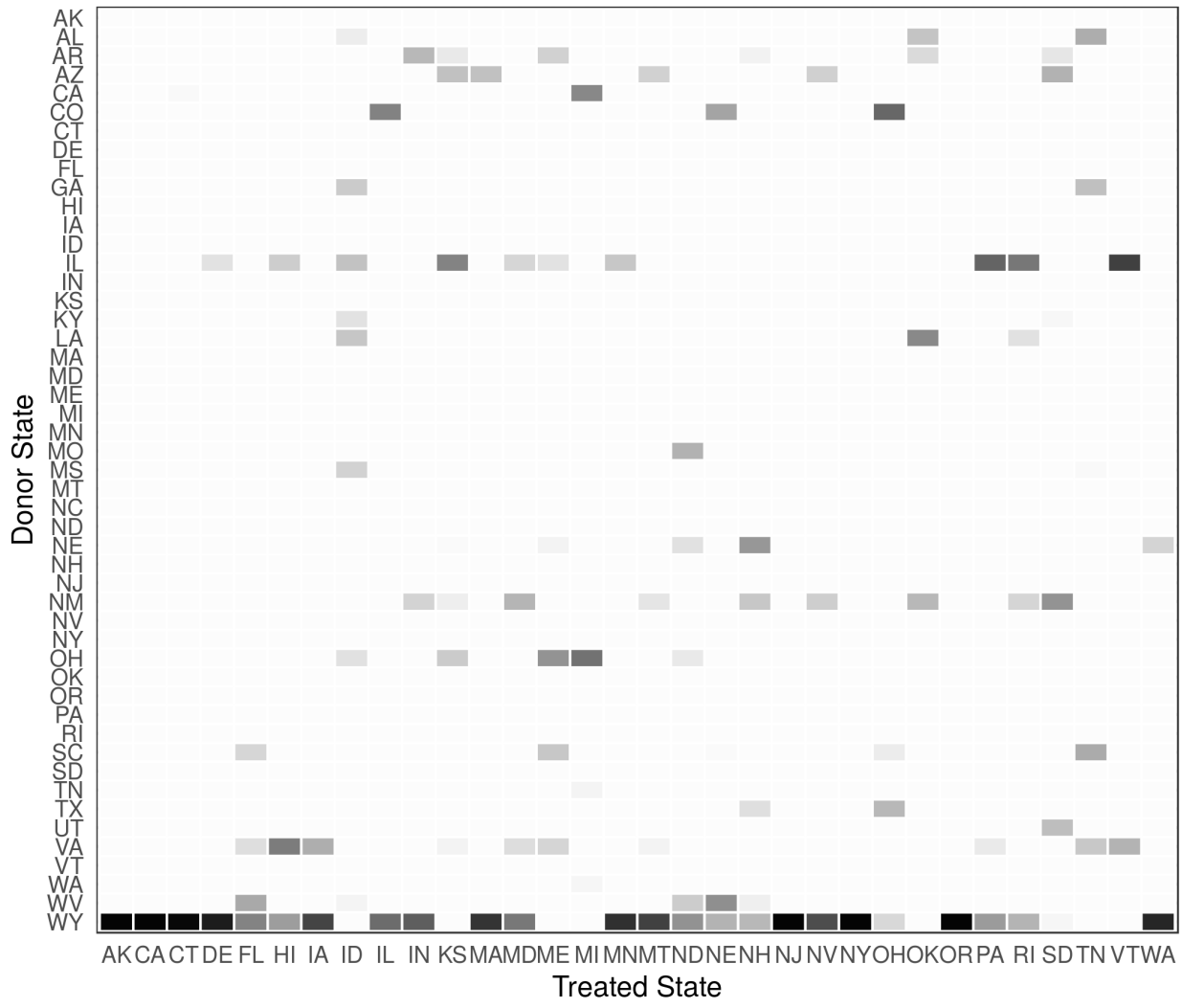


Figure B.6: Partially pooled SCM weights. White cells indicate zero weight, black cells indicate a weight of 1.

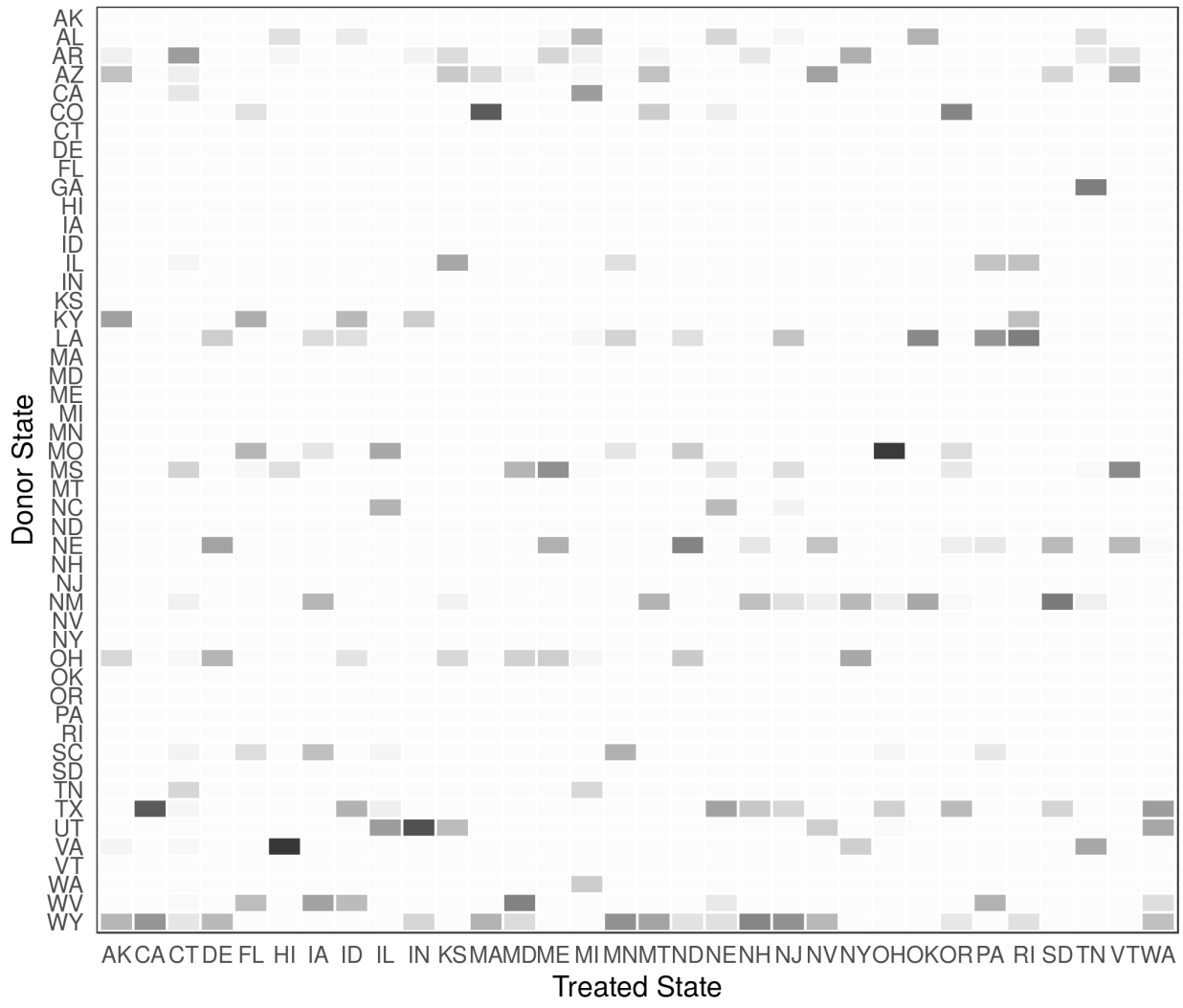


Figure B.7: Weighted event study weights. White cells indicate zero weight, black cells indicate a weight of 1.

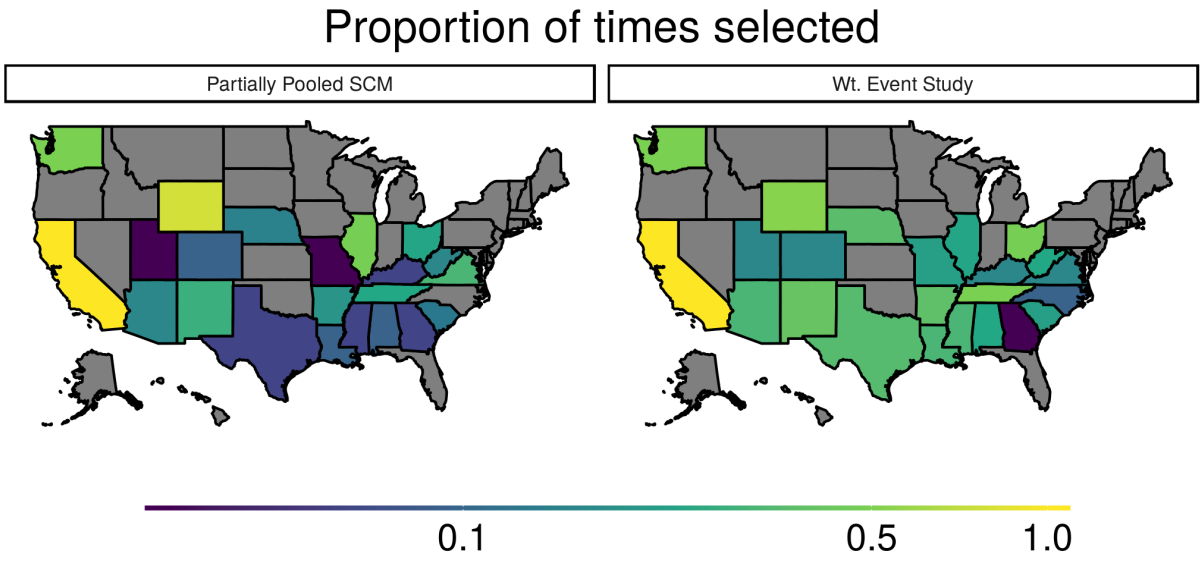
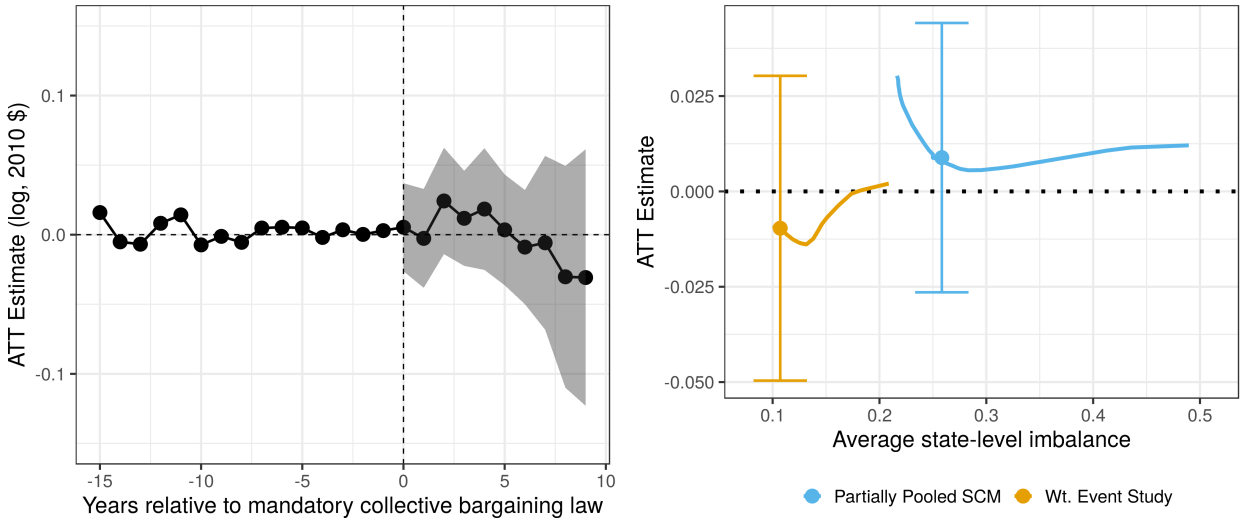


Figure B.8: Partially pooled and weighted event study weights. The number of times that each state is part of a treated state's synthetic control, normalized by the number of times it is a possible donor state. Note that California is an eligible donor state in only two cases. Colors on a log scale.



(a) Removing NY and AK ($\hat{\nu} = 0.27$)

(b) Varying ν from 0 to 1.

Figure B.9: (a) Partially pooled SCM estimates removing the two worst fit states, (b) $\widehat{\text{ATT}}$ as ν varies between 0 and 1 (plotted against q^{sep}), estimates and approximate 95% confidence intervals using $\hat{\nu}$ shown.

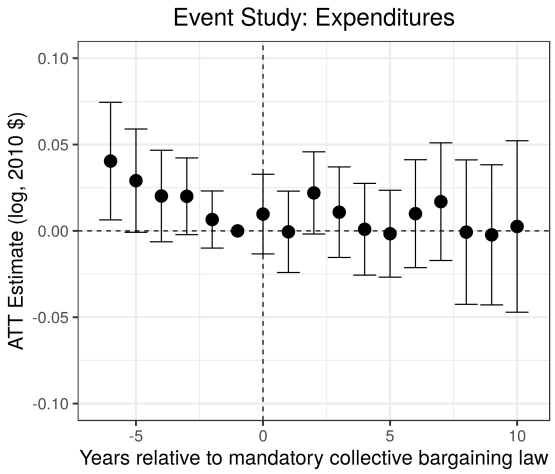


Figure B.10: Event study estimates for per pupil expenditures (log 2010 $).

(a) Event study regression

(b) Partially Pooled SCM, $\hat{\nu} = 0.50$

Figure B.11: (a) Event study and (b) partially pooled SCM estimates for average teacher salary (log, 2010 $).

B.4 Additional results and proofs

Partial pooling in dual parameters

Lemma B.1. The Lagrangian dual to Equation (3.2) is

$$\min_{\alpha,\beta} \underbrace{\sum_{j=1}^{J}\left[\sum_{W_i=0} f^*\left(\alpha_j + \sum_{\ell=1}^{L}\beta_{\ell j}Y_{i,T_j-\ell}\right) - \left(\alpha_j + \sum_{\ell=1}^{L}\beta_{\ell j}Y_{j,T_1-\ell}\right)\right]}_{\mathcal{L}(\alpha,\beta)} + \sum_{j=1}^{J}\frac{\lambda JL}{2}\|\beta_j\|_2^2, \quad \text{(B.9)}$$

where $f^*(y) = \sup_x x'y - f(x)$ is the convex conjugate of f. The resulting donor weights are $\hat{\gamma}_{ij} = f^{*\prime}\left(\hat{\alpha}_j - \sum_{\ell=1}^{L}\hat{\beta}_{\ell j}Y_{i,T_j-\ell}\right)$.

Proof of Lemma B.1. Notice that the separate synth problem (3.2) separates into J optimization problems:

$$\begin{aligned}
&\min_{\gamma_1,\ldots,\gamma_J\in\Delta_j^{\text{scm}}} \quad q^{\text{sep}}(\Gamma) \;+\; \lambda\sum_{j=1}^{J}\sum_{i=1}^{N} f(\gamma_{ij}) \\
=&\frac{1}{2J}\sum_{j=1}^{J}\min_{\gamma_j\in\Delta_j^{\text{scm}}}\left\{\left[\frac{1}{T_j-1}\sum_{\ell=1}^{T_j-1}\left(Y_{j,T_j-\ell} - \sum_{i=1}^{N}\gamma_{ij}Y_{i,T_j-\ell}\right)^2\right] + \lambda\sum_{i=1}^{N} f(\gamma_{ij})\right\}
\end{aligned} \quad \text{(B.10)}$$

Thus the Lagrangian dual objective is the sum of the Langrangian dual objectives of the individual objectives in Equation (B.10). Inserting the dual objectives derived in Appendix A.5 yields the result. □

Proof of Proposition 3.1. We start be defining auxiliary variables, $\mathcal{E}_0, \mathcal{E}_1, \ldots, \mathcal{E}_J \in \mathbb{R}^L$ where $\mathcal{E}_{j\ell} = Y_{j,T_j-\ell} - \sum_{i=1}^{N}\gamma_{ij}Y_{iT_j-\ell}$ for $j \geq 1$ and $\mathcal{E}_{0\ell} = \sum_{T_j>\ell}\left(Y_{j,T_j-\ell} - \sum_{i=1}^{N}\gamma_{ij}Y_{i,T_j-\ell}\right)$. Additionally we rescale by $\frac{1}{\lambda}$. Then we can write the partially pooled SCM problem (3.5) as

$$\begin{aligned}
\min_{\gamma_1,\ldots,\gamma_J,\mathcal{E}_0,\ldots,\mathcal{E}_J} \quad & \frac{\nu}{2L\lambda}\sum_{\ell=1}^{L}\mathcal{E}_{0\ell}^2 + \frac{1-\nu}{2J\lambda}\sum_{j=1}^{J}\frac{1}{L}\mathcal{E}_{j\ell}^2 + \sum_{j=1}^{J}\sum_{i=1}^{N} f(\gamma_{ij}) \\
\text{subject to} \quad & \mathcal{E}_{j\ell} = Y_{j,T_j-\ell} - \sum_{i=1}^{N}\gamma_{ij}Y_{iT_j-\ell} \\
& \mathcal{E}_{0\ell} = \sum_{T_j>\ell}\left(Y_{j,T_j-\ell} - \sum_{i=1}^{N}\gamma_{ij}Y_{i,T_j-\ell}\right) \\
& \gamma_j \in \Delta_j^{\text{scm}}
\end{aligned} \quad \text{(B.11)}$$

With Lagrange multipliers $\mu_\beta, \zeta_1, \ldots, \zeta_J \in \mathbb{R}^L$ and $\alpha_1, \ldots, \alpha_J \in \mathbb{R}$, the Lagrangian to Equation (B.11) is

$$\begin{aligned}
\mathcal{L}(\Gamma, \mathcal{E}_0, \ldots, \mathcal{E}_J, \alpha_1, \ldots, \alpha_J, \mu_\beta, \beta_1, \ldots, \beta_J) = \\
&\sum_{\ell=1}^{L} \left[\frac{\nu}{2L\lambda} \mathcal{E}_{0\ell}^2 - \mu_{\beta,\ell} \left(\sum_{j=1}^{J} Y_{j,T_j-\ell} - \sum_{i \in \mathcal{D}_j} \gamma_{ij} Y_{i,T_j-\ell} \right) - \mathcal{E}_{0\ell} \mu_{\beta,\ell} \right] \\
+ \quad &\sum_{j=1}^{J} \sum_{\ell=1}^{L} \left[\frac{1-\nu}{2JL\lambda} \mathcal{E}_{j\ell}^2 - \zeta \ell j \left(Y_{j,T_j-\ell} - \sum_{i \in \mathcal{D}_j} \gamma_{ij} Y_{i,T_j-\ell} \right) - \zeta_{\ell j} \mathcal{E}_{j\ell} \right] \\
+ \quad &\sum_{j=1}^{J} \sum_{i \in \mathcal{D}_j} f(\gamma_{ij}) - \alpha_j \gamma_{ij} - \alpha_j
\end{aligned} \tag{B.12}$$

Defining $\beta_j = \mu_\beta + \zeta_j$, the dual problem is:

$$\begin{aligned}
-\min_{\Gamma, \mathcal{E}_0, \mathcal{E}_1, \ldots, \mathcal{E}_J} L(\cdot) = &- \sum_{j=1}^{J} \sum_{i \in \mathcal{D}_j} \min_{\gamma_{ij}} \left\{ f(\gamma_{ij}) - \left(\alpha_j - \sum_{\ell=1}^{L} \beta_{\ell j} Y_{i,T_j-\ell} \right) \gamma_{ij} \right\} + \sum_{j=1}^{J} \alpha_j + \sum_{\ell=1}^{L} \beta_{\ell j} Y_{j,T_j-\ell} \\
&- \sum_{\ell=1}^{L} \min_{\mathcal{E}_{j\ell}} \left\{ \frac{1-\nu}{2JL\lambda} \mathcal{E}_{j\ell}^2 - \mathcal{E}_{j\ell} (\beta_{\ell j} - \mu_{\beta\ell}) \right\} \\
&- \sum_{\ell=1}^{L} \min_{\mathcal{E}_{0\ell}} \left\{ \frac{\nu}{2L\lambda} \mathcal{E}_{0\ell}^2 - \mathcal{E}_{0\ell} \mu_{\beta\ell} \right\}
\end{aligned} \tag{B.13}$$

From Lemma B.1, we see that the first term in (B.13) is $\mathcal{L}(\alpha, \beta)$ and we have the same form for the implied weights. The next two terms are the convex conjugates of a scaled L^2 norm. Using the computation that the convex conjugate of $\frac{a}{2}\|x\|_2^2$ is $\frac{1}{2a}\|x\|_2^2$. Finally, the primal problem (3.5) is still convex and a primal feasible point exists, so by Slater's condition strong duality holds. □

Duality when balancing a general number of lagged outcomes

When balancing a different number of lagged outcomes for each treated unit, the lagrangian dual problem only changes slightly. First, we immediately see that the dual to separate SCM (3.2) is nearly identical to Equation (B.9):

$$\min_{\alpha,\beta} \quad \mathcal{L}(\alpha, \beta) + \sum_{j=1}^{J} \frac{\lambda(T_j - 1)}{2} \|\beta_j\|_2^2 = \min_{\alpha,\beta} \quad \mathcal{L}(\alpha, \beta) + \sum_{j=1}^{J} \sum_{\ell=1}^{T_j-1} \frac{\lambda(T_j - 1)}{2} \beta_{\ell j}^2 \tag{B.14}$$

An important distinction is that for each treatment level j there is a different number of lagged outcomes T_j; therefore the dual parameters β_j have different lengths. Scaling the ridge penalty by T_j ensures that the dual parameters are shrunk equally towards zero, and balancing a different number of lagged outcomes for each treatment level changes the implied selection model.

As before, the average balance term induces an additional set of Lagrange multipliers $\mu_\beta \in \mathbb{R}^{T_J}$, where the varying-length β_j's are pooled towards μ_β for each $1 \leq \ell \leq T_j - 1$.

$$\min_{\alpha,\beta,\mu_\beta} \mathcal{L}(\alpha,\beta) + \sum_{j=1}^{J} \frac{\lambda(T_j-1)}{2(1-\nu)} \sum_{\ell=1}^{T_j-1} (\beta_{\ell j} - \mu_{\beta_\ell})^2 + \frac{\lambda L}{2\nu}\|\mu_\beta\|_2^2. \tag{B.15}$$

Error bounds

Time-varying AR

Theorem B.1. Let $Y_{it}(0)$ follow a time-varying AR(L) process

$$Y_{it}(0) = \sum_{\ell=1}^{L} (\bar{\rho}_\ell + \xi_{t\ell}) \rho_{t\ell} Y_{i,t-\ell} + \varepsilon_{it}, \tag{B.16}$$

where ε_{it} are independent sub-Gaussian random variables with scale parameter σ, and $S^2 \equiv \frac{1}{J}\sum_{j=1}^{J}\sum_{\ell=1}^{L} \xi_{T_j\ell}^2$. For $\hat{\gamma}_1, \ldots, \hat{\gamma}_J \in \Delta^{\text{scm}}$, the error at time T_j for treated unit j is

$$|\hat{\tau}_{j0} - \tau_{j0}| \leq \left\|\bar{\rho} + \xi_{T_j}\right\|_2 \sqrt{\sum_{\ell=1}^{L} \left(Y_{j,T_j-\ell} - \sum_{i\in\mathcal{D}_j} \hat{\gamma}_{ij} Y_{i,T_j-\ell} \right)^2} + \delta\sigma\left(1 + \|\hat{\gamma}_j\|_2\right) \tag{B.17}$$

with probability at least $1 - 2e^{-\frac{\delta^2}{2}}$. Furthermore, the error for $\widehat{\text{ATT}}_0$ is

$$\begin{aligned} \widehat{\text{ATT}}_0 - \text{ATT}_0 = \frac{1}{J}\sum_{j=1}^{J} \hat{\tau}_{j0} - \tau_{j0} = \|\bar{\rho}\|_2 & \underbrace{\sqrt{\sum_{\ell=1}^{L} \left(\frac{1}{J}\sum_{j=1}^{J} Y_{j,T_j-\ell} - \sum_{i\in\mathcal{D}_j} \hat{\gamma}_{ij} Y_{i,T_j-\ell} \right)^2}}_{\text{global fit}} \\ & + S \underbrace{\sqrt{\frac{1}{J}\sum_{j=1}^{J}\sum_{\ell=1}^{L} \left(Y_{j,T_j-\ell} - \sum_{i\in\mathcal{D}_j} \hat{\gamma}_{ij} Y_{i,T_j-\ell} \right)^2}}_{\text{individual fit}} \\ & + \frac{\delta\sigma}{\sqrt{J}}\left(1 + \|\Gamma\|_F\right) \end{aligned} \tag{B.18}$$

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

Proof of Theorem B.1. Notice that

$$\hat{\tau}_{j0} - \tau_{j0} = \sum_{\ell=1}^{L} (\bar{\rho}_\ell + \xi_{T_j \ell}) \left(Y_{j,T_j-\ell} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} Y_{i,T_j-\ell} \right) + \left(\varepsilon_{jT_j} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} \varepsilon_{iT_j} \right)$$

So by the triangle and Cauchy-Schwarz inequalities,

$$|\hat{\tau}_{j0} - \tau_{j0}| \leq \|\bar{\rho} + \xi_{T_j}\|_2 \sqrt{\sum_{\ell=1}^{L} \left(Y_{j,T_j-\ell} - \sum_{i \in \mathcal{D}_j} \gamma_{ij} Y_{i,T_j-\ell} \right)^2} + \left| \varepsilon_{jT_j} - \sum_{i \in \mathcal{D}_j} \gamma_{ij} \varepsilon_{iT_j} \right|$$

Since $\hat{\gamma}_j$ is fit on pre-T_j outcomes, the weights are not dependent on ε_{T_j}, and so the second term above is sub-Gaussian with scale parameter $\sigma\sqrt{1 + \|\hat{\gamma}_j\|_2^2} \leq \sigma(1 + \|\hat{\gamma}_j\|_2)$. This implies that

$$P\left(\left| \varepsilon_{jT_j} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} \varepsilon_{iT_j} \right| \geq \delta\sigma \left(1 + \|\hat{\gamma}_j\|_2\right) \right) \leq 2 \exp\left(-\frac{\delta^2}{2} \right)$$

This completes the proof of the first inequality. For the bound on $\widehat{\text{ATT}}_0$, notice that

$$\begin{aligned}
\widehat{\text{ATT}}_0 - \text{ATT}_0 = \frac{1}{J} \sum_{j=1}^{J} \hat{\tau}_{j0} - \tau_{j0} &= \frac{1}{J} \sum_{j=1}^{J} \left[\sum_{\ell=1}^{L} (\bar{\rho}_\ell + \xi_{T_j \ell}) \left(Y_{j,T_j-\ell} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} Y_{i,T_j-\ell} \right) + \left(\varepsilon_{jT_j} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} \varepsilon_{iT_j} \right) \right. \\
&= \sum_{\ell=1}^{L} \bar{\rho}_\ell \frac{1}{J} \sum_{j=1}^{J} \left(Y_{j,T_j-\ell} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} Y_{i,T_j-\ell} \right) \\
&\quad + \frac{1}{J} \sum_{j=1}^{J} \sum_{\ell=1}^{L} \xi_{T_j \ell} \left(Y_{j,T_j-\ell} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} Y_{i,T_j-\ell} \right) \\
&\quad + \frac{1}{J} \sum_{j=1}^{J} \left(\varepsilon_{jT_j} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} \varepsilon_{iT_j} \right)
\end{aligned} \tag{B.19}$$

By Cauchy-Schwarz the absolute value of the first term is

$$\left| \sum_{\ell=1}^{L} \bar{\rho}_\ell \frac{1}{J} \sum_{j=1}^{J} \left(Y_{j,T_j-\ell} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} Y_{i,T_j-\ell} \right) \right| \leq \|\bar{\rho}\|_2 \sqrt{\sum_{\ell=1}^{L} \left(\frac{1}{J} \sum_{j=1}^{J} \left[Y_{j,T_j-\ell} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} Y_{i,T_j-\ell} \right] \right)^2}.$$

Similarly, the absolute value of the second term is

$$\left|\frac{1}{J}\sum_{j=1}^{J}\sum_{\ell=1}^{L}\xi_{T_j\ell}\left(Y_{j,T_j-\ell}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{i,T_j-\ell}\right)\right| \leq \frac{1}{J}\sum_{j=1}^{J}\|\xi_{T_j}\|_2\sqrt{\sum_{\ell=1}^{L}\left(Y_{j,T_j-\ell}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{i,T_j-\ell}\right)^2}$$
$$\leq S\sqrt{\frac{1}{J}\sum_{j=1}^{J}\sum_{\ell=1}^{L}\left(Y_{j,T_j-\ell}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{i,T_j-\ell}\right)^2}$$

Finally, notice that $\frac{1}{J}\sum_{j=1}^{J}\varepsilon_{jT_j}$ is the average of J independent sub-Gaussian random variables and so is itself sub-Gaussian with scale parameter $\frac{\sigma}{\sqrt{J}}$. However, $\frac{1}{J}\sum_{j=1}^{J}\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}\varepsilon_{iT_j}$ is the weighted average of sub-Gaussian variables that are independent over i but not necessarily independent over j, and so the weighted average is sub-Gaussian with scale parameter $\frac{\sigma}{\sqrt{J}}\|\Gamma\|_F$. The two averages are independent of each other, so

$$P\left(\frac{1}{J}\sum_{j=1}^{J}\left(\varepsilon_{jT_j}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}\varepsilon_{iT_j}\right)\geq\frac{\delta\sigma}{\sqrt{J}}\left(1+\|\hat{\Gamma}\|_F\right)\right)\leq 2\exp\left(-\frac{\delta^2}{2}\right)$$

Putting together the pieces completes the proof. □

Linear factor model

We begin by bounding the error for the treatment effect for a single treated unit. This re-writes Theorem 2.1 with the notation in Chapter 3. Then we prove Theorem 3.1.

Proposition B.1. Let $Y_{it}(0)$ follow a linear factor model

$$Y_{it}(0)=\mu_t'\phi_i+\varepsilon_{it}, \tag{B.20}$$

where $\mu_t,\phi_i\in\mathbb{R}^F$ with $\max_t\|\mu_t\|_\infty\leq M$, and ε_{it} is independent sub-Gaussian with scale parameter σ. For $\hat{\gamma}_1,\ldots,\hat{\gamma}_J\in\Delta^{\text{scm}}$, the error at time T_j for treated unit j is

$$|\hat{\tau}_{j0}-\tau_{j0}|\leq\frac{M^2F}{\sqrt{L}}\sqrt{\sum_{t=1}^{L}\left(Y_{jt}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{it}\right)^2}+\sigma\frac{M^2F}{\sqrt{L}}\left(2\delta+\sqrt{\log\|\mathcal{D}_j\|}\right)+\delta\sigma\left(1+\|\hat{\gamma}_j\|_2\right) \tag{B.21}$$

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

Proof of Proposition B.1. First notice that

$$\hat{\tau}_{j0}-\tau_{j0}=\mu_{T_j}'\left(\phi_j-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}\phi_i\right)+\left(\varepsilon_{jt}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}\varepsilon_{iT_j}\right)$$

From the proof of Theorem B.1, we know that

$$P\left(\left|\varepsilon_{jT_j} - \sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}\varepsilon_{iT_j}\right| \geq \delta\sigma\left(1 + \|\hat{\gamma}_j\|_2\right)\right) \leq 2\exp\left(-\frac{\delta^2}{2}\right).$$

Following Abadie et al. (2010), we can re-write ϕ_i in terms of the lagged outcomes as

$$\phi_i = (\mu'_{1:L}\mu_{1:L})^{-1}\sum_{t=1}^{L}\mu_t(Y_{it} - \varepsilon_{it}) = \frac{1}{L}\sum_{t=1}^{L}\mu_t(Y_{it} - \varepsilon_{it}) \tag{B.22}$$

where $\mu_{1:L} \in \mathbb{R}^{L\times F}$ is the matrix of factors from time $t = 1, \ldots, L$. With the Cauchy-Schwarz inequality, this implies that

$$\begin{aligned}\left|\mu'_{T_j}\left(\phi_j - \sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}\phi_i\right)\right| &= \frac{1}{L}\sum_{t=1}^{L}\mu'_{T_j}\mu_t\left(Y_{jt} - \sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{it}\right) - \frac{1}{L}\sum_{t=1}^{L}\mu'_{T_j}\mu_t\left(\varepsilon_{jt} - \sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}\varepsilon_{it}\right)\\ &\leq \frac{1}{L}\|\mu'_{T_j}\mu_{1:L}\|_2\sqrt{\sum_{t=1}^{L}\left(Y_{jt} - \sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{it}\right)^2} + \frac{1}{L}\left|\sum_{t=1}^{L}\mu'_{T_j}\mu_t\varepsilon_{jt}\right| + \max_{i\in\mathcal{D}_j}\left|\sum_{t=1}^{L}\mu'_{T_j}\mu_t\varepsilon_{it}\right|\end{aligned}$$

Next, since ε_{jT_j} are independent sub-Gaussian,

$$P\left(\frac{1}{L}\left|\sum_{t=1}^{L}\mu'_{T_j}\mu_t\varepsilon_{jt}\right| \geq \frac{\delta\sigma}{L}\|\mu'_{T_j}\mu_{1:L}\|_2\right) \leq 2\exp\left(-\frac{\delta^2}{2}\right)$$

and by the standard tail bound on maxima of sub-Guassian random variables,

$$P\left(\frac{1}{L}\max_{i\in\mathcal{D}_j}\left|\sum_{t=1}^{L}\mu'_{T_j}\mu_t\varepsilon_{it}\right| \geq \frac{\sigma}{L}\|\mu'_{T_j}\mu_{1:L}\|_2\left(\sqrt{\log\|\mathcal{D}_j\|} + \delta)\right)\right) \leq 2\exp(-\frac{\delta^2}{2})$$

Now notice that $\frac{1}{L}\|\mu'_{T_j}\mu_{1:L}\|_2 \leq \frac{M^2F}{\sqrt{L}}$. Combining these bounds gives the result. □

Proof of Theorem 3.1. Using Equation (B.22), we can write the error for the ATT as

$$\begin{aligned}\widehat{\mathrm{ATT}}_k - \mathrm{ATT}_k = \frac{1}{J}\sum_{j=1}^{J}\hat{\tau}_{jk} - \tau_{jk} &= \frac{1}{JL}\sum_{j=1}^{J}\sum_{t=1}^{L}\mu'_{T_j+k}\mu_t\left(Y_{jt} - \sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{it}\right)\\ &\quad - \frac{1}{JL}\sum_{j=1}^{J}\sum_{t=1}^{L}\mu'_{T_j+k}\mu_t\left(\varepsilon_{jt} - \sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}\varepsilon_{it}\right)\\ &\quad + \frac{1}{J}\sum_{j=1}^{J}\left(\varepsilon_{jt} - \sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}\varepsilon_{iT_j}\right).\end{aligned} \tag{B.23}$$

From the proof of Theorem B.1, we can bound the final term in Equation (B.23). We now bound the first two terms. First, we decompose the first term into a time constant, and a time varying component:

$$\underbrace{\frac{1}{JL}\sum_{j=1}^{J}\sum_{t=1}^{L}\mu'_{T_j+k}\mu_t\left(Y_{jt}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{it}\right)}_{(*)}=\frac{1}{JL}\bar{\mu}'_k\sum_{t=1}^{L}\mu_t\sum_{j=1}^{J}\left(Y_{jt}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{it}\right)$$
$$+\frac{1}{JL}\sum_{j=1}^{J}\sum_{t=1}^{L}\xi'_{T_j+k}\mu_t\left(Y_{jt}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{it}\right),$$

where $\xi_{T_j+k}\equiv\mu_{T_j+k}-\bar{\mu}_k$. Now by Cauchy-Schwarz and using that $\frac{1}{L}\|\mu_{1:L}\|_2\leq\frac{M\sqrt{F}}{\sqrt{L}}$ we get that

$$|(*)|\leq M\sqrt{\frac{F}{L}}\|\bar{\mu}_k\|_2\sqrt{\sum_{t=1}^{L}\left(\frac{1}{J}\sum_{j=1}^{J}Y_{jt}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{it}\right)^2}+\frac{M}{J}\sqrt{\frac{F}{L}}\sum_{j=1}^{J}\|\xi_{T_j+k}\|_2\sqrt{\sum_{t=1}^{L}\left(Y_{jt}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{it}\right)^2}$$
$$\leq M\sqrt{\frac{F}{L}}\|\bar{\mu}_k\|_2\sqrt{\sum_{t=1}^{L}\left(\frac{1}{J}\sum_{j=1}^{J}Y_{jt}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{it}\right)^2}+M\sqrt{\frac{F}{L}}S_k\sqrt{\frac{1}{J}\sum_{j=1}^{J}\sum_{t=1}^{L}\left(Y_{jt}-\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}Y_{it}\right)^2}$$

We now turn to the second term in Equation (B.23). Since ε_{it} are independent sub-Gaussian random variables and $\frac{1}{L}\|\mu'_{T_j+k}\mu_{1:L}\|_2\leq\frac{M^2F}{\sqrt{L}}$,

$$P\left(\frac{1}{L}\left|\frac{1}{J}\sum_{j=1}^{J}\sum_{t=1}^{L}\mu'_{T_j+k}\mu_t\varepsilon_{jt}\right|\geq\frac{\delta\sigma M^2F}{\sqrt{JL}}\right)\leq 2\exp\left(-\frac{\delta^2}{2}\right)$$

Next, since $\hat{\gamma}_1,\ldots,\hat{\gamma}_J\in\Delta^{\text{scm}}$, $\frac{1}{J}\sum_{j=1}^{J}\|\hat{\gamma}_j\|_1=1$, so by Hölder's inequality

$$\left|\frac{1}{J}\sum_{j=1}^{J}\sum_{t=1}^{L}\mu'_{T_j+k}\mu_t\sum_{i\in\mathcal{D}_j}\hat{\gamma}_{ij}\varepsilon_{it}\right|\leq\max_{j\in\{1,\ldots,J\},i\in\mathcal{D}_j}\left|\sum_{t=1}^{L}\mu'_{T_j+k}\mu_t\varepsilon_{it}\right|\leq 2\frac{\sigma M^2F}{\sqrt{L}}\left(\sqrt{\log NJ}+\delta\right)$$

where the final inequality holds with probability at least $1-2\exp\left(-\frac{\delta^2}{2}\right)$ by the standard tail bound on the maximum of sub-Gaussian random variables. Putting together the pieces with a union bound gives that

$$\left|\widehat{\mathrm{ATT}}_k - \mathrm{ATT}_k\right| \leq \frac{M\sqrt{F}}{\sqrt{L}} \left(\|\bar{\mu}_k\|_2 \sqrt{\sum_{t=1}^{L} \left(\frac{1}{J} \sum_{j=1}^{J} Y_{jt} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} Y_{it} \right)^2} + S_k \sqrt{\frac{1}{J} \sum_{j=1}^{J} \sum_{t=1}^{L} \left(Y_{jt} - \sum_{i \in \mathcal{D}_j} \hat{\gamma}_{ij} Y_{it} \right)^2} \right)$$
$$+ \frac{\sigma M^2 F}{\sqrt{L}} \left(\left(2 + \frac{1}{\sqrt{J}} \right) \delta + 2\sqrt{\log NJ} \right)$$
$$+ \frac{\delta \sigma}{\sqrt{J}} \left(1 + \|\hat{\Gamma}\|_F \right)$$

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

Appendix C

Supplementary materials for Chapter 4

C.1 Within-subject comparison

We compare the weighting estimates for the effect of submitting an LOR on the second reader scores to estimates exploiting an additional feature of the pilot study. After the admissions process concluded, 10,000 applicants who submitted letters were randomly sampled and the admissions office recruited several readers to conduct additional evaluations of the applicants (Rothstein, 2017). During this supplemental review cycle, the readers were *not* given access to the letters of recommendation, but otherwise the evaluations were designed to be as similar as possible to the second reads that were part of the regular admissions cycle; in particular, readers had access to the first readers' scores.

With these third reads we can estimate the treatment effect by taking the average difference between the second read (with the letters) and the third read (without the letters). One major issue with this design is that readers might have applied different standards during the supplemental review cycle. Regardless, if the third readers applied a different standard consistently across URM and admissibility status, we can distinguish between treatment effects within these subgroups. We show the results in Figures C.9 and C.10.

C.2 Proofs

Proof of Proposition 4.1. First, we will augment the primal optimization problem in Equation (4.13) with auxiliary covariates $\mathcal{E}_1, \ldots, \mathcal{E}_j$ so that $\mathcal{E}_g = \sum_{G_i=g, W_i=0} \gamma_i \phi(X_i) - \sum_{G_i=g, W_i=1} \phi(X_i)$.

Then the optimization problem becomes:

$$\begin{aligned} \min_{\gamma} \quad & \sum_{z=1}^{J} \frac{1}{2\lambda_g} \|\mathcal{E}_j\|_2^2 + \frac{\lambda_g}{2} \sum_{Z_i=z, W_i=0} \gamma_i^2 + \mathcal{I}(\gamma_i \geq 0) \\ \text{subject to } & \sum_{W_i=0} \gamma_i \phi(X_i) = \sum_{W_i=1} \phi(X_i) \\ & \mathcal{E}_j = \sum_{G_i=g, W_i=0} \gamma_i \phi(X_i) - \sum_{G_i=g, W_i=1} \phi(X_i), \quad z = 1, \ldots, J \\ & \sum_{G_i=g, W_i=0} \gamma_i = n_{1g}, \end{aligned} \tag{C.1}$$

where $\mathcal{I}(x \geq 0) = \begin{cases} 0 & x \geq 0 \\ \infty & x < 0 \end{cases}$ is the indicator function. The first constraint induces a Lagrange multiplier μ_β, the next J constraints induce Lagrange multipliers $\delta_1, \ldots, \delta_J$, and the sum-to-one constraints induce Lagrange multipliers $\alpha_1, \ldots, \alpha_J$. Then the Lagrangian is

$$\begin{aligned} \mathcal{L}(\gamma, \mathcal{E}, \mu_\beta, \delta, \alpha) = & \sum_{z=1}^{J} \left[\frac{1}{2\lambda_g} \|\mathcal{E}_j\|_2^2 - \mathcal{E}_j \cdot \delta_j + \sum_{G_i=g, W_i=0} \frac{1}{2}\gamma_i^2 + \mathcal{I}(\gamma_i \geq 0) - \gamma_i(\alpha + (\mu_\beta + \delta_j) \cdot \phi(X_i)) \right] \\ & + \sum_{z=1}^{J} \sum_{G_i=g, W_i=1} (1 + (\mu_\beta + \delta_j) \cdot \phi(X_i)) \end{aligned} \tag{C.2}$$

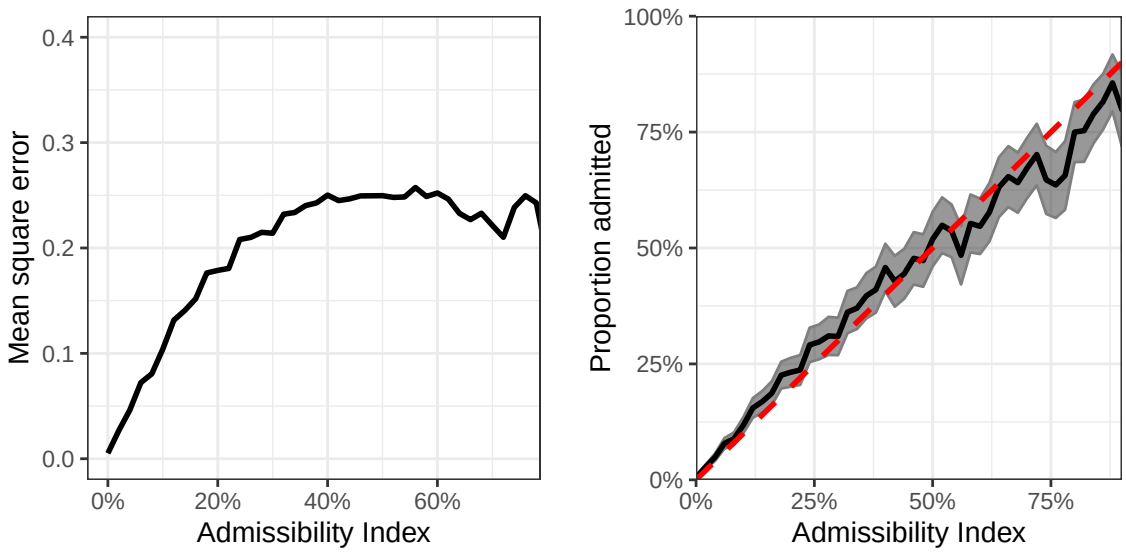
The dual objective is:

$$\begin{aligned} q(\mu_\beta, \delta, \alpha) = & \sum_{z=1}^{J} \left[\min_{\mathcal{E}_j} \left\{ \frac{1}{2\lambda_g} \|\mathcal{E}_j\|_2^2 - \mathcal{E}_j \cdot \delta_j \right\} + \sum_{G_i=g, W_i=0} \min_{\gamma_i \geq 0} \left\{ \frac{1}{2}\gamma_i^2 - \gamma_i(\alpha + (\mu_\beta + \delta_j) \cdot \phi(X_i)) \right\} \right] \\ & + \sum_{z=1}^{J} \sum_{G_i=g, W_i=1} (1 + (\mu_\beta + \delta_j) \cdot \phi(X_i)) \end{aligned} \tag{C.3}$$

Note that the inner minimization terms are the negative convex conjugates of $\frac{1}{2}\|x\|_2^2$ and $\frac{1}{2}x^2 + \mathcal{I}(X \geq 0)$, respectively. Solving these inner optimization problems yields that

$$\begin{aligned} q(\mu_\beta, \delta, \alpha) = & -\sum_{z=1}^{J} \left[\frac{\lambda_g}{2} \|\delta_j\|_2^2 + \sum_{G_i=g, W_i=0} [\alpha_j + (\mu_\beta + \delta_j) \cdot \phi(X_i)]_+^2 \right] \\ & + \sum_{z=1}^{J} \sum_{G_i=g, W_i=1} (1 + (\mu_\beta + \delta_j) \cdot \phi(X_i)) \end{aligned} \tag{C.4}$$

Now since there exists a feasible solution to the primal problem (4.13), from Slater's condition we see that the solution to the primal problem is equivalent to the solution to



(a) Brier score.

(b) Admission rates

Figure C.1: (a) Mean square error (Brier score) and (b) admission rates for the Admissibility Index predicting the 2016-2017 cycle admissions results, computed in 2% groups.

College	URM	AUC	Brier Score
Letters and Science	URM	89%	9%
	Not URM	88%	11%
Engineering	URM	92%	5%
	Not URM	89%	11%

Table C.1: AUC and Brier score for the Admissibility Index predicting the 2016-2017 cycle admissions results.

$\max_{\mu_\beta,\alpha,\delta} q(\mu_\beta, \alpha, \delta)$. Defining $\beta_j \equiv \mu_\beta + \delta_j$ gives the dual problem (4.18). Finally, note that the solution to the minimization over the weights in Equation (C.3) is $\gamma_i = [\alpha_j + \beta_j \cdot \phi(X_i)]_+$, which shows how to map from the dual solution to the primal solution. □

C.3 Additional figures and tables

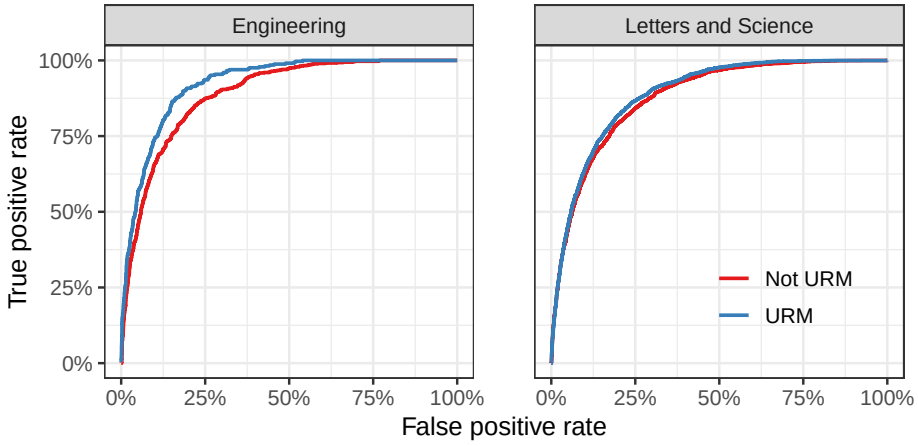


Figure C.2: ROC curve for Admissibility Index predicting the 2016-2017 cycle admissions results.

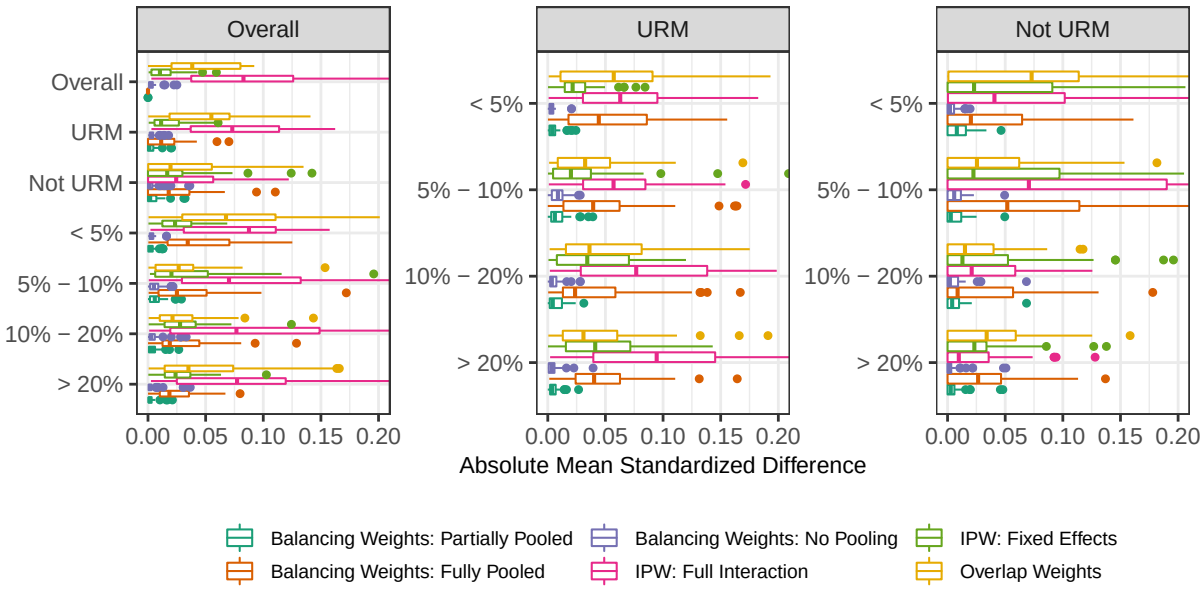
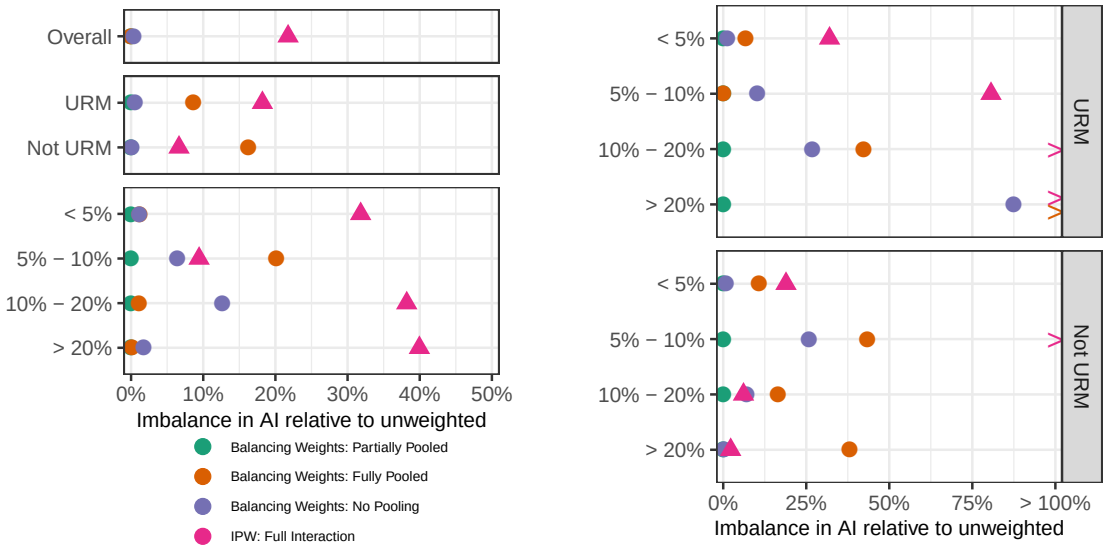


Figure C.3: Distribution of covariate balance measured by the mean standardized difference for different weighting methods.



(a) Overall and by URM status and AI. (b) By URM status interacted with AI.

Figure C.4: Imbalance in the admissibility index after weighting relative to before weighting, overall and within each subgroup. For several subgroups, the fully pooled balancing weights procedure results in *increased* imbalance in the admissibility index, denoted by an arrow.

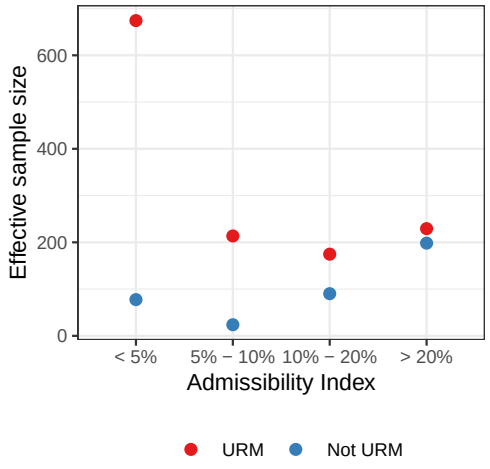
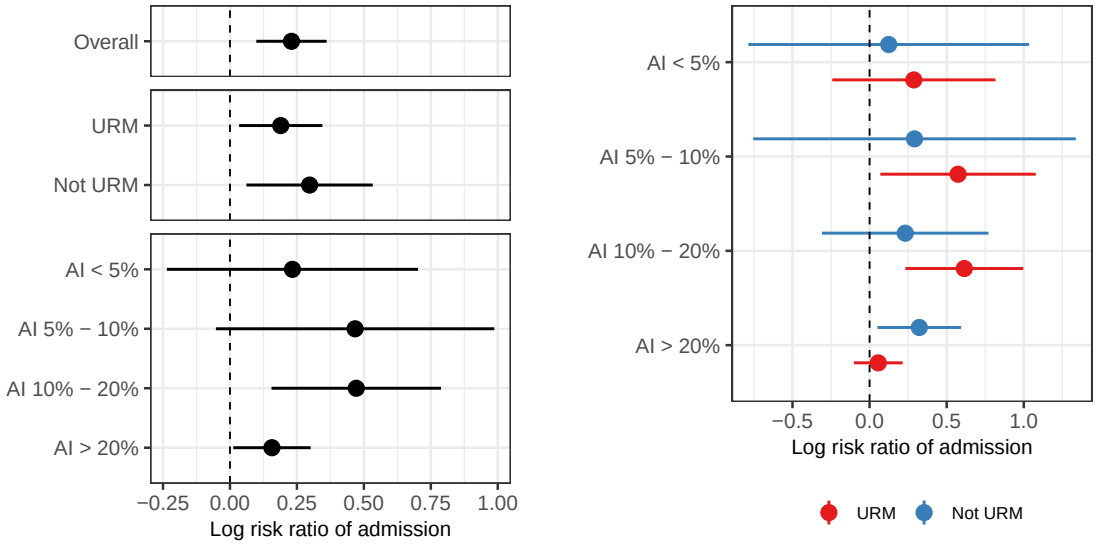
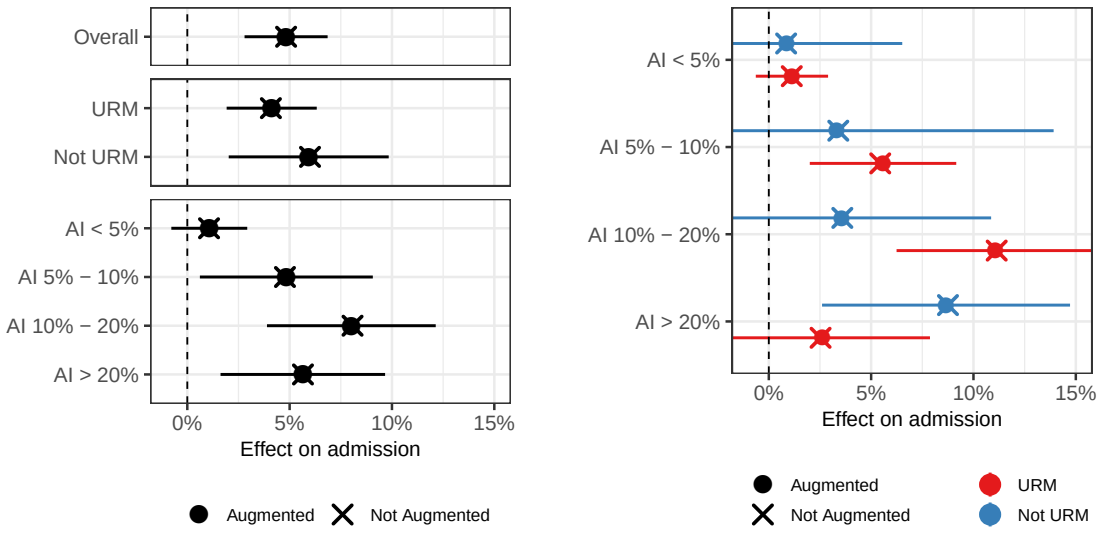


Figure C.5: Effective sample size for each subgroup, with weights solving the approximate balancing weights problem (4.13).



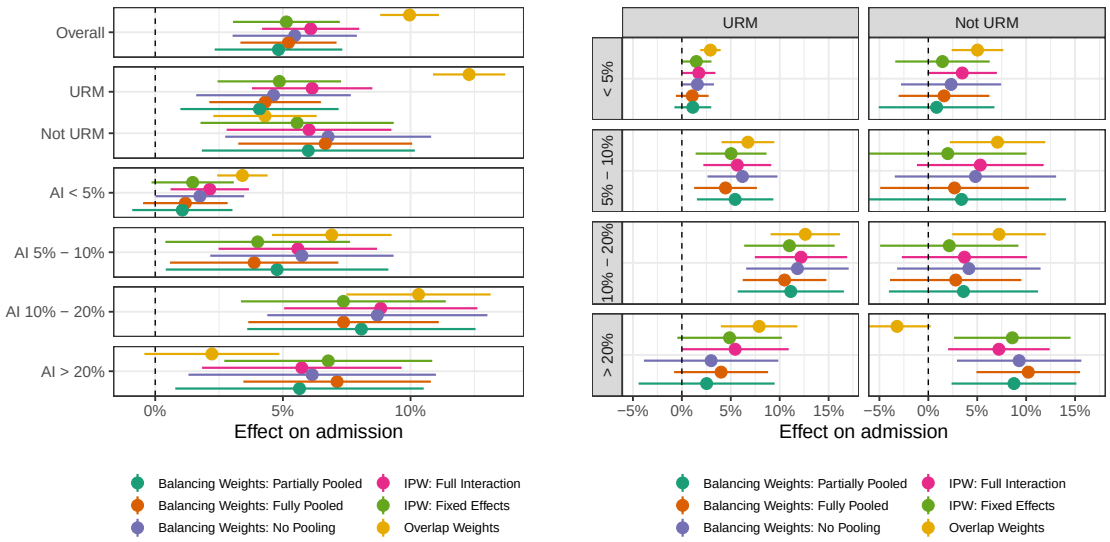
(a) Overall and by URM status and AI. (b) By URM status interacted with AI.

Figure C.6: Estimated log risk ratio of admission with and without letters of recommendation ± two standard errors computed via the delta method, overall and by URM status and AI.



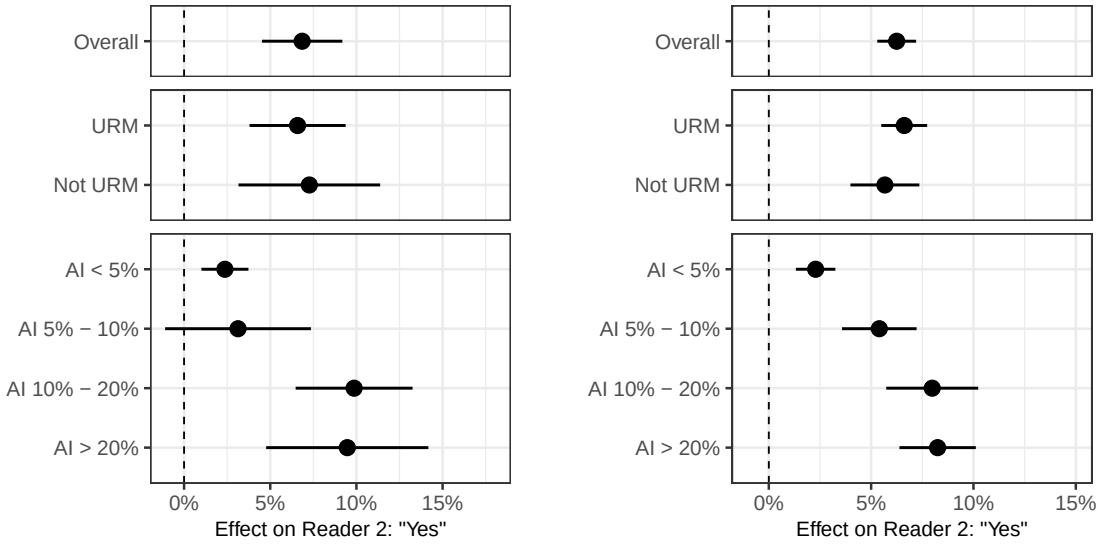
(a) Overall and by URM status and AI. (b) By URM status interacted with AI.

Figure C.7: Estimated effect of letters of recommendation on admission rates with and without augmentation via ridge regression.



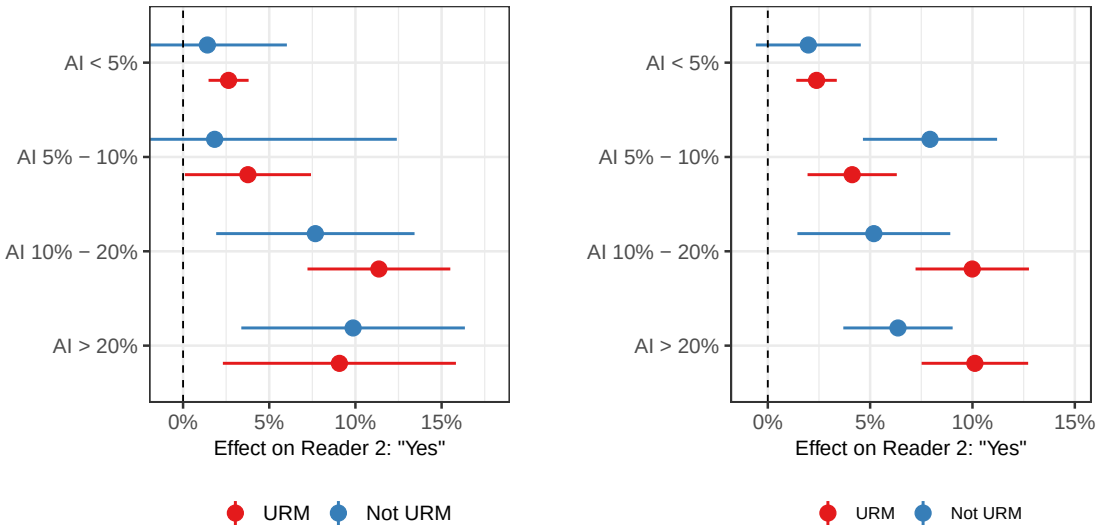
(a) Overall and by URM status and AI. (b) By URM status interacted with AI.

Figure C.8: Estimated effect of letters of recommendation on admission rates for comparable weighting estimators.



(a) Partially pooled balancing weights (b) Within-subect design

Figure C.9: Effects on second reader scores overall, by URM status, and by AI, estimated via (a) the partially pooled balancing weights estimator and (b) the within-subject design.



(a) Partially pooled balancing weights (b) Within-subject design

Figure C.10: Effects on second reader scores by URM status interacted with AI, estimated via (a) the partially pooled balancing weights estimator and (b) the within-subject design.

C.4 Additional simulation results

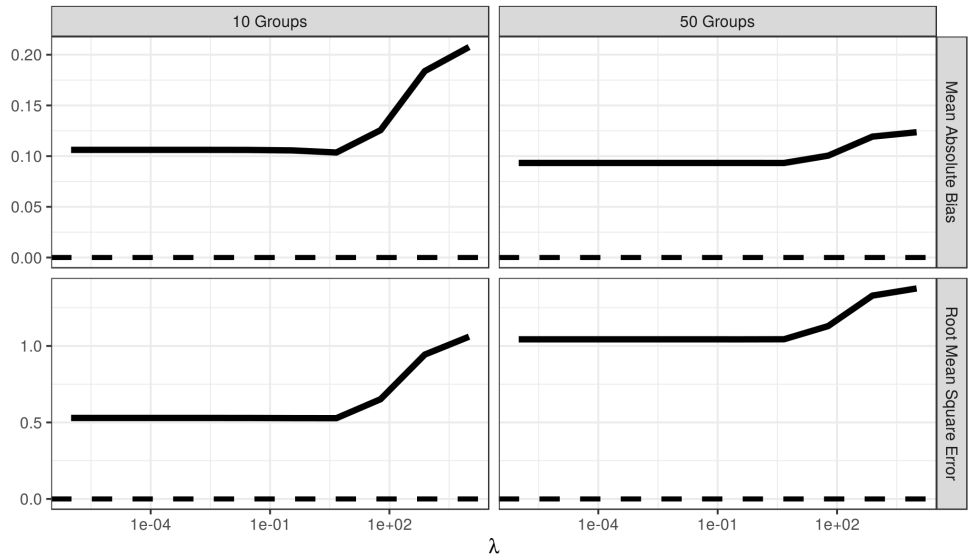


Figure C.11: Performance of approximate balancing weights for estimating subgroup treatment effects as λ varies.