Supplementary Materials for
"The Augmented Synthetic Control Method"

October 2020

A Inference

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

The approximate validity of the conformal inference procedure in Section 5.4 depends on the predictive accuracy of $\hat{Y}_{it}^{\text{aug}}(0)$ when fit using all periods $t = 1, \ldots, T$, including the post-treatment period T. Denoting $\mathbf{Y_1}. \equiv (\mathbf{X}_{1\cdot}, Y_1) \in \mathbb{R}^T$ 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

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

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 (3) in an asymptotic framework where T_0 is growing while T is fixed. We state the model and assumptions for asymptotically (in T_0) valid inference below.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

B Additional results

B.1 Specialization of Ridge ASCM results to SCM

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

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

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

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

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

We can similarly specialize Theorem 1.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

B.3 Error under a linear factor model with covariates

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\underbrace{\frac{2JM^2\sigma}{\sqrt{T_0}} \left(\sqrt{\log N_0} + \frac{\delta}{2} \right)}_{\text{SCM approximation error}} \qquad + \underbrace{\delta\sigma(1 + \|\hat{\gamma}^{\text{cov}}\|_2)}_{\text{post-treatment noise}}$$

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

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

C Simulation data generating process

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

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

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

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

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

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

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

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

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

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

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

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

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

D Proofs

D.1 Proofs for Section 4

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

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

where

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

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

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

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

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

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

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

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

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

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

Since the lagged outcomes are centered, this implies that

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

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

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

This implies that

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

Finally, the weights are the ridge weights

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

Similarly, the Lagrangian dual to (18) is

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

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

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

Finally, the weights are the ridge ASCM weights

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

□

Proof of Lemma 3. Notice that

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

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

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

□

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

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

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

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

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

□

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

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

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

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

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

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

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

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

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

Proof. First, note that the estimation error is

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

□

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

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

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

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

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

$$Y_{1T}(0) - \sum_{W_i=0} \hat{\gamma}_i Y_{iT} \leq \|\boldsymbol{\beta}\|_2 \underbrace{\left\|\text{diag}\left(\frac{\lambda}{d_j^2+\lambda}\right)\left(\widetilde{\boldsymbol{X}}_1 - \sum_{W_i=0}\hat{\gamma}_i^{\text{scm}}\widetilde{\boldsymbol{X}}_i\right)\right\|_2}_{\text{imbalance in } X} + \underbrace{2\varpi\left(\sqrt{\log 2N_0}+\frac{\delta}{2}\right)}_{\text{approximation error}}$$

$$\underbrace{(1+\delta)4\varpi\left\|\text{diag}\left(\frac{d_j}{d_j^2+\lambda}\right)\left(\widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_{0\cdot}'\hat{\gamma}^{\text{scm}}\right)\right\|_2}_{\text{excess approximation error}} + \underbrace{\delta\sigma\left(1+\|\hat{\gamma}\|_2\right)}_{\text{post-treatment noise}}, \tag{A.29}$$

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

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

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

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

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

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

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

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

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

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

Then we see that the post treatment outcomes are

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

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

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

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

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

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

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

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 $\hat{\boldsymbol{\gamma}}^{\text{aug}}$ (17) with hyperparameter $\lambda^{\text{ridge}} = N_0\lambda$ satisfy

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

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

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

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

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

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. □

D.3 Proofs for Sections 6 and B.3

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

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

Now notice that

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

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

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

The pre-treatment fit is

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

This gives the bound on the pre-treatment fit.

□

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

Then the error is

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

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

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

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

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

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

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

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

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 $\sum_{W_i=0} \sum_{t=1}^{T_0} \boldsymbol{\mu}_T' \boldsymbol{\mu}_{t\cdot} Z_{ik} \varepsilon_{it}$ 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 4 and putting together the pieces with the union bound gives the result. □

E Connection to balancing weights and IPW

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

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

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

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

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

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

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

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, 2018). This more general formulation also suggests natural extensions of both SCM and ASCM beyond the L^2 setting to other forms, especially L^1 regularization.

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

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

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

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

The Lagrangian is

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

The dual maximizes the objective

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

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

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

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

□

F Additional figures

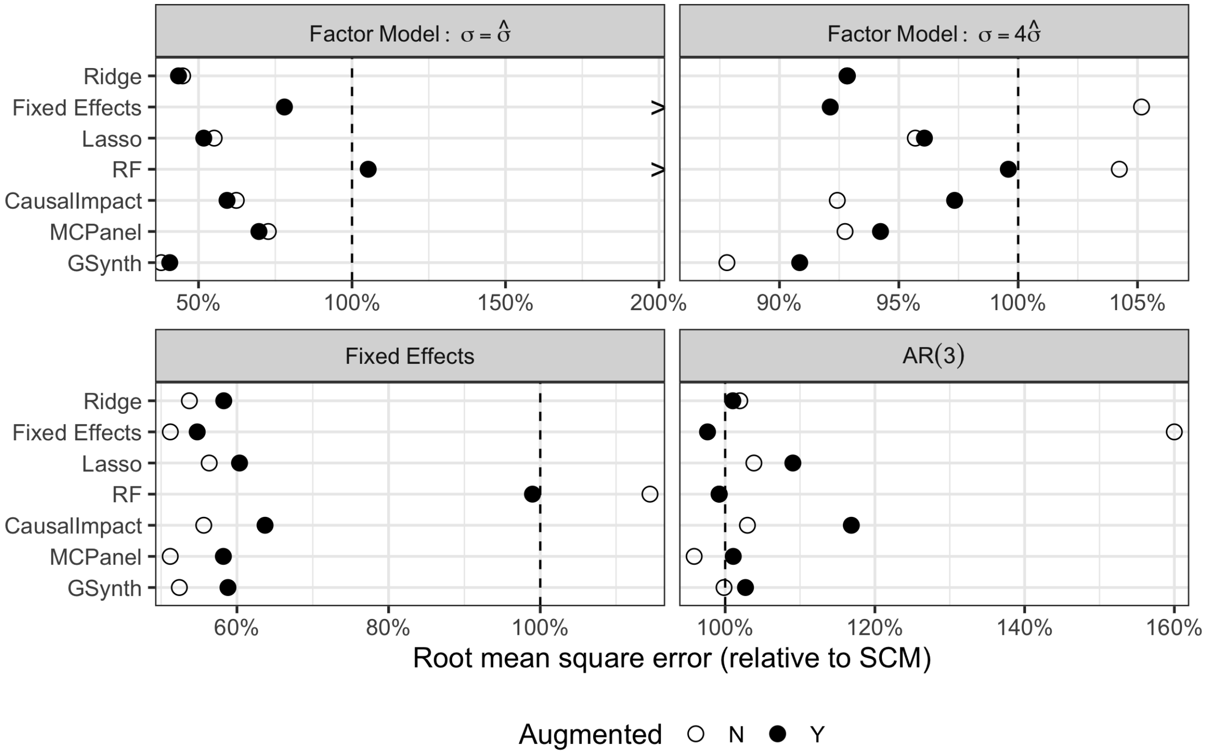

Figure F.1: RMSE for different augmented and non-augmented estimators across outcome models.

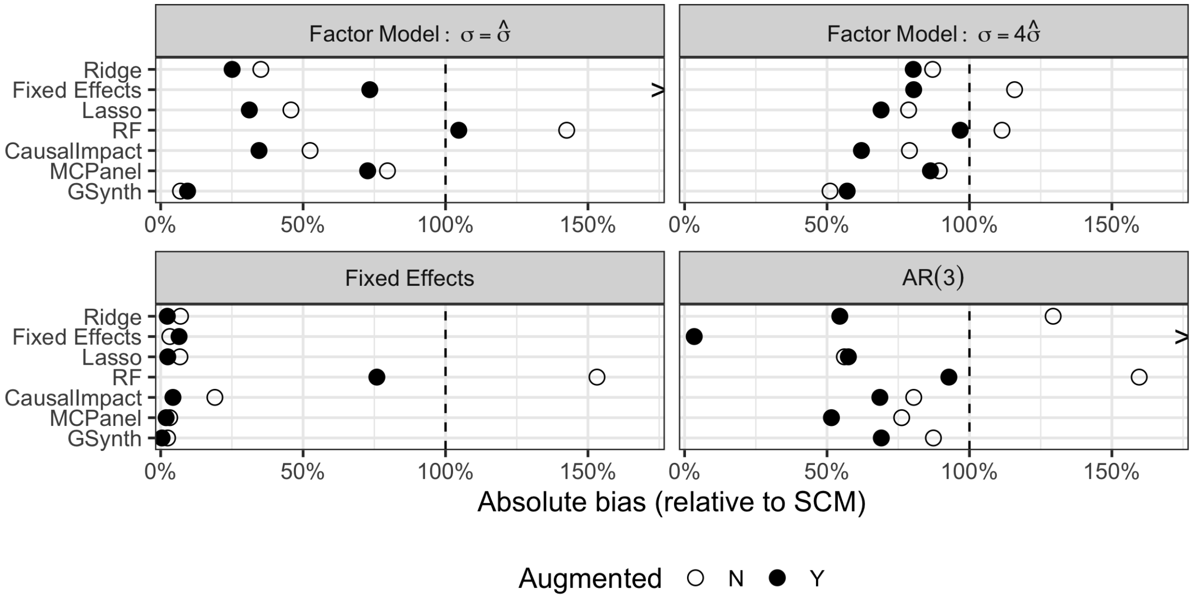

Figure F.2: Absolute bias (as a percentage of SCM bias) for ridge, fixed effects, and several machine learning and panel data outcome models, and their augmented versions using the same data generating processes as Figure 3.

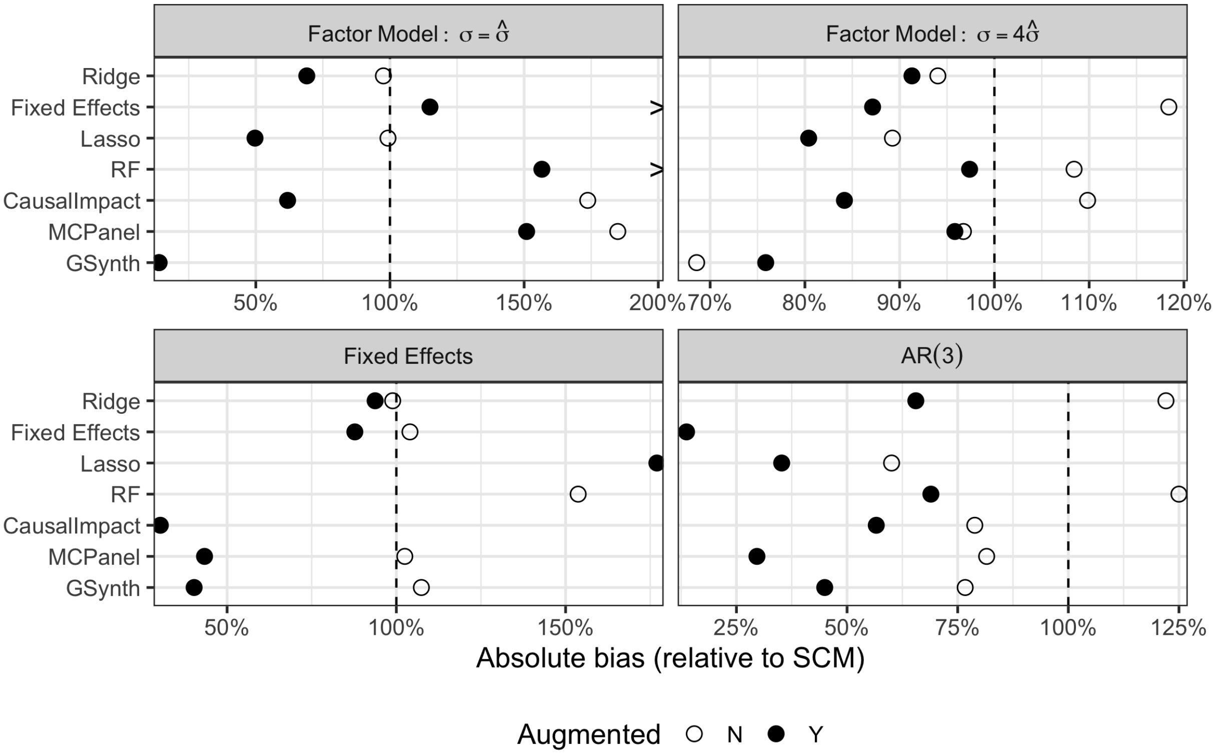

Figure F.3: Bias for different augmented and non-augmented estimators across outcome models conditioned on SCM fit in the top quintile.

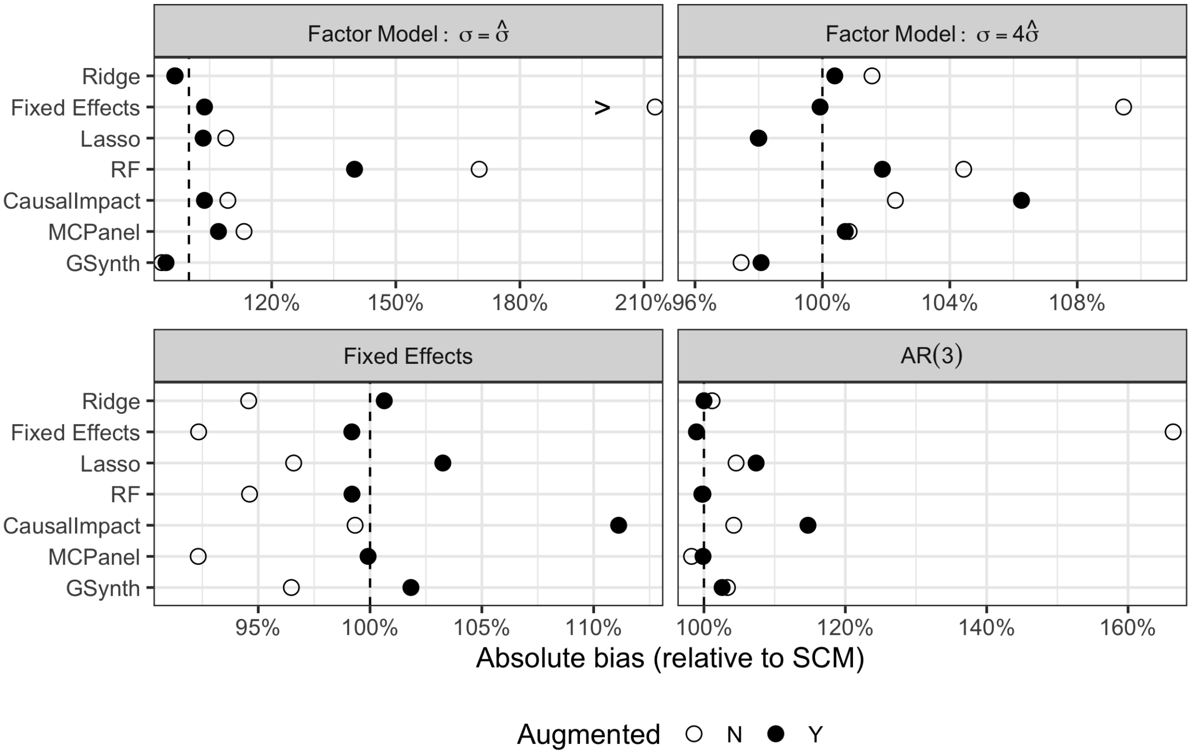

Figure F.4: RMSE for different augmented and non-augmented estimators across outcome models conditioned on SCM fit in the top quntile.

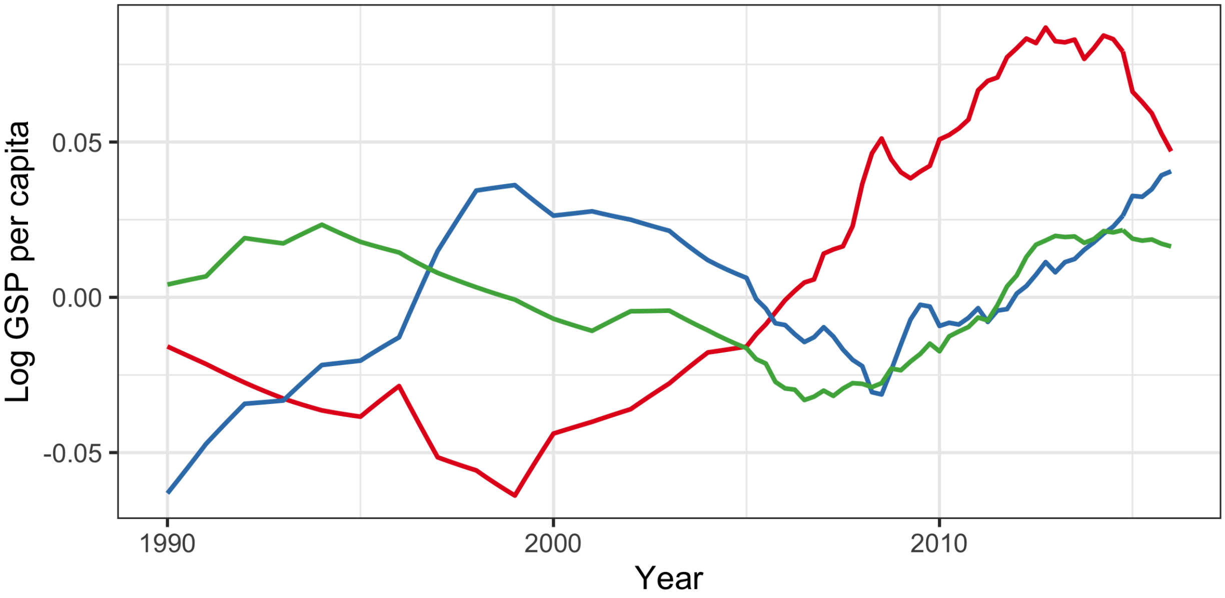

Figure F.5: Latent factors for calibrated simulation studies.

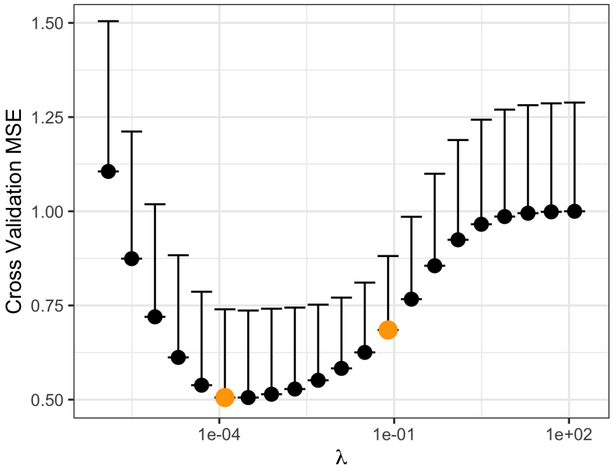

Figure F.6: Cross validation MSE and one standard error computed according to Equation (27). The minimal point, and the maximum λ within one standard error of the minimum are highlighted.

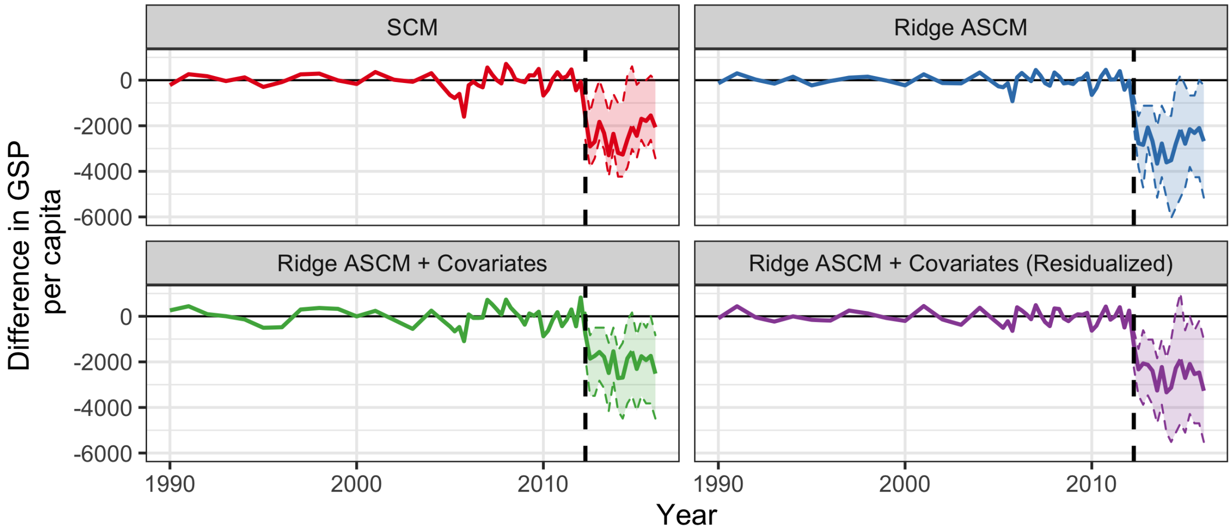

Figure F.7: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on GSP per capita using SCM, ridge ASCM, and ridge ASCM with covariates.

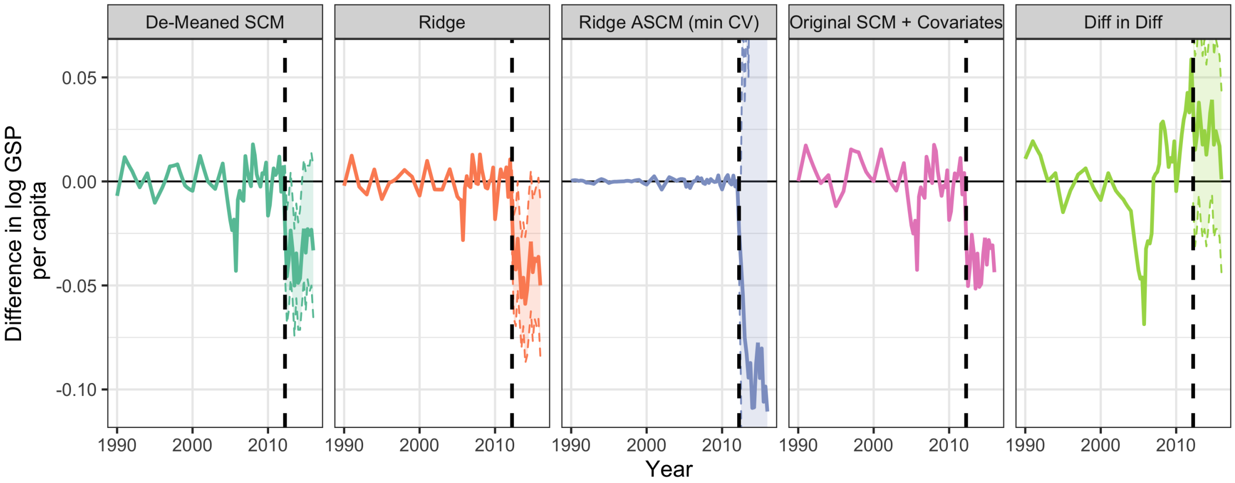

Figure F.8: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on log GSP per capita using de-meaned SCM, ridge regression alone, ridge ASCM with λ chosen to minimize the cross validated MSE, the original SCM proposal with covariates (Abadie et al.,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

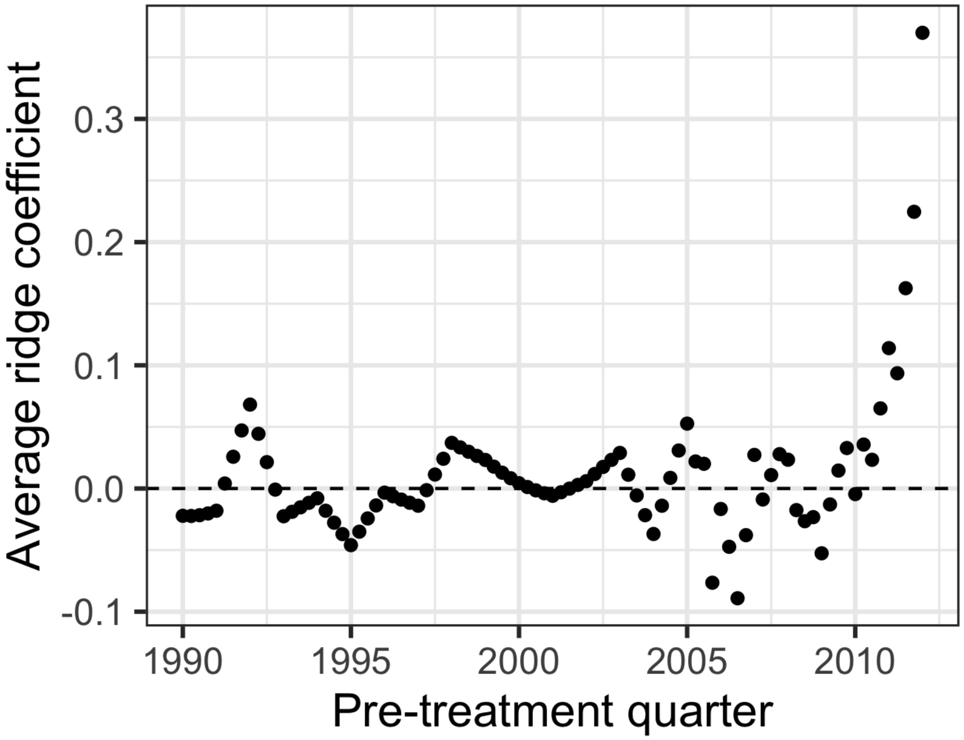

Figure F.9: Ridge regression coefficients for each pre-treatment quarter, averaged across post-treatment quarters.

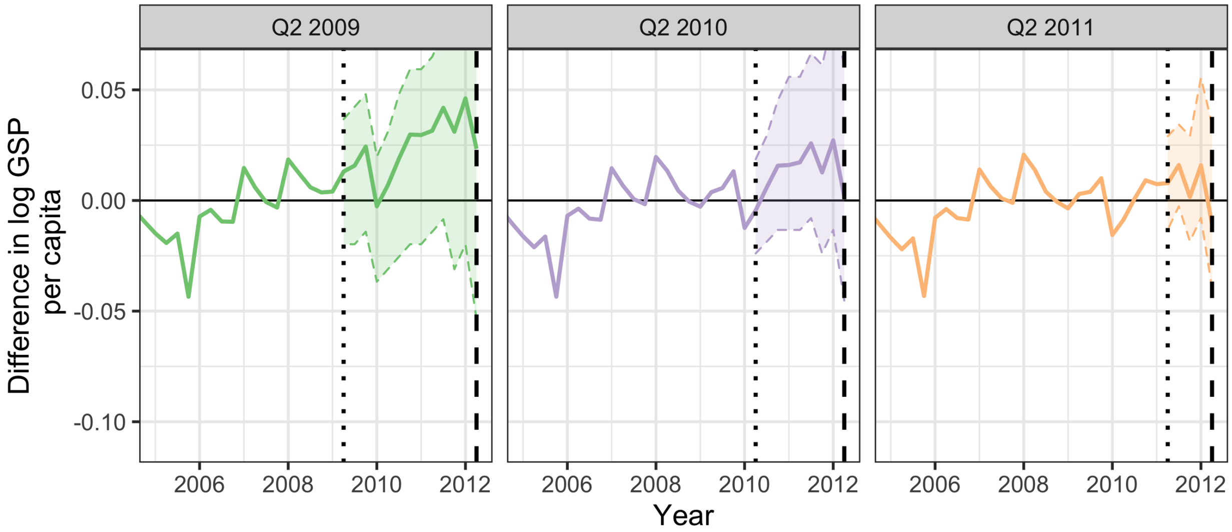

Figure F.10: Placebo point estimates along with 95% conformal confidence intervals for SCM with placebo treatment times in Q2 2009, 2010, and 2011. Scale begins in 2005 to highlight placebo estimates.

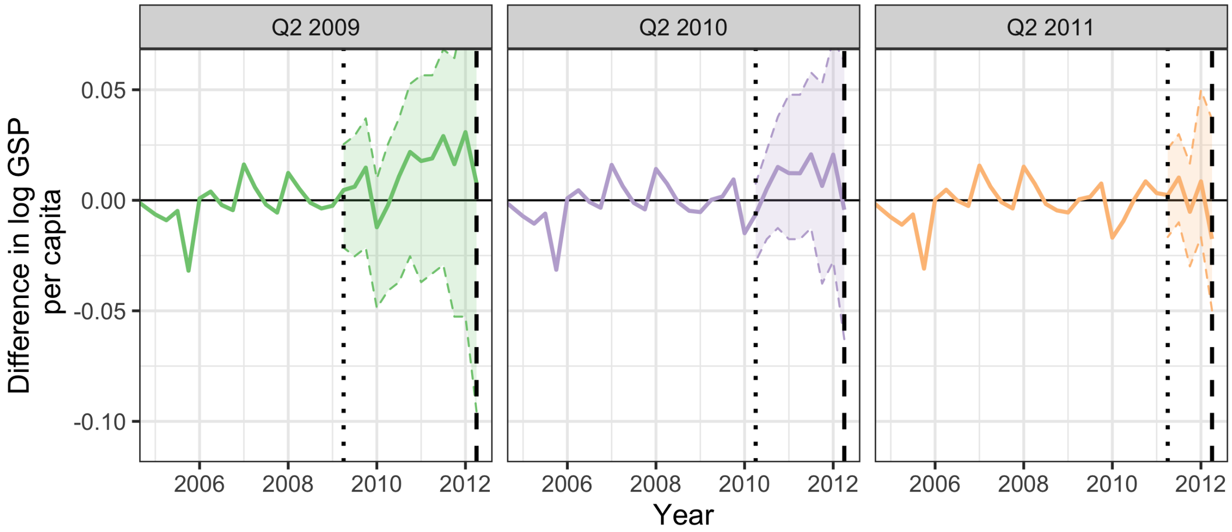

Figure F.11: Placebo point estimates along with 95% conformal confidence intervals for ridge ASCM with placebo treatment times in Q2 2009, 2010, and 2011. Scale begins in 2005 to highlight placebo estimates.

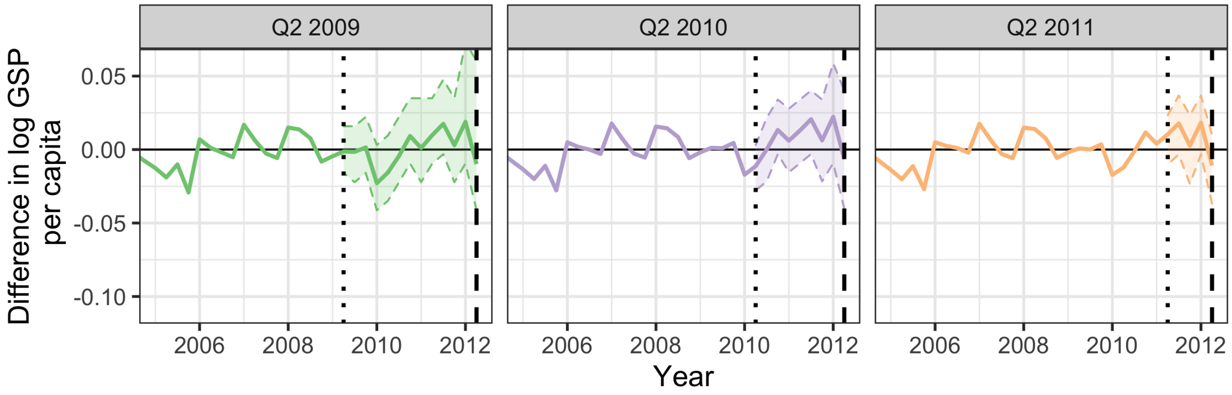

Figure F.12: Placebo point estimates along with 95% conformal confidence intervals for Ridge ASCM with covariates with placebo treatment times in Q2 2009, 2010, and 2011. The time period begins in 2005 and ends in Q1 2012 to highlight placebo estimates.

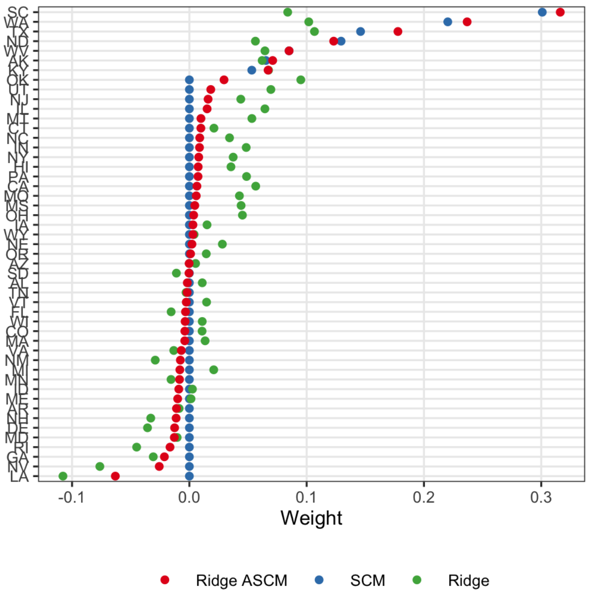

Figure F.13: Donor unit weights for SCM, ridge regression, and ridge ASCM balancing lagged outcomes.

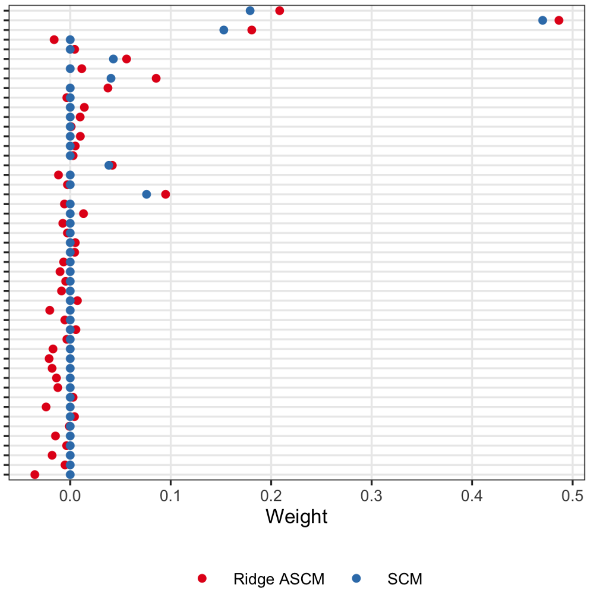

Figure F.14: Donor unit weights for SCM and ridge ASCM fit on lagged outcomes after residualizing out auxiliary covariates.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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