Regularization of Synthetic Controls for Policy Evaluation

Yi-Ting Chen
Department of Finance
National Taiwan University

This Version: February 23, 2022

Abstract

We explore an upper bound of the mean squared prediction error (MSPE) of an arbitrary synthetic control (SC) method in predicting the counterfactual of a treated unit. This potential MSPE is essential for unifying and comparing a variety of SC methods. It is established without assuming the true outcome model or imposing a combination restriction on the SC unit, and allows for the use of auxiliary models to deal with the potential imperfect matching between the treated unit and the SC unit. We further propose a generalized SC method to regularize the squared-bias and variance components of the potential MSPE. The regularized SC method encompasses several existing SC methods or their variants, and generates useful complements to existing methods. We also show the usefulness of our method by simulation and empirical illustration.

JEL Classification: C31, C54, D30.
Keywords: Synthetic control, biases, mean squared prediction error, regularization.

†Correspondence to: Yi-Ting Chen, Department of Finance, Center for Research in Econometric Theory and Applications, National Taiwan University, No.1, Sec.4, Roosevelt Rd., Taipei 10617, Taiwan. E-mail address: chenyt@ntu.edu.tw

‡Acknowledgements: This study is funded by the Ministry of Science and Technology of Taiwan (MOST-109-2410-H-002-219) and partially supported by the Center for Research in Econometric Theory and Applications (Grant no. 110L9002) from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan and MOST 109-2634-F-002-045.

1 Introduction

Policy evaluation involves estimating the effect of an intervention on certain outcome variables of the treated unit. Conceptually, the intervention effect may only be identified by contrasting the post-intervention outcomes of the treated unit with their counterfactual defined in the absence of intervention. By construction, the counterfactual is unobserved. Conventionally, it is common to estimate the intervention effect by the method of comparative case study (CCS) that identifies the counterfactual as the outcomes of an ideal controlled unit which is subjectively selected or determined by a natural experiment. It is also popular to estimate the intervention effect using the difference-in-differences method which requires the treated unit and the controlled unit to share parallel trends in the absence of intervention. Recently, the synthetic control (SC) method has been regarded as a promising alternative to these conventional methods, and has attracted a wide variety of empirical applications and theoretical extensions; see, e.g., Athey and Imbens (2017), Samartsidis et al. (2019) and Abadie (2021) for discussions.

The canonical SC method (CSC) was first introduced by Abadie and Gardeazabal (2003) in a case study. It replaces the ideal control unit of CCS, which is often infeasible in practice, by a data-driven SC unit. The CSC unit is a convex combination of a donor pool of untreated units that optimally matches the pre-intervention outcomes and features of the treated unit. The convex combination restriction is considered for avoiding the extrapolation and maintaining the sparsity (interpretability) of the composition of the CSC unit. By assuming that the CSC unit perfectly matches the treated unit in their pre-intervention outcomes and features, Abadie, Diamond and Hainmueller (2010, ADH) established the unbiasedness (a bias bound) of the counterfactual predicted by the outcomes of the CSC unit when the true outcome model is assumed to be an autoregressive model (a linear latent factor model without the parallel-trend assumption).

Nonetheless, as discussed by ADH (2010), CSC may have an "interpolation bias" if the counterfactual is a nonlinear function of the pre-intervention features. Moreover, the perfect-matching assumption fails if the pre-intervention outcomes and features of the treated unit are outside of the convex hull of their untreated counterparts. ADH (2010, p.495) addressed that CSC is not recommended for use in the presence of this poor-matching problem. Motivated by (one of) these two potential problems, several extensions of CSC have been recently proposed. In particular, Abadie and L'Hour (2019) proposed the penalized SC method (PSC) which accounts for the potential interpolation bias of

CSC. Researchers have also extended CSC with, or without, using auxiliary models for refining the poor-matching problem. Doudchenko and Imbens (2017) proposed a synthesis of related methods that determines the combination weights of a SC unit by estimating a penalized regression; see also Valero (2015) for the use of the Lasso regression. This approach deals with the poor-matching problem by relaxing the convex combination restriction of CSC. It also considers the inclusion of intercept in the penalized regression; see, e.g., Valero (2015), Doudchenko and Imbens (2017) and Ferman and Pinto (2019). This demeaned design itself is also useful for refining the poor-matching problem if the problem reflects the mean difference between the pre-intervention outcomes of the treated unit and the SC unit. Chen (2020) also considered a model-based SC method (MSC) that explains the poor-matching problem using a set of observed factors in the distributional context. In the mean context, MSC reduces to a demeaned SC method (DSC) when the observed factor degenerates to a constant. It is also related to, but different from, the augmented SC method (ASC) of Ben-Michael et al. (2021) which deals with the poor-matching problem of CSC from the aspect of bias correction; see also Abadie and L'Hour (2019), among others, for related bias-correction methods.

In this paper, we explore an upper bound of the mean squared prediction errors (MSPE) of the counterfactual predicted by an arbitrary SC method. This potential MSPE is essential for unifying and comparing various SC methods. It is established in a generalized context that does not assume the true outcome model or impose a combination restriction on the SC unit, and allows for the use of auxiliary models to refine the potential poor-matching problem. The potential MSPE is applicable to assessing the potential biases of the counterfactual predicted by an arbitrary SC method and the potential variance of the prediction errors. We also propose a generalized SC method to regularize the potential squared-bias and variance components of the potential MSPE. The regularized SC method (RSC) unifies a number of existing SC methods or their variants, and generates useful complements to existing methods. We also show the usefulness of the proposed method by comparing RSC with existing SC methods via theoretical discussions, simulation and empirical illustration.

The remainder of this paper is organized as follows. In Section 2, we define a generalized context of SC, establish the potential MSPE of an arbitrary SC method, and introduce the proposed RSC. In Section 3, we compare RSC with existing SC methods from theoretical viewpoints. Section 4 includes the simulation. Section 5 contains the empirical illustration. Section 6 concludes this paper. The mathematical proofs are collected in the Appendix.

2 The Proposed Method

Let $\{y_{it}\}$ be an outcome sequence of the ith unit at time t, for $i = 1, 2, \ldots, n+1$ and $t = 1, 2, \ldots, T$. Following Abadie and Gardeazabal (2003), ADH (2010, 2015), among many others, we focus on the context of single treated unit. Among the $n+1$ units, the first unit ($i = 1$) is the treated unit that experienced an intervention (or said a treatment) at $T_0 + 1$ for some $T_0 < T$, and the remaining ($i \geq 2$) constitutes a donor pool of n untreated units that are not affected by the intervention. Let δ_t be the intervention effect on the treated unit for $t \geq T_0 + 1$, and $y_{1t}(0)$ be the counterfactual outcome of the treated unit that would appear if the intervention was hypothetically absent for $t \geq T_0 + 1$. Denote the $T_0 \times 1$ vector of pre-intervention outcomes:

$$Y_i := (y_{i1}, \ldots, y_{iT_0})^\top,$$

for $i \geq 1$. A basic setting of the SC method is to extract $\{\delta_t\}_{t=T_0+1}^T$ from $\{y_{1t}\}_{t=T_0+1}^T$ by first matching Y_1 using a linear combination of the Y_i's of the n untreated units and then predicting $\{y_{1t}(0)\}_{t=T_0+1}^T$ using the same combination of the post-intervention outcomes $\{y_{it}\}_{t=T_0+1}^T$'s of the n untreated units.

This setting is popular in the recent SC literature; see, e.g., Doudchenko and Imbens (2017), Ferman and Pinto (2019) and Ben-Michael et al. (2021). As suggested by Doudchenko and Imbens (2017, p.20), it might be extended to include another vector of pre-intervention covariates, denoted as z_i, by replacing Y_i with the residual of the regression: Y_i on z_i in the setting; see also Abadie (2021, p.419) and Ben-Michael et al. (2021) for related discussions. In the following, we adopt this setting and consider a generalized framework to define the matching and prediction problems of the SC method.

2.1 A generalized framework of SC

Let $\mathcal{M}(\theta)$ be an "auxiliary model" which is considered for explaining the cross-sectional differences among the y_{it}'s in the absence of intervention using a set of observable covariates, $\mu_{it}(\theta_i)$ be a submodel of $\mathcal{M}(\theta)$ with the dependent variable y_{it}, and $\hat{\theta} := (\hat{\theta}_1^\top, \hat{\theta}_2^\top, \ldots, \hat{\theta}_{n+1}^\top)^\top$ be an estimator of the parameter vector $\theta := (\theta_1^\top, \theta_2^\top, \ldots, \theta_{n+1}^\top)^\top$ generated from the *pre-intervention* outcomes of the $n+1$ units:

$$\boldsymbol{Y} := (Y_1, Y_2, \ldots, Y_{n+1})$$

and possibly other pre-intervention covariates. Importantly, we do *not* assume $\mathcal{M}(\theta)$ to be the true model of the y_{it}'s in the absence of intervention. The auxiliary role of $\mathcal{M}(\theta)$ for the SC method will be explained later. Let $w := (w_2, \ldots, w_{n+1})^\top$ be a n-dimensional weighting vector in $\mathbb{R}^n$. We denote $\Delta\mu_{it}(\theta_1, \theta_i) := \mu_{1t}(\theta_1) - \mu_{it}(\theta_i)$,

$$y_t(w) := \sum_{i \geq 2} w_i y_{it}, \tag{1}$$

$$m_t(w, \theta) := \sum_{i \geq 2} w_i \Delta\mu_{it}(\theta_1, \theta_i) \tag{2}$$

and

$$y_t(w, s) := y_t(w) + s \cdot m_t(w, \theta), \tag{3}$$

for $t = 1, 2, \ldots, T$, where "$\sum_{i \geq 2}$" represents "$\sum_{i=2}^{n+1}$," and s is a selection variable which equals one if $\mathcal{M}(\theta)$ is used (otherwise, zero), and define the following $T_0 \times 1$ vectors:

$$Y(w) := (y_1(w), y_2(w), \ldots, y_{T_0}(w))^\top, \tag{4}$$

$$M(w, \theta) := (m_1(w, \theta), m_2(w, \theta), \ldots, m_{T_0}(w, \theta))^\top \tag{5}$$

and

$$Y(w, s) := Y(w) + s \cdot M(w, \theta). \tag{6}$$

In this framework, an arbitrary SC method matches the treated unit's pre-intervention outcome vector Y_1 using

$$\hat{Y}(w, s) := Y(w) + s \cdot M(w, \hat{\theta}), \tag{7}$$

and predicts the counterfactual $y_{1t}(0)$ using

$$\hat{y}_t(w, s) := y_t(w) + s \cdot m_t(w, \hat{\theta}) \tag{8}$$

and estimates the intervention effect δ_t by

$$\hat{\delta}_t = y_{1t} - \hat{y}_t(w, s), \tag{9}$$

for $t \geq T_0 + 1$, by the choice of (w, s) and the design of $\mathcal{M}(\theta)$ if $s = 1$. In the case where $s = 0$, the restrictions: $\hat{Y}(w, s) = Y(w, s) = Y(w)$ and $\hat{y}_t(w, s) = y_t(w, s) = y_t(w)$ are satisfied, and the SC method chooses a particular w to match Y_1 by $Y(w)$ and to predict

$y_{1t}(0)$ by $y_t(w)$ without using $\mathcal{M}(\theta)$. In the case where $s = 1$, $\mathcal{M}(\theta)$ is an auxiliary model which is introduced to refine the poor-matching problem that arises when $Y_1 - Y(w) \neq 0$. The refinement is based on interpreting $Y_1 - Y(w)$ using $M(w, \hat{\theta})$. We consider this generalized framework because it is flexible enough to encompass a wide variety of SC methods, as will be discussed in Section 3.

2.2 Potential MSPE

Since an arbitrary SC method estimates the intervention effect by predicting the counterfactual, it is undoubtedly essential to explore the potential properties of the prediction. Denote the unit simplex:

$$\mathbb{W} := \{w | w \in [0,1]^n \text{ and } \iota_n^\top w = 1\}, \tag{10}$$

where $\iota_n := (1, \ldots, 1)^\top$ denotes a $n \times 1$ vector of one. ADH (2010) established a bias bound of the prediction generated by CSC, which sets $s = 0$ and requires w to satisfy the convex-combination restriction: $w \in \mathbb{W}$, by assuming that the true model (of the y_{it}'s) is a linear factor model and that the perfect-matching condition:

$$Y_1 = Y(w) \tag{11}$$

holds for some $w \in \mathbb{W}$; see Botosaru and Ferman (2019) for related discussions. Several studies also explored the bias bounds or the prediction errors of their SC methods by assuming that the true model is a linear model or a linear factor model; see, e.g., Amjad et al. (2018) that relaxes the restriction: $w \in \mathbb{W}$ and Ferman and Pinto (2019) and Ben-Michael et al. (2021) that further relax the perfect-matching condition. In the following, we explore a potential MSPE of an arbitrary SC method that predicts $y_{1t}(0)$ by $\hat{y}_t(w, s)$ without assuming the true model or the perfect-matching condition and without requiring a combination restriction.

Let $\psi_t(Y_i) := \mathbb{E}[y_{it}|Y_i]$ be the *unknown* conditional mean function of y_{it} given Y_i, and D_{1t} be a binary intervention variable for the treated unit ($i = 1$) with the two potential outcomes: $D_{1t} = 0$ (untreated) and $D_{1t} = 1$ (treated) for $t \geq T_0 + 1$. Denote $D_{it} := 0$ for $t \leq T_0$ if $i = 1$ and for all t's if $i \geq 2$, a $T_0 \times 1$ vector:

$$m_{i\cdot}(\theta) := (\Delta\mu_{i1}(\theta_1, \theta_i), \ldots, \Delta\mu_{iT_0}(\theta_1, \theta_i))^\top$$

and a $n \times 1$ vector:

$$m_{\cdot t}(\theta) := (\Delta\mu_{2t}(\theta_1, \theta_2), \ldots, \Delta\mu_{n+1,t}(\theta_1, \theta_{n+1}))^\top.$$

We make the following assumptions:

Assumption 1 *For $t \geq T_0 + 1$ and for all i's,*
(i) $y_{it} = \psi_t(Y_i) + \delta_t D_{it} + \varepsilon_{it}$, where ε_{it} is a zero-mean error;
(ii) $\mathbb{E}[\psi_t(Y_i)|Y_i, D_{1t}] = \psi_t(Y_i)$;
(iii) $\mathbb{E}[\varepsilon_{it}|Y_i, D_{1t}] = 0$.

Assumption 2 *For $t \geq T_0 + 1$ and for $i \geq 2$,*
(i) $|\psi_t(\cdot)| \leq \xi_{\psi,0}$ and $\|\nabla\psi_t(\cdot)\| \leq \xi_{\psi,1}$, for some finite $\xi_{\psi,0}$ and $\xi_{\psi,1}$;
(ii) $\|m_{i\cdot}(\theta)\| \leq \xi_{m,0}$ and $\|\nabla_{\theta^\top} m_{\cdot t}(\cdot)\| \leq \xi_{m,1}$, for some finite $\xi_{m,0}$ and $\xi_{m,1}$, if $s = 1$.

Assumption 3 *For $t \geq T_0 + 1$ and for all i's,*
(i) $\mathbb{E}[\varepsilon_{it}|\boldsymbol{Y}] = 0$, $\mathbb{E}[\varepsilon_{it}^2|\boldsymbol{Y}] = \sigma_\varepsilon^2 < \infty$ and $\mathbb{E}[\varepsilon_{it}\varepsilon_{jt}|\boldsymbol{Y}] = 0$, for $j \neq i$;
(ii) $\mathbb{E}[m_t(w, \hat{\theta}) - m_t(w, \theta)|\boldsymbol{Y}] = 0$ and $\sigma_\theta^2 := \mathbb{E}[\|\hat{\theta} - \theta\|^2|\boldsymbol{Y}] < \infty$, if $s = 1$.

Assumption 1 (i) requires Y_i to be informative for predicting y_{it} in a reduced form. It also requires the treated and untreated units to be comparable in the sense that they share the same conditional mean function $\psi_t(\cdot)$ in the absence of intervention. This condition is often presented in a stricter form in related studies. As mentioned, ADH (2010) assumes the true model to be a linear factor model; see also Gobillon and Magnac (2016), Xu (2017) and Ben-Michael et al. (2021), among others, for the use of factor models. For the treated unit ($i = 1$), Assumption 1(ii) implies "no anticipation effect" of the intervention, and Assumption 1(iii) is an unconfoundedness condition. For the untreated units ($i \geq 2$), Assumption 1(ii) and (iii) imply "no interference effect" of the intervention. Such an assumption is standard in the SC literature; see ADH (2010, p.494) for related discussions. It is essential for interpreting δ_t as the average treatment effect on the treated. It is also important for the contextual evaluation of the SC method discussed by Abadie (2021, Section 5). Assumption 2 requires $\psi_t(\cdot)$ and $\mathcal{M}(\theta)$ to be bounded and differentiable. This condition is weak but essential for establishing the potential MSPE of the prediction of counterfactual generated by an arbitrary SC method. Assumption 3 is considered for simplicity. It requires the reduced-form error ε_{it} to be unpredictable by $\boldsymbol{Y}$. It also requires that, given $\boldsymbol{Y}$, ε_{it} is conditionally homoskedastic and uncorrelated to ε_{jt};

moreover, $m_t(w, \hat{\theta})$ is unbiased for $m_t(w, \theta)$ with a finite σ_θ^2 if $s = 1$. Note that σ_θ^2 is a measure for the estimation uncertainty of $\hat{\theta}$, and reduces to $\|\hat{\theta} - \theta\|^2$ if $\hat{\theta}$ is fully generated from $\boldsymbol{Y}$.

In the Appendix, we show the following decomposition:

Lemma 1 *Given Assumption 1, for $t \geq T_0 + 1$,*

$$y_{1t}(0) - \hat{y}_t(w, s) = Bias_t(w, s) + u_t(w, s), \tag{12}$$

where

$$\begin{aligned} Bias_t(w, s) &:= \psi_t(Y_1) - \sum_{i \geq 2} w_i \psi_t(Y_i) - s \cdot m_t(w, \theta) \\ &= \underbrace{(\psi_t(Y_1) - \psi_t(Y(w, s)))}_{Bias_{pre,t}(w,s)\text{: pre-intervention bias}} + \underbrace{\left(\psi_t(Y(w)) - \sum_{i \geq 2} w_i \psi_t(Y_i) \right)}_{Bias_{nl,t}(w)\text{: nonlinearity bias}} \\ &+ \underbrace{(\psi_t(Y(w, s)) - \psi_t(Y(w)) - s \cdot m_t(w, \theta))}_{Bias_{ms,t}(w,s)\text{: model specification bias}} \end{aligned} \tag{13}$$

and

$$u_t(w, s) := \underbrace{(\varepsilon_{1t} - \varepsilon_t(w))}_{e_t(w)\text{: intrinsic error}} - s \cdot \underbrace{\Big(m_t(w, \hat{\theta}) - m_t(w, \theta) \Big)}_{\text{estimation uncertainty}}, \tag{14}$$

with $\varepsilon_t(w) := \sum_{i \geq 2} w_i \varepsilon_{it}$.

This shows that an arbitrary SC method that predicts $y_{1t}(0)$ using $\hat{y}_t(w, s)$ is potentially biased because of the possible presence of the "pre-intervention bias" $Bias_{pre,t}(w, s)$, the "nonlinearity bias" $Bias_{nl,t}(w)$ and the "model specification bias" $Bias_{ms,t}(w, s)$ that are defined in (13). Note that the pre-intervention bias and the nonlinearity bias correspond to the "extrapolation bias" and the "interpolation bias" considered by Kellogg et al. (2020), respectively, *if* $s = 0$ and $w \in \mathbb{W}$. In the following, we further interpret these potential bias components.

We interpret $Bias_{pre,t}(w, s)$ as the pre-intervention bias because it is zero if the pre-intervention outcomes satisfy a weak form of the perfect-matching condition:

$$Y_1 = Y(w, s); \tag{15}$$

otherwise, it is in general non-zero. Note that (15) reduces to the perfect-matching condition (11) if $s = 0$. As mentioned, a SC method with $s = 0$ has the poor-matching problem if (11) is not satisfied, and a SC method with $s = 1$ defends against this type of bias by interpreting $Y_1 - Y(w)$ using $M(w, \hat{\theta})$.

We interpret $Bias_{nl,t}(w)$ as the nonlinearity bias because it is zero when $\psi_t(\cdot)$ is linear in the sense that

$$\psi_t(Y) = \alpha_{1t} + Y^\top \beta_{1t}, \tag{16}$$

where Y is an arbitrary $T_0 \times 1$ vector, and α_{1t} and β_{1t} are, respectively, a scalar and a $T_0 \times 1$ vector of unknowns that could be time-varying, if the convex-combination restriction: $w \in \mathbb{W}$ and condition (11) are satisfied. To see this point, note that

$$Bias_{nl,t}(w) = (1 - \iota_n^\top w)\, \psi_t(Y(w)) + \sum_{i \geq 2} w_i \left(\psi_t(Y(w)) - \psi_t(Y_i)\right). \tag{17}$$

If $w \in \mathbb{W}$, we can simplify (17) as:

$$Bias_{nl,t}(w) = \sum_{i \geq 2} w_i \left(\psi_t(Y_1) - \psi_t(Y_i)\right). \tag{18}$$

If $w \in \mathbb{W}$ and that the linearity in (16) is satisfied, we can further simplify (18) as:

$$Bias_{nl,t}(w) = \sum_{i \geq 2} w_i \left(Y_1^\top \beta_{1t} - Y_i^\top \beta_{1t}\right) = (Y_1 - Y(w))^\top \beta_{1t}, \tag{19}$$

which is zero under condition (11). In comparison, $Bias_{nl,t}(w)$ is in general non-zero when $\psi_t(\cdot)$ is nonlinear even if condition (11) holds for some $w \in \mathbb{W}$. This is consistent with the interpolation bias discussed by ADH (2010) and Kellogg et al. (2020).

We interpret $Bias_{ms,t}(w, s)$ as the model specification bias because it is zero when $s = 0$; that is, when the SC method does not involve the use of $\mathcal{M}(\theta)$. Note that, under the linearity in (16),

$$\begin{aligned} Bias_{ms,t}(w, s) &= Y(w, s)^\top \beta_t - Y(w)^\top \beta_t - s \cdot m_t(w, \theta) \\ &= s \cdot \left(M(w, \theta)^\top \beta_t - m_t(w, \theta)\right). \end{aligned}$$

Thus, $Bias_{ms,t}(w, 1)$ reduces to zero if the linearity also holds for $M(w, \theta)$ in predicting $m_t(w, \theta)$; otherwise, it is in general non-zero.

In addition to these potential bias components, the decomposition in (12) also includes

the composite error: $u_t(w,s)$ shown in (14), which comprises the intrinsic error $e_t(w)$ and the measure of estimation uncertainty σ_θ^2 if $s=1$. Importantly, this means that the choice of (w,s) might influence the theoretical properties of the counterfactual predicted by the SC method through the potential biases and the variance of prediction errors. To make this point clear, we need to further explore the conditional MSPE:

$$MSPE_t(w,s) := \mathbb{E}\left[(y_{1t}(0) - \hat{y}_t(w,s))^2 \middle| \boldsymbol{Y} \right], \tag{20}$$

which summarizes the potential biases and variance generated by predicting $y_{1t}(0)$ using $\hat{y}_t(w,s)$ for $t \geq T_0 + 1$.

Let $\|\cdot\|_1$ be the L_1 norm of a vector, and $\|\cdot\|$ be the L_2 norm of a vector or the Frobenius norm of a matrix. In particular, $\|w\|_1 := \sum_{i\geq 2} |w_i|$ and $\|w\| := \sqrt{\sum_{i\geq 2} w_i^2}$. We define the following divergence measures:

$$B_p(w) := \|Y_1 - Y(w)\|, \tag{21}$$

$$B_p(w,s) := \|Y_1 - Y(w,s)\|, \tag{22}$$

$$B_c(w) := \sum_{i\geq 2} |w_i| \|Y_1 - Y_i\| \tag{23}$$

and

$$B_a(w) := |1 - \iota_n^\top w| \,. \tag{24}$$

Note that $B_p(w)$ measures the divergence of the perfect-matching condition (11), $B_p(w,s)$ measures the divergence of a weak form of the perfect-matching condition defined in (15), $B_c(w)$ measures the divergence of the "perfect-control condition:"

$$Y_1 = Y_i, \quad \text{if } w_i \neq 0, \tag{25}$$

and $B_a(w)$ measures the divergence of the aggregation restriction: $w \in \mathbb{A}$, where

$$\mathbb{A} := \{w | w \in \mathbb{R}^n \text{ and } \iota_n^\top w = 1\} \,.$$

Also, note that $B_p(w,s) = B_p(w)$ holds if $s=0$. Moreover, condition (25) is stricter than condition (11), and the restriction: $w \in \mathbb{A}$ is weaker than the restriction: $w \in \mathbb{W}$.

In the Appendix, we show the following result:

Proposition 1 *Given Assumptions 1-3, for* $t \geq T_0 + 1$,

$$MSPE_t(w,s) = Bias_t^2(w,s) + \sigma_t^2(w,s), \tag{26}$$

$$\sigma_t^2(w,s) = \sigma_\varepsilon^2\left(1 + \|w\|^2\right) + s \cdot \mathbb{E}[(m_t(w,\hat{\theta}) - m_t(w,\theta))^2|\boldsymbol{Y}], \tag{27}$$

$$|Bias_{pre,t}(w,s)| \leq \overline{Bias}_{pre}(w,s) := \xi_{\psi,1} B_p(w,s), \tag{28}$$

$$|Bias_{nl,t}(w)| \leq \overline{Bias}_{nl}(w) := \xi_{\psi,0} B_a(w) + \xi_{\psi,1}\left(\|w\|_1 B_p(w) + B_c(w)\right), \tag{29}$$

$$|Bias_{ms,t}(w,s)| \leq \overline{Bias}_{ms}(w,s) := (1 + \xi_{\psi,1})\xi_{m,0}(s \cdot \|w\|_1), \tag{30}$$

$$\begin{aligned} |Bias_t(w,s)| \leq \overline{Bias}(w,s) := {} & \xi_{\psi,1}\left(B_p(w,s) + \|w\|_1 B_p(w) + B_c(w)\right) + \xi_{\psi,0} B_a(w) \\ & + (1 + \xi_{\psi,1})\xi_{m,0}\left(s \cdot \|w\|_1\right), \end{aligned} \tag{31}$$

$$\sigma_t^2(w,s) \leq \bar{\sigma}^2(w,s) := \sigma_\varepsilon^2 + (\sigma_\varepsilon^2 + s \cdot \xi_{m,1}^2 \sigma_\theta^2)\|w\|^2 \tag{32}$$

and

$$MSPE_t(w,s) \leq \overline{MSPE}(w,s) := \overline{Bias}^2(w,s) + \bar{\sigma}^2(w,s). \tag{33}$$

This shows that, given $\boldsymbol{Y}$, the MSPE is composed of the squared bias of $\hat{y}_t(w,s)$ and the variance of the prediction error: $\sigma_t^2(w,s)$, which is the same as $\sigma_\varepsilon^2(1+\|w\|^2)$ if $s = 0$ and influenced by the estimation uncertainty of $\hat{\theta}$ if $s = 1$. It also shows that the pre-intervention bias, the nonlinearity bias, the model specification bias, the potential bias and the variance of prediction errors are, respectively, bounded above by $\overline{Bias}_{pre}(w,s)$, $\overline{Bias}_{nl}(w)$, $\overline{Bias}_{ms}(w,s)$, $\overline{Bias}(w,s)$ and $\bar{\sigma}^2(w,s)$. If $w \in \mathbb{W}$, $\overline{Bias}_{pre}(w,s)$ reduces to $\xi_{\psi,1} B_p(w)$ if $s = 0$, and $\overline{Bias}_{nl}(w)$ degenerates to $\xi_{\psi,1} B_c(w)$ if $B_p(w) = 0$, where $B_p(w)$ and $B_c(w)$ correspond to the "interpolation measure" and the "extrapolation measure"considered by Kellogg et al. (2020), respectively. In addition, the bias of $\hat{y}_t(w,s)$, the variance $\sigma_t^2(w,s)$ and the MSPE are, respectively, bounded above by $\overline{Bias}(w,s)$, $\bar{\sigma}^2(w,s)$ and $\overline{MSPE}(w,s)$. These bounds all hold for $t \geq T_0 + 1$. Given the unknown $\xi_{\psi,0}$ and $\xi_{\psi,1}$ (and the model-specific $\xi_{m,0}$ and $\xi_{m,1}$ if $s = 1$), the bounds are determined by the choice of (w,s).

Before further discussions, it should be noted that the decomposition presented in Lemma 1 is by no means unique and that the upper bounds presented in Proposition 1

are not ensured to be the least upper bounds that are difficult to establish in our context. Nonetheless, the decomposition and these bounds have important implications on the SC method. In particular, $\overline{MSPE}(w,s)$ constitutes a potential (conservative) MSPE of an arbitrary SC method that predicts the counterfactual $y_{1t}(0)$ using $\hat{y}_t(w,s)$ for $t \geq T_0+1$. This allows us to contrast the potential differences among different SC methods in a unified framework. In addition, as shown by (31), (32) and (33), $\overline{MSPE}(w,s)$ is governed by the "matching-quality" divergences: $B_p^2(w,s)$, $\|w\|_1^2 B_p^2(w)$ and $B_c^2(w)$ and the combination measures: $B_a^2(w)$, $s\cdot\|w\|_1^2$ and $\|w\|^2$ that may be regularized by the choice of (w,s). This illustrates that the prediction problem is inseparable from the matching problem for the SC method through the choice of (w,s), and motivates us to propose a generalized SC method, that is RSC, by regularizing the squared bias and variance components of $\overline{MSPE}(w,s)$ via the choice of (w,s). As will be shown in Section 3, several existing SC methods, or their variants, amount to choosing (w,s) by minimizing a certain combination of the components of $\overline{MSPE}(w,s)$, and hence an encompassed by RSC.

2.3 The regularized SC method

To introduce RSC, note that $B_p(w,s)$ is dependent on the unknown parameter θ in the case where $s=1$. In this case, we estimate $B_p(w,s)$ using its $\hat{\theta}$-based counterpart:

$$\hat{B}_p(w,s) := \|Y_1 - \hat{Y}(w,s)\|.$$

Note that the restriction: $\hat{B}_p(w,s) = B_p(w,s) = B_p(w)$ holds if $s=0$. Since $\overline{MSPE}(w,s)$ is governed by the matching-quality divergences and the combination measures, it is sensible to choose (w,s) by minimizing a multiple-objective function:

$$Q(w,s|r) := \hat{B}_p^2(w,s) + r_1\|w\|_1^2 B_p^2(w) + r_2 B_c^2(w) + r_3 B_a^2(w) + r_4(s\cdot\|w\|_1^2) + r_5\|w\|^2, \quad (34)$$

where $r := (r_1, r_2, r_3, r_4, r_5) \geq 0$ denotes a vector of regularization parameters that controls how the divergence measures are regularized. However, this minimization problem is complicated because $Q(w,s|r)$ is highly nonlinear in terms of w given a specific (s,r).

To simplify this minimization problem as a quadratic programming problem, we adopt a sign-and-size restriction of w: $w \in \mathbb{S}(\tau)$, where

$$\mathbb{S}(\tau) := \left\{ w | w \in \mathbb{R}_+^n \text{ and } \iota_n^\top w \leq \tau, \text{ for some } \tau \geq 1 \right\}. \quad (35)$$

This restriction implies $\|w\|_1^2 B_p^2(w) \leq \tau^2 B_p^2(w)$ and $s \cdot \|w\|_1^2 \leq s \cdot \tau^2$ because $\|w\|_1 = \iota_n^\top w$ when $w \in \mathbb{R}_+^n$. It also implies $B_a^2(w) \leq b_\tau$, where $b_\tau := \max(1, (1-\tau)^2)$, because $1 - \tau \leq 1 - \iota_n^\top w \leq 1$ when $w \in \mathbb{R}_+^n$ and $\iota_n^\top w \leq \tau$. Accordingly, we may use the restriction: $w \in \mathbb{S}(\tau)$ to establish an upper bound of $Q(w, s|r)$:

$$\begin{aligned}\bar{Q}(w, s|\tau, r) &:= \hat{B}_p^2(w, s) + r_1\tau^2 B_p^2(w) + r_2 B_c^2(w) + r_3 b_\tau + r_4(s \cdot \tau^2) + r_5\|w\|^2, \\ &= \begin{cases} \hat{B}_p^2(w, 1) + r_1\tau^2 B_p^2(w) + r_2 B_c^2(w) + r_5\|w\|^2 + r_3 b_\tau + r_4\tau^2, & \text{if } s = 1, \\ (1 + r_1\tau^2) B_p^2(w) + r_2 B_c^2(w) + r_5\|w\|^2 + r_3 b_\tau, & \text{if } s = 0. \end{cases}\end{aligned} \tag{36}$$

Note that $\bar{Q}(w, s|\tau, r)$ considerably simplifies $Q(w, s|r)$. It only involves the matching-quality divergences: $\hat{B}_p^2(w, s)$, $B_c^2(w)$ and $s \cdot B_p^2(w)$ and the squared L_2 norm: $\|w\|^2$. Moreover, because $B_c(w) = \sum_{i \geq 2} w_i \|Y_1 - Y_i\|$ holds under the sign restriction: $w \in \mathbb{R}_+^n$, we may further present the minimization of $\bar{Q}(w, s|\tau, r)$ with respect to w as a quadratic programming problem for a fixed (s, τ, r).

To see this point, note that by rescaling the coefficients and removing the constants of $\bar{Q}(w, s|\tau, r)$, we transform the minimization of $\bar{Q}(w, s|\tau, r)$ with respect to w to the minimization of the following function with respect to w:

$$Q^*(w|s, \kappa) := \hat{B}_p^2(w, s) + \kappa_1 B_c^2(w) + \kappa_2\|w\|^2 + s \cdot \kappa_3 B_p^2(w) \tag{37}$$

under the constraint: $w \in \mathbb{S}(\tau)$, for a fixed (s, κ), where κ is a regularization-parameter vector such that $\kappa = (\kappa_1, \kappa_2) \geq 0$, if $s = 0$, or $\kappa = (\kappa_1, \kappa_2, \kappa_3) \geq 0$, if $s = 1$. Accordingly, we propose a generalized SC method, that is RSC, that chooses the following w:

$$w_{rsc}(s, \tau, \kappa) := \underset{w \in \mathbb{S}(\tau)}{\operatorname{argmin}} Q^*(w|s, \kappa), \tag{38}$$

for a fixed (s, τ, κ). In the Appendix, we show that $w_{rsc}(s, \tau, \kappa)$ is the solution to a quadratic programming problem:

$$w_{rsc}(s, \tau, \kappa) = \underset{w \in \mathbb{S}(\tau)}{\operatorname{argmin}} \left(\frac{1}{2} w^\top V w - v^\top w \right), \tag{39}$$

where

$$V := 2\left(\hat{\boldsymbol{Y}}_{(-1)}^{s\top} \hat{\boldsymbol{Y}}_{(-1)}^{s} + \kappa_1 \boldsymbol{D}_Y \boldsymbol{D}_Y^\top + \kappa_2 I_n + s \cdot \kappa_3 \boldsymbol{Y}_{(-1)}^\top \boldsymbol{Y}_{(-1)} \right) \tag{40}$$

and

$$v := 2\left(\hat{\boldsymbol{Y}}^{s}_{(-1)} + s \cdot \kappa_3 \boldsymbol{Y}_{(-1)}\right)^{\top} Y_1 \tag{41}$$

are defined by the following $T_0 \times n$ matrices:

$$\boldsymbol{Y}_{(-1)} := (Y_2, \ldots, Y_{n+1}),$$

$$\boldsymbol{M}_{(-1)}(\hat{\theta}) := (M_2(\hat{\theta}), \ldots, M_{n+1}(\hat{\theta}))$$

and

$$\hat{\boldsymbol{Y}}^{s}_{(-1)} := \boldsymbol{Y}_{(-1)} + s \cdot \boldsymbol{M}_{(-1)}(\hat{\theta})$$

and the $n \times 1$ vector:

$$\boldsymbol{D}_Y := (\|Y_1 - Y_2\|, \ldots, \|Y_1 - Y_{n+1}\|)^{\top}.$$

In applications, we solve $w_{rsc}(s, \tau, \kappa)$ using the `R` package "`quadprog`" for a fixed (s, τ, κ).

3 Comparison with Existing SC Methods

In this section, we illustrate that several existing SC methods, or their variants, could be interpreted as particular RSCs with different settings of (s, τ, κ), and the proposed method generates useful complements to existing SC methods.

3.1 Convex combination

We first consider the case where $s = 0$ and $\tau = 1$. In this case, $w_{rsc}(s, \tau, \kappa)$ degenerates to

$$w_{rsc}(0, 1, \kappa_1, \kappa_2) = \underset{w \in \mathbb{S}(1)}{\operatorname{argmin}} B_p^2(w) + \kappa_1 B_c^2(w) + \kappa_2 \|w\|^2. \tag{42}$$

In comparison, CSC chooses the following w:

$$w_{csc} := \underset{w \in \mathbb{W}}{\operatorname{argmin}} \|Y_1 - Y(w)\|^2. \tag{43}$$

Note that CSC is fundamentally essential for the whole SC literature. It generalizes CCS by replacing a restrictive assumption that the perfect-control condition (25) holds for a single control unit with a weaker assumption that the perfect-matching condition in (11) holds

for $w = w_{csc}$; that is, $B_p(w_{csc}) = 0$. Importantly, because $B_p(w) = \|Y_1 - Y(w)\|$, as shown in (21), and $\mathbb{W} \subset \mathbb{S}(1)$, RSC includes CSC as a special case where $w = w_{rsc}(0, 1, 0, 0)$.

Under the perfect-matching condition: $B_p(w_{csc}) = 0$, ADH (2010) established a bias bound of CSC by assuming the true model to be a linear factor model. Indeed, Proposition 1 implies that an arbitrary SC method with $s = 0$ and $w \in \mathbb{W}$ has the bias bound:

$$\overline{Bias}(w, 0) = \xi_{\psi,1} \left(2B_p(w) + B_c(w)\right); \tag{44}$$

see (31). This illustrates that the perfect-matching condition is insufficient for ensuring the unbiasedness of CSC if $\psi_t(\cdot)$ is unknown, unless $B_c(w_{csc}) = 0$. Since the nonlinearity bias of an arbitrary SC method with $s = 0$ and $w \in \mathbb{W}$ is bounded above by

$$\overline{Bias}_{nl}(w) = \xi_{\psi,1} \left(B_p(w) + B_c(w)\right), \tag{45}$$

as implied by (29), this result is consistent with the statement of ADH (2010) that CSC may have an interpolation bias if the counterfactual is a nonlinear function of the pre-intervention features. Therefore, it is theoretically important to regularize not only the pre-intervention bias but also the nonlinearity bias, which might appear if $B_c(w_{csc}) > 0$, by choosing w. This notion is an essential motivation of PSC which is originally established in the context of multiple treated units.

In the context of single treated unit, PSC sets $s = 0$ and $w \in \mathbb{W}$, and chooses the following w:

$$w_{psc} := \underset{w \in \mathbb{W}}{\operatorname{argmin}} \|Y_1 - Y(w)\|^2 + \lambda \sum_{i \geq 2} w_i \|Y_1 - Y_i\|^2, \tag{46}$$

where $\lambda \geq 0$ is a regularization parameter. This method includes CSC (a nearest-neighbor matching estimator) as a special case where $\lambda = 0$ ($\lambda \to \infty$), and regularizes not only the pre-intervention bias measured by $B_p^2(w)$ but also the interpolation bias measured by $\sum_{i \geq 2} w_i \|Y_1 - Y_i\|^2$ if $\lambda > 0$. This design is consistent with the suggestion of ADH (2010, 2015) about reducing the interpolation bias by matching the treated unit and the untreated units in a pairwise way; see also Kellogg et al. (2020) for a related model-average estimator. Importantly, (42) reduces to

$$w_{rsc}(0, 1, \kappa_1, 0) = \underset{w \in \mathbb{S}(1)}{\operatorname{argmin}} B_p^2(w) + \kappa_1 B_c^2(w) \tag{47}$$

if $\kappa_2 = 0$. By comparing (46) with (47), we observe that the RSC with $w = w_{rsc}(0, 1, \kappa_1, 0)$

amounts to a variant of PSC because $B_c^2(w)$ is quite similar to $\sum_{i\geq 2} w_i \|Y_1 - Y_i\|^2$ in measuring the pairwise-matching divergence.

Moreover, as implied by (27) and (32), an arbitrary SC method with $s = 0$ and $w \in \mathbb{W}$ has the variance of prediction errors: $\sigma_t^2(w, 0)$, which is bounded above by

$$\bar{\sigma}^2(w, 0) = \sigma_\varepsilon^2(1 + \|w\|^2). \tag{48}$$

By comparing (42) with (46), we observe that the RSC with $w = w_{rsc}(0, 1, \kappa_1, \kappa_2)$ regularizes not only the squared-bias measures: $B_p^2(w)$ and $B_c^2(w)$ but also the squared L_2-norm of w: $\|w\|^2$ in order to control for $\bar{\sigma}^2(w, 0)$. This is an essential feature of this particular RSC that is not shared by CSC and PSC.

In addition to the potential nonlinearity (interpolation) bias, CSC encounters the poor-matching problem if $B_p(w_{csc}) > 0$. This problem is not uncommon in practice. It appears when Y_1 is outside of the convex hull of $\boldsymbol{Y}_{(-1)} := (Y_2, \ldots, Y_{n+1})$. Intuitively, this problem is closely related to the fact that CSC is constrained by the setting: $s = 0$ and $w \in \mathbb{W}$. In comparison, RSC allows us to deal with this problem by relaxing the convex-combination restriction: $w \in \mathbb{W}$, by setting $s = 1$ with a suitable design of $\mathcal{M}(\theta)$, or by both.

3.2 Relaxation of convex combination

In the case where $s = 0$ and $\tau \geq 1$, $w_{rsc}(s, \tau, \kappa)$ reduces to

$$w_{rsc}(0, \tau, \kappa_1, \kappa_2) = \underset{w \in \mathbb{S}(\tau)}{\operatorname{argmin}} B_p^2(w) + \kappa_1 B_c^2(w) + \kappa_2 \|w\|^2. \tag{49}$$

Note that (49) includes

$$w_{rsc}(0, \tau, 0, 0) = \underset{w \in \mathbb{S}(\tau)}{\operatorname{argmin}} B_p^2(w) \tag{50}$$

as a special case where $\kappa_1 = \kappa_2 = 0$. The RSC with $w = w_{rsc}(0, \tau, 0, 0)$ generalizes CSC by relaxing the convex-combination restriction: $w \in \mathbb{W}$. Conceptually, this is useful for remedying the poor-matching problem of CSC because the relationship: $\mathbb{W} \subset \mathbb{S}_+(\tau)$ implies $B_p^2(w_{rsc}(0, \tau, 0, 0)) \leq B_p^2(w_{csc})$ when $\tau \geq 1$. It is essential to observe that the choice of $w = w_{rsc}(0, \tau, 0, 0)$ might be interpreted as a non-negative Lasso estimator for

the coefficient vector of the regression: Y_1 on $\boldsymbol{Y}_{(-1)}$:

$$w_{lasso(+)} := \underset{w \in \mathbb{R}^n_+}{\operatorname{argmin}} \|Y_1 - Y(w)\|^2 + \lambda \|w\|_1, \tag{51}$$

where $\lambda \geq 0$ is a penalization parameter that corresponds to τ. The choice of $w = w_{rsc}(0, \tau, 0, \kappa_2)$ further extends the former by accounting for the regularization of $\|w\|^2$. This interpretation illustrates that the RSC with $w = w_{rsc}(0, \tau, 0, \kappa_2)$ is a variant of the penalized-regression approach considered by the SC literature.

Specifically, Valero (2015) proposed replacing w_{csc} by estimating w using the Lasso regression. Doudchenko and Imbens (2017) proposed estimating w using the elastic-net regression, which encompasses the Lasso regression and the ridge regression that penalize $\|w\|_1$ and $\|w\|^2$, respectively. See also Amjad et al. (2018), Li (2020), Hollingsworth and Wing (2020), Ben-Michael et al. (2021) and Chernozhukov et al. (2021), among others, for related studies. The penalized-regression approach is in between CSC and the least squares (LS) method that minimizes $B_p^2(w)$ without imposing any restriction on w, but the LS method is infeasible if $T_0 < n$. In comparison, the RSC with $w = w_{rsc}(0, \tau, 0, \kappa_2)$ relaxes the size restriction: $\tau = 1$ but maintains the sign restriction: $w \in \mathbb{R}^n_+$ for the reasons explained in Section 2.3.

Importantly, by Proposition 1, we also observe that refining the poor-matching problem by relaxing the restriction: $w \in \mathbb{W}$ is not without costs. As implied by (31), an arbitrary SC method that sets $s = 0$ but relaxes the restriction: $w \in \mathbb{W}$ has the bias bound:

$$\overline{Bias}(w, 0) = \xi_{\psi,1} \left((1 + \|w\|_1) B_p(w) + B_c(w) \right) + \xi_{\psi,0} B_a(w) \tag{52}$$

and the variance bound $\bar{\sigma}^2(w, 0)$ shown in (48). The bias bound in (52) tends to be larger than that in (44) because it includes two additional components: $\xi_{\psi,1} \|w\|_1 B_p(w)$ and $\xi_{\psi,0} B_a(w)$. Meanwhile, $B_c(w)$ and $\bar{\sigma}^2(w, 0)$ might also increase because $\|w\|_1$ and $\|w\|^2$ tend to increase after the relaxation. In comparison, the RSC with $w = w_{rsc}(0, \tau, \kappa_1, \kappa_2)$ accounts for the potential costs of relaxing the size restriction: $\tau = 1$ by regularizing not only $B_p^2(w)$ but also $B_c^2(w)$ and $\|w\|^2$.

3.3 Model and bias correction

In the case where $s = 1$ and $\tau = 1$, $w_{rsc}(s, \tau, \kappa)$ degenerates to

$$w_{rsc}(1, 1, \kappa_1, \kappa_2, \kappa_3) = \underset{w \in \mathbb{S}(1)}{\operatorname{argmin}} \hat{B}_p^2(w, 1) + \kappa_1 B_c^2(w) + \kappa_2 \|w\|^2 + \kappa_3 B_p^2(w), \tag{53}$$

which includes

$$w_{rsc}(1, 1, 0, 0, 0) = \underset{w \in \mathbb{S}(1)}{\operatorname{argmin}} \hat{B}_p^2(w, 1) \tag{54}$$

as a special case where $\kappa_1 = \kappa_2 = \kappa_3 = 0$. The particular RSC with $w = w_{rsc}(1, 1, 0, 0, 0)$ generalizes CSC by substituting condition (15) for condition (11) and using $\mathcal{M}(\theta)$ to refine the potential poor-matching problem. This RSC includes MSC as a special case.

Specifically, MSC is established by extending ADH's (2010) linear factor model using a vector of observed factors, denoted as x_t here, to control for the poor-matching problem of CSC. Denote the *pre-intervention* LS estimator:

$$\hat{\theta}_i = \left[\sum_{t=1}^{T_0} x_t x_t^\top\right] \left[\sum_{t=1}^{T_0} x_t y_{it}\right],$$

and the residual for all (i, t)'s:

$$\tilde{y}_{it} := y_{it} - x_t^\top \hat{\theta}_i.$$

Note that $\tilde{y}_{it}$ is conceptually different from the residual of the regression: Y_i on z_i mentioned in Section 2 because, unlike x_t, z_i is considered for predicting y_{it} for $t \geq T_0 + 1$ rather than for explaining $Y_1 - Y(w)$. Denote $X := (x_1^\top, \ldots, x_{T_0}^\top)^\top$, $\tilde{Y}_i := (\tilde{y}_{i1}, \ldots, \tilde{y}_{iT_0})^\top$ and $\tilde{Y}(w) = \sum_{i \geq 2} w_i \tilde{Y}_i$. MSC chooses the following w:

$$w_{msc} := \underset{w \in \mathbb{W}}{\operatorname{argmin}} \|\tilde{Y}_1 - \tilde{Y}(w)\|^2. \tag{55}$$

Since

$$\begin{aligned} \tilde{Y}_1 - \tilde{Y}(w) &= \left(Y_1 - X\hat{\theta}_1\right) - \sum_{i \geq 2} w_i \left(Y_i - X\hat{\theta}_i\right) \\ &= Y_1 - Y(w) - \sum_{i \geq 2} w_i (X\hat{\theta}_1 - X\hat{\theta}_i), \end{aligned}$$

the objective function in (55) is the same as $\hat{B}_p^2(w,1) = \left\| Y_1 - Y(w) - M(w,\hat{\theta}) \right\|^2$ with $M(w,\hat{\theta}) = \sum_{i\geq 2} w_i (X\hat{\theta}_1 - X\hat{\theta}_i)$. Thus, the RSC with $w = w_{rsc}(1,1,0,0,0)$ includes MSC as a special case where $\mu_{it}(\theta_i) = x_t^\top \theta_i$. Note that MSC further reduces to the DSC of Ferman and Pinto (2019, Equation 7) when $x_t = 1$ and $\theta_i = \mu_i$. It also corresponds to the DSC of Doudchenko and Imbens (2017), which chooses $w = w_{csc(d)}$ as a part of the minimizer:

$$(w_{csc(d)}, \hat{\alpha}) := \underset{w\in\mathbb{W}}{\operatorname{argmin}} \left\| Y_1 - Y(w) - \alpha \cdot \iota_{T_0} \right\|^2, \tag{56}$$

with $\hat{\alpha}$ denoting an estimator for the intercept α; see also Valero (2015). The objective function in (56) is the same as $B_p^2(w,1)$ when $\mu_{it}(\theta_i) = \mu_i$ with $\mu_i := \mathbb{E}[y_{it}]$. This illustrates that DSC is useful for refining the poor-matching problem of CSC *if* the pre-intervention bias is due to the mean difference between Y_1 and $Y(w)$.

According to (2) and (8), an arbitrary SC method with $s = 1$ generates the following prediction of counterfactual:

$$\hat{y}_t(w,1) = y_t(w) + \left(\mu_{1t}(\hat{\theta}_1) - \sum_{i\geq 2} w_i \mu_{it}(\hat{\theta}_i) \right). \tag{57}$$

By introducing the choice of $w = w_{csc}$ in (57), we obtain that

$$\hat{y}_t(w_{csc},1) = y_t(w_{csc}) + \underbrace{\left(\mu_{1t}(\hat{\theta}_1) - \sum_{i\geq 2} w_{csc,i} \mu_{it}(\hat{\theta}_i) \right)}_{\text{bias-correction term}}, \tag{58}$$

where $w_{csc,i}$ is an element of w_{csc}. Importantly, although the predictions: $\hat{y}_t(w,1)$, with $w = w_{rsc}(1,1,0,0,0)$, and $\hat{y}_t(w_{csc},1)$ both set $s = 1$, they are generated from different choices of w; see also Chen (2020, p.511). In particular, $\hat{y}_t(w_{csc},1)$ amounts to a bias correction of $y_t(w_{csc})$ for the poor-matching problem of CSC. In comparison, ASC is motivated by (58), and interprets $\mu_{it}(\theta_i)$ as a linear model of $\psi_t(Y_i)$; see Ben-Michael et al. (2021, Equations 9 and 10). See also Abadie and L'Hour (2019, Equation 9) for a similar bias correction of PSC and Chernozhukov et al. (2019) and Arkhangelsky et al.(2021) for related bias-reduction methods. Specifically, ASC assumes that $\psi_t(Y_i)$ is a time-invariant linear function of Y_i, and estimates the coefficients using a ridge regression. Ben-Michael et al. (2021, Equation 18) showed that, by this assumption and relaxing the sign restriction: $w \in \mathbb{R}_+^n$, this ASC amounts to a SC method that sets $s = 0$ and chooses

the following w:

$$w_{asc} := \underset{w \in \mathbb{A}}{\operatorname{argmin}} \frac{1}{2\lambda} \|Y_1 - Y(w)\|^2 + \frac{1}{2} \|w - w_{csc}\|^2, \tag{59}$$

where $\|w - w_{csc}\|^2$ is interpreted as the "level of extrapolation," and the regularization parameter $\lambda > 0$ controls for the deviation of w from w_{csc}.

As implied by (31) and (32), an arbitrary SC method with $s = 1$ has the bias bound:

$$\overline{Bias}(w, 1) = \xi_{\psi,1} \left(B_p(w, 1) + B_p(w) + B_c(w) \right) + (1 + \xi_{\psi,1}) \xi_{m,0} \tag{60}$$

and the variance bound:

$$\bar{\sigma}^2(w, 1) = \sigma_\varepsilon^2 + (\sigma_\varepsilon^2 + \xi_{m,1}^2 \sigma_\theta^2) \|w\|^2, \tag{61}$$

respectively. By comparing (44) with (60) and comparing (48) with (61), we may observe that sets $s = 1$ and uses $\mathcal{M}(\theta)$ to refine the pre-intervention bias of CSC is at the cost of generating an additional bias component $(1 + \xi_{\psi,1})\xi_{m,0}$ and an additional variance component $\xi_{m,1}^2 \sigma_\theta^2 \|w\|^2$. The result in (60) also reminds us that the "bias-corrected" prediction in (58) is not necessarily unbiased. Compared to the aforementioned model-based methods, the RSC with $w = w_{rsc}(1, 1, \kappa_1, \kappa_2, \kappa_3)$ regularizes not only $\hat{B}_p^2(w, 1)$ and $\|w\|^2$ but also $B_c^2(w)$.

In addition, RSC is also applicable to the case where $s = 1$ and $\tau \geq 1$. In this case, RSC refines the poor-matching problem by relaxing the convex-combination restriction and using the auxiliary model $\mathcal{M}(\theta)$ simultaneously. In particular, the RSC with $w = w_{rsc}(1, \tau, 0, 0, 0)$ extends the non-negative Lasso estimator or MSC by choosing the following w:

$$w_{lasso(+)} := \underset{w \in \mathbb{R}_+^n}{\operatorname{argmin}} \|\tilde{Y}_1 - \tilde{Y}(w)\|^2 + \lambda \|w\|_1, \tag{62}$$

where $\lambda \geq 0$ corresponds to τ. In the case where $\mu_{it} = \mu_i$, this particular RSC corresponds to a modified SC method considered by Li (2020, Equation 4).

3.4 Selection of regularization parameters

Like the penalized regressions, PSC, ASC and RSC all involve certain regularization parameters to be selected in their general forms. In the literature, it is common to select

the regularization parameter(s) of a SC method by minimizing a validation criterion. For the RSC method, we let $V(s,\tau,\kappa)$ be such a validation criterion which is dependent on the choice of (s,τ,κ). Following Abadie and L'Hour (2019), we set $V(s,\tau,\kappa)$ to be the empirical MSPE of the RSC method in a validation period before the intervention; see also ADH (2015), Amjad et al.(2018, p.10) and Abadie (2021, p.397) for this setting. Specifically, we split the pre-intervention period into the training period: $t \in [1,R]$ and the validation period: $t \in [R+1,T_0]$. In addition, we let $\boldsymbol{Y}_R$ be the first $R \times (n+1)$ submatrix of $\boldsymbol{Y}$ that comprises the pre-intervention outcomes in the training period, and $\hat{\theta}_R$ and $w_{R,rsc}(s,\tau,\kappa)$ be, respectively, the counterparts of $\hat{\theta}$ and $w_{rsc}(s,\tau,\kappa)$ obtained by replacing $\boldsymbol{Y}_R$ with $\boldsymbol{Y}$. Accordingly, we set $V(s,\tau,\kappa)$ to be the empirical MSPE:

$$V(s,\tau,\kappa) = \frac{1}{T_0 - R + 1} \sum_{t=R+1}^{T_0} \left(y_{1t} - \hat{y}_t(w,s)\right)^2, \tag{63}$$

with $\hat{y}_t(w,s) = y_t(w) + s \cdot m_t(w,\hat{\theta}_R)$ and $w = w_{R,rsc}(s,\tau,\kappa)$, and select (s,τ,κ) as the following minimizer:

$$(s^*,\tau^*,\kappa^*) = \underset{s,\tau,\kappa}{\operatorname{argmin}}\, V(s,\tau,\kappa). \tag{64}$$

For particular RSC methods, PSC and ASC, we also select the regularization parameters by minimizing the associated empirical MSPEs in a similar way. In applications, we set R to be the integer part of $\frac{2}{3}T_0$ for simplicity. Theoretically, one might also replace the aforementioned validation criterion by a cross-validation criterion. However, the computational cost of cross validation could be prohibitive for the most general form of RSC.

4 Monte Carlo Simulation

In this simulation, we consider the following data generating processes (DGP):

$$y_{it}^* = h_{it} + \lambda_i^\top f_t + \delta_t D_{it} + \varepsilon_{it},$$

where $h_{1t} = a + bt$, $h_{it} = 0$ for $i \geq 2$, $\lambda_i = (\lambda_{1i}, \lambda_{2i})^\top$, $f_t = (1,t)^\top$, $\delta_t = 0$ if $t \leq T_0$, $\delta_t = t$ if $t \geq T_0 + 1$, and $\lambda_{1i} \sim N(0,1)$ $\lambda_{2i} \sim U(0,0.2)$ and $\varepsilon_{it} \sim N(0,\sigma^2)$ are independently and identically distributed random variables, and generate y_{it} from the transformation: $y_{it} = y_{it}^*$ if $c = 1$, and $y_{it} = |y_{it}^*|^{1.5}$ if $c = 2$. To generate different types of data, we consider the following settings of (a,b,c,σ^2):

- DGP1: $(a, b, c, \sigma^2) = (0, 0, 1, 1)$;
- DGP2: $(a, b, c, \sigma^2) = (5, 0, 1, 1)$;
- DGP3: $(a, b, c, \sigma^2) = (5, 0.2, 1, 1)$.

Note that the designed context of CSC holds under DGP1, but fails under DGP2 (DGP3) because of the poor-matching problem caused by the mean (-and-trend) difference between the treated unit and the untreated units. We also let DGP4-6 be the nonlinear counterparts of DGP1-3, respectively, that are defined by replacing $c = 1$ with $c = 2$, and DGP7-12 be the counterparts of DGP1-6, respectively, that are defined by replacing $\sigma^2 = 1$ with $\sigma^2 = 2$.

In addition, we consider the existing SC methods: CSC, DSC, MSC, ASC and PSC and the following RSCs:

- RSC_p: $w = w_{rsc}(s, \tau, \kappa_1, \kappa_2) = w_{rsc}(0, 1, 0, 0)$,
- RSC_c: $w = w_{rsc}(s, \tau, \kappa_1, \kappa_2) = w_{rsc}(0, 1, \kappa_1, 0)$,
- RSC_w: $w = w_{rsc}(s, \tau, \kappa_1, \kappa_2) = w_{rsc}(0, 1, 0, \kappa_2)$,
- RSC_g: $w = w_{rsc}(s, \tau, \kappa_1, \kappa_2) = w_{rsc}(0, \tau, \kappa_1, \kappa_2)$,
- $RSC_p(D)$: $w = w_{rsc}(s, \tau, \kappa_1, \kappa_2, \kappa_3) = w_{rsc}(1, 1, 0, 0, 0)$,
- $RSC_c(D)$: $w = w_{rsc}(s, \tau, \kappa_1, \kappa_2, \kappa_3) = w_{rsc}(1, 1, \kappa_1, 0, 0)$,
- $RSC_w(D)$: $w = w_{rsc}(s, \tau, \kappa_1, \kappa_2, \kappa_3) = w_{rsc}(1, 1, 0, \kappa_2, 0)$,
- $RSC_g(D)$: $w = w_{rsc}(s, \tau, \kappa_1, \kappa_2, \kappa_3) = w_{rsc}(1, \tau, \kappa_1, \kappa_2, \kappa_3)$,
- $RSC_p(M)$: $w = w_{rsc}(s, \tau, \kappa_1, \kappa_2, \kappa_3) = w_{rsc}(1, 1, 0, 0, 0)$,
- $RSC_c(M)$: $w = w_{rsc}(s, \tau, \kappa_1, \kappa_2, \kappa_3) = w_{rsc}(1, 1, \kappa_1, 0, 0)$,
- $RSC_w(M)$: $w = w_{rsc}(s, \tau, \kappa_1, \kappa_2, \kappa_3) = w_{rsc}(1, 1, 0, \kappa_2, 0)$,
- $RSC_g(M)$: $w = w_{rsc}(s, \tau, \kappa_1, \kappa_2, \kappa_3) = w_{rsc}(1, \tau, \kappa_1, \kappa_2, \kappa_3)$.

Among these SC methods, RSC_p, $RSC_p(D)$ and $RSC_p(M)$ correspond to CSC, DSC and MSC, respectively, and RSC_c is a variant of PSC. Given $s = 0$, the subscript "p" of RSC_p means that this RSC regularizes $B_p(w)$, the subscript "c" of RSC_c means that

this RSC regularizes not only $B_p(w)$ but also $B_c(w)$, the subscript "w" of RSC_w means that this RSC regularizes not only $B_p(w)$ but also $\|w\|^2$, and the subscript "g" of RSC_g means that RSC_g is the general form of RSC that regularizes $B_p^2(w)$, $B_c^2(w)$ and $\|w\|^2$ by allowing $\tau \geq 1$. Given $s = 1$, $RSC_c(D)$, $RSC_w(D)$ and $RSC_g(D)$ are, respectively, the counterparts of RSC_c, RSC_w and RSC_g that use $\mathcal{M}(\theta)$ with $x_t = 1$, and $RSC_c(M)$, $RSC_w(M)$ and $RSC_g(M)$ are, respectively, the counterparts of RSC_c, RSC_w and RSC_g that use $\mathcal{M}(\theta)$ with $x_t = (1, t)^\top$. Correspondingly, DSC and MSC use the same $\mathcal{M}(\theta)$'s as $RSC_p(D)$ and $RSC_p(M)$, respectively. In this simulation, we set $\lambda = \kappa_1$ for PSC, and multiply the objective function in (59) by 2λ, with $\lambda = \kappa_2$, for ASC, and select the regularization parameters of RSC using the validation method discussed in Section 3.4 based on the settings: $\tau = 1, 1.1, \ldots, 1.5$, $\kappa_1 = 0, 1, \ldots, 5$, $\kappa_2 = 0, 100, 200, \ldots, 5000$ and $\kappa_3 = 0, 1, \ldots, 5$.

Let $\hat{y}_{1t}^{(b)}(w, s)$ be the $\hat{y}_{1t}(w, s)$ of a SC method generated by the bth replication of the simulation for $b = 1, 2, \ldots, B$ with B denoting the number of replications. We measure the finite-sample performance of a SC method using the average of the absolute biases:

$$|bias| := \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} |bias_{B,t}|,$$

where $bias_{B,t} := \hat{\delta}_{B,t} - \delta_t$, $\hat{\delta}_{B,t} := B^{-1} \sum_{b=1}^{B} \hat{\delta}_t^{(b)}$ and $\hat{\delta}_t^{(b)} := \hat{y}_{1t}^{(b)}(w) - y_{1t}(0)$, and the average of the root mean squared errors (RMSEs):

$$RMSE := \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} RMSE_{B,t},$$

where $RMSE_{B,t} := \sqrt{B^{-1} \sum_{b=1}^{B} (\hat{\delta}_t^{(b)} - \delta_t)^2}$. In Table 1, we report these two performance measures of the SC methods for the settings: $n = 50$, $T = T_0 + 10$, $T_0 = 50$ or 100, and $B = 1000$. The main simulation findings are summarized as follows.

Firstly, CSC performs quite well in its designed context (DGP1 or DGP7), but the performance of CSC is worsened beyond its designed context (under the other DGPs). More specifically, given $T_0 = 50$, CSC has $|bias| = 0.03$ and $RMSE = 1.203$ under DGP1 that increase to 0.182 and 2.531 (9.905 and 10.578) under DGP2 (DGP3), respectively. The $|bias|$ of CSC further increases to 0.213, 1.812 and 61.038 under DGP4-6, respectively, and the $RMSE$ of CSC increase to 4.195, 12.386 and 66.009 under these DGPs. This shows

Table 1: Averages of the absolute biases and the RMSEs of the SC methods.

			existing SC methods					$s=0$				$s=1(D)$				$s=1(M)$			
	T_0	DGP	CSC	DSC	MSC	ASC	PSC	RSC_p	RSC_c	RSC_w	RSC_g	RSC_p	RSC_c	RSC_w	RSC_g	RSC_p	RSC_c	RSC_w	RSC_g
		1	0.030	0.033	0.029*	0.025**	0.031	0.030*	0.112	5.527	0.042	0.028	0.195	0.111	0.026*	0.029	0.145	0.098	0.025**
		2	0.182	0.033	0.029**	0.957	0.167	0.182*	0.391	10.527	0.832	0.032	0.102	0.031*	0.429	0.029**	0.091	0.098	0.120
		3	9.905	3.347	0.029*	0.299	9.904	9.905	9.946	21.627	4.375*	3.347	3.348	3.367	1.086*	0.029	0.027**	0.071	0.405
		4	0.213	0.126	0.096	0.089**	0.337	0.236*	0.326	15.178	0.433	0.138*	0.846	0.433	0.347	0.093*	0.715	0.269	0.133
		5	1.812	0.195*	0.927	3.809	1.741	1.812*	2.089	35.593	2.203	0.195*	0.911	0.362	1.519	0.926	0.128**	0.748	0.589
		6	61.038	30.508	3.533*	3.754	61.031	61.038	61.121	101.618	41.244*	30.508	30.510	30.615	2.093*	3.533	3.535	3.709	1.661**
	50	7	0.038	0.044	0.041	0.035**	0.038	0.045*	0.179	5.529	0.070	0.038	0.278	0.174	0.037*	0.041*	0.235	0.144	0.041*
		8	0.227	0.044	0.041*	1.157	0.201	0.227*	0.468	10.529	0.819	0.043	0.167	0.035**	0.447	0.041*	0.153	0.115	0.149
		9	9.936	3.489	0.041**	0.390	9.930	9.936	9.989	21.629	4.540*	3.489	3.489	3.518	1.095*	0.041**	0.042	0.081	0.469
		10	0.252	0.162	0.137	0.135**	0.527	0.280*	0.392	15.413	0.637	0.175*	1.148	0.548	0.463	0.135**	0.978	0.298	0.231
		11	2.086	0.423*	1.098	5.810	1.996	2.086*	2.442	35.729	30.217	0.423*	1.350	0.640	1.535	1.098	0.230**	0.906	0.748
absolute		12	61.137	31.212	3.683*	4.211	61.110	61.137*	61.270	101.717	91.361	31.212	31.212	31.342	2.170*	3.683	3.745	3.845	1.588**
bias		1	0.047	0.048	0.046**	0.052	0.052	0.067*	0.077	10.718	0.075	0.061*	0.138	0.107	0.062	0.046**	0.072	0.077	0.060
		2	0.497	0.048	0.046**	0.414	0.520	0.496	0.441	15.718	0.294*	0.050*	0.144	0.051	0.080	0.046**	0.067	0.071	0.061
		3	15.559	5.914	0.046**	0.391	15.555	15.559	15.597	36.818	5.752*	5.914	5.914	5.921	1.823*	0.046**	0.046**	0.057	0.117
		4	0.394	0.352	0.325	0.264**	0.406	0.405*	0.419	39.992	0.561	0.374*	0.832	0.478	0.510	0.338*	0.973	0.424	0.360
		5	0.854	0.407	0.366*	1.834	0.985	0.854	0.809*	66.178	0.819	0.406	1.187	0.366*	0.434	0.365**	0.867	0.396	0.675
		6	127.475	67.837	8.437	2.869*	127.474	127.475	127.501	225.916	79.398*	67.837	67.837	67.903	2.225**	8.437	8.437	8.499	8.037*
	100	7	0.067	0.068	0.065**	0.073	0.071	0.097*	0.123	10.720	0.117	0.090	0.209	0.157	0.086*	0.065**	0.105	0.108	0.088
		8	0.507	0.068	0.065**	0.651	0.500	0.506	0.435	15.720	0.323*	0.072*	0.257	0.073	0.086	0.065**	0.109	0.092	0.084
		9	15.582	6.011	0.065**	0.428	15.581	15.582	15.628	36.820	5.856*	6.011	6.011	6.023	1.816*	0.065**	0.065**	0.075	0.189
		10	0.546	0.502	0.413	0.380**	0.630	0.554*	0.570	40.167	0.805	0.535*	1.162	0.671	0.692	0.428*	1.357	0.531	0.522
		11	0.892	0.405*	0.641	3.788	0.931	0.892	0.845*	66.292	55.622	0.404**	1.794	0.404**	0.504	0.639*	1.117	0.653	0.857
		12	127.619	68.325	9.185	5.689*	127.601	127.619*	127.656	225.994	209.780	68.325	68.325	68.406	2.293**	9.185	9.185	9.253	6.291*
		1	1.203	1.164	1.122*	1.165	1.261	1.182	1.149	6.530	1.131*	1.156	1.206	1.122*	1.136	1.122	1.198	1.091**	1.161
		2	2.531	1.164	1.122*	19.021	2.581	2.531	2.380	11.087	1.560*	1.163	1.250	1.155*	1.349	1.122	1.185	1.087**	1.270
		3	10.578	3.914	1.122*	8.272	10.579	10.578	10.614	21.905	5.683*	3.914	3.914	3.926	1.765*	1.122	1.138	1.081**	1.508
		4	4.195	4.120	4.101*	4.146	4.384	4.186	4.115	19.196	4.074*	4.116	4.322	4.016*	4.094	4.100	4.319	3.954**	4.155
		5	12.386	6.156	5.517*	9.012	12.531	12.386	12.209	39.342	7.789*	6.156	6.415	6.058*	6.479	5.517	5.856	5.321**	6.119
		6	66.009	34.506	8.594*	12.634	66.012	66.009	66.064	104.460	48.540*	34.506*	34.506*	34.570	11.733	8.594	8.599	8.575**	10.340
	50	7	1.670	1.643	1.586*	1.635	1.743	1.645	1.594	6.610	1.581*	1.633	1.682	1.577*	1.595	1.586	1.694	1.545**	1.640
		8	2.813	1.643	1.586*	2.775	2.853	2.813	2.668	11.135	1.932*	1.641	1.708	1.621*	1.775	1.586	1.675	1.532**	1.755
		9	10.691	4.175	1.586*	15.686	10.692	10.691	10.736	21.930	5.951*	4.175	4.175	4.189	2.175*	1.586	1.616	1.527**	1.967
		10	5.820	5.781	5.720*	5.804	6.071	5.806	5.697*	19.645	5.700	5.776	5.986	5.628*	5.734	5.720	5.980	5.521**	5.866
		11	13.774	8.350	7.714*	11.491	13.944	13.774	13.579*	39.757	33.603	8.350	8.583	8.221*	8.690	7.714	8.151	7.476**	8.547
		12	66.643	35.959	11.482*	15.857	66.645	66.643*	66.731	104.790	94.448	35.959	35.959	36.013	15.000*	11.482	11.564	11.398**	13.676
RMSE		1	1.148	1.097	1.071*	1.096	1.189	1.121	1.101	12.433	1.081*	1.086	1.126	1.066*	1.080	1.071	1.097	1.046**	1.077
		2	2.488	1.097	1.071*	1.423	2.579	2.487	2.430	16.934	1.384*	1.095	1.192	1.078*	1.256	1.071	1.098	1.040**	1.130
		3	16.838	6.806	1.071*	1.597	16.837	16.838	16.874	37.353	8.058*	6.806	6.806	6.811	2.418*	1.071	1.071	1.037**	1.243
		4	5.588	5.444	5.378*	5.487	5.805	5.579	5.506	49.953	5.395*	5.440	5.663	5.350*	5.430	5.374	5.662	5.279**	5.456
		5	14.648	7.334	6.404*	9.390	15.060	14.648	14.586	75.784	8.596*	7.333	8.003	7.233*	7.572	6.404	6.873	6.310**	6.830
		6	140.031	76.647	13.503*	14.898	140.036	140.031	140.042	233.030	97.862*	76.647	76.647	76.692	19.018*	13.503	13.503	13.492**	16.171
	100	7	1.590	1.549	1.515*	1.550	1.639	1.555	1.526	12.475	1.518*	1.533	1.588	1.503*	1.517	1.515	1.551	1.482**	1.526
		8	2.762	1.549	1.515*	2.006	2.832	2.761	2.697	16.965	1.762*	1.545	1.669	1.522*	1.676	1.515	1.564	1.472**	1.585
		9	16.915	6.968	1.515*	2.184	16.916	16.915	16.954	37.368	8.232*	6.968	6.968	6.975	2.726*	1.515	1.515	1.468**	1.745
		10	7.754	7.642	7.562*	7.784	8.030	7.741	7.635	50.284	7.573*	7.636	7.883	7.517*	7.626	7.557	7.855	7.417**	7.660
		11	16.317	9.865	9.031*	12.982	16.679	16.317	16.229*	76.105	63.940	9.864	10.468	9.743*	10.199	9.031	9.561	8.891**	9.506
		12	140.604	77.724	16.913*	20.563	140.603	140.604*	140.612	233.284	216.481	77.724*	77.724*	77.762	22.775	16.913	16.913	16.865**	19.632

Note: For RSC, "$s=0$" includes RSC_p, RSC_c, RSC_w and RSC_g; "$s=1(D)$" includes $RSC_p(D)$, $RSC_c(D)$, $RSC_w(D)$ and $RSC_g(D)$; "$s=1(M)$" includes $RSC_p(M)$, $RSC_c(M)$, $RSC_w(M)$ and $RSC_g(M)$. For each pair of T_0 and DGP, "*" means the minimum value of the absolute bias or the RMSE among the existing SC methods, the RSCs with $s=0$, the RSCs with $s=1(D)$ or the RSCs with $s=1(M)$, and "**" means the minimum values of the absolute bias or the RMSE among all the SC methods.

that the performance of CSC is damaged by the poor-matching problem which is due to the mean (and trend) difference under DGP2 (DGP3) and further complicated by the nonlinearity under DGP4-6. This result is even more evident when $T_0 = 100$. In addition, the performance of CSC is slightly worsened by the increase of σ^2.

Secondly, focusing on the existing methods, DSC has similar performance to CSC under DGP1 (or DGP7), but outperforms CSC under the other DGPs. Given $T_0 = 50$, DSC has $|bias| = 0.033$ (0.044) and $RMSE = 1.164$ (1.586) under DGP1-2 (DGP7-8). This shows that DSC is useful for refining the poor-matching problem of CSC caused by the mean difference. However, DSC is generally outperformed by MSC and ASC except for DGP5 that comprises the mean difference and the nonlinearity. Given $T_0 = 50$, compared to DSC that has $|bias| = 3.347$ (30.508) and $RMSE = 3.914$ (34.506) under DGP3 (DGP6) that comprises the mean-and-trend difference, MSC has $|bias| = 0.029$ (3.533) and $RMSE = 1.122$ (8.594) and that ASC has $|bias| = 0.209$ (3.754) and $RMSE = 8.272$ (12.634) under the same DGP. Meanwhile, the $|bias|$ and $RMSE$ of MSC are invariant under DGP1-3 (DGP7-9). This illustrates that MSC outperforms DSC because it uses a more flexible $\mathcal{M}(\theta)$ to deal with the poor-matching problem of CSC. We also observe that MSC has smaller $|bias|$'s relative to ASC in several cases as $T_0 = 50$, but the results are reversed as $T_0 = 100$. In addition, MSC has smaller $RMSE$'s relative to ASC for all cases considered. In comparison, PSC has similar performance to CSC in our simulation.

Thirdly, as expected, RSC_p, $RSC_p(D)$ and $RSC_p(M)$ are essentially identical to CSC, DSC and MSC in terms of their performance, and RSC_c has very similar performance to PSC for all DGPs. This is consistent with the theoretical relationships among these SC methods, and shows that RSC is useful for unifying existing methods. In this unified framework, $RSC_p(D)$ and $RSC_p(M)$ generalize CSC using auxiliary models, and RSC_g generalizes CSC by setting $s = 0$ but relaxing the restriction: $\tau = 1$. The simulation shows that $RSC_p(D)$ and $RSC_p(M)$ considerably outperform RSC_g in most cases, and suggests that it is better to refine the potential biases of CSC using suitable auxiliary models rather than simply relaxing the aggregation restriction.

Fourthly, RSC_g substantially outperforms DSC and MSC in terms of $|bias|$ under DGP6 (DGP12) which comprises a complicated divergence from the designed context of CSC. Under DGP6 (DGP12), DSC and MSC are of the $|bias|$'s: 30.508 (31.212) and 3.533 (3.683), respectively, and $RSC_g(D)$ and $RSC_g(M)$ are, respectively, of much smaller $|bias|$'s: 2.093 (2.170) and 1.661 (1.588). It is useful to observe that DSC has the same performance as $RSC_p(D)$, and $RSC_g(D)$ shares the same $\mathcal{M}(\theta)$ as $RSC_p(D)$. Thus,

the aforementioned result indeed reflects that $RSC_g(D)$ remedies the bias of DSC under DGP6 (DGP12) by suitably selecting the regularization parameters. The same interpretation also applies to the relative performance of the MSC method and the $RSC_g(M)$ method. Moreover, we observe that $RSC_w(M)$ uniformly outperforms the other SC methods in terms of $RMSE$ for all DGPs and for both T_0's considered in this simulation.

Generally speaking, these results show that RSC does not only encompass certain existing SC methods but also generate useful complements to existing methods. In particular, $RSC_w(M)$ compares favourably with the other SC methods in terms of $RMSE$. The design of $RSC_w(M)$ reflects the importance of first dealing with the pre-intervention bias of CSC using a relatively flexible $\mathcal{M}(\theta)$ and then suitably regularizing $\|w\|^2$. Note that the regularization of $\|w\|^2$ is beyond the consideration of MSC, and is useful for controlling the potential cost of using $\mathcal{M}(\theta)$ discussed in Section 3.3.

5 Empirical Illustration

In this section, we further compare the performance of different SC methods using two case studies. The first one is the case study considered by ADH (2010), and the second one is a hypothetical case study based on the former. The motivation and design of the hypothetical case study will be explained later.

5.1 Actual case study

In the case study of ADH (2010), the outcome variable y_{it} is the cigarette sales of a state in the tth year of the sampling period: 1970-2000, the intervention variable D_{it} is defined by California's Tobacco Control Program (CTCP) that took place in 1989 ($T_0 = 20$), California is the treated unit ($i = 1$), and the donor pool of untreated units comprises 37 states ($n = 37$) that did not have a similar program during the sampling period; see Figure 1 for the outcome sequences $\{y_{it}\}$'s. ADH (2010) found that, with the use of certain pre-intervention covariates, the cigarette sales of the synthetic California, which is defined by the convex combination of the Y_i's, for $i \geq 2$, based on CSC, closely matches Y_1 in the pre-intervention period: 1970-1988, and the counterfactual cigarette sales predicted by CSC are considerably lower than the y_{1t}'s in the post-intervention period: 1989-2000. This shows the effectiveness of CTCP for reducing California's cigarette sales.

In the following, we apply not only CSC but also the other SC methods considered in Section 4 to estimating the intervention effects of CTCP. We also utilize this case study to

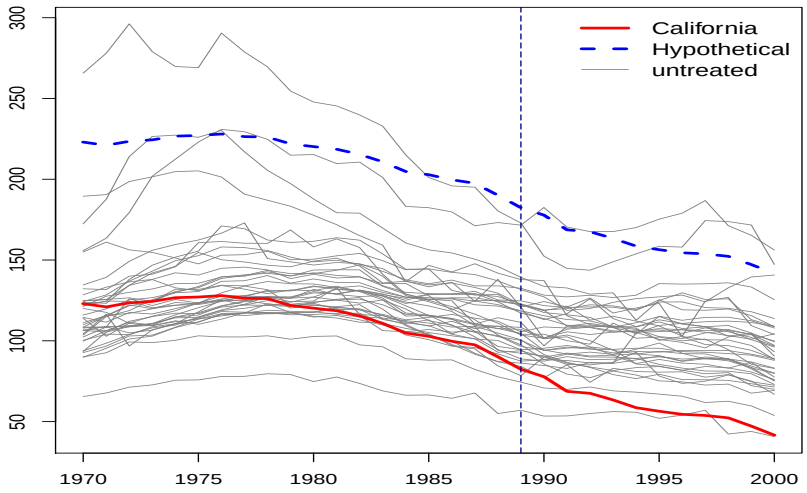

Figure 1: The outcome sequences of California (red), the hypothetical treated unit (blue) and the untreated units (gray).

assess the relative performance of the SC methods in real data. For the SC methods with regularization parameters, the parameters are selected using the same validation method and settings as the simulation, and the validation method is implemented by splitting the pre-intervention period into the training period: 1970-1981 ($R = 12$) and the validation period: 1982-1988. Note that the general form of RSC involves 1836 different settings of $(\tau, \kappa_1, \kappa_2)$ when $s = 0$ and 11016 different settings of $(\tau, \kappa_1, \kappa_2, \kappa_3)$ when $s = 1$. In Table 2, we report the minimum, the maximum and the deciles of the validation criterion values and the associated regularization parameters among these settings. The minimum corresponds to RSC_g if $s = 0$ and $RSC_g(D)$ or $RSC_g(M)$ if $s = 1$. The criterion values and the regularization parameters of RSC_p, RSC_c, RSC_w, $RSC_p(D)$, $RSC_c(D)$, $RSC_w(D)$, $RSC_p(M)$, $RSC_c(M)$ and $RSC_w(M)$ are reported in the same table. In Figure 2, we compare the outcome sequence of California $\{y_{1t}\}_{t=1}^{T}$ with the synthetic outcome sequences $\{\hat{y}_{1t}(w,s)\}_{t=1}^{T}$'s generated by the SC methods considered in the simulation. In Table 3, we further report the pre-intervention RMSEs:

$$RMSE_{pre} := \sqrt{\frac{1}{T_0}\sum_{t=1}^{T_0}\left(y_{1t} - \hat{y}_{1t}(w,s)\right)^2} = T_0^{-1/2}\hat{B}_p(w,s)$$

Table 2: Validation criterion values and regularization parameters of RSC.

	$s=0$				$s=1(D)$					$s=1(M)$				
	V	τ	κ_1	κ_2	V	τ	κ_1	κ_2	κ_3	V	τ	κ_1	κ_2	κ_3
(a) Actual														
10%	21.221	1.0	4	700	24.675	1.2	0	1900	3	2.612	1.2	0	1000	2
20%	23.771	1.4	5	0	32.259	1.2	0	3000	2	5.996	1.1	2	4600	3
30%	36.594	1.1	3	700	38.671	1.0	0	2600	1	8.117	1.4	2	2900	1
40%	47.547	1.0	5	3400	45.368	1.2	0	4300	1	9.368	1.4	3	4300	4
50%	50.733	1.5	5	4100	50.159	1.3	2	2400	5	10.884	1.2	3	3200	2
60%	53.505	1.2	5	5000	53.191	1.2	4	4800	4	12.194	1.5	5	4200	4
70%	58.456	1.0	3	3000	56.080	1.2	2	2900	3	14.455	1.2	3	2500	4
80%	65.352	1.3	2	2800	58.665	1.5	5	2600	2	21.337	1.4	5	1900	5
90%	76.222	1.1	2	4500	64.330	1.1	1	4100	5	40.498	1.1	0	3600	0
max	94.271	1.5	1	5000	97.664	1.5	1	5000	0	75.470	1.5	5	300	0
RSC_p	31.376	1.0	0	0	13.545	1.0	0	0	0	28.423	1.0	0	0	0
RSC_c	18.913	1.0	2	0	13.545	1.0	0	0	0	28.423	1.0	0	0	0
RSC_w	20.168	1.0	0	700	5.361	1.0	0	400	0	28.423	1.0	0	0	0
RSC_g	14.190	1.0	5	600	5.361	1.0	0	400	0	0.193	1.0	0	2400	1
(b) Hypothetical														
10%	246.861	1.2	0	2900	391.730	1.5	0	3900	3	266.047	1.5	0	900	5
20%	834.537	1.1	1	3800	435.769	1.2	0	1300	1	306.520	1.0	0	2900	1
30%	995.658	1.4	1	800	488.354	1.0	0	700	5	386.375	1.3	4	3800	2
40%	1205.623	1.4	2	2500	542.671	1.0	1	5000	1	461.029	1.0	5	700	5
50%	1399.580	1.5	3	4400	562.003	1.1	3	300	3	496.031	1.1	1	3800	4
60%	1525.764	1.5	3	3300	574.723	1.4	2	2200	5	524.667	1.1	2	3300	5
70%	1704.009	1.0	2	1200	601.841	1.2	4	4000	1	563.594	1.5	3	5000	5
80%	1842.213	1.0	3	3600	658.143	1.1	3	400	5	592.009	1.1	3	3700	4
90%	1903.964	1.0	3	1000	842.679	1.1	2	2900	0	855.081	1.1	1	4900	0
max	1955.163	1.5	5	5000	469.551	1.5	5	5000	0	1211.832	1.5	5	5000	0
RSC_p	552.156	1.0	0	0	13.542	1.0	0	0	0	30.716	1.0	0	0	0
RSC_c	552.156	1.0	0	0	13.542	1.0	0	0	0	30.716	1.0	0	0	0
RSC_w	519.274	1.0	0	100	5.397	1.0	0	400	0	30.716	1.0	0	0	0
RSC_g	32.707	1.5	0	0	5.397	1.0	0	400	0	30.714	1.1	0	0	0

Note: "Actual" and "Hypothetical" represent the actual case study and the hypothetical case study, respectively. "V" represents the validation criterion value; that is, the MSPE in the validation period. The entries are the deciles and the maximum of the criterion values and the associated $(\tau, \kappa_1, \kappa_2)$ of the RSCs with $s = 0$, the RSCs with "$s = 1(D)$" that sets $s = 1$ and uses $\mathcal{M}(\theta)$ with $x_t = 1$, and the RSCs with "$s = 1(M)$" that sets $s = 1$ and uses $\mathcal{M}(\theta)$ with $x_t = (1, t)^\top$. The minimum corresponds to RSC_g. "$s = 0$" includes RSC_p, RSC_c, RSC_w and RSC_g; "$s = 1(D)$" includes $RSC_p(D)$, $RSC_c(D)$, $RSC_w(D)$ and $RSC_g(D)$; "$s = 1(M)$" includes $RSC_p(M)$, $RSC_c(M)$, $RSC_w(M)$ and $RSC_g(M)$.

and the estimated intervention effects $\{\hat{\delta}_t\}_{t=T_0+1}^{T}$'s of the SC methods. The main empirical findings are summarized as follows.

Firstly, Table 2 shows that although CSC performs reasonably well, it is outperformed by certain RSCs in the validation period. Recall that CSC corresponds to RSC_p. Table 2 shows that RSC_p has the criterion value: 31.376, which is between the 20% quantile: 23.771 and the 30% quantile: 36.594 of the criterion values when $s = 0$. In comparison, $RSC_g(M)$, defined by the choice of $(\tau, \kappa_1, \kappa_2, \kappa_3) = (1, 0, 2400, 1)$, has the minimum criterion value: 0.193 among all the RSC methods. Like CSC, $RSC_g(M)$ also sets $\tau = 1$ and $\kappa_1 = 0$. Importantly, unlike CSC, $RSC_g(M)$ selects $s = 1$ by using $\mathcal{M}(\theta)$ with $x_t = (1, t)^\top$, and regularizes $\|w\|^2$ and $B_p(w)$ by setting $\kappa_2 = 2400$ and $\kappa_3 = 1$.

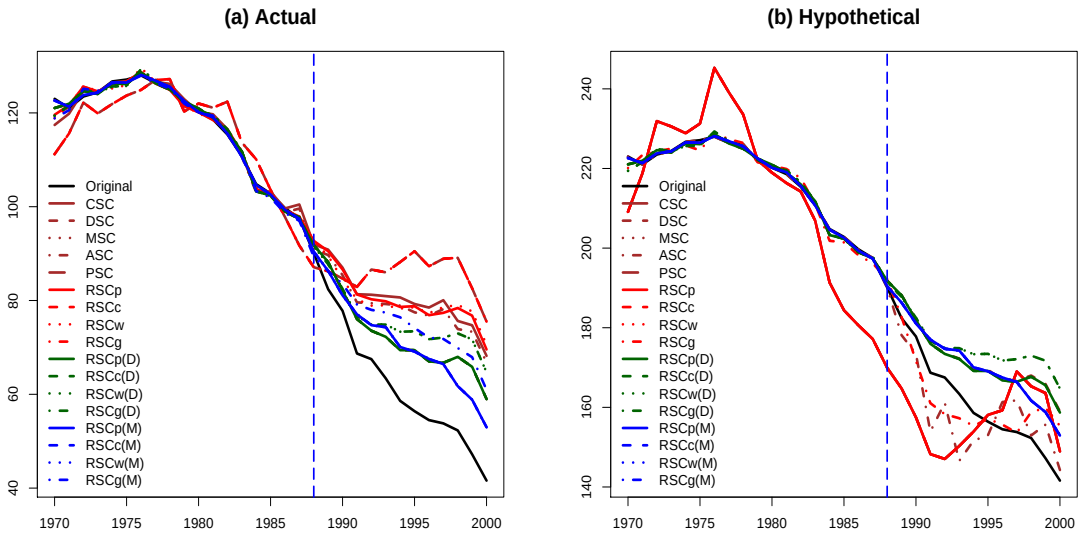

Figure 2: The outcome sequences of (a) California and (b) the hypothetical treated unit and the associated synthetic outcome sequences generated by different SC methods. The blue dashed line is evaluated at $T_0 = 19$ (that is, 1989).

Secondly, Table 2 also shows that $RSC_c(D)$ reduces to $RSC_p(D)$ by setting $\kappa_1 = 0$, and $RSC_c(M)$ and $RSC_w(M)$ both degenerate to $RSC_p(M)$ by setting $\tau = 1$ and $\kappa_1 = \kappa_2 = 0$. Thus, the former two SC methods and the latter three SC methods are, respectively, of the same performance in this case study. From Table 3, we also observe that PSC, RSC_c and RSC_g have the same performance in this case study.

Thirdly, Figure 2 shows that the $Y(w)$'s generated by most of the SC methods closely match Y_1. However, the $Y(w_{psc})$ generated by PSC (RSC_c or RSC_g) does not suitably match Y_1. This reflects the trade off between minimizing $B_p(w)$ and regularizing $B_c(w)$ in this case study. In comparison, it is visually difficult to distinguish CSC from the other SC methods regarding their performance in matching Y_1. Nonetheless, Table 3 shows that $RSC_p(M)$, $RSC_c(M)$ or $RSC_w(M)$ has the minimum $RMSE_{pre}$: 0.938 among the SC methods. In comparison, CSC has a larger $RMSE_{pre}$: 2.484.

Fourthly, Figure 2 also shows that the counterfactuals $\hat{y}_{1t}(w, s)$'s predicted by the SC methods are consistently greater than the y_{1t}'s in the post-intervention period. Accordingly, we obtain the same conclusion as ADH (2010) regarding the effectiveness of CTCP for reducing the treated state's cigarette sales. However, our conclusion is a consensus resulted from different SC methods. In addition, Figure 2 and Table 3 show that different

Table 3: Empirical performance measures of various SC methods.

		existing SC methods					$s=0$				$s=1(D)$				(d) $s=1(M)$			
		CSC	DSC	MSC	ASC	PSC	RSC_p	RSC_c	RSC_w	RSC_g	RSC_p	RSC_c	RSC_w	RSC_g	RSC_p	RSC_c	RSC_w	RSC_g
(a) Actual case	RMSE	2.484	1.594	0.939	2.020	4.444	2.179	4.444	2.128	4.444	1.594	1.594	1.691	1.691	0.938	0.938	0.938	1.454
	1990	-9.206	-4.300	-3.290	-7.594	-6.900	-8.816	-6.900	-6.643	-6.900	-4.301	-4.301	-4.692	-4.692	-3.289	-3.289	-3.289	-5.719
	1991	-12.634	-7.332	-8.355	-10.891	-14.200	-12.592	-14.200	-11.543	-14.200	-7.335	-7.335	-7.945	-7.945	-8.351	-8.351	-8.351	-10.440
	1992	-13.728	-6.051	-7.245	-11.930	-19.100	-12.757	-19.100	-11.409	-19.100	-6.053	-6.053	-7.391	-7.391	-7.241	-7.241	-7.241	-10.506
	1993	-17.533	-8.856	-10.910	-15.853	-22.600	-16.429	-22.600	-15.574	-22.600	-8.859	-8.859	-11.456	-11.456	-10.907	-10.907	-10.907	-14.019
	1994	-22.049	-10.805	-11.488	-20.327	-29.600	-20.035	-29.600	-19.672	-29.600	-10.808	-10.808	-14.740	-14.740	-11.486	-11.486	-11.486	-17.818
	1995	-22.857	-13.016	-12.716	-21.083	-34.100	-22.388	-34.100	-22.167	-34.100	-13.020	-13.020	-17.035	-17.035	-12.710	-12.710	-12.710	-17.742
	1996	-23.997	-12.529	-13.040	-22.273	-32.800	-22.373	-32.800	-22.795	-32.800	-12.533	-12.533	-17.250	-17.250	-13.035	-13.035	-13.035	-17.558
	1997	-26.261	-12.910	-12.648	-24.369	-35.100	-23.573	-35.100	-24.798	-35.100	-12.913	-12.913	-18.302	-18.302	-12.639	-12.639	-12.639	-17.999
	1998	-23.337	-15.690	-9.442	-21.612	-36.800	-26.108	-36.800	-26.731	-36.800	-15.694	-15.694	-20.690	-20.690	-9.438	-9.438	-9.438	-17.561
	1999	-27.520	-18.653	-11.652	-26.146	-35.400	-29.571	-35.400	-30.482	-35.400	-18.658	-18.658	-24.502	-24.502	-11.648	-11.648	-11.648	-20.715
	2000	-26.596	-17.382	-11.395	-25.195	-33.900	-27.961	-33.900	-28.948	-33.900	-17.386	-17.386	-23.102	-23.102	-11.382	-11.382	-11.382	-19.374
	sum	18.810	10.627	9.348	17.273	25.042	18.550	25.042	18.397	25.042	10.630	10.630	13.925	13.925	9.344	9.344	9.344	14.121
(b) Hypothetical case	RMSE	12.009	1.594	0.939	0.955	12.009	12.009	12.009	12.010	1.580	1.587	1.587	1.692	1.692	0.936	0.936	0.936	0.936
	1990	20.325	-4.300	-3.290	4.372	20.325	20.325	20.325	20.328	6.160	-4.129	-4.129	-4.699	-4.699	-3.273	-3.273	-3.273	-3.273
	1991	20.441	-7.332	-8.355	14.792	20.441	20.441	20.441	20.444	7.645	-7.263	-7.263	-7.951	-7.951	-8.334	-8.334	-8.334	-8.334
	1992	20.463	-6.051	-7.245	6.356	20.463	20.463	20.463	20.466	9.216	-5.925	-5.925	-7.396	-7.396	-7.216	-7.216	-7.216	-7.216
	1993	13.120	-8.856	-10.910	16.868	13.120	13.120	13.120	13.123	6.131	-8.765	-8.765	-11.461	-11.461	-10.883	-10.883	-10.883	-10.883
	1994	4.690	-10.805	-11.488	6.967	4.690	4.690	4.690	4.693	3.263	-10.569	-10.569	-14.746	-14.746	-11.444	-11.444	-11.444	-11.444
	1995	-1.698	-13.016	-12.716	3.392	-1.698	-1.698	-1.698	-1.693	-0.715	-12.739	-12.739	-17.040	-17.040	-12.656	-12.656	-12.656	-12.656
	1996	-4.766	-12.529	-13.040	-6.822	-4.766	-4.766	-4.766	-4.762	-1.160	-12.269	-12.269	-17.255	-17.255	-12.982	-12.982	-12.982	-12.982
	1997	-15.198	-12.910	-12.648	-7.962	-15.198	-15.198	-15.198	-15.191	0.187	-12.623	-12.623	-18.307	-18.307	-12.577	-12.577	-12.577	-12.577
	1998	-12.990	-15.690	-9.442	-0.638	-12.990	-12.990	-12.990	-12.984	-6.501	-15.397	-15.397	-20.695	-20.695	-9.396	-9.396	-9.396	-9.396
	1999	-16.371	-18.653	-11.652	-8.458	-16.371	-16.371	-16.371	-16.367	-12.683	-18.346	-18.346	-24.507	-24.507	-11.592	-11.592	-11.592	-11.592
	2000	-7.360	-17.382	-11.395	-2.664	-7.360	-7.360	-7.360	-7.359	-13.745	-17.091	-17.091	-23.107	-23.107	-11.317	-11.317	-11.317	-11.317
	sum	11.452	10.627	9.348	6.608	11.452	11.452	11.452	11.451	5.617	10.426	10.426	13.930	13.930	9.306	9.306	9.306	9.306
(b)-(a) Differences	RMSE	9.526	0.000	0.000	-1.065	7.565	9.830	7.565	9.882	-2.864	-0.007	-0.007	0.001	0.001	-0.001	-0.001	-0.001	-0.517
	1990	29.531	0.000	0.000	11.966	27.225	29.141	27.225	26.971	13.060	0.173	0.173	-0.006	-0.006	0.016	0.016	0.016	2.446
	1991	33.075	0.000	0.000	25.683	34.641	33.033	34.641	31.987	21.845	0.072	0.072	-0.006	-0.006	0.017	0.017	0.017	2.107
	1992	34.191	0.000	0.000	18.286	39.563	33.220	39.563	31.875	28.316	0.128	0.128	-0.006	-0.006	0.025	0.025	0.025	3.290
	1993	30.653	0.000	0.000	32.721	35.720	29.549	35.720	28.697	28.731	0.094	0.094	-0.006	-0.006	0.024	0.024	0.024	3.136
	1994	26.738	0.000	0.000	27.295	34.290	24.724	34.290	24.365	32.863	0.239	0.239	-0.005	-0.005	0.043	0.043	0.043	6.375
	1995	21.159	0.000	0.000	24.476	32.402	20.691	32.402	20.474	33.385	0.281	0.281	-0.005	-0.005	0.054	0.054	0.054	5.086
	1996	19.231	0.000	0.000	15.451	28.034	17.607	28.034	18.033	31.640	0.264	0.264	-0.005	-0.005	0.054	0.054	0.054	4.577
	1997	11.063	0.000	0.000	16.407	19.902	8.375	19.902	9.607	35.287	0.291	0.291	-0.005	-0.005	0.062	0.062	0.062	5.422
	1998	10.347	0.000	0.000	20.974	23.810	13.118	23.810	13.747	30.299	0.296	0.296	-0.005	-0.005	0.042	0.042	0.042	8.165
	1999	11.148	0.000	0.000	17.688	19.029	13.200	19.029	14.116	22.717	0.312	0.312	-0.005	-0.005	0.056	0.056	0.056	9.123
	2000	19.236	0.000	0.000	22.531	26.540	20.600	26.540	21.588	20.155	0.295	0.295	-0.005	-0.005	0.065	0.065	0.065	8.057
	sum	-7.358	0.000	0.000	-10.665	-13.590	-7.098	-13.590	-6.946	-19.425	-0.204	-0.204	0.005	0.005	-0.038	-0.038	-0.038	-4.815

Note: For RSC, "$s=0$" includes RSC_p, RSC_c, RSC_w and RSC_g; "$s=1(D)$" includes $RSC_p(D)$, $RSC_c(D)$, $RSC_w(D)$ and $RSC_g(D)$; "$s=1(M)$" includes $RSC_p(M)$, $RSC_c(M)$, $RSC_w(M)$ and $RSC_g(M)$.

SC methods might generate different $\hat{\delta}_t$'s. In particular, after excluding PSC (RSC_c or RSC_g), the absolute intervention effects estimated by other SC methods are obviously smaller than those estimated by CSC (RSC_p or RSC_w).

5.2 Hypothetical case study

In the following, we consider a hypothetical case study that has the same intervention effects as, but a different type of data from, the (previous) actual case study. The only difference between the two cases is due to the design that the hypothetical case is defined by replacing the outcome sequence of the treated unit $\{y_{1t}\}_{t=1}^{T}$ with a hypothetical outcome sequence $\{y_{1t}^*\}_{t=1}^{T}$, where $y_{1t}^* := \alpha + y_{1t}$ and $\alpha = 100$. As shown by Figure 2, unlike the actual outcome sequence $\{y_{1t}\}_{t=1}^{T_0}$ which is surrounded by its untreated counterparts, the hypothetical outcome sequence $\{y_{1t}^*\}_{t=1}^{T_0}$ is essentially above all but one of the untreated counterparts. By this design, the hypothetical case mimics a poor-matching problem that does not appear in the actual case. It should be noted that, because the hypothetical case is artificial, the associated results do not have direct empirical interpretations. However, this design is useful for assessing the potential robustness of the SC methods. Specifically, although the poor-matching problem is not uncommon in real data, the intervention effects across different real-data case studies are typically incomparable. In comparison, because the hypothetical case shares the same intervention effects as the actual case, the design allows us to assess the potential robustness of a SC method by comparing the intervention effects estimated from the actual and hypothetical cases. For ease of comparison, we plot $\{y_{1t}^*\}_{t=T_0+1}^{T}$ and the associated synthetic outcome sequences in Figure 2, and report the validation criterion values and the regularization parameters (the pre-intervention RMSEs and the estimated intervention effects) of the hypothetical case in Table 2 (Table 3). The main findings are summarized as follows.

Firstly, as shown by Table 2, the validation performance of CSC (RSC_p) is considerably worsened in the hypothetical case because of the poor-matching problem. Specifically, RSC_p has the validation criterion value: 552.156, which is substantially greater than its actual counterpart: 31.376. In comparison, RSC_g, defined by setting $\tau = 1.5$ and $\kappa_1 = \kappa_2 = 0$, has the criterion value: 32.707 in the hypothetical case. This shows that relaxing the restriction: $\tau = 1$ is useful for refining the poor-matching problem of the CSC method to some extent. Nonetheless, RSC_g is considerably outperformed by $RSC_w(D)$, defined by setting $\tau = 1$ and $(\kappa_1, \kappa_2, \kappa_3) = (0, 400, 0)$, which has the minimum criterion

value: 5.397 in this case. This shows again the importance of refining the poor-matching problem by first using a suitable $\mathcal{M}(\theta)$ and then regularizing $\|w\|^2$.

Secondly, Table 2 shows that RSC_c reduces to RSC_p by setting $\kappa_1 = 0$. This result is interesting. It reflects that $B_p(w)$ dominates $B_c(w)$ in the regularization problem of RSC_c when the poor-matching problem is obvious. Table 2 also shows that $RSC_g(D)$ reduces to $RSC_w(D)$ by setting $(\kappa_1, \kappa_3) = (0, 0)$, and $RSC_c(D)$ and $RSC_w(D)$ both degenerate to $RSC_p(D)$ by setting $\kappa_1 = \kappa_2 = 0$ in this case. Table 3 further shows that RSC_w has similar performance to RSC_p (RSC_g), $RSC_p(D)$ and $RSC_c(D)$ are identical, and $RSC_p(M)$, $RSC_c(M)$, $RSC_w(M)$ and $RSC_g(M)$ have the same performance in this case.

Thirdly, Figure 2 shows that CSC (RSC_p or RSC_c) is obviously different from the other SC methods in matching $\{y_{1t}^*\}_{t=1}^{T_0}$. Specifically, CSC (RSC_p or RSC_c) is unable to match the pre-intervention outcomes of the hypothetical treated unit. By contrast, the other SC methods still match the pre-intervention outcomes quite well. Specifically, Table 3 shows that CSC (PSC) has $RMSE_{pre} = 12.009$, while $RSC_p(M)$, $RSC_c(M)$, $RSC_w(M)$ or $RSC_g(M)$ has the minimum $RMSE_{pre}$: 0.936 in this case.

Fourthly, Figure 2 shows that, unlike the actual case, the counterfactuals predicted by the SC methods are not consistently greater than the post-intervention outcomes of the hypothetical treated unit, even though the hypothetical case shares the same intervention effects as the actual case. This finding illustrates that some of the SC methods are not robust in estimating the intervention effects. To see this point, we further report the difference between the RMSEs and the difference between the estimated intervention effects of the two cases for each SC method in Table 3. The differences are essential zero for DSC and MSC, and are also close to zero for $RSC_w(D)$, $RSC_g(D)$, $RSC_p(M)$, $RSC_c(M)$ and $RSC_w(M)$. By contrast, the absolute intervention effects estimated by the other SC methods in the hypothetical case are obviously lower than their actual counterparts, especially for CSC, ASC, PSC and the particular RSCs without using auxiliary models by setting $s = 0$.

6 Conclusions

The notion of SC has been widely regarded as essential for policy evaluation. In the recent literature, researchers have proposed various extensions of CSC that choose the weighting vector w in different ways, including PSC that considers the potential interpolation bias of CSC, different penalized-regression methods that account for the potential poor-

matching problem of CSC without using auxiliary models ($s = 0$), and DSC, MSC and ASC that deal with the potential poor-matching problem of CSC using auxiliary models ($s = 1$). In this paper, we propose a unified approach to compare and generate useful complements of existing methods in a generalized context where the true outcome model is unknown. This approach is established by first exploring an upper bound of the MSPE of the counterfactual predicted by an arbitrary SC and then proposing a generalized SC method, that is RSC, to regularize the components of this potential MSPE by the choice of (w, s) under a sign-and-size restriction: $w \in \mathbb{S}(\tau)$ for $\tau \geq 1$. The components include the matching-quality divergences: $\hat{B}_p^2(w, s)$, $B_c^2(w)$ and $B_p^2(w)$ (if $s = 1$) and the squared L_2-norm: $\|w\|^2$. We illustrate that RSC includes several existing SC methods, or their variants, as particular examples with different restrictions of regularization parameters. By this unified approach, we assess the potential biases and MSPEs of a particular SC method and generate useful new SC methods. In particular, we observe that the RSC that regularizes both $\hat{B}_p^2(w, s)$ and $\|w\|^2$ is a useful complement of MSC because MSC regularizes $\hat{B}_p^2(w, s)$ but overlooks $\|w\|^2$ which is essential for determining the potential variance of its prediction error. The simulation shows that this type of RSC performs favorably in comparison with MSC and many other SC methods. We also illustrate the usefulness of our method using the case study considered by ADH (2010), and assess the robustness of SC methods in estimating the intervention effects by comparing this case study with an associated hypothetical case study.

References

[1] Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the Basque Country, *American Economic Review*, **93**, 113-132.

[2] Abadie, A. and J. L'Hour (2019). A penalized synthetic control estimator for disaggregated data, Working paper, Massachusetts Institute of Technology.

[3] 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**, 493-505.

[4] Abadie, A., A. Diamond and J. Hainmueller (2015). Comparative politics and the synthetic control method, *American Journal of Political Science*, **59**, 495-510.

[5] Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects, *Journal of Economic Literature*, **59**, 391-425.

[6] Amjad, M., D. Shah and D. Shen (2018). Robust synthetic control, *Journal of Machine Learning Research*, **19**, 1-51.

[7] Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens and S. Wager (2021). Synthetic difference in differences, *American Economic Reviews*, **111**, 4088-4118.

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

[9] Ben-Michael, E., A. Feller and J. Rothstein (2021). The augmented synthetic control method, *Journal of the American Statistical Association*, forthcoming.

[10] Botosaru, I. and B. Ferman (2019). On the role of covariates in the synthetic control method, *Econometrics Journal*, **22**, 117-130.

[11] Chen, Y.-T. (2020). A distributional synthetic control method for policy evaluation, *Journal of Applied Econometrics*, **35**, 505-525.

[12] Chernozhukov, V., K. Wüthrich and Y. Zhu (2019). Practical and robust t-test based inference for synthetic control and related methods, Working paper, arXiv.

[13] Chernozhukov, V., K. Wüthrich and Y. Zhu (2021). An exact and robust conformal inference method for counterfactual and synthetic controls, *Journal of the American Statistical Association*, forthcoming.

[14] Doudchenko and Imbens (2017). Balancing, regression, difference-in-differences and synthetic control methods: A Synthesis, Working paper, NBER.

[15] Ferman, B. and C. Pinto (2019). Synthetic controls with imperfect pre-treatment fit, Working paper, Sao Paulo School of Economics.

[16] Gobillon, L. and T. Magnac (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls, *Review of Economics and Statistics*, **98**, 535-51.

[17] Hollingsworth, A. and C. Wing (2020). Tactics for design and inference in synthetic control studies: An applied example using high-dimensional data, *Working paper*, Indiana University.

[18] Kellogg, M., M. Mogstad, G. Pouliot and A. Torgovitsky (2020). Combining matching and synthetic controls to trade off biases from extrapolation and interpolation, Working paper, NBER.

[19] Li, K. T. (2020). Statistical inference for average treatment effects estimated by synthetic control methods, *Journal of the American Statistical Association*, **115**, 2068-83.

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

[21] Valero, R. (2015). Synthetic control method versus standard statistical techniques: A comparison for labor market reforms, Working paper, University of Alincante.

[22] Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models, *Political Analysis*, **25**, 57-76.

Supplementary Appendix of "Regularization of Synthetic Controls for Policy Evaluation"

Yi-Ting Chen
Department of Finance
National Taiwan University

This appendix presents the mathematical proofs of the paper.

Proof of Lemma 1

Under Assumption 1, the counterfactual $y_{1t}(0)$ is evaluated at the potential outcome of the intervention variable: $D_{1t} = 0$, and has the form:

$$y_{1t}(0) = \psi_t(Y_1) + \varepsilon_{1t}, \tag{A1}$$

for $t \geq T_0 + 1$. According to (3) and (A1), we have

$$\begin{aligned}
y_{1t}(0) - \hat{y}_t(w, s) &= (\psi_t(Y_1) + \varepsilon_{1t}) - \left(y_t(w) + s \cdot m_t(w, \hat{\theta})\right) \\
&= (\psi_t(Y_1) + \varepsilon_{1t}) - \left(\sum_{i \geq 2} w_i y_{it} + s \cdot m_t(w, \hat{\theta})\right) \\
&= (\psi_t(Y_1) + \varepsilon_{1t}) - \left(\sum_{i \geq 2} w_i (\psi_t(Y_i) + \varepsilon_{it}) + s \cdot m_t(w, \hat{\theta})\right) \\
&= \psi_t(Y_1) - \sum_{i \geq 2} w_i \psi_t(Y_i) - s \cdot m_t(w, \theta) \\
&+ \left(\varepsilon_{1t} - \sum_{i \geq 2} w_i \varepsilon_{it}\right) - s \cdot \left(m_t(w, \hat{\theta}) - m_t(w, \theta)\right),
\end{aligned} \tag{A2}$$

for $t \geq T_0 + 1$, in which the second equality is due to (1), the third equality is due to Assumption 1(i) for $i \geq 2$. We obtain this lemma from (A2). □

Proof of Proposition 1

To see (26), note that Assumption 3 implies

$$\mathbb{E}[e_t(w)|\boldsymbol{Y}] = 0$$

and

$$\mathbb{E}[u_{it}(w,s)|\boldsymbol{Y}] = 0.$$

Thus,

$$MSPE_t(w,s) = Bias_t^2(w,s) + \sigma_t^2(w,s),$$

where

$$\sigma_t^2(w,s) := \text{var}[u_t(w,s)|\boldsymbol{Y}] = \mathbb{E}[e_t^2(w)|\boldsymbol{Y}] + s \cdot \mathbb{E}[(m_t(w,\hat{\theta}) - m_t(w,\theta))^2|\boldsymbol{Y}],$$

and

$$\mathbb{E}[e_t^2(w)|\boldsymbol{Y}] = \sigma_\varepsilon^2 \left(1 + \|w\|^2\right).$$

Thus, we have 26).

To show (28), note that

$$Bias_{pre,t}(w,s) = \psi_t(Y_1) - \psi_t(Y(w,s)).$$

Under Assumption 2(i), we may use the mean-value expansion to show that

$$\psi_t(Y_1) = \psi_t(Y(w,s)) + \nabla\psi_t^\top(Y_{1w}^\dagger)(Y_1 - Y(w,s)),$$

where $Y_{1w}^\dagger$ is a mean value such that $\|Y_{1w}^\dagger - Y(w,s)\| \leq \|Y_1 - Y(w,s)\|$. Given this result, we may use the Cauchy-Schwarz inequality to further show that

$$\begin{aligned} |Bias_{pre,t}(w,s)| &\leq \|\nabla\psi_t(Y_{1w}^\dagger)\|\|Y_1 - Y(w,s)\| \\ &\leq \xi_{\psi,1}\|Y_1 - Y(w,s)\|, \end{aligned} \tag{A3}$$

where the last inequality is due to Assumption 2(i). Thus, we have (28).

To show (29), note that

$$Bias_{nl,t}(w) = \psi_t(Y(w)) - \sum_{i \geq 2} w_i \psi_t(Y_i). \tag{A4}$$

Under Assumption 2(i), we may also use the mean-value expansion to show that

$$\psi_t(Y(w)) = \psi_t(Y_i) + \nabla \psi_t^\top (Y_{wi}^\dagger)(Y(w) - Y_i),$$

where $Y_{wi}^\dagger$ is a mean value such that $\|Y_{wi}^\dagger - Y_i\| \leq \|Y(w) - Y_i\|$. By the Cauchy-Schwarz inequality, we may further show that

$$\begin{aligned} |\psi_t(Y(w)) - \psi_t(Y_i)| &\leq \|\nabla \psi_t(Y_{wi}^\dagger)\| \|Y(w) - Y_i\| \\ &\leq \xi_{\psi,1} \|Y(w) - Y_i\|. \end{aligned} \tag{A5}$$

Given the decomposition:

$$Y(w) - Y_i = (Y(w) - Y_1) + (Y_1 - Y_i)\,, \tag{A6}$$

we may use the triangle inequality to show that

$$\begin{aligned} \|Y(w) - Y_i\| &\leq \|Y_1 - Y_i\| + \|Y_1 - Y(w)\| \\ &= \|Y_1 - Y_i\| + B_p(w). \end{aligned} \tag{A7}$$

By introducing (A7) in (A5), we have

$$|\psi_t(Y(w)) - \psi_t(Y_i)| \leq \xi_{\psi,1} \left(\|Y_1 - Y_i\| + B_p(w) \right). \tag{A8}$$

By applying the triangle inequality to (17), we further obtain

$$\begin{aligned} |Bias_{nl,t}| &\leq |1 - \iota_n^\top w| \, |\psi_t(Y(w))| + \sum_{i \geq 2} |w_i| |\psi_t(Y(w)) - \psi_t(Y_i)| \\ &\leq \xi_{\psi,0} B_a(w) + \sum_{i \geq 2} |w_i| |\psi_t(Y(w)) - \psi_t(Y_i)|, \end{aligned} \tag{A9}$$

where the last inequality is due to Assumption 2(i). By introducing (A8) in (A9), we have (29).

To show (30), note that

$$Bias_{ms,t}(w,s) = \psi_t(Y(w,s)) - \psi_t(Y(w)) - s \cdot m_t(w,\theta). \tag{A10}$$

By applying the triangle inequality to (A10), we have

$$|Bias_{ms,t}(w,s)| \leq |\psi_t(Y(w,s)) - \psi_t(Y(w))| + s \cdot |m_t(w,\theta)|. \tag{A11}$$

Under Assumption 2(i), we may use the mean-value expansion to show that

$$\begin{aligned}\psi_t(Y(w,s)) &= \psi_t(Y(w)) + \nabla\psi_t^\top(Y_{ws}^\dagger)(Y(w,s) - Y(w)),\\ &= \psi_t(Y(w)) + \nabla\psi_t^\top(Y_{ws}^\dagger)(s \cdot M(w,\theta)),\end{aligned}$$

where $Y_{ws}^\dagger$ is a mean value such that $\|Y_{ws}^\dagger - Y(w)\| \leq \|Y(w,s) - Y(w)\|$, and the second equality is due to (6). By the Cauchy-Schwarz inequality, we may further show that

$$\begin{aligned}|\psi_t(Y(w,s)) - \psi_t(Y(w))| &\leq \|\nabla\psi_t(Y_{ws}^\dagger)\|\|s \cdot M(w,\theta)\|\\ &\leq s \cdot \xi_{\psi,1}\|M(w,\theta)\|.\end{aligned} \tag{A12}$$

According to (2) and (5), we can write that

$$M(w,\theta) = \sum_{i\geq 2} w_i m_{i\cdot}(\theta). \tag{A13}$$

By applying the triangle inequality to (A13), we have

$$\|M(w,\theta)\| \leq \sum_{i\geq 2} |w_i|\|m_{i\cdot}(\theta)\| \leq \sum_{i\geq 2} |w_i|\xi_{m,0}^2 = \xi_{m,0}\|w\|_1, \tag{A14}$$

where the second inequality is due to Assumption 2(ii). By introducing (A14) in (A12), we have

$$|\psi_t(Y(w,s)) - \psi_t(Y(w))| \leq \xi_{\psi,1}\xi_{m,0} s \cdot \|w\|_1. \tag{A15}$$

The triangle inequality also implies

$$|m_t(w,\theta)| \leq \sum_{i\geq 2} |w_i||\mu_{1t}(\theta_1) - \mu_{it}(\theta_i))| \leq \xi_{m,0}\|w\|_1. \tag{A16}$$

By introducing (A15) and (A16) in (A11), we obtain

$$|Bias_{ms,t}(w,s)| \leq (1+\xi_{\psi,1})\xi_{m,0}(s\cdot\|w\|_1). \tag{A17}$$

This shows (30).

In addition, given (13), we obtain (31) from (28) and (29).

To show (32), note that, according to (2),

$$\begin{aligned} m_t(w,\hat{\theta}) - m_t(w,\theta) &= \sum_{i\geq 2} w_i \left(\Delta\mu_{it}(\hat{\theta}_1,\hat{\theta}_i) - \Delta\mu_{it}(\theta_1,\theta_i)\right) \\ &= w^\top (m_{\cdot t}(\hat{\theta}) - m_{\cdot t}(\theta)). \end{aligned}$$

Accordingly, we may use the Cauchy-Schwartz inequality to show that

$$\left(m_t(w,\hat{\theta}) - m_t(w,\theta)\right)^2 \leq \|w\|^2 \|m_{\cdot t}(\hat{\theta}) - m_{\cdot t}(\theta)\|^2. \tag{A18}$$

By the mean-value expansion, we have

$$m_{\cdot t}(\hat{\theta}) = m_{\cdot t}(\theta) + \nabla_{\theta^\top} m_{\cdot t}(\theta^\dagger)(\hat{\theta}-\theta),$$

where $\theta^\dagger$ is a mean value such that $\|\theta^\dagger - \theta\| \leq \|\hat{\theta}-\theta\|$. Thus,

$$\begin{aligned} \|m_{\cdot t}(\hat{\theta}) - m_{\cdot t}(\theta)\| &= \|\nabla_{\theta^\top} m_{\cdot t}(\theta^\dagger)(\hat{\theta}-\theta)\| \leq \|\nabla_{\theta^\top} m_{\cdot t}(\theta^\dagger)\| \|\hat{\theta}-\theta\| \\ &\leq \xi_{m,1}\|\hat{\theta}-\theta\|, \end{aligned} \tag{A19}$$

where the first inequality is due to the Schwarz matrix inequality, and the second inequality is due to Assumption 2(ii). From (A18) and (A19), we have

$$\left(m_t(w,\hat{\theta}) - m_t(w,\theta)\right)^2 \leq \xi_{m,1}^2 \|w\|^2 \|\hat{\theta}-\theta\|^2. \tag{A20}$$

According to (A20) and Assumption 3(ii), we obtain

$$\mathbb{E}\left[\left(m_t(w,\hat{\theta}) - m_t(w,\theta)\right)^2 \middle| \boldsymbol{Y}\right] \leq \xi_{m,1}^2 \sigma_\theta^2 \|w\|^2. \tag{A21}$$

The result in (32) is due to (27) and (A21).

In addition, given (26) we may obtain (33) from (31) and (32). □

Derivation of (39)

To show (39), note that

$$\begin{aligned}\hat{B}_p^2(w,s) &= \|Y_1 - \hat{\boldsymbol{Y}}(w,s)\|^2 = \|Y_1 - \hat{\boldsymbol{Y}}^s_{(-1)}w\|^2 \\ &= (Y_1^\top - w^\top \hat{\boldsymbol{Y}}^{s\top}_{(-1)})(Y_1 - \hat{\boldsymbol{Y}}^s_{(-1)}w) \\ &= Y_1^\top Y_1 - 2w^\top \hat{\boldsymbol{Y}}^{s\top}_{(-1)} Y_1 + w^\top \hat{\boldsymbol{Y}}^{s\top}_{(-1)} \hat{\boldsymbol{Y}}^s_{(-1)} w,\end{aligned}$$

$$B_c(w) = \left(\sum_{i\geq 2} w_i \|Y_1 - Y_i\|\right)^2 = (\boldsymbol{D}_Y^\top w)^\top (\boldsymbol{D}_Y^\top w) = w^\top \boldsymbol{D}_Y \boldsymbol{D}_Y^\top w,$$

and

$$B_p^2(w) = Y_1^\top Y_1 - 2w^\top \boldsymbol{Y}^\top_{(-1)} Y_1 + w^\top \boldsymbol{Y}^\top_{(-1)} \boldsymbol{Y}_{(-1)} w.$$

Accordingly, we can rewrite (37) as:

$$\begin{aligned}Q^*(w,s|\kappa) &= \left(Y_1^\top Y_1 - 2w^\top \hat{\boldsymbol{Y}}^{s\top}_{(-1)} Y_1 + w^\top \hat{\boldsymbol{Y}}^{s\top}_{(-1)} \hat{\boldsymbol{Y}}^s_{(-1)} w\right) + \kappa_1 \left(w^\top \boldsymbol{D}_Y \boldsymbol{D}_Y^\top w\right) \\ &\quad + \kappa_2 w^\top w + s\cdot\kappa_3 \left(Y_1^\top Y_1 - 2w^\top \boldsymbol{Y}^\top_{(-1)} Y_1 + w^\top \boldsymbol{Y}^\top_{(-1)} \boldsymbol{Y}_{(-1)} w\right) \\ &= (1 + s\cdot\kappa_3) Y_1^\top Y_1 \\ &\quad + w^\top \left(\hat{\boldsymbol{Y}}^{s\top}_{(-1)} \hat{\boldsymbol{Y}}^s_{(-1)} + \kappa_1 \boldsymbol{D}_Y \boldsymbol{D}_Y^\top + \kappa_2 I_n + s\cdot\kappa_3 \boldsymbol{Y}^\top_{(-1)} \boldsymbol{Y}_{(-1)}\right) w \\ &\quad - 2w^\top \left(\hat{\boldsymbol{Y}}^s_{(-1)} + s\cdot\kappa_3 \boldsymbol{Y}_{(-1)}\right)^\top Y_1 \\ &= (1 + s\cdot\kappa_3) Y_1^\top Y_1 + \left(\frac{1}{2} w^\top V w - v^\top w\right),\end{aligned} \tag{A22}$$

where V and v are, respectively, defined in (40) and (41). Since $(1 + s\cdot\kappa_3) Y_1^\top Y_1$ is free of w, we obtain (39) from (38) and (A22). □