Asymptotic Properties of the Synthetic Control Method[*]

Xiaomeng Zhang[†], Wendun Wang[‡], Xinyu Zhang[§]

November 22, 2022

Abstract

This paper provides new insights into the asymptotic properties of the synthetic control method (SCM). We show that the synthetic control (SC) weight converges to a limiting weight that minimizes the mean squared prediction risk of the treatment-effect estimator when the number of pretreatment periods goes to infinity, and we also quantify the rate of convergence. Observing the link between the SCM and model averaging, we further establish the asymptotic optimality of the SC estimator under imperfect pretreatment fit, in the sense that it achieves the lowest possible squared prediction error among all possible treatment effect estimators that are based on an average of control units, such as matching, inverse probability weighting and difference-in-differences. The asymptotic optimality holds regardless of whether the number of control units is fixed or divergent. Thus, our results provide justifications for the SCM in a wide range of applications. The theoretical results are verified via simulations.

JEL classification: C13, C21, C23
Keywords: Synthetic control method; Model averaging; Asymptotic optimality; Linear factor model; Policy evaluation

1 Introduction

The synthetic control method (SCM), proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010), has become one of the most popular approaches for policy evaluation. The idea of the SCM is to construct a synthetic control (SC) unit intended to mimic the behavior of pretreatment outcomes of the treated unit to the greatest possible, such that one can use the posttreatment outcome of the synthetic control as the counterfactual of the treated outcome. Thus, in a seminal paper, Abadie et al. (2010) regards a good pretreatment fit as a prerequisite for the standard SCM to work. Despite its wide range of applications and generalizations, the theoretical (asymptotic) properties of the SCM, especially under imperfect pretreatment fit, are relatively less studied; see Abadie (2021) for a review.

Assuming that outcomes are generated from a linear factor model, Ferman and Pinto (2021) investigates the asymptotic behavior of the SC weight when the pretreatment fit is not perfect. They demonstrate that the SC weight converges to a limit that generally does not recover the factor structure of the treated unit as the number of pretreatment periods increases, implying an asymptotic bias of the SC estimate of the treatment effect. Ferman (2021) further examines the case of imperfect pretreatment fit when the number of control units goes to infinity and derives the conditions under which the factor structure of the treated unit can be asymptotically recovered by the SC unit, such that the SC estimator is asymptotically unbiased. A crucial condition here is that there exist weights that are diluted among an increasing number of control units and at the same time can produce an SC unit whose factor structure asymptotically reconstructs that of the treated unit. While the analysis is very informative, the existence of such weights is not automatically guaranteed. It is not yet clear in which cases these presumed diluting weights exist and which quantities affect the convergence of the SC weight. Moreover, it is also unclear how the SC estimator performs compared to other popularly used treatment-effect estimators when the pretreatment fit is imperfect.

In this paper, we provide new insights into the asymptotic properties of the SCM. First, we show that the SC weight converges to a limiting weight that minimizes the mean squared prediction risk of the treatment-effect estimator, and we also quantify the rate of convergence. Under a widely studied linear factor model (see, e.g., Abadie et al., 2010; Hsiao and Zhou, 2019; Botosaru and Ferman, 2019), we demonstrate that our limiting weight is compatible

with the infeasible weight studied in Ferman and Pinto (2021), and thus it also balances the two parts of errors in fitting the factor structure and idiosyncratic shocks. This result confirms the finding in Ferman and Pinto (2021) that the SC estimator is asymptotically biased under imperfect pretreatment fit, and is also in line with the argument of Bottmer et al. (2021) that the SC estimator is generally biased in a design-based framework. The property of being biased while asymptotically reaching the minimum mean squared prediction risk suggests that there is a bias-variance tradeoff in the SC estimator under imperfect pretreatment fit. We also complement Ferman and Pinto (2021) by quantifying the rate of convergence. We find that a better pre- and posttreatment fit both facilitate the convergence of the SC weight as expected. The role of the number of pretreatment periods is mixed, depending on the goodness of fit before and after the treatment. We also find that a larger number of control units is associated with a slower convergence rate.

Second, our derivation also offers a method to verify the existence of the diluting weight assumed in Ferman (2021). We provide a sufficient condition under which our limiting weight is diluted and can asymptotically reconstruct the factor structure of the treated unit when the number of control units diverges. Intuitively, it requires that the information contained in the factor structure neither diminishes nor diverges. Nondiminishing guarantees that the factor structure of the treated unit can be reconstructed by the SC unit and nondivergence guarantees that weights can dilute among control units.

Finally, motivated by the bias-variance tradeoff of the SC estimator, we explore its (expected) mean squared prediction error (MSPE), namely the (expected) mean squared loss of the SC estimator in the posttreatment periods.[1] We show that under imperfect pretreatment fit[2], the SC estimator is asymptotically optimal in the sense that it achieves the lowest possible prediction risk (and loss) among all possible treatment effect estimators that are based on a (weighted) average of control units, when the number of pretreatment periods goes to infinity. Importantly, the asymptotic optimality holds regardless of whether the number of control units is fixed or divergent. In fact, under a fixed number of control units, the SCM makes the optimal tradeoff between bias and variance. With an increasing number of control units, the SC estimator is asymptotically unbiased, and thus the variance of the SC estimator

---

[1]We emphasize the prediction aspect of errors to highlight the out-of-sample feature of the treatment-effect estimate as it employs pretreatment information to extrapolate posttreatment counterfactual outcomes.

[2]Here an imperfect pretreatment fit refers to the situation where there are no weights such that the outcome and observed covariates of the SC unit precisely equal those of the treated unit at each pretreatment period. Further discussion on the definition of imperfect fit is provided below Condition 9.

converges to its lower bound. Note that there are several treatment effect estimators that are based on a (weighted) average of the control units, such as the matching estimator, the inverse probability weighting (IPW), and difference-in-differences (DID) when an intercept is introduced (Doudchenko and Imbens, 2016), our result implies that these methods cannot outperform the SCM in terms of MSPE under imperfect pretreatment fit, at least asymptotically. The asymptotic optimality of SCM provides theoretical foundation for the numerical finding of Bottmer et al. (2021) that SC estimators produce lower root mean squared errors than difference-in-means or DID in their simulation studies. It is worth noting that while we base on linear factor models to demonstrate the main results, the asymptotic optimality continues to hold in a model-free setup, namely without assuming that outcomes are generated by a linear factor model (see Section 3.3). This generalization alleviates the concern that linear factor models are not sufficiently general to depict the data generating process (DGP) of potential outcomes, and provides guarantee for the SCM in a broad range of situations.

Our analysis offers a justification for the SCM under imperfect pretreatment fit. Abadie et al. (2010) requires perfect pretreatment fit to guarantee the unbiasedness of the SC estimator. Botosaru and Ferman (2019) relaxes the perfect fit of covariates and shows that the asymptotic unbiasedness still holds as long as the pretreatment fit of outcomes is perfect. This relaxation comes at the cost of imposing stronger assumptions on the effects of covariates. Ferman and Pinto (2021) shows that the SC estimator is biased under imperfect pretreatment fit, and Ferman (2021) further argues that this bias is diminishing as the number of control units increases. Our result of asymptotic optimality of the SCM does not rely on the divergence of the number of control units or the effect of covariates. It suggests that although the SC estimator may be biased under imperfect pretreatment fit, it is still the best choice among all other estimators of a similar construction. Considering the fact that an absolutely perfect fit is hardly achievable in real data, our result significantly widens the applicability of the SCM.

Our analysis of asymptotic optimality is inspired by and contributes to the optimal model averaging literature (see, e.g., Hansen, 2007; Wan et al., 2010; Liu and Okui, 2013; Zhang, 2021, among many others), which seeks to achieve the best prediction by optimally combining estimators obtained from candidate models (with different specifications). Observing the link between the SCM and optimal model averaging, we build upon the asymptotic optimality of averaging estimators to examine the properties of the SC estimator. Unlike model

4

averaging that combines the estimates of candidate models, the SCM synthesizes the realized (posttreatment) observations of the control units. Thus, our analysis does not constrain the behavior of candidate controls. An important contribution of this paper to the model averaging literature is that we provide the first attempt to prove the asymptotic optimality of the out-of-sample *prediction* risk, which accounts for the randomness of both outcomes and weights. In contrast, the majority of optimal averaging studies examine the risk while assuming that the weights are fixed. The only exception is Zhang (2021), who analyze the risk accounting for the randomness of weights, but they only consider the in-sample risk.

In an independent and parallel study, Chen (2022) associates SCM with online learning and shows that SCM can perform almost as well as the best weighted average in a worse-case scenario. While their overall conclusion is somewhat related with our asymptotic optimality, we differ from this study from several major perspectives. First, to link with online learning, Chen (2022) considers a thought experiment, where the outcomes are generated by an adversary and become sequentially available over time, such that the SC weights are calculated in a time-varying manner. While the adversarial framework is useful in several senses, it restricts the analysis to one-step-ahead prediction of the outcome, which is regarded as an undesirable departure from the standard SCM as the author points out in his conclusion. In contrast, we consider the standard setup of SCM with pretreatment outcomes fully accessible for prediction. This framework facilitates multiple-step-ahead prediction, and thus allows estimation of the treatment effect in multiple posttreatment periods simultaneously, a more common practice for SCM. Second, the targeting best weight in Chen (2022) is defined to minimize the mean squared loss over both pre- and posttreatment periods. Such an objective function also deviates from the goal of SCM (and other methods for treatment evaluation) that focuses on only the posttreatment prediction accuracy. With the pretreatment fit also accounted for in the evaluation, the overall regret guarantee in Chen (2022) does not necessarily imply the minimum loss in the posttreatment periods. In contrast, our asymptotic optimality explicitly concerns the mean squared loss in the posttreatment periods. Finally, Chen (2022) illustrates the advantage of SCM in terms of the regret bound, while our asymptotic optimality concerns the ratio of (expected) MSPE over its infimum. Under certain regularity conditions, the asymptotic optimality can imply the bound convergence of the corresponding regret (see the end of Section 3.2.1 for details).

The remainder of this paper is organized as follows. Section 2 describes the model

setup. Section 3 presents the main theoretical results, where we examine the convergence and establish the asymptotic optimality. The theory is verified via simulations in Section 4, and Section 5 concludes the paper. The proofs of the main theorems are provided in the Appendix. The Online Appendix contains additional theoretical results.

2 Model framework

Our framework considers the canonical SCM panel data, where there are $i \in \{0, 1, \ldots, J\}$ units with $i = 0$ being the only treated unit and the remaining $\{1, \ldots, J\}$ being the control units. The set of potential control units is also referred to as the "donor pool". Suppose that the treatment is assigned after time 0; then, we denote $\mathcal{T}_0 = \{-T_0 + 1, \ldots, 0\}$ and $\mathcal{T}_1 = \{1, \ldots, T_1\}$ as T_0 pretreatment and T_1 posttreatment periods, respectively. Potential outcomes are denoted by $y_{i,t}^I$ when unit i is treated at time t and by $y_{i,t}^N$ when not treated.

Following Ferman and Pinto (2021), we assume that the potential outcomes are generated from a factor structure, i.e.,

$$
\begin{align}
y_{i,t}^N &= \boldsymbol{\theta}_t^\top \mathbf{Z}_i + \boldsymbol{\lambda}_t^\top \boldsymbol{\mu}_i + \delta_t + c_i + \epsilon_{i,t}, \tag{1} \\
y_{i,t}^I &= \alpha_{i,t} + y_{i,t}^N, \tag{2}
\end{align}
$$

where $\boldsymbol{\lambda}_t = (\lambda_{1,t}, \ldots, \lambda_{F,t})^\top$ is an $F \times 1$ vector of unobserved common factors with unknown factor loadings $\boldsymbol{\mu}_i = (\mu_{i,1}, \ldots, \mu_{i,F})^\top$ $(F \times 1)$, $\mathbf{Z}_i$ is an $r \times 1$ vector of observed time-invariant covariates that are not affected by the treatment, $\boldsymbol{\theta}_t$ $(r \times 1)$ represents the associated unknown parameters, δ_t can be interpreted as an unknown common factor with homogeneous loadings across units, c_i captures the unobserved individual heterogeneity, and $\epsilon_{i,t}$ is the idiosyncratic shock. Here both δ_t and c_i can be included in the linear factor structure $\boldsymbol{\lambda}_t^\top \boldsymbol{\mu}_i$; thus, our setup also encompasses Abadie et al. (2010), Botosaru and Ferman (2019), Powell (2022) and Ferman (2021), among others. In this article, we treat r and F as fixed constants. We denote $\mathbf{M} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_J)$, $\boldsymbol{\epsilon}_t = (\epsilon_{0,t}, \ldots, \epsilon_{J,t})^\top$ and $\mathbf{c} = (c_1, \ldots, c_J)^\top$. All matrices and vectors are marked in bold. The linear factor model seems a general and benchmark setup for most theoretical analysis of SCM and are particularly useful to illustrate some properties. Nonetheless, considering that the SCM may still be applicable even without specifying an outcome model, we shall relax this model assumption and analyze the properties of SCM in

a model-free setup in Section 3.3.

The interest is in estimating the treatment effect $\alpha_{i,t}$ of the single treated unit $(i = 0)$ for $t \in \mathcal{T}_1$. However, in practice, one cannot observed all potential outcomes but only the realized outcomes given as

$$y_{i,t} = d_{i,t} y_{i,t}^I + (1 - d_{i,t}) y_{i,t}^N, \quad i \in \{0, 1, \ldots, J\} \text{ and } t \in \mathcal{T}_0 \cup \mathcal{T}_1, \tag{3}$$

with $d_{i,t}$ being a dummy that equals 1 for the treated unit $i = 0$ at $t \in \mathcal{T}_1$, zero otherwise. The idea of the SCM is to construct an SC unit out of the donor pool, whose outcome serves as a proxy for the counterfactual outcome of the treated unit, i.e.,

$$\widehat{\alpha}_{0,t}(\widehat{\mathbf{w}}) = y_{0,t}^I - \widehat{y}_{0,t}^N(\widehat{\mathbf{w}}) = y_{0,t} - \sum_{j=1}^{J} \widehat{w}_j y_{j,t}, \quad t \in \mathcal{T}_1,$$

where $\widehat{\mathbf{w}}$ is the SC weight determined by some predictors. We focus on the popular specification where the predictors include all pretreatment outcomes (see, e.g., Doudchenko and Imbens, 2016; Ferman and Pinto, 2021; Ferman, 2021), and we also allow for observed covariates. Let $\mathbf{X}_i = \left(y_{i,-T_0+1}^N, \ldots, y_{i,0}^N, \mathbf{Z}_i^\top \right)^\top = \left(y_{i,-T_0+1}, \ldots, y_{i,0}, \mathbf{Z}_i^\top \right)^\top$ be a $(T_0+r) \times 1$ vector of the pretreatment characteristics of unit i for all i, and define $\boldsymbol{\mathcal{X}}_c = (\mathbf{X}_1, \ldots, \mathbf{X}_J)$ with subscript c representing control units. The original SCM restricts the weights to be in the set $\mathcal{H}_{\text{orig}} = \left\{ \mathbf{w} = (w_1, \ldots, w_J)^\top \in [0,1]^J \mid \sum_{j=1}^J w_j = 1 \right\}$, such that control units are combined convexly. To allow for possible extrapolation, we relax the positivity constraint and consider $\mathcal{H} = \left\{ \mathbf{w} = (w_1, \ldots, w_J)^\top \in [-C_L, C_U]^J \mid \sum_{j=1}^J w_j = 1 \right\}$, where C_L and C_U are two nonnegative constants. Clearly, $\mathcal{H} \supseteq \mathcal{H}_{\text{orig}}$, and thus our results also hold for the original SC weight. Note that since $\mathcal{H}$ is bounded, the range of extrapolation it permits is bounded, and the sum-up-to-unity restriction still plays a crucial role.

For any given positive definite matrix $\mathbf{V}$, the SC weight can be obtained by solving the following optimization:

$$\widehat{\mathbf{w}}(\mathbf{V}) = \arg\min_{\mathbf{w} \in \mathcal{H}_{\text{orig}}} \sqrt{(\mathbf{X}_0 - \boldsymbol{\mathcal{X}}_c \mathbf{w})^\top \mathbf{V} (\mathbf{X}_0 - \boldsymbol{\mathcal{X}}_c \mathbf{w})}.$$

For the sake of simplicity, we follow Ferman and Pinto (2021) and Ferman (2021) to set $\mathbf{V} = \mathbf{I}_{T_0+r}$, where $\mathbf{I}_{T_0+r}$ denotes the $(T_0+r) \times (T_0+r)$ identity matrix. Then, the SC weight

7

can be obtained by

$$
\begin{aligned}
\widehat{\mathbf{w}} &= \underset{\mathbf{w}\in\mathcal{H}}{\arg\min} L_{T_0}(\mathbf{w}) \\
&= \underset{\mathbf{w}\in\mathcal{H}}{\arg\min} \frac{1}{T_0} \|\mathbf{X}_0 - \boldsymbol{\mathcal{X}}_c\mathbf{w}\|^2 \\
&= \underset{\mathbf{w}\in\mathcal{H}}{\arg\min} \frac{1}{T_0} \left\{ \sum_{t\in\mathcal{T}_0} \left(y_{0,t} - \sum_{j=1}^{J} w_j y_{j,t} \right)^2 + \left\| \mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j \right\|^2 \right\},
\end{aligned}
\tag{4}
$$

where $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top \mathbf{x}}$ for any vector $\mathbf{x}$. We first analyze the properties of the SC weight defined as above. In Section 3.2.2, we consider adding an intercept to (4), such that the SCM can also be compared with the difference-in-differences (DID) estimator.

3 Theoretical results

This section presents the main theoretical results. Considering the fact that MSPE is widely used to evaluate the accuracy of treatment effect estimates, we focus on the (expectation of) MSPE and study how the SC estimator and weights are related to their optimal counterparts in terms of minimizing the MSPE.

We first show that the SC weights converge to the infeasible optimal weights that minimize the risk of MSPEs. Our convergence results suggest that the SCM generally leads to a biased treatment effect estimator but potentially with a reduced variance. We also complement the existing asymptotic analysis of the SC weight by quantifying the rate of convergence. Furthermore, we establish the asymptotic optimality of the SC estimator in the sense that it achieves the lowest possible loss among all possible averaging estimators of the control units. Unless otherwise stated, all limiting properties hold when the number of pretreatment periods goes to infinity, i.e., $T_0 \to \infty$.

3.1 Convergence of the SC weight

To show the convergence, we assume that the following regularity conditions hold.

Condition 1.

(i) We treat $\{\boldsymbol{\mu}_i, c_i, \mathbf{Z}_i \mid i \in \{0, 1, \ldots, J\}\}$, $\{\boldsymbol{\lambda}_t, \delta_t \mid t \in \mathcal{T}_0 \cup \mathcal{T}_1\}$ and $\{\alpha_{0,t} \mid t \in \mathcal{T}_1\}$ as fixed and $\{\epsilon_{i,t} \mid i \in \{0, 1, \ldots, J\}, t \in \mathcal{T}_0 \cup \mathcal{T}_1\}$ as stochastic.

(ii) $\mathbb{E}\epsilon_{i,t} = 0$ for $i \in \{0, 1, \ldots, J\}$ and $t \in \mathcal{T}_0 \cup \mathcal{T}_1$.

Condition 1 concerns the sampling of data. Condition 1 (i) is to simplify the proof, and a similar assumption is also used in Ferman and Pinto (2021, Assumption 2.1) and Ferman (2021, Assumption 2). Condition 1 (ii) requires the idiosyncratic shocks to have a zero-mean, which appears as a rather standard assumption in the SCM literature (see, e.g., Xu, 2017; Botosaru and Ferman, 2019; Ferman, 2021; Ben-Michael et al., 2021, among others). It can be interpreted as "selection on unobservables" because it allows for any form of dependence between the treatment assignment and the factor structure as long as the assignment is uncorrelated with idiosyncratic shocks (Ferman and Pinto, 2021).

Condition 2.

(i) $T_1^{-1} \sum_{t \in \mathcal{T}_1} \left(T_0^{-1} \sum_{k \in \mathcal{T}_0} \boldsymbol{\lambda}_k^\top \boldsymbol{\lambda}_k - \boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_t \right) = O\left(T_0^{-1/2} \right).$

(ii) $T_1^{-1} \sum_{t \in \mathcal{T}_1} \left(T_0^{-1} \sum_{k \in \mathcal{T}_0} \delta_k^2 - \delta_t^2 \right) = O\left(T_0^{-1/2} \right).$

Condition 2 requires that the variation of the common factors does not change substantially after treatment. This condition is in line with the rationality of the SCM since it implies that the main difference between the pre- and posttreatment outcomes is exclusively due to the treatment effect. Only in this case can the pretreatment data be used to construct the SC unit and estimate the treatment effect for the posttreatment periods. If we treat $\{\boldsymbol{\lambda}_t\}$ and $\{\delta_t\}$ as stochastic and assume T_1 to be divergent at rate $O(T_0)$, then using the central limit theorem for dependent observations (see, e.g., Theorems 5.16 and 5.20 in White, 1984), we can obtain that

$$\frac{1}{\sqrt{T_0}} \sum_{t \in \mathcal{T}_0} \boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_t = T_0^{-1/2} \sum_{t \in \mathcal{T}_0} \mathbb{E}\boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_t + O_p(1), \quad \frac{1}{\sqrt{T_0}} \sum_{t \in \mathcal{T}_1} \boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_t = T_1^{-1/2} \sum_{t \in \mathcal{T}_1} \mathbb{E}\boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_t + O_p(1),$$

$$\frac{1}{\sqrt{T_0}} \sum_{t \in \mathcal{T}_0} \delta_t^2 = T_0^{-1/2} \sum_{t \in \mathcal{T}_0} \mathbb{E}\delta_t^2 + O_p(1), \qquad \frac{1}{\sqrt{T_0}} \sum_{t \in \mathcal{T}_1} \delta_t^2 = T_1^{-1/2} \sum_{t \in \mathcal{T}_1} \mathbb{E}\delta_t^2 + O_p(1).$$

In this case, Condition 2 is satisfied in probability under certain stability assumptions, such as Assumptions 4–5 in Ferman and Pinto (2021). This condition also allows for "divergent"

factors, i.e., $T_0^{-1} \sum_{t \in \mathcal{T}_0} \boldsymbol{\lambda}_t \to \infty$. To see this more explicitly, consider a simple example where $T_1 = T_0$, $F = 1$ and $\lambda_{1,t} = |t - t_0|^{1/2}$ for $t \in \mathcal{T}_0 \cup \mathcal{T}_1$, where t_0 denotes the time of treatment and $t_0 = 0$ with our notation. In this example, $T_0^{-1} \sum_{t \in \mathcal{T}_0} \boldsymbol{\lambda}_t = T_0^{-1} \sum_{t=1}^{T_0} t^{1/2}$ goes to infinity, and thus $\boldsymbol{\lambda}_t$ is a divergent factor. The condition is satisfied because $T_1^{-1} \sum_{t \in \mathcal{T}_1} \left(T_0^{-1} \sum_{k \in \mathcal{T}_0} \boldsymbol{\lambda}_k^\top \boldsymbol{\lambda}_k - \boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_t \right) = 0$. Nevertheless, not all kinds of divergent factors satisfy this condition, e.g., $\lambda_{1,t} = t$.

Condition 3. *There exists a constant C_0 such that $\mu_{il} < C_0$ and $c_i < C_0$ for $i \in \{0, 1, \ldots, J\}$ and $l \in \{1, \ldots, F\}$.*

Condition 3 requires the uniform boundedness of the factor loadings and time-invariant fixed effects. This condition easily holds under the often assumed conditions for factor identifiability, e.g., $\mathbf{M}\mathbf{M}^\top$ being a diagonal matrix (Xu, 2017) or $T_0^{-1}\mathbf{M}\mathbf{M}^\top \to \mathbf{I}_F$ (Stock and Watson, 2002), where we recall that $\mathbf{M} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_J)$. The same condition is also used in Ferman (2021, Assumption 3.2(b)). The uniform boundedness of c_i is also a mild condition once we note that c_i can be regarded as the loading of the factor $\lambda_t = 1$.

Condition 4.

(i) $T_1^{-1} \sum_{t \in \mathcal{T}_1} \left(T_0^{-1} \sum_{k \in \mathcal{T}_0} \boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k - \boldsymbol{\theta}_t^\top \boldsymbol{\theta}_t \right) = O(T_0^{-1/2})$.

(ii) *There exists a constant C_z such that $\|\mathbf{Z}_i\| < C_z$ for $i \in \{0, 1, \ldots, J\}$.*

Condition 4 concerns the variability of the observed covariates and their corresponding coefficients, and Conditions 4 (i) and (ii) can be viewed as covariate versions of Conditions 2 and 3, respectively. Specifically, Condition 4 (i) requires that the change in the coefficients before and after the treatment is limited, and it obviously holds for any time-invariant coefficients. Condition 4 (ii) requires uniformly bounded covariates across units. Both parts of Condition 4 can be justified in the same way as Conditions 2 and 3. Note that Ferman (2021) assumes that the observed covariates and their coefficients satisfy the same set of conditions for factors and loadings when he proves the results with covariates (see Section A.2.5 of Supplementary Material in Ferman (2021)), which resembles our strategy here.

We also need some restrictions on the relation between the idiosyncratic shock of the treated and control units. Let $e_{t,\epsilon}^{(i)} = \epsilon_{0,t} - \epsilon_{i,t}$ for $i \in \{1, \ldots, J\}$ and $t \in \mathcal{T}_0 \cup \mathcal{T}_1$, $R_{T_0}(\mathbf{w}) = \mathbb{E}L_{T_0}(\mathbf{w})$ and $\xi_{T_0} = \inf_{\mathbf{w} \in \mathcal{H}} R_{T_0}(\mathbf{w})$. Intuitively, we can interpret ξ_{T_0} as a measure of pre-treatment fit.

10

Condition 5. $\sup_{\mathbf{w} \in \mathcal{H}} \left| T_0^{-1} \sum_{t \in \mathcal{T}_0} \mathbb{E} \left(\sum_{j=1}^J w_j e_{t,\epsilon}^{(j)} \right)^2 - T_1^{-1} \sum_{t \in \mathcal{T}_1} \mathbb{E} \left(\sum_{j=1}^J w_j e_{t,\epsilon}^{(j)} \right)^2 \right| = o\left(\xi_{T_0} \right).$

This condition means that, with respect to the goodness of pretreatment fit, the difference in idiosyncratic shocks between the treated and any weighted average of control units does not change substantially after treatment. Generally, it is more likely to hold when the idiosyncratic shocks are more stable over time or more alike across units. For example, in the setting of Ferman (2021) where $\{\epsilon_{0,t}, \epsilon_{1t}, \dots, \epsilon_{j,t}\}$ are of zero-mean and independent across units, we have

$$\mathbb{E} \left(\sum_{j=1}^J w_j e_{t,\epsilon}^{(j)} \right)^2 = \mathbb{E} \left(\epsilon_{0,t}^2 \right) + \sum_{j=1}^J w_j^2 \mathbb{E} \left(\epsilon_{j,t}^2 \right), \quad t \in \mathcal{T}_0 \cup \mathcal{T}_1.$$

If we further assume that the sequence $\{\epsilon_{i,t}\}_{t \in \mathcal{T}_0 \cup \mathcal{T}_1}$ is stationary, then Condition 5 holds, because

$$\sup_{\mathbf{w} \in \mathcal{H}} \left| \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \mathbb{E} \left(\sum_{j=1}^J w_j e_{t,\epsilon}^{(j)} \right)^2 - \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \mathbb{E} \left(\sum_{j=1}^J w_j e_{t,\epsilon}^{(j)} \right)^2 \right|$$

$$= \sup_{\mathbf{w} \in \mathcal{H}} \left| \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \mathbb{E} \left(\epsilon_{0,t}^2 \right) + \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \sum_{j=1}^J w_j^2 \mathbb{E} \left(\epsilon_{j,t}^2 \right) \right.$$

$$\left. - \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \mathbb{E} \left(\epsilon_{0,t}^2 \right) - \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \sum_{j=1}^J w_j^2 \mathbb{E} \left(\epsilon_{j,t}^2 \right) \right|$$

$$= 0.$$

To state the next condition, define $\mathbf{Y}_i = (y_{i,-T_0+1}, \dots, y_{i,0})^\top$ for $i \in \{0, 1, \dots, J\}$, $\boldsymbol{\mathcal{Y}}_c = (\mathbf{Y}_1, \dots, \mathbf{Y}_J)$ and $\boldsymbol{\Sigma} = T_0^{-1} \mathbb{E} \left(\boldsymbol{\mathcal{Y}}_c^\top \boldsymbol{\mathcal{Y}}_c \right)$, where the subscript "$c$" indicates control units. We use $\lambda_{min}(\cdot)$ and $\lambda_{max}(\cdot)$ to represent the minimum and maximum eigenvalue of a matrix.

Condition 6. *There exist constants κ_1 and κ_2 such that $0 < \kappa_1 \leq \lambda_{min}(\boldsymbol{\Sigma}) \leq \lambda_{max}(\boldsymbol{\Sigma}) \leq \kappa_2$.*

This condition bounds the variability of the pretreatment outcomes of control units from both below and above. It can also be satisfied under many standard setups of the SCM. For example, consider pure factor models as the DGP, as in Hsiao and Zhou (2019) and Ben-Michael et al. (2021), i.e., $y_{i,t}^N = \boldsymbol{\lambda}_t^\top \boldsymbol{\mu}_i + \epsilon_{i,t}$, and it can be written in a matrix form as $\boldsymbol{\mathcal{Y}}_c = \boldsymbol{\Lambda}^{-\top} \mathbf{M} + \boldsymbol{\epsilon}^-$, where $\boldsymbol{\Lambda}^- = (\boldsymbol{\lambda}_{-T_0+1}, \dots, \boldsymbol{\lambda}_0)$, $\mathbf{M} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_J)$ and $\boldsymbol{\epsilon}^- = (\boldsymbol{\epsilon}_{-T_0+1}, \dots, \boldsymbol{\epsilon}_0)$

being a $T_0 \times J$ matrix of idiosyncratic shocks with $\boldsymbol{\epsilon}_t = (\epsilon_{0,t}, \ldots, \epsilon_{J,t})^\top$. In this case, we have $\boldsymbol{\Sigma} = T_0^{-1}\mathbf{M}^\top \boldsymbol{\Lambda}^-\boldsymbol{\Lambda}^{-\top}\mathbf{M} + T_0^{-1}\mathbb{E}\boldsymbol{\epsilon}^{-\top}\boldsymbol{\epsilon}^-$. Note that

$$\lambda_{min}(\mathbf{A}) + \lambda_{min}(\mathbf{B}) \le \lambda_{min}(\mathbf{A} + \mathbf{B}) \le \lambda_{max}(\mathbf{A} + \mathbf{B}) \le \lambda_{max}(\mathbf{A}) + \lambda_{max}(\mathbf{B}),$$

for any Hermite matrices $\mathbf{A}$ and $\mathbf{B}$ of the same order. Thus, we only need to study $\lambda_{min}(T_0^{-1}\mathbf{M}^\top\boldsymbol{\Lambda}^-\boldsymbol{\Lambda}^{-\top}\mathbf{M})$ and $\lambda_{min}(T_0^{-1}\mathbb{E}\boldsymbol{\epsilon}^{-\top}\boldsymbol{\epsilon}^-)$ for the lower bound of $\lambda_{min}(\boldsymbol{\Sigma})$ and study $\lambda_{max}(T_0^{-1}\mathbf{M}^\top\boldsymbol{\Lambda}^-\boldsymbol{\Lambda}^{-\top}\mathbf{M})$ and $\lambda_{max}(T_0^{-1}\mathbb{E}\boldsymbol{\epsilon}^{-\top}\boldsymbol{\epsilon}^-)$ for the upper bound of $\lambda_{max}(\boldsymbol{\Sigma})$. Under Condition 1 (ii), $\lambda_{min}(T_0^{-1}\mathbb{E}\boldsymbol{\epsilon}^{-\top}\boldsymbol{\epsilon}^-)$ and $\lambda_{max}(T_0^{-1}\mathbb{E}\boldsymbol{\epsilon}^{-\top}\boldsymbol{\epsilon}^-)$ are both finite as long as the idiosyncratic shock has a finite variance. For the other two terms involving factors and loadings, it is common to assume that $T_0^{-1}\boldsymbol{\Lambda}^-\boldsymbol{\Lambda}^{-\top} = \mathbf{I}_F$ when the factor is not divergent (see, e.g., Bai, 2009; Xu, 2017; Ben-Michael et al., 2021). Thus, $\lambda_{min}(T_0^{-1}\mathbf{M}^\top\boldsymbol{\Lambda}^-\boldsymbol{\Lambda}^{-\top}\mathbf{M}) = \lambda_{min}(\mathbf{M}^\top\mathbf{M})$ and $\lambda_{max}(T_0^{-1}\mathbf{M}^\top\boldsymbol{\Lambda}^-\boldsymbol{\Lambda}^{-\top}\mathbf{M}) = \lambda_{max}(\mathbf{M}^\top\mathbf{M})$. Therefore, Condition 6 can be simplified to stating that there exist constants $\widetilde{\kappa}_1$ and $\widetilde{\kappa}_2$ such that

$$0 < \widetilde{\kappa}_1 \le \lambda_{min}(\mathbf{M}^\top\mathbf{M}) \le \lambda_{max}(\mathbf{M}^\top\mathbf{M}) < \widetilde{\kappa}_2,$$

which is often used in factor models and resembles the rank condition in standard regressions (Bai, 2009; Xu, 2017). As for the diverging factors, Condition 6 can be satisfied with a more restrictive assumption on the loadings than $T_0^{-1}\boldsymbol{\Lambda}^-\boldsymbol{\Lambda}^{-\top} = \mathbf{I}_F$.

To evaluate the performance of the SC treatment effect estimator, we consider the MSPE for some weight $\mathbf{w} \in \mathcal{H}$, defined as

$$
\begin{aligned}
L_{T_1}(\mathbf{w}) &= \frac{1}{T_1}\sum_{t \in \mathcal{T}_1}\{\alpha_{0,t} - \widehat{\alpha}_{0,t}(\mathbf{w})\}^2 \\
&= \frac{1}{T_1}\sum_{t \in \mathcal{T}_1}\{y_{0,t}^N - \widehat{y}_{0,t}^N(\mathbf{w})\}^2 \\
&= \frac{1}{T_1}\sum_{t \in \mathcal{T}_1}\left(y_{0,t}^N - \sum_{j=1}^J w_j y_{j,t}\right)^2,
\end{aligned}
\tag{5}
$$

and its risk is $R_{T_1}(\mathbf{w}) = \mathbb{E}L_{T_1}(\mathbf{w})$. The optimal weight vector for a given T_1 is defined as the minimizer of the risk, i.e.,

$$\mathbf{w}_{T_1}^{\text{opt}} = \arg\min_{\mathbf{w} \in \mathcal{H}} R_{T_1}(\mathbf{w}).
\tag{6}$$

Theorem 1. *Given any T_1, if $\mathbf{w}_{T_1}^{opt}$ is an interior point of $\mathcal{H}$ and Conditions 1–6 hold, then*

$$\left\|\widehat{\mathbf{w}} - \mathbf{w}_{T_1}^{opt}\right\| = O_p\left(T_0^{\nu}\xi_{T_0}^{1/2} + T_0^{\nu}\xi_{T_1}^{1/2} + T_0^{-1/4+\nu}J\right),$$

where $\nu > 0$ is a sufficiently small constant, $\xi_{T_0} = \inf_{\mathbf{w}\in\mathcal{H}} R_{T_0}(\mathbf{w})$, and $\xi_{T_1} = \inf_{\mathbf{w}\in\mathcal{H}} R_{T_1}(\mathbf{w})$.

Theorem 1 shows that as T_0 and $T_1 \to \infty$, the SC weight $\widehat{\mathbf{w}}$ converges to the optimal weight vector sequence $\mathbf{w}_{T_1}^{\mathrm{opt}}$ at a rate that depends on ξ_{T_0}, ξ_{T_1} and J.[3] We discuss the roles of ξ_{T_0}, ξ_{T_1}, T_0 and J in turn. First, a faster rate of ξ_{T_0} and ξ_{T_1} going to zero implies quicker convergence of $\widehat{\mathbf{w}}$. Recall that ξ_{T_0} is a measure of pretreatment fit, and thus the theorem clearly links a good pretreatment fit with accurate weight estimation. Second, the number of pretreatment periods T_0 plays a mixed role in the convergence rate, depending on the goodness of pre- and posttreatment fit. If the fit is good in both regimes, then the dominant term is $T_0^{-1/4+\nu}J$, and an increase in T_0 improves the accuracy of estimated weights. However, when the fit is poor, the net effect of increasing T_0 may be detrimental to weight convergence. This result is not surprising because a poor fit implies that the predictors are not informative for constructing the SC unit. Thus, increasing the sample size of uninformative or even misleading predictors drives the SC weight further from optimality. Finally, a larger J is associated with a slower convergence rate. Note that this result does not conflict with the asymptotic unbiasedness of the SC estimator when $J \to \infty$ shown by Ferman (2021), because the SC weight may still converge to the unbiased limiting weight but just at a lower rate (see below for further detail on the relation with Ferman (2021)). Considering that J corresponds to the number of weight parameters to be estimated, less accurate estimates are expected when the dimension of parameters increases. The negative role of J in the rate of convergence is consistent with the conjecture of Ferman (2021) that stronger assumptions on the moments of $\epsilon_{i,t}$ are required when J diverges at a faster rate.

We discuss how Theorem 1 is related to the asymptotic result of Ferman and Pinto (2021). First, we show that the limit of the SC weight defined in (6) is compatible with the infeasible weight studied in Ferman and Pinto (2021), which we denoted as $\mathbf{w}_{\mathrm{FP}}^*$ and satisfies $(\boldsymbol{\mu}_0^\top, c_0)^\top \neq (\mathbf{M}^\top, \mathbf{c})^\top \mathbf{w}_{\mathrm{FP}}^*$. To remain close to the benchmark setup of Ferman and Pinto (2021), we consider a simplified version of our DGP without observed covariates and assume that $T_0^{-1}\sum_{t\in\mathcal{T}_0} \boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^\top \to_p \sigma_{\boldsymbol{\epsilon}}^2\mathbf{I}_{J+1}$ and $T_0^{-1}\sum_{t\in\mathcal{T}_0} \boldsymbol{\lambda}_t\boldsymbol{\lambda}_t^\top \to_p \boldsymbol{\Omega}_0$, where $\boldsymbol{\Omega}_0$ is a positive semidefinite

---

[3]The posttreatment periods T_1 plays a role in convergence via ξ_{T_1}.

13

matrix. Ferman and Pinto (2021) shows that the original SC weight ($\mathbf{w} \in \mathcal{H}_{\text{orig}}$) converges in probability to $\mathbf{w}_{\text{FP}}^*$ that minimizes the following quantity

$$Q_0(\mathbf{w}) = \left\{ \left(c_0 - \mathbf{c}^\top \mathbf{w}\right)^2 + \left(\boldsymbol{\mu}_0 - \mathbf{M}\mathbf{w}\right)^\top \boldsymbol{\Omega}_0 \left(\boldsymbol{\mu}_0 - \mathbf{M}\mathbf{w}\right) \right\} + \sigma_\epsilon^2 \left(1 + \mathbf{w}^\top \mathbf{w}\right).$$

Recall that the limiting weight we study minimizes $R_{T_1}(\mathbf{w})$. Thus, we examine how $R_{T_1}(\mathbf{w})$ is related to $Q_0(\mathbf{w})$. Denote $\boldsymbol{\epsilon}_{c,t} = (\epsilon_{1,t}, \ldots, \epsilon_{J,t})^\top$ as the error vector of control units. Then, $R_{T_1}(\mathbf{w})$ can be decomposed as

$$
\begin{aligned}
R_{T_1}(\mathbf{w}) &= \left\{ \left(c_0 - \mathbf{c}^\top \mathbf{w}\right)^2 + \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \left(\boldsymbol{\mu}_0 - \mathbf{M}\mathbf{w}\right)^\top \boldsymbol{\lambda}_t \boldsymbol{\lambda}_t^\top \left(\boldsymbol{\mu}_0 - \mathbf{M}\mathbf{w}\right) \right\} \\
&\quad + \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \mathbb{E} \left(\epsilon_{0,t} - \boldsymbol{\epsilon}_{c,t}^\top \mathbf{w}\right)^2.
\end{aligned}
\tag{7}
$$

We can show that $R_{T_1}(\mathbf{w}) \to_p Q_0(\mathbf{w})$ for any $\mathbf{w} \in \mathcal{H}$ if $\{\boldsymbol{\lambda}_t\}$ and $\{\epsilon_{i,t}\}$ are both stationary for all t such that $T_1^{-1} \sum_{t \in \mathcal{T}_1} \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t^\top \to_p \sigma_\epsilon^2 \mathbf{I}_{J+1}$ and $T_1^{-1} \sum_{t \in \mathcal{T}_1} \boldsymbol{\lambda}_t \boldsymbol{\lambda}_t^\top \to_p \boldsymbol{\Omega}_0$. This result implies that minimizing $R_{T_1}(\mathbf{w})$ is asymptotically equivalent to minimizing $Q_0(\mathbf{w})$, and thus our weight limit $\mathbf{w}_{T_1}^{\text{opt}}$ is asymptotically identical to the limit $\mathbf{w}_{\text{FP}}^*$ considered in Ferman and Pinto (2021) if the factors and idiosyncratic shocks are stationary. Due to this asymptotic equivalence, $\mathbf{w}_{T_1}^{\text{opt}}$ also fails to recover the factor structure. To see this more explicitly, note from (7) that $R_{T_1}(\mathbf{w})$ is composed of two parts of errors when approximating the treated unit using the SC unit: the error of approximating the factors and the error of approximating the idiosyncratic shock, similar to the decomposition of $Q_0(\mathbf{w})$. As a result, $\mathbf{w}_{T_1}^{\text{opt}}$ needs to balance the two parts of errors in $R_{T_1}(\mathbf{w})$, thus deviating from the minimizer of the first part and failing to recover the true factor structure, i.e., $(\boldsymbol{\mu}_0^\top, c_0)^\top \neq (\mathbf{M}^\top, \mathbf{c})^\top \mathbf{w}_{T_1}^{\text{opt}}$. It further implies that the resulting SC estimator generally does not converge to the targeted treatment effect $\alpha_{0,t}$, confirming the conclusion of Ferman and Pinto (2021). We complement Ferman and Pinto (2021) by quantifying the rate of convergence.

The result in Theorem 1 also offers a verification of the important assumption (Assumption 3.2) of Ferman (2021) to guarantee the asymptotic unbiasedness of the SC estimator, i.e., there exists a weight vector $\mathbf{w}^* \in \mathcal{H}_{\text{orig}}$ such that $\|\mathbf{w}^*\| \to_p 0$ and $\|\boldsymbol{\mu}_0 - \mathbf{M}\mathbf{w}^*\| \to_p 0$. We show that as J diverges, the optimal weight $\mathbf{w}_{T_1}^{\text{opt}}$ defined by (6) is a candidate choice that could satisfy Assumption 3.2 of Ferman (2021), i.e., $\left\|\mathbf{w}_{T_1}^{\text{opt}}\right\| \to 0$ and $\left\|\boldsymbol{\mu}_0 - \mathbf{M}\mathbf{w}_{T_1}^{\text{opt}}\right\| \to 0$. To see this more explicitly, we focus on the weight $\mathbf{w}_{T_1}^{\text{opt}} \in \mathcal{H}_{\text{orig}}$ and follow Ferman (2021) to

assume that $\{\epsilon_{i,t}\}_{t \in \mathcal{T}_0 \cup \mathcal{T}_1}$ are independent across i and $\text{var}(\epsilon_{i,t}) = \sigma_\epsilon^2$ for the sake of simplification. We can (re-)define the factors and loadings, such that c_i is absorbed into the factor structure $\boldsymbol{\lambda}_t^\top \boldsymbol{\mu}_i$ as Ferman (2021). Thus, $R_{T_1}(\mathbf{w})$ can be written as

$$R_{T_1}(\mathbf{w}) \;=\; \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} (\boldsymbol{\mu}_0 - \mathbf{Mw})^\top \boldsymbol{\lambda}_t \boldsymbol{\lambda}_t^\top (\boldsymbol{\mu}_0 - \mathbf{Mw}) + \sigma_\epsilon^2 \left(1 + \mathbf{w}^\top \mathbf{w}\right).$$

Denote $\mathbf{Q} = T_1^{-1} \sum_{t \in \mathcal{T}_1}^\top \boldsymbol{\lambda}_t \boldsymbol{\lambda}_t^\top$. We can analytically obtain the solution of $\mathbf{w}_{T_1}^{\text{opt}}$ by solving the optimization (6) with Karush-Kuhn-Tucker conditions as

$$\mathbf{w}_{T_1}^{\text{opt}} = \left(\mathbf{M}^\top \mathbf{Q} \mathbf{M} + \sigma_\epsilon^2 \mathbf{I}_J\right)^{-1} \left(\mathbf{M}^\top \mathbf{Q} \boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{\rho}_1 - \frac{1}{2}\rho_2 \boldsymbol{\iota}_J\right), \tag{8}$$

where $\boldsymbol{\rho}_1$ is a nonnegative $J \times 1$ constant vector, ρ_2 is a constant, and $\boldsymbol{\iota}_J$ is a $J \times 1$ vector of ones. In the Online Appendix, we show that $\left\|\boldsymbol{\mu}_0 - \mathbf{Mw}_{T_1}^{\text{opt}}\right\| \to 0$ and $\left\|\mathbf{w}_{T_1}^{\text{opt}}\right\| \to 0$ hold if there exist positive constants c_1, c_2, c_3 and c_4 such that

$$c_1 \leq \lambda_{min}\left(J^{-1}\mathbf{MM}^\top\right) \leq \lambda_{max}\left(J^{-1}\mathbf{MM}^\top\right) \leq c_2 \tag{9}$$

and

$$c_3 \leq \lambda_{min}\left(\mathbf{Q}\right) \leq \lambda_{max}\left(\mathbf{Q}\right) \leq c_4. \tag{10}$$

Hence, our analysis of the convergence of SC weights provides conditions under which Assumption 3.2 of Ferman (2021) holds such that the asymptotic unbiasedness of the SC estimator can be achieved when J diverges. Intuitively, it requires that the information contained in the factor structure neither diminishes nor diverges. Such upper and lower bounds guarantee that the factors are informative (to be able to reconstruct) and do not dominate the idiosyncratic shocks (so that the weights can dilute among control units), respectively.

Theorem 1 also provides a new perspective for understanding the role of covariates. Botosaru and Ferman (2019) shows that a perfect fit of observed covariates is not essential to achieve the asymptotic unbiasedness of SC estimators as long as the fit of outcomes is good, but a better fit of covariatesis associated with tighter bounds. Theorem 1 illustrates the role of covariates via the convergence rate, and it suggests that a better fit of covariates (hence a smaller ξ_{T_0}) helps promote the convergence of the SC weight. This result is in line

with the study of bounds by Botosaru and Ferman (2019).

To conclude this subsection, the above discussion shows that the SC estimator constructed using the limiting optimal weight $\mathbf{w}_{T_1}^{\mathrm{opt}}$ minimizes the expected MSPE but also suffers from an asymptotic bias under fixed J. This result suggests a bias-variance tradeoff when using the averaged outcome of the control units for treatment effect evaluation and motivates us to further study how the SC weight $\widehat{\mathbf{w}}$ compare with other weighting schemes in terms of MSPE.

3.2 Asymptotic optimality of SC estimators

In this subsection, we investigate how the SC weights balance the bias and variance. We first establish the asymptotic optimality for SC estimators without an intercept and then consider the case with an intercept. We also compare the SC estimator with other treatment effect estimators that also involve a weighted average of control units.

3.2.1 Asymptotic optimality of SCM without an intercept

We need some additional conditions.

Condition 7. *For any $i \in \{0, 1, \ldots, J\}$, $\{\epsilon_{i,t}\}$ is either α-mixing with the mixing coefficient $\alpha = -r/(r-2)$ or ϕ-mixing with the mixing coefficient $\phi = -r/(2r-1)$ for $r \geq 2$.*

Condition 7 restricts the dependence of the idiosyncratic shocks. A similar assumption is needed in Ferman (2021) (see Assumption 3.1 (b)), while Abadie et al. (2010) imposes a stronger requirement that $\{\epsilon_{i,t}\}$ are independent across units and over time.

Condition 8.

(i) *There exists a constant C_1 such that $\mathbb{E}\epsilon_{i,t}^4 \leq C_1 < \infty$ for $i \in \{0, 1, \ldots, J\}$ and $t \in \mathcal{T}_0 \cup \mathcal{T}_1$.*

(ii) *There exists a constant C_2 such that $\mathrm{var}\left(T_0^{-1/2} \sum_{t \in \mathcal{T}_0} e_{t,\epsilon}^{(i)} e_{t,\epsilon}^{(j)}\right) \geq C_2 > 0$ for all T_0 sufficiently large and any $i, j \in \{1, \ldots, J\}$.*

(iii) *There exists a constant C_3 such that $\mathrm{var}\left(T_1^{-1/2} \sum_{t \in \mathcal{T}_1} e_{t,\epsilon}^{(i)} e_{t,\epsilon}^{(j)}\right) \geq C_3 > 0$ for all T_1 sufficiently large and any $i, j \in \{1, \ldots, J\}$.*

Condition 8 provides a set of regularity conditions to apply the central limit theorem for the dependent process; see Schönfeld (1971), Scott (1973) and Wooldridge and White (1988). Condition 8 (i) requires that all idiosyncratic shocks should not have heavy tails such that their fourth moments can be uniformly bounded. The same condition is also imposed in Xu (2017, Assumption 4), Botosaru and Ferman (2019, Assumption 3.1 (c)) and Ferman (2021, Assumption 3.1 (c)). Conditions 8 (ii)–(iii) concern the difference between the idiosyncratic shock of the treated and control units. They guarantee that the variances of shocks do not degenerate as T_0 and T_1 increase, such that the asymptotic distributions can be properly defined. If the variances are degenerating, then the sample MS(P)E of the SC estimator $L_{T_0}(\mathbf{w})$ and $L_{T_1}(\mathbf{w})$ converge to their expectation $R_{T_0}(\mathbf{w})$ and $R_{T_1}(\mathbf{w})$, respectively, at a faster rate (see (B.1) and (B.5) in the Appendix), which cannot be quantified by the standard central limit theorems, but the conclusion is then expected to hold more easily.

Condition 9. $\xi_{T_0}^{-1} T_0^{-1/2} J^2 = o(1)$.

Condition 9 restricts the relative rate of several quantities going to infinity, i.e., ξ_{T_0}, T_0 and J. Importantly, note that this condition implies that $\xi_{T_0} \neq 0$, which turns out to be a crucial condition to establish the asymptotic optimality of the SC weight. Intuitively, $\xi_{T_0} \neq 0$ means that it is not possible to perfectly fit the pretreatment outcomes and observed covariates of the treated unit using a linear combination of the covariates and outcomes of the control units, and we refer to this situation as imperfect pretreatment fit. Imperfect pretreatment fit can result from multiple sources. To see this, note that $\xi_{T_0} = \inf_{\mathbf{w} \in \mathcal{H}} R_{T_0}(\mathbf{w})$, and we can decompose $R_{T_0}(\mathbf{w})$ as

$$
\begin{aligned}
R_{T_0}(\mathbf{w}) = & \left(\widetilde{\boldsymbol{\mu}}_0 - \sum_{j=1}^{J} w_j \widetilde{\boldsymbol{\mu}}_j \right)^{\top} \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \widetilde{\boldsymbol{\lambda}}_t \widetilde{\boldsymbol{\lambda}}_t^{\top} \left(\widetilde{\boldsymbol{\mu}}_0 - \sum_{j=1}^{J} w_j \widetilde{\boldsymbol{\mu}}_j \right) \\
& + \frac{1}{T_0} \mathbb{E} \left\| \mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j \right\|^2 + \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \mathbb{E} \left(\epsilon_{0,t} - \sum_{j=1}^{J} w_j \epsilon_{j,t} \right)^2,
\end{aligned}
$$

where $\widetilde{\boldsymbol{\lambda}}_t = (\delta_t, 1, \boldsymbol{\theta}_t^{\top}, \boldsymbol{\lambda}_t^{\top})^{\top}$ for $t \in \mathcal{T}_0 \cup \mathcal{T}_1$ and $\widetilde{\boldsymbol{\mu}}_i = (1, c_i, \mathbf{Z}_i^{\top}, \boldsymbol{\mu}_i^{\top})^{\top}$ for $i \in \{0, \ldots, J\}$. A poor fit of any of the three components in $R_{T_0}(\mathbf{w})$ can lead to an imperfect pretreatment fit. For example, an irrelevant donor pool can lead to a poor reconstruction of the factor structure and covariates, which further deteriorates the fit. A poor pretreatment fit can also be caused by sizeable idiosyncratic shocks since the last part of $R_{T_0}(\mathbf{w})$ is substantial when

the variance of shocks is large. Moreover, imperfect fit also occurs when a weight vector does not simultaneously kill the approximation errors of factor structure, covariates and shocks.

We also discuss how our definition of imperfect pretreatment fit is related to those in the literature. This concept was first formally presented by Abadie et al. (2010), in which they define a perfect pretreatment fit as the existence of a weight $\mathbf{w} \in \mathcal{H}_{\text{orig}}$ that satisfies $y_{0,t} = \sum_{j=1}^{J} w_j y_{j,t}$ for all $t \in \mathcal{T}_0$ and $\mathbf{Z}_0 = \sum_{j=1}^{J} w_j \mathbf{Z}_j$, implying that $\xi_{T_0} = 0$. The same definition is also used by Botosaru and Ferman (2019) and Ben-Michael et al. (2021), among others. Thus our definition is in line with these studies. Ferman and Pinto (2021) defines "imperfect pre-treatment fit" as the (possible) nonexistence of $\mathbf{w}^* \in \mathcal{H}_{\text{orig}}$ that satisfies $y_{0,t} = \sum_{j=1}^{J} w_j^* y_{jt}$ for every $t \in \mathcal{T}_0$, implying that ξ_{T_0} can be non-zero. Hence, our definition is also compatible with theirs.

Theorem 2. *If T_1 is finite, then under Conditions 1–5, 7, 8 (i)–(ii) and 9, we have*

$$\frac{R_{T_1}(\widehat{\mathbf{w}})}{\inf_{\mathbf{w} \in \mathcal{H}} R_{T_1}(\mathbf{w})} \xrightarrow{p} 1. \tag{11}$$

If T_1 diverges at rate $O(T_0)$, then under Conditions 1–5 and 7–9, we have (11) and

$$\frac{L_{T_1}(\widehat{\mathbf{w}})}{\inf_{\mathbf{w} \in \mathcal{H}} L_{T_1}(\mathbf{w})} \xrightarrow{p} 1; \tag{12}$$

furthermore, if $\xi_{T_1}^{-1} \{L_{T_1}(\widehat{\mathbf{w}}) - \xi_{T_1}\}$ is uniformly integrable, then

$$\frac{\mathbb{E} L_{T_1}(\widehat{\mathbf{w}})}{\inf_{\mathbf{w} \in \mathcal{H}} R_{T_1}(\mathbf{w})} \to 1. \tag{13}$$

Theorem 2 establishes the asymptotic optimality of the SC estimator, and the form differs depending on whether T_1 is divergent and the randomness of weights are incorporated. Specifically, when T_1 is finite, (11) shows that the SC weight is asymptotically optimal among all possible weighting schemes in the sense that the risk of the SC estimator $R_{T_1}(\widehat{\mathbf{w}})$ is asymptotically identical to that of the infeasible best estimator. When T_1 goes to infinity at the same rate as T_0, we can state a similar optimality but in terms of the squared error $L_{T_1}(\mathbf{w})$, a sample counterpart of $R_{T_1}(\mathbf{w})$, as in (12). Further examination of the proof reveals that (11) is one of the sufficient conditions of (12). Finally, note that $R_{T_1}(\widehat{\mathbf{w}})$ and $L_{T_1}(\widehat{\mathbf{w}})$ are both obtained by replacing the unknown weight with the estimated SC weight $\widehat{\mathbf{w}}$, i.e.,

$R_{T_1}(\widehat{\mathbf{w}}) = R_{T_1}(\mathbf{w})\big|_{\mathbf{w}=\widehat{\mathbf{w}}}$ and $L_{T_1}(\widehat{\mathbf{w}}) = L_{T_1}(\mathbf{w})\big|_{\mathbf{w}=\widehat{\mathbf{w}}}$, and thus neither (11) nor (12) accounts for the randomness of $\widehat{\mathbf{w}}$. Therefore, (13) establishes the asymptotic optimality regarding $\mathbb{E}L_{T_1}(\widehat{\mathbf{w}})$, where the expectation is taken with respect to $y_{i,t}^N$ and $\widehat{\mathbf{w}}$, such that the randomness in $\widehat{\mathbf{w}}$ is explicitly incorporated. Overall, the result in (13) shows that the expected squared error of the SC estimator, accounting for the randomness of the SC weight, is asymptotically identical to the minimum risk achieved by the infeasible best weight among all possible weighting schemes in the set $\mathcal{H}$. Theorem 2 also holds in the absence of $\mathbf{Z}$, implying that balancing between pretreatment outcomes and covariates is not essential as long as $\xi_{T_0} \neq 0$.

The asymptotic optimality of the SC weight is inspired by optimal model averaging, once we observe the link between the SC estimator and the model averaging estimator. Optimal model averaging concerns the bias-variance tradeoff in the presence of model uncertainty and is intended to obtain the best prediction by optimally combining estimators obtained from candidate models with different specifications. While asymptotic optimality is one of the most important properties in optimal model averaging studies, almost all works focus on the risk assuming that the weights and data are fixed. The only exception is Zhang (2021), which analyzes the risk while accounting for the randomness of weights, but they only consider in-sample risk. Since we average the *posttreatment* outcomes of control units, which are treated as random and can be regarded as an out-of-sample extension of the pretreatment outcomes, we need to incorporate the randomness of data and weights and study out-of-sample prediction risk. Thus, we contribute to the model averaging literature by providing the first out-of-sample asymptotic optimality accounting for the randomness of data and weights. Moreover, we also generalize existing optimality analysis in the model averaging by allowing for negative weights. See Radchenko et al. (2022) for detailed discussions on the effect of negative weights in the combination.

In practice, there are other treatment effect estimators that construct the counterfactual outcome also using a weighted average of the control units. Two popular examples include the matching estimator and IPW. Specifically, the matching estimator constructs the counterfactual outcome of the treated unit based on a set of matched units (see, e.g., Rosenbaum and Rubin, 1983; Dehejia and Wahba, 2002; Abadie and Imbens, 2006). In the case of only one treated unit, denote the fixed constant K as the number of matches, and let $\mathcal{J}_K$ be the set of matches for the treated unit (determined based on, e.g., covariates or propensity scores). Then, the matching counterfactual estimator can be written as

$\widehat{y}_{0,t}^{N}(\widehat{\mathbf{w}}^{\mathrm{match}}) = \sum_{j=1}^{J} \widehat{w}_{j}^{\mathrm{match}} y_{j,t}^{N}$, where

$$\widehat{w}_{j}^{\mathrm{match}} = \begin{cases} 1/K & \text{if } j \in \mathcal{J}_{K}(0) \\ 0 & \text{otherwise} \end{cases} .$$

Obviously, $\widehat{\mathbf{w}}^{\mathrm{match}}$ also belongs to $\mathcal{H}$.

The normalized IPW method constructs the weights based on the propensity score (see, e.g., Imbens, 2004). Denote $\mathcal{S}$ as the set of indices of treated units and $\mathbb{I}(\cdot)$ as an indicator function. In our framework, $\mathcal{S} = \{0\}$ and $\mathbb{I}(i \in \mathcal{S}) = 1$ only when $i = 0$. For some prespecified characteristics $\mathbf{X}_i$ for unit i, e.g., $\mathbf{X}_i = \left(y_{i,-T_0+1}, \ldots, y_{i,0}, \mathbf{Z}_i^{\top}\right)^{\top}$ in our setup, the propensity score of unit i is defined as $\pi(\mathbf{X}_i) = \Pr\{\mathbb{I}(i \in \mathcal{A}) \mid \mathbf{X}_i\}$, that is, the conditional probability of unit i being treated. Then, the IPW estimator can be written as $\widehat{y}_{0,t}^{N}(\widehat{\mathbf{w}}^{\mathrm{IPW}}) = \sum_{j=1}^{J} \widehat{w}_{j}^{\mathrm{IPW}} y_{j,t}^{N}$, where

$$\begin{aligned} \widehat{w}_{j}^{\mathrm{IPW}} &= \left\{ \sum_{i=1}^{J} \frac{(1 - d_{i,t})\pi(\mathbf{X}_i)}{1 - \pi(\mathbf{X}_i)} \right\}^{-1} \frac{(1 - d_{j,t})\pi(\mathbf{X}_j)}{1 - \pi(\mathbf{X}_j)} \\ &= \left\{ \sum_{i=1}^{J} \frac{\pi(\mathbf{X}_i)}{1 - \pi(\mathbf{X}_i)} \right\}^{-1} \frac{\pi(\mathbf{X}_j)}{1 - \pi(\mathbf{X}_j)}, \end{aligned} \tag{14}$$

and this weight also satisfies $\widehat{\mathbf{w}}^{\mathrm{IPW}} \in \mathcal{H}$. From Theorem 2, we know that the SC weight is potentially more desirable than other weights if one's aim is to achieve the minimum MSPE. This implies that matching and IPW estimators cannot outperform the SC estimator in terms of achieving the lowest risk, at least in the asymptotic sense.

Theorem 2 provides another justification for the SC estimator, especially under imperfect pretreatment fit and a finite number of control units. Specifically, Ferman and Pinto (2021) shows that the SC estimator is biased under imperfect pretreatment fit, and Ferman (2021) shows that this bias disappears only when the number of control units goes to infinity (see also the discussion in Section 3.1). Our result in Theorem 2 suggests that although the SC estimator is biased, it is asymptotically optimal in terms of achieving the minimum (expected) squared prediction error among all possible estimators that construct counterfactual outcomes based on averaging control units. Such asymptotic optimality holds regardless of whether the number of control units J is finite. Thus, our results significantly widen the range of applicability of the SC estimator, showing that it is still a recommended method

even under imperfect pretreatment fit with a finite number of control units. Note further that the MSPE can be decomposed into the variance and the square of bias. Thus, in the case where the SC estimator is unbiased due to diverging J (as shown by Ferman (2021) and Theorem 1), the asymptotic optimality in Theorem 2 implies that the variance of the SC estimator converges to the lower bound. Related to the regret analysis in Chen (2022), our asymptotic optimality also implies the bound convergence of the corresponding regret (with a potentially different rate from Chen (2022)) in some situations. This is because the denominators in (11)–(13), i.e., the infimum of the (expected) MSPEs, are typically of a constant (or even lower) order under certain regularity conditions, and thus the convergence of the ratios of (expected) MSPEs in (11)–(13) implies that the bound of the corresponding regret, e.g., $R_{T_1}(\widehat{\mathbf{w}}) - \inf_{\mathbf{w} \in \mathcal{H}} R_{T_1}(\mathbf{w})$, converges. Besides, Theorem 2 also provides theoretical foundation for the numerical finding of Bottmer et al. (2021) that the root mean squared errors of SC estimators are substantially lower than difference-in-means in their simulation studies, but they do not theoretically prove this.

3.2.2 Asymptotic optimality of SCM with an intercept

Doudchenko and Imbens (2016) argues that many of the treatment effect estimators in the literature employ a common linear structure to construct the counterfactual outcome of the treated unit, i.e.,

$$\widehat{y}_{0,t}^N(\mathbf{w}, d) = \sum_{j=1}^{J} w_j y_{j,t}^N + d, \tag{15}$$

where $\mathbf{w} = (w_1, \ldots, w_J)^\top \in \mathcal{H}$ and d is an intercept whose parameter space is $\mathcal{D}$. In this framework, the standard SC estimator is to set $d = 0$ and chooses the optimal weight within $\mathcal{H}_{\text{orig}}$. An alternative popular estimator is DID, which relaxes the restriction of $d = 0$ but imposes that all weights are equal across control units (Athey and Imbens, 2006; Doudchenko and Imbens, 2016), i.e., $\widehat{w}_j^{\text{DID}} = 1/J$ for $j \in \{1, \ldots, J\}$ and

$$\widehat{d}^{\text{DID}} = \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} y_{0,t}^N - \frac{1}{T_0 J} \sum_{t \in \mathcal{T}_0} \sum_{j=1}^{J} y_{j,t}^N.$$

To enjoy the flexibility of non-zero intercepts in DID but still maintain the advantage of

data-driven weights in the SCM, Doudchenko and Imbens (2016) and Ferman and Pinto (2021) propose a demeaned SC (DSC) estimator, given by

$$\widehat{\alpha}_{0,t}^{\mathrm{DSC}} = y_{0,t} - \sum_{j=1}^{J} \widehat{w}_j^{\mathrm{DSC}} y_{j,t} - \widehat{d}^{\mathrm{DSC}},$$

where $\widehat{w}_j^{\mathrm{DSC}}$ and $\widehat{d}^{\mathrm{DSC}}$ are the demeaned version of the SC weight and intercept and can be obtained by

$$(\widehat{w}_j^{\mathrm{DSC}}, \widehat{d}^{\mathrm{DSC}}) = \operatorname*{arg\,min}_{\mathbf{w} \in \mathcal{H}, \; d \in \mathcal{D}} L_{T_0}(\mathbf{w}, d), \tag{16}$$

where

$$L_{T_0}(\mathbf{w}, d) = \frac{1}{T_0} \left\{ \sum_{t \in \mathcal{T}_0} \left(y_{0,t} - \sum_{j=1}^{J} w_j y_{j,t} - d \right)^2 + \left\| \mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j - d\boldsymbol{\iota}_r \right\|^2 \right\},$$

with $\boldsymbol{\iota}_r$ being an $r \times 1$ vector of ones. Note that the resulting intercept estimate from (16) coincides the estimated intercept of Ferman and Pinto (2021) in standard cases.[4]

Ferman and Pinto (2021) shows that incorporating the intercept relaxes the condition for the SC estimator to be unbiased. Specifically, when the treatment assignment is not correlated with time-varying unobservables, the DSC estimator is asymptotically unbiased like DID, despite that its weights still fail to recover the time-invariant fixed effects of the treated unit. The DSC estimator also leads to a lower MSPE than DID. In contrast, when the treatment assignment does correlate with time-varying unobservables with $\mathbb{E}\boldsymbol{\lambda}_t \neq 0$ for $t \in \mathcal{T}_1$, neither the DSC nor the DID estimator is asymptotically unbiased. Ferman and Pinto (2021) also claims that in general it is impossible to rank these two estimators in terms of bias and MSPE in this case.

In this section, we investigate the asymptotic optimality of the DSC estimator. Some

---

[4]To illustrate the relation between the intercept estimated by (16) and that of Ferman and Pinto (2021), we assume that $\mathcal{D} = [D_{\mathrm{L}}, D_{\mathrm{U}}]$ with D_{L} and D_{U} being some constants. Ferman and Pinto (2021) calculate $\widehat{w}_j^{\mathrm{DSC}}$ by $\min_{\mathbf{w} \in \mathcal{H}} 1/T_0 \sum_{t \in \mathcal{T}_0} \left[y_{0,t} - \sum_{j=1}^{J} w_j y_{j,t} - \left(y_0^- - \sum_{j=1}^{J} w_j y_j^- \right) \right]^2$, with $\bar{y}_i^- = T_0^{-1} \sum_{t \in \mathcal{T}_0} y_{i,t}$. In the absence of covariates, this is equivalent to solving (16) subject to $d = y_0^- - \sum_{j=1}^{J} w_j y_j^-$ as long as $D_{\mathrm{L}} < \widehat{d}^{\mathrm{DSC}} < D_{\mathrm{U}}$. Since it is not difficult to define D_{L} and D_{U} such that $\widehat{d}^{\mathrm{DSC}}$ is an interior point of $\mathcal{D}$, the equivalence holds in most standard cases.

additional conditions are needed.

Condition 10. *There exists a sufficiently large constant C_d such that $|d| \leq C_d$ for any $d \in \mathcal{D} \subset \mathbb{R}$.*

This condition restricts the range of extrapolation caused by allowing for the intercept term. Since C_d can be sufficiently large, this restriction is mild. Additionally, note that this boundedness requirement can be satisfied under a compact parameter space, which is commonly used in the econometric literature.

Condition 11.

(i) *There exists a constant $\widetilde{C}_2$ such that $\mathrm{var}\left\{T_0^{-1/2} \sum_{t \in \mathcal{T}_0} e_{t,\epsilon}^{(j)}\right\} \geq \widetilde{C}_2 > 0$ for all T_0 sufficiently large and any $j \in \{1, \ldots, J\}$.*

(ii) *There exists a constant $\widetilde{C}_3$ such that $\mathrm{var}\left\{T_1^{-1/2} \sum_{t \in \mathcal{T}_1} e_{t,\epsilon}^{(j)}\right\} \geq \widetilde{C}_3 > 0$ for all T_1 sufficiently large and any $j \in \{1, \ldots, J\}$.*

This condition resembles Conditions 8 (ii)–(iii), needed for the central limit theorem for dependent processes. It guarantees that the variance of the (difference in) idiosyncratic shocks is nondegenerate as T_0 and T_1 increase.

Condition 12.

(i) $T_1^{-1} \sum_{t \in \mathcal{T}_1} \left(T_0^{-1} \sum_{k \in \mathcal{T}_0} \boldsymbol{\lambda}_k - \boldsymbol{\lambda}_t\right) = O\left(T_0^{-1/2}\right).$

(ii) $T_1^{-1} \sum_{t \in \mathcal{T}_1} \left(T_0^{-1} \sum_{k \in \mathcal{T}_0} \delta_k - \delta_t\right) = O\left(T_0^{-1/2}\right).$

(iii) $T_1^{-1} \sum_{t \in \mathcal{T}_1} \left(T_0^{-1} \sum_{k \in \mathcal{T}_0} \boldsymbol{\theta}_k - \boldsymbol{\theta}_t\right) = O\left(T_0^{-1/2}\right).$

Condition 12 requires that the factor loadings, time fixed effects and the slope coefficients do not change substantially after the treatment. It plays a similar role as Conditions 2 and 4 (i). Note that the "diverging" factors in the example discussed below Condition 2 also satisfy Condition 12.

Let $R_{T_0}(\mathbf{w}, d) = \mathbb{E}L_{T_0}(\mathbf{w}, d)$ and $\widetilde{\xi}_{T_0} = \inf_{\mathbf{w} \in \mathcal{H}, d \in \mathcal{D}} R_{T_0}(\mathbf{w}, d).$

Condition 13. $\widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w} \in \mathcal{H}} \left| T_0^{-1} \sum_{t \in \mathcal{T}_0} \mathbb{E} \left(\sum_{j=1}^J w_j e_{t,\epsilon}^{(j)} \right)^2 - T_1^{-1} \sum_{t \in \mathcal{T}_1} \mathbb{E} \left(\sum_{j=1}^J w_j e_{t,\epsilon}^{(j)} \right)^2 \right| = o(1).$

Condition 14. $\widetilde{\xi}_{T_0}^{-1} T_0^{-1/2} J^2 = o(1).$

Conditions 13 and 14 are demeaned version of Conditions 5 and 9, respectively. Note that $\widetilde{\xi}_{T_0} = \inf_{\mathbf{w} \in \mathcal{H}, d \in \mathcal{D}} R_{T_0}(\mathbf{w}, d) \leq \inf_{\mathbf{w} \in \mathcal{H}} R_{T_0}(\mathbf{w}, 0) = \xi_{T_0}$, and thus Conditions 13 and 14 are both slightly stronger than their counterparts without demeaning. Nevertheless, most standard setups of the SCM would still satisfy these two conditions, including the example described below Condition 5.

Theorem 3. *If T_1 is finite and Conditions 1–4, 7, 8 (i)–(ii), 10, 11 (i) and 12–14 hold, then*

$$\frac{R_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}}, \widehat{d}^{\mathrm{DSC}})}{\inf_{\mathbf{w} \in \mathcal{H}, d \in \mathcal{D}} R_{T_1}(\mathbf{w}, d)} \xrightarrow{p} 1. \tag{17}$$

If T_1 diverges at rate $O(T_0)$ and Conditions 1–4, 7–8 and 10–14 hold, then we have (17) and

$$\frac{L_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}}, \widehat{d}^{\mathrm{DSC}})}{\inf_{\mathbf{w} \in \mathcal{H}, d \in \mathcal{D}} L_{T_1}(\mathbf{w}, d)} \xrightarrow{p} 1; \tag{18}$$

furthermore, if $\widetilde{\xi}_{T_1}^{-1} \left\{ L_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}}, \widehat{d}^{\mathrm{DSC}}) - \widetilde{\xi}_{T_1} \right\}$ is uniformly integrable, then

$$\frac{\mathbb{E} L_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}}, \widehat{d}^{\mathrm{DSC}})}{\inf_{\mathbf{w} \in \mathcal{H}, d \in \mathcal{D}} R_{T_1}(\mathbf{w}, d)} \to 1. \tag{19}$$

Theorem 3 establishes the asymptotic optimality of the DSC estimator. It shows that the DSC estimator can still achieve the minimum (expected) squared prediction risk (and loss) asymptotically even if it is biased when the treatment assignment is correlated with time-varying unobservables. Moreover, it also suggests that the DSC estimator is not worse than the DID estimator in terms of the squared prediction error, at least asymptotically. Note that this result does not conflict with the statement of Ferman and Pinto (2021) that it is in general impossible to compare the bias and MSPE between the DSC and the DID estimators. Our optimality is in the asymptotics, and we focus on the squared error, and the finite sample bias and MSPE comparison between these two estimators remains unclear.

Intuitively, the asymptotic optimality of the SC estimator (with or without an intercept) is a consequence of the fact that the objective function to estimate the SC weight is consistent

with the evaluation criterion, namely the (expected) squared error, while other estimators, such as matching, IPW and DID, construct weights using a different objective function or simply impose equal weights. For this reason, the SCM can also be viewed as a task-based approach, whose advantages in prediction are illustrated in (Donti et al., 2017). The property of asymptotic optimality is also precisely aligned with the goal of the SCM—to synthesize a good control unit to predict the counterfactual of the treated—and it provides the conditions under which the SCM may outperform other estimators.

3.3 Asymptotic properties in a model-free setup

While the linear factor model covers a wide range of random processes to generate potential outcomes, one may still be reluctant to completely disregard other DGPs. Yet, the applicability of SCM does not seem to be restrained by certain specific forms of models. To reconcile these two perspectives, we investigate the asymptotic behavior of SCM without specifying a linear factor model as the outcome process. We show that the convergence result and the asymptotic optimality of SCM continue to hold in a model-free setup. By imposing assumptions directly on outcomes (rather than factor structures), some arguments even simplify. To save space, we only provide the assumptions and (re)state the general versions of Theorems 1 and 2 here but relegate the detailed discussions, the case with intercepts, and proofs to the Online Appendix.

First, we provide conditions needed to examine the convergence of SC weights in a model-free steup.

Condition 15. *We treat* $\{\mathbf{Z}_i \mid i \in \{0, 1, \dots, J\}\}$ *as fixed and* $\{y_{i,t}^N \mid i \in \{0, 1, \dots, J\}, t \in \mathcal{T}_0 \cup \mathcal{T}_1\}$ *as stochastic.*

Condition 16. $\sup\limits_{\mathbf{w} \in \mathcal{H}} \left| T_0^{-1} \sum_{t \in \mathcal{T}_0} \mathbb{E} \left(y_{0,t}^N - \sum_{j=1}^J w_j y_{j,t}^N \right)^2 - R_{T_1}(\mathbf{w}) \right| = O(T_0^{-1/2} J^2) + o(\xi_{T_0}).$

Condition 15 is a general version of Condition 1 (i). Condition 16 restricts the difference between the fits in the pre- and posttreatment periods. Its direct implication is that the main difference between the pre- and posttreatment outcomes is exclusively due to the treatment effect. This condition can be viewed as a generalization of Conditions 2–5.

Theorem 4. *Given any T_1, if $\mathbf{w}_{T_1}^{opt}$ is an interior point of $\mathcal{H}$ and Conditions 4 (ii), 6 and 15–16 hold, then $\left\| \widehat{\mathbf{w}} - \mathbf{w}_{T_1}^{opt} \right\| = O_p \left(T_0^\nu \xi_{T_0}^{1/2} + T_0^\nu \xi_{T_1}^{1/2} + T_0^{-1/4+\nu} J \right)$, where $\nu > 0$ is a sufficiently small constant.*

This theorem is a restatement of the convergence of SC weights (Theorem 1) without assuming a linear factor model as the DGP.

Next, we provide additional conditions needed for the asymptotic optimality in a model-free setup.

Condition 17. *For any $i \in \{0, 1, \ldots, J\}$, $\{y_{i,t}^N\}$ is either α-mixing with the mixing coefficient $\alpha = -r/(r-2)$ or ϕ-mixing with the mixing coefficient $\phi = -r/(2r-1)$ for $r \geq 2$.*

Denote $e_{t,y^N}^{(i)} = y_{0,t}^N - y_{i,t}^N$ for $i \in \{1, \ldots, J\}$ and $t \in \mathcal{T}_0 \cup \mathcal{T}_1$.

Condition 18.

(i) *There exists a constant C_4 such that $\mathbb{E} \left(y_{i,t}^N \right)^4 \leq C_4 < \infty$ for $i \in \{0, 1, \ldots, J\}$ and $t \in \mathcal{T}_0 \cup \mathcal{T}_1$.*

(ii) *There exists a constant C_5 such that $\mathrm{var} \left(T_0^{-1/2} \sum_{t \in \mathcal{T}_0} e_{t,y^N}^{(i)} e_{t,y^N}^{(j)} \right) \geq C_5 > 0$ for all T_0 sufficiently large and any $i, j \in \{1, \ldots, J\}$.*

(iii) *There exists a constant C_6 such that $\mathrm{var} \left(T_1^{-1/2} \sum_{t \in \mathcal{T}_1} e_{t,y^N}^{(i)} e_{t,y^N}^{(j)} \right) \geq C_6 > 0$ for all T_1 sufficiently large and any $i, j \in \{1, \ldots, J\}$.*

Conditions 17 and 18 generalize the factor-model-based Conditions 7 and 8, respectively.

Theorem 5. *If T_1 is finite, then under Conditions 4 (ii), 9, 15–17 and 18 (i)–(ii), we have (11); If T_1 diverges at rate $O(T_0)$, then under Conditions 4 (ii), 9 and 15–18, we have (11) and (12); Furthermore, if $\xi_{T_1}^{-1} \{L_{T_1}(\widehat{\mathbf{w}}) - \xi_{T_1}\}$ is uniformly integrable, then we have (13).*

This theorem is a generalized version of Theorem 2, relaxing the linear factor model assumption.

4   Simulation

In this section, we verify the theory via simulation. We first examine the convergence of the SC weight and then compare the SC estimators with popular competing methods to verify their asymptotic optimality.

4.1   Simulation design

We follow Hsiao and Zhou (2019) to generate the data from the following pure factor model:

$$y_{i,t}^N = \gamma_{1,i} f_{1,t} + \gamma_{2,i} f_{2,t} + u_{i,t},$$

where the common factors $f_{s,t}$ and the factor loadings $\gamma_{s,i}$, $s \in \{1,2\}$ are both drawn independently from $\mathcal{N}(0,1)$. The idiosyncratic shock $u_{i,t}$ is weakly cross-sectionally dependent, generated by

$$u_{i,t} \;=\; (1+b^2)v_{i,t} + bv_{i+1,t} + bv_{i-1,t},$$

where $v_{i,t} \sim$ i.i.d. $\mathcal{N}(0,\sigma_i^2)$, $b = 1$ and σ_i^2 are drawn independently from $0.5\left(\chi^2(1)+1\right)$ for all i. As above, we set unit 0 as the only treated unit, and the remaining $\{1,\dots,J\}$ units are the control. Since the value of $\alpha_{i,t}$ does not influence the estimation and evaluation procedure based on (4) and (6), we follow the literature on the in SCM not assigning $\alpha_{i,t}$ a value. We set $J \in \{30,50\}$, the number of pretreatment periods $T_0 \in \{50,100,200,400\}$ and the number of posttreatment periods $T_1 = 10$. The number of replications is $R = 1000$.

4.2   Simulation results on weight convergence

To investigate the convergence of the SC weight, we need to know the limit of the SC weight, which further requires the knowledge of $R_{T_1}(\mathbf{w})$. If we denote $\boldsymbol{\eta}_t = \left(y_{0,t}^N,\dots,y_{J,t}^N\right)^\top$, then $\boldsymbol{\eta}_t$ follows $\mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\eta}_t}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}_t}\right)$, where $\boldsymbol{\mu}_{\boldsymbol{\eta}_t}$ is a $(J+1) \times 1$ vector with its $(i+1)$-th element being $\mu_{\boldsymbol{\eta}_t, i+1} = \gamma_{1,i} f_{1,t} + \gamma_{2,i} f_{2,t}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_t}$ is a $(J+1) \times (J+1)$ matrix with its element in the

$(i + 1)$-th row and $(j + 1)$-th column being

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}_t, i+1, j+1} = \begin{cases} (1 + b^2)^2 \sigma_0^2 + b^2 \sigma_1^2 & if\ i = j = 0, \\ (1 + b^2)^2 \sigma_i^2 + b^2 \left(\sigma_{i-1}^2 + \sigma_{i+1}^2 \right) & if\ 1 \leq i = j \leq J - 1, \\ (1 + b^2)^2 \sigma_J^2 + b^2 \sigma_{J-1}^2 & if\ i = j = J, \\ b \left(1 + b^2 \right) \left(\sigma_i^2 + \sigma_j^2 \right) & if\ |i - j| = 1, \\ b^2 \sigma_{(i+j)/2}^2 & if\ |i - j| = 2, \\ 0 & \text{otherwise.} \end{cases}$$

Note further that $\{\boldsymbol{\eta}_t \mid t \in \mathcal{T}_0 \cup \mathcal{T}_1\}$ are independent over t. With $\boldsymbol{\mu}_{\boldsymbol{\eta}_t}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_t}$ at hand, one can compute $R_{T_1}(\mathbf{w})$ as

$$\begin{aligned} R_{T_1}(\mathbf{w}) &= \mathbb{E}L_{T_1}(\mathbf{w}) \\ &= \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \mathbb{E} \left(y_{0,t}^N - \sum_{j=1}^J w_j y_{j,t}^N \right)^2 \\ &= \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \mathbb{E} \left\{ \boldsymbol{\eta}_t^\top \left(1, -\mathbf{w}^\top \right)^\top \left(1, -\mathbf{w}^\top \right) \boldsymbol{\eta}_t \right\} \\ &= \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \boldsymbol{\mu}_{\boldsymbol{\eta}_t}^\top \left(1, -\mathbf{w}^\top \right)^\top \left(1, -\mathbf{w}^\top \right) \boldsymbol{\mu}_{\boldsymbol{\eta}_t} + \text{tr} \left\{ \left(1, -\mathbf{w}^\top \right)^\top \left(1, -\mathbf{w}^\top \right) \boldsymbol{\Sigma}_{\boldsymbol{\eta}_t} \right\}. \quad (20) \end{aligned}$$

When an intercept is allowed for, the generalized version of $R_{T_1}(\mathbf{w}, d)$ can be written as

$$\begin{aligned} R_{T_1}(\mathbf{w}, d) &= \mathbb{E}L_{T_1}(\mathbf{w}, d) = \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \mathbb{E} \left(y_{0,t}^N - \sum_{j=1}^J w_j y_{j,t}^N - d \right)^2 \\ &= \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \left(\boldsymbol{\mu}_{\boldsymbol{\eta}_t}^\top, 1 \right) \left(1, -\mathbf{w}^\top, -d \right)^\top \left(1, -\mathbf{w}^\top, -d \right) \left(\boldsymbol{\mu}_{\boldsymbol{\eta}_t}^\top, 1 \right)^\top \\ &\quad + \text{tr} \left\{ \left(1, -\mathbf{w}^\top \right)^\top \left(1, -\mathbf{w}^\top \right) \boldsymbol{\Sigma}_{\boldsymbol{\eta}_t} \right\}. \end{aligned}$$

In the simulation, we search for the weight in the original weight set $\mathcal{H}_{\text{orig}}$, i.e., setting $C_{\text{L}} = 0$ and $C_{\text{U}} = 1$. The SC weight is estimated by

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathcal{H}_{\text{orig}}} \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \left(y_{0,t} - \sum_{j=1}^J w_j y_{j,t} \right)^2,$$

and its limit is obtained by $\mathbf{w}_{T_1}^{\mathrm{opt}} = \arg\min_{\mathbf{w} \in \mathcal{H}_{\mathrm{orig}}} R_{T_1}(\mathbf{w})$ with $R_{T_1}(\mathbf{w})$ computed from (20). Allowing for negative weights does not qualitatively change the results.

Figure 1 plots the vector norm of the difference between the SC weight and its limit, $\left\| \widehat{\mathbf{w}} - \mathbf{w}_{T_1}^{\mathrm{opt}} \right\|$, averaged over the 1000 replications, as T_0 increases. The solid line indicates that $J = 30$, and the dashed line is when $J = 50$. Under both sample sizes, $\left\| \widehat{\mathbf{w}} - \mathbf{w}_{T_1}^{\mathrm{opt}} \right\|$ is monotonically decreasing as T_0 increases, which confirms the convergence result in Theorem 1. Comparing the values obtained under different numbers of control units, we find that $\widehat{\mathbf{w}}$ converges faster when $J = 30$ than $J = 50$, which again confirms that the rate of convergence slows when J increases as stated in Theorem 1.
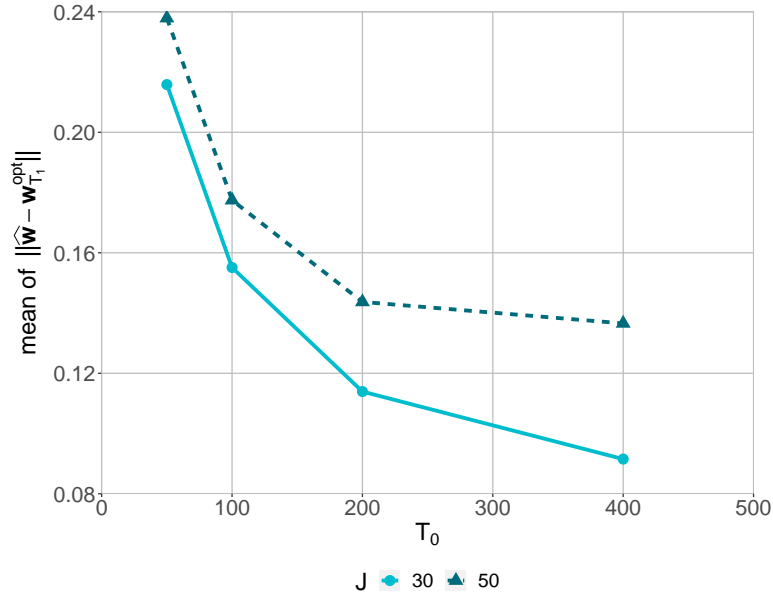

Figure 1: Average of $\left\| \widehat{\mathbf{w}} - \mathbf{w}_{T_1}^{\mathrm{opt}} \right\|$ over 1000 replications

4.3 Simulation results on asymptotic optimality

To examine the asymptotic optimality of the SC estimator, we compare it with alternative estimators that also construct counterfactual outcomes by averaging control units. We consider two versions of SCM, the standard SCM and the SCM with an intercept (denoted as DSC). We compare them with five alternative methods, i.e., propensity score matching (PSM), IPW, equal-weight averaging (Equal), best control selection (Sel) and DID.

The standard SCM and DSC are implemented as (4) and (16), respectively. PSM is one of the most popular matching methods, first proposed by Rosenbaum and Rubin (1983). It is based on the propensity score, $\widehat{\pi}(\mathbf{X}_i)$ estimated from a logistic regression with $\mathbf{X}_i$ as regressors. The weights are computed by (14), where we set $K = 1$ and

$$\mathcal{J}_K(0) = \left\{ j \in \{1, \ldots, J\} \mid \sum_{k=1}^{J} \mathbb{I}\Big(|\widehat{\pi}(\mathbf{X}_k) - \widehat{\pi}(\mathbf{X}_0)| \leq |\widehat{\pi}(\mathbf{X}_j) - \widehat{\pi}(\mathbf{X}_0)| \Big) \leq K \right\},$$

with $\mathbb{I}(\cdot)$ being an indicator function. IPW computes the weight from (14), also using the estimated propensity score $\widehat{\pi}(\mathbf{X}_i)$ as above. Equal-weight averaging constructs the counterfactuals by using simple average of all control units. The best control selection selects a single control unit based on minimizing the in-sample mean squared error. It can be viewed as a special case of best subset selection; see Doudchenko and Imbens (2016) for further explanation. Furthermore, it can also be viewed as a special case of averaging controls, since the weight equals 1 for the selected best control and zeros for the remaining controls, i.e.,

$$\widehat{w}_j^{\text{Sel}} = \begin{cases} 1 & \text{if } j = \arg\min_{m_0 \in \{1, \ldots, J\}} T_0^{-1} \sum_{t \in \mathcal{T}_0} \left(y_{0,t}^N - y_{m_0,t}^N \right)^2, \\ 0 & \text{otherwise.} \end{cases}$$

We evaluate all methods by $R_{T_1}(\mathbf{w})$ computed from (20).

Figure 2 plots the ratio of risk, $R_{T_1}(\mathbf{w})/\inf_{\mathbf{w}^* \in \mathcal{H}} R_{T_1}(\mathbf{w}^*)$ for the methods without intercepts and $R_{T_1}(\mathbf{w}, d)/\inf_{\mathbf{w}^* \in \mathcal{H}, d^* \in \mathcal{D}} R_{T_1}(\mathbf{w}^*, d^*)$ for the methods with intercepts, averaged over replications. SCM and DSC generally perform quite similarly, and both methods exhibit clear superiority as their ratios of the risk are much lower than those of the other methods for all T_0 and J. Moreover, the curves of SCM and DSC both monotonically decrease toward 1 as T_0 increases, implying that the risk of the SCM and DSC estimators converges to the lowest possible risk as the number of pretreatment periods increases. This result precisely coincides with the asymptotic optimality stated in Theorems 2 and 3. Regarding the competing methods, we note that PSM and Sel produce a very similar ratio of risk because they typically select the same control unit to construct the counterfactual. The performance of IPW, Equal and DID are also similar. The similarity between IPW and Equal can be explained by the fact that our DGP generates all J control units from the same distribution, resulting in $\widehat{\mathbf{w}}^{\text{IPW}} \approx 1/J$, and the similarity between Equal and DID is also expected because both impose equal weights with the only difference being in the intercept. Figure 2

shows that the ratios of all competing methods do not converge, implying that none of these methods achieves asymptotic optimality in terms of risk.
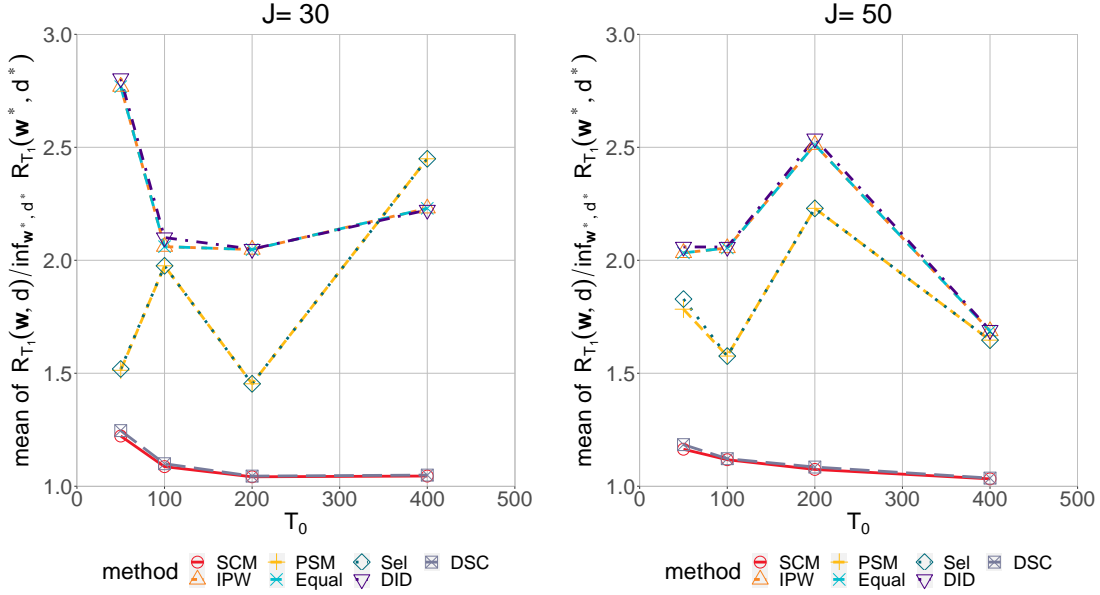

Figure 2: Average of $R_{T_1}(\mathbf{w}, d)/\inf_{\mathbf{w}^* \in \mathcal{H}, d^* \in \mathcal{D}} R_{T_1}(\mathbf{w}^*, d^*)$ over 1000 replications

5 Conclusion

This paper investigates the asymptotic properties of the SCM as the number of the pretreatment periods diverges. We show that the SC weight converges to the limit that may not recover the factor structure but minimizes the (expected) mean squared prediction error. We quantify the rate of convergence, from which one can better understand how the number of control units and the pre- and posttreatment fit influence the convergence rate. Our convergence results also verify under which conditions the SC weight can dilute and reconstruct the factor structure as the number of controls increases, so that the unbiasedness of the SC estimator can be achieved.

Furthermore, we establish the asymptotic optimality of both the standard SCM and DSC estimators. We provide conditions under which the two versions of SC estimators are asymptotically optimal among the class of estimators that construct counterfactual outcomes using an average of control units. Our asymptotic optimality suggests that if there is no

weight to produce a perfect pretreatment fit, then the (expected) squared prediction error of the SC estimator converges to the lowest possible risk, implying that the SC estimator asymptotically dominates other estimators in this class, such as matching, IPW and DID. This result justifies the use of the SC estimator even when there are unobserved confounders, pretreatment fit is not perfect, and the number of control units is finite. These properties are also free of the underlying assumptions on the DGP of potential outcomes.

Our theoretical analysis suggests that the SCM provides theoretical guarantees and can be used in a wide range of applications, regardless of whether pretreatment fit is perfect. When pretreatment fit is perfect, the SCM estimator can be interpreted as an unbiased estimator of the treatment effect; in the presence of imperfect pretreatment fit, the SCM can still be applied as an asymptotically minimum-MSPE estimator of the treatment effect.

References

A. Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59:391–425, 2021.

A. Abadie and J. Gardeazabal. The economic costs of conflict: A case study of the Basque country. *American Economic Review*, 93:113–132, 2003.

A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235–267, 2006.

A. Abadie, A. Diamond, and J. Hainmueller. 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, 2010.

S. Athey and G. W. Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74:431–497, 2006.

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

E. Ben-Michael, A. Feller, and J. Rothstein. The augmented synthetic control method. *Journal of the American Statistical Association*, 116:1789–1803, 2021.

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

L. Bottmer, G. W. Imbens, J. Spiess, and M. J. Warnick. A design-based perspective on synthetic control methods. Working paper, 2021.

J. Chen. Synthetic control as online linear regression. Working paper, 2022.

R. H. Dehejia and S. Wahba. Propensity score-matching methods for monexperimental causal studies. *The Review of Economics and Statistics*, 84:151–161, 2002.

P. Donti, B. Amos, and J. Z. Kolter. Task-based end-to-end model learning in stochastic optimization. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates Inc., 2017.

N. Doudchenko and G. W. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Working Paper 22791, National Bureau of Economic Research, 2016.

J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32:928–961, 2004.

B. Ferman. On the properties of the synthetic control estimator with many periods and many controls. *Journal of the American Statistical Association*, 116:1764–1772, 2021.

B. Ferman and C. Pinto. Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, 12:1197–1221, 2021.

Y. Gao, X. Zhang, S. Wang, T. T.-l. Chong, and G. Zou. Frequentist model averaging for threshold models. *Annals of the Institute of Statistical Mathematics*, 71:275–306, 2019.

B. E. Hansen. Least squares model averaging. *Econometrica*, 75:1175–1189, 2007.

C. Hsiao and Q. Zhou. Panel parametric, semiparametric, and nonparametric construction of counterfactuals. *Journal of Applied Econometrics*, 34:463–481, 2019.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86:4–29, 2004.

Q. Liu and R. Okui. Heteroskedasticity-robust C_p model averaging. *The Econometrics Journal*, 16:463–472, 2013.

X. Lu and L. Su. Jackknife model averaging for quantile regressions. *Journal of Econometrics*, 188:40–58, 2015.

D. Powell. Synthetic control estimation beyond comparative case studies: Does the minimum wage reduce employment? *Journal of Business & Economic Statistics*, 40:1302–1314, 2022.

P. Radchenko, V. A. L., and W. Wang. Too similar to combine? On negative weights in forecast combination. *International Journal of Forecasting*, forthcoming, 2022.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

P. Schönfeld. A useful central limit theorem for m-dependent variables. *Metrika*, 17:116–128, 1971.

D. J. Scott. Central limit theorems for martingales and for processes with stationary increments using a Skorokhod representation approach. *Advances in Applied Probability*, 5: 119–137, 1973.

J. H. Stock and M. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97:1167–1179, 2002.

A. T. Wan, X. Zhang, and G. Zou. Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156:277–283, 2010.

H. White. *Asymptotic theory for econometricians*. Academic Press, Orlando, 1984.

J. M. Wooldridge and H. White. Some invariance principles and central limit theorems for dependent heterogenrous processes. *Econometric Therory*, 4:210–230, 1988.

Y. Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25:57–76, 2017.

X. Zhang. A new study on asymptotic optimality of least squares model averaging. *Econometric Theory*, 37:388–407, 2021.

APPENDIX

A Proof of Theorem 1

Denote $\tau_{T_0} = T_0^\nu \xi_{T_0}^{1/2} + T_0^\nu \xi_{T_1}^{1/2} + T_0^{-1/4+\nu} J$, and let $\mathbf{u} = (u_1, \ldots, u_J)^\top$. According to Fan and Peng (2004) and Lu and Su (2015), to prove Theorem 1, it suffices to show that, for any $\varepsilon > 0$, there exists a constant C_ε, such that

$$\Pr \left\{ \inf_{\|\mathbf{u}\| = C_\varepsilon, \left(\mathbf{w}_{T_1}^{\text{opt}} + \tau_{T_0} \mathbf{u}\right) \in \mathcal{H}} L_{T_0}(\mathbf{w}_{T_1}^{\text{opt}} + \tau_{T_0} \mathbf{u}) \geq L_{T_0}(\mathbf{w}_{T_1}^{\text{opt}}) \right\} > 1 - \varepsilon, \qquad (A.1)$$

under a given T_1 and any sufficient large T_0. (A.1) implies that there exists a $\mathbf{w}^\star$ in the bounded closed domain set $\left\{ \mathbf{w}_{T_1}^{\text{opt}} + \tau_{T_0} \mathbf{u} \mid \left(\mathbf{w}_{T_1}^{\text{opt}} + \tau_{T_0} \mathbf{u}\right) \in \mathcal{H}, \|\mathbf{u}\| \leq C_\varepsilon \right\}$, such that it (locally) minimizes $L_{T_0}(\mathbf{w})$ and $\left\| \mathbf{w}^\star - \mathbf{w}_{T_1}^{\text{opt}} \right\| = O_p(\tau_{T_0})$. From the convexity of $L_{T_0}(\mathbf{w})$ and $\mathcal{H}$, $\mathbf{w}^\star$ is also the unique global minimizer, i.e., $\mathbf{w}^\star = \widehat{\mathbf{w}}$.

Define $D(\mathbf{u}) = L_{T_0}(\mathbf{w}_{T_1}^{\text{opt}} + \tau_{T_0} \mathbf{u}) - L_{T_0}(\mathbf{w}_{T_1}^{\text{opt}})$. Then, we can decompose $D(\mathbf{u})$ as

$$\begin{aligned}
D(\mathbf{u}) &= \frac{1}{T_0} \left\| \mathbf{Y}_0 - \boldsymbol{\mathcal{Y}}_c(\mathbf{w}_{T_1}^{\text{opt}} + \tau_{T_0} \mathbf{u}) \right\|^2 - \frac{1}{T_0} \left\| \mathbf{Y}_0 - \boldsymbol{\mathcal{Y}}_c \mathbf{w}_{T_1}^{\text{opt}} \right\|^2 \\
&= -\frac{2\tau_{T_0}}{T_0} \left(\mathbf{Y}_0 - \boldsymbol{\mathcal{Y}}_c \mathbf{w}_{T_1}^{\text{opt}} \right)^\top \boldsymbol{\mathcal{Y}}_c \mathbf{u} + \frac{\tau_{T_0}^2}{T_0} \left\| \boldsymbol{\mathcal{Y}}_c \mathbf{u} \right\|^2 \\
&\equiv \Delta_1 + \Delta_2,
\end{aligned}$$

where $\Delta_1 = -2\tau_{T_0} T_0^{-1} \left(\mathbf{Y}_0 - \boldsymbol{\mathcal{Y}}_c \mathbf{w}_{T_1}^{\text{opt}} \right)^\top \boldsymbol{\mathcal{Y}}_c \mathbf{u}$ and $\Delta_2 = \tau_{T_0}^2 T_0^{-1} \left\| \boldsymbol{\mathcal{Y}}_c \mathbf{u} \right\|^2$. We show that Δ_2 is the dominant term of $D(\mathbf{u})$ as follows.

We first consider Δ_2. From Condition 6, we have that, with probability approaching 1,

$$\kappa_1 \leq \lambda_{min} \left(\frac{1}{T_0} \boldsymbol{\mathcal{Y}}_c^\top \boldsymbol{\mathcal{Y}}_c \right) \leq \lambda_{max} \left(\frac{1}{T_0} \boldsymbol{\mathcal{Y}}_c^\top \boldsymbol{\mathcal{Y}}_c \right) \leq \kappa_2. \qquad (A.2)$$

This further implies that, with probability approaching 1,

$$\Delta_2 \geq \frac{\tau_{T_0}^2}{T_0} \lambda_{min} \left(\boldsymbol{\mathcal{Y}}_c^\top \boldsymbol{\mathcal{Y}}_c \right) \|\mathbf{u}\|^2 \geq \kappa_1 \tau_{T_0}^2 \|\mathbf{u}\|^2. \qquad (A.3)$$

Next, we consider Δ_1. To simplify the notation, denote $\widetilde{\boldsymbol{\lambda}}_t = (\delta_t, 1, \boldsymbol{\theta}_t^\top, \boldsymbol{\lambda}_t^\top)^\top$ for $t \in \mathcal{T}_0 \cup \mathcal{T}_1$ and $\widetilde{\boldsymbol{\mu}}_i = (1, c_i, \mathbf{Z}_i^\top, \boldsymbol{\mu}_i^\top)^\top$ for $i \in \{0, \ldots, J\}$. Let $\mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(i)} = \widetilde{\boldsymbol{\mu}}_0 - \widetilde{\boldsymbol{\mu}}_i$ for $i \in \{0, \ldots, J\}$. Under our linear factor structure (1) and Condition 1, we have that

$$\sup_{\mathbf{w} \in \mathcal{H}} |R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w})|$$

$$
\begin{aligned}
= \; & \sup_{\mathbf{w} \in \mathcal{H}} \left| \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \left(\sum_{j=1}^{J} w_j \widetilde{\boldsymbol{\lambda}}_t^\top \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)} \right)^2 + \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \mathbb{E} \left(\sum_{j=1}^{J} w_j e_{t,\epsilon}^{(j)} \right)^2 + \frac{1}{T_0} \left\| \mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j \right\|^2 \right. \\
& \left. - \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \left(\sum_{j=1}^{J} w_j \widetilde{\boldsymbol{\lambda}}_t^\top \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)} \right)^2 - \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \mathbb{E} \left(\sum_{j=1}^{J} w_j e_{t,\epsilon}^{(j)} \right)^2 \right| \\
\leq \; & \sup_{\mathbf{w} \in \mathcal{H}} \left| \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \left(\sum_{j=1}^{J} w_j \widetilde{\boldsymbol{\lambda}}_t^\top \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)} \right)^2 - \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \left(\sum_{j=1}^{J} w_j \widetilde{\boldsymbol{\lambda}}_t^\top \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)} \right)^2 \right| + \frac{1}{T_0} \sup_{\mathbf{w} \in \mathcal{H}} \left\| \mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j \right\|^2 \\
& + \sup_{\mathbf{w} \in \mathcal{H}} \left| \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \mathbb{E} \left(\sum_{j=1}^{J} w_j e_{t,\epsilon}^{(j)} \right)^2 - \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \mathbb{E} \left(\sum_{j=1}^{J} w_j e_{t,\epsilon}^{(j)} \right)^2 \right| \\
\equiv \; & \mathbb{I}_1 + \mathbb{I}_2 + \mathbb{I}_3.
\end{aligned}
\tag{A.4}
$$

We analyze $\mathbb{I}_1$, $\mathbb{I}_2$ and $\mathbb{I}_3$ in turn. First, we analyze $\mathbb{I}_1$. From Conditions 2–4, we have

$$
\frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \left(\frac{1}{T_0} \sum_{k \in \mathcal{T}_0} \widetilde{\boldsymbol{\lambda}}_k^\top \widetilde{\boldsymbol{\lambda}}_k - \widetilde{\boldsymbol{\lambda}}_t^\top \widetilde{\boldsymbol{\lambda}}_t \right) = O(T_0^{-1/2}),
$$

and the components of $\widetilde{\boldsymbol{\mu}}_i$ are bounded; hence,

$$
\begin{aligned}
\mathbb{I}_1 \; = \; & \sup_{\mathbf{w} \in \mathcal{H}} \left| \sum_{i=1}^{J} \sum_{j=1}^{J} w_i w_j \left\{ \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \widetilde{\boldsymbol{\lambda}}_t^\top \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(i)} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)\top} \widetilde{\boldsymbol{\lambda}}_t - \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \widetilde{\boldsymbol{\lambda}}_t^\top \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(i)} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)\top} \widetilde{\boldsymbol{\lambda}}_t \right\} \right| \\
\leq \; & \sup_{\mathbf{w} \in \mathcal{H}} \sum_{i=1}^{J} \sum_{j=1}^{J} |w_i| |w_j| \left| \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \widetilde{\boldsymbol{\lambda}}_t^\top \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(i)} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)\top} \widetilde{\boldsymbol{\lambda}}_t - \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \widetilde{\boldsymbol{\lambda}}_t^\top \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(i)} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)\top} \widetilde{\boldsymbol{\lambda}}_t \right| \\
\leq \; & [\max\{C_{\mathrm{L}}, C_{\mathrm{U}}\}]^2 \sum_{i=1}^{J} \sum_{j=1}^{J} \left| \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \widetilde{\boldsymbol{\lambda}}_t^\top \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(i)} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)\top} \widetilde{\boldsymbol{\lambda}}_t - \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \widetilde{\boldsymbol{\lambda}}_t^\top \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(i)} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)\top} \widetilde{\boldsymbol{\lambda}}_t \right| \\
= \; & [\max\{C_{\mathrm{L}}, C_{\mathrm{U}}\}]^2 \sum_{i=1}^{J} \sum_{j=1}^{J} \left| \mathrm{tr} \left(\frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(i)} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)\top} \widetilde{\boldsymbol{\lambda}}_t \widetilde{\boldsymbol{\lambda}}_t^\top \right) - \mathrm{tr} \left(\frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(i)} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)\top} \widetilde{\boldsymbol{\lambda}}_t \widetilde{\boldsymbol{\lambda}}_t^\top \right) \right| \\
= \; & [\max\{C_{\mathrm{L}}, C_{\mathrm{U}}\}]^2 \sum_{i=1}^{J} \sum_{j=1}^{J} \left| \mathrm{tr} \left\{ \left(\frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \widetilde{\boldsymbol{\lambda}}_t \widetilde{\boldsymbol{\lambda}}_t^\top - \frac{1}{T_1} \sum_{t \in \mathcal{T}_1} \widetilde{\boldsymbol{\lambda}}_t \widetilde{\boldsymbol{\lambda}}_t^\top \right) \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(i)} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)\top} \right\} \right| \\
= \; & O(T_0^{-1/2} J^2).
\end{aligned}
\tag{A.5}
$$

Here, the last equality is satisfied due to the fixed r and F. We then examine $\mathbb{I}_2$. From

Condition 4 (ii), we have that

$$
\begin{aligned}
\mathbb{I}_2 &\leq \frac{1}{T_0} \sup_{\mathbf{w} \in \mathcal{H}} \left\{ \sum_{j=1}^{J} \|w_j (\mathbf{Z}_0 - \mathbf{Z}_j)\| \right\}^2 \\
&\leq [\max\{C_{\mathrm{L}}, C_{\mathrm{U}}\}]^2 \, T_0^{-1} \left\{ \sum_{j=1}^{J} \|\mathbf{Z}_0 - \mathbf{Z}_j\| \right\}^2 \\
&= O_p(T_0^{-1} J^2).
\end{aligned}
\tag{A.6}
$$

Finally, we consider $\mathbb{I}_3$. From Condition 5, it is obvious that $\mathbb{I}_3 = o(\xi_{T_0})$. Combining with Inequalities (A.4)–(A.6), it implies that

$$
\sup_{\mathbf{w} \in \mathcal{H}} |R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w})| = O(T_0^{-1/2} J^2) + o(\xi_{T_0}). \tag{A.7}
$$

The above equation holds for $\mathbf{w}_{T_1}^{\mathrm{opt}}$, and note that $R_{T_1}(\mathbf{w}_{T_1}^{\mathrm{opt}}) = \xi_{T_1}$. Thus, we have that

$$
R_{T_0}(\mathbf{w}_{T_1}^{\mathrm{opt}}) - \xi_{T_1} = O(T_0^{-1/2} J^2) + o(\xi_{T_0}). \tag{A.8}
$$

Since $\mathbb{E} \left\| \mathbf{Y}_0 - \boldsymbol{\mathcal{Y}}_c \mathbf{w}_{T_1}^{\mathrm{opt}} \right\|^2 \leq T_0 R_{T_0}(\mathbf{w}_{T_1}^{\mathrm{opt}})$, using (A.8), we can obtain that

$$
\left\| \mathbf{Y}_0 - \boldsymbol{\mathcal{Y}}_c \mathbf{w}_{T_1}^{\mathrm{opt}} \right\| = O_p(T_0^{1/2} \xi_{T_1}^{1/2}) + O_p(T_0^{1/4} J) + o_p(T_0^{1/2} \xi_{T_0}^{1/2}). \tag{A.9}
$$

Therefore, we have that

$$
\begin{aligned}
|\Delta_1| &\leq \frac{2\tau_{T_0}}{T_0} \left\| \mathbf{Y}_0 - \boldsymbol{\mathcal{Y}}_c \mathbf{w}_{T_1}^{\mathrm{opt}} \right\| \|\boldsymbol{\mathcal{Y}}_c \mathbf{u}\| \\
&\leq \frac{2\tau_{T_0}}{T_0} \left\| \mathbf{Y}_0 - \boldsymbol{\mathcal{Y}}_c \mathbf{w}_{T_1}^{\mathrm{opt}} \right\| \sqrt{\lambda_{max}\left(\boldsymbol{\mathcal{Y}}_c^{\top} \boldsymbol{\mathcal{Y}}_c\right)} \|\mathbf{u}\| \\
&= O_p(\tau_{T_0} \xi_{T_1}^{1/2}) \|\mathbf{u}\| + O_p(\tau_{T_0} T_0^{-1/4} J) \|\mathbf{u}\| + o_p(\tau_{T_0} \xi_{T_0}^{1/2}) \|\mathbf{u}\|,
\end{aligned}
$$

where the second inequality is due to (A.2) and the last equality is due to (A.9). This equation together with (A.3) shows that Δ_2 asymptotically dominates Δ_1. Therefore, $D(\mathbf{u}) \geq 0$ in probability for any $\mathbf{u}$ that satisfies $\|\mathbf{u}\| = C_\varepsilon$ and $\left(\mathbf{w}_{T_1}^{\mathrm{opt}} + \tau_{T_0} \mathbf{u}\right) \in \mathcal{H}$. This completes the proof of (A.1), and thus Theorem 1 holds.

B   Proof of Theorem 2

We first prove (11) in Theorem 2. To this end, we decompose $L_{T_0}(\mathbf{w})$ as

$$
L_{T_0}(\mathbf{w}) = R_{T_1}(\mathbf{w}) + L_{T_0}(\mathbf{w}) - R_{T_0}(\mathbf{w}) + R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w}).
$$

By Lemma 1 in Gao et al. (2019), it suffices to show that

$$\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{L_{T_0}(\mathbf{w})-R_{T_0}(\mathbf{w})}{R_{T_0}(\mathbf{w})}\right|=o_p(1) \tag{B.1}$$

and

$$\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{R_{T_0}(\mathbf{w})-R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})}\right|=o(1). \tag{B.2}$$

We first verify (B.1). Note that

$$
\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{L_{T_0}(\mathbf{w})-R_{T_0}(\mathbf{w})}{R_{T_0}(\mathbf{w})}\right|\\
\leq{}& \xi_{T_0}^{-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|L_{T_0}(\mathbf{w})-R_{T_0}(\mathbf{w})\right|\\
={}& \xi_{T_0}^{-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\left\{\left(y_{0,t}-\sum_{j=1}^{J}w_jy_{j,t}\right)^2-\mathbb{E}\left(y_{0,t}-\sum_{j=1}^{J}w_jy_{j,t}\right)^2\right\}\right|\\
={}& \xi_{T_0}^{-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\sum_{i=1}^{J}\sum_{j=1}^{J}w_iw_j\left\{(y_{0,t}-y_{i,t})(y_{0,t}-y_{j,t})-\mathbb{E}(y_{0,t}-y_{i,t})(y_{0,t}-y_{j,t})\right\}\right|\\
\leq{}& \xi_{T_0}^{-1}\sup_{\mathbf{w}\in\mathcal{H}}\sum_{i=1}^{J}\sum_{j=1}^{J}|w_i||w_j|\left|\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\left\{(y_{0,t}-y_{i,t})(y_{0,t}-y_{j,t})-\mathbb{E}(y_{0,t}-y_{i,t})(y_{0,t}-y_{j,t})\right\}\right|\\
\leq{}& [\max\{C_{\mathrm{L}},C_{\mathrm{U}}\}]^2\,\xi_{T_0}^{-1}\sum_{i=1}^{J}\sum_{j=1}^{J}\left|\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\left\{(y_{0,t}-y_{i,t})(y_{0,t}-y_{j,t})-\mathbb{E}(y_{0,t}-y_{i,t})(y_{0,t}-y_{j,t})\right\}\right|\\
={}& [\max\{C_{\mathrm{L}},C_{\mathrm{U}}\}]^2\,\xi_{T_0}^{-1}T_0^{-1/2}\sum_{i=1}^{J}\sum_{j=1}^{J}\Psi_{T_0}(i,j),
\end{aligned}
\tag{B.3}
$$

where

$$\Psi_{T_0}(i,j)=\left|\frac{1}{\sqrt{T_0}}\sum_{t\in\mathcal{T}_0}\left\{(y_{0,t}-y_{i,t})(y_{0,t}-y_{j,t})-\mathbb{E}(y_{0,t}-y_{i,t})(y_{0,t}-y_{j,t})\right\}\right|.$$

Under the linear factor structure (1) and Condition 1, we can rewrite $\Psi_{T_0}(i,j)$ as a function of idiosyncratic shocks as

$$\Psi_{T_0}(i,j)=\left|\frac{1}{\sqrt{T_0}}\sum_{t\in\mathcal{T}_0}\left\{e_{t,\epsilon}^{(i)}e_{t,\epsilon}^{(j)}-\mathbb{E}e_{t,\epsilon}^{(i)}e_{t,\epsilon}^{(j)}\right\}\right|,$$

where we recall that $e_{t,\epsilon}^{(i)}=\epsilon_{0,t}-\epsilon_{i,t}$. From Condition 7 and Theorem 3.49 in White (1984),

$e_{t,\epsilon}^{(i)} e_{t,\epsilon}^{(j)}$ for $i, j \in \{1, \ldots, J\}$ is either an α-mixing sequence with the mixing coefficient $\alpha = -r/(r-2)$ or a ϕ-mixing sequence with the mixing coefficient $\phi = -r/(2r-1)$, $r \geq 2$. Moreover, $\mathrm{var}\left(e_{t,\epsilon}^{(i)} e_{t,\epsilon}^{(j)}\right)$ can be uniformly bounded, as a result of Condition 8 (i). Using Theorem 5.20 in White (1984), such properties of $e_{t,\epsilon}^{(i)} e_{t,\epsilon}^{(j)}$ together with Condition 8 (ii) imply that

$$\sum_{i=1}^{J} \sum_{j=1}^{J} \Psi_{T_0}(i, j) = O_p(J^2). \tag{B.4}$$

Combining (B.3), (B.4) and Condition 9, we can obtain (B.1).

Then, we show that (B.2) holds. Condition 1 implies (A.4), and Conditions 2–4 imply Inequalities (A.5) and (A.6). With (A.4)–(A.6) and Condition 5, we have that

$$\begin{aligned}
&\sup_{\mathbf{w} \in \mathcal{H}} \left| \frac{R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \\
\leq\ & \xi_{T_0}^{-1} \sup_{\mathbf{w} \in \mathcal{H}} |R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w})| \\
=\ & O(\xi_{T_0}^{-1} T_0^{-1/2} J^2) + o(1).
\end{aligned}$$

This equation, combined with Condition 9, leads to (B.2).

Note that the proof above holds regardless of whether T_1 is finite or divergent, as does (11) in Theorem 2.

Next, we prove (12) in Theorem 2 when T_1 diverges at rate $O(T_0)$. Since

$$L_{T_0}(\mathbf{w}) = L_{T_1}(\mathbf{w}) + L_{T_0}(\mathbf{w}) - R_{T_0}(\mathbf{w}) + R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w}) + R_{T_1}(\mathbf{w}) - L_{T_1}(\mathbf{w})$$

and (B.1)–(B.2) hold when T_1 diverges at rate $O(T_0)$, by Lemma 1 in Gao et al. (2019), it suffices to show that for a divergent T_1 at rate $O(T_0)$,

$$\sup_{\mathbf{w} \in \mathcal{H}} \left| \frac{L_{T_1}(\mathbf{w}) - R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| = o_p(1). \tag{B.5}$$

Similar to the arguments of (B.3)–(B.4), under Conditions 1, 7, 8 (i) and 8 (iii) and a divergent T_1 at rate $O(T_0)$, we can obtain that

$$\sup_{\mathbf{w} \in \mathcal{H}} \left| \frac{L_{T_1}(\mathbf{w}) - R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| = O_p(\xi_{T_0}^{-1} T_0^{-1/2} J^2).$$

This equation together with Condition 9 leads to (B.5), which further verifies (12).

Finally, we prove (13) in Theorem 2 when T_1 is divergent at rate $O(T_0)$. We can bound

$\xi_{T_1}^{-1} L_{T_1}(\widehat{\mathbf{w}})$ from both above and below as

$$
\begin{aligned}
& \xi_{T_1}^{-1} L_{T_1}(\widehat{\mathbf{w}}) \\
=\ & \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\widehat{\mathbf{w}})}{L_{T_1}(\mathbf{w})} \frac{L_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \frac{R_{T_0}(\mathbf{w})}{R_{T_1}(\mathbf{w})} \right| \\
\leq\ & \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\widehat{\mathbf{w}})}{L_{T_1}(\mathbf{w})} \right| \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{R_{T_0}(\mathbf{w})}{R_{T_1}(\mathbf{w})} \right| \\
\leq\ & \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\widehat{\mathbf{w}})}{L_{T_1}(\mathbf{w})} \right| \left\{ 1 + \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\mathbf{w}) - R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \right\} \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{R_{T_0}(\mathbf{w})}{R_{T_1}(\mathbf{w})} \right| && \text{(B.6)} \\
=\ & \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\widehat{\mathbf{w}})}{L_{T_1}(\mathbf{w})} \right| \left\{ 1 + \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\mathbf{w}) - R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \right\} \left\{ \inf_{\mathbf{w}\in\mathcal{H}} \left| \frac{R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \right\}^{-1} \\
\leq\ & \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\widehat{\mathbf{w}})}{L_{T_1}(\mathbf{w})} \right| \left\{ 1 + \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\mathbf{w}) - R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \right\} \left\{ 1 - \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \right\}^{-1},
\end{aligned}
$$

and

$$
\begin{aligned}
& \xi_{T_1}^{-1} L_{T_1}(\widehat{\mathbf{w}}) \\
=\ & \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\widehat{\mathbf{w}})}{L_{T_1}(\mathbf{w})} \frac{L_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \frac{R_{T_0}(\mathbf{w})}{R_{T_1}(\mathbf{w})} \right| \\
\geq\ & \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\widehat{\mathbf{w}})}{L_{T_1}(\mathbf{w})} \right| \inf_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \inf_{\mathbf{w}\in\mathcal{H}} \left| \frac{R_{T_0}(\mathbf{w})}{R_{T_1}(\mathbf{w})} \right| \\
\geq\ & \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\widehat{\mathbf{w}})}{L_{T_1}(\mathbf{w})} \right| \left\{ 1 - \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\mathbf{w}) - R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \right\} \inf_{\mathbf{w}\in\mathcal{H}} \left| \frac{R_{T_0}(\mathbf{w})}{R_{T_1}(\mathbf{w})} \right| && \text{(B.7)} \\
=\ & \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\widehat{\mathbf{w}})}{L_{T_1}(\mathbf{w})} \right| \left\{ 1 - \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\mathbf{w}) - R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \right\} \left\{ \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \right\}^{-1} \\
\geq\ & \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\widehat{\mathbf{w}})}{L_{T_1}(\mathbf{w})} \right| \left\{ 1 - \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{L_{T_1}(\mathbf{w}) - R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \right\} \left\{ 1 + \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| \right\}^{-1}.
\end{aligned}
$$

Based on (12), (B.2) and (B.5), these bounds suggest that $\xi_{T_1}^{-1} L_{T_1}(\widehat{\mathbf{w}}) = 1 + o_p(1)$, which further leads to $\xi_{T_1}^{-1} \{L_{T_1}(\widehat{\mathbf{w}}) - \xi_{T_1}\} = o_p(1)$. Hence, based on the uniform integrability of $\xi_{T_1}^{-1} \{L_{T_1}(\widehat{\mathbf{w}}) - \xi_{T_1}\}$, we can show (13) in Theorem 2 for a divergent T_1 at rate $O(T_0)$. This completes the proof of Theorem 2. $\qquad\square$

C   Proof of Theorem 3

The proof of (17) in Theorem 3 resembles that of (11) in Theorem 2, namely, we need to show

$$\sup_{\mathbf{w}\in\mathcal{H}, d\in\mathcal{D}} \left| \frac{L_{T_0}(\mathbf{w}, d) - R_{T_0}(\mathbf{w}, d)}{R_{T_0}(\mathbf{w}, d)} \right| = o_p(1), \tag{C.1}$$

and

$$\sup_{\mathbf{w}\in\mathcal{H}, d\in\mathcal{D}} \left| \frac{R_{T_0}(\mathbf{w}, d) - R_{T_1}(\mathbf{w}, d)}{R_{T_0}(\mathbf{w}, d)} \right| = o(1). \tag{C.2}$$

We first show (C.1). We use the same notations $\widetilde{\boldsymbol{\lambda}}_t$, $\widetilde{\boldsymbol{\mu}}_i$, $\mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(i)}$ and $e_{t,\epsilon}^{(i)}$ defined as above. Under the linear factor structure (1), we have

$$
\begin{aligned}
& L_{T_0}(\mathbf{w}, d) \\
&= \frac{1}{T_0} \left\{ \sum_{t\in\mathcal{T}_0} \left(y_{0,t} - \sum_{j=1}^{J} w_j y_{j,t} - d \right)^2 + \left\| \mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j - d\boldsymbol{\iota}_r \right\|^2 \right\} \\
&= L_{T_0}(\mathbf{w}) + d^2 - \frac{2d}{T_0} \sum_{t\in\mathcal{T}_0} \left(y_{0,t} - \sum_{j=1}^{J} w_j y_{j,t} \right) + \frac{rd^2}{T_0} - \frac{2d}{T_0}\boldsymbol{\iota}_r^{\top} \left(\mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j \right) \\
&= L_{T_0}(\mathbf{w}) + \frac{d^2(T_0 + r)}{T_0} - \frac{2d}{T_0} \sum_{t\in\mathcal{T}_0}\sum_{j=1}^{J} w_j \left\{ \widetilde{\boldsymbol{\lambda}}_t^{\top} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)} + e_{t,\epsilon}^{(j)} \right\} - \frac{2d}{T_0}\boldsymbol{\iota}_r^{\top} \left(\mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j \right),
\end{aligned}
$$

for any $\mathbf{w}\in\mathcal{H}$ and $d\in\mathcal{D}$. Then under Condition 1, we have

$$
\begin{aligned}
& R_{T_0}(\mathbf{w}, d) \\
&= \mathbb{E}L_{T_0}(\mathbf{w}, d) \\
&= R_{T_0}(\mathbf{w}) + \frac{d^2(T_0 + r)}{T_0} - \frac{2d}{T_0} \sum_{t\in\mathcal{T}_0}\sum_{j=1}^{J} w_j \widetilde{\boldsymbol{\lambda}}_t^{\top} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)} - \frac{2d}{T_0}\boldsymbol{\iota}_r^{\top} \left(\mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j \right). \quad \text{(C.3)}
\end{aligned}
$$

This further suggests that

$$
\begin{aligned}
& \sup_{\mathbf{w}\in\mathcal{H}, d\in\mathcal{D}} \left| \frac{L_{T_0}(\mathbf{w}, d) - R_{T_0}(\mathbf{w}, d)}{R_{T_0}(\mathbf{w}, d)} \right| \\
&\leq \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w}\in\mathcal{H}, d\in\mathcal{D}} \left| L_{T_0}(\mathbf{w}, d) - R_{T_0}(\mathbf{w}, d) \right| \\
&= \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w}\in\mathcal{H}, d\in\mathcal{D}} \left| L_{T_0}(\mathbf{w}) - R_{T_0}(\mathbf{w}) - \frac{2d}{T_0} \sum_{t\in\mathcal{T}_0}\sum_{j=1}^{J} w_j e_{t,\epsilon}^{(j)} \right|
\end{aligned}
$$

41

$$
\begin{aligned}
\leq \quad & \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w} \in \mathcal{H}} |L_{T_0}(\mathbf{w}) - R_{T_0}(\mathbf{w})| + 2C_d \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w} \in \mathcal{H}} \left| \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \sum_{j=1}^{J} w_j e_{t,\epsilon}^{(j)} \right| \\
\leq \quad & \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w} \in \mathcal{H}} |L_{T_0}(\mathbf{w}) - R_{T_0}(\mathbf{w})| + 2C_d \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w} \in \mathcal{H}} \sum_{j=1}^{J} |w_j| \left| \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} e_{t,\epsilon}^{(j)} \right| \\
\leq \quad & \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w} \in \mathcal{H}} |L_{T_0}(\mathbf{w}) - R_{T_0}(\mathbf{w})| + 2C_d \max\{C_\mathrm{L}, C_\mathrm{U}\} \widetilde{\xi}_{T_0}^{-1} T_0^{-1/2} \sum_{j=1}^{J} \Phi_{T_0}(j), \quad \text{(C.4)}
\end{aligned}
$$

where $\Phi_{T_0}(j) = \left| T_0^{-1/2} \sum_{t \in \mathcal{T}_0} e_{t,\epsilon}^{(j)} \right|$. Under Condition 7 and Theorem 3.49 in White (1984), $\left\{ e_{t,\epsilon}^{(j)} \right\}$ for $j \in \{1, \ldots, J\}$ is either an α-mixing sequence with the mixing coefficient $\alpha = -r/(r-2)$ or a ϕ-mixing sequence with the mixing coefficient $\phi = -r/(2r-1)$, $r \geq 2$. Moreover, $\mathrm{var}\left\{ e_{t,\epsilon}^{(j)} \right\}$ is uniformly bounded for $j \in \{1, \ldots, J\}$, based on Condition 8 (i). Furthermore, using Condition 11 (i) and Theorem 5.20 in White (1984), we can obtain that

$$
\sum_{i=1}^{J} \Phi_{T_0}(j) = O_p(J).
$$

Due to Conditions 1, 7, and 8 (i)–(ii), Formulas (B.3)–(B.4) hold, and thus we have

$$
\widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w} \in \mathcal{H}} |L_{T_0}(\mathbf{w}) - R_{T_0}(\mathbf{w})| = O_p(\widetilde{\xi}_{T_0}^{-1} T_0^{-1/2} J^2). \quad \text{(C.5)}
$$

Formulas (C.4) and (C.5) together with Condition 14 imply that (C.1) holds.

We then show (C.2). Similar to the statement in (A.6), we can obtain that

$$
\sup_{\mathbf{w} \in \mathcal{H}} \left| \boldsymbol{\iota}_r^\top \left(\mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j \right) \right| \leq \sup_{\mathbf{w} \in \mathcal{H}} \|\boldsymbol{\iota}_r\| \left\| \mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j \right\| = O_p\left(J^2 \right), \quad \text{(C.6)}
$$

where the last equality is guaranteed by Condition 4 (ii) and that r is a fixed number. Similar to the analysis of (A.4)–(A.7), we can show that

$$
\sup_{\mathbf{w} \in \mathcal{H}} |R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w})| = O_p(T_0^{-1/2} J^2) + o(\widetilde{\xi}_{T_0}), \quad \text{(C.7)}
$$

under Conditions 1–4 and 13. Similar to (C.3), Conditions 3 and 12 imply that

$$
R_{T_1}(\mathbf{w}, d) = R_{T_1}(\mathbf{w}) + d^2 - \frac{2d}{T_1} \sum_{t \in \mathcal{T}_1} \sum_{j=1}^{J} w_j \widetilde{\boldsymbol{\lambda}}_t^\top \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)}.
$$

Thus, we obtain that

$$
\begin{aligned}
& \sup_{\mathbf{w}\in\mathcal{H},d\in\mathcal{D}} \left| \frac{R_{T_0}(\mathbf{w},d) - R_{T_1}(\mathbf{w},d)}{R_{T_0}(\mathbf{w},d)} \right| \\
\leq\ & \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w}\in\mathcal{H},d\in\mathcal{D}} |R_{T_0}(\mathbf{w},d) - R_{T_1}(\mathbf{w},d)| \\
\leq\ & 2C_d \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w}\in\mathcal{H}} \left| \sum_{j=1}^{J} w_j \left(\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\widetilde{\boldsymbol{\lambda}}_t - \frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\widetilde{\boldsymbol{\lambda}}_t \right)^{\top} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)} \right| \\
& + \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w}\in\mathcal{H}} |R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w})| + rC_d^2\widetilde{\xi}_{T_0}^{-1}T_0^{-1} + 2C_d\widetilde{\xi}_{T_0}^{-1}T_0^{-1} \sup_{\mathbf{w}\in\mathcal{H}} \left| \boldsymbol{\iota}_r^{\top}\left(\mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j \right) \right| \\
\leq\ & 2C_d \max\{C_{\mathrm{L}}, C_{\mathrm{U}}\} \widetilde{\xi}_{T_0}^{-1} \sum_{j=1}^{J} \left| \left(\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\widetilde{\boldsymbol{\lambda}}_t - \frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\widetilde{\boldsymbol{\lambda}}_t \right)^{\top} \mathbf{e}_{\widetilde{\boldsymbol{\mu}}}^{(j)} \right| \\
& + \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w}\in\mathcal{H}} |R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w})| + rC_d^2\widetilde{\xi}_{T_0}^{-1}T_0^{-1} + 2C_d\widetilde{\xi}_{T_0}^{-1}T_0^{-1} \sup_{\mathbf{w}\in\mathcal{H}} \left| \boldsymbol{\iota}_r^{\top}\left(\mathbf{Z}_0 - \sum_{j=1}^{J} w_j \mathbf{Z}_j \right) \right| \\
=\ & O\left(\widetilde{\xi}_{T_0}^{-1} T_0^{-1/2} J^2 \right) + o(1), \tag{C.8}
\end{aligned}
$$

where the last equality holds due to (C.6)–(C.7) together with Conditions 3 and 12 (i). Combining (C.8) with Condition 14, we can obtain (C.2). Thus, (17) holds.

Next, we show (18) in Theorem 3. It suffices to show that

$$
\sup_{\mathbf{w}\in\mathcal{H},d\in\mathcal{D}} \left| \frac{L_{T_1}(\mathbf{w},d) - R_{T_1}(\mathbf{w},d)}{R_{T_0}(\mathbf{w},d)} \right| = o_p(1). \tag{C.9}
$$

This equation holds because when T_1 is divergent at rate $O(T_0)$,

$$
\sup_{\mathbf{w}\in\mathcal{H},d\in\mathcal{D}} \left| \frac{L_{T_1}(\mathbf{w},d) - R_{T_1}(\mathbf{w},d)}{R_{T_0}(\mathbf{w},d)} \right| = O_p(\widetilde{\xi}_{T_0}^{-1}T_0^{-1/2} J^2),
$$

under Conditions 1, 7, 8 (i), 8 (iii) and 11 (ii), similar to the analyses of (C.4)–(C.5).

Finally, we show (19) in Theorem 3. Similar to the arguments of (B.6)–(B.7), we can show that

$$
\begin{aligned}
\widetilde{\xi}_{T_1}^{-1} L_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}}, \widehat{d}^{\mathrm{DSC}}) \leq\ & \sup_{\mathbf{w}\in\mathcal{H},d\in\mathcal{D}} \left| \frac{L_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}}, d)}{L_{T_1}(\mathbf{w},d)} \right| \left\{ 1 + \sup_{\mathbf{w}\in\mathcal{H},d\in\mathcal{D}} \left| \frac{L_{T_1}(\mathbf{w},d) - R_{T_1}(\mathbf{w},d)}{R_{T_0}(\mathbf{w},d)} \right| \right\} \\
& \left\{ 1 - \sup_{\mathbf{w}\in\mathcal{H},d\in\mathcal{D}} \left| \frac{R_{T_0}(\mathbf{w},d) - R_{T_1}(\mathbf{w},d)}{R_{T_0}(\mathbf{w},d)} \right| \right\}^{-1}, \tag{C.10}
\end{aligned}
$$

and

$$\widetilde{\xi}_{T_1}^{-1} L_{T_1}(\widehat{\mathbf{w}}^{\text{DSC}}, \widehat{d}^{\text{DSC}}) \geq \sup_{\mathbf{w} \in \mathcal{H}, d \in \mathcal{D}} \left| \frac{L_{T_1}(\widehat{\mathbf{w}}^{\text{DSC}}, d)}{L_{T_1}(\mathbf{w}, d)} \right| \left\{ 1 - \sup_{\mathbf{w} \in \mathcal{H}, d \in \mathcal{D}} \left| \frac{L_{T_1}(\mathbf{w}, d) - R_{T_1}(\mathbf{w}, d)}{R_{T_0}(\mathbf{w}, d)} \right| \right\}$$

$$\left\{ 1 + \sup_{\mathbf{w} \in \mathcal{H}, d \in \mathcal{D}} \left| \frac{R_{T_0}(\mathbf{w}, d) - R_{T_1}(\mathbf{w}, d)}{R_{T_0}(\mathbf{w}, d)} \right| \right\}^{-1}. \tag{C.11}$$

By (18), (C.2) and (C.9)–(C.11), we obtain that $\widetilde{\xi}_{T_1}^{-1} L_{T_1}(\widehat{\mathbf{w}}^{\text{DSC}}, \widehat{d}^{\text{DSC}}) = 1 + o_p(1)$, which further implies that $\widetilde{\xi}_{T_1}^{-1} \left\{ L_{T_1}(\widehat{\mathbf{w}}^{\text{DSC}}, \widehat{d}^{\text{DSC}}) - \widetilde{\xi}_{T_1} \right\} = o_p(1)$. Hence, by the uniform integrability of $\widetilde{\xi}_{T_1}^{-1} \left\{ L_{T_1}(\widehat{\mathbf{w}}^{\text{DSC}}, \widehat{d}^{\text{DSC}}) - \widetilde{\xi}_{T_1} \right\}$, we show (19) in Theorem 3 when T_1 is divergent at rate $O(T_0)$. This completes the proof of Theorem 3. $\qquad\square$

Online Appendix to

Asymptotic Properties of the Synthetic Control Method

This file contains additional theoretical discussions and proofs. Specifically, Section S.4 provides further discussion on the sufficient condition of Assumption 3.2 in Ferman (2021). Section S.5 relaxes the assumption of linear factor DGPs and proves the convergence of SC weight and the asymptotic optimality of SCM in a model-free setup.

All limiting processes in the theoretical results below correspond to $n \to \infty$ unless stated otherwise.

S.1 Explanation of the sufficient condition of Assumption 3.2 in Ferman (2021)

In this section, we prove that (9) and (10) are sufficient to guarantee Assumption 3.2 of Ferman (2021), i.e., $\left\| \boldsymbol{\mu}_0 - \mathbf{M}\mathbf{w}_{T_1}^{\text{opt}} \right\| \to 0$ and $\left\| \mathbf{w}_{T_1}^{\text{opt}} \right\| \to 0$.

Denote $\{\Lambda_1, \ldots, \Lambda_J\}$ as the singular value of matrix $\mathbf{M}^\top \mathbf{Q}\mathbf{M}$, and use $\zeta_{min}(\cdot)$ and $\zeta_{max}(\cdot)$ to represent the minimum and maximum singular value of a matrix. From (9) and (10), we have

$$
\begin{aligned}
&\min\{\Lambda_1, \ldots, \Lambda_J\} \\
={}& J\zeta_{min}\left(J^{-1}\mathbf{M}^\top \mathbf{Q}\mathbf{M}\right) \\
\geq{}& J\zeta_{min}\left(J^{-1}\mathbf{M}\mathbf{M}^\top\right)\zeta_{min}(\mathbf{Q}) \\
={}& J\lambda_{min}\left(J^{-1}\mathbf{M}\mathbf{M}^\top\right)\lambda_{min}(\mathbf{Q}) \\
\geq{}& c_1 c_3 J
\end{aligned}
\tag{S.1}
$$

and

$$
\zeta_{max}(\mathbf{M}) = \sqrt{J\lambda_{max}\left(J^{-1}\mathbf{M}^\top \mathbf{M}\right)} = O(J^{1/2}). \tag{S.2}
$$

With the analytical solution of $\mathbf{w}_{T_1}^{\text{opt}}$, we can obtain that

$$
\boldsymbol{\mu}_0 - \mathbf{M}\mathbf{w}_{T_1}^{\text{opt}} = \left\{ \mathbf{I}_F - \mathbf{M}\left(\mathbf{M}^\top \mathbf{Q}\mathbf{M} + \sigma_\epsilon^2 \mathbf{I}_J\right)^{-1}\mathbf{M}^\top \mathbf{Q} \right\} \boldsymbol{\mu}_0
$$

$$-\frac{1}{2}\mathbf{M}\left(\mathbf{M}^{\top}\mathbf{Q}\mathbf{M} + \sigma_{\epsilon}^{2}\mathbf{I}_{J}\right)^{-1}\left(\boldsymbol{\rho}_{1} - \rho_{2}\boldsymbol{\iota}_{J}\right).$$

Let $\mathbf{P} = \left\{\mathbf{I}_{F} - \mathbf{M}\left(\mathbf{M}^{\top}\mathbf{Q}\mathbf{M} + \sigma_{\epsilon}^{2}\mathbf{I}_{J}\right)^{-1}\mathbf{M}^{\top}\mathbf{Q}\right\}$.

We first consider the case where $\boldsymbol{\rho}_{1} - \rho_{2}\boldsymbol{\iota}_{J} = 0$. Using (S.1) and Condition 3, we have that

$$
\begin{aligned}
& \left\|\boldsymbol{\mu}_{0} - \mathbf{M}\mathbf{w}_{T_{1}}^{\text{opt}}\right\| \\
={}& \left\|\mathbf{P}\boldsymbol{\mu}_{0}\right\| \\
\leq{}& F^{1/2}C_{0}\zeta_{max}\left(\mathbf{P}\right) \\
\leq{}& F^{1/2}C_{0}\left\{1 - \zeta_{min}\left(\mathbf{M}\left(\mathbf{M}^{\top}\mathbf{Q}\mathbf{M} + \sigma_{\epsilon}^{2}\mathbf{I}_{J}\right)^{-1}\mathbf{M}^{\top}\mathbf{Q}\right)\right\} \\
={}& F^{1/2}C_{0}\left\{1 - \zeta_{min}\left(\left(\mathbf{M}^{\top}\mathbf{Q}\mathbf{M} + \sigma_{\epsilon}^{2}\mathbf{I}_{J}\right)^{-1}\mathbf{M}^{\top}\mathbf{Q}\mathbf{M}\right)\right\} \\
={}& F^{1/2}C_{0}\left\{1 - \frac{\min\{\Lambda_{1},\ldots,\Lambda_{J}\}}{\min\{\Lambda_{1},\ldots,\Lambda_{J}\} + \sigma_{\epsilon}^{2}}\right\} \\
\leq{}& \frac{F^{1/2}C_{0}\sigma_{\epsilon}^{2}}{c_{1}c_{3}J + \sigma_{\epsilon}^{2}}.
\end{aligned}
\tag{S.3}
$$

Thus, $\left\|\boldsymbol{\mu}_{0} - \mathbf{M}\mathbf{w}_{T_{1}}^{\text{opt}}\right\| \to 0$ holds as J diverges. Moreover, $\boldsymbol{\rho}_{1} - \rho_{2}\boldsymbol{\iota}_{J} = 0$ also implies that

$$\mathbf{w}_{T_{1}}^{\text{opt}} = \sigma_{\epsilon}^{-2}\mathbf{M}^{\top}\mathbf{Q}\left(\boldsymbol{\mu}_{0} - \mathbf{M}\mathbf{w}_{T_{1}}^{\text{opt}}\right).$$

Thus, along with (10), (S.2) and (S.3), we obtain that

$$\left\|\mathbf{w}_{T_{1}}^{\text{opt}}\right\| \leq \sigma_{\epsilon}^{-2}\zeta_{max}\left(\mathbf{M}\right)\zeta_{max}\left(\mathbf{Q}\right)\left\|\boldsymbol{\mu}_{0} - \mathbf{M}\mathbf{w}_{T_{1}}^{\text{opt}}\right\| = O\left(J^{-1/2}\right).$$

This implies that $\left\|\mathbf{w}_{T_{1}}^{\text{opt}}\right\| \to 0$ as J diverges.

Next, we consider $\boldsymbol{\rho}_{1} - \rho_{2}\boldsymbol{\iota}_{J} \neq 0$. In this case, one may need stronger conditions to guarantee $\left\|\mathbf{w}_{T_{1}}^{\text{opt}}\right\| \to 0$ and $\left\|\boldsymbol{\mu}_{0} - \mathbf{M}\mathbf{w}_{T_{1}}^{\text{opt}}\right\| \to 0$. For example, $\left\|\mathbf{w}_{T_{1}}^{\text{opt}}\right\|$ and $\left\|\boldsymbol{\mu}_{0} - \mathbf{M}\mathbf{w}_{T_{1}}^{\text{opt}}\right\|$ both converge to zero when $\lambda_{min}\left(\mathbf{M}^{\top}\mathbf{Q}\mathbf{M} + \sigma_{\epsilon}^{2}\mathbf{I}_{J}\right)$ diverges faster than $\zeta_{max}(\mathbf{M})\left\|\boldsymbol{\rho}_{1} - \rho_{2}\boldsymbol{\iota}_{J}\right\|$ as J increases, and this condition can be interpreted as an increasing amount of information in the factors and loadings as J diverges. Intuitively, this condition means that the factor structure can be recovered if the newly added control units provide additional information on the factors and loadings.

S.2 Relaxing the linear factor model DGP

In this section, we examine the asymptotic behavior of SC weights and asymptotic optimality of SC estimators without assuming the DGP to be a linear factor model. For coherence, we list the theorems again.

S.2.1   Convergence rate of the SC weight

Theorem 4. *Given any T_1, if $\mathbf{w}_{T_1}^{opt}$ is an interior point of $\mathcal{H}$ and Conditions 4 (ii), 6 and 15–16 hold, then*

$$\left\|\widehat{\mathbf{w}} - \mathbf{w}_{T_1}^{opt}\right\| = O_p\left(T_0^\nu \xi_{T_0}^{1/2} + T_0^\nu \xi_{T_1}^{1/2} + T_0^{-1/4+\nu} J\right),$$

where $\nu > 0$ is a sufficiently small constant, $\xi_{T_0} = \inf_{\mathbf{w}\in\mathcal{H}} R_{T_0}(\mathbf{w})$, and $\xi_{T_1} = \inf_{\mathbf{w}\in\mathcal{H}} R_{T_1}(\mathbf{w})$.

This theorem generalizes Theorem 1 in a model-free setup.

Proof of Theorem 4. We follow the proof of Theorem 1 to verify Theorem 4. Thus, it suffices to show that for any $\varepsilon > 0$, there exists a constant C_ε, such that (A.1) holds, under a given T_1 and any sufficient large T_0, without assuming the DGP to be a linear factor structure. Recall that $\tau_{T_0} = T_0^\nu \xi_{T_0}^{1/2} + T_0^\nu \xi_{T_1}^{1/2} + T_0^{-1/4+\nu} J$, and $\mathbf{u} = (u_1, \ldots, u_J)^\top$ satisfying $\|\mathbf{u}\| = C_\varepsilon$ and $\left(\mathbf{w}_{T_1}^{\mathrm{opt}} + \tau_{T_0}\mathbf{u}\right) \in \mathcal{H}$. Based on the analysis in Appendix A, we need to investigate the order of $\Delta_1 = -2\tau_{T_0} T_0^{-1}\left(\mathbf{Y}_0 - \boldsymbol{\mathcal{Y}}_c \mathbf{w}_{T_1}^{\mathrm{opt}}\right)^\top \boldsymbol{\mathcal{Y}}_c \mathbf{u}$ and $\Delta_2 = \tau_{T_0}^2 T_0^{-1} \left\|\boldsymbol{\mathcal{Y}}_c \mathbf{u}\right\|^2$ and then verify that Δ_2 asymptotically dominates Δ_1.

First, we consider the order of Δ_2. Based on Condition 6, we can obtain that (A.2)–(A.3) hold with probablity approaching to 1, without specifying the model for the DGP. Next, we consider the order of Δ_1. From Conditions 4 (ii) and 15, we can also obtain (A.6). Then, with (A.6) and Condition 16, we obtain that

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}} \left|R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w})\right| \\
&= \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{1}{T_0} \sum_{t\in\mathcal{T}_0} \mathbb{E}\left(y_{0,t}^N - \sum_{j=1}^J w_j y_{j,t}^N\right)^2 + \frac{1}{T_0}\left\|\mathbf{Z}_0 - \sum_{j=1}^J w_j \mathbf{Z}_j\right\|^2 - R_{T_1}(\mathbf{w}) \right| \\
&= \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{1}{T_0} \sum_{t\in\mathcal{T}_0} \mathbb{E}\left(y_{0,t}^N - \sum_{j=1}^J w_j y_{j,t}^N\right)^2 + \frac{1}{T_0}\left\|\mathbf{Z}_0 - \sum_{j=1}^J w_j \mathbf{Z}_j\right\|^2 - R_{T_1}(\mathbf{w}) \right| \\
&\leq \sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{1}{T_0} \sum_{t\in\mathcal{T}_0} \mathbb{E}\left(y_{0,t}^N - \sum_{j=1}^J w_j y_{j,t}^N\right)^2 - R_{T_1}(\mathbf{w}) \right| + \sup_{\mathbf{w}\in\mathcal{H}} \frac{1}{T_0}\left\|\mathbf{Z}_0 - \sum_{j=1}^J w_j \mathbf{Z}_j\right\|^2 \\
&= O(T_0^{-1/2} J^2) + o(\xi_{T_0}). \tag{S.4}
\end{aligned}$$

Similar to the analyses of (A.8)–(A.9), combined with (A.2), we can derive the order of Δ_1 as

$$\begin{aligned}
|\Delta_1| &\leq \frac{2\tau_{T_0}}{T_0}\left\|\mathbf{Y}_0 - \boldsymbol{\mathcal{Y}}_c \mathbf{w}_{T_1}^{\mathrm{opt}}\right\|\left\|\boldsymbol{\mathcal{Y}}_c \mathbf{u}\right\| \\
&= O_p(\tau_{T_0}\xi_{T_1}^{1/2})\|\mathbf{u}\| + O_p(\tau_{T_0} T_0^{-1/4} J)\|\mathbf{u}\| + o_p(\tau_{T_0}\xi_{T_0}^{1/2})\|\mathbf{u}\|.
\end{aligned}$$

This formula, together with (A.3), indicates that Δ_2 asymptotically dominates Δ_1. This completes the proof of Theorem 4. $\qquad\square$

S.2.2   Asymptotic optimality of SCM without an intercept

Theorem 5. *If T_1 is finite, then under Conditions 4 (ii), 9, 15–17 and 18 (i)–(ii), we have*

$$\frac{R_{T_1}(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{H}} R_{T_1}(\mathbf{w})} \xrightarrow{p} 1. \tag{S.5}$$

If T_1 diverges at rate $O(T_0)$, then under Conditions 4 (ii), 9 and 15–18, we have (S.5) and

$$\frac{L_{T_1}(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{H}} L_{T_1}(\mathbf{w})} \xrightarrow{p} 1; \tag{S.6}$$

furthermore, if $\xi_{T_1}^{-1}\left\{L_{T_1}(\widehat{\mathbf{w}}) - \xi_{T_1}\right\}$ is uniformly integrable, then

$$\frac{\mathbb{E}L_{T_1}(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{H}} R_{T_1}(\mathbf{w})} \to 1. \tag{S.7}$$

The results in (S.5)–(S.7) are the same in (11)–(13) in Theorem 2, but relax the assumption on the linear factor model as the outcome process.

Proof of Theorem 5. We follow the proof of Theorem 2 to verify Theorem 5. Based on the analyses in Appendix B, it suffices to show (B.1)–(B.2) in order to prove (S.5). Hence, we first prove (B.1), without the assumption of a linear factor DGP. Note that (B.3) continues to hold regardless of whether the DGP is a linear factor model. With $e_{t,y^N}^{(i)} = y_{0,t}^N - y_{i,t}^N$ for $i \in \{1, \dots, J\}$ and $t \in \mathcal{T}_0 \cup \mathcal{T}_1$, we can rewrite $\Psi_{T_0}(i,j)$ in (B.3) as

$$\Psi_{T_0}(i,j) = \left| \frac{1}{\sqrt{T_0}} \sum_{t\in\mathcal{T}_0} \left\{ e_{t,y^N}^{(i)} e_{t,y^N}^{(j)} - \mathbb{E}e_{t,y^N}^{(i)} e_{t,y^N}^{(j)} \right\} \right|.$$

Similar to the statements of (B.4), from Condition 17 and Conditions 18 (i)–(ii), we can obtain that

$$\sum_{i=1}^{J}\sum_{j=1}^{J} \Psi_{T_0}(i,j) = O_p(J^2). \tag{S.8}$$

Combining (B.3), (S.8) and Condition 9, we can obtain (B.1). Next, we consider the proof of (B.2). With Conditions 4 (ii), 15 and 16, we obtain (S.4); thus, (B.2) follows from (S.4) and Condition 9. Note that the proof above is valid regardless of whether T_1 is finite or

divergent. Therefore, we complete the proof of (S.5) under a fixed T_1, as well as when T_1 diverges.

Next, we prove (S.6) when T_1 diverges at rate $O(T_0)$. From the arguments in the proof of (12) in Appendix B, we need to show (B.5) without assuming the DGP to be a factor model. Similar to the arguments of (B.3) and (S.8), under Conditions 17, 18 (i) and 18 (iii) and a divergent T_1 at rate $O(T_0)$, we can obtain that

$$\sup_{\mathbf{w} \in \mathcal{H}} \left| \frac{L_{T_1}(\mathbf{w}) - R_{T_1}(\mathbf{w})}{R_{T_0}(\mathbf{w})} \right| = O_p(\xi_{T_0}^{-1} T_0^{-1/2} J^2).$$

This equation together with Condition 9 leads to (B.5), which further proves (S.6).

Finally, we verify (S.7) when T_1 diverges at rate $O(T_0)$. As (B.6)–(B.7) continue to hold without a linear factor DGP, combined with (B.2) and (B.5), we can obtain that $\xi_{T_1}^{-1} L_{T_1}(\widehat{\mathbf{w}}) = 1 + o_p(1)$, which further leads to $\xi_{T_1}^{-1} \{L_{T_1}(\widehat{\mathbf{w}}) - \xi_{T_1}\} = o_p(1)$. Moreover, by the uniform integrability of $\xi_{T_1}^{-1} \{L_{T_1}(\widehat{\mathbf{w}}) - \xi_{T_1}\}$, we can prove (S.7) for a divergent T_1 at rate $O(T_0)$. This completes the proof of Theorem 5. $\qquad \square$

S.2.3 Asymptotic optimality of SCM with an intercept

Condition S2.1.

(i) There exists a constant $\widetilde{C}_4$ such that $\mathrm{var}\left\{ T_0^{-1/2} \sum_{t \in \mathcal{T}_0} e_{t,y^N}^{(j)} \right\} \geq \widetilde{C}_4 > 0$ for all T_0 sufficiently large and any $j \in \{1, \ldots, J\}$.

(ii) There exists a constant $\widetilde{C}_5$ such that $\mathrm{var}\left\{ T_1^{-1/2} \sum_{t \in \mathcal{T}_1} e_{t,y^N}^{(j)} \right\} \geq \widetilde{C}_5 > 0$ for all T_1 sufficiently large and any $j \in \{1, \ldots, J\}$.

Condition S2.2. $T_1^{-1} \sum_{t \in \mathcal{T}_1} \left(T_0^{-1} \sum_{k \in \mathcal{T}_0} \mathbb{E}y_{i,k}^N - \mathbb{E}y_{i,t}^N \right) = O\left(T_0^{-1/2}\right)$ for $i \in \{0, 1, \ldots, J\}$ uniformly.

Condition S2.3. $\widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w} \in \mathcal{H}} \left| T_0^{-1} \sum_{t \in \mathcal{T}_0} \mathbb{E} \left(\sum_{j=1}^{J} w_j e_{t,y^N}^{(j)} \right)^2 - T_1^{-1} \sum_{t \in \mathcal{T}_1} \mathbb{E} \left(\sum_{j=1}^{J} w_j e_{t,y^N}^{(j)} \right)^2 \right| = o(1)$.

Condition S5.1 guarantees the application of the central limit theorem for dependent processes, as Condition 11. Compared to Condition 11, this condition directly concerns the potential outcome $y_{i,t}^N$ to allow for a model-free setup. Condition S5.2 can be viewed as a stationary assumption that generalizes Conditions 4 (ii) and 12–13. Condition S5.3 is a generalized version of Condition 13.

Theorem 6. *If T_1 is finite and Conditions 4 (ii), 10, 14–15, 17, 18 (i)–(ii), S5.1 (i) and S5.2–S5.3 hold, then*

$$\frac{R_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}}, \widehat{d}^{\mathrm{DSC}})}{\inf_{\mathbf{w}\in\mathcal{H}, d\in\mathcal{D}} R_{T_1}(\mathbf{w}, d)} \xrightarrow{p} 1. \tag{S.9}$$

If T_1 diverges at rate $O(T_0)$ and Conditions 4 (ii), 10, 14–15 and 17–S5.3 hold, then we have (S.9) and

$$\frac{L_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}}, \widehat{d}^{\mathrm{DSC}})}{\inf_{\mathbf{w}\in\mathcal{H}, d\in\mathcal{D}} L_{T_1}(\mathbf{w}, d)} \xrightarrow{p} 1; \tag{S.10}$$

furthermore, if $\widetilde{\xi}_{T_1}^{-1}\left\{L_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}}, \widehat{d}^{\mathrm{DSC}}) - \widetilde{\xi}_{T_1}\right\}$ is uniformly integrable, then

$$\frac{\mathbb{E}L_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}}, \widehat{d}^{\mathrm{DSC}})}{\inf_{\mathbf{w}\in\mathcal{H}, d\in\mathcal{D}} R_{T_1}(\mathbf{w}, d)} \to 1. \tag{S.11}$$

The results in (S.9)–(S.11) generalize (17)–(19) in Theorem 3, respectively, in a model-free setup.

Proof of Theorem 6. We follow the proof of Theorem 3 to verify Theorem 6. Based on the analysis in Appendix C, to prove (S.9), it suffices to show that (C.1)–(C.2) hold. First, we give the proof of (C.1) without assuming a factor model DGP. Note that

$$L_{T_0}(\mathbf{w}, d)$$
$$= \frac{1}{T_0}\left\{\sum_{t\in\mathcal{T}_0}\left(y_{0,t}^N - \sum_{j=1}^J w_j y_{j,t}^N - d\right)^2 + \left\|\mathbf{Z}_0 - \sum_{j=1}^J w_j \mathbf{Z}_j - d\boldsymbol{\iota}_r\right\|^2\right\}$$
$$= L_{T_0}(\mathbf{w}) + \frac{d^2(T_0+r)}{T_0} - \frac{2d}{T_0}\sum_{t\in\mathcal{T}_0}\sum_{j=1}^J w_j e_{t,y^N}^{(j)} - \frac{2d}{T_0}\boldsymbol{\iota}_r^\top\left(\mathbf{Z}_0 - \sum_{j=1}^J w_j \mathbf{Z}_j\right),$$

for any $\mathbf{w} \in \mathcal{H}$ and $d \in \mathcal{D}$; thus, with Condition 15,

$$R_{T_0}(\mathbf{w}, d)$$
$$= \mathbb{E}L_{T_0}(\mathbf{w}, d)$$
$$= R_{T_0}(\mathbf{w}) + \frac{d^2(T_0+r)}{T_0} - \frac{2d}{T_0}\sum_{t\in\mathcal{T}_0}\sum_{j=1}^J w_j \mathbb{E}e_{t,y^N}^{(j)} - \frac{2d}{T_0}\boldsymbol{\iota}_r^\top\left(\mathbf{Z}_0 - \sum_{j=1}^J w_j \mathbf{Z}_j\right). \tag{S.12}$$

Similiar to (C.4), we have that

$$\sup_{\mathbf{w}\in\mathcal{H}, d\in\mathcal{D}}\left|\frac{L_{T_0}(\mathbf{w}, d) - R_{T_0}(\mathbf{w}, d)}{R_{T_0}(\mathbf{w}, d)}\right|$$

$$\leq \quad \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w}\in\mathcal{H}, d\in\mathcal{D}} \left| L_{T_0}(\mathbf{w}) - R_{T_0}(\mathbf{w}) - \frac{2d}{T_0}\sum_{t\in\mathcal{T}_0}\sum_{j=1}^{J} w_j \left\{ e_{t,y^N}^{(j)} - \mathbb{E}e_{t,y^N}^{(j)} \right\} \right|$$

$$\leq \quad \widetilde{\xi}_{T_0}^{-1} \sup_{\mathbf{w}\in\mathcal{H}} |L_{T_0}(\mathbf{w}) - R_{T_0}(\mathbf{w})| + 2C_d \max\{C_{\mathrm{L}}, C_{\mathrm{U}}\}\widetilde{\xi}_{T_0}^{-1} T_0^{-1/2}\sum_{j=1}^{J}\widetilde{\Phi}_{T_0}(j), \quad \text{(S.13)}$$

where $\widetilde{\Phi}_{T_0}(j) = \left| T_0^{-1/2}\sum_{t\in\mathcal{T}_0}\left\{ e_{t,y^N}^{(j)} - \mathbb{E}e_{t,y^N}^{(j)} \right\}\right|$. Based on Conditions 17, 18 (i) and S5.1 (i), we can obtain that

$$\sum_{j=1}^{J}\widetilde{\Phi}_{T_0}(j) = O_p(J). \quad \text{(S.14)}$$

Note that (B.3) still holds without assuming a linear factor DGP. With Conditions 17 and 18 (i)–(ii), (S.8) holds; thus, combined with (B.3), we can obtain (C.5). Formulas (C.5) and (S.13)–(S.14), together with Condition 14, guarantee that (C.1) holds without assuming a linear factor DGP.

Then, we consider (C.2). From Condition S5.2, we have

$$\sup_{\mathbf{w}\in\mathcal{H}} \left| \sum_{j=1}^{J} w_j \left(\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\mathbb{E}e_{t,y^N}^{(j)} - \frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\mathbb{E}e_{t,y^N}^{(j)} \right)\right|$$

$$= \quad \sup_{\mathbf{w}\in\mathcal{H}} \left| \sum_{j=1}^{J} w_j \left\{ \left(\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\mathbb{E}y_{0,t}^N - \frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\mathbb{E}y_{0,t}^N \right) - \left(\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\mathbb{E}y_{j,t}^N - \frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\mathbb{E}y_{j,t}^N \right) \right\}\right|$$

$$\leq \quad \sup_{\mathbf{w}\in\mathcal{H}} \left| \sum_{j=1}^{J} w_j \left(\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\mathbb{E}y_{0,t}^N - \frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\mathbb{E}y_{0,t}^N \right)\right|$$

$$+ \sup_{\mathbf{w}\in\mathcal{H}} \left| \sum_{j=1}^{J} w_j \left(\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\mathbb{E}y_{j,t}^N - \frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\mathbb{E}y_{j,t}^N \right)\right|$$

$$= \quad O(T_0^{-1/2}). \quad \text{(S.15)}$$

Since r is fixed, from Condition 4 (ii), we obtain (C.6). Similar to the arguments in the proof of (S.4), we can obtain that

$$\sup_{\mathbf{w}\in\mathcal{H}} |R_{T_0}(\mathbf{w}) - R_{T_1}(\mathbf{w})| = O_p(T_0^{-1/2}J^2) + o(\widetilde{\xi}_{T_0}), \quad \text{(S.16)}$$

under Conditions 4 (ii), 15 and S5.3. Since

$$R_{T_1}(\mathbf{w}, d) = R_{T_1}(\mathbf{w}) + d^2 - \frac{2d}{T_1}\sum_{t\in\mathcal{T}_1}\sum_{j=1}^{J} w_j\mathbb{E}e_{t,y^N}^{(j)},$$

with (C.6), (S.12) and (S.15)–(S.16), we have that

$$
\sup_{\mathbf{w}\in\mathcal{H},d\in\mathcal{D}}\left|\frac{R_{T_0}(\mathbf{w},d)-R_{T_1}(\mathbf{w},d)}{R_{T_0}(\mathbf{w},d)}\right|
$$

$$
\leq\ \widetilde{\xi}_{T_0}^{-1}\sup_{\mathbf{w}\in\mathcal{H},d\in\mathcal{D}}|R_{T_0}(\mathbf{w},d)-R_{T_1}(\mathbf{w},d)|
$$

$$
\leq\ 2C_d\widetilde{\xi}_{T_0}^{-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\sum_{j=1}^{J}w_j\left(\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\mathbb{E}e_{t,y^N}^{(j)}-\frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\mathbb{E}e_{t,y^N}^{(j)}\right)\right|+\widetilde{\xi}_{T_0}^{-1}\sup_{\mathbf{w}\in\mathcal{H}}|R_{T_0}(\mathbf{w})-R_{T_1}(\mathbf{w})|
$$

$$
+rC_d^2\widetilde{\xi}_{T_0}^{-1}T_0^{-1}+2C_d\widetilde{\xi}_{T_0}^{-1}T_0^{-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\boldsymbol{\iota}_r^{\top}\left(\mathbf{Z}_0-\sum_{j=1}^{J}w_j\mathbf{Z}_j\right)\right|
$$

$$
\leq\ 2C_d\max\{C_{\mathrm{L}},C_{\mathrm{U}}\}\widetilde{\xi}_{T_0}^{-1}\sum_{j=1}^{J}\left|\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\mathbb{E}e_{t,y^N}^{(j)}-\frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\mathbb{E}e_{t,y^N}^{(j)}\right|+\widetilde{\xi}_{T_0}^{-1}\sup_{\mathbf{w}\in\mathcal{H}}|R_{T_0}(\mathbf{w})-R_{T_1}(\mathbf{w})|
$$

$$
+rC_d^2\widetilde{\xi}_{T_0}^{-1}T_0^{-1}+2C_d\widetilde{\xi}_{T_0}^{-1}T_0^{-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\boldsymbol{\iota}_r^{\top}\left(\mathbf{Z}_0-\sum_{j=1}^{J}w_j\mathbf{Z}_j\right)\right|
$$

$$
=\ O\left(\widetilde{\xi}_{T_0}^{-1}T_0^{-1/2}J^2\right)+o(1). \tag{S.17}
$$

Combining (S.17) with Condition 14, we can obtain (C.2) without assuming a linear factor DGP. Thus, (S.9) holds.

Next, we prove (S.10). Based on the arguments in the proof of (18) in Appendix C, we need to show (C.9) without assuming a linear factor DGP. Similar to the arguments of (S.13)–(S.14), when T_1 is divergent at rate $O(T_0)$,

$$
\sup_{\mathbf{w}\in\mathcal{H},d\in\mathcal{D}}\left|\frac{L_{T_1}(\mathbf{w},d)-R_{T_1}(\mathbf{w},d)}{R_{T_0}(\mathbf{w},d)}\right|=O_p(\widetilde{\xi}_{T_0}^{-1}T_0^{-1/2}J^2) \tag{S.18}
$$

holds under Conditions 17, 18 (i), 18 (iii) and S5.1 (ii). Combining (S.18) with Condition 14, we can obtain (C.9). Thus, (S.10) is verified.

Finally, we consider the proof of (S.11) when T_1 diverges at rate $O(T_0)$. As (C.10)–(C.11) hold without assuming a linear factor DGP, combined with (C.2), (C.9) and (S.10), we obtain $\widetilde{\xi}_{T_1}^{-1}L_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}},\widehat{d}^{\mathrm{DSC}})=1+o_p(1)$, which further implies $\widetilde{\xi}_{T_1}^{-1}\left\{L_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}},\widehat{d}^{\mathrm{DSC}})-\widetilde{\xi}_{T_1}\right\}=o_p(1)$. Hence, by the uniform integrability of $\widetilde{\xi}_{T_1}^{-1}\left\{L_{T_1}(\widehat{\mathbf{w}}^{\mathrm{DSC}},\widehat{d}^{\mathrm{DSC}})-\widetilde{\xi}_{T_1}\right\}$, we show (S.11) in Theorem 6 when T_1 is divergent at rate $O(T_0)$. This completes the proof of Theorem 6. $\quad\square$