A k-fold Inference for Treatment Effects

Dennis Lim*

September 23, 2024

Abstract

In this paper, we introduce and study a k-fold cross-fitting procedure for synthetic control, aimed at conducting valid inference on individual treatment effects. Our proposed estimator generalizes several well-known methods, including the synthetic control method (SCM) developed by Abadie and Gardeazabal (2003) and the difference-in-difference SCM introduced by Ferman and Pinto (2021). Notably, when k is set to the number of pre-treatment periods, our estimator encompasses these established approaches. The motivation behind the k-fold procedure arises from the bias present in existing estimators when treatment assignment is correlated with unobserved confounders. To address this issue, we propose a test for treatment effects that is asymptotically valid, independent of such assumptions. We demonstrate the favorable properties of our test through comprehensive simulation results.

Keywords: Synthetic Control, Treatment Effect, Linear Factor Model

JEL Classification: C13, C21, C23

1 Introduction

In addressing the challenge of estimating treatment effects when the number of treated units is limited, traditional methods often struggle due to insufficient asymptotic approximation. The Synthetic Control Method (SCM), as proposed by Abadie and Gardeazabal (2003), offers a promising alternative. SCM operates by constructing a counterfactual outcome for a treated unit using a weighted average of control units in the pre-treatment period. This approach hinges on the assumption that it is possible to achieve a 'perfect pre-treatment fit' (PPTF), where the outcomes of the treated unit can be exactly replicated by a weighted combination of the control units' outcomes during the pre-treatment period.

Abadie, Diamond, and Hainmueller (2010) provided theoretical justifications for SCM, including necessary conditions for its effectiveness. The key requirement is that weights exist such that a weighted average of control units (J) can perfectly match the treated unit's outcomes in the pre-treatment period. Subsequent work by Doudchenko and Imbens (2016), and Abadie, Diamond,

*Singapore Management University. E-mail address: dennis.lim.2019@phdecons.smu.edu.sg.

and Hainmueller (2015) has upheld this assumption. Under the PPTF condition, these studies directly estimate the treatment effect for the post-treatment period without needing to recover the specific weights, which is challenging due to unobserved factor-loadings. Abadie et al. (2010) also proposed an approximately unbiased estimator due to the difficulty in recovering exact weights. This approach differs from our work in several ways. First, their method assumes an autoregressive factor model, as seen in equation (5) of their paper; second, it requires that the matrix $\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\lambda_t\lambda_t'$ be non-singular, where λ_t is the vector of time-varying common factors. Our method does not impose these assumptions.

For cases with a growing number of control units, Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021) offer an alternative method for inference on treatment effects but assume Gaussian errors that are identically and independently distributed. Xu (2017) relaxes this assumption to allow for weak serial dependence but still requires cross-sectional independence and homoscedasticity of errors. Ben-Michael, Feller, and Rothstein (2021) [Lemma 3] derived finite-sample bias bounds for SCM and demonstrated that these bounds do not converge to zero as the number of control units remains fixed while the number of pre-treatment periods grows. Ferman and Pinto (2021) further showed that SCM cannot asymptotically recover the true weights, demonstrating the inherent limitations in achieving zero asymptotic bias.

In response to these limitations, Amjad, Shah, and Shen (2018) proposed a two-step de-noising algorithm that, while effective, requires careful hyperparameter tuning and assumes serially uncorrelated data. Our method, by contrast, only requires weak dependence in the idiosyncratic errors to provide an unbiased estimate of the treatment effect. Additionally, our results build on and complement the works of Bai (2003) (who argued that estimating factor-loadings consistently when J is fixed under generally requires strong assumptions on the idiosyncratic shocks), Abadie et al. (2010) (they showed that recovery of factor-loadings close enough to the oracle/true values were possible whenever the idiosyncratic variance is relatively small; our proposed method allows for significant idiosyncratic shocks) and Ben-Michael et al. (2021) (they require the number of control units to diverge) by addressing the challenge of recovering pre-treatment weights with a fixed number of control units under PPTF. This is our first contribution.

Table 1: size control with 5,000 replications using different weight estimators in the literature. θ is the nominal-size. SCM is based on Abadie et al. (2010). See Section 4 for more details

DGP3, $\rho_e = 0.9$	Our estimator	Diff-in-diff	SCM	Constrained-LASSO	Ferman and Pinto (2021)
$\theta = \mathbf{0.05}$	0.066	0.07	0.12	0.095	0.08
$\theta = \mathbf{0.1}$	0.122	0.154	0.214	0.161	0.146

An important contribution by Ferman and Pinto (2021) is the introduction of “imperfect pre-treatment fit"(IPTF), where pre-treatment weights cannot be recovered.[1] However, they are still able to provide valid inference for treatment effect by "skipping" the step of estimating the pre-treatment weights by demeaning (as in difference-in-difference method) and assuming that time-invariant fixed effect is stationary. We build on their results and show that our inference procedure remains valid under IPTF whenever the time-invariant fixed effect is stationary. Furthermore, whenever time-invariant fixed effect is non-stationary, our inference procedure remains valid under PPTF while Ferman and Pinto (2021) becomes invalid. This is our second contribution. We provide a brief summary in Table 1.

Our third and final contribution is in studying and deriving the properties of a k-fold cross-fitting procedure, which encompasses well-known estimators as special cases. This is intended to assist future researchers in their work.

Pre-Treatment fit	SCM in Literature	Ferman and Pinto (2021)	Our Method
PPTF, λ_t stationary			
Diverging control-units	Unbiased	Unbiased	Unbiased
Fixed control-units	**Biased**	Unbiased	Unbiased
PPTF, λ_t non-stationary			
Diverging control-units	Unbiased	Unbiased	Unbiased
Fixed control-units	**Biased**	**Biased**	Unbiased
IPTF, λ_t stationary			
Diverging control-units	Unbiased	Unbiased	Unbiased
Fixed control-units	**Biased**	Unbiased	Unbiased
IPTF, λ_t non-stationary			
Diverging control-units	**Biased**	**Biased**	**Biased**
Fixed control-units	**Biased**	**Biased**	**Biased**

Table 2: Bias of SCM estimators in literature, including those discussed above; λ_t is the time-invariant fixed effect

Notations: We write ι to be a vector of ones, where the dimensions will be clear from the context. We write $\xrightarrow{p}$ to denote convergence in probability and $\mathbb{N}_+$ to mean $\{1, 2, ...\}$. For any $A, B \in \mathbb{R}$, we write $A \lesssim B$ to mean than $A \leq KB$, where K is some universal constant.

Structure of Paper: Section 2 provides the model setup of our paper. Section 3 derives the main theoretical results. Section 4 provides simulation results. Proofs of Theorems and Corollaries are contained in the Appendix.

[1]See their Proposition 1

2 Model and Test Statistics

2.1 Model and Motivation

We assume that we observe a panel of $j \in \{0, 1, ..., J+1\}$ individuals for time $t \in \mathcal{T}_0 \cup \mathcal{T}_1$, where $\mathcal{T}_0$ is the set of periods where no individuals are treated and $\mathcal{T}_1$ is the set of periods where individual $j = 0$ is treated with the remaining $j = 1, ..., J+1$ individuals still untreated. We let $T_0 := card(\mathcal{T}_0)$ and $T_1 := card(\mathcal{T}_1)$ be the number of periods associated with $\mathcal{T}_0$ and $\mathcal{T}_1$ respectively; furthermore denote $T := T_0 + T_1$ to be the total number of periods under consideration. We will analyze the following factor model, which has been extensively studied.[2]

Assumption 1 (Factor Model Potential Outcome[3])**.** *The potential outcome for unit j at time t for the treated (y^I_{jt}) and non-treated (y^N_{jt}) are given by*

$$\begin{aligned} y^N_{jt} &= c_j + \delta_t + \lambda'_t \mu_j + \varepsilon_{jt} \\ y^I_{jt} &= \alpha_{jt} + y^N_{jt} \end{aligned} \tag{2.1}$$

where δ_t is an unknown common factor with constant factor loadings across units, c_j is an unknown time-invariant fixed effect, λ_t is a $(F\times 1)$ vector of common factors, μ_j is a $(F\times 1)$ vector of unknown factor loadings, and the error terms ε_{jt} are unobserved idiosyncratic shocks

We are interested in testing the treatment effect for individual $j = 0$

$$H_0 : \alpha_0 = \alpha \quad \text{versus} \quad H_1 : \alpha_0 \neq \alpha$$

where $\alpha_0 = \{\alpha_{0t}\}_{t\in\mathcal{T}_1}$ is some sequence of true parameter and $\alpha = \{\alpha_t\}_{t\in\mathcal{T}_1}$ is a sequence of parameters we hypothesize to be true. We assume throughout this paper that $T_0 \to \infty$; the setting with $T_1 \to \infty$ has been studied extensively by Chernozhukov, Wuthrich, and Zhu (2022) – in particular they introduce a t-test that provides valid inference for the average treatment effect (i.e. $\frac{1}{T_1}\sum_{t\in\mathcal{T}_1} \alpha_{0t}$). Instead, our focus will be on T_1 being fixed, so that the central-limit-theorem argument used to derive their t-test is no longer valid. We assume that our sampling procedure follows the given structure.

Assumption 2 (Sampling structure)**.** *We observe realizations of $\{y_{0t}, \ldots, y_{Jt}\}_{t\in\mathcal{T}_0\cup\mathcal{T}_1}$, where $y_{jt} = d_{jt}y^I_{jt} + (1 - d_{jt})\, y^N_{jt}$, while $d_{jt} = 1$ if $j = 0$ and $t \in \mathcal{T}_1$, and zero otherwise. Potential outcomes are determined by assumption 1. We treat $\{c_j, \mu_j\}_{j=0}^J$ as fixed, $\{\lambda_t, \delta_t\}_{t\in\mathcal{T}_0\cup\mathcal{T}_1}$ and $\{\varepsilon_{jt}\}_{t\in\mathcal{T}_0\cup\mathcal{T}_1, j\in\{0,1,\ldots,J+1\}}$ as stochastic*

To motivate the problem, consider the synthetic control method weights (SCM) introduced by

[2] see Bai (2003) Abadie et al. (2010), Ferman and Pinto (2021) among many others

[3] We assume without loss of generality that $c_{J+1} = 0$, since we can always replace δ_t by $\delta_t + c_{J+1}$.

Abadie et al. (2010)

$$\widehat{W}^{SCM} := \underset{W:\sum_{j=1}^{J+1} W_j=1, W_j\geq 0}{\arg\min} \frac{1}{T_0}\sum_{t\in\mathcal{T}_0}(y_{0t}-y_t'W)^2. \tag{2.2}$$

where $y_t = (y_{1t},...,y_{J+1\,t})' \in \mathbb{R}^{J+1}$. Fixing some weight W such that $\sum_{j=1}^{J+1} W_j = 1$, we can write

$$\begin{aligned}
\widehat{Q}_{T_0}(W) &:= \frac{1}{T_0}\sum_{t\in\mathcal{T}_0}(y_{0t}-y_t'W)^2 \\
&= \frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\left\{c_0+\delta_t+\lambda_t'\mu_0+\varepsilon_{0t}-(c'W+\delta_t\iota'W+\lambda_t'\mu W+\varepsilon_t'W)\right\}^2 \\
&= \frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\left\{(c_0-c'W)+\lambda_t'(\mu_0-\mu W)+(\varepsilon_{0t}-\varepsilon_t'W)\right\}^2 \qquad (2.3)
\end{aligned}$$

where $\varepsilon_t = (\varepsilon_{1t},...,\varepsilon_{J+1\,t})'$, $c := (c_1,...,c_{J+1})'$, $\mu := (\mu_1',...,\mu_{J+1}')' \in \mathbb{R}^{F\times J+1}$ and the last equality follows from $\iota'W = 1$. Under some mild assumptions we can show that[4]

$$\widehat{Q}_{T_0}(W) \xrightarrow{p} \sigma_\varepsilon^2(1+W'W)+\left[(c_0-c'W)^2+(\mu_0-\mu W)'\Omega_0(\mu_0-\mu W)\right] =: Q_0(W), \tag{2.4}$$

where we assume for the moment that $\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\widetilde{\varepsilon}_t\widetilde{\varepsilon}_t' \xrightarrow{p} \sigma_\varepsilon^2 I_{J+2}$ for $\widetilde{\varepsilon}_t := (\varepsilon_{0t},\varepsilon_t')'$ and some constant σ_ε^2. Then it can be shown that

$$\widehat{W}^{SCM} \xrightarrow{p} W^\star \equiv \underset{W\in\mathbb{R}^{J+1}:W'\iota=1,W_j\geq 0}{\arg\min} Q_0(W)$$

The usual SCM estimator for α_{0t} can be given as

$$\widehat{\alpha}_{0t}^{SCM} := y_{0t}-y_t'\widehat{W}^{SCM} \xrightarrow{p} \alpha_{0t}+\lambda_t(\mu_0-\mu'W^\star)+(c_0-c'W^\star)+(\varepsilon_{0t}-\varepsilon_t'W^\star) \tag{2.5}$$

Note that the asymptotic variance of the estimator $\widehat{\alpha}_{0t}^{SCM}$ is $Q_0(W^\star)$, so that the limit of $\widehat{W}^{SCM}$ is actually the minimizer argument for the asymptotic variance of the treatment effect estimator, i.e. $W^\star \in \arg\min_W avar(\widehat{\alpha}_{0t}^{SCM})$

We see that the variance σ_ε^2 given in (2.4) prevents us from recovering the pre-treatment weights when these weights exist, i.e. $\mu W^\star$ is generally not equal to μ_0 and $c'W^\star$ is not equal to c_0. Ferman and Pinto (2021) explain that the only way to fully recover the pre-treatment weights is when $\sigma_\varepsilon^2 = 0$. In view of this short-coming, Ferman (2021) suggests that "when the number of control units increases, the importance of the variance of this weighted average of the idiosyncratic shocks vanishes if it is possible to recover the factor-loadings of the treated unit with weights that are

[4]See Ferman and Pinto (2021)[assumption 4]

diluted among an increasing number of control units". Indeed, when

$$(A) \quad \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \lambda_t \varepsilon_t \xrightarrow{p} 0$$

$$(B) \quad \text{the number of control-units increases}$$

then Ferman (2021) showed that the weights $W^\star$ has the property that $\mu W^\star = \mu_0$ and $c' W^\star = c_0$. Since (A) and (B) may not hold in some settings, we aim to construct an appropriate weight estimator that consistently recovers the true weight without these restrictions.

To achieve this, we address the error term σ_ε^2 by partitioning the non-treated sample period $\mathcal{T}_0$ into sub-samples (or blocks). We then estimate the synthetic control weights by aggregating across these blocks, leveraging the mean-zero property of the error terms to render them negligible. We call this new estimator $\widetilde{W}_T^{DBSCM}$, and we can obtain

$$\arg\min_W \widehat{Q}_{T_0}(\widetilde{W}_T^{DBSCM}) \xrightarrow{p} \arg\min_W \left[(c_0 - c'W)^2 + (\mu_0 - \mu W)' \Omega_0 (\mu_0 - \mu W) \right] \tag{2.6}$$

which holds whenever the RHS of (2.6) has a unique solution. When this uniqueness does not hold, a strictly-convex penalty term can be introduced to induce a unique solution, which is given by $f(\cdot)$ in the next section. Intuitively, the blocking approach removes the need for the number of control units to increase since the variance σ_ε^2 is completely eliminated. Furthermore, under PPTF, we are able to completely recover the true weights. We explain this intuition in more detail in section 2.2 below.

2.2 Test Statistic

Consider a general k-fold cross-fitting procedure, where $k \in \mathbb{N}_+$ and we define $\Delta := \lfloor \frac{T_0}{k} \rfloor$. Then Δ can be seen as the number of elements we want to fit in a single block. Then for any $j \in \{0, 1, ..., J+1\}$, $s \in \{1, ..., k\}$ and $q \in \{1, ..., \Delta\}$ define $\varepsilon_{jq}^s := \varepsilon_{j, s\Delta+q}$ and $\overline{\varepsilon}^s := \frac{1}{\Delta}(\varepsilon_{s\Delta+1} + \varepsilon_{s\Delta+2} + ... + \varepsilon_{(s+1)\Delta})$. Furthermore, write $\overline{y}^s := \frac{1}{\Delta}\left(y_{s\Delta+1} + y_{s\Delta+2} + ... + y_{(s+1)\Delta}\right)$ and $\overline{y}_0^s := \frac{1}{\Delta}\sum_{q=1}^{\Delta} y_{0q}$. Denote $Y_{jt} := y_{jt} - y_{J+1,t}$ for $j = 0, 1, ..., J$, and define the demeaned-block-synthetic-control-method (DBSCM) weight as[5]

$$\widetilde{W}_T^{DBSCM}(f, \Lambda) := \arg\min_{W \in \Delta_\eta^J} \left\{ \frac{1}{k} \sum_{s=1}^{k} \left\{ \overline{Y}_0^s - (\overline{Y}^s)'W - (\overline{Y}_0 - \overline{Y}'W) \right\}^2 + f(W, \Lambda) \right\} \tag{2.7}$$

[5]Note that our weight space Δ_η^J allows for negative weights, which differs from Abadie et al. (2010), Abadie et al. (2015) in that their weights are assumed to be in a simplex.

where $\Lambda \geq 0$ is some given value, $\Delta_\eta^J := \{W \in \mathbb{R}^J : ||W||_1 \leq \eta\}$ is the set containing vectors of weight W whose Frobenius-norm is bounded by some fixed $\eta > 0$ and $f(W, \Lambda)$[6] is some non-negative penalty term with the property that it (is)

1. Strictly-convex in $W \in \mathbb{R}^J$
2. Equi-continuous in $(W, \Lambda) \in \Delta_\eta^J \times [0, 1]$ for $\Lambda \geq 0$
3. Converges to zero for any $0 \leq \Lambda \downarrow 0$
4. Equals zero whenever Λ equals zero

and $(\overline{Y}_0^s, \overline{Y}^s)$ is defined in a similar manner to $(\overline{y}_0^s, \overline{y}^s)$. When $k = T_0$, $\Lambda \equiv 0$, $\eta = 1$ and restricting W to non-negatives, we obtain the estimator

$$\widehat{W}^{FP} := \underset{W \in \Delta_1^J, W_j \geq 0}{\arg\min} \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \left\{ Y_{0t} - Y_t' W - (\overline{Y}_0 - \overline{Y}' W) \right\}^2 \tag{2.8}$$

which was provided by Ferman and Pinto (2021). If we define $f(W, \Lambda) := \Lambda ||W||_1$, then we have the ℓ_1-regularized penalty-term; if instead we take $f(W, \Lambda) := \Lambda ||W||_2^2$ then we have a ridge penalty-term. The main result of the paper is that under the correct null of $\alpha_0 = \alpha$, for any $\theta \in (0, 1)$ we will have

$$\lim_{T \to \infty} \mathbb{P}(\widehat{p}(\widetilde{W}_T^{DBSCM}(f, \Lambda_T)) \leq \theta) = \theta$$

for any sequence of $0 < \Lambda_T \downarrow 0$ and $\widehat{p}(\cdot)$ is defined in section 3.3; this result is given in Corollary 3.4, which allows us to conduct inference on α_0 for any subset of periods in $\mathcal{T}_1$.

To see why the blocking/k-fold approach works, suppose for simplicity that errors ε_{jt} are independent with homoskedastic variance σ_ε^2. Then the variance of each block is approximately $\sigma_\varepsilon^2 / \Delta$, which converges to zero as $\Delta \to \infty$. This removes the influence of the error term in (2.3), allowing us to obtain (2.6).

3 Main Results

Throughout the remainder of Section 3, unless stated otherwise, we will assume that t represents a fixed value within $\mathcal{T}_1$. In Section 3.1, we present a basic version of our proposed weight estimator (2.7) under the Perfect Pre-Treatment Fit (PPTF) condition. This serves to motivate and develop the results in a clear and instructive manner. Section 3.2 explores the performance of the estimator

[6] In Theorem 3 we consider $f(\Lambda, W) := \Lambda \left(W'W + (\sum_{j=1}^J W_j - 1)^2 \right)$.

given in (2.7) under the Imperfect Pre-Treatment Fit (IPTF) setting. Finally, Section 3.3 describes the approach for obtaining valid inference for individual treatment effects.

Assumption 3. *1. T_1 is fixed and $T_0 \to \infty$*

2. *$\{\varepsilon_{0t}, \varepsilon_t'\}_{t\in\mathcal{T}_0\cup\mathcal{T}_1}$ is a stationary and α-mixing process,*[7] *with mixing-coefficient $\alpha(\tau) \lesssim \tau^{-p}$ for some $2 < p < \infty$*

3. *$\max_{j\in\{0,\ldots,J+1\}, t\in\mathcal{T}_0\cup\mathcal{T}_1} \left(\mathbb{E}||\lambda_t||^2 + \mathbb{E}\delta_t^2 + \mathbb{E}|\varepsilon_{jt}|^{2+p}\right) \lesssim 1$*

4. *$\frac{1}{T_0}\sum_{t\in\mathcal{T}_0} \varepsilon_t\varepsilon_t' \xrightarrow{p} \Sigma$, $\frac{1}{T_0}\sum_{t=1}^{T_0}\lambda_t \xrightarrow{p} 0$ and $\frac{1}{r}\sum_{s=1}^{r} \overline{\lambda}^s(\overline{\lambda}^s)' \xrightarrow{p} \Omega_0$, where Σ and Ω_0 are positive semi-definite matrices*

5. *$\Delta \to \infty$ and k can be fixed or diverging*

Assumption 3.2 allows for errors to be dependent. Note that the mixing-coefficient is weaker than the exponential speed. Assumption 3.3 requires higher moments of the stochastic processes $(\lambda_t, \delta_t, \varepsilon_t)$ to be bounded; this could hold for instance if the random variables were sub-exponential.[8] Assumption 3.4 is similar to Ferman and Pinto (2021)[Assumption 4], except we do not require $\frac{1}{T_0}\sum_{t\in\mathcal{T}_1}\lambda_t\varepsilon_t = o_p(1)$ to derive our asymptotic results – these assumptions could be satisfied under stronger conditions such as independence of λ_t and ε_t, and λ_t is α-mixing with exponential speed; instead, assumption 3.3 only implies that this term is $O_p(1)$, which is a weaker requirement.[9]

We have chosen to pick weights defined over the space of Δ_η^J in (2.7), which differs slightly from the usual SCM in two ways: the usual SCM has the constraint that (1) $\sum_{j=1}^{J+1} W_j = 1$ and (2) $W_j \geq 0$ for $j = 1, \ldots, J+1$. The first constraint is called the "adding-up" constraint, while the second is the "non-negativity" constraint. These constraints retain interpret-ability of the SCM weights (see Abadie et al. (2010)). However, if the unit of interest is an outlier relative to the control-units, then the adding-up constraint may fail. For instance, Abadie et al. (2010) studied the effect of the tobacco control program in California[10] by using per capita smoking as outcome. However, if the chosen treated unit is heavily dependent on tobacco, then this constraint might fail. Next, the non-negativity constraint ensures the existence of a unique solution in the minimization problem, as well as reduces the deviation of the estimated weights to the true weights by limiting the sum of squared weights entering into the variance under estimation. This often ensures that the weights are non-zero only for a small subset of the control units which makes the weights easier to interpret. In many cases raw-correlations between the treated and control-units are positive;

[7] We can and will assume without loss of generality that by stationarity, $\mathbb{E}\varepsilon_{jt} = 0$ for every $t \in \mathcal{T}_0 \cup \mathcal{T}_1$ and $j = \{0, \ldots, J+1\}$.

[8] See Theorem 2.13 of Wainwright (2019)

[9] Suppose $\frac{1}{T_0}\sum_{t=1}^{T_0}\lambda_t \xrightarrow{p} w_0 \neq 0$. Then we can consider an observably equivalent model with $w_0 = 0$ by adjusting c_j for each $j = 0, \ldots, J+1$

[10] This program was called proposition 99

however, this does not mean that the correlation between the treated and every control-unit must be non-negative. Therefore, allowing for negative weights can improve out-of-sample predictions.

3.1 Perfect Pre-Treatment Fit

We begin by defining $Y_{jt} := y_{jt} - y_{J+1,t}$ for $j = 0, 1, ..., J$ so that we rewrite (2.1) as

$$\begin{aligned} Y_{jt}^N &= C_j + \lambda_t' M_j + u_{jt} \\ Y_{jt}^I &= \alpha_{jt} + Y_{jt}^N \end{aligned} \tag{3.1}$$

where $C_j := c_j - c_{J+1} = c_j$,[11] $M_j := \mu_j - \mu_{J+1}$ and $u_{jt} := \varepsilon_{jt} - \varepsilon_{J+1,t}$. We denote $C := (C_1, ..., C_J)'$ and $M := (M_1, ..., M_J)'$ as the $J \times 1$ vector and $J \times F$ matrix. Furthermore, define $\Delta_\eta^J := \{W \in \mathbb{R}^J : ||W|| \leq \eta\}$ for some given $\eta > 0$. Furthermore, define the Block synthetic control method weight (BSCM) as

$$\widetilde{W}_T^{BSCM}(f, \Lambda) := \underset{W \in \Delta_\eta^J}{\arg\min} \left\{ \frac{1}{k} \sum_{s=1}^{k} \left\{ \overline{Y}_0^s - (\overline{Y}^s)' W) \right\}^2 + f(W, \Lambda) \right\} \tag{3.2}$$

where f satisfies conditions 1–4 and $\overline{Y}_0^s$ and $\overline{Y}^s$ is defined similar to $\overline{y}_0^s$ and $\overline{y}^s$, given in section 2.2. Note that if we restrict W to non-negative elements, set $k = T_0$, $\Lambda \equiv 0$ and $\eta = 1$, (3.2) reduces to the SCM estimator given in (2.2).

Definition 3.1. *We say that we have perfect pre-treatment fit whenever there exists some weight* W^* *and* $\eta > 0$ *such that* $W^* \in \Phi := \{W \in \Delta_\eta^J : W'C = C_0 \text{ and } W'M = M_0\}$

The term "perfect pre-treatment fit" (or PPTF as we have denoted it) usually refers to the case where some linear combination of controls units (in this case the $j = 1, ..., J+1$ units) can perfectly recover the time-invariant fixed effect of the treated individual c_0 and its unknown factor-loadings μ_0 in (2.1). Formally, the conventional definition of PPTF is the following: that there exists some $W^\dagger \in \widetilde{\Phi} := \{W \in \widetilde{\Delta}_1 : W'c = c_0 \ \ \textit{and} \ \ W'\mu = \mu_0\}$, where $\widetilde{\Delta}_1 := \{W \in \mathbb{R}^{J+1} : \sum_{j=1}^{J+1} W_j = 1 \ \textit{and} \ W_i \geq 0 \ \textit{for} \ i = 1, ..., J+1\}$ is the simplex.[12] This is slightly different from definition 3.1. However, whenever we have PPTF in the usual sense, we will also have PPTF under definition 3.1. Formally we have the following:

Lemma 3.1. *Suppose there exists some* $W^\dagger \in \widetilde{\Phi}$. *Then there exists some* $\eta > 0$ *and some* $W^* \in \Phi$. *In particular, we can let* $\eta = 1$.

Definition 3.1 can therefore be seen as a more general version of the usual PPTF. We will work with this more general definition throughout this paper as it allows for extrapolation out of the

[11] Recall that we can and will assume that $c_{J+1} = 0$ without loss of generality

[12] Note that $W^* \in \mathbb{R}^J$ while $W^\dagger \in \mathbb{R}^{J+1}$

usual simplex, possibly enabling better pre-treatment fit and ultimately out-of-sample prediction. We have an alternative characterization of PPTF: define

$$G_0 := \begin{pmatrix} C_0 \\ M_0 \end{pmatrix} \quad and \quad G := \begin{pmatrix} C' \\ M \end{pmatrix}$$

Then we will have PPTF whenever there exists a $W^* \in \Phi$ such that $GW^* = G_0$. If G has full column-rank, $W^* \equiv (G'G)^{-1}G'(C_0, M_0')'$ is the solution to $GW = G_0$. In general, we can choose η to be large enough to ensure that at least one solution falls within the prescribed Δ_η^J. The only time PPTF does not hold is exactly when G_0 cannot be written as a linear combination of G; for instance, when the number of control-units J is smaller than the $1+F$ features coming from $(C_j, M_j')' \in \mathbb{R}^{1+F}$.[13] In this case, our proposed SCM weights $\widetilde{W}_T^{BSCM}(f,\Lambda)$ defined in (3.2) converges to weight W^{min} such that W^{min} minimizes the distance between GW and G_0 subject to some weighing matrix Ω_0 defined below. This can be interpreted as our method trying to search for weights that best approximate G_0 by a linear combination of G as best as possible. Formally, we have the following result.

Theorem 1 (Uniform approximation of BSCM estimator)**.** *Suppose assumptions 1, 2 and 3 holds. Then for any fixed $\gamma > 0$ and any f that satisfies conditions (1)-(4), we have*

$$\sup_{\Lambda\in[\gamma,1]} |\widetilde{W}_T^{BSCM}(f,\Lambda) - \overline{W}(\Lambda)| = o_p(1)$$

where $\overline{W}(\Lambda) := \arg\min_{W\in\Delta_\eta^J} \mathcal{A}(f,W,\Lambda)$ and $\mathcal{A}(f,W,\Lambda) := (C_0 - C'W)^2 + (M_0 - MW)'\Omega_0(M_0 - MW) + f(W,\Lambda)$; moreover, we can take γ to be zero whenever $(CC' + M'\Omega_0 M)$ is positive-definite. As $0 < \Lambda \downarrow 0$,

$$\overline{W}(\Lambda) \to \overline{W}(0)$$

where $\overline{W}(0) \in \min\arg\min_{W\in\Delta_\eta^J}(GW - G_0)'\widetilde{V}(GW - G_0)$, $\widetilde{V} := diag(1,\Omega_0)$ is a weighing matrix and $\min\arg\min(\cdot)$ is the minimum-norm vector in the space of $\arg\min(\cdot)$, where $\arg\min(\cdot)$ is the argument that minimizes $(\cdot)$.

Remark 1. *The reason we consider $\Lambda \downarrow 0$ instead of setting $\Lambda = 0$ and solving it directly is due to the fact that the objective function $\mathcal{A}(f,W,0)$ may not be strictly convex (by Ω_0 only being positive semi-definite), and therefore not unique. The solution set in this case could have cardinality greater than two, which renders the Newey and McFadden (1994) approach redundant. If Ω_0 is positive semi-definite and so $(GW - G_0)'\widetilde{V}(GW - G_0)$ does not have a unique solution, we could in principle set $\Lambda = 0$ to obtain some solution set for $\arg\min_{W\in\Delta_\eta^J} \frac{1}{r}\sum_{s=1}^{r}\{\overline{y}_0^s - (\overline{y}^s)'W)\}^2$, and then pick any*

[13]When the number of control-units diverge as in Ferman (2021), so that $J >> 1+F$, we can expect G to have full-column rank, so that PPTF exists.

of these solution, say $\check{W}$. However, this is not the most efficient choice: by recalling (2.4) and (2.5), the asymptotic variance of $\widehat{\alpha}_{0t}^{SCM}(\check{W})$ is at least as large as $\widehat{\alpha}_{0t}^{SCM}(\overline{W}(0))$, i.e.

$$\begin{aligned} &avar(\widehat{\alpha}_{0t}^{SCM}(\check{W})) = Q_0(\check{W}) = \mathcal{A}(f,\check{W},0) + \sigma_\varepsilon^2(1+\check{W}'\check{W}) \\ &\geq \mathcal{A}(f,\overline{W}(0),0) + \sigma_\varepsilon^2(1+\overline{W}(0)'\overline{W}(0)) = Q_0(\overline{W}(0)) = avar(\widehat{\alpha}_{0t}^{SCM}(\overline{W}(0))) \end{aligned} \tag{3.3}$$

if we assume that errors are homoskedastic (i.e. $\Sigma = \sigma_\varepsilon^2 I_J$)[14]*, where we recall that $\check{W} \in \arg\min \mathcal{A}(f,W,0)$ and $\overline{W}(0) \in \min\arg\min \mathcal{A}(f,W,0)$.*

The difficulty in considering $\gamma = 0$ stems from the assumption that $\Omega_0 \equiv Plim_{T_0\to\infty}\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\lambda_t\lambda_t'$ is only positive semi-definite, which implies that the solution set of the probability limit of $\widetilde{W}_T^{SC}(0)$ may not be unique. The implication is that $\widetilde{W}_T^{SC}(0)$ can potentially converge in probability to any W that solves $\Omega_0(M_0 - MW) = 0$, due to this lack of unique identification. To be precise, note that for any $\Lambda > 0$, we can express

$$\overline{W}(\Lambda) = (\Lambda I + cc' + \mu'\Omega_0\mu)^{-1}(\mu'\Omega_0\mu_0 + c_0 c).$$

As $\Lambda \downarrow 0$, the term $(\Lambda I + cc' + \mu'\Omega_0\mu)^{-1}$ could potentially diverge to infinity. This prevents $\overline{W}(\Lambda)$ from being equi-continuous in $\Lambda \in (0,1]$; consequently we are not able select a finite number of points $\Lambda_i \in (0,1]$ such that the union of balls around Λ_i covers the interval $(0,1]$ and the probability that any $\Lambda \in (0,1]$ is covered by one of the balls Lipschitz continuous. If we instead strengthen Ω_0 in assumption 3 to being positive-definite instead, then Theorem 1 implies that for **any** sequence of $0 \leq \Lambda_T \downarrow 0$,

$$\widetilde{W}_T^{BSCM}(f,\Lambda_T) \xrightarrow{p} \overline{W}(0)$$

An application of Theorem 4 below yields exact asymptotic size-control under the correct null.

Under PPTF, $\overline{W}(0)$ will be the pre-treatment weights that we seek. Note that $\overline{W}(0)$ is unique by Lemma A.1, even when Ω_0 is only positive semi-definite. In general we can ensure that $\overline{W}(0) \in \Phi$ whenever $\arg\min_{W\in\Delta_\eta^J}(GW - G_0)'\widetilde{V}(GW - G_0)$ has a single solution. A sufficient condition is for Ω_0 to be positive-definite. This usually occurs whenever $\{\lambda_t\}_{t\in\mathcal{T}_0\cup\mathcal{T}_1}$ is a non-stationary process.

Example 1. *consider*

$$\{\lambda_1,\lambda_2,...,\lambda_9\} = \{(1,0)',(1,0)',(0,1)',(0,1)',(1,1)',(1,1)',(1,1)',(1,1)',(1,1)'\}$$

[14] The homoskedastic assumption is made for simplification of argument; it can be shown with some algebra that the inequality between the left and right-hand-side of (3.3) still holds under general heteroskedastic errors.

and this time-dependent common factor repeats itself. Define

$$\omega_0 := \lim_{T_0 \to \infty} \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \lambda_t = \frac{2}{9}(1,0)' + \frac{2}{9}(0,1)' + \frac{5}{9}(1,1)' = (0.777, 0.777)'$$

so that defining $\widetilde{\lambda}_t := \lambda_t - \omega_0$*, we have*

$$\lim_{T_0 \to \infty} \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \widetilde{\lambda}_t = 0,$$

hence satisfying assumption 3. Consider a 3*-fold cross fitting. Some algebraic manipulation yields*

$$\overline{\lambda}^1 = (0.222, -0.777)', \qquad \overline{\lambda}^2 = (-0.777, 0.222)' \quad \textit{and} \quad \overline{\lambda}^3 = (0.222, 0.222)'$$

where $\overline{\lambda}_i := \frac{1}{\Delta_i} \sum_{t \in \Delta_i} \widetilde{\lambda}_t$ *for* $i = 1, 2, 3$*. Then*

$$\frac{1}{k} \sum_{s=1}^{k} (\overline{\lambda}^s)(\overline{\lambda}^s)' \to \begin{pmatrix} 0.567 & -0.432 \\ -0.432 & 0.567 \end{pmatrix},$$

which is positive-definite. □

A direct implication of Theorem 1 is the following:

Corollary 3.1. *Suppose assumption 1, 2 and 3 holds. Further assume that* f *satisfies conditions* (1)-(4)*. Then for any* $\xi > 0$*, there exists a* $\Lambda(\xi) > 0$ *such that for any fixed* $0 < \Lambda \leq \Lambda(\xi)$*,*

$$|\widetilde{W}^{BSCM}(f, \Lambda) - \overline{W}(0)| \leq \xi + o_p(1)$$

Corollary 3.1 assures us that we can obtain as close an approximation to $\overline{W}(0)$. Therefore there exists some sequence of Λ_T such that $\widetilde{W}_T^{SC}(\Lambda_T)$ consistently estimates $\overline{W}(0)$. This is formalized below.

Corollary 3.2. *Suppose assumption 1, 2, 3 holds and* f *satisfies conditions* (1)-(4)*. Then there exists a sequence* $0 < \Lambda_T \downarrow 0$ *such that*

$$\widetilde{W}_T^{BSCM}(f, \Lambda_T) = \overline{W}(0) + o_p(1)$$

If Ω_0 *is positive-definite, then any sequence of* $0 < \Lambda_T \downarrow 0$ *will satisfy the preceding equation.*

The block synthetic control-based weights $\widetilde{W}_T^{BSCM}(f, \Lambda_T)$ therefore provides a way to recover the pre-treatment weights whenever such weights exist within or even outside the simplex. In general, without making more assumptions it is impossible to obtain the sequence of penalty terms Λ_T given in corollary 3.2. Therefore, in applications, corollary 3.1 is more useful, and we simply choose an arbitrarily small $\Lambda > 0$.

3.2 Imperfect Pre-Treatment Fit

In the previous section, under perfect pre-treatment fit, we can "almost recover" the weights used to conduct unbiased inference. In this section we discuss the implications of our estimator under imperfect pre-treatment fit (IPTF), i.e. when $\nexists W^* \in \Phi$.

3.2.1 IPTF under Stationarity

We consider the case when $\{\lambda_t\}_{t\in\mathcal{T}_0\cup\mathcal{T}_1}$ is a stationary process, so that Ω_0 will generally be a positive semi-definite matrix. In such a case we want our estimator to be at least efficient in the following sense: consider any fixed $t \in \mathcal{T}_1$ and let $\overline{W}(0)$ be the limit of $\widetilde{W}_T^{BSCM}(f, \Lambda_T)$ for some sequence of $0 < \Lambda_T \downarrow 0$, so that we recall from section 3.1 that

$$\widehat{\alpha}_{0t}^{SCM}(\widetilde{W}_T^{BSCM}(f, \Lambda_T)) \xrightarrow{p} \alpha_{0t} + (C_0 - C'\overline{W}(0)) + \lambda_t'(M_0 - M\overline{W}(0)) + (u_{0t} - u_t'\overline{W}(0)). \quad (3.4)$$

where $\widehat{\alpha}_{0t}^{SCM}(\cdot)$ was defined in (2.5). If Ω_0 is positive-definite, we will have by Theorem 1 that $\overline{W}(0) \in \Phi$, implying that the first term $C_0 - C'\overline{W}(0) = 0$; therefore $\widehat{\alpha}_{0t}^{SCM}(\widetilde{W}_T^{BSCM}(f, \Lambda_T))$ is an unbiased estimator for α_{0t} whenever $\mathbb{E}[\lambda_t] = 0$, as is the case under a stationary λ_t process. However, if instead Ω_0 is only positive-semi definite, the weights obtained from our BSCM procedure may not cancel out the first term (i.e. $C_0 \neq C'\overline{W}(0)$), making $\widehat{\alpha}_{0t}^{SCM}(\widetilde{W}_T^{BSCM}(f, \Lambda_T))$ a biased estimate. To overcome this, we can demean our BSCM and construct a "demeaned-version" of the BSCM defined as the DBSCM-based weights given in (2.7). Note that by Lemma A.1 this weight is uniquely defined. Then we have the following result:

Theorem 2 (Uniform approximation of DBSCM estimator)**.** *Suppose assumption 1, 2 and 3 holds. Then for any fixed $\gamma > 0$ and any f that satisfies conditions (1)-(4), we have*

$$\sup_{\Lambda\in[\gamma,1]} \left|\widetilde{W}_T^{DBSCM}(f, \Lambda) - \overline{W}^0(\Lambda)\right| = o_p(1)$$

where $\overline{W}^0(\Lambda) := \arg\min_{W\in\Delta_\eta^J} \mathcal{B}(f, W, \Lambda)$ and $\mathcal{B}(f, W, \Lambda) := (M_0 - MW)'\Omega_0(M_0 - MW) + f(W, \Lambda)$; moreover, we can take γ to be zero whenever Ω_0 is a positive-definite matrix. As $0 < \Lambda \downarrow 0$,

$$\overline{W}^0(\Lambda) \to \overline{W}^0(0)$$

where $\overline{W}^0(0) \in \min \arg\min_{W\in\Delta_\eta^J}(M_0 - MW)'\Omega_0(M_0 - MW)$

As a result of the preceding theorem, we can obtain a more efficient estimator of α_{0t} than using $\widehat{\alpha}_{0t}^{SCM}(\cdot)$ defined in (2.5). To this end, define the demeaned synthetic control-based estimator (which we denote as DSCM) for α_{0t} as

$$\widehat{\alpha}_{0t}^{DSCM}(W) := Y_{0t} - Y_t'W - \left(\overline{Y}_0 - \overline{Y}'W\right) \tag{3.5}$$

for any $t \in \mathcal{T}_0 \cup \mathcal{T}_1$. Then our demeaned block synthetic control method estimator (DBSCM) is given as

$$\widehat{\alpha}_{0t}^{DBSCM}(f,\Lambda) := \widehat{\alpha}_{0t}^{DSCM}\left(\widetilde{W}_T^{DBSCM}(f,\Lambda)\right) \xrightarrow{p} \alpha_{0t} + \lambda_t'(M_0 - M\overline{W}^0(\Lambda)) + (u_{0t} - u_t'\overline{W}^0(\Lambda)) \tag{3.6}$$

as $\Lambda_T \downarrow 0$. In this case, $\widehat{\alpha}_{0t}^{DBSCM}(f,\Lambda_T)$ is an unbiased estimator for α_{0t} whenever $\mathbb{E}[\lambda_t] = 0$, despite the existence of an imperfect pre-treatment fit. Furthermore, it is clear that for any $W \in \mathbb{R}^J$, $avar(\widehat{\alpha}_{0t}^{SCM}(W)) = avar(\widehat{\alpha}_{0t}^{DSCM}(W))$ by simply observing that $\widehat{\alpha}_{0t}^{SCM}(W)$ has an extra non-random term $(C_0 - C'W)$ in the limit (compare (3.4) and (3.6)). Therefore, by using $\widehat{\alpha}_{0t}^{DBSCM}(f,\Lambda_T)$ instead of $\widehat{\alpha}_{0t}^{SCM}(\widetilde{W}_T(f,\Lambda_T))$, we can retain efficiency while obtaining an unbiased estimator.

Remark 2. *As in Remark 1, it is more efficient to pick* $\overline{W}^0(0)$ *by letting* $\overline{W}^0(\Lambda) \to \overline{W}^0(0)$ *as* $0 < \Lambda \downarrow 0$ *instead of picking some arbitrary* $\check{W} \in \arg\min_{W \in \Delta_\eta^J} \frac{1}{r}\sum_{s=1}^r \{\overline{y}_0^s - (\overline{y}^s)'W - (\overline{y}_0 - \overline{y}'W)\}^2$. *To see this, simply observe that*

$$\begin{aligned} avar(\widehat{\alpha}_{0t}^{DSCM}(\check{W})) &= \mathcal{B}(f,\check{W},0) + \sigma_\varepsilon^2(1 + \check{W}'\check{W}) \\ &\geq \mathcal{B}(f,\overline{W}^0(0),0) + \sigma_\varepsilon^2(1 + \overline{W}^0(0)'\overline{W}^0(0)) = avar(\widehat{\alpha}_{0t}^{DBSCM}) \end{aligned} \tag{3.7}$$

if we assume that errors are homoskedastic (i.e. $\Sigma = \sigma_\varepsilon^2 I_J$*)*[15]*, where we recall that* $\check{W} \in \arg\min \mathcal{B}(f,W,0)$ *and* $\overline{W}^0(0) \in \min\arg\min \mathcal{B}(f,W,0)$.

3.2.2 IPTF under Non-Stationarity and Efficiency of Existing Estimators

In this section we will discuss the implication of existing estimators under Imperfect Pre-Treatment Fit without the assumed stationarity of λ_t. We will also explore the efficiency and the bias of our proposed estimator together with existing estimators. To this end, note that the difference-in-difference (DID) estimator can be written as

$$\widehat{\alpha}_{0t}^{DID} \equiv \widehat{\alpha}_{0t}^{DSCM}\left(\frac{1}{J}\iota\right)$$

[15]The homoskedastic assumption is made for simplification of argument; it can be shown with some algebra that the inequality between the left-hand-side and right-hand-side in (3.7) still holds under general heteroskedastic errors.

where $\iota = (1, ..., 1)' \in \mathbb{R}^J$ with $\widehat{\alpha}_{0t}^{DSCM}(\cdot)$ is defined in (3.5). In general, one can always consider the estimator $\widehat{\alpha}_{0t}^{DSCM}(W)$ defined in (2.5) for any given $W \in \mathbb{R}^J$ so that the DID-estimator is just a special case. Ferman and Pinto (2021)[Proposition 3] showed that their FP-based estimator $\widehat{\alpha}_{0t}^{FP}$ is more efficient (in the sense of weakly smaller asymptotic variance) than the $\widehat{\alpha}_{0t}^{DSCM}(\cdot)$ estimator, where

$$\widehat{\alpha}_{0t}^{FP} := \widehat{\alpha}_{0t}^{DSCM}\left(\widehat{W}^{FP}\right)$$

with $\widehat{W}^{FP}$ defined in (2.8). We will begin by showing that the DBSCM estimator $\widehat{\alpha}_{0t}^{DBSCM}(f, \Lambda)$ defined in (3.6) is more efficient than the FP-based estimator for some penalty term $f(W, \Lambda)$ and parameter $\Lambda \geq 0$, and therefore also more efficient than the DID-based estimator under homoskedasticity or error terms and stationarity of λ_t. We begin with an assumption. which is consistent to Assumption 5 of Ferman and Pinto (2021), called the 'Stability in the pre- and post-treatment period' assumption.

Assumption 4. *For $t \in \mathcal{T}_1$, $cov(\lambda_t, (\varepsilon_{0t}, \varepsilon_t')') = 0$, $\mathbb{E}\lambda_t\lambda_t' = \Omega_0$ and $\mathbb{E}(\varepsilon_{0t}, \varepsilon_t')(\varepsilon_{0t}, \varepsilon_t')' = \sigma_\varepsilon^2 I_{J+1}$*

We have the following result:

Theorem 3 (Efficiency of estimator)**.** *Under Assumptions 1–4 and $\mathbb{E}[\lambda_t] = 0$, if we define $\overline{f}(\Lambda, W) := \Lambda(1 + W'W + (W'\iota - 1)^2)$ and let $\eta \geq 1$, then the demeaned-block-synthetic-control method estimator $\widehat{\alpha}_{0t}^{DBSCM}(\overline{f}, \sigma_\varepsilon^2)$ defined in (3.6) dominates both $\widehat{\alpha}_{0t}^{FP}$ and $\widehat{\alpha}_{0t}^{DID}$ in terms of asymptotic mean-squared-error (MSE).*

Note that the assumption of λ_t being is a stationary process automatically implies $\mathbb{E}[\lambda_t] = 0$ and $\mathbb{E}\lambda_t\lambda_t' = \Omega_0$ for every $t \in \mathcal{T}_0 \cup \mathcal{T}_1$. Theorem 3 tells us that under such stationarity, there is a special case for which our estimator has weakly smaller MSE. Furthermore, note that

$$\widehat{\alpha}_{0t}^{FP} \xrightarrow{p} \tau_{0t}\left(W_{-(J+1)}^{FP}\right) \qquad \text{and} \qquad \widehat{\alpha}_{0t}^{DBSCM}(f, \Lambda) \xrightarrow{p} \tau_{0t}\left(\overline{W}^0(\Lambda)\right)$$

where $W^{FP} = (W_{-(J+1)}^{FP}, W_{J+1}^{FP}) = Plim\widehat{W}^{FP}$ with $W_{-(J+1)}^{FP}$ denoting the first J terms (and omitting the $J+1$-th term) and

$$\tau_{0t}(W) := \alpha_{0t} + \lambda_t'(M_0 - MW) + \left(\varepsilon_{0t} + \sum_{j=1}^{J} W_j\varepsilon_{jt} - (1 - \sum_{j=1}^{J} W_j)\varepsilon_{J+1,t}\right) \tag{3.8}$$

When under non-stationarity of λ_t (so that $\mathbb{E}\lambda_t \neq 0$), then $\widehat{\alpha}_{0t}^{FP}$ and $\widehat{\alpha}_{0t}^{DBSCM}(f, \Lambda)$ is biased for any fixed $\Lambda > 0$ unless their respective weights cancel out the second term on the right-side of $\tau_{0t}(W)$ (i.e. unbiased only if $M_0 - MW = 0$). Note that the FP weights $\widehat{W}^{FP}$ come from

$$\underset{W \in \mathbb{R}^J : \Delta_1^J \cap W_i \geq 0}{\arg\min} \quad \mathcal{B}(\overline{f}, W, \sigma_\varepsilon^2) \tag{3.9}$$

where $\overline{f}(\cdot)$ is defined as in Theorem 3 and $\mathcal{B}(\cdot,\cdot,\cdot)$ is defined in Theorem 2, under homoskedastic error[16] so that in general the weights W^{FP} do not solve $MW = M_0$ even when such weights exist (i.e. $\exists W^* \in \Phi$), due to the fact that σ_ε^2 is present in the minimization problem of (3.9), leading $\widehat{\alpha}_{0t}^{FP}$ to be a biased estimator for α_{0t}; to see this, observe that WPA1,

$$\mathbb{E}[\widehat{\alpha}_{0t}^{FP}] = \mathbb{E}[\tau_{0t}(W_{-(J+1)}^{FP})] = \alpha_{0t} + \mathbb{E}[\lambda_t]'(M_0 - MW_{-(J+1)}^{FP}) \neq \alpha_{0t}$$

In spite of this, the non-stationarity of λ_t allows us to exploit the positive-definiteness of Ω_0, which by Theorem 2 allows us to recover $\overline{W}^0(0)$ whenever PPTF holds (i.e. $M_0 - M\overline{W}^0(0) = 0$). This implies that $\widehat{\alpha}_{0t}^{DBSCM}(f, \Lambda_T)$ is an unbiased estimator of α_{0t} under non-stationarity for any sequence of $\Lambda_T \downarrow 0$. It is also clear that even under stationarity, $\widehat{\alpha}_{0t}^{DBSCM}(f, \Lambda_T)$ remains an unbiased estimator (by $\mathbb{E}\lambda_t = 0$ and (3.6)). Fornally, we have the following.

Corollary 3.3. *Under assumptions 1–3 and assuming f satisfies 1–4, if either (i) λ_t is stationary or (ii) λ_t is non-stationary, Ω_0 is positive-definite and there exists some $\overline{W} \in \Phi$ (i.e. there exists some perfect pre-treatment fit), then $\widehat{\alpha}_{0t}^{DBSCM}(f, \Lambda_T)$ is an unbiased estimator of α_{0t} for any sequence of* $0 < \Lambda_T \downarrow 0$

So far we we have seen that our estimator $\widehat{\alpha}_{0t}^{DBSCM}(f, \Lambda_T)$ is an unbiased estimator for α_{0t} whenever we have either (i) Perfect Pre-Treatment Fit or (i) Imperfect Pre-Treatment Fit with $\{\lambda_t\}_{t\in\mathcal{T}_0\cup\mathcal{T}_1}$ being a stationary process. However, when we have Imperfect Pre-Treatment Fit under non-stationary λ_t, then it is not possible to obtain an unbiased estimator. Intuitively, this is because unbiasedness requires the existence of some W^* such that the term $\mathbb{E}\left[\lambda_t'\right](M_0 - MW^*) = 0$ of (3.8); this is satisfied if either $\mathbb{E}[\lambda]_t = 0$ or $M_0 - MW^* = 0$. Under Imperfect Fit, $M_0 - MW^* \neq 0$ for any W^*, yet the lack of stationarity property from λ_t prevents us from estimating $\mathbb{E}[\lambda_t]$ for any fixed $t \in \mathcal{T}_1$ since we cannot "learn" from previous observations in $\mathcal{T}_0$. Despite this impossibility, we have shown that (A) existing estimators may be biased under the setting of (ii) in Corollary 3.3 and (B) our estimator is at least as efficient as existing estimators, meriting the use of the DBSCM estimator over other existing estimators.

Recall from Theorem 3 that the DBSCM estimator is efficient under some penalty term and fixed $\Lambda \equiv \sigma_\varepsilon^2$, whenever we have homoskedastic error and stationarity of λ_t. However, note that in general, for any f satisfying conditions 1–4,

$$\begin{aligned} avar\left(\widehat{\alpha}_{0t}^{DBSCM}(f, \Lambda_T))\right) &= \mathcal{B}(f, \overline{W}^0(0), 0) + \sigma_\varepsilon^2(1 + \overline{W}^0(0)'\overline{W}^0(0)) \\ &\geq \mathcal{B}(f, \overline{W}^0(\sigma_\varepsilon^2), 0) + \sigma_\varepsilon^2(1 + \overline{W}^0(\sigma_\varepsilon^2)'\overline{W}^0(\sigma_\varepsilon^2)) = avar\left(\widehat{\alpha}_{0t}^{DBSCM}(\overline{f}, \sigma_\varepsilon^2))\right) \end{aligned}$$

so that $\widehat{\alpha}_{0t}^{DBSCM}(f, \Lambda_T)$ may not necessarily be more efficient than either $\widehat{\alpha}_{0t}^{FP}$ or $\widehat{\alpha}_{0t}^{DID}$. There is

[16] The assumption of homoskedastic error is made in order to simplify the arguments. Under heteroskedastic error the argument will still hold

therefore a bias-variance tradeoff in that the "cost" to obtain an unbiased estimator $\widehat{\alpha}_{0t}^{DBSCM}(f, \Lambda_T)$ is an increase in variance. We summarize this in Table 3.

	Asymptotic Mean	Asymptotic Variance
$\widehat{\alpha}^{DBSCM}(f, \Lambda_T)$		
λ_t stationary, PPTF	α_{0t}	$\sigma_\varepsilon^2(1 + \overline{W}^0(0)'\overline{W}^0(0))$
λ_t stationary, IPTF	α_{0t}	$\sigma_\varepsilon^2(1 + \overline{W}^0(0)'\overline{W}^0(0)) + \mathcal{B}(\overline{W}^0(0))$
λ_t non-stationary, PPTF	α_{0t}	$\sigma_\varepsilon^2(1 + \overline{W}^0(0)'\overline{W}^0(0))$
λ_t non-stationary, IPTF	$\alpha_{0t} + \mathbb{E}[\lambda_t]'\left(M_0 - M\overline{W}^0(0)\right)$	$\sigma_\varepsilon^2(1 + \overline{W}^0(0)'\overline{W}^0(0)) + \mathcal{B}(\overline{W}^0(0))$
$\widehat{\alpha}^{DBSCM}(\overline{f}, \sigma_\varepsilon^2)$		
λ_t stationary, PPTF	α_{0t}	$\sigma_\varepsilon^2(1 + \overline{W}^0(\sigma_\varepsilon^2)'\overline{W}^0(\sigma_\varepsilon^2)) + \mathcal{B}(\overline{W}^0(\sigma_\varepsilon^2))$
λ_t stationary, IPTF	α_{0t}	$\sigma_\varepsilon^2(1 + \overline{W}^0(\sigma_\varepsilon^2)'\overline{W}^0(\sigma_\varepsilon^2)) + \mathcal{B}(\overline{W}^0(\sigma_\varepsilon^2))$
λ_t non-stationary, PPTF	$\alpha_{0t} + \mathbb{E}[\lambda_t]'\left(M_0 - M\overline{W}^0(\sigma_\varepsilon^2)\right)$	$\sigma_\varepsilon^2(1 + \overline{W}^0(\sigma_\varepsilon^2)'\overline{W}^0(\sigma_\varepsilon^2)) + \mathcal{B}(\overline{W}^0(\sigma_\varepsilon^2))$
λ_t non-stationary, IPTF	$\alpha_{0t} + \mathbb{E}[\lambda_t]'\left(M_0 - M\overline{W}^0(\sigma_\varepsilon^2)\right)$	$\sigma_\varepsilon^2(1 + \overline{W}^0(\sigma_\varepsilon^2)'\overline{W}^0(\sigma_\varepsilon^2)) + \mathcal{B}(\overline{W}^0(\sigma_\varepsilon^2))$
$\widehat{\alpha}_{0t}^{FP}$		
λ_t stationary, PPTF	α_{0t}	$\sigma_\varepsilon^2(1 + (W^{FP})'W^{FP}) + \mathcal{B}(W^{FP})$
λ_t stationary, IPTF	α_{0t}	$\sigma_\varepsilon^2(1 + (W^{FP})'W^{FP}) + \mathcal{B}(W^{FP})$
λ_t non-stationary, PPTF	$\alpha_{0t} + \mathbb{E}[\lambda_t]'\left(M_0 - MW^{FP}(\sigma_\varepsilon^2)\right)$	$\sigma_\varepsilon^2(1 + (W^{FP})'W^{FP}) + \mathcal{B}(W^{FP})$
λ_t non-stationary, IPTF	$\alpha_{0t} + \mathbb{E}[\lambda_t]'\left(M_0 - MW^{FP}(\sigma_\varepsilon^2)\right)$	$\sigma_\varepsilon^2(1 + (W^{FP})'W^{FP}) + \mathcal{B}(W^{FP})$
$\widehat{\alpha}_{0t}^{DID}$		
λ_t stationary, PPTF	α_{0t}	$\sigma_\varepsilon^2(1 + (W^{DID})'W^{DID}) + \mathcal{B}(W^{DID})$
λ_t stationary, IPTF	α_{0t}	$\sigma_\varepsilon^2(1 + (W^{DID})'W^{DID}) + \mathcal{B}(W^{DID})$
λ_t non-stationary, PPTF	$\alpha_{0t} + \mathbb{E}[\lambda_t]'\left(M_0 - MW^{DID}(\sigma_\varepsilon^2)\right)$	$\sigma_\varepsilon^2(1 + (W^{DID})'W^{DID}) + \mathcal{B}(W^{DID})$
λ_t non-stationary, IPTF	$\alpha_{0t} + \mathbb{E}[\lambda_t]'\left(M_0 - MW^{DID}(\sigma_\varepsilon^2)\right)$	$\sigma_\varepsilon^2(1 + (W^{DID})'W^{DID}) + \mathcal{B}(W^{DID})$

Table 3: Summary of different tests. We write $\mathcal{B}(\cdot) \equiv \mathcal{B}(f, \cdot, 0)$

3.3 Inference for Treatment Effect

In the previous sections we discussed how we can obtain an unbiased estimator even under non-stationarity, as long as PPTF exists. In this section we discuss how to make inference on $\alpha_0 = \{\alpha_{0t}\}_{t\in\mathcal{T}_1}$. Unfortunately, if $T_1 := card(\mathcal{T}_1)$ does not diverge to infinity, then the usual method of applying some normalization to the test statistic in order to obtain a central-limit-theorem to conduct inference would fail. To overcome this, we appeal to conformal-inference.

Suppose first that we have an estimator $\widetilde{W}$ of some sort, such that $\widetilde{W} \xrightarrow{P} \overline{W}$, with $\overline{W} \in \Phi$. Then recall from (3.5) and (3.8) that as $T_0 \to \infty$, given some estimator $\widetilde{W}$, for any fixed $t \in \mathcal{T}_1$,

$$\widehat{\alpha}_{0t}^{DSCM}(\widetilde{W}) \xrightarrow{p} \alpha_{0t} + \lambda_t\left(M_0 - M'\overline{W}\right) + \left(\varepsilon_{0t} + \sum_{j=1}^{J}\overline{W}_j\varepsilon_{jt} - (1 - \sum_{j=1}^{J}\overline{W}_j)\varepsilon_{J+1,t}\right)$$

Following the notations of Chernozhukov, Wuthrich, and Zhu (2021), let $\mathcal{T}_0 = \{1, ..., T_0\}$ and $\mathcal{T}_1 = \{T_0+1, ..., T\}$. We define the moving block permutation for $m \in \{0, 1, ..., T-1\}$ as $\Pi := \{\pi_m\}_{m=0}^{T-1}$, where

$$\pi_m(i) = \begin{cases} i+m & \text{if } i+m \leq T \\ i+m-T & \text{otherwise} \end{cases}$$

For notational simplicity, define $\varepsilon_t := (\varepsilon_{1t}, ..., \varepsilon_{Jt})'$ and $y_t := (y_{1t}, ..., y_{Jt})'$. Then for any $t \in \mathcal{T}_0 \cup \mathcal{T}_1$, define for any $\widetilde{W}$

$$\widehat{P}_t^N(\widetilde{W}) := \widehat{\alpha}_{0t}^{DSCM}(\widetilde{W}) - \alpha_t =: -\widehat{v}_t(\widetilde{W})$$

where $\{\alpha_t\}_{t\in\mathcal{T}_1}$ is the hypothesized value of the treatment effect $\{\alpha_{0t}\}_{t\in\mathcal{T}_1}$ and $\widehat{\alpha}_{0t}^{DSCM}(\cdot)$ was defined in (3.5). By convention, we define $\alpha_t \equiv \alpha_{0t} \equiv 0$ for every $t \in \mathcal{T}_0$. The p-value is defined as

$$\widehat{p}(\widetilde{W}) := \frac{1}{|\Pi|} \sum_{\pi\in\Pi} \mathbb{1}\{S(\widehat{v}_\pi(\widetilde{W})) \geq S(\widehat{v}(\widetilde{W}))\} \tag{3.10}$$

where

$$S(\widehat{v}(\widetilde{W})) := \frac{\sum_{t=T_0+1}^{T} |\widehat{v}_t(\widetilde{W})|}{\sqrt{T_1}},$$
$$\widehat{v}(\widetilde{W}) = (\widehat{v}_1(\widetilde{W}), ..., \widehat{v}_T(\widetilde{W}))'$$

and $\widehat{v}_\pi(\widetilde{W}) = (\widehat{v}_{\pi(1)}(\widetilde{W}), ..., \widehat{v}_{\pi(T)}(\widetilde{W}))$ is the permutation of $\widehat{v}(\widetilde{W})$ by $\pi \in \Pi$. Then we have the following result.

Theorem 4. *Suppose $\widetilde{W} \xrightarrow{p} \overline{W}$ and assumptions 1–3 holds. If either*

1. *$\{\lambda_t\}_{t\in\mathcal{T}_0\cup\mathcal{T}_1}$ is stationary or*
2. *$\{\lambda_t\}_{t\in\mathcal{T}_0\cup\mathcal{T}_1}$ is non-stationary and $\overline{W} \in \Phi$*

holds, then under the correct null of $\{\alpha_t\}_{t\in\mathcal{T}_1} = \{\alpha_{0t}\}_{t\in\mathcal{T}_1}$, for any $\theta \in (0, 1)$,

$$\left|\mathbb{P}(\widehat{p}(\widetilde{W}) \leq \theta) - \theta\right| = o_p(1)$$

Theorem 2 and Corollary 3.3 gives us the immediate results, with states that the p-value based on the DBSCM-derived weights $\widetilde{W}_T^{DBSCM}(f, \Lambda_T)$ yield asymptotically valid results under the null.

Corollary 3.4. *Suppose assumptions 1–3 holds. If either*

1. $\{\lambda_t\}_{t \in \mathcal{T}_0 \cup \mathcal{T}_1}$ *is stationary or*

2. $\{\lambda_t\}_{t \in \mathcal{T}_0 \cup \mathcal{T}_1}$ *is non-stationary,* Ω_0 *is positive-definite and* $\exists W^* \in \Phi$*,*

then under the correct null of $\{\alpha_t\}_{t\in\mathcal{T}_1} = \{\alpha_{0t}\}_{t\in\mathcal{T}_1}$ *and any* f *satisfying conditions 1–4, for large enough* $\eta > 0$ *and any* $\theta \in (0,1)$*,*

$$\left|\mathbb{P}\left(\widehat{p}(\widetilde{W}_T^{DBSCM}(f, \Lambda_T)\right) \leq \theta) - \theta\right| = o_p(1)$$

for any sequence of $\Lambda_T \downarrow 0$*.*

Remark 3. *Under the conditions of Theorem 4.2, the DID and FP-based weights* $\widehat{W}^{FP}$ *and* $\widehat{W}^{DID}$ *defined in section 3.2.2 will generally not recover the true weights, i.e.* $\mathbb{P}(\widehat{p}(W) \leq \theta) \not\to \theta$ *for* $W \in \{\widehat{W}^{FP}, \widehat{W}^{DID}\}$*. We can see this from section 4, where there is some over-sizing.*

4 Simulation Study

In this section we provide simulation results for section 3.3, by comparing against existing tests in the literature. We consider the factor model of Chernozhukov et al. (2021)[section G] with minor changes in order to demonstrate the improvement of our test. The model can be written as

$$\begin{aligned}
&y_{jt}^N = c_j + \delta_t + \lambda_t' \mu_j + \varepsilon_{jt} \\
&c_j = (j+1)/J, \ \mu_j = \sqrt{(j+1)/J} \\
&\varepsilon_{jt} = \rho_\varepsilon \varepsilon_{j,t-1} + \xi_{jt}, \quad \xi_{jt} \overset{iid}{\sim} \mathcal{N}(0, 1-\rho_\varepsilon^2) \\
&\delta_t \overset{iid}{\sim} \frac{1}{\sqrt{|\mathcal{N}(t,1)|}}
\end{aligned}$$

The potential outcome is generated using

$$\begin{aligned}
&y_{0t} = \begin{cases} \sum_{j=1}^J W_j^{oracle} y_{jt}^N + e_t & \text{if } t > T_0 \\ \alpha_{0t} + \sum_{j=1}^J W_j^{oracle} y_{jt}^N + e_t & \text{if } t \leq T_0 \end{cases} \\
&e_t = \rho_e e_{t-1} + v_t, \quad v_t \overset{iid}{\sim} \mathcal{N}(0, 1-\rho_e^2)
\end{aligned}$$

We consider the following common-factors

$$\lambda_t \overset{iid}{\sim} \mathcal{N}\left(\frac{2t}{T}, 1\right)$$

where we recall that $T = T_0 + T_1$. We let $\rho_\varepsilon = 0.6$ and vary $\rho_e = \{0, 0.5, 0.9\}$. We let $J = 20$, $T_0 = 100$ and $T_1 = 1$ and consider the following four DGPs, which consists of both sparse and dense weights:

	Weight specification	Correctly specified model
DGP 1	$W^{oracle} = (\frac{1}{J}, ..., \frac{1}{J})'$	DID, SCM, Constrained Lasso, FP, DBSCM
DGP 2	$W^{oracle} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, ..., 0)'$	SCM, Constrained Lasso, FP, DBSCM
DGP 3	$W^{oracle} = -(\frac{1}{J}, ..., \frac{1}{J})'$	Constrained Lasso, DBSCM
DGP 4	$W^{oracle} = (\frac{1}{2}, -\frac{1}{2}, 0, ..., 0)'$	Constrained Lasso, DBSCM

We choose $\Lambda = 0.01$ and let $k = 3$. We consider and compare several existing weight estimators and apply them into (3.10):

1. Difference-in-difference estimator, where $\widetilde{W} = \widehat{W}^{DID} \equiv (1/J, ..., 1/J)'$
2. Ferman and Pinto (2021)'s weight estimator $\widetilde{W} \equiv \widehat{W}^{FP}$ as defined in (2.8)
3. Abadie et al. (2010)'s SCM weights where $\widetilde{W} \equiv \widehat{W}^{SCM}$, defined in (2.2)
4. Our proposed estimator $\widetilde{W} \equiv \widetilde{W}_T^{DBSCM}(f, \Lambda)$ defined in (2.7) with $f(W, \Lambda) \equiv \Lambda \|W\|_1$ and $\Lambda \equiv 0.01$. We set $\eta = 1$
5. Constrained LASSO, which we denote by $\widehat{W}^{CLASSO}$. We still apply (3.10) by defining

$$\widehat{v}_t(\widehat{W}^{CLASSO}) := y_{0t}^N - \left(\widehat{H} + \sum_{j=1}^{J} \widehat{W}_j^{CLASSO} y_{jt}^N \right),$$

where

$$y_{0t}^N = \begin{cases} y_{0t} & \text{if } t \leq T_0 \\ y_{0t} - \alpha_t & \text{if } t > T_0 \end{cases}$$

and

$$(\widehat{H}, \widehat{W}^{CLASSO}) = \underset{(H,W) \in \mathbb{R} \times \mathbb{R}^J}{\arg\min} \sum_{t=1}^{T} \left(y_{0t}^N - H - \sum_{j=1}^{J} W_j y_{jt} \right)^2 \quad s.t. \quad \|W\|_1 \leq 1$$

A few remarks are in order. First, by corollary 3.4 and remark 3, we expect that the DID and FP estimator would not be able to recover the oracle weights, leading to under or over-sizing. This is consistent with what we have seen in Table 6 under DGP2, which reveals that these two estimators lack size-control under non-stationarity of λ_t. In contrast, the DBSCM has some size-control. Second, the size for DGP1 in Table 6 is bad for all estimators, including the oracle estimator W^{DID}. When we consider either $\rho_e = 0.5$ or 0, this size issue goes away for all the estimators. Third, the DID, SCM and FP estimators do not do very well under DGP3 and DGP4 for any $\rho_e = \{0, 0.5, 0.9\}$. This is to be expected, since the weights are required to be in a simplex, while the oracle weights are out of the simplex, preventing these estimators from recovering the true weights. Finally, the constrained Lasso generally does quite well. However, when ρ_e is high as in Table 6, its size can be worse than the simplex weights under DGP3 or DGP4. The reason that the DBSCM does better is that the "blocking-process" helps remove the variance (e_t, ε_{jt}) when estimating its weights, similar to removing the σ_ε^2 in (2.4) when estimating weights, as explained in (2.6).

Table 4: $\rho_e = 0$, nominal level θ, results based on 5000 replications. **Bold** when size is 3% above nominal level

		DGP 1			
θ	Diff-in-Diffs	Synthetic Control	Constrained Lasso	Ferman and Pinto	DBSCM
5%	0.043	0.047	0.046	0.047	0.056
10%	0.096	0.093	0.092	0.091	0.11
15%	0.147	0.141	0.144	0.142	0.162
20%	0.194	0.194	0.187	0.193	0.212
		DGP 2			
θ	Diff-in-Diffs	Synthetic Control	Constrained Lasso	Ferman and Pinto	DBSCM
5%	0.05	0.04	0.044	0.043	0.051
10%	0.102	0.093	0.092	0.094	0.1
15%	0.156	0.148	0.148	0.149	0.15
20%	0.208	0.193	0.194	0.194	0.2
		DGP 3			
θ	Diff-in-Diffs	Synthetic Control	Constrained Lasso	Ferman and Pinto	DBSCM
5%	0.078	**0.111**	0.043	0.067	0.051
10%	**0.139**	**0.2**	0.092	0.125	0.095
15%	**0.204**	**0.281**	0.142	**0.185**	0.148
20%	**0.262**	**0.345**	0.194	**0.241**	0.2
		DGP 4			
θ	Diff-in-Diffs	Synthetic Control	Constrained Lasso	Ferman and Pinto	DBSCM
5%	0.06	0.068	0.045	0.05	0.051
10%	0.123	**0.13**	0.09	0.109	0.099
15%	0.174	**0.182**	0.145	0.157	0.153
20%	0.0223	**0.237**	0.194	0.21	0.203

Table 5: $\rho_e = 0.5$, nominal level θ, results based on 5000 replications. **Bold** when size is 3% above nominal level

		DGP 1			
θ	Diff-in-Diffs	Synthetic Control	Constrained Lasso	Ferman and Pinto	DBSCM
5%	0.047	0.053	0.054	0.053	0.063
10%	0.093	0.103	0.106	0.107	0.116
15%	0.147	0.154	0.1456	0.16	0.17
20%	0.199	0.21	0.207	0.212	0.221
		DGP 2			
θ	Diff-in-Diffs	Synthetic Control	Constrained Lasso	Ferman and Pinto	DBSCM
5%	0.055	0.045	0.05	0.051	0.056
10%	0.109	0.099	0.103	0.102	0.108
15%	0.158	0.147	0.158	0.153	0.156
20%	0.211	0.203	0.204	0.204	0.206
		DGP 3			
θ	Diff-in-Diffs	Synthetic Control	Constrained Lasso	Ferman and Pinto	DBSCM
5%	**0.076**	**0.115**	0.048	0.069	0.058
10%	**0.144**	**0.204**	0.099	**0.129**	0.11
15%	**0.198**	**0.278**	0.154	**0.186**	0.161
20%	**0.255**	**0.346**	0.21	**0.241**	0.211
		DGP 4			
θ	Diff-in-Diffs	Synthetic Control	Constrained Lasso	Ferman and Pinto	DBSCM
5%	0.063	0.066	0.051	0.054	0.055
10%	0.119	**0.128**	0.102	0.109	0.102
15%	**0.175**	**0.186**	0.154	0.159	0.157
20%	**0.0232**	**0.246**	0.2	0.211	0.207

Table 6: $\rho_e = 0.9$, nominal level θ, results based on 5000 replications. **Bold** when size is 3% above nominal level

DGP 1					
θ	Diff-in-Diffs	Synthetic Control	Constrained Lasso	Ferman and Pinto	DBSCM
5%	**0.084**	**0.081**	**0.095**	**0.099**	0.074
10%	**0.148**	**0.138**	**0.162**	**0.161**	**0.135**
15%	**0.2**	**0.196**	**0.218**	**0.219**	**0.189**
20%	**0.246**	**0.247**	**0.272**	**0.27**	**0.244**

DGP 2					
θ	Diff-in-Diffs	Synthetic Control	Constrained Lasso	Ferman and Pinto	DBSCM
5%	0.079	0.073	**0.091**	**0.091**	0.065
10%	**0.137**	**0.131**	**0.155**	**0.161**	0.121
15%	**0.188**	**0.183**	**0.213**	**0.216**	0.173
20%	**0.247**	**0.236**	**0.267**	**0.266**	0.222

DGP 3					
θ	Diff-in-Diffs	Synthetic Control	Constrained Lasso	Ferman and Pinto	DBSCM
5%	0.07	**0.12**	**0.095**	**0.08**	0.066
10%	**0.154**	**0.214**	**0.161**	**0.146**	0.122
15%	**0.212**	**0.287**	**0.222**	**0.207**	0.176
20%	**0.274**	**0.358**	**0.277**	**0.265**	0.229

DGP 4					
θ	Diff-in-Diffs	Synthetic Control	Constrained Lasso	Ferman and Pinto	DBSCM
5%	0.077	**0.082**	**0.09**	0.078	0.063
10%	**0.137**	**0.146**	**0.156**	**0.14**	0.124
15%	**0.199**	**0.208**	**0.209**	**0.198**	0.177
20%	**0.0251**	**0.266**	**0.259**	**0.253**	0.227

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association 105*(490), 493–505.

Abadie, A., A. Diamond, and J. Hainmueller (2015). Comparative politics and the synthetic control method. *American Journal of Political Science 59*(2), 495–510.

Abadie, A. and J. Gardeazabal (2003, March). The economic costs of conflict: A case study of the basque country. *American Economic Review 93*(1), 113–132.

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

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021, December). Synthetic difference-in-differences. *American Economic Review 111*(12), 4088â€“4118.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Ben-Michael, E., A. Feller, and J. Rothstein (2021). The augmented synthetic control method. *Journal of the American Statistical Association 116*(536), 1789–1803.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association 116*(536), 1849–1864.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2022). A t-test for synthetic controls.

Doudchenko, N. and G. W. Imbens (2016, October). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. (22791).

Durrett, R. (2019). *Probability: Theory and Examples (5th ed.)*. Cambridge: Cambridge University Press.

Ferman, B. (2021). On the properties of the synthetic control estimator with many periods and many controls. *Journal of the American Statistical Association 116*(536), 1764–1772.

Ferman, B. and C. Pinto (2021). Synthetic controls with imperfect pretreatment fit. *Quantitative Economics 12*(4), 1197–1221.

Hall, P. and C. Heyde (1980). *Martingale limit theory and its application.* Academic Press New York.

Newey, W. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica 59*(4), 1161–67.

Newey, W. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. Engel and D. McFadden (Eds.), *Handbook of Econometrics.* Elsvier Science, Amsterdam: Elsvier Science.

Wainwright, M. J. (2019). *High-dimensional Statistics: A Non-Asymptotic Viewpoint*, Volume 48. Cambridge university press.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis 25*(1), 57–76.

Contents

A Proof of Lemmas

A.1 Auxiliary Lemma

Lemma A.1. *Let $A \in \mathbb{R}^{m\times n}$, so that by singular value decomposition we can write $A = U\Sigma V'$, where $\Sigma \in \mathbb{R}^{m\times n}$ has non-zero elements except possibly only its diagonals, with these values denoted as $\sigma_1, ..., \sigma_r$. The minimum-norm least squares solution to the linear equation $AX = b$, that is, the shortest vector X that achieves*

$$\min_X ||AX - b||^2 \equiv \sum_{i=r+1}^{n} (U_i'b)^2$$

is unique, given by

$$\widehat{X} = V\Sigma^{\dagger}U'b$$

where

$$\Sigma^{\dagger} = \begin{pmatrix} 1/\sigma_1 & & & & 0 & \cdots & 0 \\ & 1/\sigma_2 & & & & & \\ & & \ddots & & \vdots & & \vdots \\ & & & 1/\sigma_r & \vdots & & \vdots \\ & & & & 0 & & \\ & & & & \vdots & \ddots & \\ & & & & 0 & \cdots & 0 \end{pmatrix}$$

Also, $||\widehat{X}||^2 = \sum_{i=1}^{r} (U_i'b/\sigma_i)^2$

Lemma A.1:
The least square solution to $AX = b$ can be written as

$$\min_X ||U\Sigma V'X - b|| = \min_X ||U(\Sigma V'X - U'b)|| \stackrel{(i)}{=} \min_X ||(\Sigma V'X - U'b)|| \stackrel{(ii)}{=} \min_y ||(\Sigma y - c)||$$

where (i) follows from the fact that U is orthogonally-normalized so that the euclidean-norm remains unchanged; (ii) follows by defining $y := V'X$ and $c := U'b$. We want to minimize the vector

$$\begin{pmatrix} \sigma_1 & 0 & & \cdots & 0 & 0 \\ & \ddots & & & & 0 \\ & & \sigma_r & & & \vdots \\ & & & 0 & & \vdots \\ & & & & \ddots & \vdots \\ & & & & & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_r \\ y_{r+1} \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} c_1 \\ \vdots \\ \vdots \\ c_r \\ c_{r+1} \\ \vdots \\ c_n \end{pmatrix}$$

which leads to the solution

$$y_i = \frac{c_i}{\sigma_i} \quad for\ i \in 1, ..., r$$

with the choice of y_i to be any number for $i \in r+1, ..., n$. However, note that by $VV' = I$, we have $||X|| = ||V'X|| = ||y||$. In order to minimize $||X||$ we have to minimize $||y||$, which forces us to choose $y_i := 0$ for $i \in r+1, ..., n$, i.e. $y = \Sigma^{\dagger} c$ is the unique solution to the minimum-norm least square problem. Solving for X yields

$$\widehat{X} = Vy = V\Sigma^{\dagger} c = V\Sigma^{\dagger} U'b$$

It is clearly unique. Furthermore, since $y_i \equiv 0$ for $i = r+1, ..., n$

$$\min_X ||AX - b|| = ||A\widehat{X} - b|| = ||(-c_{r+1}, \cdots, -c_n)|| = \sum_{i=r+1}^{n} (U_i'b)^2.$$

Finally,

$$||\widehat{X}||^2 = ||Vy||^2 = ||y||^2 = \sum_{i=1}^{r} (c_i/\sigma_i)^2 = \sum_{i=1}^{r} (U_i'b/\sigma_i)^2$$

$\square$

A.2 Proof of Lemma 3.1

Define $W_{-(J+1)} := (W_1, ..., W_J)'$ and $W := (W_1, ..., W_J, W_{J+1})'$. Then by some algebraic manipulation we can obtain

$$M_0 - MW_{-(J+1)} = (\mu_0 - \mu W) - (1 - \sum_{j=1}^{J+1} W_j)\mu_{J+1} \tag{A.1}$$

and

$$C_0 - C'W_{-(J+1)} = (c_0 - c'W) + c_{J+1}(W_{J+1} - 1) = (c_0 - c'W) \tag{A.2}$$

where we recall that $c_{J+1} \equiv 0$. If there exists some $W^\dagger = (W_1^\dagger, ..., W_J^\dagger, W_{J+1}^\dagger)' \in \widetilde{\Phi}$, then by recalling that $\sum_{j=1}^{J+1} W_j^\dagger = 1$ and observing both (A.1) and (A.2), we have that $W^* := W_{-(J+1)}^\dagger \in \Phi$. Finally, note $\|W^*\| \leq 1$.

B Proof of Theorems

B.1 Proof of Theorem 1

We will prove Theorem 1 for $\widetilde{W}_T^{BSCM}(f,\lambda)$ defined with (y_{0t}, y_t) instead of (Y_{0t}, Y_t) as defined in (3.2), i.e. we will prove that Theorem 1 holds for

$$\widetilde{W}_T^{BSCM}(f,\Lambda) := \underset{W\in\Delta_\eta^J}{\arg\min}\left\{\frac{1}{k}\sum_{s=1}^{k}\left\{\overline{y}_0^s - (\overline{y}^s)'W)\right\}^2 + f(W,\Lambda)\right\}$$

instead. The proof using (3.2) is similar, with the difference of additional notations.

Step 1: We show that for any $W \in \Delta_\eta^J$,

$$\frac{1}{k\Delta^2}\sum_{s=1}^{k}\left\{\sum_{q=1}^{\Delta}(\varepsilon_{0q}^s - (\varepsilon_q^s)'W)\right\}^2 = o_p(1) \tag{B.1}$$

Fix any $W \in \Delta_\eta^J$ and observe

$$\begin{aligned}
&\frac{1}{k\Delta^2}\sum_{s=1}^{k}\left\{\sum_{q=1}^{\Delta}(\varepsilon_{0q}^s - (\varepsilon_q^s)'W)\right\}^2 \\
&= \frac{1}{k\Delta^2}\sum_{s=1}^{k}\sum_{q=1}^{\Delta}(\varepsilon_{0q}^s - (\varepsilon_q^s)'W)^2 + 2\frac{1}{k\Delta^2}\sum_{s=1}^{k}\sum_{\ell=1}^{\Delta}\sum_{q=1}^{\ell-1}(\varepsilon_{0\ell}^s - (\varepsilon_\ell^s)'W)(\varepsilon_{0q}^s - (\varepsilon_q^s)'W) \\
&= \frac{1}{T_0\Delta}\sum_{t=1}^{T_0}(\varepsilon_{0t} - \varepsilon_t'W)^2 + 2\frac{1}{k\Delta^2}\sum_{s=1}^{k}\sum_{\ell=1}^{\Delta}\sum_{q=1}^{\ell-1}(\varepsilon_{0\ell}^s - (\varepsilon_\ell^s)'W)(\varepsilon_{0q}^s - (\varepsilon_q^s)'W) \equiv A_1 + A_2.
\end{aligned}$$

We will show that $A_1, A_2 = o_p(1)$. Noting the simple inequality of $(a+b)^2 \le 2a^2 + 2b^2$,

$$\begin{aligned}
A_1 &\le \frac{1}{T_0\Delta}\sum_{t=1}^{T_0}\varepsilon_{0t}^2 + \frac{1}{\Delta}W'(\frac{1}{T_0}\sum_{t=1}^{T_0}\varepsilon_t\varepsilon_t')W - \frac{2}{T_0\Delta}\sum_{t=1}^{T_0}\varepsilon_{0t}\varepsilon_t'W \\
&= \frac{1}{\Delta}O_p(1) + \frac{1}{\Delta}W'(\Sigma + o_p(1))W + O_p(\Delta^{-1}) = o_p(1)
\end{aligned}$$

as $\Delta \to \infty$, where the second last equality follows from assumptions 3.3 and 3.4. To deal with A_2, for notational simplicity, define $X_{0,\ell}^s := \sum_{q=1}^{\ell-1}(\varepsilon_{0\ell}^s - (\varepsilon_\ell^s)'W)(\varepsilon_{0q}^s - (\varepsilon_q^s)'W)$. Then we have

$$\mathbb{E}|A_2| = \mathbb{E}\left|\frac{1}{k\Delta^2}\sum_{s=1}^{k}\sum_{\ell=1}^{\Delta}X_{0,\ell}^s\right| \overset{(i)}{\lesssim} \frac{1}{k\Delta^2}\sum_{s=1}^{k}\sum_{\ell=1}^{\Delta}(1+\Delta^{1-p/3}) = \Delta^{-1} + \Delta^{-p/3} = o(1)$$

where (i) follows from

$$\mathbb{E}|X_{0,\ell}^s| \overset{(i.1)}{\leq} 8\sum_{q=1}^{\ell-1}(\mathbb{E}|\varepsilon_{0\ell}^s-(\varepsilon_\ell^s)'W|^3)^{1/3}\cdot(\mathbb{E}|\varepsilon_{0q}^s-(\varepsilon_q^s)'W|^3)^{1/3}\cdot\alpha(\ell-q)^{1/3} \overset{(i.2)}{\lesssim} \sum_{q=1}^{\ell-1}(\ell-q)^{-p/3}$$
$$=\sum_{q=1}^{\ell-1}q^{-p/3}\leq 1+\int_1^{\ell-1}x^{-p/3}dx = 1+\frac{(\ell-1)^{1-p/3}-1}{1-p/3}\lesssim 1+(\ell-1)^{1-p/3}\lesssim 1+\Delta^{1-p/3}$$

$(i.1)$ follows from Durrett (2019)[Lemma 8.3.6]; $(i.2)$ follows from assumptions 3.2 and 3.3; hence $A_2=o_p(1)$ by Markov inequality. We conclude that (B.1) is shown.

Step 2: We write $\widetilde{\mathcal{A}}_{T_0}(W,\Lambda)\equiv\widetilde{\mathcal{A}}_{T_0}(f,W,\Lambda)$ where f is a known fixed function, for notational simplicity. Then define

$$\widetilde{\mathcal{A}}_{T_0}(W,\Lambda):=\frac{1}{k}\sum_{s=1}^{k}\{\overline{y}_0^s-(\overline{y}^s)'W\}^2+f(W,\Lambda)$$

and

$$\mathcal{A}(W,\Lambda):=(c_0-c'W)^2+(\mu_0-\mu W)'\Omega_0(\mu_0-\mu W)+f(W,\Lambda),$$

we want to show that

$$\sup_{(W,\Lambda)\in\Delta_\eta^J\times[0,1]}\left|\widetilde{\mathcal{A}}_{T_0}(W,\Lambda)-\mathcal{A}(W,\Lambda)\right|=o_p(1) \tag{B.2}$$

First we require a lemma:

Lemma B.1. *(Corollary 2.2 of Newey (1991)) Assume (1) $\Delta_\eta^J\times[\gamma,1]$ is compact, (2) $\widetilde{\mathcal{A}}_{T_0}(W,\Lambda)\xrightarrow{p}\mathcal{A}(W,\Lambda)$ for every $(W,\Lambda)\in\Delta_\eta^J\times[\gamma,1]$, (3) $\Delta_\eta^J\times[\gamma,1]$ is a metric space and (4) there is a B_{T_0} such that $B_{T_0}=O_p(1)$ and for all $(W_1,\Lambda_1),(W_2,\Lambda_2)\in\Delta_\eta^J$, $|\widetilde{\mathcal{A}}_{T_0}(W_1,\Lambda_1)-\widetilde{\mathcal{A}}_{T_0}(W_2,\Lambda_2)|\leq B_{T_0}||(W_1,\Lambda_1)-(W_2,\Lambda_2)||$ and (5) $\{\mathcal{A}(W,\Lambda)\}_{(W,\Lambda)\in\Delta_\eta^J\times[\gamma,1]}$ is equi-continuous. Then $\widetilde{\mathcal{A}}_{T_0}(W,\Lambda)\xrightarrow{p}\mathcal{A}(W,\Lambda)$ uniformly over $(W,\Lambda)\in\Delta_\eta^J\times[\gamma,1]$*

Fixing any $(W,\Lambda)\in\Delta_\eta^J\times[\gamma,1]$, we have

$$\begin{aligned}\widetilde{\mathcal{A}}_{T_0}(W,\Lambda)&=\frac{1}{k}\sum_{s=1}^{k}\Big[(c_0-c'W)+(\overline{\lambda}^s)'(\mu_0-\mu'W)+(\overline{\varepsilon}_0^s-(\overline{\varepsilon}^s)'W)\Big]^2+f(W,\Lambda)\\&=(c_0-c'W)^2+(\mu_0-\mu'W)'\left(\frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^s(\overline{\lambda}^s)'\right)(\mu_0-\mu'W)+\frac{1}{k}\sum_{s=1}^{k}(\overline{\varepsilon}_0^s-(\overline{\varepsilon}^s)'W)^2\\&+2(c_0-c'W)\left(\frac{1}{k}\sum_{s=1}^{k}(\overline{\lambda^s})'\right)(\mu_0-\mu'W)+2(c_0-c'W)\left(\frac{1}{k}\sum_{s=1}^{k}(\overline{\varepsilon}_0^s-(\overline{\varepsilon}^s)'W)\right)\\&+2(\mu_0-\mu'W)'\left(\frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^s(\overline{\varepsilon}_0^s-(\overline{\varepsilon}^s)'W)\right)-(\overline{y}_0-\overline{y}'W)^2+f(W,\Lambda),\end{aligned}$$

so that by

$$
\begin{aligned}
&(a) \quad \frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^{s}(\overline{\lambda}^{s})' = \Omega_0 + o_p(1) \\
&(b) \quad \frac{1}{k}\sum_{s=1}^{k}(\overline{\varepsilon}_0^s - (\overline{\varepsilon}^s)'W)^2 = \frac{1}{k\Delta^2}\sum_{s=1}^{k}\left\{\sum_{q=1}^{\Delta}(\varepsilon_{0q}^s - (\varepsilon_q^s)'W)\right\}^2 = o_p(1) \quad \text{by (B.1)} \\
&(c) \quad \frac{1}{k}\sum_{s=1}^{k}(\overline{\lambda^s})' = \frac{1}{T_0}\sum_{t=1}^{T_0}(\lambda_t)' = o_p(1) \quad \text{by assumption 3.4} \\
&(d) \quad \frac{1}{k}\sum_{s=1}^{k}(\overline{\varepsilon}_0^s - (\overline{\varepsilon}^s)'W) = \frac{1}{T_0}\sum_{t=1}^{T_0}\varepsilon_{0t} - \frac{1}{T_0}\sum_{t=1}^{T_0}\varepsilon_t'W = o_p(1) \quad \text{by assumption 3.2} \\
&(e) \quad \frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^{s}(\overline{\varepsilon}_0^s - (\overline{\varepsilon}^s)'W) = \frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^{s}\overline{\varepsilon}_0^s - \frac{1}{T_0}\sum_{t=1}^{T_0}(\varepsilon_t)'W = o_p(1)
\end{aligned}
$$

where the last equality in (e) follows from

$$
\left\|\frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^{s}\overline{\varepsilon}_0^s\right\| \leq \frac{1}{k}\sum_{s=1}^{k}\|\overline{\lambda}^{s}\| \cdot \|\overline{\varepsilon}_0^s\| \overset{(i)}{=} \frac{1}{k}\sum_{s=1}^{k}\|\overline{\lambda}^{s}\|o_p(1) \overset{(ii)}{=} \frac{1}{k}\sum_{s=1}^{k}O_p(1)o_p(1) = o_p(1),
$$

where (i) follows from Hall and Heyde (1980)[Corollary 5.1] and the fact that $\sum_{\tau=1}^{\infty}\alpha(\tau)^{p/(2+p)} < \infty$ by assumption 3.2; (ii) follows from assumption 3.3 while the last equality follows from a fixed number of k. Therefore we have shown

$$
\widetilde{\mathcal{A}}_{T_0}(W,\Lambda) \xrightarrow{p} \mathcal{A}(W,\Lambda) \tag{B.3}
$$

It is clear that $\Delta_\eta^J \times [0,1]$ is compact, so condition (1) of Lemma B.1 is satisfied. It is clear that condition (3) and (5) also holds, i.e. $\{\mathcal{A}(W,\Lambda)\}_{(W,\Lambda)\in\Delta_\eta^J\times[0,1]}$ is equi-continuous. Condition (2) follows from (B.3). To show (B.2), it remains to prove that condition (4) of Lemma B.1 holds, which is what we do now.

We can remove the common time-effect from $\widetilde{\mathcal{A}}_{T_0}(W)$ by defining $\widetilde{y}_0^s := \overline{y}_0^s - \overline{\delta}^s$ and $\widetilde{y}^s := \overline{y}^s - \iota\overline{\delta}^s$, so that

$$
\widetilde{\mathcal{A}}_{T_0}(W,\Lambda) = \frac{1}{k}\sum_{s=1}^{k}\{\widetilde{y}_0^s - (\widetilde{y}^s)'W + \overline{\delta}^s(1-\iota'W)\}^2 + f(W,\Lambda)
$$

Then using mean value theorem, for any $(W_1,\Lambda_1),(W_2,\Lambda_2) \in \Delta_\eta^J \times [0,1]$, there exists a $(W_3,\Lambda_3) \in \Delta_\eta^J \times [0,1]$ such that

$$
\left|\widetilde{\mathcal{A}}_{T_0}(W_1,\Lambda_1) - \widetilde{\mathcal{A}}_{T_0}(W_2,\Lambda_2)\right|
$$

$$
\begin{aligned}
&= \left| \left(\frac{2}{k} \sum_{s=1}^{k} \{\widetilde{y}_0^s - (\widetilde{y}^s)' W_3\}(-\widetilde{y}^s - \overline{\delta}^s \iota) + ||W_3||^2 + 2\Lambda_3 W_3 \right) \cdot ||(W_1, \Lambda_1) - (W_2, \Lambda_2)|| \right| \\
&= B_{T_0} ||(W_1, \Lambda_1) - (W_2, \Lambda_2)||
\end{aligned}
$$

with

$$
\begin{aligned}
B_{T_0} :=& \left|\left| \left(\frac{2}{k} \sum_{s=1}^{k} \{\widetilde{y}_0^s - (\widetilde{y}^s)' W_3\}(-\widetilde{y}^s - \overline{\delta}^s \iota) + ||W_3||^2 + 2\Lambda_3 W_3 \right) \right|\right| \\
\leq& \left|\left| \left(\frac{2}{k} \sum_{s=1}^{k} \{\widetilde{y}_0^s - (\widetilde{y}^s)' W_3\}(\widetilde{y}^s) \right) \right|\right| + \left|\left| \left(\frac{2}{k} \sum_{s=1}^{k} \{\widetilde{y}_0^s - (\widetilde{y}^s)' W_3\}\overline{\delta}^s \right) \right|\right| \cdot ||\iota|| + ||W_3||^2 + 2\Lambda_3 ||W_3|| \\
\leq& \left|\left| \frac{2}{k} \sum_{s=1}^{k} \widetilde{y}_0^s \widetilde{y}^s \right|\right| + \left|\left| \frac{2}{k} \sum_{s=1}^{k} \widetilde{y}^s (\widetilde{y}^s)' \right|\right| \times ||W_3|| + \left|\left| \frac{2}{k} \sum_{s=1}^{k} \widetilde{y}_0^s \overline{\delta}^s \right|\right| + \left|\left| \frac{2}{k} \sum_{s=1}^{k} \widetilde{y}^s \overline{\delta}^s \right|\right| \times ||W_3|| \times ||\iota|| + \eta^2 + 2\eta \\
=& ||2A_1|| + ||2A_2|| \times \eta + ||2A_3|| + ||2A_4|| \times \sqrt{J}\eta + \eta^2 + 2\eta
\end{aligned}
$$

where the second last inequality follows from $||W_3|| \leq \eta$ and $\Lambda \leq 1$. We will show that $B_{T_0} = O_p(1)$ by showing that each term $A_1, ..., A_4$ is $O_p(1)$. Observe first that

$$
\begin{aligned}
&(a) \quad \frac{1}{k} \sum_{s=1}^{k} \overline{\lambda}^s (\overline{\lambda}^s)' = \Omega_0 \\
&(b) \quad \frac{1}{k} \sum_{s=1}^{k} \overline{\varepsilon}^s (\overline{\varepsilon}^s)' = O_p(1) \\
&(c) \quad \frac{1}{k} \sum_{s=1}^{k} \overline{\lambda}^s = \frac{1}{T_0} \sum_{t=1}^{T_0} \lambda_t = o_p(1) \\
&(d) \quad \frac{1}{k} \sum_{s=1}^{k} (\overline{\varepsilon}^s)' = \frac{1}{T_0} \sum_{t=1}^{T_0} (\varepsilon_t)' = o_p(1) \\
&(e) \quad \frac{1}{k} \sum_{s=1}^{k} \overline{\lambda}^s (\overline{\varepsilon}^s)' = O_p(1) \\
&(f) \quad \frac{1}{k} \sum_{s=1}^{k} \overline{\varepsilon}_0^s (\overline{\varepsilon}^s)' = O_p(1) \\
&(g) \quad \frac{1}{k} \sum_{s=1}^{k} \overline{\delta}^s = \frac{1}{T_0} \sum_{t=1}^{T_0} \delta_t = O_p(1) \\
&(h) \quad \frac{1}{k} \sum_{s=1}^{k} \overline{\delta}^s (\overline{\lambda}^s)' = O_p(1) \\
&(i) \quad \frac{1}{k} \sum_{s=1}^{k} \overline{\delta}^s (\overline{\varepsilon}^s)' = O_p(1)
\end{aligned}
$$

where (c) and (d) follows from the assumptions, (a) follows from Markov-inequality and

$$\mathbb{E}\left|\left|\frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^{s}(\overline{\lambda}^{s})'\right|\right| \le \frac{1}{k\Delta^2}\sum_{s=1}^{k}\mathbb{E}\left|\left|\sum_{m=1}^{\Delta}\lambda_m^s\right|\right|^2 \le \frac{1}{k\Delta^2}\sum_{s=1}^{k}\sum_{m=1}^{\Delta}\sum_{\ell=1}^{\Delta}\mathbb{E}\left(||\lambda_m^s||\cdot||\lambda_\ell^s||\right)$$
$$\le \frac{2}{k\Delta^2}\sum_{s=1}^{k}\sum_{m=1}^{\Delta}\sum_{\ell=1}^{\Delta}\left(\mathbb{E}||\lambda_m^s||^2+\mathbb{E}||\lambda_\ell^s||^2\right) \lesssim 1,$$

(b) follows in an analogous manner, (e) follows from Markov inequality and

$$\mathbb{E}\left|\left|\frac{1}{r}\sum_{s=1}^{r}\overline{\lambda}^{s}(\overline{\varepsilon}^{s})'\right|\right| \le \frac{1}{r}\sum_{s=1}^{r}\mathbb{E}\left(||\frac{1}{\Delta}\sum_{m=1}^{\Delta}\lambda_m^s||\cdot||\frac{1}{\Delta}\sum_{m=1}^{\Delta}\varepsilon_m^s||\right)$$
$$\le \frac{1}{r\Delta^2}\sum_{s=1}^{r}\sum_{m=1}^{\Delta}\sum_{\ell=1}^{\Delta}\mathbb{E}\left(||\lambda_m^s||\cdot||\varepsilon_m^s||\right) \lesssim 1,$$

(f) follows from Markov inequality and

$$\mathbb{E}\left|\left|\frac{1}{k}\sum_{s=1}^{k}\overline{\varepsilon}_0^{s}(\overline{\varepsilon}^{s})'\right|\right| \le \frac{1}{k}\sum_{s=1}^{k}\mathbb{E}\left(||\frac{1}{\Delta}\sum_{m=1}^{\Delta}\varepsilon_{0,m}^s||\cdot||\frac{1}{\Delta}\sum_{m=1}^{\Delta}\varepsilon_m^s||\right)$$
$$\le \frac{1}{k\Delta^2}\sum_{s=1}^{k}\sum_{m=1}^{\Delta}\sum_{\ell=1}^{\Delta}\mathbb{E}\left(||\varepsilon_{0,m}^s||\cdot||\varepsilon_m^s||\right) \lesssim 1,$$

(g) follows from Markov inequality and bounded second moment of δ_t, both (h) and (i) follows in the same way as (e).

We can show $A_1, ..., A_4 = O_p(1)$ by writing

$$\begin{aligned}
A_1 &= \frac{1}{k}\sum_{s=1}^{k}(c_0+\mu_0'\overline{\lambda}^s+\overline{\varepsilon}_0^s)(c+\mu'\overline{\lambda}^s+\overline{\varepsilon}^s)' \\
&= c_oc + c_0\mu\cdot\frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^s + c_0(\frac{1}{k}\sum_{s=1}^{k}\overline{\varepsilon}^s) + c\mu_0'(\frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^s) + \mu_0'(\frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^s(\overline{\lambda}^s)')\mu \\
&+ \mu_0'(\frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^s(\overline{\varepsilon}^s)') + (\frac{1}{k}\sum_{s=1}^{k}\overline{\varepsilon}_0^s)c' + (\frac{1}{k}\sum_{s=1}^{k}\overline{\varepsilon}_0^s(\overline{\lambda}^s)')\mu + \frac{1}{k}\sum_{s=1}^{k}\overline{\varepsilon}_0^s(\overline{\varepsilon}^s)', \\
A_2 &= \frac{1}{k}\sum_{s=1}^{k}(c+\mu'\overline{\lambda}^s+\overline{\varepsilon}^s)\cdot(c+\mu'\overline{\lambda}^s+\overline{\varepsilon}^s)' \\
&= cc' + \mu'\left(\frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^s(\overline{\lambda}^s)\right)\mu + \frac{1}{k}\sum_{s=1}^{k}\overline{\varepsilon}^s(\overline{\varepsilon}^s)' \\
&+ 2c\left(\frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^s\right)\mu + 2c\frac{1}{k}\sum_{s=1}^{k}(\overline{\varepsilon}^s)' + 2\mu'\left(\frac{1}{k}\sum_{s=1}^{k}\overline{\lambda}^s(\overline{\varepsilon}^s)'\right),
\end{aligned}$$

$$A_3 = \frac{1}{k}\sum_{s=1}^{k}\overline{\delta}^s(c + \mu'\overline{\lambda}^s + \overline{\varepsilon}^s)' = (\frac{1}{k}\sum_{s=1}^{k}\overline{\delta}^s)c' + (\frac{1}{k}\sum_{s=1}^{k}\overline{\delta}^s(\overline{\lambda}^s)')\mu + \frac{1}{k}\sum_{s=1}^{k}\overline{\delta}^s(\overline{\varepsilon})'$$
$$A_4 = \frac{1}{k}\sum_{s=1}^{k}(c + \mu'\overline{\lambda}^s + \overline{\varepsilon}^s)\overline{\delta}^s = c(\sum_{s=1}^{k}\overline{\delta}^s) + \mu'(\sum_{s=1}^{k}\overline{\lambda}^s\overline{\delta}^s) + \sum_{s=1}^{k}\overline{\varepsilon}^s\overline{\delta}^s$$

and then applying $(a) - (i)$. Therefore condition (4) of Lemma B.1 is shown, implying (B.2).

Step 3: For notational simplicity, we write $\widetilde{W}(\Lambda) \equiv \widetilde{W}^{BSCM}(f, \Lambda)$. We show that for any fixed $\gamma > 0$,

$$\sup_{\Lambda\in[\gamma,1]}\left|\widetilde{W}(\Lambda) - \overline{W}(\Lambda)\right| = o_p(1), \tag{B.4}$$

where $\overline{W}(\Lambda) = \arg\min_{W\in\Delta_\eta^J}\mathcal{A}(W, \Lambda)$.

Lemma B.2. *(Newey and McFadden (1994)[Theorem 2.1]) Suppose there is a function $Q_0(\theta)$ such that it is (i) uniquely minimized at θ_0; (ii) Θ is compact, where $\theta \in \Theta$; (iii) $Q_0(\theta)$ is continuous and (iv) $\sup_{\theta\in\Theta}|\widehat{Q}(\theta) - Q_0(\theta)| = o_p(1)$. Then for $\widehat{\theta} := \arg\min\widehat{Q}(\theta)$, we have $\widehat{\theta} \xrightarrow{p} \theta_0$*

We first show point-wise convergence of $\widetilde{W}(\Lambda)$ for every $\Lambda \in [\gamma, 1]$. Replace Θ by Δ_η^J, which is compact. We fix any Λ and replace $Q_0(\theta)$ by $\mathcal{A}(\Lambda, W)$. Since $\mathcal{A}(\Lambda, W)$ is strictly convex (for $\Lambda > 0$), it has a uniquely-minimized solution. Clearly $\mathcal{A}(W, \Lambda)$ is continuous in Θ and condition (iv) of Lemma B.2 follows from equation (B.2), with $\widehat{Q}(\theta)$ as $\widehat{\mathcal{A}}_{T_0}(W, \Lambda)$. Therefore we have

$$\widetilde{W}(\Lambda) \xrightarrow{p} \overline{W}(\Lambda) \tag{B.5}$$

for every fixed $\Lambda \in [\gamma, 1]$. This satisfies condition (2) of Lemma B.1. Since $[\gamma, 1]$ is compact, condition (1) is satisfied. Condition (3) is clear. To show that $\overline{W}(\Lambda)_{\Lambda\in[\gamma,1]}$ is equi-continuous, first observe that

$$\overline{W}(\Lambda) = (\Lambda I + cc' + \mu'\Omega_0\mu)^{-1}(\mu'\Omega_0\mu_0 + c_0c)$$

We can take the spectral decomposition of $cc' + \mu'\Omega_0\mu = VDV'$, where $VV' = I = V'V$ and D is the diagonal matrix with non-negative eigenvalues $(d_1, ..., d_J)$ as its elements. Define $D^\Lambda = D + \Lambda I$. Then for any $\Lambda_1, \Lambda_2 \in [\gamma, 1]$,

$$\begin{aligned}
\left|\left|\overline{W}(\Lambda_1) - \overline{W}(\Lambda_2)\right|\right| &\leq ||(\Lambda_1 I + cc' + \mu'\Omega_0\mu)^{-1} - (\Lambda_2 I + cc' + \mu'\Omega_0\mu)^{-1}||_\infty \cdot ||\mu'\Omega_0\mu_0 + c_0c||_1 \\
&= ||V((D^{\Lambda_1})^{-1} - (D^{\Lambda_2})^{-1})V'||_\infty \cdot ||\mu'\Omega\mu_0 + c_0c||_1 \\
&\overset{(i)}{=} ||(D^{\Lambda_1})^{-1} - (D^{\Lambda_2})^{-1}||_\infty \cdot ||\mu'\Omega_0\mu_0 + c_0c||_1 \\
&= ||\mu'\Omega_0\mu_0 + c_0c||_1 \max_{i=1,\dots,J}\frac{|\Lambda_1 - \Lambda_2|}{(d_i + \Lambda_1)(d_i + \Lambda_2)} \\
&\leq \frac{||\mu'\Omega_0\mu_0 + c_0c||_1}{\gamma^2}|\Lambda_1 - \Lambda_2|
\end{aligned}$$

where (i) follows from V being orthogonal. Condition (5) of Lemma B.1 is shown. For any

$\Lambda_1, \Lambda_2 \in [\gamma, 1]$, we can diagonalize $\frac{1}{r}\sum_{s=1}^{r} \overline{y}^s(\overline{y}^s)' = V_T D_T V_T'$ and define $D_T^\Lambda := D_T + \Lambda I$, so that

$$\begin{aligned}\left|\left|\widetilde{W}(\Lambda_1) - \widetilde{W}(\Lambda_2)\right|\right| &\leq \left\|\frac{1}{k}\sum_{s=1}^{k} \overline{y}_0^s \cdot \overline{y}^s\right\|_1 \cdot \max_{i=1,\ldots,J} \frac{|\Lambda_1 - \Lambda_2|}{(d_{i,T}+\Lambda_1)(d_{i,T}+\Lambda_2)} \\ &\leq \frac{||\frac{1}{k}\sum_{s=1}^{k} \overline{y}_0^s \cdot \overline{y}^s||_1}{\gamma^2} \cdot |\Lambda_1 - \Lambda_2| =: B_{T_0} \cdot |\Lambda_1 - \Lambda_2|\end{aligned}$$

where

$$\begin{aligned}&\gamma^2 \cdot \mathbb{E}(B_{T_0}) \leq \frac{1}{k}\sum_{j=1}^{J}\sum_{s=1}^{k} \mathbb{E}(|\overline{y}_0| \cdot |\overline{y}_j^s|) \\ &\leq \frac{1}{k}\frac{1}{\Delta^2}\sum_{j=1}^{J}\sum_{s=1}^{k} \mathbb{E}\left(\sum_{\ell=1}^{\Delta}(|c_0| + |\delta_{s,\ell}| + |\lambda_{s,\ell}'\mu_0| + |\varepsilon_{0,s\Delta+\ell}|)\right) \cdot \left(\sum_{\ell=1}^{\Delta}(|c_j| + |\delta_{s,\ell}| + |\lambda_{s,\ell}'\mu_j| + |\varepsilon_{j,s\Delta+\ell}|)\right) \\ &\lesssim \frac{1}{k}\sum_{j=1}^{J}\sum_{s=1}^{k}(1 + |c_0| + |c_j| + ||\mu_0|| + ||\mu_j||) = J(1 + |c_0| + |c_j| + ||\mu_0|| + ||\mu_j||) = O(1)\end{aligned}$$

where the last inequality follows from the bounded second moments of δ_t, λ_t and ε_{jt} by assumption. By Markov-inequality, condition (4) of Lemma B.1 is satisfied, so that we obtain (B.4).

Step 4: In the case where $cc' + \mu'\Omega_0\mu$ is positive definite, then (B.5) holds for $\Lambda = 0$ by $\mathcal{A}(0, W)$ being strictly convex. We can repeat the proof as in **step 3** with $\gamma = 0$ and show that for any $\Lambda_1, \Lambda_2 \in [0, 1]$,

$$\begin{aligned}||\overline{W}(\Lambda_1) - \overline{W}(\Lambda_2)|| &\leq ||\mu'\Omega_0\mu_0 + c_0 c||_1 \max_{i=1,\ldots,J} \frac{|\Lambda_1 - \Lambda_2|}{(d_i+\Lambda_1)(d_i+\Lambda_2)} \\ &\leq ||\mu'\Omega_0\mu_0 + c_0 c||_1 \max\{d_1^{-1}, ..., d_J^{-1}\}|\Lambda_1 - \Lambda_2|\end{aligned}$$

where we note that $\max\{d_1^{-1}, ..., d_J^{-1}\} < \infty$ by $\{d_i\}_{i=1}^J$ being the eigenvalues of $cc' + \mu'\Omega_0\mu$, a positive-definite matrix. In this case (B.4) holds with $\gamma = 0$.

Step 5: We show that $\overline{W}(\Lambda) \to W^* \equiv \overline{W}(0)$ as $0 < \Lambda \downarrow 0$

For any $\Lambda > 0$, $\mathcal{A}(W, \Lambda)$ is a strictly convex and continuous function, so that $\overline{W}(\Lambda) := \arg\min_{W \in \Delta_\eta^J} \mathcal{A}(W, \Lambda)$ is unique.[17] Recall that W^* is the unique solution that minimizes $\mathcal{H}(W) := (c_0 - c'W)^2 + (\mu_0 - \mu W)'\Omega_0(\mu_0 - \mu W)$ over $W \in \Delta_\eta^J$; this uniqueness follows from Lemma A.1. Note that for any other $W \in \Delta_\eta^J$ with the property that $\mathcal{H}(W) = \mathcal{H}(W^*)$, it must be that $||W^*|| < ||W||$ by the uniqueness property.

For any $W^\dagger \in \Delta_\eta^J$ with $||W^\dagger|| \geq ||W^*||$, we have $\mathcal{A}(W^*, \Lambda) < \mathcal{A}(W^\dagger, \Lambda)$ since $\mathcal{A}(W, \Lambda) \equiv \mathcal{H}(W) + \Lambda||W||^2$. Therefore $\overline{W}(\Lambda) \neq W^*$ whenever $||\overline{W}(\Lambda)|| \leq ||W^*||$.

[17] The strict-convexity follows from $f(W, \Lambda)$ being strictly-convex for $\Lambda > 0$

Furthermore, we know that any $W \in \Delta_\eta^J \backslash \{W^*\}$ such that $||W|| \leq ||W^*||$ has the property that $\mathcal{H}(W) > 0$, since W^* is the unique minimum-norm solution. Define $\Delta(W) := W - W^*$ and consider any fixed $\delta > 0$. Then consider the open ball around W^*, defined as $B_\delta(W^*) \equiv \{W : ||\Delta(W)|| < \delta\}$. Since $\widetilde{\Delta} := \left\{W \in \Delta_\eta^J : ||W|| \leq ||W^*||\right\}$ is compact, then $\widetilde{\Delta} \cap B_\delta^c(W^*)$ is compact, where $B_\delta^c(W^*)$ is the complement set to $B_\delta(W^*)$, which is closed. By Weierstrass extreme-value-theorem, there exists a $W^\ddagger \in \widetilde{\Delta} \cap B_\delta^c(W^*)$ with $\mathcal{H}(W^\ddagger) \equiv \inf_{W \in \widetilde{\Delta} \cap B_\delta^c(W^*)} \mathcal{H}(W)$. Note that $0 < \mathcal{H}(W^\ddagger)$ by $W^\ddagger \in \widetilde{\Delta}$. Hence there must be a $\overline{c}(\delta) > 0$ such that whenever Λ is chosen such that $0 < \Lambda < \overline{c}(\delta)$, then $\Lambda\eta < \mathcal{H}(W^\ddagger) - \mathcal{H}(W^*)$.[18] We will show that

$$\overline{W}(\Lambda) \in \widetilde{\Delta} \cap B_\delta(W^*). \tag{B.6}$$

First note that any $W \in \Delta_\eta^J \backslash \{\widetilde{\Delta}\}$ cannot possibly minimize $\mathcal{A}(W, \Lambda)$, since W^* can always be chosen. Therefore $\overline{W}(\Lambda) \in \widetilde{\Delta}$. Second, by contradiction assume $\overline{W}(\Lambda) \in B_\delta^c(W^*)$. Then

$$\mathcal{A}(W^*, \Lambda) = \mathcal{H}(W^*) + \Lambda||W^*||^2 \leq \mathcal{H}(W^*) + \Lambda\eta < \mathcal{H}(W^\ddagger) \leq \mathcal{H}(\overline{W}(\Lambda)) \leq \mathcal{A}(\overline{W}(\Lambda), \Lambda),$$

which cannot be true since $\overline{W}(\Lambda)$ is the unique minimizer of $\mathcal{A}(W, \Lambda)$. Thus $\overline{W}(\Lambda) \in B_\delta(W^*)$, proving (B.6). Finally, we see that $\overline{c}(\delta) \downarrow 0$ as $\delta \downarrow 0$ because $\inf_{W \in \widetilde{\Delta} \cap B_\delta^c(W^*)} \mathcal{H}(W)$ is non-increasing with δ. This implies that as $0 < \Lambda \downarrow 0$, $\overline{W}(\Lambda) \to W^*$.

B.2 Proof of Theorem 2

The proof follows in the exact same way as the proof of Theorem 1. For the sake of completeness, the only difference in proof comes from (B.3), i.e. we will instead show that

$$\widetilde{\mathcal{B}}_T(f, W, \Lambda) \xrightarrow{p} \mathcal{B}(f, W, \Lambda) \tag{B.7}$$

for any fixed $(W, \Lambda) \in \Delta_\eta^J \times [0, 1]$,

$$\widetilde{\mathcal{B}}_T(f, W, \Lambda) := \frac{1}{k} \sum_{s=1}^{k} \{\overline{y}_0^s - (\overline{y}^s)'W - (\overline{y}_0 - \overline{y}'W)\}^2 + f(W, \Lambda)$$

Using the notations of **Step 3** in Theorem 1, we can express

$$\begin{aligned} \widetilde{\mathcal{B}}_T(f, W, \Lambda) &= \widetilde{\mathcal{A}}_T(f, W, \Lambda) - (\overline{y}_0 - \overline{y}'W)^2 \overset{(i)}{=} \mathcal{A}(f, W, \Lambda) - (\overline{y}_0 - \overline{y}'W)^2 + o_p(1) \\ &\overset{(ii)}{=} \mathcal{A}(f, W, \Lambda) - (c_0 - c'W)^2 + o_p(1) = \mathcal{B}(f, W, \Lambda) + o_p(1), \end{aligned}$$

where (i) follows from (B.3) and (ii) follows from

$$(\overline{y}_0 - \overline{y}'W)^2 = \left((C_0 - C'W) + \overline{\lambda}(M_0 - MW) + (\overline{u}_0 - \overline{u})\right)^2 = (C_0 - C'W)^2 + o_p(1)$$

[18]Note $\mathcal{H}(W^\ddagger) > \mathcal{H}(W^*)$, because $W^\ddagger \in \widetilde{\Delta}$.

B.3 Proof of Theorem 3

The proof that $\widehat{\alpha}_{0t}^{FP}$ is more efficient than $\widehat{\alpha}_{0t}^{DID}$ is shown in Ferman and Pinto (2021)[Proposition 3]. It remains to show that $\widehat{\alpha}_{0t}^{DBSCM}$ is asymptotically more efficient than $\widehat{\alpha}_{0t}^{FP}$, which we do now. Observe

$$\begin{aligned}\widehat{\alpha}_{0t}^{FP} &\xrightarrow{p} \alpha_{0t} + \lambda_t'(\mu_0 - \mu W^{FP}) + (\varepsilon_{0t} - \varepsilon_t' W^{FP}) \\ &= \alpha_{0t} + \lambda_t'(M_0 - MW_{-(J+1)}^{FP}) + (\varepsilon_{0t} - \varepsilon_t' W^{FP})\end{aligned}$$

where $W^{FP} = (W_{-(J+1)}^{FP}, W_{J+1}^{FP}) = Plim \widehat{W}^{FP}$ and the last equation follows from $\sum_{j=1}^{J+1} W_j^{FP} = 1$. The asymptotic variance is therefore

$$\begin{aligned}&avar(\widehat{\alpha}_{0t}^{FP}) = (M_0 - MW_{-(J+1)}^{FP})'\Omega_0(M_0 - MW_{-(J+1)}^{FP}) + \sigma_\varepsilon^2(1 + ||W^{FP}||_F^2) \\ &= (M_0 - MW_{-(J+1)}^{FP})'\Omega_0(M_0 - MW_{-(J+1)}^{FP}) + \sigma_\varepsilon^2 \left(1 + (W_{-(J+1)}^{FP})' W_{-(J+1)}^{FP} + (\sum_{j=1}^{J} W_j^{FP} - 1)^2\right) \\ &= \mathcal{B}(W_{-(J+1)}^{FP}, \sigma_\varepsilon^2)\end{aligned}$$

where

$$\mathcal{B}(W, \Lambda) := (M_0 - MW)'\Omega_0(M_0 - MW) + \Lambda\left(1 + W'W + (W'\iota - 1)^2\right).$$

Furthermore, by (3.6) and Assumption 4 we have that

$$\begin{aligned}\widehat{\alpha}_{0t}^{DBSCM}(\overline{f}, \sigma_\varepsilon^2) &\xrightarrow{p} \alpha_{0t} + \lambda_t'(M_0 - M\overline{W}^0(0)) + (u_{0t} - u_t'\overline{W}^0(0)) \\ &= \alpha_{0t} + \lambda_t'(M_0 - M\overline{W}^0(0)) + (\varepsilon_{0t} - \varepsilon_t'\overline{W}^0(0) + (\sum_{j=1}^{J} \overline{W}_j^0(0) - 1)\varepsilon_{J+1,t})\end{aligned}$$

so that

$$\begin{aligned}&avar(\widehat{\alpha}_{0t}^{DBSCM}(\overline{f}, \sigma_\varepsilon^2)) \\ &= (M_0 - M\overline{W}^0(\sigma_\varepsilon^2))'\Omega_0(M_0 - M\overline{W}^0(\sigma_\varepsilon^2)) + \sigma_\varepsilon^2\left(1 + (\overline{W}^0(\sigma_\varepsilon^2))'\overline{W}^0(\sigma_\varepsilon^2) + (\sum_{j=1}^{J} \overline{W}_j^0(\sigma_\varepsilon^2) - 1)^2\right) \\ &\equiv \mathcal{B}(\overline{W}^0(\sigma_\varepsilon^2), \sigma_\varepsilon^2)\end{aligned}$$

By the fact that $\overline{W}^0(\sigma_\varepsilon^2) \in \arg\min_{W \in \Delta_\eta^J} \mathcal{B}(W, \sigma_\varepsilon^2)$, we have that $\mathcal{B}(\overline{W}^0(\sigma_\varepsilon^2), \sigma_\varepsilon^2) \leq \mathcal{B}(W_{-(J+1)}^{FP}, \sigma_\varepsilon^2)$.

B.4 Proof of Theorem 4

We begin by fixing any $t \in \mathcal{T}_0 \cup \mathcal{T}_1$ and writing in terms of Chernozhukov et al. (2021)'s notation:

$$P_t^N := Y_{0t}^N - \overline{W}'Y_t^N - \left(\overline{Y}_0^N - \overline{Y}^N\overline{W}\right) + \left(\overline{\lambda}'(M_0 - M\overline{W}) + (\overline{u}_0 - \overline{W}'\overline{u})\right) =: -v_t$$

$$Y_{0t}^N = \begin{cases} Y_{0t} - \alpha_{0t} & \text{for} \quad t \in \mathcal{T}_1 \\ Y_{0t} & \text{for} \quad t \in \mathcal{T}_0 \end{cases}$$

Furthermore, we have

$$\begin{aligned} 0 &= P_t^N + v_t \\ \alpha_t &= P_t^N + \alpha_{0t} + v_t \end{aligned}$$

under the correct null, where $\alpha_{0t} \equiv 0$ for every $t \in \mathcal{T}_1$. Note that

$$\begin{aligned} P_t^N \equiv -v_t &= \Big\{(C_0 - C'\overline{W}) + \lambda_t'(M_0 - \overline{W}'M) + (u_{0t} - \overline{W}'u_t)\Big\} \\ &\quad - \Big\{(C_0 - C'\overline{W}) + \overline{\lambda}'(M_0 - \overline{W}'M) + (\overline{u}_0 - \overline{W}'\overline{u})\Big\} + \Big(\overline{\lambda}'(M_0 - M\overline{W}) + (\overline{u}_0 - \overline{W}'\overline{u})\Big) \\ &= \lambda_t'(M_0 - \overline{W}'M) + (u_{0t} - \overline{W}'u_t) \end{aligned}$$

Then under either case 4.1 or 4.2,

$$v_t \text{ is a mean-zero stationary process} \tag{B.8}$$

since $\{(u_{0t}, u_t)\}_{t \in \mathcal{T}_0 \cup \mathcal{T}_1}$ is stationary by assumption 3. Furthermore, under the correct null we have

$$P_t^N - \widehat{P}_t^N = -\lambda_t'(\overline{W} - \widetilde{W})'M) - (\overline{W} - \widetilde{W})'u_t + \Big(\overline{\lambda}'(M_0 - M\widetilde{W}) + (\overline{u}_0 - \widetilde{W}'\overline{u})\Big)$$

Therefore, writing $\widehat{P}^N := (\widehat{P}_1^N, ..., \widehat{P}_T^N)$ and $P^N := (P_1^N, ..., P_T^N)$, by noting the simple inequality of $(a+b+c+d)^2 \le 8(a^2+b^2+c^2+d^2)$,

$$\begin{aligned} ||\widehat{P}^N - P^N||_2^2/T &= \frac{1}{T}\sum_{t=1}^{T}(\widehat{P}_t^N - P_t^N)^2 \\ &\le \frac{8}{T}\sum_{t=1}^{T}\Big\{||M'(\overline{W} - \widetilde{W})||_2^2||\lambda_t||_2^2 + ||\overline{W} - \widetilde{W}||_2^2||u_t||_2^2 + ||\overline{\lambda}||_2^2||M_0 - M\widetilde{W}||_2^2 + ||\overline{u}_0 - \widetilde{W}'\overline{u}||_2^2\Big\} \\ &\overset{(i)}{=} o_p(1) \cdot \frac{8}{T}\sum_{t=1}^{T}\{||\lambda_t||_2^2 + ||u_t||_2^2\} + o_p(1) + o_p(1) \overset{(ii)}{=} o_p(1)O_p(1) + o_p(1) + o_p(1) = o_p(1) \end{aligned} \tag{B.9}$$

where (i) follows from $\widetilde{W} \xrightarrow{p} \overline{W}$ and assumption 3; (ii) follows from assumption 3. Finally, for any $t_1 \in \mathcal{T}_1$,

$$(\widehat{P}_{t_1}^N - P_{t_1}^N)^2 \le 2||M'(\overline{W} - \widetilde{W})||_2^2||\lambda_{t_1}||_2^2 + 2||\overline{W} - \widetilde{W}||_2^2||u_t||_2^2 = o_p(1)O_p(1) = o_p(1). \tag{B.10}$$

Equations (B.8), (B.9) and (B.10) satisfy Assumptions 1,2 and 3 of Chernozhukov et al. (2021), so that an application of Chernozhukov et al. (2021)[Theorem 1] yields the result.

C Proof of Corollaries

C.1 proof of corollary 3.1

By Theorem 1, there exists a $\Lambda(\xi) > 0$ such that for any $0 < \Lambda \leq \Lambda(\xi)$, we have $|\overline{W}(\Lambda) - \overline{W}(0)| \leq \xi$. Define $\gamma := \Lambda$. Then by triangle inequality,

$$|\widetilde{W}^{BSCM}(f, \Lambda) - \overline{W}(0)| \leq |\widetilde{W}^{BSCM}(f, \Lambda) - \overline{W}(\Lambda)| + |\overline{W}(\Lambda) - \overline{W}(0)| \leq o_p(1) + \xi$$

so that the result is shown.

C.2 Proof of corollary 3.2

Consider any positive decreasing sequence $(\xi_m)_{m=1}^{\infty}$ that converges to 0. By corollary 3.1, for ξ_1, there is some $m_0(\xi_1) \in \mathbb{N}$ and $\Lambda(\xi_1) > 0$ such that for any $T \geq m_0(\xi_1)$, with probability at least $1 - \xi_1$ we have

$$|\widetilde{W}_T^{BSCM}(f, \Lambda(\xi_1)) - \overline{W}(0)| \leq \xi_1$$

Moving to ξ_2, there exists $m_0(\xi_2) > m_0(\xi_1)$ and $\Lambda(\xi_2) > 0$ such that for any $T \geq m_0(\xi_2)$, with probability at least $1 - \xi_2$,

$$|\widetilde{W}_T^{SC}(\Lambda(\xi_1)) - W^*| \leq \xi_2$$

We can express this recursively, by taking $\Lambda_T \equiv \Lambda(\xi_1)$ for $T = 1, ..., m_0(\xi_2)$, $\Lambda_T \equiv \Lambda(\xi_2)$ for $T = m_0(\xi_2) + 1, ..., m_0(\xi_3)$, so on and so forth. Then we see that the first part of the result holds.

For the second part, when Ω_0 is positive-definite, simply apply Theorem 1 and note that γ can be taken to be zero in this case, i.e. for any sequence of $\Lambda_T \downarrow 0$,

$$\begin{aligned}
\left|\widetilde{W}_T^{BSCM}(f, \Lambda_T) - \overline{W}(0)\right| &\leq \left|\widetilde{W}_T^{BSCM}(f, \Lambda_T) - \overline{W}(\Lambda_T)\right| + \left|\overline{W}(\Lambda_T) - \overline{W}(0)\right| \\
&\leq \sup_{\Lambda \in [0,1]} \left|\widetilde{W}_T^{BSCM}(f, \Lambda) - \overline{W}(\Lambda)\right| + \left|\overline{W}(\Lambda_T) - \overline{W}(0)\right| \\
&= o_p(1) + o(1) = o_p(1)
\end{aligned}$$

C.3 Proof of Corollary 3.3

(i) follows from (3.6), noting $\mathbb{E}[\lambda_t] = 0$ under the stationary case. (ii) follows from Theorem 2 and (3.6), which implies that $\overline{W}^0(0) \in \Phi$, i.e. $M_0 - M\overline{W}^0(0) = 0$.

C.4 Proof of Corollary 3.4

By Theorem 2 we know that $\widetilde{W}_T^{DBSCM}(f, \Lambda_T) \xrightarrow{p} \overline{W}^0(0)$, so that by Theorem 4 it suffices to show that $\overline{W}^0(0) \in \Phi$ whenever $\exists W^* \in \Phi$, but this follows again from Theorem 2.