Supplementary Materials to "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods"

October 27, 2019

The supplementary materials contain seven appendices with detailed proofs, additional numerical simulations and more robustness checks for the empirical application. The supplementary Appendices A - G are only for referees' convenience, and they are not for publication. They will be made available online to all readers.

Appendix A: Proofs of Theorems 3.1, 3.2 and 4.1

A.1 Proof of Theorem 3.1

The constrained estimator is defined by

$$\hat{\beta}_{T_1} = \arg\min_{\beta\in\Lambda}(\beta - \hat{\beta}_{OLS})'(X'X/T_1)(\beta - \hat{\beta}_{OLS}). \tag{A.1}$$

Thus, $\hat{\beta}_{T_1}$ is the projection of $\hat{\beta}_{OLS}$ onto Λ with respect to the norm $\|a\| = \sqrt{a'(X'X/T_1)a}$ which is random, rendering the theory in Fang and Santos (2018) not directly applicable. However, since $X'X/T_1 \xrightarrow{p} E(X_tX_t')$, we show that one can replace $X'X/T_1$ by $E(X_tX_t')$ without affecting the asymptotic results. Define the following "infeasible estimator" (it is infeasible because $E(X_t'X_t)$ is unknown in practice):

$$\tilde{\beta}_{T_1} = \arg\min_{\beta\in\Lambda}(\beta - \hat{\beta}_{OLS})'E(X_tX_t')(\beta - \hat{\beta}_{OLS}) = \Pi_\Lambda\hat{\beta}_{OLS}, \tag{A.2}$$

where Π_Λ is the projection onto Λ with respect to the norm $\|a\| = \sqrt{a'E(X_tX_t')a}$, i.e., $\Pi_\Lambda\beta = \arg\min_{\lambda\in\Lambda}(\beta-\lambda)'E(X_tX_t')(\beta-\lambda)$. By Lemma 4.6 of Zarantonello (1971) and Proposition 4.1 of Fang and Santos (2018), we know that

$$\begin{aligned}
\sqrt{T_1}(\tilde{\beta}_{T_1} - \beta_0) &= \sqrt{T_1}(\Pi_\Lambda\hat{\beta}_{OLS} - \Pi_\Lambda\beta_0) \\
&= \sqrt{T_1}\Pi_{T_{\Lambda,\beta_0}}(\hat{\beta}_{OLS} - \beta_0) + o_p(1) \\
&= \Pi_{T_{\Lambda,\beta_0}}\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0) + o_p(1) \\
&\xrightarrow{d} \Pi_{T_{\Lambda,\beta_0}} Z_1, \qquad \text{(A.3)}
\end{aligned}$$

where the first equality follows from $\tilde{\beta}_{T_1} = \Pi_\Lambda\hat{\beta}_{OLS}$ and $\beta_0 \in \Lambda$ so that $\beta_0 = \Pi_\Lambda\beta_0$.

We give some explanations of the above derivations. Hilbert Space projection onto convex sets was studied by Zarantonello (1971) and extended to general econometric model settings by Fang and Santos (2018). The projection operator Π_Λ: $\mathcal{R}^N \to \Lambda$ (Λ is a convex subset in $\mathcal{R}^N$) can be viewed as a functional mapping. Zarantonello (1971) showed that Π_Λ is (Hadamard) directional differentiable, and its directional derivative at $\beta_0 \in \Lambda$ is $\Pi_{T_{\Lambda,\beta_0}}$, which is the projection onto the tangent cone of Λ at β_0. Hence, the second equality of (A.3) follows from a functional Taylor expansion, the third equality follows from the fact that T_{Λ,β_0} is positive homogenous of degree one, i.e., for $\alpha \geq 0$, $\alpha\, T_{\Lambda,\beta_0}\,\theta = T_{\Lambda,\beta_0}\,\alpha\,\theta$ for all $\theta \in \mathcal{R}^N$, and the last line follows from $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0) \xrightarrow{d} Z_1$ and the continuous mapping theorem because projection is a continuous mapping.

We can see that the term 'tangent cone' is analogous to referring to the derivative of a function at a given point as a 'tangent line' of the function (at the given point). Now, the functional derivative of the mapping Π_Λ is a projection onto the cone $\Pi_{T_{\Lambda,\beta_0}}$ (rather than a line). Therefore, it is called the 'tangent cone' of Λ at β_0 and is denoted as T_{Λ,β_0}. For readers' convenience, we give the formal definition of tangent cone of Λ at $\theta \in \mathcal{R}^N$ below:

$$T_{\Lambda,\theta} = \overline{\cup_{\alpha\geq 0}\alpha\{\Lambda - \Pi_\Lambda\theta\}}, \qquad \text{(A.4)}$$

where for any set $A \in \mathcal{R}^N$, $\overline{A}$ is the closure of A ($\overline{A}$ is the smallest closed set that contains A).

Using the above definition one can easily check that for our synthetic control estimation problem, the tangent cone of Λ at β_0 is the same as the asymptotic range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$.

In Lemma C.1 of Appendix C, we show that

$$\hat{\beta}_{T_1} = \tilde{\beta}_{T_1} + o_p(T_1^{-1/2}) = \Pi_\Lambda \hat{\beta}_{OLS} + o_p(T_1^{-1/2}). \tag{A.5}$$

Theorem 3.1 follows from (A.3) and (A.5).

A.2 Proof of Theorem 3.2

First, we write $\hat{A} = \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$ defined in (4.2) as $\hat{A} = \hat{A}_1 + \hat{A}_2$, where

$$\hat{A}_1 = -\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'\right]\sqrt{\frac{T_2}{T_1}}\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0), \qquad \hat{A}_2 = \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t}. \tag{A.6}$$

We know that $\hat{A}_2 \xrightarrow{d} Z_2$ by assumption 2, where Z_2 is distributed as $N(0, \Sigma_v)$. By Theorem 3.1 and assumption 1, we have $\hat{A}_1 \xrightarrow{d} A_1 = -\phi E(x_t')\Pi_{T_{\Lambda,\beta_0}} Z_1$, where $\phi = \lim_{T_1,T_2\to\infty}\sqrt{T_2/T_1}$ and Z_1 is the weak limit of $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0)$, i.e., $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0) \xrightarrow{d} Z_1$. Also, by Lemma A.1 and Theorem 3.2 of Li and Bell (2017), we know that Z_1 and Z_2 are asymptotically independent with each other. This implies that $A_1 = -\phi E(x_t)\Pi_{T_{\Lambda,\beta_0}} Z_1$ is asymptotically independent of Z_2. Hence, we have $\hat{A} \xrightarrow{d} -\phi E(x_t')\Pi_{T_{\Lambda,\beta_0}} Z_1 + Z_2$.

A.3 Proof of Theorem 4.1

The proof that $\hat{A}^*$ can be used to approximate the distribution of $\hat{A}$ consists of the following arguments. First, we show that one can consistently estimate Σ_v by $\hat{\Sigma}_v = T_2^{-1}\sum_{t=T_1+1}^{T}\hat{v}_{1t}^2$ (when v_{1t} is serially uncorrelated), where $\hat{v}_{1t} = \hat{\Delta}_{1t} - \hat{\Delta}_1$. From $\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^0 = x_t'(\beta_0 - \hat{\beta}_{T_1}) + \Delta_{1t} + u_{1t} = \Delta_{1t} + u_{1t} + O_p(T_1^{-1/2})$ and $\hat{\Delta}_1 = \bar{x}'(\beta_0 - \hat{\beta}_{T_1}) + \bar{\Delta}_1 + \bar{u}_1 = \Delta_1 + O_p(T_1^{-1/2} + T_2^{-1/2})$, we have $\hat{\Sigma}_v = \frac{1}{T_2}\sum_{t=T_1+1}^{T}(\Delta_{1t} + u_{1t} - \Delta_1)^2 + O_p(T_1^{-1/2} + T_2^{-1/2}) = \Sigma_v + O_p(T_1^{-1/2} + T_2^{-1/2})$.

Next, it follows that $T_2^{-1/2}\sum_{t=T_1+1}^{T} v_{1t}^* \overset{d}{\sim} T_2^{-1/2}\sum_{t=T_1+1}^{T} v_{1t} \xrightarrow{d} Z_2$, where $A \overset{d}{\sim} B$ means that A and B have the same asymptotic distribution. By the conditions that $m \to \infty$, $m/T_1 \to 0$ as $T_1 \to \infty$ and the weak convergence result of Theorem 3.1, we know that $\sqrt{m}(\hat{\beta}_m^* - \hat{\beta}_{T_1}) \overset{d}{\sim} \sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ by Theorem 2.2.1 of Politis, Romano, and Wolf (1999). It follows that $\hat{A}^*$ defined in (4.3) and $\hat{A}$ defined in (4.2) have the same asymptotic distribution.

Appendix B: Uniqueness of the SC (MSC) estimator

B.1 A projection of the unconstrained estimator

We write the regression model in matrix form: $Y = X\beta_0 + u$, where Y and u are both $T_1 \times 1$ vectors, X is of dimension $T_1 \times N$ and has a full column rank, and β_0 is of dimension $N \times 1$. We assume that the true parameter $\beta_0 \in \Lambda$, where Λ is a closed and convex set ($\Lambda = \Lambda_{SC}$ or Λ_{MSC} in our applications). We denote the constrained least squares estimator as $\hat{\beta}_{T_1}$, i.e.,

$$\hat{\beta}_{T_1} = \arg\min_{\beta\in\Lambda}(Y - X\beta)'(Y - X\beta) \equiv \arg\min_{\beta\in\Lambda} \|Y - X\beta\|^2,$$

where $||A||^2 = A'A$ for a vector A.

We denote the unconstrained least squares estimator as $\hat{\beta}_{OLS} = arg\min_{\beta\in\mathcal{R}^N}(Y - X\beta)'(Y - X\beta)$, i.e., $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$. By the definition of $\hat{\beta}_{OLS}$, we may write $Y = X\hat{\beta}_{OLS} + \hat{u}$, where $\hat{u} = Y - X\hat{\beta}_{OLS}$. It follows that

$$\begin{aligned} f(\beta) &\stackrel{def}{=} \|Y - X\beta\|^2 \\ &= \|X(\hat{\beta}_{OLS} - \beta) + \hat{u}\|^2 \\ &= \|X(\hat{\beta}_{OLS} - \beta)\|^2 + \|\hat{u}\|^2 \\ &\equiv (\hat{\beta}_{OLS} - \beta)'X'X(\hat{\beta}_{OLS} - \beta) + \|\hat{u}\|^2, \end{aligned} \tag{B.1}$$

where we dropped a cross term in the third equality because $\hat{u}'X = 0$ (least squares residual $\hat{u}$ is orthogonal to X). Since $\|\hat{u}\|^2$ is unrelated to β, the minimizer of $f(\beta)$ is identical to the minimizer of $(\hat{\beta}_{OLS} - \beta)'X'X(\hat{\beta}_{OLS} - \beta)$. Thus, we have

$$\begin{aligned} \hat{\beta}_{T_1} &= \arg\min_{\beta\in\Lambda}(\hat{\beta}_{OLS} - \beta)'X'X(\hat{\beta}_{OLS} - \beta) \\ &= \arg\min_{\beta\in\Lambda}(\hat{\beta}_{OLS} - \beta)'(X'X/T_1)(\hat{\beta}_{OLS} - \beta) \\ &= \arg\min_{\beta\in\Lambda} ||\hat{\beta}_{OLS} - \beta||_X^2, \end{aligned}$$

where the second equality follows since $T_1 > 0$.

B.2 The uniqueness of the (modified) synthetic control estimator

We first give the definition of a strictly convex function. A function f is said to be *strictly convex* if $f(\alpha x+(1-\alpha)y)) < \alpha f(x)+(1-\alpha)f(y)$ for all $0<\alpha<1$ and for all $x \neq y$, $x, y \in D$, where D is the domain of f.

Under the assumption that the data matrix $X_{T_1 \times N}$ has a full column rank, we show below that $f(\beta) \stackrel{def}{=} \sum_{t=1}^{T_1}(y_{1t} - x_t'\beta)^2$ is a strictly convex function. Since the objective function is a convex function and the constrained domains for β, Λ_{SC} and Λ_{MSC}, are convex sets, then the constrained minimization problem has a unique (global) minimizer. To see this, we argue by contradiction. Suppose that we have two local minimizers $z_1 \neq z_2$. Then for any convex combination $z_3 = \alpha z_1 + (1-\alpha)z_2$, we have $f(z_3) < \alpha f(z_1) + (1-\alpha)f(z_2)$ for all $\alpha \in (0,1)$. This contradicts the fact that z_1 and z_2 are two minimizers. Hence, we must have $z_1 = z_2$ and the minimizer is unique.

It remains to show that $f(\beta) = (\hat{\beta}_{OLS} - \beta)'X'X(\hat{\beta}_{OLS} - \beta)$ is a strictly convex function (we ignore the irrelevant constant term $\|\hat{u}\|^2$ in $f(\beta)$ defined in (B.1)). We first establish an intermediate result. For β, $\gamma \in \mathcal{R}^N$ with $\beta \neq \gamma$, because $A \equiv X'X$ is positive definite, we have

$$\begin{aligned}
0 &< (\beta-\gamma)'A(\beta-\gamma) \\
&= ((\beta-\hat{\beta}_{OLS}) - (\gamma-\hat{\beta}_{OLS}))'A((\beta-\hat{\beta}_{OLS}) - (\gamma-\hat{\beta}_{OLS})) \\
&= (\beta-\hat{\beta}_{OLS})'A(\beta-\hat{\beta}_{OLS}) + (\gamma-\hat{\beta}_{OLS})'A(\gamma-\hat{\beta}_{OLS}) - 2(\beta-\hat{\beta}_{OLS})'A(\gamma-\hat{\beta}_{OLS}) \\
&= f(\beta) + f(\gamma) - 2(\hat{\beta}_{OLS}-\beta)'A(\hat{\beta}_{OLS}-\gamma). \qquad \text{(B.2)}
\end{aligned}$$

Then for all $\alpha \in (0,1)$, we have

$$\begin{aligned}
f(\alpha\beta+(1-\alpha)\gamma) &= (\hat{\beta}_{OLS} - (\alpha\beta+(1-\alpha)\gamma))'A(\hat{\beta}_{OLS} - (\alpha\beta+(1-\alpha)\gamma)) \\
&= (\alpha(\hat{\beta}_{OLS}-\beta) + (1-\alpha)(\hat{\beta}_{OLS}-\gamma))'A(\alpha(\hat{\beta}_{OLS}-\beta) + (1-\alpha)(\hat{\beta}_{OLS}-\gamma)) \\
&= \alpha^2(\hat{\beta}_{OLS}-\beta)'A(\hat{\beta}_{OLS}-\beta) + (1-\alpha)^2(\hat{\beta}_{OLS}-\gamma)'A(\hat{\beta}_{OLS}-\gamma) \\
&\quad +2\alpha(1-\alpha)(\hat{\beta}_{OLS}-\beta)'A(\hat{\beta}_{OLS}-\gamma) \\
&= \alpha^2 f(\beta) + (1-\alpha)^2 f(\gamma) + 2\alpha(1-\alpha)(\hat{\beta}_{OLS}-\beta)'A(\hat{\beta}_{OLS}-\gamma) \\
&< \alpha^2 f(\beta) + (1-\alpha)^2 f(\gamma) + \alpha(1-\alpha)[f(\beta)+f(\gamma)] \\
&= \alpha f(\beta) + (1-\alpha) f(\gamma), \qquad \text{(B.3)}
\end{aligned}$$

where the inequality follows from (B.2). Eq. (B.3) shows that $f(\cdot)$ is a strictly convex function.

Appendix C: Three useful lemmas

In this supplementary appendix, we prove two lemmas that are used to prove Theorem 3.1.

Lemma C.1 *Under the same conditions as in Theorem 3.1, we have*

$$\hat{\beta}_{T_1} = \tilde{\beta}_{T_1} + o_p(T_1^{-1/2}) = \Pi_\Lambda \hat{\beta}_{OLS} + o_p(T_1^{-1/2}).$$

Proof: For any fixed $\epsilon > 0$, suppose that $\sqrt{T_1}\|\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}\| > \epsilon$. Then we have

$$\sqrt{T_1}(\hat{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \hat{\beta}_{OLS}) < \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS}), \tag{C.1}$$

where the strict inequality is due to uniqueness of the projection and the assumption that $\epsilon > 0$ which implies that $\hat{\beta}_{T_1} \neq \tilde{\beta}_{T_1}$. By simple algebra (adding/subtracting terms), we have:

$$\begin{aligned}
& \sqrt{T_1}(\hat{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \hat{\beta}_{OLS}) \\
= \; & \sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1} + \tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1} + \tilde{\beta}_{T_1} - \hat{\beta}_{OLS}) \\
= \; & \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS}) \\
& + \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})'(X'X/T_1)\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}) \\
& + 2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}).
\end{aligned} \tag{C.2}$$

By (C.1) and (C.2), we know that the sum of the last two terms in (C.2) is negative, i.e.,

$$\begin{aligned}
D_{T_1} \; & \stackrel{def}{=} \; \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})' \left(\frac{1}{T_1} X'X\right) \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}) + 2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})' \left(\frac{1}{T_1} X'X\right) \sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) \\
& \equiv \; D_{1,T_1} + D_{2,T_1} < 0.
\end{aligned} \tag{C.3}$$

Let $\mathcal{S}^N = \{a \in \mathcal{R}^N : \|a\| = 1\}$ denote the unit sphere in $\mathcal{R}^N$. We have

$$\begin{aligned}
D_{1,T_1} \; & = \; \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})' \left(\frac{1}{T_1} X'X\right) \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}) \\
& = \; \|\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})\|^2 \left[\frac{\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})'}{\|\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})\|} \left(\frac{1}{T_1} X'X\right) \frac{\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})}{\|\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})\|}\right]
\end{aligned}$$

$$
\begin{aligned}
&\geq T_1\|\tilde{\beta}_{T_1}-\hat{\beta}_{T_1}\|^2 \inf_{a\in\mathcal{S}^N} a'\left(\frac{1}{T_1}X'X\right)a \\
&= T_1\|\tilde{\beta}_{T_1}-\hat{\beta}_{T_1}\|^2\lambda_{\min}\left(\frac{1}{T_1}X'X\right) \\
&\geq \epsilon^2\lambda_{\min}\left(\frac{1}{T_1}X'X\right) \\
&\xrightarrow{p} \epsilon^2\lambda_{\min}[E(X_tX_t')]>0, \qquad \text{(C.4)}
\end{aligned}
$$

because $\sqrt{T_1}\|\tilde{\beta}_{T_1}-\hat{\beta}_{T_1}\|\geq\epsilon$ and $E(X_tX_t')$ is nonsingular. The minimum eigenvalue of a square matrix A is denoted by $\lambda_{\min}(A)$. The third equality uses Lemma C.2 which is proved at the end of this appendix.

By writing $(X'X/T_1)=E(X_tX_t')+(X'X/T_1)-E(X_tX_t')$, the second term in (C.3) can be rewritten as

$$
\begin{aligned}
D_{2,T_1} &= 2\sqrt{T_1}(\tilde{\beta}_{T_1}-\hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1}-\tilde{\beta}_{T_1}) \\
&= 2\sqrt{T_1}(\tilde{\beta}_{T_1}-\hat{\beta}_{OLS})'[E(X_tX_t')]\sqrt{T_1}(\hat{\beta}_{T_1}-\tilde{\beta}_{T_1}) \\
&\quad +2\sqrt{T_1}(\tilde{\beta}_{T_1}-\hat{\beta}_{OLS})'(X'X/T_1-E[X_tX_t'])\sqrt{T_1}(\hat{\beta}_{T_1}-\tilde{\beta}_{T_1}) \\
&= D_{2,1,T_1}+D_{2,2,T_1}. \qquad \text{(C.5)}
\end{aligned}
$$

By the definition of $\tilde{\beta}_{T_1}$ and Lemma 1.1 in Zarantonello (1971)

$$
D_{2,1,T_1}=\sqrt{T_1}(\tilde{\beta}_{T_1}-\hat{\beta}_{OLS})'[E(X_tX_t')]\sqrt{T_1}(\hat{\beta}_{T_1}-\tilde{\beta}_{T_1})\geq 0. \qquad \text{(C.6)}
$$

By a law of large numbers, $X'X/T_1-E(X_tX_t')=o_p(1)$. Also, $\sqrt{T_1}(\tilde{\beta}_{T_1}-\hat{\beta}_{OLS})=O_p(1)$ and $\sqrt{T_1}(\hat{\beta}_{T_1}-\tilde{\beta}_{T_1})=O_p(1)$ because

$$
\begin{aligned}
\|\sqrt{T_1}(\hat{\beta}_{T_1}-\tilde{\beta}_{T_1})\| &\leq \|\sqrt{T_1}(\hat{\beta}_{T_1}-\beta_0)\|+\|\sqrt{T_1}(\tilde{\beta}_{T_1}-\beta_0)\| \\
&= \|\sqrt{T_1}(\Pi_{\Lambda,T_1}\hat{\beta}_{OLS}-\beta_0)\|+\|\sqrt{T_1}(\Pi_{\Lambda}\hat{\beta}_{OLS}-\beta_0)\| \\
&\leq \sqrt{T_1}\|\hat{\beta}_{OLS}-\beta_0\|_{T_1}+\sqrt{T_1}\|\hat{\beta}_{OLS}-\beta_0\|=O_p(1),
\end{aligned}
$$

where we used the Lipschitz continuity of projection operators, and Π_{Λ,T_1} is the projection onto Λ with respect to the aforementioned random norm $\|a\|_{T_1}=\sqrt{a'(X'X/T_1)a}$ (Zarantonello,

1971). Hence, we have $D_{2,2,T_1} = o_p(1)$. Combining $D_{2,2,T_1} = o_p(1)$ and (C.6), we obtain

$$D_{2,T_1} \geq o_p(1). \tag{C.7}$$

Thus, we have shown that if $\sqrt{T_1}\|\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}\| > \epsilon$, then $D_{T_1} < 0$. This implies that (if A implies B, then $P(A) \leq P(B)$, this argument is used twice in (C.8) below)

$$\begin{aligned} P(\sqrt{T_1}\|\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}\| > \epsilon) &\leq P(D_{T_1} < 0) \\ &\leq P(o_p(1) + \epsilon^2 \lambda_{\min}\left(\frac{1}{T_1}X'X\right) < 0) \\ &\to P(\epsilon^2 \lambda_{\min}\left(E(X_t X_t')\right) \leq 0) \\ &= 0, \end{aligned} \tag{C.8}$$

where the second inequality above follows from $D_{T_1} = D_{1,T_1} + D_{2,T_1} \geq \epsilon^2 \lambda_{min}(X'X/T_1) + o_p(1)$ by (C.4) and (C.7). Hence, $D_{T_1} < 0$ implies $\epsilon^2 \lambda_{min}(X'X/T_1) + o_p(1) < 0$.

Equation (C.8) is equivalent to $\hat{\beta}_{T_1} - \tilde{\beta}_{T_1} = o_p(T_1^{-1/2})$ or

$$\hat{\beta}_{T_1} = \Pi_\Lambda \hat{\beta}_{OLS} + o_p(T_1^{-1/2}) \ . \tag{C.9}$$

This concludes the proof of Lemma C.1.

Lemma C.2 *Let A be an $N \times N$ positive definite matrix, and $\mathcal{S}^N = \{a \in \mathcal{R}^N : \|a\| = 1\}$ denotes the unit sphere in $\mathcal{R}^N$. Then we have* $\inf_{a \in \mathcal{S}^N} a'Aa = \lambda_{min}(A)$.

Proof: Let $v_1, ..., v_N$ be N eigen-vectors of A with corresponding eigen-values $\lambda_1, ..., \lambda_N$ so that $Av_j = \lambda_j v_j$ for $j = 1, ..., N$. Then since $v_1, ..., v_N$ form an orthonormal basis for $\mathcal{S}^N$, we have for any $a \in \mathcal{S}^N$, $a = \sum_{i=1}^N c_i v_i$ with $\sum_{i=1}^N c_i^2 = 1$ since $a'a = 1$ and $v_i'v_j = \delta_{ij}$ (the Kronecker delta). Then we have

$$\begin{aligned} a'Aa &= \sum_{i=1}^N \sum_{j=1}^N c_i v_i' A c_j v_j = \sum_{i=1}^N \sum_{j=1}^N c_i v_i' c_j A v_j = \sum_{i=1}^N \sum_{j=1}^N c_i c_j \lambda_j v_i' v_j \\ &= \sum_{i=1}^N \lambda_j c_j^2 \geq \lambda_{min} \sum_{j=1}^N c_j^2 = \lambda_{min}, \end{aligned} \tag{C.10}$$

which implies (i) $\inf_{a \in \mathcal{S}^N} a'Aa \geq \lambda_{min}$.

On the other hand, pre-multiplying $Av_j = \lambda_j v_j$ by v_j', we get $\lambda_j = v_j' A v_j \geq \inf_{a \in \mathcal{S}^N} a'Aa$ for all $j = 1, ..., N$, which implies (ii) $\lambda_{min} \geq \inf_{a \in \mathcal{S}^N} a'Aa$. Combining (i) and (ii), we finish the proof of Lemma C.2.

Lemma C.3 (stability property) *Theorem 4.1 holds uniformly locally at β_{0,T_1} as β_{0,T_1} approaches the boundary of the set Λ (as $T_1 \to \infty$).*

The theoretical result presented in Theorem 4.1 is pointwise. That is, Theorem 4.1 holds true for a fixed vector $\beta_0 \in \Lambda$. However, one may be concerned whether Theorem 4.1 also holds uniformly locally at β_{0,T_1} as β_{0,T_1} approaches the boundary of the set Λ (as $T_1 \to \infty$). If the limiting distribution of $\hat{A}$ depends discontinuously on β_0 when β_0 is at the boundary of Λ, then the test may fail to adequately control for size when β_0 is close to the boundary of Λ. In the case of the MSC method, β_0 is at the boundary of Λ if $\beta_{0,j} = 0$ for some $2 \leq j \leq N$. To examine this issue we consider a sequence of distributions in the form of $\beta_{0,T_1} = \beta_0 + c/\sqrt{T_1} \in \Lambda$, where $\beta_{0,j} \geq 0$ for all $j \in \{2, ..., N\}$, $\beta_{0,i} = 0$ for at least one $i \in \{2, ..., N\}$, and $c_j \geq 0$ for all $j \in \{2, ..., N\}$. By Proposition 4.2 of Fang and Santos (2018), we know that the projection mapping (on to Λ) is convex. Then by Corollary 3.2 or Corollary S.1.1 of the supplementary appendix (for general dependent data case) of Fang and Santos (2018), we know that under the above null hypothesis H_0,

$$\limsup_{T_1 \to \infty} P_{\beta_0 + c/\sqrt{T_1}} \left(\hat{A} > \hat{c}_{1-\alpha} \right) \leq \alpha, \tag{C.11}$$

where $\hat{A}$ is defined in (4.2), and $P_{\beta_0 + c/\sqrt{T_1}}$ indicates the distribution of the data associated with $\beta = \beta_0 + c/\sqrt{T_1} \in \Lambda$ and that β_0 is at the boundary of Λ. Equation (C.11) proves Lemma C.3 and it implies that our analysis delivers inference procedures with reliable size control.

Appendix D: Asymptotic theory with non-stationary data

D.1 The trend stationary data

The trend-stationary data generating process can also be motivated using a factor model framework. Let $\{y_{it}^0\}$, for $i = 1, ..., N$ and $t = 1, ..., T$, be generated by some common factors with one of the factors being a time trend and the remaining factors being weakly dependent stationary

variables. Following Hsiao, Ching, and Wan (2012), we assume that $y_t^0 = (y_{1t}^0, y_{2t}^0, ..., y_{Nt}^0)'$ is generated via a factor model

$$y_t^0 = \delta_0 + Bf_t + \epsilon_t, \tag{D.1}$$

where $\delta_0 = (\delta_{01}, ..., \delta_{0N})'$ is an $N \times 1$ vector of intercepts, B is an $N \times K$ factor loading matrix, $f_t = (f_{1t}, ..., f_{Kt})'$ is a $K \times 1$ vector of common factors, and $\epsilon_t = (\epsilon_{1t}, ..., \epsilon_{Nt})'$ is an $N \times 1$ vector of idiosyncratic errors. We assume that $f_{1t} = t$ and all other factors are stationary variables. Also, ϵ_t is a zero mean, weakly dependent stationary process with finite fourth moment. Hence, y_t^0 follows a trend-stationary process.

Hsiao, Ching, and Wan (2012) and Li and Bell (2017) show that, under the condition that $rank(B) = K$, one can replace the unobservable factor f_t by $x_t = (1, y_{2t}, ..., y_{Nt})'$ to estimate the counterfactual outcome y_{1t}^0. Specifically, one can estimate the following regression model

$$y_{1t} = x_t'\delta + u_{1t}, \qquad (t = 1, ..., T_1), \tag{D.2}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})'$ and $\delta = (\delta_1, ..., \delta_N)'$.

To facilitate the asymptotic analysis, we consider the time trend component explicitly. We write $y_{jt} = c_{0,j} + c_{1,j}t + \eta_{jt}$, where η_{jt} is a weakly dependent stationary process (de-trended from y_{jt}) for $j = 2, ..., N$. Let $\tilde{y}_t = (y_{2t}, ..., y_{Nt})'$ and $\tilde{\delta} = (\delta_2, ..., \delta_N)'$. Then in vector notation, we have $\tilde{y}_t = \tilde{c}_0 + \tilde{c}_1 t + \tilde{\eta}_t$, $\tilde{c}_0 = (c_{0,2}, ..., c_{0,N})'$, $\tilde{c}_1 = (c_{1,2}, ..., c_{1,N})'$ and $\tilde{\eta} = (\eta_{2t}, ..., \eta_{Nt})'$. Then we can write $\tilde{y}_t'\tilde{\delta} = (\tilde{c}_0 + \tilde{c}_1 t + \tilde{\eta}_t)'\tilde{\delta}$. Hence, we can re-write (D.2) as

$$\begin{aligned} y_{1t} &= \delta_1 + \tilde{y}_t'\tilde{\delta} + u_{1t} \\ &= \alpha_0 t + \beta_1 + \tilde{\delta}'\tilde{\eta}_t + u_{1t} \\ &= \alpha_0 t + z_t'\beta_0 + u_{1t} \qquad\qquad t = 1, ..., T_1, \end{aligned} \tag{D.3}$$

where $\alpha_0 = \tilde{c}_1'\tilde{\delta}$, $\beta_1 = \delta_1 + \tilde{c}_0'\tilde{\delta}$, $\beta_0 = (\beta_1, \tilde{\delta}')'$ and $z_t = (1, \tilde{\eta}')' \equiv (1, \eta_{2t}, ..., \eta_{Nt})'$.

Below we derive the asymptotic distribution of the ATE estimator $\hat{\Delta}_1$ defined in (3.7). For the post-treatment period, we have $y_{1t}^1 = y_{1t}^0 + \Delta_{1t}$. Hence, we have for $t = 1, ..., T$,

$$y_{1t} = \alpha t + z_t'\beta + d_t\Delta_{1t} + v_{1t}, \tag{D.4}$$

where $d_t = 0$ for $t \leq T_1$ and $d_t = 1$ for $t \geq T_1 + 1$.

Let $\hat{\alpha}_{T_1}$ and $\hat{\beta}_{T_1}$ be the SC/MSC estimators of α_0 and β_0 based on (D.3). Then it is to show that $\hat{\alpha}_{T_1} - \alpha = O_p(T_1^{-3/2})$ and $\hat{\beta}_{T_1} - \beta = O_p(T_1^{-1/2})$. Thus, using (3.7) and (D.4), we have

$$\begin{aligned}
\hat{\Delta}_1 - \Delta_1 &= \frac{1}{T_2} \sum_{t=T_1+1}^{T} \left[y_{1t} - \hat{y}_{1t}^0 \right] - \Delta_1 \\
&= -\frac{1}{T_2} \sum_{t=T_1+1}^{T} \left[(\hat{\alpha}_{T_1} - \alpha_0)t - z_t'(\hat{\beta}_{T_1} - \beta_0) + \Delta_{1t} - \Delta_1 + v_{1t} \right] \\
&= -\left[\frac{2T_1 + T_2 + 1}{2} \right] (\hat{\alpha}_{T_1} - \alpha) - [E(z_t') + o_p(1)](\hat{\beta}_{T_1} - \beta) + \frac{1}{T_2} \sum_{t=T_1+1}^{T} v_{1t}, \quad \text{(D.5)}
\end{aligned}$$

where we used $\sum_{t=T_1+1}^{T} t = (T_1 + 1 + T)T_2/2 = (2T_1 + T_2 + 1)T_2/2$ and $v_{1t} = \Delta_{1t} - \Delta_1 + u_{1t}$.

Hence,

$$\begin{aligned}
\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) &= -\sqrt{T_2/T_1} \left[\frac{2 + T_2/T_1}{2} \right] \sqrt{T_1^3}(\hat{\alpha}_{T_1} - \alpha_0) - \sqrt{T_2/T_1} E(z_t') \sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) \\
&\quad + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\
&= -\left(\sqrt{T_2/T_1}(2 + T_2/T_1)/2 \,,\ \sqrt{T_2/T_1} E(z_t') \right) \begin{pmatrix} \sqrt{T_1^3}(\hat{\alpha}_{T_1} - \alpha_0) \\ \sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) \end{pmatrix} \\
&\quad + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\
&= -c' M_{T_1}(\hat{\gamma}_{T_1} - \gamma_0) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} + o_p(1), \quad \text{(D.6)}
\end{aligned}$$

where $c = (\sqrt{\phi}(2 + \phi)/2, \sqrt{\phi}E(z_t'))'$, $\phi = \lim_{T_1, T_2 \to \infty} T_2/T_1$, $\hat{\gamma}_{T_1} = (\hat{\alpha}_{T_1}, \hat{\beta}_{T_1}')'$, $\gamma_0 = (\alpha_0, \beta_0')'$, $M_{T_1} = \sqrt{T_1} diag(T_1, 1, ..., 1)$ which is a $(N+1) \times (N+1)$ diagonal matrix with the first diagonal element equal to $T_1^{3/2}$ and all other diagonal elements equal to $\sqrt{T_1}$.

To establish the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$, we make the following assumptions.

Assumption D1. Let $z_t = (1, \eta_{2t}, ..., \eta_{Nt})'$. We assume that (i) $\{z_t\}_{t=1}^{T}$ is a weakly dependent and weakly stationary process, $T_1^{-1} \sum_{t=1}^{T_1} z_t z_t' \xrightarrow{p} E(z_t z_t')$ as $T_1 \to \infty$, and $[E(z_t z_t')]$ is invertible; (ii) $M_{T_1}(\hat{\gamma}_{OLS} - \gamma) \xrightarrow{d} N(0, \Omega)$, where Ω is a positive definite matrix.

Assumption D2. Let $v_{1t} = \Delta_{1t} - \Delta_1 + v_{1t}$. Then $T_2^{-1/2} \sum_{t=T_1+1}^{T} v_{1t} \xrightarrow{d} N(0, \Sigma_v)$ as $T_2 \to \infty$, where $\Sigma_v = \lim_{T_2 \to \infty} T_2^{-1} \sum_{t=T_1+1}^{T} \sum_{s=T_1+1}^{T} E(v_{1t} v_{1s})$.

Assumption D3. Let $w_t = (v_{1t}, \eta_{2t}, ..., \eta_{Nt})'$. We assume that w_t is a ρ-mixing process where the mixing coefficient $\rho(\tau)$ satisfies the condition: $\rho(\tau) \leq C\,\lambda^{\tau}$ for some finite positive constants $C > 0$ and $0 < \lambda < 1$, where $\rho(\tau) = \max_{1\leq i,j\leq N} |Cov(w_{it}, w_{j,t+\tau})|/\sqrt{Var(w_{it})Var(w_{j,t+\tau})}$, and w_{it} is the i^{th} component of w_t for $i = 1, ..., N$.

Assumptions D1 and D2 are not restrictive. They require that (z_t, v_{1t}) be a weakly dependent stationary process so that law of large numbers and central limit theorem hold for their (partial) sums. If $E(z_t z_t')$ is not invertible, we can remove the linearly dependent regressors and redefine z_t as a subset of $(1, \eta_{2t}, ..., \eta_{Nt})'$ such that assumption 1 holds. Assumption D3 further imposes an exponential decay rate for the ρ-mixing processes. Many ARMA processes are known to be ρ-mixing with exponential decay rate.

By Appendix B.1 of this supplementary Appendix, we know that ($\Lambda = \Lambda_{SC}$ or $\Lambda = \Lambda_{MSC}$)

$$\hat{\gamma}_{T_1} = \arg\min_{\gamma\in\Lambda}(\gamma - \hat{\gamma}_{OLS,T_1})'X'X(\gamma - \hat{\gamma}_{OLS,T_1}) \equiv \arg\min_{\gamma\in\Lambda} A(\gamma), \tag{D.7}$$

where $A(\gamma) = (\gamma - \hat{\gamma}_{OLS,T_1})'X'X(\gamma - \hat{\gamma}_{OLS,T_1})$, X is the $T_1 \times (N+1)$ matrix with its t^{th}-row given by (t, z_t'), and $z_t = (\eta_{2t}, ..., \eta_{Nt})'$.

Our derivation below is based on Andrews (1999). By adding/subtracting γ_0 and inserting identity matrix $I_{N+1} = M_{T_1}M_{T_1}^{-1}$, we can write $A(\gamma)$ as:

$$\begin{aligned}
A(\gamma) &= (\gamma - \hat{\gamma}_{OLS,T_1})'X'X(\gamma - \hat{\gamma}_{OLS,T_1}) \\
&= [(\gamma - \gamma_0) - (\hat{\gamma}_{OLS,T_1} - \gamma_0)]'X'X[(\gamma - \gamma_0) - (\hat{\gamma}_{OLS,T_1} - \gamma_0)] \\
&= [(\gamma - \gamma_0) - (\hat{\gamma}_{OLS,T_1} - \gamma_0)]'M_{T_1}M_{T_1}^{-1}X'XM_{T_1}^{-1}M_{T_1}[(\gamma - \gamma_0) - (\hat{\gamma}_{OLS,T_1} - \gamma_0)] \\
&= \{M_{T_1}[(\gamma - \gamma_0) - (\hat{\gamma}_{OLS,T_1} - \gamma_0)]\}'[M_{T_1}^{-1}X'XM_{T_1}^{-1}]\{M_{T_1}[(\gamma - \gamma_0) - (\hat{\gamma}_{OLS,T_1} - \gamma_0)]\} \\
&= [M_{T_1}(\gamma - \gamma_0) - Z_{T_1}]'\, J_{T_1}\, [M_{T_1}(\gamma - \gamma_0) - Z_{T_1}] \\
&= [\lambda_{T_1} - Z_{T_1}]'\, J_{T_1}\, [\lambda_{T_1} - Z_{T_1}],
\end{aligned} \tag{D.8}$$

where the fourth equality follows from $(AB)' = B'A'$, $Z_{T_1} = M_{T_1}(\hat{\gamma}_{OLS,T_1} - \gamma_0)$, $\lambda_{T_1} = M_{T_1}(\gamma - \gamma_0)$, $J_{T_1} = M_{T_1}^{-1}X'XM_{T_1}^{-1}$.

We know that (Hamilton, 1994) that $Z_{T_1} \xrightarrow{d} Z_3$, where Z_3 is a zero mean, finite variance, $(N+1) \times 1$ vector of normal random variable. It is easy to show that $J_{T_1} \xrightarrow{p} J_{tr}$, where J_{tr} is

an $(N+1)\times(N+1)$ positive definite matrix defined by

$$J_{tr} = \begin{pmatrix} 1/3 & (1/2)E(z_t') \\ (1/2)E(z_t) & E(z_t z_t') \end{pmatrix}. \tag{D.9}$$

From (D.8) we can see that choosing $\gamma \in \Lambda$ to minimize $A(\gamma)$ is equivalent to choosing $\lambda_{T_1} = M_{T_1}(\gamma - \gamma_0) \in M_{T_1}(\Lambda - \gamma_0) \to T_{\Lambda,\gamma_0}$ as $T_1 \to \infty$ (T_{Λ,γ_0} is the tangent cone of Λ at γ_0) to minimize $A(\gamma)$.

Since $\hat{\gamma}_{T_1}$ minimizes $A(\gamma)$ subject to $\gamma \in \Lambda$, we know that $\hat{\lambda}_{T_1} \overset{def}{=} M_{T_1}(\hat{\gamma}_{T_1} - \gamma_0)$ also minimizes $A(\gamma)$ subject to $\hat{\lambda}_{T_1} \in M_{T_1}(\Lambda - \gamma_0)$. Hence, if we take the limit of $T_1 \to \infty$ and let $\hat{\lambda}$ denote the limiting distribution of $\hat{\lambda}_{T_1}$, because $Z_{T_1} \xrightarrow{d} Z_3$ and $J_{T_1} \xrightarrow{d} J_{tr}$, we see that $\hat{\lambda}$ satisfies that

$$\hat{\lambda} = \arg\min_{\lambda \in T_{\Lambda,\gamma_0}} (\lambda - Z_3)' J_{tr} (\lambda - Z_3) \overset{def}{=} \Pi^{tr}_{T_{\Lambda,\gamma_0}} Z_3, \tag{D.10}$$

where Z_3 is the limiting distribution of $\hat{\lambda}_{T_1} = M_{T_1}(\hat{\gamma}_{OLS,T_1} - \gamma_0)$. Note that the last equal sign in (D.10) defines a projection. That is, for the time trend model, the projection of $\theta \in \mathcal{R}^{N+1}$ onto a convex set Λ is defined as

$$\Pi^{tr}_{\Lambda}\theta = \arg\min_{\lambda \in \Lambda} (\lambda - \theta)' J_{tr} (\lambda - \theta). \tag{D.11}$$

Thus, we just showed that

$$\hat{\lambda}_{T_1} \overset{def}{=} M_{T_1}(\hat{\gamma}_{T_1} - \gamma_0) \xrightarrow{d} \hat{\lambda} = \Pi^{tr}_{T_{\Lambda,\gamma_0}} Z_3. \tag{D.12}$$

By Assumption D3 and the proof of Theorem 3.2 and Lemma 1 in Li and Bell (2017), we know that $\hat{\gamma} - \gamma$ is asymptotic independent with $T_2^{-1/2} \sum_{t=T_1+1}^{T} v_{1t}$. Therefore, applying the projection theory to (D.6) we immediately have the following result.

Under assumptions D1 to D3 and noting that $\gamma_0 \in \Lambda$, we have

$$\begin{aligned} \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) &= -c' M_{T_1}(\hat{\gamma}_{T_1} - \gamma_0) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} \\ &\xrightarrow{d} -c' \Pi^{tr}_{T_{\Lambda,\gamma_0}} Z_3 + Z_2, \end{aligned} \tag{D.13}$$

by (D.12), where Z_3 is the weak limit of $M_{T_1}(\hat{\gamma}_{OLS} - \gamma_0)$ as described in Assumption C1, and Z_2 is independent with Z_3 and is normally distributed with a zero mean and variance Σ_v.

D.2 The unit-root non-stationary data

Here we only consider unit-root processes without drifts because the asymptotic theory for a unit-root process a drift is the same as the trend-stationary data case due to the fact that the drift term leads to a time trend component which dominates other components. Therefore, we assume that, in the absence of treatment, the outcome variables follow unit-root processes without drifts:

$$y_{jt}^0 = y_{j,t-1}^0 + \eta_{jt}, \qquad j = 1, ..., N; t = 1, ...T,$$

where η_{jt} is a zero mean, weakly dependent stationary process that satisfies Assumption D4 below.

Define $\tilde{x}_t = (y_{2t}, ..., y_{Nt})'$. Then we have $x_t = (1, y_{2t}, ..., y_{Nt})' = (1, \tilde{x}_t')'$. We assume that

Assumption D4.

(i) $T^{-2}\sum_{t=1}^{T} \tilde{x}_t\tilde{x}_t' \xrightarrow{d} \int_0^1 W_\eta(r)W_\eta(r)'dr \equiv W_{\eta,2}$, where $W_\eta(r) = V_\eta B_\eta(r)$, B_η is a $(N-1)\times 1$ vector of standard Brownian motion, $V_\eta = \Sigma_\eta^{1/2}$ and $\Sigma_\eta = \lim_{T_1\to\infty} T_1^{-1}\sum_{t=1}^{T_1}\sum_{s=1}^{T_1} E(\eta_{it}\eta_{js}')$.

(ii) $T^{-3/2}\sum_{t=1}^{T} \tilde{x}_t \xrightarrow{d} \int_0^1 W_\eta(r)dr \equiv W_{\eta,1}$.

(iii) $T^{-1}\sum_{t=1}^{T} \tilde{x}_t u_{1t} \xrightarrow{d} \int_0^1 W_\eta(r)dW_u(r) \equiv W_{\eta,u}$, where $W_u(r) = V_u B_u(r)$, B_u is a (scalar) standard Brownian motion generated by partial sum of u_{1t}'s (B_u is independent of B_η), $V_u = \Sigma_u^{1/2}$ with $\Sigma_u = \lim_{T_1\to\infty} T_1^{-1}\sum_{t=1}^{T_1}\sum_{s=1}^{T_1} E(u_{1t}u_{1s})$.

(iv) $T_1^{-1/2}\sum_{t=1}^{T_1} u_{1t} \xrightarrow{d} W_u(1)$.

(v) $T_2^{-1/2}\sum_{t=T_1+1}^{T} v_{1t} \xrightarrow{d} N(0, \Sigma_v)$, where $\Sigma_v^2 = \lim_{T_2\to\infty} T_2^{-1}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1}^{T} E(v_{1t}v_{1s})$.

Assumption D5. (i) The convergence results presented at Assumption D4 hold jointly, then by the continuous mapping theorem we have

$$D_{T_1}(\hat{\beta}_{OLS} - \beta_0) \xrightarrow{d} \begin{pmatrix} 1 & W_{\eta,1}' \\ W_{\eta,1} & W_{\eta,2} \end{pmatrix}^{-1} \begin{pmatrix} W_u(1) \\ W_{\eta,u} \end{pmatrix} \equiv Z_4, \tag{D.14}$$

where $D_{T_1} = T_1\, Diag(T_1^{-1/2}, 1, ..., 1)$ is the $N \times N$ diagonal matrix defined in Section 3.4.

(ii) Let $w_t = (v_{1t}, \eta_{1t}, ..., \eta_{Nt})'$. We assume that w_t is a ρ-mixing process with the mixing coefficient $\rho(\tau)$ satisfies the condition: $\rho(\tau) \le C\,\lambda^\tau$ for some finite positive constants $C > 0$ and $0 < \lambda < 1$, where $\rho(\tau) = \max_{1\le i,j\le N} |Cov(w_{it}, w_{j,t+\tau})|/\sqrt{Var(w_{it})Var(w_{j,t+\tau})}$, and w_{it} is the i^{th} component of w_t for $i = 1, ..., N$.

Remark D.1 *Co-integration theory is well developed in the literature. Primitive conditions that ensure that Assumption D4 and D5 (i) hold can be found in many published papers, e.g., Stock and Watson (1993).*

Recall that $\phi = \lim_{T_1,T_2\to\infty} T_2/T_1$. It can be shown that when $\phi = 0$, $\hat{A}_1 = o_p(1)$. Therefore, we only need to consider the case that $\phi > 0$. By Assumption D4 (ii) and noting that $(T_1 + 1)/T \approx T_1/T = (T/T_1)^{-1} \to (1+\phi)^{-1}$, we get

$$\frac{1}{T^{3/2}} \sum_{t=T_1+1}^{T} \tilde{x}_t \quad \xrightarrow{d} \quad \int_{1/(1+\phi)}^{1} W_\eta(r)dr. \tag{D.15}$$

For the unit-root data process, define $J_{I,T_1} = D_{T_1}^{-1}(X'X)D_{T_1}^{-1}$, then we have

$$J_{I,T_1} \xrightarrow{d} J_I \equiv \begin{pmatrix} 1 & \int_0^1 W_\eta(r)'dr \\ \int_0^1 W_\eta(r)dr & \int_0^1 W_\eta(r)W_\eta(r)'dr \end{pmatrix}, \tag{D.16}$$

because

$$\begin{aligned} J_{I,T_1} &= D_{T_1}^{-1}(X'X)D_{T_1}^{-1} = D_{T_1}^{-1}\begin{pmatrix} \sum_{t=1}^{T_1} 1 & \sum_{t=1}^{T_1} \tilde{x}_t' \\ \sum_{t=1}^{T_1} \tilde{x}_t & \sum_{t=1}^{T_1} \tilde{x}_t\tilde{x}_t' \end{pmatrix} D_{T_1}^{-1} \\ &= \begin{pmatrix} T_1^{-1}\sum_{t=1}^{T_1} 1 & T_1^{-3/2}\sum_{t=1}^{T_1} \tilde{x}_t' \\ T_1^{-3/2}\sum_{t=1}^{T_1} \tilde{x}_t & T_1^{-2}\sum_{t=1}^{T_1} \tilde{x}_t\tilde{x}_t' \end{pmatrix} \\ &= \begin{pmatrix} 1 & T_1^{-1}\sum_{t=1}^{T_1}(\tilde{x}_t/\sqrt{T_1})' \\ T_1^{-1}\sum_{t=1}^{T_1}(\tilde{x}_t/\sqrt{T_1}) & T_1^{-1}\sum_{t=1}^{T_1}(\tilde{x}_t/\sqrt{T_1})(\tilde{x}_t/\sqrt{T_1})' \end{pmatrix} \\ &\xrightarrow{d} \begin{pmatrix} 1 & \int_0^1 W_\eta(r)'dr \\ \int_0^1 W_\eta(r)dr & \int_0^1 W_\eta(r)W_\eta(r)'dr \end{pmatrix} \\ &= J_I. \end{aligned} \tag{D.17}$$

Similar to the derivation to (D.11), we can show that, for the unit-root process, the projection of $\theta \in \mathcal{R}^N$ onto a convex set Λ is defined as

$$\Pi_\Lambda^I \theta = \arg\min_{\lambda\in\Lambda} (\lambda - \theta)' J_I (\lambda - \theta). \tag{D.18}$$

Similar to the derivations of (D.9) and (D.12), we can show that

$$D_{T_1}(\hat{\beta}_{T_1} - \beta_0) \xrightarrow{d} \Pi_{T_\Lambda,\beta_0}^I Z_4. \tag{D.19}$$

By noting that $T/T_1 = 1 + T_2/T_1 \to 1 + \phi$, we have

$$\begin{aligned}
\hat{A}_1 &= -T_2^{-1/2} \sum_{t=T_1+1}^{T} x_t'(\hat{\beta}_{T_1} - \beta_0) \\
&= -T_2^{-1/2} \sum_{t=T_1+1}^{T} x_t' D_{T_1}^{-1} D_{T_1}(\hat{\beta}_{T_1} - \beta_0) \\
&= -\left((T_2/T_1)^{1/2}, (T_1/T_2)^{1/2}(T/T_1)^{3/2} T^{-3/2} \sum_{t=T_1+1}^{T} \tilde{x}_t'\right) D_{T_1}(\hat{\beta}_{T_1} - \beta_0) \\
&\xrightarrow{d} -\left(\sqrt{\phi}, \phi^{-1/2}(1+\phi)^{3/2} \int_{1/(1+\phi)}^{1} W_\eta(r)' dr\right) \Pi^I_{T_\Lambda, \beta_0} Z_4 \\
&\equiv Z_5 \Pi^I_{T_\Lambda, \beta_0} Z_4, \qquad \text{(D.20)}
\end{aligned}$$

by (D.19), where $Z_5 = -(\sqrt{\phi},\ \phi^{-1/2}(1+\phi)^{3/2} \int_{1/(1+\phi)}^{1} W_\eta(r)' dr)$.

By Assumption D5 (v), we have $\hat{A}_2 = T_2^{-1/2} \sum_{t=T_1+1}^{T} v_{1t} \xrightarrow{d} Z_2$. It can be shown that $\hat{A}_1$ and $\hat{A}_2$ are asymptotically independent with each other. This completes the proof of Theorem 3.5.

Appendix E: Additional simulation results

In this supplementary appendix, we report some additional simulation results. In Section 5.4, we compute $MSE(\hat{\Delta}_1)$ for four different methods. We also compute squared biases and variances of these estimators. The results show that variances dominate biases in the sense that more than 96% of the MSEs come from variances. We show the results for the modified synthetic control (MSC) and HCW methods in Table 8. The results for the original synthetic control (OSC) and the synthetic control (SC) are similar and will not be presented here.

E.1 Estimation and inference for large N

In Section 5.4, we report $MSE(\hat{\Delta}_1) = M^{-1} \sum_{j=1}^{M} (\hat{\Delta}_{1,j} - \Delta_1)^2$. We also computed squared bias and variance of $\hat{\Delta}_1$, where $Bias(\hat{\Delta}_1) = \bar{\hat{\Delta}}_1 - \Delta_1$, $Var(\hat{\Delta}_1) = M^{-1} \sum_{j=1}^{M} (\hat{\Delta}_{1,j} - \bar{\hat{\Delta}}_1)^2$, $\bar{\hat{\Delta}}_1 = M^{-1} \sum_{j=1}^{M} \hat{\Delta}_{1,j}$, $M = 10,000$ is the number of simulations. It is easy to check that the identity $MSE(\hat{\Delta}_1) = (Bias(\hat{\Delta}_1))^2 + Var(\hat{\Delta}_1)$ holds. To save space, we report the ratios of $Var(\cdot)/MSE(\cdot)$ for the modified synthetic control (MSC) and HCW methods as they dominate the original synthetic control (OSC) and the synthetic control (SC) methods in most cases.

Table 8 reports the variance to MSE ratios for the case that u_{it} (defined in (5.1)) is uniformly distributed. We see from Table 8 that ratios of $Var(\cdot)/MSE(\cdot)$ are greater than 99% for all cases. Therefore, the squared biases are negligible compared to variances.

The negligible squared biases may be partly due to symmetric distribution of u_{it}. Thus, next we replace u_{it} by an asymmetric χ^2_1 distribution (normalized to have zero mean and unit variance). The variance to MSE ratios for chi-square distributed u_{it} case are given in Table 9 where we see that variance to MSE ratios indeed drop but still the ratios are greater than 96% for all cases considered. The results show that variance is the main component of MSE.

Table 8: Ratio $Var(\hat{\Delta}_1)/MSE(\hat{\Delta}_1)$, $u_{it} \sim$ Uniform$[-\sqrt{3}, \sqrt{3}]$

N	11	21	31	51	81	11	21	31	51	81
	DGP5					DGP6				
MSC	1.0000	1.0000	0.9999	1.0000	0.9996	1.0000	1.0000	0.9999	1.0000	0.9992
HCW	1.0000	1.0000	0.9999	1.0000	0.9999	0.9999	1.0000	1.0000	0.9999	0.9998
	DGP7					DGP8				
MSC	1.0000	1.0000	1.0000	0.9996	0.9999	0.9999	1.0000	0.9999	1.0000	1.0000
HCW	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000	1.0000	1.0000	0.9997	0.9999

Table 9: Ratio $Var(\hat{\Delta}_1)/MSE(\hat{\Delta}_1)$, $u_{it} \sim (\chi^2_1 - 1)/\sqrt{2}$

N	11	21	31	51	81	11	21	31	51	81
	DGP5					DGP6				
MSC	0.9980	0.9940	0.9970	0.9899	0.9818	0.9788	0.9717	0.9723	0.9772	0.9658
HCW	0.9994	0.9975	0.9997	0.9994	0.9998	0.9996	0.9984	0.9979	0.9989	1.0000
	DGP7					DGP8				
MSC	0.9743	0.9693	0.9657	0.9762	0.9861	0.9997	0.9996	0.9999	1.0000	0.9989
HCW	0.9732	0.9649	0.9596	0.9705	0.9905	0.9991	0.9983	0.9991	0.9997	0.9999

We also computed $MSE(\hat{y}^0_1) = M^{-1}\sum_{j=1}^{N} T_2^{-1}\sum_{t=T_1+1}^{T}(\hat{y}^0_{1t} - y^0_{1t})^2$, where $y^0_{1t,j}$ and $\hat{y}^0_{1t}$ are the generated outcome data and its estimator at the j^{th} replication. The results are given in Table 10. We see the same ranking as in the case of $MSE(\hat{\Delta}_1)$ reported in Table 3 that only for DGP6 with for $N = 11$ and $N = 21$, the HCW has smaller MSE than the modified synthetic control (MSC). For all other cases, the modified synthetic control (MSC) has smaller MSE than HCW.

We report estimated coverage probabilities of the modified synthetic control (MSC) and HCW methods for DGP7 and DGP8 discussed in Section 5.5. The results are given in Table

Table 10: MSE of $\hat{y}_1^0$

N	11	21	31	51	81	11	21	31	51	81
	DGP5					DGP6				
MSC	1.142	1.179	1.225	1.358	1.641	1.834	1.555	1.459	1.441	1.674
HCW	1.196	1.343	1.560	2.360	11.23	1.193	1.344	1.566	2.351	11.31
	DGP7					DGP8				
MSC	3.925	2.864	2.513	2.264	2.167	1.057	1.062	1.055	1.061	1.075
HCW	3.974	3.033	2.896	3.620	14.85	1.153	1.321	1.533	2.345	11.17

11. They are similar to the cases of DGP5 and DGP6. While HCW CIs significantly over-cover Δ_1 for large N, the modified synthetic control (MSC) method has more accurate coverage probabilities than the HCW method.

Table 11: Coverage probabilities for large N

	DGP7									
	Modified SC control							HCW		
N	N=31			N=51			N=81	N=31	N=51	N=81
m	40	60	90	70	80	90	90	90	90	90
50%	.481	.477	.521	.504	.498	.509	.447	.591	859	791
80%	.798	.802	.802	.797	.796	.793	.772	.865	.997	.998
90%	.883	.898	.896	.900	.883	.891	.881	.940	.998	1.00
95%	.935	.950	.941	.938	.937	.948	.935	.979	1.00	1.00
	DGP8									
	Modified SC control							HCW		
N	N=31			N=51			N=81	N=31	N=51	N=81
m	40	60	90	70	80	90	90	90	90	90
50%	.498	.437	.487	.477	.510	.491	.487	.585	.820	.731
80%	.781	.773	.793	.786	.797	.783	.798	.864	.996	.985
90%	.881	.867	.897	.879	.919	.880	.892	.948	1.00	1.00
95%	.937	.933	.937	.935	.961	.924	.931	.983	1.00	1.00

E.2 Inferences when T_2 is small

In this section, we consider the case of large T_1 $(100, 200)$ and small T_2 $(3, 5)$. We use Andrews' (2003) end-of-sample instability to test the null hypothesis H_0: $\Delta_{1t} = 0$ ($\Delta_{1,0} = 0$) against the one-sided alterative H_1: $\Delta_{1t} > 0$ for all $t = T_1 + 1, ..., T$. The data is generated by the three factor model (DGP1) as discussed in section 5.1, and the treatment effects are generated via (5.2) with $\alpha_0 = 0$ under H_0, and $\alpha_0 = 0.5$, 1 under H_1. The number of simulations is $10,000$.

The simulation results are reported in Table 12.

Table 12: Coverage probabilities for DGP1 (Andrews' (2003) instability test)

	H_0: $\alpha_0 = 0$					
	$T_2 = 3$			$T_2 = 5$		
T_1	5%	10%	20%	5%	10%	20%
100	0.0849	0.1362	0.2366	0.0935	0.1497	0.2440
200	0.0652	0.1161	0.2191	0.0711	0.1250	0.2273
	H_1: $\alpha_0 = 0.5$					
	$T_2 = 3$			$T_2 = 5$		
T_1	5%	10%	20%	5%	10%	20%
100	0.2892	0.4076	0.6656	0.3492	0.4753	0.6985
	H_1: $\alpha_0 = 1$					
	$T_2 = 3$			$T_2 = 5$		
T_1	5%	10%	20%	5%	10%	20%
100	0.5416	0.6573	0.7937	0.6994	0.7939	0.8853

Andrews' (2003) test is expected to give good estimated sizes when T_1 is large. As expected, we see from Table 12 that the test is oversized for $T_1 = 100$. Its estimated sizes improve as T_1 increases to 200. Another result from Table 12 is that, if we fix T_1, the estimated sizes deteriorate as T_2 increases. That is understandable because this test is designed for large T_1 and small T_2.

Recall that a test is said to be a consistent test if, when the null hypothesis is false, the probability of rejecting the (false) null hypothesis converges to one as sample size goes to infinity ($T_2 \to \infty$). As Andrews (2003) points out, this statistic is not a consistent test for small values of T_2. While a large T_1 helps to give better estimated sizes, it does not increase the power of the test. Therefore, we only consider $T_1 = 100$ for power calculations because for $T_1 = 200$ or even larger T_1, the powers of the test are similar. When T_1 is large, the power of the test increases with T_2 and also depends on the magnitude of $\sum_{t=T_1+1}^{T}(\Delta_{1t} - \Delta_{1,0})$ under H_1. From Table 12, we see that the estimated power increases with T_2 as well as with α_0 (the magnitude of Δ_{1t}). However, a large T_2 adversely affects the estimated sizes of Andrews' (2003) test.

We also conducted simulations of Andrews' (2003) test under DGP1 using $T_1 = 90$ and $T_2 = 20$ (the same T_1 and T_2 as in our empirical data). Based on 10,000 simulations with $\alpha_0 = 0$, the estimated sizes are 0.1660 and 0.1964 for nominal levels 5% and 10%, respectively. We see that for the $T_2 = 20$ and $T_1 = 90$ case is not large enough for the test to have good estimated

sizes because an error term of order $\sqrt{T_2/T_1}$ is not negligible, which causes Andrews' (2003) test invalid in our context. Therefore, the end-of-sample stability testing and the subsampling testing procedures are complements to each other. The former can be used when T_2 is small while the later is preferred when T_2 is not small.

Remark E.1 *For our (modified) synthetic control ATE estimator with panel data, large T_2 invalidates Andrews' (2003) test due an error term of order $\sqrt{T_2/T_1}$ becoming non-negligible. This differs from the time series model considered by Andrews (2003), where when T_2 is also large, testing a possible structural break at T_1 becomes a simple and standard problem.*

Appendix F: Explanation of subsampling method works for a wide range of subsample sizes

In this appendix, we explain why the subsampling method works well for our estimated ATE estimator for a wide range of subsample size m values.

F.1 A simple example from Andrews (2000)

We consider a simple example as considered in Andrews (2000). For $i = 1, ..., n$, Y_i is iid $N(\mu_0, 1)$ with $\mu_0 \geq 0$. I.e., $Y_i = \mu_0 + u_i$ with u_i iid $N(0,1)$ and $\mu_0 \in \Lambda = \mathcal{R}^+ \stackrel{def}{=} \{y : y \geq 0\}$. The constrained least squares estimator of μ_0 is $\hat{\mu}_n = \max\{\bar{Y}_n, 0\}$, where $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$. It is easy to show that

$$\hat{S}_n \stackrel{def}{=} \sqrt{n}(\hat{\mu}_n - \mu_0) \stackrel{d}{\to} \begin{cases} Z & \text{if } \mu_0 > 0 \\ \max\{Z, 0\} & \text{if } \mu_0 = 0, \end{cases} \tag{F.1}$$

where Z denotes a standard normal random variable. Let Y_i^* be random draws from $\{Y_j\}_{j=1}^n$. Then a bootstrap analogue of $(F.1)$ is $\sqrt{n}(\tilde{\mu}_n^* - \hat{\mu}_n)$, where $\hat{\mu}_n^* = \max\{\bar{Y}_n^*, 0\}$ and $\bar{Y}_n^* = n^{-1}\sum_{i=1}^n Y_i^*$. Andrews (2000) shows that this standard resampling bootstrap method as well as several parametric bootstrap methods do not work in the sense that, when $\mu_0 = 0$, $\tilde{S}_n^* = \sqrt{n}(\tilde{\mu}_n^* - \hat{\mu}_n)$ will not converge to $\max\{Z, 0\}$, the limiting distribution of $\hat{S}_n$. In fact, Andrews (2000) shows that $\hat{S}_n^*$ converges to a distribution that is to the left of $\max\{Z, 0\}$.

Andrews (2000) also suggests a few re-sampling methods that overcome the problem. One particular easy-to-implement method is a parametric subsampling method. Specifically, for values of m that satisfy $m \to \infty$ and $m/n \to 0$ as $n \to \infty$, one can use $\tilde{S}_m^* = \sqrt{m}(\hat{\mu}_m^* - \hat{\mu}_n)$ to approximate the distribution of $\sqrt{n}(\hat{\mu}_n - \mu_0)$. Here $\hat{\mu}_m = \max\{\bar{Y}_m^*, 0\}$ and $\bar{Y}_m^* = m^{-1}\sum_{i=1}^m Y_i^*$ with Y_i^* being iid draws from $N(\bar{Y}_n, 1)$. I.e., $Y_i^* = \bar{Y}_n + u_i^*$ with u_i^* iid $N(0,1)$. To see that the subsampling method indeed works, we have that, conditional on $\{Y_i\}_{i=1}^n$,

$$\begin{aligned}
\hat{S}_m^* &\stackrel{def}{=} \sqrt{m}(\hat{\mu}_m^* - \hat{\mu}_n) \\
&= \max\left\{\sqrt{m}\,\bar{Y}_m^*, 0\right\} - \sqrt{m}\,\hat{\mu}_n \\
&= \max\left\{\sqrt{m}\,\bar{Y}_m^*, 0\right\} - \sqrt{m}\,\mu_0 - \sqrt{m}(\hat{\mu}_n - \mu_0) \\
&= \max\left\{\sqrt{m}(\bar{Y}_m^* - \bar{Y}_n + \bar{Y}_n - \mu_0), -\sqrt{m}\,\mu_0\right\} - \sqrt{m}(\hat{\mu}_n - \mu_0) \\
&= \max\left\{\sqrt{m}(\bar{Y}_m^* - \bar{Y}_n) + \sqrt{m/n}\sqrt{n}(\bar{Y}_n - \mu_0), -\sqrt{m}\,\mu_0\right\} - \sqrt{m/n}\sqrt{n}(\hat{\mu}_n - \mu_0) \\
&= \max\left\{\sqrt{m}(\bar{Y}_m^* - \bar{Y}_n) + o_p(1), -\sqrt{m}\,\mu_0\right\} + o_p(1) \\
&\stackrel{d}{\to} \begin{cases} Z & \text{if } \mu_0 > 0 \\ \max\{Z, 0\} & \text{if } \mu_0 = 0, \end{cases}
\end{aligned} \tag{F.2}$$

where the second equality follows from the definition of $\hat{\mu}_m^*$, the third equality follows from adding and subtracting $\sqrt{m}\,\mu_0$, the fourth equality follows from $\max\{a,b\} - c = \max\{a-c, b-c\}$, the sixth equality follows from $m/n = o(1)$, $\sqrt{n}(\bar{Y}_n - \mu_0) = O_p(1)$ and $o(1)O_p(1) = o_p(1)$. The last equality follows from the fact that $Y_i^* - \bar{Y}_n = u_i^*$ is iid $N(0,1)$. Hence, $\sqrt{m}(\hat{Y}_m^* - \bar{Y}_n) \stackrel{d}{\sim} N(0,1) \equiv Z$ for any value of m. If $\{Y_i^*\}_{i=1}^m$ is iid with mean $\bar{Y}_n$ and unit variance but is not normally distributed, then we need m to be large so that $\sqrt{m}(\hat{Y}_m^* - \bar{Y}_n) \stackrel{d}{\to} N(0,1) \equiv Z$ by virtue of a central limit theorem argument (as $m \to \infty$).

Comparing (F.1) and (F.2), we see that subsampling method works under very mild conditions that $m \to \infty$ and $m/n \to 0$ as $n \to \infty$.

F.2 Testing for zero ATE by subsampling method

We conduct simulations to examine the finite sample performances of the subsampling method. We generate Y_i iid $N(0,1)$ (i.e., $\mu_0 = 0$) for $i = 1, ..., n$ and we choose $n = 100$ and conduct 5000 simulations. Within each simulation, we generate 2000 subsampling samples with subsample

sizes $m \in \{5, 10, 20, 30, 50, 100\}$. Note that we select the largest $m = n = 100$ because we want to show numerically that the standard bootstrap method does not work. For each fixed value m, we sort the 2000 subsampling statistics in ascending order such that $\hat{S}^*_{m,(1)} \leq \hat{S}^*_{m,(2)} \leq ... \leq \hat{S}^*_{m,(2000)}$. Then we get right-tail α-percentile value by $\hat{S}^*_{((1-\alpha)(2000))}$. We record rejection rate as the percentage that $\hat{S}$ is greater or equal to $\hat{S}^*_{((1-\alpha)(2000))}$ for $\alpha \in \{0.01, 0.05, 0.1, 0.2\}$. We consider two cases: (i) We generate Y_i iid $N(0,1)$ and $Y^*_i = \bar{Y}_n + v_i$ with v_i iid $N(0,1)$; and (ii) We generate Y_i uniformly distributed over $[-\sqrt{3}, \sqrt{3}]$ (so that it has zero mean and unit variance) and $Y^*_i = \bar{Y}_n + v_i$ with v_i iid uniformly distributed over $[-\sqrt{3}, \sqrt{3}]$. The results for the two cases are almost identical. For brevity, we only report the normally distributed v_i case in Table 13.

Table 13: Estimated sizes ($Y^*_i \sim N(\bar{Y}_n, 1)$)

	m=5	m=10	m=20	m=30	m=50	m=100
1%	.0132	. 0126	.0124	.0130	.0136	.0248
5%	.0516	.0518	.0518	.0532	.0658	.1032
10%	.0960	.0968	.1006	.1104	.1346	.2014
20%	.1936	.2004	.2278	.2588	.3164	.4020

First, we see that the subsampling method with $5 \leq m \leq 20$ seem to work well. Second, we see clearly that using $m = n$ or m close to n ($m \geq 50$) do not work. For example, when $m = n$, it gives estimated rejection rates double that of the nominal levels. Andrews (2000) shows that the distribution of $\sqrt{n}(\hat{\mu}^*_n - \hat{\mu}_n)$ is to the left of that of $\sqrt{n}(\hat{\mu}_n - \mu_0)$. Hence, the bootstrap method will lead to over rejection of the null hypothesis. Our simulation results verifies Andrews' theoretical analysis.

The simulation results seem contradict to the simulation results reported in Section 5 where even for $m = n$, the subsampling method seems to be fine. We explain the seemingly contradictory result in the next subsection.

F.3 Not all parameters are at the boundary

Our simulations reported in Section 5 correspond to the case of $\beta_{0,j} > 0$ for $j = 2, ..., 7$ and $\beta_{0,j} = 0$ for $j = 8, ..., 11$. The constrained estimators $\hat{\beta}_{T_1,j}$ ($\hat{\beta}^*_{m,j}$) for $j = 8, 9, 10, 11$ can cause problems for the standard bootstrap method. However, notice that our ATE estimator also

depends on $\hat{\beta}_{T_1,j}$ ($\hat{\beta}^*_{m,j}$) for $j = 1, ..., 7$, which does not take boundary value 0. This helps to improve subsampling method for large value of m. More importantly, our ATE estimator also contains a term not related to $\hat{\beta}_{T_1}$ (see the second term at the right hand side of (4.5) and the existence of this term further improves the performance of the subsampling method when m is close to or equal to n. This is the reason why in our simulations even when $m = n$, the subsampling method seems to work fine. To numerically verify this conjecture, we generate a sequence of iid $Z_1, Z_2 \sim N(0, \sigma_v^2)$ random variables and add them to $\hat{S}_n$ and $\hat{S}^*_m$, i.e., $\tilde{S}_n = \hat{S}_n + Z_1$ and $\tilde{S}^*_m = \hat{S}^*_m + Z_2$. We then repeat the simulations to compute the estimated sizes. The results for $\sigma_v = 1$ and 5 are reported in Table 14. We observe that the performance of the subsampling statistic $\tilde{S}^*_m$ has significant improvements over $\hat{S}^*_m$ for $m = 50$ and 100. Consider the case of $\sigma_v = 1$ and $m = n$. The rejection rates based on $\tilde{S}^*_m$ is about 20% higher than that of the nominal levels whereas it was 100% higher than that of nominal levels based on $\hat{S}^*_m$.

From Table 14, we see that when σ_v^2 is large, Z_1 and Z_2 becomes the dominating components of $\tilde{S}_n$ and $\tilde{S}^*_m$. Therefore, the subsampling method works well for all values of m including $m = n$. The estimated sizes for $\sigma_v^2 = 1$ are only slightly oversized compared to $\sigma_v^2 = 25$. This shows that the significant improvements in the estimated sizes (over the case of $\sigma_v^2 = 0$) does not require adding a regular component with large dominating variance.

Table 14: Estimated sizes: Adding a $N(0, \sigma_v^2)$ to $\hat{S}_n$ and $\hat{S}^*_m$

	m=5	m=10	m=20	m=30	m=50	m=100
	$\sigma_v = 1$					
1%	.0104	.0110	.0112	.0128	.0122	.0114
5%	.0550	.0562	.0562	.0590	.0600	.0648
10%	.1066	.1098	.1140	.1168	.1198	.1236
20%	.2170	.2244	.2320	.2372	.2440	.2520
	$\sigma_v = 5$					
1%	.0112	.0116	.0116	.0110	.0124	.0128
5%	.0518	.0521	.0528	.0530	.0542	.0556
10%	.1030	.1044	.1046	.1048	.1060	.1074
20%	.2070	.2082	.2030	.2102	.2126	.2160

Appendix G: Additional robustness check results

G.1 Comparison with the unconstrained estimator (OLS)

In this subsection, we consider using the ordinary least squares method (we interchangeably use ordinary least squares, HCW and unconstrainted estimator) to estimate the counterfactual outcome. Let $\hat{\beta}_{OLS}$ denote the least squares estimator of β using the pre-treatment sample. Then the counterfactual outcome is estimated by $\hat{y}_t^0 = x_t'\hat{\beta}_{OLS}$ (e.g., Hsiao, Ching, and Wan (2012)). Applying this method to the Columbus data gives an estimated ATE of \$645.3 increase in weekly sales after the opening of a showroom in Columbus. While this number is close to the ATE estimation result of \$673.91 by the modified synthetic control, we would like to compare the out-of-sample forecasting performances of the two estimation methods in order to judge which method gives a more accurate ATE estimation result.

The difference between the least squares method and our modified synthetic control method is that the synthetic control method imposes a non-negativity restriction on the slope coefficients when estimating the regression model using the pre-treatment data. The rationale for imposing the non-negativity constraints is that outcome variables from treated and control units are driven by some common factors and therefore, they are more likely to move up and down together. Imposing a correct restriction can improve out-of-sample forecast. Therefore, we compare the out-of-sample forecast performances of the modified synthetic control method and the least squares method. We choose a value $T_0 \in (1, T_1) = (1, 90)$ to estimate the regression model. Then we forecast outcome y_{1t} for $t = T_0 + 1, ..., T_1$. Since there is no treatment prior to T_1, we can compare the average prediction squared error over the period $t = T_0 + 1, ..., T_1$. Specifically, we estimate the following model

$$y_t = x_t'\beta + u_{1t}, \qquad t = 1, ..., T_0 \tag{G.1}$$

by the modified synthetic control and the least squares method. Let $\hat{\beta}_{T_0}$ and $\hat{\beta}_{OLS}$ denote the resulting estimators using the two methods, respectively. We predict y_{1t}^0 by $\hat{y}_{1t,MSC}^0 = x_t'\hat{\beta}_{T_0}$ and $\hat{y}_{1t,OLS}^0 = x_t'\hat{\beta}_{OLS}$ for $t = T_0 + 1, ..., T_1$. Then we compute the prediction MSEs by $PMSE_{MSC} = (T_1 - T_0)^{-1} \sum_{t=T_0+1}^{T_1} (y_{1t} - \hat{y}_{1t,MSC}^0)^2$ and $PMSE_{OLS} = (T_1 - T_0)^{-1} \sum_{t=T_0+1}^{T_1} (y_{1t} - \hat{y}_{1t,OLS}^0)^2$. As in Li and Bell (2017), we consider the cases where the 'pre-treatment' estimation

sample is larger than the 'post-treatment' evaluation sample. We choose six different values for $T_0 = \{60, 65, 70, 75, 80, 85\}$. The corresponding evaluation sample sizes are $T_1 - T_0 = \{30, 25, 20, 15, 10, 5\}$. We report the ratio of PMSE as $PMSE_{OLS}/PMSE_{MSC}$. The results are reported in Table 15.

Table 15: Out-of-sample Prediction MSE ratio

T_0	60	65	70	75	80	85
$\frac{PMSE_{OLS}}{PMSE_{MSC}}$	1.680	1.104	1.020	1.273	1.188	1.143

From Table 15 we observe that the least squares method has larger PMSE than the modified synthetic control method for all cases. The PMSE for the former ranges from 2% to 68% larger than the later. Thus, the empirical example shows that, in order to more accurately predict the counterfactual outcomes for the treated unit, it is helpful to impose non-negativity restriction on the slope coefficients when estimating model (G.1).

G.2 Adding Covariates

We collect monthly data on unemployment rate (Unemp), labor force (LF) and average weekly earnings (Inc) for Columbus and linearly extrapolate them to weekly data. The data is downloaded from the Bureau of Labor Statistics website (bls.gov). The estimation model is

$$y_{1t} = x_t'\beta_0 + z_{1t}'\gamma_0 + u_{1t}, \qquad t = 1, ..., T_1 \tag{G.2}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})'$, we consider three cases of adding covariates: (i) $z_{1t} = (Unemp_t, LF_t, Inc_t)'$, i.e., add the three covariates linearly to the regression model; (ii) add both the three covariates and their square terms, i.e., add a total of six additional regressors; (iii) add three more cross product terms of the three covariates, i.e., add a total of nine additional regressors (3 linear, 6 quadratic terms), γ_0 is a $k \times 1$ vector of parameters, where k is the dimensional of z_{1t}. Since opening a showroom has no (or negligible) effect on z_{1t}, we can use the above model to predict post-treatment counterfactual sales for the treated city. Specifically, we estimate model (G.2) under the restriction $\beta_j \geq 0$ for $j \geq 2$ using the pre-treatment data $t = 1, ..., T_1$ (there are no restrictions for the other parameters). Let $\hat{\beta}_{T_1}$ and $\hat{\gamma}_{T_1}$ denote the corresponding estimators.

We estimate the counterfactual outcome y_{1t}^0 by

$$\hat{y}_{1t}^0 = x_t'\hat{\beta}_{T_1} + z_{1t}'\hat{\gamma}_{T_1} \tag{G.3}$$

for $t = T_1 + 1, ..., T$ and estimate ATE by $T_2^{-1}\sum_{t=T_1+1}^{T}(y_{1t} - \hat{y}_{1t}^0)$. Note that in (G.3) we use the treated unit's covariates z_{1t} in estimating the counterfactual outcome y_{1t}^0. We do not need to use control units' covariates to form a synthetic path for z_{1t} because z_{1t} is exogenous in the sense that the treatment event will not affect (or its effect on z_{1t} is negligible) covariates' evolution of the treated unit.

Figure 5: Columbus: Modified synthetic control ATE, add Covariates

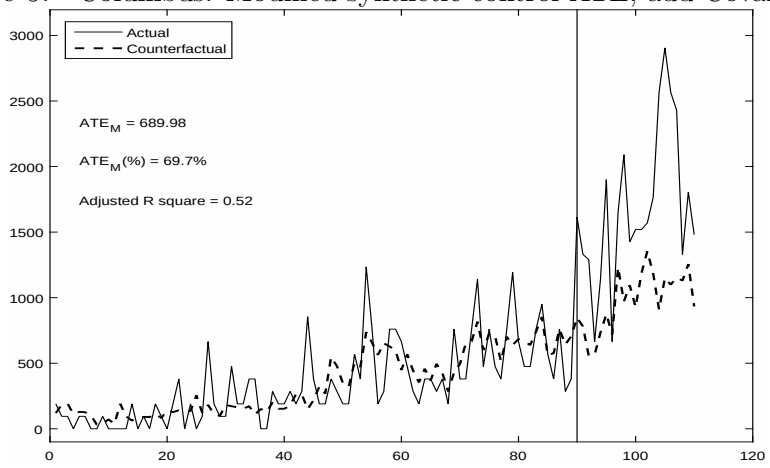


Figure 5 plots the estimation result for Columbus with three covariates added to the regression model linearly, i.e., z_{1t} is of dimensional three. The ATE becomes 69.7% which is quite close to the original result of 67%. However, the adjusted R^2 decreased slightly from 0.528 to 0.520, indicating that the three covariates do not have additional explanatory power to explain sales. Obtaining virtually the same ATE estimation result even with added covariates supports our original ATE estimation result. For cases (ii) and (iii), z_{1t} is of dimensional six and nine, the resulting adjusted R^2 are reduced to .495 and .478, respectively. Therefore, adding quadratic terms of the three covariates do not give additional prediction power to Columbus' sales.

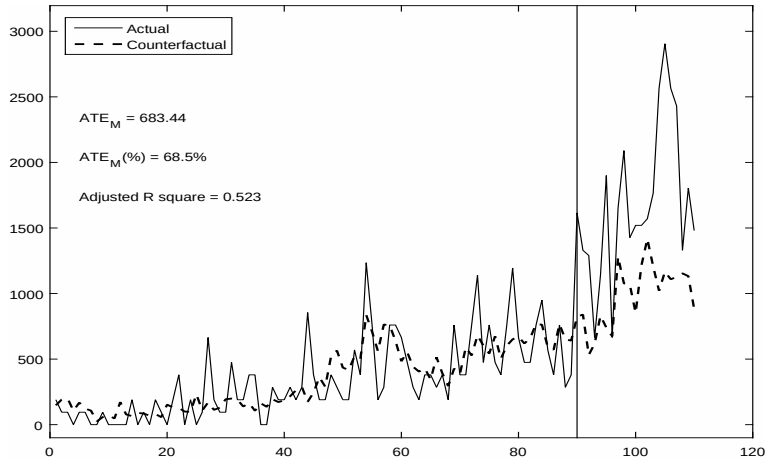
G.3 Selecting control units based on covariate matching

In this subsection, we first select cities whose covariates are close to the covariates of the treated city. Then we select the number of control cities by comparing adjusted R^2. Finally we estimate

ATE using the selected control units. We explain this procedure in more detail below.

For each $j = 1, 2, 3$ (corresponding to Unemp, LF, Inc), we regress $z_{1,jt}$ on $z_{i,jt}$ using the pre-treatment data and obtain the goodness-of-fit $R^2_{i,j}$ for $i = 2, ..., 11$. We obtain a total R-square for city i by $R^2_i = R^2_{i,1} + R^2_{i,2} + R^2_{i,3}$. We sort them in a non-increasing order: $R^2_{(2)} \geq R^2_{(3)} \geq ... \geq R^2_{(11)}$. Their corresponding sales are denoted by $y_{(2),t}$,...,$y_{(11),t}$ for $t = 1, ..., T_1$. Next, we regress y_{1t} on $y_{(2),t}$ and obtain an adjusted $\bar{R}^2_{(2)}$. Then, we regress y_{1t} on $(y_{(2),t}, y_{(3),t})$ and obtain an adjusted $\bar{R}^2_{(2),(3)}$. We continue this way until we regress y_{1t} on all $(y_{(2),t}, ..., y_{(11),t})$. We choose a model with the largest adjusted $\bar{R}^2$. For Columbus, the method that selects seven cities (Portland, Houston and Atlanta are not selected) gives the largest adjusted $\bar{R}^2$. Using the seven selected cities as control group, the modified synthetic control method's estimation result is plotted in Figure 6. The ATE estimation result is 68.5% which is quite close to the original result of 67%. The robustness check shows that our ATE estimation result is not sensitive to the selection of different control units.

Figure 6: Columbus: ATE Estimation Based on Covariates Matching



G.4 Allowing for v_{1t} to be serially correlated

As discussed in Section 6.2, when testing the null that v_{1t} is serially uncorrelated, we obtain a p-value of 0.0963. It is not strong evidence supporting the null hypothesis. In this section, we allow for v_{it} to follow an AR(1) process: $v_{1t} = \rho_v v_{1,t-1} + \xi_t$, where ξ_t is serially uncorrelated. Since v_{1t} enters the term $\hat{A}_2$, this only changes our calculation of $\hat{A}^*_2$. The steps of generating $\hat{A}^*_2$ are as follows: First, one obtains $\hat{\rho}_v$ by regressing $\hat{v}_{1t}$ on $\hat{v}_{1,t-1}$ with $t = T_1 + 1, ..., T$.

Then one estimates ξ_t by $\hat{\xi}_t = \hat{v}_{1t} - \hat{\rho}_v \hat{v}_{1,t-1}$ and compute $\hat{\sigma}_\xi^2 = T_2^{-1} \sum_{t=T_1+2}^{T} \hat{\xi}_t^2$. Next, one generates $\xi_t^* \sim$ iid $N(0, \hat{\sigma}_\xi^2)$ and $v_{1t}^* = \hat{\rho}_v v_{1,t-1}^* + \xi_t^*$ for $t = T_1 + 1, ..., T$, where $v_{1,T_1}^* \sim$ iid $N(0, \hat{\sigma}_\xi^2/(1 - \hat{\rho}_v^2))$. Finally, one obtains $\hat{A}_2^* = T_2^{-1/2} \sum_{t=T_1+1}^{T} v_{1t}^*$. Note that $\hat{A}_1^*$ is generated the same way as discussed in Section 4.1 and is $\hat{A}^* = \hat{A}_1^* + \hat{A}_2^*$. The above steps are repeated J times, and the remaining steps as how to obtain the $1 - \alpha$ confidence interval for Δ_1 are the same as discussed in Section 4.1.

The estimated confidence intervals are given in Table 16. Comparing Table 16 with Table 5, we observe the results are similar although the estimated confidence intervals reported in Table 16 are wider than those in Table 5.

Table 16: Confidence intervals (MSC, v_{1t} follows an AR(1) process)

	m=20	m=40	m=60	m=80	m=90
80% CI	[471.2, 897.6]	[465.8, 8906.7]	[468.4, 893.7]	[470.5, 892.7]	[462.1, 888.7]
90% CI	[415.5, 959.6]	[411.9, 951.2]	[408.7, 952.1]	[408.2, 957.1]	[403.2, 953.9]
95% CI	[367.7, 1152.3]	[361.3, 1009.3]	[359.2, 1006.7]	[361.0, 1009.9]	[357.4, 1001.3]
99% CI	[262.7, 1125.8]	[246.7, 1157.4]	[254.2, 1105.8]	[261.6, 1121.7]	[261.7, 1106.5]

References

Andrews, D. W. K. (1999): "Estimation when a parameter is on a boundary," *Econometrica*, 67(6), 1341–1383.

——— (2000): "Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space," *Econometrica*, 68(2), 399–405.

——— (2003): "End-of-Sample Instability Tests," *Econometrica*, 71(6), 1661–1694.

Fang, Z., and A. Santos (2018): "Inference on directionally differentiable functions," *forthcoming in Review of Economic Studies*.

Hsiao, C., S. Ching, and K. S. Wan (2012): "A panel data approach for program evaluation: measuring the benefits of political and economic integration of Hong Kong with mainland China," *Journal of Applied Econometrics*, 27(5), 705–740.

Li, K. T., and D. R. Bell (2017): "Estimation of average treatment effects with panel data: Asymptotic theory and implementation," *Journal of Econometrics*, 197(1), 65–75.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*. Springer Series in Statistics. Berlin: Springer.

Stock, J. H., and M. W. Watson (1993): "A simple estimator of cointegrating vectors in higher order integrated systems," *Econometrica*, 3(5), 783–820.

Zarantonello, E. H. (1971): "Projections on Convex Sets in Hilbert Space and Spectral Theory: Part I. Projections on Convex Sets: Part II. Spectral Theory," in *Contributions to Nonlinear Functional Analysis*, pp. 237–424. Elsevier.