Inference Theory for Treatment Effects with Nonlinear Trending Data of Unknown Form

Kathleen T. Li
McCombs School of Business, University of Texas
2110 Speedway Stop B6700, Austin TX 78712
kathleen.li@mccombs.utexas.edu

Venkatesh Shankar
The Mays Business School, Texas A&M University
College Station, TX 77843
vshankar@mays.tamu.edu

May 27, 2020

Abstract

The synthetic control (SC) method is a valuable technique to estimate causal effects, in particular, the average treatment effect on the treated (ATT), in quasi-experimental data. The modified synthetic control (MSC) method is a powerful method for estimating causal effects when the synthetic control parallel trends assumption does not hold. Much of the inference theory for the SC and the MSC methods has been developed for stationary data and non-stationary data of known form. However, in many real world applications, the data follow unknown, nonlinear trends. For example, marketing data are typically nonlinear and nonstationary. We fill the research void and extend the literature by developing the inference theory for the SC/MSC ATT estimate when the outcome variables exhibit nonlinear trend of unknown form. With the new inference theory and results we provide in this paper, researchers can use the powerful SC/MSC methods to estimate the ATT and conduct inference for a variety of data types, including stationary, unit-root non-stationary, and nonlinear trend non-stationary processes.

Keywords: quasi-experiment; treatment effects, nonlinear trend, causal inference, synthetic control.

JEL Number: C1, C13, C18, M3, M30, M31

1 Introduction

The synthetic control (SC) method is a valuable technique to estimate causal effects in quasi-experimental data (Abadie et al., 2010; Abadie and Gardeazabal, 2003). The modified synthetic control (MSC), which relaxes the zero intercept and the 'weight sum to one' constraints behind the SC parallel trends assumption, is a powerful method for estimating causal effects when the SC parallel trends assumption does not hold (Doudchenko and Imbens, 2016). Much of the inference theory for the SC and the MSC methods has been developed for stationary data and nonstationary data of known form. However, in many real world applications, the data follow unknown, nonlinear trends. For example, marketing data are typically nonlinear and nonstationary (e.g., Dekimpe and Hanssens, 1995). The inference theory for causal estimation using synthetic control for such data is absent.

In this paper, we fill this void and develop the missing inference theory by extending the existing SC/MSC inference theory to allow the outcome variables to exhibit nonlinear trend of unknown form. Because the SC/MSC average treatment effect on the treated (ATT) estimate is related to the least squares based ATT estimation method (the HCW method, Hsiao, Ching, and Wan, 2012) via projection (Li, 2020), our inference theory also extends the HCW method to cover nonlinear trend data. Our inference theory is based on a large number of pre- and post-treatment periods with one (or a few) treatment unit(s) and a fixed number of control units. Through simulations, we also show that our theory performs well even when the number of pre- and post-treatment time periods is such that the number of control units (as regressors) is similar to the pre- and post-treatment sample size.

We extend the growing literature on synthetic control and related methods to estimate the ATT. Athey et al. (2017) consider a general class of the ATT estimation methods that include the SC and the MSC as special cases. They allow both the number of time series observations and the number of control units to be large and derive bounds for the ATT estimate. However, they do not provide any inference theory for the estimate. Chernozhukov, Wuthrich, and Zhu (2019) propose a general inference procedure that covers the difference-in-differences (DID) method, the SC method, and a factor-model-based method. However, they use a strong exchangeability assumption that may not be plausible in many applications where treatment and control units follow different distributions. To relax the strong exchangeability condition, Chernozhukov et al. (2019) impose a condition of a small post-treatment sample size. The small post-treatment sample size assumption is not always empirically common. Li (2020) derives the large sample distribution theory for the SC/MSC ATT and considers the case of long pre- and post-treatment time periods, one treatment unit and a fixed number of controls, while allowing the data to be stationary, trend stationary and unit root non-stationary. However, Li (2020) does not cover nonlinear

trend data of unknown form. We fill this research gap and extend the literature and by developing the inference theory for the SC/MSC ATT estimate when the outcome variables exhibit a nonlinear trend of unknown form. With the new inference theory and results we provide in this paper, researchers can use the powerful SC/MSC methods to estimate the ATT and conduct inference for a variety of data types, including stationary, unit-root non-stationary, and nonlinear trend non-stationary processes.

2 The Model and Inference

2.1 Estimation of the ATT

In this section, we discuss the notation to estimate the ATT. We use y_{it} to denote the outcome variable for unit i at time t for $i = 1, ..., N$ and $t = 1, ..., T$. We consider the case where only one unit receives a treatment at time $T_1 + 1$ with $1 < T_1 < T - 1$. Without loss of generality, we assume the first unit is the treatment unit ($i = 1$) and all others are control units ($i = 2, ...N$). We use y_{1t}^1 and y_{1t}^0 to denote the outcome of the treatment unit at time t with and without treatment, respectively. Then the treatment effect for first unit at time t is $\Delta_{1t} = y_{1t}^1 - y_{1t}^0$. However, the fundamental problem of causal inference is that we observe only $y_{1t} = y_{1t}^1$ for $t > T_1$. Therefore, we need to estimate the counterfactual y_{1t}^0 using control units' information. We assume that, in the absence of treatment, the outcome variables are generated by a nonlinear trend factor, plus a stationary term as follows:

$$y_{jt}^0 = c_j + d_j f_t + u_{jt}, \qquad j = 1, ..., N; \ t = 1, ..., T, \tag{2.1}$$

where u_{jt} are zero mean, finite variance, and serially uncorrelated stationary idiosyncratic errors. The factor f_t can be random, and f_t and u_{jt} are uncorrelated for all $j = 1, ..., N$, $t = 1, ..., T$. We assume that f_t has a nonlinear trend of unknown form. For example, $f_t = b(t) + \eta_t$, where η_t is a zero mean, finite variance stationary process; $b(t) = \sum_{l=1}^{m} a_l t^{\alpha_l}$ consists of m polynomial terms in t with $0 \leq \alpha_1 < \alpha_2 < ... < \alpha_m$. We do not need α_l to be integers but require that $m \geq 2$ so that $b(t)$ has a trend component.

The factor model structure is a convenient, useful, and natural way to generate cross sectionally dependent data without imposing any parametric distribution assumptions on the data. We allow f_t to have unknown nonlinear trend and do not require the distribution of η_{jt} to be known. An advantage of the factor model structure is that it is flexible and can incorporate any specific parametric distribution, e.g., normal distribution. To see this, consider a simple case of $N = 3$ and assume that $f_t = b(t) + v_t$, where $b(t)$ is a deterministic function of t, v_t iid $N(0, \sigma_v^2)$, $u = (u_{1t}, u_{2t}, u_{3t})'$ is a zero mean normal vector $u_t \sim N(\mathbf{0}_{3\times 1}, V)$, where V is a 3×3 positive definite matrix. We allow u_{it} and u_{jt} to be correlated with $j \neq i$ so that V is not restricted to be a diagonal matrix. Then from (2.1), it is easy to see that

$y_t = (y_{1t}, y_{2t}, y_{3t})'$ is normally distributed with a flexible cross sectional dependence structure.

$$y_t = N\left(\begin{pmatrix} c_1 + d_1 b(t) \\ c_2 + d_2 b(t) \\ c_3 + d_3 b(t) \end{pmatrix},\ V + \sigma_v^2 \begin{pmatrix} d_1^2 & d_1 d_2 & d_1 d_3 \\ d_2 d_1 & d_2^2 & d_2 d_3 \\ d_3 d_1 & d_3 d_2 & d_3^2 \end{pmatrix}\right), \tag{2.2}$$

where $V = Var(u_t)$ is 3×3 positive definite matrix. Therefore, commonly used parametric cross sectional dependent structures can be easily incorporated into a factor model framework. In addition, a factor model can generate more general dependence structures as we do not need to assume that v_t and u_{jt} are normally distributed.

To estimate the ATT, we need to construct a counterfactual outcome for the treatment unit during the posttreatment time period. We mainly consider the SC and the MSC methods that estimate the following linear regression model under certain constraints using the pretreatment data:

$$y_{1t} = x_t'\beta + e_{1t}, \qquad t = 1, ..., T_1, \tag{2.3}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})$, and e_{1t} is a zero mean stationary idiosyncratic error term. The SC/MSC method estimates β by minimizing $\sum_{t=1}^{T_1}(y_{1t} - x_t'\beta)^2$ subject to some constraints on β. Specifically, the SC method imposes three restrictions: $\beta_1 = 0$ (no intercept), (ii) $\sum_{j=2}^{N}\beta_j = 1$ (weight sum to one), and (iii) $\beta_j \geq 0$ for $j \geq 2$ (non-negativity weight). The MSC method removes the first two constrains and only keeps the non-negativity weight restriction. A closely related method proposed by Hsiao, Ching and Wan (2012) suggests dropping all the three restrictions imposed by the SC method and simply estimating β using the unconstrained least squares method. Using $\hat{\beta}$ to denote a generic estimator of β from (2.3), we can estimate the counterfactual outcome using $\hat{y}_{1t}^0 = x_t'\hat{\beta}$ and estimate the ATT using

$$\hat{\Delta}_1 = \frac{1}{T_2}\sum_{t=T_1+1}^{T}(y_{1t} - \hat{y}_{1t}^0) = \frac{1}{T_2}\sum_{t=T_1+1}^{T}(y_{1t} - x_t'\hat{\beta}). \tag{2.4}$$

Since we focus on one treatment unit, the average is over the number of posttreatment time periods.

2.2 The Least Squares Theory

Because the ATT contains $\hat{\beta}$ as seen from (2.4), the inference theory for the ATT estimate $\hat{\Delta}_1$ depends on the inference theory of $\hat{\beta}$. In addition, the inference theory of the SC/MSC estimate is related to the least squares theory because the constrained estimator of β can be represented as a projection of the least squares estimator onto a convex cone (Li, 2020). Therefore, we need to first examine the distribution theory for the least squares estimator of β. A difficulty arises because control units' outcome variables share a common nonlinear trend factor f_t, which is the dominating component for all the regressors, $y_{2t}, ..., y_{Nt}$. These regressors are asymptotically collinear. As a result, the standard large sample (for large T_1) least squares inference theory breaks down. This would not be hard to fix if we know the functional form of f_t.

For example, if $f_t = t$, we can simply replace y_{jt} (for $j = 2, ..., N$) by its de-trended version and add a linear trend t as an additional regressor to (2.3). This exercise results in a linear trend regression model, for which the least squares theory is well established (e.g., Hamilton 1994, Chapter 16). However, marketing data typically have complicated nonlinear trends, so a misspecified linear trend model can severely bias the ATT. Therefore, we do not make any functional form assumption for the factor f_t. To our knowledge, the least squares theory for Model (2.3) when f_t is a nonlinear trend process with unknown form, is not explored in the literature.

We fill this gap by developing an inference theory for the least squares estimator of β based on (2.3) with outcome variables generated by (2.1). This extension is important because the large sample behaviors of the SC and the MSC estimates are projections of the least squares estimates onto some convex cones (Li, 2020). Therefore, it is necessary to develop the inference theory for the least squares estimator before we can use the result to derive the inference theory for the SC and the MSC ATT estimators.

To derive the limiting distribution of the least squares estimator of β, we impose some conditions on the nonlinear trend function f_t in Assumption 2 of Appendix A. We allow f_t to belong to a wide range of functions including polynomials in t (permitting non-integer powers of t). See discussions that follow Assumption 2 in Appendix A for more details. We use a monotone increasing function $g(t)$ to characterize the speed that f_t diverges to infinity and the fast rate of convergence of the trending variable coefficient estimate. We require that the $g(\cdot)$ function satisfies Assumption 2 in Appendix A. One of the conditions is

$$\frac{T_1^{-1}\sum_{t=1}^{T_1} f_t}{g(T_1)} \to b_1 \qquad \text{as } T_1 \to \infty, \tag{2.5}$$

where b_1 is a (non-zero) constant. In many examples, $g(t)$ can be chosen as the leading component of f_t. For instance, if $f_t = \sum_{j=1}^{L} a_j t^{\nu_j}$ with $0 \le \nu_1 < ... < \nu_L = \nu$ (ν is not necessarily an integer). We can choose $g(t) = t^{\nu}$. Then $T_1^{-1}\sum_{t=1}^{T_1} f_t/g(T_1) \approx T_1^{-(\nu+1)}\sum_{t=1}^{T_1} a_L t^{\nu} \to a_L/(\nu+1)$. Hence, (2.5) holds. Here, the notation $A(T_1) \approx B(T_1)$ means that $B(T_1)$ is the leading term of $A(T_1)$, or $A(T_1)/B(T_1) \to 1$ as $T_1 \to \infty$. In practice, we do not need to know the functional form of f_t provided that it satisfies Assumption 2 of Appendix A. We present the limiting distribution result for the least squares estimator of β in the following Theorem.

Theorem 2.1 *If the outcome variables are generated by the factor model (2.1) for $t = 1, ..., T_1$, and Assumptions 1 and 2 in Appendix A hold, for $N \ge 3$,*

(i) $\hat{\beta}_{OLS,T_1,j} - \beta_j$ converges to zero at the rate $T_1^{-1/2}$ for $j = 1, ..., N$; and $\sum_{j=2}^{N} d_j(\hat{\beta}_{OLS,T_1,j} - \beta_j)$ converges to zero at a rate $(\sqrt{T_1}\, g(T_1))^{-1}$, where $g(t)$ is an unbounded monotone increasing function satisfies Assumption 2;

(ii) $C_{T_1}(\hat{\beta}_{OLS,T_1} - \beta) \xrightarrow{d} N(0, \Omega)$,

where $\hat{\beta}_{OLS,T_1}$ *is the least squares estimator of* β *based on (2.3) using the pretreatment data,* C_{T_1} *is an* $N \times N$ *(invertible) transformation matrix defined in the Web Appendix C (and in Appendix B for* $N = 3$ *case),* Ω *is an* $N \times N$ *positive definite matrix.*

Proof of Theorem 2.1 appears in Appendix B for the simple case of $N = 3$ and in the Web Appendix C for the general case.

At first glance, Theorem 2.1 (i) may be unexpected because it is well known that for a regression model with a time trend regressor, the least squares coefficient estimator converges to the true value at a rate much faster than $T_1^{-1/2}$ (the rate for stationary data). The intuition for this result is that because $y_{2t}, ..., y_{Nt}$ are asymptotically collinear, the identification and estimation of β actually come from u_{jt}, the stationary components in y_{jt}. The collinearity prevents $\hat{\beta}_{OLS,T_1}$ from converging to β faster than $T_1^{-1/2}$. This is the reason why Theorem 2.1 holds only when $N \geq 3$. Because for $N = 2$, there is only one trending regressor y_{2t} at the right-hand-side of Equation (2.3) and collinearity problem disappears. It can be easily shown that, when $N = 2$, the least squares estimator of β_2 converges to the true value at a rate faster than $T_1^{-1/2}$

Theorem 2.1 claims that $\hat{\beta}_{OLS,T_1,j}$ converges to β_j at the $T_1^{-1/2}$ rate for all $j = 1, ..., N$, while a linear combination of them converges to zero at a rate faster than $T_1^{-1/2}$. This faster rate of convergence is actually quite intuitive to understand. For e_{1t} defined in (2.3) to be a stationary process, the trend component f_t's coefficients from both sides of (2.3) must cancel each other. After substituting (2.1) into (2.3), the coefficient of f_t on the left-hand-side is d_1, while on the right-hand-side, the coefficient is $\sum_{j=2}^{N} \beta_j d_j$ so that we must have $d_1 - \sum_{j=2}^{N} \beta_j d_j = 0$. Moreover, because it is the coefficient of the trend variable f_t, it is well established that its estimate converges to zero faster than $T_1^{-1/2}$. Therefore, $\sum_{j=2}^{N} d_j(\hat{\beta}_{OLS,T_1,j} - \beta_j)$ converges to zero faster than $T_1^{-1/2}$. For example, if $f_t = t$, the rate of convergence of a (linear) time trend coefficient estimate is $T_1^{-3/2}$. This is consistent with our Theorem 2.1 because when $f_t = t$, $g(t) = t$ as $g(t)$ is the leading component of f_t, and $(\sqrt{T_1} g(T_1))^{-1} = T_1^{-3/2}$ gives us the correct rate of convergence. See Web Appendix E for simulation evidence supporting the theoretical prediction of Theorem 2.1.

Wooldridge (1991) considers the problem of estimating a contegration equation and shows that when there is more than one cointegration relationship, the cointegration coefficients are uniquely determined by the stationary components of the unit root processes. Wooldridge (1991) proves the consistency of the cointegration coefficient estimate but does not provide large sample distribution theory for the cointegration vector estimator (see also Hamilton, 1994, pages 590-591). By combining the methods of Wooldridge (1991) and this paper, and under the assumption that the outcome variables are generated by a factor model,

we are able to derive the asymptotic distribution of cointegration coefficient estimate when more than one cointegration vector exists.

The normalization matrix C_{T_1} in Theorem 2.1 involves some unobservable parameters such as $g(T_1)$ which is the leading term of f_t and is unknown in practice. These unknown parameters do not prevent us from utilizing the result of Theorem 2.1 to derive the limiting distribution of HCW ATT estimate (see Theorem 3.1 and the proof in Appendix A). However, in other contexts such as hypothesis testings, a known normalization factor is helpful. The next theorem presents a result with a known normalization $\sqrt{T_1}$.

Theorem 2.2 *Under the same conditions as in Theorem 2.1,*

$$\sqrt{T_1}(\hat{\beta}_{OLS,T_1} - \beta) \xrightarrow{d} N(0, \Sigma),$$

where Σ is an $N \times N$ positive semidefinite matrix with rank $N-1$.

Proof of Theorem 2.2 appears in Appendix B for the simple case of $N = 3$ and in the Web Appendix C for the general case.

The asymptotic variance Σ is not full rank, which implies that Σ is not invertible. However, this issue does not affect the usefulness of Theorem 2.2. For example, if we want to test some linear restrictions imposed on β. E.g., the form of $R\beta - q = 0$, where R and q are $J \times N$ and $J \times 1$ known matrices, respectively, with $1 \leq J < N$. Let $\hat{d} = R\hat{\beta}_{OLS,T_1} - q$. Then, $Var(\sqrt{T_1}\hat{d}) = R\,Var(\sqrt{T_1}\hat{\beta}_{OLS,T_1})\,R' \to R\Sigma R'$ as $T_1 \to \infty$, which is invertible in general even Σ does not have a full column rank. From a practical point of view, the variance of $\sqrt{T_1}(\hat{\beta}_{OLS,T_1} - \beta)$ is always invertible in finite sample applications. We can define X as the $T_1 \times N$ matrix with its t^{th} row given by $x_t = (1, y_{2t}, ..., y_{Nt})'$. Then, $X'X$ matrix is always invertible in practice because if X is not full rank (in finite sample applications), we can remove the redundant control units and still consistently estimate the ATT. Because our regression model is a linear projection model, regardless of which control units we use as x_t, the idiosyncratic error e_{1t} is orthogonal to x_t by the property of the linear projection.

3 Inference Theory for the ATT Estimator

Using the large sample theory for the least squares estimator of β developed in the previous section, we investigate the inference theory for different ATT estimators. Using the definition of the treatment effects $\Delta_{1t} = y_{1t}^1 - y_{1t}^0$ and (2.3), we can express the treatment unit's outcome in the posttreatment period as

$$y_{1t} = x_t'\beta + \Delta_{1t} + e_{1t}, \qquad t = T_1 + 1, ..., T. \tag{3.1}$$

Substituting (3.1) into (2.4) and using $\hat{y}_{1t}^0 = x_t'\hat{\beta}$, we obtain

$$\hat{\Delta}_1 = \bar{\Delta}_1 - \frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'(\hat{\beta}-\beta) + \frac{1}{T_2}\sum_{t=T_1+1}^{T} e_{1t}, \tag{3.2}$$

where $\bar{\Delta}_1 = T_2^{-1}\sum_{t=T_1+1}^{T}\Delta_{1t}$.

We decompose the ATT estimator into two terms. Let $A = \sqrt{T_2}(\hat{\Delta}_1 - \bar{\Delta}_1)$ and rearrange terms in (3.2) yields

$$\begin{aligned} A &= \sqrt{T_2}(\hat{\Delta}_1 - \bar{\Delta}_1) \\ &= -\frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} x_t'(\hat{\beta}-\beta) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} e_{1t} \\ &= A_1 + A_2, \end{aligned} \tag{3.3}$$

where $A_1 = -T_2^{-1/2}\sum_{t=T_1+1}^{T} x_t'(\hat{\beta}-\beta)$ and $A_2 = T_2^{-1/2}\sum_{t=T_1+1}^{T} e_{1t}$.

Our inference method is based on approximating the distributions of A_1 and A_2 defined in Equation (3.3). Because the inference theory for the SC and the MSC methods rely on the large sample theory for the least squares based the ATT estimator, we will first present the inference procedure for the least squares method based (HCW) ATT estimator before discussing the SC/MSC method.

3.1 Inference Procedure for the HCW Method

Replacing $\hat{\beta}$ in (3.3) by $\hat{\beta}_{OLS,T_1}$ results in

$$\begin{aligned} A_{HCW} &\equiv \sqrt{T_2}(\hat{\Delta}_{HCW,1} - \bar{\Delta}_1) \\ &= -\frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} x_t'(\hat{\beta}_{OLS,T_1}-\beta) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} e_{1t} \\ &= A_{HCW,1} + A_{HCW,2}, \end{aligned} \tag{3.4}$$

where $A_{HCW,1} = -T_2^{-1/2}\sum_{t=T_1+1}^{T} x_t'(\hat{\beta}_{OLS,T_1}-\beta)$ and $A_{HCW,2} = T_2^{-1/2}\sum_{t=T_1+1}^{T} e_{1t}$.

The large sample theory of the HCW ATT estimator is presented in the next theorem whose proof appears in Appendix A.

Theorem 3.1 *Under the same conditions as in Theorem 2.1,*

$$A_{HCW} \xrightarrow{d} N(0, V),$$

where $V = B\Omega B' + \sigma_e^2$, B is the probability limit of $T_2^{-1/2}\sum_{t=T_1+1} x_t'(C_{T_1}')^{-1}$, $\sigma_e^2 = E(e_{1t}^2)$ and Ω is defined in Theorem 2.1.

In Appendix A we show that a consistent estimator of the asymptotic variance of A_{HCW} is

$$\hat{V} = \hat{\sigma}_e^2\psi'(X'X)^{-1}\psi/T_2 + \hat{\sigma}_e^2, \tag{3.5}$$

with $\psi = \sum_{t=T_1+1}^{T} x_t$, $\hat{\sigma}_e^2 = \sum_{t=1}^{T_1} \hat{e}_{1t}^2/(T_1 - N)$, $\hat{e}_{1t} = y_{1t} - x_t'\hat{\beta}_{OLS,T_1}$, $\psi = \sum_{t=T_1+1}^{T} x_t$.

Therefore, the following standardized ATT estimator is approximately standard normally distributed.

$$\frac{\sqrt{T_2}(\hat{\Delta}_{HCW,1} - \bar{\Delta}_1)}{\sqrt{\hat{V}}} \overset{d}{\sim} N(0,1). \tag{3.6}$$

Using (3.6), we can calculate the $(1-\alpha)^{th}$ confidence interval of $(\hat{\Delta}_{HCW,1} - \Delta_1)/\sqrt{\hat{V}}$ for $\alpha \in (0,1)$ as follows:

$$P\left(c_{\alpha/2} \leq (\hat{\Delta}_{HCW,1} - \bar{\Delta}_1)/\sqrt{\hat{V}/T_2} \leq c_{1-\alpha/2}\right), \tag{3.7}$$

where c_α is the α-th quantile of a standard normal random variable, i.e., $P(N(0,1) \leq c_\alpha) = \alpha$. Therefore, the asymptotic $(1-\alpha)^{th}$ confidence interval of $\bar{\Delta}_1$ is:

$$\left[\ \hat{\Delta}_{HCW,1} - c_{1-\alpha/2}\sqrt{\hat{V}/T_2},\ \ \hat{\Delta}_{HCW,1} - c_{\alpha/2}\sqrt{\hat{V}/T_2}\ \right]. \tag{3.8}$$

For example, for $\alpha = 0.10$, the 90% confidence interval for $\bar{\Delta}_1$ is

$$\left[\ \hat{\Delta}_{HCW,1} - 1.645\sqrt{\hat{V}/T_2},\ \ \hat{\Delta}_{HCW,1} + 1.645\sqrt{\hat{V}/T_2}\ \right], \tag{3.9}$$

because the critical values are $c_{0.95} = 1.645$ and $c_{0.05} = -1.645$.

An key insight is that we do not need to know the specific trending functional form of f_t because our inference procedure is invariant to the nonlinear functional form of f_t. Simulations in Section 4 confirm our theoretical analysis and show that the inference procedure based on (3.8) works well.

3.2 Large Sample Theory for the MSC/SC ATT Estimator

The large sample analysis for the SC and the MSC methods are very similar. We will mainly focus on the inference procedure for the MSC method. Let $\hat{\beta}_{MSC,T_1}$ denote the MSC estimator of β which can be obtained as a minimizer of $\sum_{t=1}^{T_1}(y_{1t} - x_t'\beta)^2$ subject to $\beta_j \geq 0$ for $j = 2, ..., N$. Li (2020) shows that the limiting distribution of $\sqrt{T_1}(\hat{\beta}_{MSC,T_1} - \beta)$ can be represented as a projection of the limiting normal distribution of $\sqrt{T_1}(\hat{\beta}_{OLS,T_1} - \beta)$ onto a convex cone.[1] Therefore, by combining Theorem 2.1 and Theorem 3.1 of Li (2020), we can obtain the limiting distribution of $\sqrt{T_1}(\hat{\beta}_{MSC,T_1} - \beta)$.

Let Z_{OLS} denote the limiting distribution of $C_{T_1}(\hat{\beta}_{OLS,T_1} - \beta)$ as presented in Theorem 2.1. Because both $\hat{\beta}_{MSC,T_1}$ and β take values in the constrained set $\Lambda_{MSC} \overset{def}{=} \{\beta \in \mathcal{R}^N,\ \ \beta_j \geq 0 \text{ for } j = 2, ..., N.\}$, we can show that $C_{T_1}(\hat{\beta}_{MSC,T_1} - \beta)$ takes values in a convex cone, which is the asymptotic range of $C_{T_1}(\hat{\beta}_{MSC,T_1} - \beta)$ (as $T_1 \to \infty$). We use $T_{\Lambda_{MSC},\beta}$ to denote this convex cone.[2] We use $\Pi_D\theta$ to denote a projection of $\theta \in \mathcal{R}$ onto a convex set D. See Appendix A for the detailed definition of this projection. The

[1] See Li (2020) for a detailed definition of the projection operator and the related convex cone.

[2] In the statistical literature, it is called the Tangent cone of Λ_{MSC} evaluated at θ.

large sample distribution of $A_{MSC} = \sqrt{T_2}(\hat{\Delta}_{MSC,1} - \bar{\Delta}_1)$, where $\Delta_{MSC,1} = T_2^{-1} \sum_{t=T_1+1}^{T} (y_{1t} - x_t' \hat{\beta}_{MSC,T_1})$, is stated in the following theorem.

Theorem 3.2 *Let Z_{OLS} denote the limiting normal distribution of $\sqrt{T_1}(\hat{\beta}_{OLS,T_1} - \beta)$. Under the same conditions as in Theorem 2.1,*

$$A_{MSC} \xrightarrow{d} -B\,\Pi_{\Lambda_{MSC},\beta} Z_{OLS} + Z_{MSC,2},$$

where B is defined in Theorem 3.1, $Z_{HCW,2}$ is a normal random variable with zero mean, variance σ_e^2 and is independent of Z_{OLS}.

Theorem 3.2 provides the large sample distribution theory for the MSC ATT estimator. However, this limiting distribution is non-standard and we cannot determine its critical values because the distribution depends a convex cone that is unknown in practice.[3] In addition, bootstrap method is known to be invalid when some β_j take boundary value zero. Fortunately, in these cases, the subsampling method can still be used. Politis, Romano and Wolf (1999) show that the subsampling method can be used to approximate a well defined limiting distribution under very weak conditions which only requires that the subsampling sample size increases with the original sample size but at a slower growth rate. Using m for the subsampling size, we only need $m \to \infty$ and $m/T_1 \to 0$ as $T_1 \to \infty$. Alternatively, we can also use a numerical bootstrap method to approximate the non-standard limiting distribution or use a numerical delta method to estimate the projection operator (functional derivative). Refer to Hong and Li (2018, 2020) for detailed discussion on using numerical bootstrap method to approximation non-standard distributions when bootstrap method fails.

In this paper we suggest using a parametric subsampling method for inference. First, let us examine the MSC ATT estimator. Replacing $\hat{\beta}$ by $\hat{\beta}_{MSC,T_1}$ in (3.3), we obtain the MSC ATT statistic:

$$\begin{aligned} A_{MSC} &\equiv \sqrt{T_2}(\hat{\Delta}_{MSC,1} - \bar{\Delta}_1) \\ &= -\frac{1}{\sqrt{T_1 T_2}} \sum_{t=T_1+1}^{T} x_t' \sqrt{T_1}(\hat{\beta}_{MSC,1} - \beta) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} e_{1t} \\ &\equiv A_{MSC,1} + A_{MSC,2}. \end{aligned} \tag{3.10}$$

We include a parametric bootstrap procedure as a special case of our subsampling procedure (when the subsample size $m = T_1$). Although the bootstrap method does not provide asymptotically valid inference when some β_j takes boundary value zero, Li (2020) shows that the bias in estimating treatment effects confidence interval is quite small in practice. Therefore, bootstrap methods often produce reasonably accurate estimated confidence intervals for the MSC ATT estimator. We refer readers to the Web Appendix

[3] The convex cone depends on how many β_j takes the boundary value of zero which is unknown in applications.

F in Li (2020) for a detailed argument and simulation evidence supporting the use of bootstrap methods in inference for the MSC ATT estimator. Therefore, we propose the following parametric subsampling-bootstrap method. We call it a subsampling-bootstrap approach because on one part of the statistic, $A_{MSC,1}$, we use a subsampling method, and on the other part of the statistic, $A_{MSC,2}$, we use a bootstrap method. In addition, we allow the subsampling size m to be the same as the sample size T_1. Therefore, we allow the use of bootstrap method to both terms, $A_{MSC,1}$ and $A_{MSC,2}$.

Notice that the only term that prevents bootstrap methods from delivering valid inference is $(\hat{\beta}_{MSC,T_1} - \beta)$ in $A_{MSC,1}$ which has a non-standard distribution. Following the standard subsampling principle, we replace $\sqrt{T_1}(\hat{\beta}_{MSC,T_1} - \beta)$ by $\sqrt{m}(\hat{\beta}^*_{MSC,m} - \hat{\beta}_{MSC,T_1})$, where $\hat{\beta}^*_{MSC,m}$ is computed using a subsample size m and will be explained below. For $A_{MSC,2}$, we can use a simple parametric bootstrap method and replace it by $A^*_2 = T_2^{-1}\sum_{t=T_1+1}^{T} e^*_{1t}$, where e^*_{1t} is iid $N(0, \hat{\sigma}^2_e)$, $\hat{\sigma}^2_e$ is defined in the subsampling step (i) below. Thus, our subsampling-bootstrap version of A_{MSC} is defined as follows:

$$\begin{aligned} A^*_{MSC} &= -\frac{1}{\sqrt{T_1 T_2}}\sum_{t=T_1+1}^{T} x_t'\sqrt{m}(\hat{\beta}^*_{MSC,m} - \hat{\beta}_{MSC,T_1}) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} e^*_{1t} \\ &\equiv A^*_{MSC,1} + A^*_{MSC,2}. \end{aligned} \tag{3.11}$$

We explain how to generate $\hat{\beta}^*_{MSC,m}$ in the subsampling procedure step (ii) below.

A Parametric Subsampling Procedure for the MSC Method

Step (i) Generate e^*_{1t} iid $N(0, \hat{\sigma}^2_e)$ for $t = 1, ..., T$, where $\hat{\sigma}^2_e = \frac{1}{T_1 - N}\sum_{t=1}^{T_1} \hat{e}^2_{1t}$ and $\hat{e}_{1t} = y_{1t} - x_t'\hat{\beta}_{MSC,T_1}$. Compute $A^*_{MSC,2} = T_2^{-1}\sum_{t=T_1+1}^{T} e^*_{1t}$.

Step (ii) Choose the last m period's x-data from the pretreatment sample as our subsample x. That is, $(x^*_1, ..., x^*_{m-1}, x^*_m) = (x_{T_1-(m-1)}, ..., x_{T_1-1}, x_{T_1})$. Generate $y^*_{1t} = x^{*'}_t\hat{\beta}_{MSC,T_1} + e^*_{1t}$ for $t = 1, ..., m$, where e^*_{1t} is generated in step (i). Use the subsampling sample $\{x^*_t, y^*_{1t}\}_{t=1}^m$ to estimate β by the MSC method to obtain $\hat{\beta}^*_{MSC,m}$, i.e., $\hat{\beta}^*_{MSC,m}$ minimizes $\sum_{t=1}^m (y^*_{1t} - x^{*'}_t\beta)^2$, subject to the constraint that all the slope coefficients are non-negative. With $\hat{\beta}^*_{MSC,m}$, compute $A^*_{MSC,1}$ as defined in (3.11). This term, together with $A^*_{MSC,2}$ from step (i), results in A^*_{MSC} as described in (3.11).

Step (iii) Repeat steps (i) and (ii) B times to get $\{A^*(b)\}_{b=1}^B$. Use the empirical distribution of $\{A^*(b)\}_{b=1}^B$ to approximate the distribution of $A_{MSC} = \hat{\Delta}_{MSC,1} - \bar{\Delta}_1$. Sort $\{A^*(b)\}_{b=1}^B$ in an ascending order such that $A^*_{(1)} \leq A^*_{(1)} \leq ... \leq A^*_{(B)}$, for $\alpha \in (0,1)$, the $(1-\alpha)$ confidence interval of $\bar{\Delta}_1$ is:

$$[\hat{\Delta}_{MSC,1} - A^*_{(((1-\alpha/2)M))},\ \ \hat{\Delta}_{MSC,1} - A^*_{((\alpha/2)M)}]. \tag{3.12}$$

The subsampling procedure for the SC method is similar to that of the MSC method except we replace

$\hat{\beta}_{MSC,T_1}$ and $\hat{\beta}^*_{MSC,m}$ by $\hat{\beta}_{SC,T_1}$ and $\hat{\beta}^*_{SC,m}$, respectively in the subsampling steps. We use simulation to examine the finite sample performance of both the asymptotic inference theory for A_{HCW} and the subsampling-bootstrap procedure for A_{HCW} and A_{SC}, where $A_{SC} = \sqrt{T_2}(\hat{\Delta}_{SC,1} - \bar{\Delta}_1)$.

4 Simulation

In this section, we use simulations to study the finite sample performances of the SC, the MSC and the HCW methods when data have a nonlinear trend. We examine the accuracy of the ATT estimation mean squared errors (MSE) and the coverage probability, i.e., the probability that our estimated confidence interval covers the true ATT. We show that the asymptotic inference theory works well for HCW ATT estimator, and the subsampling-bootstrap procedure provides satisfactory finite sample approximation for the SC and the MSC ATT estimators.

4.1 Simulation Setup

We study the finite sample performances of the ATT estimators and inference procedures as discussed in Section 2. Hsiao, Ching and Wan (2012) use three stationary factors in their simulation studies. We replace the HCW's first factor (a stationary AR(1) factor) by a nonlinear trend factor. Therefore, we use the following three factors to generate the outcome variables:

$$\begin{aligned} f_{1t} &= 0.2t - 0.8\sqrt{t} + 0.8f_{1t-1} + \epsilon_{1t}, \\ f_{2t} &= -0.6f_{1t-1} + \epsilon_{2t} + 0.8\epsilon_{2t-1}, \\ f_{3t} &= \epsilon_{3t} + 0.9\epsilon_{3t-1} + 0.4\epsilon_{3t-2}, \end{aligned} \tag{4.1}$$

where ϵ_{it} is iid $N(0,1)$. We also used other distributions such as a uniform distribution or a center and scale adjusted χ^2_1 distribution to replace the standard normal distribution in generating ϵ_{it}. The results are virtually identical, so we only report the standard normal distribution case for brevity.

The first factor f_{1t} is a nonlinear trend process while the second and third factors, f_{2t} and f_{3t}, are stationary factors. In the absence of treatment, the outcome variable is generated by the three factors and an error term.

$$y^0_t = a + Bf_t + u_t, \qquad t = 1, ..., T \tag{4.2}$$

where $y^0_t = (y^0_{1t}, y^0_{2t}, ..., y^0_{Nt})'$, $a = (a_1, a_2, ..., a_N)'$, and $u_t = (u_{1t}, u_{2t}, ..., u_{Nt})'$ are all $N \times 1$ vectors, $B = (b_1, b_2, ..., b_N)'$ is the $N \times 3$ loading matrix where b_j is a 3×1 loading vector for unit j, and $f_t = (f_{1t}, f_{2t}, f_{3t})'$. We choose $(a_1, a_2, ..., a_N) = (1, 1, ..., 1)$, u_{it} is iid $N(0,1)$.

We consider $T_1 \in \{20, 40, 80\}$, $T_2 \in \{10, 20, 40\}$ and $N \in \{11, 31\}$ (10 or 30 control units). We set 3×1 vector factor loading $b_i = c_i \mathbf{1}_{3\times 1}$ where $\mathbf{1}_{3\times 1} = (1, 1, 1)'$ and c_i is a scalar. Equation (4.2) implies that

$$y_{it}^0 = 1 + c_i \mathbf{1}_{3\times 1}' f_t + u_{it}, \qquad i = 1, ..., N;\ t = 1, ..., T. \tag{4.3}$$

For the choice of factor loadings, we set $b_1 = c_1 \mathbf{1}_{3\times 1}$, $b_i = c_2 \mathbf{1}_{3\times 1}$ for $i = 2, ..., (N+1)/2$ and $b_i = c_3 \mathbf{1}_{3\times 1}$ for $i = (N+3)/2, ..., N$. Following Li (2020), we consider the following four sets of (c_1, c_2, c_3).

$$\begin{aligned} DGP1: &\qquad (c_1, c_2, c_2) = (1, 1, 1), \\ DGP2: &\qquad (c_1, c_2, c_2) = (0.2, 1, 0.5), \\ DGP3: &\qquad (c_1, c_2, c_2) = (1, 2, -0, 5), \\ DGP4: &\qquad (c_1, c_2, c_2) = (1, -2, 0.5). \end{aligned} \tag{4.4}$$

For DGP1, the treatment and control units are random draws from a common distribution. For the other three DGPs, the treatment and controls are draws from different distributions. However, for DGP3, $c_3 < c_1 < c_2$ so that the treatment unit's outcome is within the convex hull of control units' outcomes. For DGP2 and DGP4, the treatment outcome is outside the convex hull of controls' outcome because c_1 is either smaller than $\min\{c_2, c_3\}$ or larger than $\max\{c_2, c_3\}$. The SC parallel trend is violated for DGP2 and DGP4. Since both the MSC and the HCW methods do not impose the restriction that the treatment unit's outcome lies inside the convex hull of the control units, their ATT estimation results are expected to be robust regardless of whether the treatment unit's outcome rests inside the convex hull of the controls. As discussed in Li (2020), DGP4 favors the HCW method because half of the control units are strongly negatively correlated with the treatment unit. The HCW can leverage the strongly negatively correlated controls to accurately predict the treatment unit's counterfactual outcome as long as the number of regressors is not too large.

4.2 Comparison of MSEs by Method

In this section, we compare the ATT estimation mean squared errors (MSE) for the SC, the MSC and the HCW methods. The MSE is:

$$MSE(\hat{\Delta}_1) = \frac{1}{M} \sum_{j=1}^{M} (\hat{\Delta}_{1j} - \bar{\Delta}_1)^2, \tag{4.5}$$

where $\hat{\Delta}_{1j}$ is the ATT estimator $\hat{\Delta}_1$ using the j^{th} simulation data, and $M = 10,000$ is the number of simulation replications. We consider $N \in \{11, 31\}$, $T_1 \in \{20, 40, 80\}$ and $T_2 \in \{10, 20, 40\}$. Table 1 presents the estimation results for $N = 11$.

Table 1: Comparison of MSEs across the Methods when $N = 11$

(T_1, T_2)	(20,10)	(40,10)	(80,10)	(20,20)	(40,20)	(80,20)	(20,40)	(40,40)	(80,40)
	SC Method								
DGP1	0.1302	0.1221	0.1189	0.0641	0.0608	0.0589	0.0317	0.0307	0.0288
DGP2	0.5578	5.7950	46.57	1.466	8.525	54.85	5.849	16.06	73.93
DGP3	0.2239	0.3630	0.1911	0.3762	0.4347	0.1465	1.308	0.7127	0.1497
DGP4	0.9009	15.82	129.1	2.565	23.45	152.3	10.45	44.36	205.3
	MSC Method								
DGP1	0.4945	0.3690	0.2056	0.8832	0.4376	0.1636	2.635	0.7689	0.1766
DGP2	0.4675	0.3092	0.1716	0.8748	0.3732	0.1429	2.667	0.6295	0.1603
DGP3	0.5548	0.3668	0.1937	0.9988	0.4477	0.1568	2.985	0.7531	0.1755
DGP4	1.2206	1.0156	0.3751	2.6079	1.4253	0.3343	8.271	2.661	0.3955
	HCW Method								
DGP1	0.9644	0.4466	0.2154	1.7396	0.5369	0.1732	5.176	0.9397	0.1924
DGP2	0.8374	0.4026	0.1957	1.4861	0.4715	0.1591	4.398	0.8247	0.1726
DGP3	0.9152	0.4246	0.2038	1.6539	0.5048	0.1653	4.991	0.8814	0.1813
DGP4	0.9163	0.4265	0.2051	1.4861	0.4715	0.1591	4.935	0.8743	0.1814

For DGP1 and DGP3, the SC parallel trends assumption holds. As expected, the SC method has the smallest MSE for DGP1 and DGP3 for all (T_1, T_2) combinations. For DGP2 and DGP4, the SC parallel trend assumption is violated, and the SC method can have a substantially large MSE due to its enormous estimation bias. The MSC method performs best for DGP2 while the HCW outperforms its competitors for DGP4. This result is also as expected because DGP4 is designed as a case that favors the HCW method because half of the control units are strongly negatively correlated with the treatment unit while the other half are weakly positively correlated with the treatment unit. While the MSC method drops the negatively correlated controls, the HCW method takes the advantage of these strongly negatively correlated controls to more accurately estimate the ATT.

An interesting result is that, the MSE differences among the different methods for different DGPs are much larger for the nonlinear trend data than those for the stationary data reported in Li (2020). For example, for DGP1 with $(T_1, T_2) = (20, 10)$, the SC MSE is 0.1302, which is less than 1/3 of the MSC MSE (0.4945), and MSC MSE is about half of the HCW MSE of 0.9644. These results imply that it is very important to select the correct method to accurately estimate the ATT when the data have nonlinear trends.

Next, we increase N from 11 to 31 (30 controls) and examine how this affects the relative performance of the estimation methods. Since we need $T_1 > N$ for the HCW method, we choose $T_1 \in \{40, 80\}$ and still keep $T_2 \in \{10, 20, 40\}$. The MSE results appear in Table 2.

While the SC method still outperforms others for DGP1 and DGP3, the MSC now performs the best for both DGP2 and DGP4. Interestingly, the HCW method dominates the MSC for DGP4 when $N = 11$. However, when N is large ($N = 31$), the HCW method has large estimation variance due to using too

Table 2: Comparison of MSEs across the Methods when $N = 31$

(T_1, T_2)	(40,10)	(80,10)	(40,20)	(80,20)	(40,40)	(80,40)
	SC Method					
DGP1	0.1165	0.1125	0.0581	0.0575	0.0294	0.0284
DGP2	5.796	46.58	8.478	54.85	15.98	73.87
DGP3	0.3437	0.1800	0.4000	0.1408	0.6998	0.1364
DGP4	15.76	129.0	23.46	152.3	44.35	205.3
	MSC Method					
DGP1	0.3673	0.1940	0.4361	0.1603	0.7842	0.1750
DGP2	0.3768	0.1812	0.4572	0.1465	0.8347	0.1599
DGP3	0.4628	0.2060	0.5477	0.1711	0.9761	0.1781
DGP4	0.5394	0.2484	0.6640	0.2095	1.158	0.2265
	HCW Method					
DGP1	1.458	0.2848	1.734	0.2329	3.078	0.2533
DGP2	1.415	0.2778	1.653	0.2186	2.913	0.2451
DGP3	1.419	0.2758	1.747	0.2267	2.987	0.2386
DGP4	1.461	0.2787	1.723	0.2299	2.994	0.2465

many regressors to estimate in the regression model. The MSC drops many negatively correlated units and the effective number of regressors is greatly reduced, resulting in smaller estimation variance and MSE. Therefore, when N is large, the MSC outperforms the HCW method.

4.3 Comparison of Coverage Probabilities

In this section, we compute the estimated coverage probabilities of the confidence intervals. We consider three different values for T_1, fix $T_2 = 20$ and $N = 11$ and choose $\alpha = 0.05$, 0.1, 0.2 and 0.5, to compute the coverage probability that the $(1-\alpha)^{th}$ confidence intervals contains the true $\bar{\Delta}_1$. The true treatment effects is set to be $\Delta_{1t} = 0$ for $t = T_1 + 1, ..., T$.

4.3.1 Coverage Probability for the HCW Method

For the HCW method, We consider two different values for $T_1 \in \{40, 80\}$, fix $T_2 = 20$ and $N = 11$. Since we have a standard normal distribution for the HCW ATT estimator, we set the number of replications to be 10,000 for all cases. The estimated coverage probabilities for the HCW method appear in Table 3. We observe that the estimated coverage probabilities are very close to their nominal levels for all cases and for all DGPs. One important advantage of the HCW method is that it is computationally simple. It requires only an unconstrained least squares estimation.

4.3.2 Coverage Probability for the MSC Method

This section reports the estimated confidence intervals using the MSC method. We fix $T_2 = 20$, $N = 11$ and consider $T_1 \in \{40, 80\}$. Because the subsample size m must be between N and T_1, when $T_1 = 40$,

Table 3: Coverage Probabilities for the HCW Method ($T_2 = 20$, $N = 11$)

	DGP1				DGP2			
Cov. pr.	50%	80%	90%	95%	50%	80%	90%	95%
$T_1 = 40$	0.5042	0.8033	0.9012	0.9482	0.5124	0.8125	0.9074	0.9534
$T_1 = 80$	0.5122	0.8092	0.9053	0.9529	0.5114	0.8099	0.9062	0.9530
	DGP3				DGP4			
Cov. prob.	50%	80%	90%	95%	50%	80%	90%	95%
T_1 =40	0.5135	0.8111	0.9060	0.9527	0.5120	0.8088	0.9052	0.9515
$T_1 = 80$	0.5134	0.8099	0.9054	0.9522	0.5146	0.8106	0.9061	0.9528

we choose $m \in \{20, 40\}$ and when $T_1 = 80$, we choose $m \in \{20, 40, 60, 80\}$. The number of simulations is 500. Within each simulation, we generate 400 subsampling statistics and use its empirical distribution to approximate the finite sample distribution of the ATT estimator. The results for the MSC inferential method appear in Table 4.

Table 4: Coverage Probabilities for the MSC method ($T_2 = 20$, $N = 11$)

Cov. pr.	50%	80%	90%	95%	50%	80%	90%	95%
	$T_1 = 40$							
m	DGP1				DGP2			
20	0.506	0.772	0.864	0.906	0.486	0.755	0.862	0.914
40	0.464	0.756	0.868	0.928	0.530	0.818	0.898	0.942
	DGP3				DGP4			
20	0.534	0.834	0.906	0.952	0.450	0.758	0.860	0.930
40	0.524	0.824	0.918	0.948	0.392	0.692	0.820	0.900
	$T_1 = 80$							
m	DGP1				DGP2			
20	0.556	0.840	0.924	0.972	0.538	0.816	0.888	0.926
40	0.499	0.801	0.900	0.946	0.532	0.812	0.914	0.956
60	0.508	0.789	0.893	0.952	0.520	0.800	0.890	0.942
80	0.492	0.778	0.888	0.938	0.508	0.782	0.878	0.924
	DGP3				DGP4			
20	0.586	0.860	0.940	0.970	0.582	0.866	0.954	0.976
40	0.510	0.792	0.890	0.938	0.540	0.824	0.920	0.962
60	0.508	0.810	0.906	0.946	0.490	0.798	0.888	0.938
80	0.504	0.814	0.926	0.964	0.502	0.780	0.884	0.942

For $T_1 = 40$ and $m = 20$, estimated coverage probabilities are slightly smaller than their nominal levels for most cases, but distortions are moderate. Next, for $m = 40$, except for DGP4 which has undercoverage on the true ATT, the subsampling-bootstrap procedure works reasonably well for other DGPs. Recall that DGP4 is a case that least favors the MSC method because half of the control units are strongly negatively correlated with the treatment unit. In this case, the MSC bootstrap method suffers from relatively large bias in estimating ATT confidence intervals, but even so the distortions are moderate.

From Table 4, we observe that the estimation accuracy improves as T_1 is doubled. When $T_1 = 80$,

all the values of m, including $m = 80$, result in estimated coverage probabilities that are close to their nominal values. The bootstrap method (when $m = T_1$) works well because although some components of $\hat{\beta}_{MSC,j}$ may take the boundary value of zero, other components of $\hat{\beta}_{MSC,j}$ take positive values and the non-negativity constraints are no longer binding for those components. In addition, the term A_2 defined in (3.10) is unrelated to the subsample size m. All these factors help reduce estimation bias for the bootstrap method (refer to Appendix F of Li 2020 for a detailed explanation for why bootstrap often do not produce large bias in estimating confidence intervals). In practice, we recommend using more than one value of m as a robustness check for inference. For example, we can use $m = T_1/2$ and $m = T_1$. If the estimated confidence intervals are not sensitive to different m values, then it is reassuring that the subsampling method yields reliable estimated confidence intervals.

The result in Table 4 is consistent with the stationary data result reported in Li (2020): the bootstrap method (when $m = T_1$) often only leads to mild distortions in estimating the confidence intervals. Please refer to the supplementary Appendix F of Li (2020) for a more detailed explanation and simulation evidence supporting the argument that the bootstrap method usually does not give rise to large distortions.

4.3.3 Coverage Probability for the SC Method

For the SC control method, the parallel trend assumption is violated for DGP2 and DGP4 because the treatment is outside the convex hull of the control units. Therefore, the SC ATT is severely biased and the confidence intervals do not cover the true ATT (all estimated coverage probabilities are zero). As a result, we show only the estimated coverage probabilities for DGP1 and DGP3. For both DGP1 and DGP3, the SC parallel trend assumption holds because the treatment is within the convex hull of control units.

Table 5 shows the estimated coverage probabilities for the SC method for DGP1 and DGP3. First, we focus on $T_1 = 40$. For DGP1, both $m = 20$ and $m = 40$ produce accurate coverage probabilities. However, for DGP3, the $m = 40$ case performs better than the $m = 20$ case. When T_1 is increased to 80, the subsampling method yields satisfactory estimation results for all m values for DGP1. As to DGP3, the estimation results are adequate for $m = 60$ and $m = 80$, but there are undercoverage for $m = 20$ and $m = 40$. The key insight is that when T_1 is moderate, a larger $m \leq T_1$ is a better choice because if m is small and close to N, the subsampling method suffers from a large estimation variance. When T_1 is large, $m = T_1/2$ to $m = T_1$ are reasonable choices for m.

5 Conclusion

We filled a research void in inference theory for the SC and the MSC methods when data are nonstationary with nonlinear trends. We developed the inference theory for the SC/MSC ATT estimator when the

Table 5: Coverage Probabilities for the SC Method ($T_2 = 20$, $N = 11$)

Cov. Pr.	50%	80%	90%	95%	50%	80%	90%	95%
	$T_1 = 40$							
m	DGP1				DGP3			
20	0.512	0.820	0.900	0.944	0.432	0.698	0.832	0.892
40	0.508	0.812	0.886	0.936	0.524	0.800	0.900	0.946
	$T_1 = 80$							
	DGP1				DGP3			
20	0.476	0.774	0.892	0.944	0.418	0.694	0.824	0.894
40	0.488	0.784	0.890	0.932	0.488	0.742	0.856	0.912
60	0.524	0.832	0.916	0.950	0.528	0.762	0.878	0.936
80	0.528	0.800	0.900	0.948	0.530	0.824	0.908	0.962

outcome variables exhibit nonlinear trend of unknown form. Our inference theory covers different numbers of pre- and post-treatment periods and treatment and control units. With the new inference theory and results derived in this paper, researchers can use the SC and the MSC methods to estimate the ATT and conduct inference for a variety of data types, including stationary, unit-root non-stationary, and nonlinear trend non-stationary processes.

References

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

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

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

Athey S, Imbens G, Pham T, Wager S (2017) Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review* 107(5):278-281.

Chernozhukov V, Wuthrich K, Zhu Y (2019) An exact and robust conformal inference method for counterfactual and synthetic controls arXiv:1712.09089.

Dekimpe MG, Hanssens DM (1995) Empirical generalizations of market evolution and stationarity. *Marketing Science* 14(3):G109-G121.

Doudchenko N, Imbens G (2016) Balancing, regression, difference-in-differences and synthetic control meth-

ods: A synthesis. Working Paper. NBER.

Fang, Z, Santos, A (2018) Inference on Directionally Differentiable Functions. The Review of Economic Studies, 86, 377–412.

Hamilton, J (1994). *Time Series Analysis*. Princeton University Press.

Hong, H, Li J (2018). The numerical delta method. *Journal of Econometrics* 206 (2), 379-394.

Hong, H, Li J (2020). The numerical bootstrap. Forthcoming in *Annals of Statistics*.

Hsiao C, Ching HS, Wan, SK (2012) A panel data approach for program evaluation: Measuring the benefit of political and economic integration of Hong Kong with mainland China. *Journal of Applied Econometrics* 27, 705-740.

Li K (2020) Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of American Statistical Association*, forthcoming.

Politis DN, Romano JP, and Wolf M (1999) Subsampling. Springer.

Wooldridge, J (1991) *Notes on regression with difference-stationary data*. Michigan State University.

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. New York: Elsevier.

Appendix A: Proofs of Theorems 3.1 and 3.2

A.1 Assumptions

Before delving into the proofs of the theorems, we first make some assumptions:

Assumption 1 (i) y_{it}^0 is generated by (2.1) for $i = 1, ..., N$ and $t = 1, ..., T$; (ii) u_{jt} is a zero mean, serially uncorrelated stationary process (in t) with finite fourth moment (i.e., $E(u_{jt}^4)$ is finite), and f_t and u_{js} are uncorrelated for all $t, s \in \{1, ..., T\}$ and $j \in \{1, ..., N\}$; (iii) $e_{1t} = y_{1t}^0 - x_t'\beta_0$ is a zero mean, finite variance stationary process and that $T^{-1/2}\sum_{t=1}^{T} e_{1t} \xrightarrow{d} N(0, \sigma_e^2)$ as $T \to \infty$; (iv) Let $\eta_T = T_2/T_1 \to \eta$ as $T_1, T_2 \to \infty$, where $\eta \geq 0$ is a negative finite constant.

Assumption 2 (i) The nonlinear trend factor f_t satisfies the condition that for $j = 1, 2$, $T_1^{-1}\sum_{t=1}^{T_1} f_t^j/(g(T_1))^j \to b_j$ in probability as $T_1 \to \infty$, where $b_j > 0$ is a positive constant, and $b_2 > b_1^2$, $g(t)$ is an unbounded monotone increasing function; (ii) $\lim_{T_1,T_2\to\infty} T(g(T) - g(T_1))/[\sqrt{T_1T_2}g(T_1)] \to c_g$, where $c_g \geq 0$ is a finite constant.

Assumption 1 is quite standard. We provide a few examples for f_t that satisfies Assumption 2. A general principle is to choose the leading term of f_t to be $g(t)$. For example, if t^ν is the leading term of f_t, where $0 < \nu < \infty$ is an arbitrary positive constant.[4] We choose $g(t) = t^\nu$. Assumption 2 (i) holds because using $g(T_1) = T_1^\nu$ and $f_t \approx t^\nu$, we obtain $(T_1 g^j(T_1))^{-1}\sum_{t=1}^{T_1} f_t^j \approx T_1^{-(j\nu+1)}\sum_{t=1}^{T_1} t^{j\nu} \to 1/(1+j\nu)$ for $j = 1, 2$. So that $b_2 - b_1^2 = 1/(1+2\nu) - 1/(1+\nu)^2 = \nu^2/[(1+2\nu)(1+\nu)^2] > 0$. To check for Assumption 2 (ii) we write $T = T_1 + T_2 = T_1(1 + \eta_T)$, where $\eta_T = T_2/T_1$. We first consider the case that $\eta_T \to \eta > 0$. For $g(t) = t^\nu$,

$$\begin{aligned} \frac{T(g(T) - g(T_1))}{\sqrt{T_1T_2}g(T_1)} &= \frac{(1+\eta_T)}{\sqrt{\eta_T}}\left(\frac{(1+\eta_T)^\nu T_1^\nu - T_1^\nu}{T_1^\nu}\right) \\ &= \frac{(1+\eta_T)((1+\eta_T)^\nu - 1)}{\sqrt{\eta_T}} \\ &\to \frac{(1+\eta)((1+\eta)^\nu - 1)}{\sqrt{\eta}} \end{aligned} \tag{A.1}$$

as $T_1, T_2 \to \infty$.

Next, we consider the case $\eta = 0$. In this case, using L'Hospital's rule, it is straightforward to show that the above limit, as $\eta_T \to 0$, is zero. Therefore, for polynomial class (the power index can be non-integer) f_t, Assumption 2 holds.

Inspecting the proof of Theorem 2.1 we notice that Assumption 2 can be relaxed to allow that, for $j = 1, 2$, $T_1^{-1}\sum_{t=1}^{T_1} f_t^j/(g(T_1))^j$ converges to well defined random variable so that drift-less unit root

[4] We only need the leading term of f_t to be t^ν. For example, we can have $f_t = \sum_{j=1}^{L} a_j t^{\nu_j} + \epsilon_t$, where for all $j = 1, ..., L$, a_j is a constant with $a_L \neq 0$, without loss of generality we set $a_L = 1$, $0 \leq \nu_1 < \nu_2 < ... < \nu_L = \nu$ (ν_j is a constant but not necessarily an integer), ϵ_t is a stationary process with zero mean and finite variance. As $t \to \infty$, t^ν is the leading term of f_t.

factors are permitted. But this will make the proofs substantially more involved. Therefore, we choose the current Assumption 2 for brevity of proofs. We would like to mention that Assumption 2 allows f_t to be unit root process with drift. Because a unit root process with drift has a linear time trend as its dominating component. Therefore, Assumption 2 holds with $g(t) = t$.

We present a Lemma below which will be used to prove Theorems 3.1 and 3.2.

Lemma A.1 *Let C_{T_1} be the $N \times N$ invertible matrix defined in Web Appendix C (also in Appendix B for the $N = 3$ case). Under Assumptions 1 and 2,*

(i) $(C'_{T_1})^{-1} \sum_{t=1}^{T_1} x_t x_t' C_{T_1}^{-1} \xrightarrow{p} J$, where J is an $N \times N$ positive definite matrix;

(ii) $T_2^{-1} \sum_{t=T_1+1} x_t' C_{T_1}^{-1} \to B$, where B is an $1 \times N$ row vector of constants.

Proof of Lemma A.1 (i):

For expositional brevity, we consider the case of $N = 3$ and derive an explicit expression for J. Assuming that y_{jt} is generated by (B.1) with $f_t = t$. We can check that

$$J = \lim_{T_1 \to \infty} J_{T_1} = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 2\sigma_u^2 & 0 \\ 1/2 & 0 & 1/3 \end{pmatrix}$$

because (see Appendix B for the definition of C_{T_1} and $C_{T_1}^{-1}$, $\sum \equiv \sum_{t=1}^{T_1}$ below)

$$\begin{aligned}
J_{T_1} &= (C'_{T_1})^{-1} \sum_{t=1}^{T_1} x_t x_t' C_{T_1}^{-1} \\
&= \frac{1}{\sqrt{T_1}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & \frac{1}{T_1} \end{pmatrix} \begin{pmatrix} T_1 & \sum y_{2t} & \sum y_{3t} \\ \sum y_{2t} & \sum y_{2t}^2 & \sum y_{2t} y_{3t} \\ \sum y_{3t} & \sum y_{3t} y_{2t} & \sum y_{3t}^2 \end{pmatrix} \frac{1}{\sqrt{T_1}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & \frac{1}{T_1} \end{pmatrix} \\
&= \frac{1}{\sqrt{T_1}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & \frac{1}{T_1} \end{pmatrix} \begin{pmatrix} T_1 & \sum(t+u_{2t}) & \sum(t+u_{3t}) \\ \sum(t+u_{2t}) & \sum(t+u_{2t})^2 & \sum(t+u_{2t})(t+u_{3t}) \\ \sum(t+u_{3t}) & \sum(t+u_{2t})(t+u_{3t}) & \sum(t+u_{3t})^2 \end{pmatrix} \frac{1}{\sqrt{T_1}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & \frac{1}{T_1} \end{pmatrix} \\
&= \frac{1}{T_1} \begin{pmatrix} T_1 & \sum(u_{2t}-u_{3t}) & T_1^{-1}\sum(t+u_{3t}) \\ \sum(u_{2t}-u_{3t}) & \sum(u_{2t}-u_{3t})^2 & T_1^{-1}\sum(t+u_{3t})(u_{2t}-u_{3t}) \\ T_1^{-1}\sum(t+u_{3t}) & T_1^{-1}\sum(t+u_{3t})(u_{2t}-u_{3t}) & T_1^{-2}\sum(t+u_{3t})^2 \end{pmatrix} \\
&\xrightarrow{p} \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 2\sigma_u^2 & 0 \\ 1/2 & 0 & 1/3 \end{pmatrix} \\
&\equiv J, \qquad \text{(A.2)}
\end{aligned}$$

where the last convergence result follows from $T_1^{-2} \sum_{t=1}^{T_1} t \to 1/2$ and $T_1^{-2} \sum_{t=1}^{T_1} t^2 \to 1/3$.

The above derivation replies on $f_t = t$. Next, we consider a general f_t. From Appendix B we know that $C_{T_1}^{-1} = \frac{1}{\sqrt{T_1}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & \frac{1}{g(T_1)} \end{pmatrix}$. To compute J we use $y_{jt} = f_t + u_{jt}$. Therefore, for a general f_t, (A.2) is modified to

$$J_{T_1} = (C'_{T_1})^{-1} \sum_{t=1}^{T_1} x_t x_t' C_{T_1}^{-1}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{T_1}}\begin{pmatrix}1&0&0\\0&1&-1\\0&0&\frac{1}{g(T_1)}\end{pmatrix}\begin{pmatrix}T_1 & \sum(t+u_{2t}) & \sum(t+u_{3t})\\ \sum(f_t+u_{2t}) & \sum(f_t+u_{2t})^2 & \sum(f_t+u_{2t})(f_t+u_{3t})\\ \sum(f_t+u_{3t}) & \sum(f_t+u_{2t})(f_t+u_{3t}) & \sum(f_t+u_{3t})^2\end{pmatrix}\frac{1}{\sqrt{T_1}}\begin{pmatrix}1&0&0\\0&1&0\\0&-1&\frac{1}{g(T_1)}\end{pmatrix}\\
&= \frac{1}{T_1}\begin{pmatrix}T_1 & \sum(u_{2t}-u_{3t}) & \frac{1}{g(T_1)}\sum(t+u_{3t})\\ \sum(u_{2t}-u_{3t}) & \sum(u_{2t}-u_{3t})^2 & \frac{1}{g(T_1)}\sum(t+u_{3t})(u_{2t}-u_{3t})\\ \frac{1}{g(T_1)}\sum(t+u_{3t}) & \frac{1}{g(T_1)}\sum(t+u_{3t})(u_{2t}-u_{3t}) & \frac{1}{g^2(T_1)}\sum(t+u_{3t})^2\end{pmatrix}\\
&\xrightarrow{p} \begin{pmatrix}1&0&b_1\\0&2\sigma_u^2&0\\b_1&0&b_2\end{pmatrix}\\
&\equiv J, \qquad \text{(A.3)}
\end{aligned}
$$

because $(T_1 g(T_1))^{-1}\sum_{t=1}^{T_1} f_t \to b_1$ and $(T_1 g^2(T_1))^{-1}\sum_{t=1}^{T_1} f_t^2 \to b_2$ by Assumption 2 (i). This proves Lemma A.1 (i).

Proof of Lemma A.1 (ii):

Lemma A.1 (ii) states that $T_2^{-1/2}\sum_{t=T_1+1}^{T} x_t' C_{T_1}^{-1} \xrightarrow{p} B$, where B is a $1\times N$ row vector of constants. For expositional simplicity, we consider the simple case of $N=3$ and that y_{jt} is generated by (B.1). We first consider the case that $f_t = t$ (then $g(t)=t$ satisfies Assumption 2). We will provide an explicit expression for B and C_{T_1} to this simple case.

Define a diagonal matrix $D_{T_1} = \sqrt{T_1}\, diag(1,1,g(T_1)) = \sqrt{T_1}\, diag(1,1,T_1)$ because $g(T_1)=T_1$. Using the 3×3 transformation matrix A defined in (B.7) and notice that $C_{T_1} = D_{T_1} A$, we obtain

$$
\begin{aligned}
C_{T_1}^{-1} &= A^{-1} D_{T_1}^{-1}\\
&= \begin{pmatrix}1&0&0\\0&1&0\\0&-1&1\end{pmatrix}\frac{1}{\sqrt{T_1}}\begin{pmatrix}1&0&0\\0&1&0\\0&0&\frac{1}{T_1}\end{pmatrix}\\
&= \frac{1}{\sqrt{T_1}}\begin{pmatrix}1&0&0\\0&1&0\\0&-1&\frac{1}{T_1}\end{pmatrix}. \qquad \text{(A.4)}
\end{aligned}
$$

Using (A.4) we obtain

$$
\begin{aligned}
\frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} x_t' C_{T_1}^{-1} &= \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} (\,1,\ y_{2t},\ y_{3t}\,)\ C_{T_1}^{-1}\\
&= \frac{1}{\sqrt{T_2 T_1}}\sum_{t=T_1+1}^{T} (\,1,\ u_{2t}-u_{3t},\ y_{3t}/T_1\,)\\
&\approx \left(\sqrt{T_2/T_1},\, O_p(T_1^{-1/2}),\, \frac{(2+T_2/T_1)T_2}{2\sqrt{T_1 T_2}}\right)\\
&\xrightarrow{p} (\,\sqrt{\eta},\ 0,\ \sqrt{\eta}\,(2+\eta)/2\,)\\
&\equiv B, \qquad \text{(A.5)}
\end{aligned}
$$

where $\eta = \lim_{T_1,T_2\to\infty} T_2/T_1$, the third approximation relationship follows from $\sum_{t=T_1+1}^{T} y_{3t} \approx \sum_{t=T_1+1}^{T} t \approx (1/2)(T^2 - T_1^2) = (1/2)(T+T_1)T_2 = (2T_1+T_2)T_2/2$ because $f_t = t$ is the leading term of y_{3t}.

Next, we check the general f_t case but still focus on $N = 3$ case for brevity. In Appendix B we show that C_{T_1} defined in (A.4) needs to be modified by replacing the third diagonal element $\sqrt{T_1^3}$ by $\sqrt{T_1}g(T_1)$. This amounts to change y_{3t}/T_1 in (A.5) by $y_{3t}/g(T_1)$. Since f_t is the leading term of y_{3t}, we obtain the leading term of the third column of (A.5) as (we use $B_{3,T}$ to denote it):

$$\begin{aligned} B_{3,T} &= \frac{1}{\sqrt{T_2 T_1}g(T_1)} \sum_{t=T_1+1}^{T} f_t \\ &= \frac{1}{\sqrt{T_2 T_1}g(T_1)} \left(\sum_{t=1}^{T} f_t - \sum_{t=1}^{T_1} f_t \right) \\ &\approx \frac{b_1}{\sqrt{T_2 T_1}g(T_1)} \left(Tg(T) - T_1 g(T_1) \right) \\ &= \frac{b_1}{\sqrt{T_2 T_1}g(T_1)} \left(T(g(T) - g(T_1)) + T_2 g(T_1) \right) \\ &\to b_1(c_g + \sqrt{\eta}) \end{aligned} \tag{A.6}$$

where the approximation "$\approx$" follows from Assumption 2 (i), and the last convergence result is due to Assumption 2 (ii). Therefore, for a general $g(t)$,

$$B = (\, \sqrt{\eta},\, 0,\, b_1(c_g + \sqrt{\eta})\,). \tag{A.7}$$

Thus, we have shown that Assumption 2 implies Lemma A.1 (ii). Note that the second element of B being zero is due to our choice that $c_j = 0$ for all $j = 1, ..., N$ in generating y_{jt}. It is easy to show that if $c_j \neq 0$, the second element of B is different from zero in general.

A.2 Proof of Theorem 3.1

From (3.4),

$$\begin{aligned} A_{HCW} &= \sqrt{T_2}(\hat{\Delta}_{HCW,1} - \bar{\Delta}_1) \\ &= -\frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} x_t'(\hat{\beta}_{OLS,T_1} - \beta) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} e_{1t} \\ &= A_{HCW,1} + A_{HCW,2}, \end{aligned} \tag{A.8}$$

where

$$\begin{aligned} A_{MSC,2} &= \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} e_{1t} \\ &\xrightarrow{d} N(0, \sigma_e^2) \end{aligned} \tag{A.9}$$

by Assumption 1 (iv), and

$$A_{HCW,1} = -\frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} x_t'(\hat{\beta}_{OLS,T_1} - \beta)$$

$$= \quad -\frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} x_t' C_{T_1}^{-1} C_{T_1} (\hat{\beta}_{OLS,T_1} - \beta)$$

$$\xrightarrow{d} \quad -B\, N(0,\Omega) = N(0, V_1), \tag{A.10}$$

by Lemma A.1 (ii) that $T_2^{-1/2} \sum_{t=T_1+1}^{T} x_t' C_{T_1}^{-1} \xrightarrow{p} B$ and by Theorem 2.1 (ii) that $C_{T_1}(\hat{\beta}_{OLS,T_1} - \beta) \xrightarrow{d} N(0,\Omega)$, where $V_1 = B\Omega B'$. See Appendices B and C for a proof of Theorem 2.1 for $N = 3$ and the general cases, respectively.

Finally, $A_{HCW,1}$ and $A_{HCW,2}$ are uncorrelated and therefore, asymptotically independent. By virtue of a central limit theore, we obtain that

$$A_{HCW} = \sqrt{T_2}(\hat{\Delta}_{1,HCW} - \bar{\Delta}_1) \xrightarrow{d} N(0, V), \tag{A.11}$$

where $V = V_1 + \sigma_e^2$. This completes the proof of 3.1.

To construct an estimate for V, we first make a strong assumption that e_{1t} is iid $N(0, \sigma_e^2)$ and is independent of X (this assumption will be relaxed later). Conditional on X, the standard least squares theory yields

$$A_{HCW,1}|X \overset{d}{\sim} N(0, \sigma_e^2 \psi'(X'X)^{-1}\psi/T_2),$$

where $\psi = \sum_{t=T_2+1}^{T} x_t$.

Therefore,

$$\frac{A_{HCW,1}}{\sqrt{\sigma_e^2 \psi'(X'X)^{-1}\psi}/T_2}|X \overset{d}{\sim} N(0,1). \tag{A.12}$$

Now because the right-hand-side of (A.12) is unrelated to X, unconditionally,

$$\frac{A_{HCW,1}}{\sqrt{\sigma_e^2 \psi'(X'X)^{-1}\psi/T_2}} \overset{d}{\sim} N(0,1). \tag{A.13}$$

Next, we can see that $A_{HCW,2} \overset{d}{\sim} N(0, \sigma_e^2)$. Therefore,

$$\frac{A_{HCW,2}}{\sqrt{\sigma_e^2}} \overset{d}{\sim} N(0,1). \tag{A.14}$$

Finally, since $A_{HCW,1}$ and $A_{HCW,2}$ are independent with each other,

$$\frac{A_{HCW,1} + A_{HCW,2}}{\sqrt{\sigma_e^2 \psi'(X'X)^{-1}\psi/T_2 + \sigma_e^2}} \overset{d}{\sim} N(0,1). \tag{A.15}$$

The independence assumption can be relaxed to be uncorrelated between x_t and e_{1t}. The normality assumption can be removed. When T_1 and T_2 are large, both $A_{HCW,1}$ and $A_{HCW,2}$ are approximately normally distributed by virtue of central limit theorem arguments.

To make (A.15) operational, we replace unknown quantity σ_e^2 by a consistent estimator $\hat{\sigma}_e^2$. This step leads to

$$\frac{\hat{\Delta}_{HCW,1} - \bar{\Delta}_1}{\sqrt{\hat{V}}} \xrightarrow{d} N(0,1), \tag{A.16}$$

where $\hat{V} = \hat{\sigma}_e^2 \psi'(X'X)^{-1}\psi/T_2 + \hat{\sigma}_e^2$.

A.3 Proof of Theorem 3.2

In this Section we derive the limiting distribution of MSC ATT estimator. The projection theory we use was first developed by Zarantonello (1971) and recently extended to more general setting by Fang and Santos (2018). Let $C_{T_1} = D_{T_1} A$ be defined as in Appendix B for $N = 3$ case and in Web Appendix C for the general case. We use $\Lambda_{MSC} = \{\beta \in \mathcal{R}^N : \quad \beta_j \geq 0 \text{ for } j = 2, ..., N\}$ to denote the MSC constrained set for β, and $T_{\Lambda_{MSC},\beta}$ to denote the asymptotic range (as $T_1 \to \infty$) of $C_{T_1}(\hat{\beta}_{MSC,T_1} - \beta)$ (the so called Tangent cone of Λ_{MSC} evaluated at β). We introduce some more notation. Denote by $Z_{OLS,T_1} = C_{T_1}(\hat{\beta}_{OLS,T_1} - \beta_0)$, $Z_{MSC,T_1} = C_{T_1}(\hat{\beta}_{MSC,T_1} - \beta_0)$, and we use Z_{OLS} and Z_{MSC} to denote the limiting (as $T_1 \to \infty$) distribution of Z_{OLS,T_1} and Z_{MSC,T_1}, respectively. We use Π to denote projection and define a projection of $\lambda \in \mathcal{R}^N$ onto a convex set D as $\Pi_D \lambda = \arg\min_{\theta \in D}(\theta - \lambda)' J (\theta - \lambda)$, where $J = plim_{T_1 \to \infty} J_{T_1}$, and $J_{T_1} = (C'_{T_1})^{-1} X'X C_{T_1}^{-1}$ is an $N \times N$ positive definite matrix. Li (2019) shows that $\hat{\beta}_{MSC,T_1}$ and $\hat{\beta}_{OLS,T_1}$ are connected by another projection with a finite sample weight $X'X$, rather than the limiting quantity J as a weight function in the following projection relationship:

$$\begin{aligned}
\hat{\beta}_{MSC,T_1} &= \arg\min_{\lambda \in \Lambda_{MSC}} (\lambda - \hat{\beta}_{OLS,T_1})' X'X (\lambda - \hat{\beta}_{OLS,T_1}) \\
&= \arg\min_{\lambda \in \Lambda_{MSC}} (\lambda - \hat{\beta}_{OLS,T_1})' C'_{T_1} (C'_{T_1})^{-1} X'X C_{T_1}^{-1} C_{T_1} (\lambda - \hat{\beta}_{OLS,T_1}) \\
&= \arg\min_{\lambda \in \Lambda_{MSC}} (C_{T_1}(\lambda - \beta_0) - C_{T_1}(\hat{\beta}_{OLS,T_1} - \beta_0))' J_{T_1} (C_{T_1}(\lambda - \beta_0) - C_{T_1}(\hat{\beta}_{OLS,T_1} - \beta_0)) \\
&= \arg\min_{\lambda \in \Lambda_{MSC}} (Z_\lambda - Z_{OLS,T_1})' J_{T_1} (Z_\lambda - Z_{OLS,T_1}), \qquad \text{(A.17)}
\end{aligned}$$

where $Z_\lambda = C_{T_1}(\lambda - \beta_0)$. Let $Z_{MSC,T_1} = C_{T_1}(\hat{\beta}_{MSC,T_1} - \beta_0)$. Since Z_λ takes value in $T_{\Lambda_{MSC},\beta_0}$, we can re-write (A.17) as

$$Z_{MSC,T_1} = \arg\min_{Z_\lambda \in T_{\Lambda_{MSC},\beta_0}} (Z_\lambda - Z_{OLS,T_1})' J_{T_1} (Z_\lambda - Z_{OLS,T_1}). \qquad \text{(A.18)}$$

Letting $T_1 \to \infty$ in (A.18) we obtain

$$\begin{aligned}
Z_{MSC} &= \arg\min_{Z_\lambda \in T_{\Lambda_{MSC},\beta_0}} (Z_\lambda - Z_{OLS})' J (Z_\lambda - Z_{OLS}) \\
&= \Pi_{T_{\Lambda_{MSC},\beta_0}} Z_{OLS}, \qquad \text{(A.19)}
\end{aligned}$$

where $Z_{OLS} \overset{d}{\sim} N(0, \Omega)$ is the limiting distribution of $\sqrt{T_1}(\hat{\beta}_{OLS,T_1} - \beta_0)$ in Theorem 2.1. This proves Theorem 3.2. Our above proof method deriving (A.19) is similar to Andrews (1999) who considers asymptotic theory for a general class of constrained estimators. However, our nonlinear trend process with unknown form is not covered in Andrews (1999).

Appendix B: Proof of Theorems 2.1 and 2.2 with $N = 3$

In Appendices B.1 and B.2, we provide proofs of Theorems 2.1 and 2.2 for the simple case of $N = 3$. The purpose of presenting separate proofs for $N = 3$ is to provide simple proofs with much less notation. These proofs are easy to follow and provide good intuition for understanding the underlying theory. We provide proofs for the general case in the Web Appendix C.

B.1 Proof of Theorem 2.1 for $N = 3$

We consider a simple case where $N = 3$ and y_{jt} are nonlinear trend processes for $j = 1, 2, 3$:

$$y_{jt} = f_t + u_{jt}, \qquad i = 1, 2, 3;\ t = 1, ..., T_1, \tag{B.1}$$

where u_{jt} is iid $(0, \sigma^2)$. Compared to the general setup in (C.1), here we choose $c_j = 0$ and $d_j = 1$ purely for expositional simplicity.

We consider the following regression model[5]

$$y_{1t} = x_t'\beta + e_t, \qquad t = 1, ..., T_1, \tag{B.2}$$

where $x_t = (1, y_{2t}, y_{3t})$, and e_t is a zero mean stationary idiosyncratic error term.

Let $\hat{\beta} = \hat{\beta}_{OLS,T_1}$ denote the least squares estimator of β. We want to find the limiting distribution of $\sqrt{T_1}(\hat{\beta} - \beta)$. However, since the regressors $y_{2t} = f_t + u_{2t}$ and $y_{3t} = f_t + u_{3t}$ are asymptotically collinear, the standard least squares theory does not apply here.

Consider the least squares objective function. From (B.1), the objective function becomes:

$$\begin{aligned} S(\beta) &\equiv \sum_{t=1}^{T_1}(y_{1t} - x_t'\beta)^2 = \sum_{t=1}^{T}\left(f_t + u_{1t} - \left(\beta_1 + \sum_{j=2}^{3}\beta_j(f_t + u_{jt})\right)\right)^2 \\ &= \sum_{t=1}^{T}(-\beta_1 - (\beta_2 + \beta_3 - 1)f_t + u_{1t} - (\beta_2 u_{2t} + \beta_3 u_{3t}))^2 \\ &= \sum_{t=1}^{T_1}(-\beta_1 - \alpha f_t + \epsilon_t)^2 \end{aligned} \tag{B.3}$$

where $\alpha = (\beta_2 + \beta_3) - 1$, $\epsilon_t = u_{1t} - \beta_2 u_{2t} - \beta_3 u_{3t}$. We reparametrize $(\beta_1, \beta_2, \beta_3)$ to $\delta = (\beta_1, \beta_2, \alpha)'$ with $\alpha = \beta_2 + \beta_3 - 1$.

We need to replace β_3 by $\beta_3 = \alpha - \beta_2 + 1$ in (B.3). Doing this leads to

$$S(\delta) \equiv \sum_{t=1}^{T_1}(y_{1t} - x_t'\beta)^2$$

[5] Appendix C provides a proof that β is uniquely defined. Using (B.1) one can show that $\beta = (0, 1/2, 1/2)'$.

$$
\begin{aligned}
&= \sum_{t=1}^{T_1} (u_{1t} - u_{3t} - \beta_1 - \beta_2(u_{2t} - u_{3t}) - \alpha(f_t + u_{3t}))^2 \\
&= \sum_{t=1}^{T_1} (z_t - v_t'\delta)^2 , \qquad \text{(B.4)}
\end{aligned}
$$

where $z_t = u_{1t} - u_{3t}$, $v_t = (1, u_{2t} - u_{3t}, f_t + u_{3t})'$, and $\delta = (\beta_1, \beta_2, \alpha)'$. To obtain a concrete result without loss of generality, we assume that $f_t = t$. Let $D_{T_1} = \sqrt{T_1}\, diag(1,\, 1,\, T_1)$. It is well established (e.g., Hamilton 1994, Chapter 16) that

$$
D_{T_1}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \Omega), \qquad \text{(B.5)}
$$

where Ω is a 3×3 positive definite matrix.

It can be easily checked that for a general f_t, the diagonal matrix D_{T_1} needs to be modified to $D_{T_1} = \sqrt{T_1}\, diag(1,\, 1,\, g(T_1))$. Then (B.5) holds, and of course with a difference expression for Ω because the specific definition of the positive definite matrix Ω depends on the functional form of $g(t)$, which in turn depends on the nonlinear functional form of f_t.

Note that the transformation from $\beta = (\beta_1, \beta_2, \beta_3)'$ to $\delta = (\beta_1, \beta_2, \alpha)'$ is very simple, we only change $\beta_3 \mapsto \alpha = \beta_2 + \beta_3 - 1$. Its inverse transformation is $\alpha \mapsto \beta_3 = \alpha - \beta_2 + 1$. To facility the derivation of the large sample theory of $\hat{\beta}$, we write this transformation formally below. The reparameterization transformation can be written as:

$$
\delta = A\beta + h \qquad \text{and} \qquad \beta = A^{-1}(\delta - h), \qquad \text{(B.6)}
$$

where A, h and A^{-1} are defined by

$$
A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \qquad h = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} \qquad \text{and} \qquad A^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}. \qquad \text{(B.7)}
$$

Define a 3×3 non-singular matrix $C_{T_1} = D_{T_1} A$. Therefore, $C_{T_1}^{-1} = A^{-1} D_{T_1}^{-1}$. Then,

$$
\begin{aligned}
C_{T_1}(\hat{\beta} - \beta) &= D_{T_1} A(\hat{\beta} - \beta) \\
&= D_{T_1}(\hat{\delta} - \delta) \\
&\xrightarrow{d} N(0, \Omega), \qquad \text{(B.8)}
\end{aligned}
$$

by (B.5), where the second equality follows from $\hat{\delta} = A\hat{\beta} + h$ and $\delta = A\beta + h$.

Because $\alpha = \beta_2 + \beta_3 - 1$, we have $\hat{\alpha} - \alpha = (\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3)$. Applying this result to (B.5) (with $f_t = t$) leads to

$$
D_{T_1}(\hat{\delta} - \delta) = \begin{pmatrix} \sqrt{T_1}(\hat{\beta}_1 - \beta_1) \\ \sqrt{T_1}(\hat{\beta}_2 - \beta_2) \\ \sqrt{T_1^3}\left((\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3)\right) \end{pmatrix} \xrightarrow{d} N(0, \Omega). \qquad \text{(B.9)}
$$

So that our theory predict that while $\hat{\beta}_j - \beta_j$ converges to zero at the rate of $T_1^{-1/2}$ for $j = 1, 2, 3$,[6] $(\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3)$ converges to zero at a much faster rate of $T_1^{-3/2}$ (when $f_t = t$). We numerically verify this theoretical analysis in Web Appendix E. Specifically, from (B.1) and (B.2) and by the uniqueness formula for β developed in Web Appendix D, it is easy to check that the true value of $\beta = (0, 1/2, 1/2)'$. Our theory predicts that $\hat{\beta}_1 - \beta_1 = \hat{\beta}_1 = O_p(T_1^{-1/2})$, $\hat{\beta}_j - \beta_j = \hat{\beta}_j - 1/2 = O_p(T_1^{-1/2})$ for $j = 2, 3$, and $(\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3) = (\hat{\beta}_2 + \hat{\beta}_3) - 1 = O_p(T_1^{-3/2})$. Simulation results reported in Web Appendix E strongly support our theoretical prediction.

We make a comment on the above proof of Theorem 2.1. The proof does not rely on our simplifying assumption $f_t = t$. We can replace the linear trend assumption $f_t = t$ by any nonlinear trend of unknown form provided that Assumption 2 holds. The above derivation goes through with the following modifications: (i) The diagonal matrix D_{T_1} becomes $D_{T_1} = \sqrt{T_1}\, diag(\,1,\, 1,\, g(T_1)\,)$ and $\sqrt{T_1^3}$ in (B.9) needs to be replaced by $\sqrt{T_1} g(T_1)$. The rate of convergence is changed to $(\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3) = (\hat{\beta}_2 + \hat{\beta}_3) - 1 = O_p((\sqrt{T_1}\, g(T_1))^{-1}) = o_p(T_1^{-1/2})$.

B.2 Proof of Theorem 2.2 for $N = 3$

Define $\delta_* = (\delta_1, \delta_2)'$ and $\beta_* = (\beta_1, \beta_2)'$. For our example, $\delta_* \equiv \beta_*$ because $\delta_j = \beta_j$ for $j = 1, 2$. Therefore, (B.5) implies that

$$\sqrt{T_1}(\hat{\beta}_* - \beta_*) \equiv \sqrt{T_1}(\hat{\delta}_* - \delta_*) \xrightarrow{d} N(0, \Sigma_*), \tag{B.10}$$

where Σ_* is a 2×2 positive definite matrix. This relationship implies that $\text{rank}(\Sigma) \geq 2$ because $\text{rank}(\Sigma) \geq \text{rank}(\Sigma_*) = 2$.

From $\hat{\beta}_3 = \hat{\alpha} + 1 - \hat{\beta}_2$, we obtain

$$\begin{aligned} \sqrt{T_1}(\hat{\beta}_3 - \beta_3) &= \sqrt{T_1}(\hat{\alpha} - \alpha) - \sqrt{T_1}(\hat{\beta}_2 - \beta_2) \\ &= -\sqrt{T_1}(\hat{\beta}_2 - \beta_2) + o_p(T_1^{-1/2}), \end{aligned} \tag{B.11}$$

because $\hat{\alpha} - \alpha = O_p((T_1 g(T_1))^{-1/2}) = o_p(T_1^{-1/2})$.

Denote by $Z \overset{d}{\sim} N(0, M)$ ($M > 0$ is a constant) as the limiting normal distribution of $\sqrt{T_1}(\hat{\beta}_2 - \beta_2)$. From (B.11),

$$\begin{pmatrix} \sqrt{T_1}(\hat{\beta}_2 - \beta_2) \\ \sqrt{T_1}(\hat{\beta}_3 - \beta_3) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z \\ -Z \end{pmatrix} \tag{B.12}$$

whose limiting variance is not full rank because we use $M > 0$ to denote the variance of Z. Thus,

$$Var \begin{pmatrix} \sqrt{T_1}(\hat{\beta}_2 - \beta_2) \\ \sqrt{T_1}(\hat{\beta}_3 - \beta_3) \end{pmatrix} \to Var \begin{pmatrix} Z \\ -Z \end{pmatrix} = M \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \tag{B.13}$$

[6] Theorem 2.2 implies that $\hat{\beta}_3 - \beta_3$ converges to zero at the rate $T_1^{-1/2}$ as well.

which has rank one and is not invertible. This condition implies that $\Sigma < 3$.

Summarizing the above results, we know that $\text{rank}(\Sigma) \geq 2$ and $\text{rank}(\Sigma) < 3$. Therefore, $\text{rank}(\Sigma) = 2 = N - 1$ because $N = 3$.

Web Appendix C: Proof of Theorems 2.1 and 2.2 (general case)

C.1 Proof of Theorem 2.1 for the General Case

In the absence of treatment, the outcome variables are generated by a nonlinear factor model:

$$y_{jt}^0 = c_j + d_j f_t + u_{jt}, \qquad j = 1, ..., N;\ t = 1, ..., T, \tag{C.1}$$

where u_{jt} are zero mean, finite variance and serially uncorrelated stationary idiosyncratic error terms, we allow f_t to be random, and f_t and u_{jt} are uncorrelated for all $j = 1, ..., N$, $t = 1, ..., T$. We assume that f_t has a nonlinear trend of unknown form. For example, $f_t = a + bt^\nu + \eta_t$, where a, b and ν ($\nu > 0$) are constants, η_t is a zero mean, finite variance stationary process. We do not impose linear trend restriction in (C.1).

We estimate β from the following regression:

$$y_{1t} = x_t'\beta + e_{1t}, \qquad t = 1, ..., T_1, \tag{C.2}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})$, and e_{1t} is a zero mean stationary idiosyncratic error term. Because y_{jt} shares the common trend f_t, which is the dominating component to all $y_{2t}, ..., y_{Nt}$, these regressors are asymptotically collinear. The standard least squares theory does not apply to model (C.2).

From (C.2), we have $e_{1t} = e_{1t}(\beta) = y_{1t} - x_t'\beta$. Using (C.1) and (C.2), we can represent e_{1t} in terms of f_t and u_{jt} for $j = 1, ..., N$ as follows:

$$\begin{aligned}
e_{1t} &= y_{1t} - x_t'\beta \\
&= y_{1t} - (\beta_1 + \sum_{j=2}^{N} \beta_j y_{jt}) \\
&= (c_1 + d_1 f_t + u_{1t}) - (\beta_1 + \sum_{j=2}^{N} \beta_j (c_j + d_j f_t + u_{jt})) \\
&= -(\beta_1 + \sum_{j=2}^{N} \beta_j c_j - c_1) - (\sum_{j=2}^{N} \beta_j d_j - d_1) f_t + u_{1t} - \sum_{j=2}^{N} \beta_j u_{jt} \\
&= -\alpha_1 - \alpha_2 f_t + \epsilon_t,
\end{aligned} \tag{C.3}$$

where

$$\alpha_1 = \beta_1 + \sum_{j=2}^{N} \beta_j c_j - c_1 \tag{C.4}$$

$$\alpha_2 = \sum_{j=2}^{N} \beta_j d_j - d_1 \tag{C.5}$$

$$\epsilon_t = u_{1t} - \sum_{j=2}^{N} \beta_j u_{jt}. \tag{C.6}$$

From (C.3), the least squares method involves choosing β to minimize

$$\sum_{t=1}^{T_1}(y_{1t} - x_t'\beta)^2 = \sum_{t=1}^{T_1} e_{1t}^2(\beta) = \sum_{t=1}^{T_1}(-\alpha_1 - \alpha_2 f_t + \epsilon_t)^2, \tag{C.7}$$

where $\alpha_1 = \alpha(\beta)$, $\alpha_2 = \alpha_2(\beta)$ and $\epsilon_t = \epsilon_t(\beta)$ are all linear in β and are defined in (C.4) to (C.6). Let $\hat{\beta} \equiv \hat{\beta}_{OLS,T_1}$ be the least squares estimator of β based on (C.7). The least squares theory for $\hat{\beta}$ is difficult due to asymptotic collinearity of different components of x_t (y_{jt}, $j = 2, ..., N$). However, in the last expression of (C.7), there is no collinearity problem because f_t is the only nonlinear trend process. All other variables are stationary. This condition suggests that we can re-parameterize

$$\beta = (\beta_1, \beta_2, \beta_3, ..., \beta_N)' \qquad \text{to} \qquad \delta = (\alpha_1, \alpha_2, \beta_3, \beta_4, ..., \beta_N)' \tag{C.8}$$

and consider the least squares estimator of the new $N \times 1$ vector parameter δ. Because both $\alpha_1 = \alpha_1(\beta)$ and $\alpha_2 = \alpha_2(\beta)$ are linear in β, the transformation from β to δ is linear. I.e.,

$$\delta = A\beta + h, \tag{C.9}$$

where $h = (-c_1, -d_1, \mathbf{0}_{1\times(N-2)})'$ is an $N \times 1$ vector with $\mathbf{0}_{1\times(N-2)}$ being a $1 \times (N-2)$ row vector of zeros, and A is an $N \times N$ matrix defined by

$$A = \begin{pmatrix} A_{11} & A_{12} \\ \mathbf{0}_{(N-2)\times 2} & I_{N-2} \end{pmatrix} \tag{C.10}$$

where I_{N-2} is an $(N-2) \times (N-2)$ identity matrix, $\mathbf{0}_{(N-2)\times 2}$ is a $(N-2) \times 2$ matrix of zeros, and

$$A_{11} = \begin{pmatrix} 1 & c_2 \\ 0 & d_2 \end{pmatrix} \quad \text{and} \quad A_{12} = \begin{pmatrix} c_3 & \dots & c_N \\ d_3 & \dots & d_N \end{pmatrix}. \tag{C.11}$$

We show in Lemma C.1 that A is invertible so that

$$\beta = A^{-1}(\delta - h), \tag{C.12}$$

which is linear in δ. To minimize (C.7) with respect to δ, we need first replace β_2 in term of δ. From (C.5) we can express (without loss of generality, we assume $d_2 \neq 0$)

$$\beta_2 \quad = \quad \frac{1}{d_2}\left[\alpha_2 + d_1 - \sum_{j=3}^{N} d_j\beta_j\right]. \tag{C.13}$$

Replacing β_2 in (C.7) by the right-hand-side of (C.13) yields

$$\sum_{t=1}^{T_1}(z_t - v_t'\delta)^2 \tag{C.14}$$

where $z_t = u_{1t} - (d_1/d_2)u_{2t}$, and

$$v_t = (\,1,\, f_t + u_{2t}/d_2,\, 1 + u_{2t}/d_2, u_{3t} - (d_3/d_2)u_{2t}, ..., u_{Nt} - (d_N/d_2)u_{2t}\,)'.$$

Let $\hat{\delta} = (\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_3, ..., \hat{\beta}_N)'$ denote the least squares estimator of δ that minimizes (C.14). Because only the second component of v_t has a trend component f_t, it is easy to show that $\hat{\alpha}_2 - \alpha_2$ converges to zero faster than $T_1^{-1/2}$. For example, if $f_t = t$, it is well established that $\hat{\alpha}_2 - \alpha_2 = O_p(T_1^{-3/2})$ and all other parameter estimates have the usual $O_p(T_1^{-1/2})$ convergence rate (Hamilton 1994, Chapter 16). For the general f_t case, it can be shown that $\hat{\alpha}_2 - \alpha_2$ converges to zero at the rate of $(T_1 g(T_1))^{-1/2}$. Also since e_{1t} has no trend component, this implies that the true value of $\alpha_2 = 0$. From (C.5) we obtain $0 = \alpha_2 = \sum_{j=2}^N \hat{\beta}_j d_j - d_1$. Therefore, $d_1 = \sum_{j=2}^N \hat{\beta}_j d_j$. This leads to $\hat{\alpha}_2 = \sum_{j=2}^N \hat{\beta}_j d_j - d_1 = \sum_{j=2}^N d_j(\hat{\beta}_j - \beta_j) = O_p((T_1 g(T_1))^{-1/2})$. All individual component $\hat{\beta}_j - \beta_j = O_p(T_1^{-1/2})$ for $j = 1, ..., N$. This finishes the proof of Theorem 2.1 (i).

Define an $N \times N$ diagonal matrix[7]

$$D_{T_1} = \sqrt{T_1}\, diag(1, g(T_1), 1, ..., 1), \tag{C.15}$$

which has $(T_1 g(T_1))^{1/2}$ as its second diagonal element and all other diagonal elements are $T_1^{1/2}$. Similar to a linear trend model case (Hamilton (1994, Chapter 16), we can show that

$$D_{T_1}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \Omega), \tag{C.16}$$

where $\hat{\delta} = \hat{\delta}_{OLS,T_1}$ is the ordinary least squares estimator of δ using pretreatment data $t = 1, ..., T_1$, and Ω is a $N \times N$ positive definite matrix.

Define an invertible $N \times N$ matrix $C_{T_1} = D_{T_1} A$, where A is defined in (C.10). Using (C.9) we obtain

$$\begin{aligned} C_{T_1}(\hat{\beta} - \beta) &= D_{T_1} A(\hat{\beta} - \beta) \\ &= D_{T_1}(\hat{\delta} - \delta) \\ &\xrightarrow{d} N(0, \Omega), \end{aligned} \tag{C.17}$$

where the last convergence result follows from (C.16). This completes the proof of Theorem 2.1.

We now make a prediction based on the above theoretical analysis. If we let $f_t = t$ so that $g(T_1) = T_1$, and $c_j = 0$ and $d_j = 1$ for all $j = 1, ..., N$. Equations (C.4) and (C.5) ensure that e_{1t} to be a zero mean stationary process become $\beta_1 = 0$ and $\sum_{j=2}^N \beta_j = 1$. In addition, let $\hat{\beta}_j$ denote the OLS estimate of β_j for $j = 1, ..., N$. Our theory predict that $\hat{\beta}_j - \beta_j = O_p(T_1^{-1/2})$ and that $\sum_{j=2}^N \hat{\beta}_j - 1 = O_p(T_1^{-3/2})$. This is because $\hat{\alpha}_2 - \alpha_2 = O_p(T_1^{-3/2})$ and $\alpha_2 = \sum_{j=2}^N \beta_j - 1$ so that $\hat{\alpha}_2 - \alpha_2 = \sum_{j=2}^N (\hat{\beta}_j - \beta_j) = O_p(T_1^{-3/2})$. Simulation results reported in Web Appendix E strongly support our theoretical prediction.

[7] Here, normalization matrix D_{T_1} has $(T_1 g(T_1))^{1/2}$ as its 2nd diagonal element. This differs from Appendix B (for $N = 3$ case) where we assumed that the third diagonal element of D_{T_1} to be $(T_1 g(T_1))^{1/2}$. For the general N case, our choice of D_{T_1} here leads to simple derivations.

C.2 Proof of Theorem 2.2 for the general case

Using (C.16), we are now ready to prove Theorem 2.2: $\sqrt{T_1}(\hat{\beta}-\beta) \xrightarrow{d} N(0,\Sigma)$ and $\text{rank}(\Sigma) = N-1$. Define $\delta_* = (\alpha_1, \beta_3, ..., \beta_N)$ and $\beta_* = (\beta_1, \beta_3, ..., \beta_N)'$, i.e., we remove α_2 from δ and remove β_2 from β. By virtue of Ω being positive definite, (C.16) implies that

$$\sqrt{T_1}(\hat{\delta}_* - \delta_*) \xrightarrow{d} N(0, \Omega_*), \tag{C.18}$$

where Ω_* is an $(N-1)\times(N-1)$ positive definite matrix.

From (C.12), we know that there exists an invertible $(N-1)\times(N-1)$ matrix A_* and an $(N-1)\times 1$ vector h_* such that $\delta = A_*\beta_* + h_*$. This reasoning leads to

$$\beta_* = A_*^{-1}\delta_* - A_*^{-1}h_*. \tag{C.19}$$

From (C.18) and (C.19), we immediately have

$$\sqrt{T_1}(\hat{\beta}_* - \beta_*) \xrightarrow{d} N(0, \Sigma_*), \tag{C.20}$$

where $\Sigma_* = A_*^{-1}\Omega_* A_*^{-1}$ is an $(N-1)\times(N-1)$ positive definite matrix. This condition implies that $\text{rank}(\Sigma_*) = N-1$, which further implies that $\text{rank}(\Sigma) \geq N-1$ because $\text{rank}(\Sigma) \geq \text{rank}(\Sigma_*)$.

Therefore, the proof of Theorem 2.1 will be completed if we can show that $\hat{\beta}_2 - \beta_2$ is asymptotically nonlinear with $\hat{\beta}_* - \beta_*$. From (C.5), we can express

$$\hat{\beta}_2 = \frac{1}{d_2}\left[\hat{\alpha}_2 + d_1 - \sum_{j=3}^{N} d_j\hat{\beta}_j\right]. \tag{C.21}$$

From (C.21), we obtain

$$\begin{aligned}
\sqrt{T_1}(\hat{\beta}_2 - \beta_2) &= -\frac{1}{d_2}\sum_{j=3}^{N} d_j\sqrt{T_1}(\hat{\beta}_j - \beta_j) + \frac{1}{d_2}\sqrt{T_1}(\hat{\alpha}_2 - \alpha_2) \\
&= -\frac{1}{d_2}\left[\sum_{j=3}^{N} d_j\sqrt{T_1}(\hat{\beta}_j - \beta_j)\right] + O_p(T_1^{-1}) \\
&= B_a'\sqrt{T_1}(\hat{\beta}_* - \beta_*) + O_p(T_1^{-1}),
\end{aligned} \tag{C.22}$$

where $B_a = -(1/d_2)(0, d_3, ..., d_N)'$, and the second equality follows from $\sqrt{T_1}(\hat{\alpha}_2 - \alpha_2) = \sqrt{T_1}O_p(T_1^{-3/2}) = O_p(T_1^{-1})$. Equation (C.22) implies that $\sqrt{T_1}(\hat{\beta}_2 - \beta_2)$ is asymptotically collinear with $\sqrt{T_1}(\hat{\beta}_* - \beta_*)$, which further implies that Σ is not full rank. We already know that the rank of Σ is at least $N-1$. Therefore, the rank of Σ is $N-1$ and this concludes the proof of Theorem 2.1.

Σ has reduced rank because $\sqrt{T_1}(\hat{\beta}_2 - \beta_2)$ is asymptotically collinear with $\sqrt{T_1}(\hat{\beta}_* - \beta_*)$. This in turn depends crucially on $\sqrt{T_1}(\hat{\alpha}_2 - \alpha_2) = o_p(1)$, or equivalently, $\hat{\alpha}_2 - \alpha_2 = o_p(T_1^{-1/2})$. Therefore, as long as f_t contains a (nonlinear) trend component, $\hat{\alpha}_2 - \alpha_2 = o_p(T_1^{-1/2})$ holds.

Lemma C.1 *Let A be as defined in (C.10). If $d_j \neq 0$ for at least one $j \in \{2, ..., N\}$, A is invertible.*

Proof: Without loss of generality, we can assume that $d_2 \neq 0$ (d_2 is the factor loading of y_{2t}). From (C.9) and (C.10), we know that $\delta = A\beta + h$, where $h = (-c_1, -d_1, \mathbf{0}_{1\times(N-2)})'$ and A is

$$A = \begin{pmatrix} A_{11} & A_{12} \\ \mathbf{0}_{(N-2)\times 2} & I_{N-2} \end{pmatrix}, \tag{C.23}$$

where A_{11} and A_{12} are defined in (C.11). By elementary manipulation in computing the determinant of a matrix, we can show that $\det(A)$ will not change when we replace the A_{12} matrix with $\mathbf{0}_{2\times(N-2)}$. This condition results in a block diagonal matrix with the top block as A_{11} and the bottom block as I_{N-2}. Thus, the determinant becomes:

$$\det(A) = \det(A_{11})\det(I_{N-2}) \;=\; \det(A_{11}) = d_2 \neq 0. \tag{C.24}$$

Web Appendix D: Uniqueness of the Projection Coefficient β_0

We use β_0 to represent the $N \times 1$ vector of projection coefficient defined in (D.1).

$$y_{1t} = x_t'\beta_0 + e_{1t}, \qquad\qquad t = 1, ..., T_1, \tag{D.1}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})$, and e_{1t} is a zero mean stationary idiosyncratic error term. We will prove that β_0 is uniquely determined and derive a closed form expression for β_0.

We consider the case where for $j = 1, ..., N$, y_{jt}^0 are nonlinear trend processes generated by

$$y_{jt}^0 = c_j + d_j f_t + u_{jt}, \qquad\qquad i = 1, ..., N;\ t = 1, ..., T, \tag{D.2}$$

where u_{jt} is iid $(0, \sigma^2)$. We allow f_t to be random. We further assume that f_t contains a nonlinear trend component, whose functional form is unknown, and that f_t and u_{jt} are uncorrelated.

The least squares theory of $\hat{\beta}_{OLS,T_1}$ developed in Theorem 2.1 depends on a well defined $N \times 1$ vector β_0. Therefore, we need to show that β_0 defined in (D.1) is uniquely defined. We show that β_0 is uniquely defined by presenting a closed form expression of β_0. In (D.1), β_0 is defined as the unique $N \times 1$ vector β that minimizes $E(e_{1t}^2(\beta))$. We will show below that this condition implies that e_{1t} is a zero mean stationary process. From (D.1), we have $e_{1t} = e_{1t}(\beta) = y_{1t} - x_t'\beta$. Using (D.2), we can represent e_{1t} in terms of f_t and u_{jt} for $j = 1, ..., N$:

$$\begin{aligned} e_{1t} &= y_{1t} - x_t'\beta \\ &= y_{1t} - (\beta_1 + \sum_{j=2}^{N} \beta_j y_{jt}) \end{aligned}$$

$$
\begin{aligned}
&= (c_1 + d_1 f_t + u_{1t}) - (\beta_1 + \sum_{j=2}^{N} \beta_j (c_j + d_j f_t + u_{jt})) \\
&= -(\beta_1 + \sum_{j=2}^{N} \beta_j c_j - c_1) - (\sum_{j=2}^{N} \beta_j d_j - d_1) f_t + u_{1t} - \sum_{j=2}^{N} \beta_j u_{jt} \\
&= -\alpha_1 - \alpha_2 f_t + \epsilon_t, \qquad \text{(D.3)}
\end{aligned}
$$

where

$$
\alpha_1 = \beta_1 + \sum_{j=2}^{N} \beta_j c_j - c_1 \qquad \text{(D.4)}
$$

$$
\alpha_2 = \sum_{j=2}^{N} \beta_j d_j - d_1 \qquad \text{(D.5)}
$$

$$
\epsilon_t = u_{1t} - \sum_{j=2}^{N} \beta_j u_{jt}. \qquad \text{(D.6)}
$$

Since ϵ_t has zero mean and uncorrelated with f_t, we obtain from (D.3) that

$$
E(e_{1t}^2) = \alpha_1^2 + \alpha_2^2 F(f_t^2) + 2\alpha_1 \alpha_2 E(f_t) + E(\epsilon_t^2). \qquad \text{(D.7)}
$$

It is clear that to minimize (D.7), we must have $\alpha_2 = 0$ because f_t is non-stationary. From (D.5) this condition implies that $\sum_{j=2}^{N} \beta_j d_j = d_1$. Equation (D.7) simplifies to $E(e_t^2) = \alpha_1^2 + E(\epsilon_t^2)$. Let $\tilde{\beta} = (\beta_2, ..., \beta_N)'$. Note that $\epsilon_t = \epsilon_t(\tilde{\beta})$ (ϵ_t does not depend on β_1) so that we can choose $\tilde{\beta}$ to minimize $E(\epsilon_t^2(\tilde{\beta}))$. We select β_1 to make α_1 to be zero. Therefore, the restrictions imposed on β leading $\alpha_1 = 0$ and $\alpha_2 = 0$ are:

$$
\beta_1 = c_1 - \sum_{j=2}^{N} \beta_j c_j, \qquad \text{(D.8)}
$$

$$
\sum_{j=2}^{N} \beta_j d_j = d_1. \qquad \text{(D.9)}
$$

Thus, the unique $N \times 1$ β_0 vector that minimizes $E(e_{1t}^2)$ is

$$
\tilde{\beta}_0 = \arg\min_{\tilde{\beta} \in D} E[\epsilon_t^2(\tilde{\beta})] = \arg\min_{\tilde{\beta} \in D} E[(u_{1t} - \sum_{j=2}^{N} \beta_j u_{jt})^2], \qquad \text{(D.10)}
$$

where

$$
D = \{\tilde{\beta} \in \mathcal{R}^{N-1} : \quad \tilde{\beta} \text{ satisfies } \textstyle\sum_{j=2}^{N} \beta_j d_j = d_1\},
$$

and $\beta_{01} = c_1 - \sum_{j=2}^{N} \beta_{0j} c_j$ is determined by (D.8). Recall that $\tilde{\beta} = (\beta_2, ..., \beta_N)'$).

We can solve the equality-constrained minimization problem by the Lagrangian multiplier method. We write the constraint $\sum_{j=2}^{N} \beta_j d_j - d_1 = 0$ in a standard form:

$$
R\tilde{\beta} - q = 0, \qquad \text{(D.11)}
$$

where $R = (d_2, d_3, ..., d_N)$ is a $1 \times (N-1)$ row vector and $q = d_1$. Recall that $\tilde{\beta} = (\beta_2, ..., \beta_N)'$.

In Lemma D.1, we show that the unique minimizer to the constrained minimization problem (D.10) has the following closed form expression

$$\tilde{\beta}_0 = \tilde{\beta}_* + E[(V_t V_t')]^{-1} R' \{ R E[(V_t V_t')]^{-1} R' \}^{-1} (R\tilde{\beta}_* - q), \tag{D.12}$$

where $\tilde{\beta}_* = [E(V_t V_t')]^{-1} E(V_t u_{1t})$, $V_t = (u_{2t}, ..., u_{Nt})'$,

Thus, $\tilde{\beta}_0 = (\beta_{02}, ..., \beta_{0N})'$ is determined by (D.12), and β_{01} is (D.8) with β_j replaced by β_{j0} in (D.8). Therefore, (D.12) and (D.8) present the unique closed form expression for β_0.

Lemma D.1 *Let* $\tilde{\beta}_* = \arg\min_{\tilde{\beta} \in \mathcal{R}^{N-1}} E[(u_{1t} - V_t'\tilde{\beta})^2] = [E(V_t V_t')]^{-1} E(V_t u_{1t})$. *Then*

$$\tilde{\beta}_0 = \tilde{\beta}_* - [E(V_t V_t')]^{-1} R' \{ R[E(V_t V_t')]^{-1} R' \}^{-1} (R\tilde{\beta}_* - q).$$

Proof: By (D.10) and (D.11), we want to select $\tilde{\beta} \in \mathcal{R}^{N-1}$ to minimize $E[(u_{1t} - \sum_{j=2}^N \beta_j u_{jt})^2] = E[(u_{1t} - V_t'\tilde{\beta})^2]$ subject to $R\tilde{\beta} - q = 0$. The Lagrangian multiplier is 2λ. The Lagrangian function is $\mathcal{L} = E[(u_{1t} - V_t'\tilde{\beta})^2] + 2\lambda'(R\tilde{\beta} - q)$. The first order conditions are

$$\begin{aligned} \frac{\partial}{\partial \tilde{\beta}} \mathcal{L} &= -2E[V_t(u_{1t} - V_t'\tilde{\beta})] + 2R'\lambda = 0, & \text{(D.13)} \\ \frac{\partial}{\partial \lambda} \mathcal{L} &= R\tilde{\beta} - q = 0. & \text{(D.14)} \end{aligned}$$

Pre-multiplying (D.13) by $R[E(V_t V_t')]^{-1}$ and solving for λ leads to

$$\begin{aligned} \lambda_0 &= \{ R\,[E(V_t V_t')]^{-1} R' \}^{-1} R(\tilde{\beta}_* - \tilde{\beta}) \\ &= \{ R\,[E(V_t V_t')]^{-1} R' \}^{-1} (R\tilde{\beta}_* - q), & \text{(D.15)} \end{aligned}$$

where $\tilde{\beta}_* = [E(V_t V_t')]^{-1} E(V_t u_{1t})$ and the second equality follows from $R\tilde{\beta} = q$. Substituting (D.15) back to (D.13) and solving for $\tilde{\beta}$ yields

$$\begin{aligned} \tilde{\beta}_0 &= [E(V_t V_t')]^{-1} E(V_t u_{1t}) - [E(V_t V_t')]^{-1} R' \lambda_0 \\ &= \tilde{\beta}_* - [E(V_t V_t')]^{-1} R' \{ R\,[E(V_t V_t')]^{-1} R' \}^{-1} (R\tilde{\beta}_* - q) & \text{(D.16)} \end{aligned}$$

by the definition of $\tilde{\beta}^*$ and the solution to λ_0 in (D.15).

Web Appendix E: Verifying theoretical prediction by simulation

In this Appendix we report simulation results examining our theoretical prediction that while each component $\hat{\beta}_{OLS,T_1,j} - \beta_j$ converges to zero at the rate $T_1^{-1/2}$, the linear combination $\sum_{j=2}^N d_j(\hat{\beta}_{OLS,T_1,j} - \beta_j)$

converges to zero at a much faster rate. We numerically verify this prediction using simulations. We choose $f_t = t$, $c_j = 0$, $d_j = 1$ and u_{jt} is iid $N(0,1)$ for all $j = 1, ..., N$. Under this set up and using the result of Appendix D, it can be easily shown that $\beta_1 = 0$ and $\beta_j = 1/(N-1)$ for $j = 2, ..., N$. Therefore, $\sum_{j=2}^{N} \beta_j = 1$. The fast-rate convergence linear combination (L.C.) estimator becomes:

$$\sum_{j=2}^{N} d_j(\hat{\beta}_{OLS,T_1,j} - \beta_j) = \sum_{j=2}^{N}(\hat{\beta}_{OLS,T_1,j} - 1). \tag{E.1}$$

We select $N = 3$ and $N = 4$. The sample size $T_1 \in \{50,\ 100,\ 200,\ 500,\ 1000,\ 10,000\}$. We choose large sample sizes so that we can more accurately examine the asymptotic behavior of $\hat{\beta}_{OLS,T_1}$. The number of simulation replication is 100,000 for each case. We compute estimation mean squared errors (MSE) defined by

$$MSE_j = MSE(\hat{\beta}_{OLS,T_1,j}) = \frac{1}{M}\sum_{l=1}^{M}\left(\hat{\beta}_{OLS,T_1,j,l} - \beta_j\right)^2,$$

for $j = 1, ..., N$, where $\hat{\beta}_{OLS,T_1,j,l}$ is the estimated value of β_j using the l^{th} iteration data, $\beta_1 = 0$ and $\beta_j = 1/(N-1)$ for $j = 2, ..., N$, $M = 100,000$ is the number of replications.

Similarly, we compute the MSE of the linear combination (L.C.) estimator defined in (E.1) by

$$MSE_{L.C.} = \frac{1}{M}\sum_{l=1}^{M}\left(\sum_{j=2}^{N}\hat{\beta}_{OLS,T_1,j,l} - 1\right)^2.$$

In additional to calculating MSE for $\hat{\beta}_{OLS,T_1}$. We also compute MSE of $\hat{\sigma}_e^2$. This helps examine performance of predicted value $\hat{y}_{1t} = x_t'\hat{\beta}_{OLS,T_1}$ for y_{1t}. Because x_t contains explosive nonlinear trend and it is asymptotically collinear, we want to see whether collinearity problem has any (negative) effects on predicting $x_t'\beta$ using $x_t'\hat{\beta}_{OLS,T_1}$. Therefore, we also calculate

$$MSE_{\hat{\sigma}^2} = \frac{1}{M}\sum_{l=1}^{M}(\hat{\sigma}_{e,l}^2 - \sigma_e^2)^2,$$

where $\hat{\sigma}_{e,l}^2 = T_1^{-1}\sum_{t=1}^{T_1}\hat{e}_{1t}^2$, $\hat{e}_{1t} = y_{1t} - x_t'\hat{\beta}_{OLS,T_1}$ is the estimated residual using the l^{th} replication data, for $l = 1, ..., M$ ($M = 100,000$ is the number of replications). For our data generating process, it can be shown that $\sigma_e^2 = 1 + 1/(N-1)$ because from the Web Appendix D we know that σ_e^2 is the variance of $e_{1t}(\beta) = u_{1t} - \sum_{j=1}^{N}\beta_j u_{jt} = u_{1t} - \frac{1}{N-1}\sum_{j=2}^{N} u_{jt}$ and u_{jt} is iid $N(0,1)$.

We multiply MSE_j and $MSE_{\hat{\sigma}^2}$ by T_1 so that if $T_1 \times MSE_j$ and $T_1 \times MSE_{\hat{\sigma}^2}$ converge to some positive constants as T_1 increases, it implies that MSE_j and $MSE_{\hat{\sigma}^2}$ converge to zero at the rate $1/T_1$. Similarly, we multiply $MSE_{L.C.}$ by T_1^3. Therefore, if $T_1^3 \times MSE_{L.C.}$ converges to a positive constant as T_1 rises, this implies that $MSE_{L.C.}$ converges to zero at the rate of $1/T_1^3$. The estimation results appear in Tables E.1 and E.2, for $N = 3$ and $N = 4$, respectively, strong support our theoretical analysis. All rows in Tables

E.1 and E.2 are close to some positive constants. Therefore, the results support the MSE convergence rates predicted by our theory. In particular, the linear combination estimator $\sum_{j=2}^{N} \hat{\beta}_{OLS,T_1,j}$ converges to $\sum_{j=2}^{N} \beta_j$ at the rate of $T_1^{-3/2}$ (taking a square root of the MSE rate), while all other estimators converge to their respective true values at the rate of $T_1^{-1/2}$.

Table E.1: Scale factor adjusted MSE for $N = 3$

T_1	50	100	200	500	1000	10000
$T_1 \times \text{MSE}_1$	6.444	6.253	6.123	6.035	6.011	6.027
$T_1 \times \text{MSE}_2$	0.8143	0.7822	0.7675	0.7512	0.7542	0.7503
$T_1 \times \text{MSE}_3$	0.8139	0.7827	0.7673	0.7512	0.7541	0.7503
$T_1^3 \times \text{MSE}_{\text{L.C.}}$	18.87	18.53	18.28	18.05	18.03	18.02
$T_1 \times \text{MSE}_{\hat{\sigma}^2}$	4.683	4.585	4.513	4.514	4.502	4.505

Table E.2: Scale factor adjusted MSE for $N = 4$

T_1	50	100	200	500	1000	10000
$T_1 \times \text{MSE}_1$	5.813	5.560	5.455	5.393	5.411	5.320
$T_1 \times \text{MSE}_2$	0.9858	0.9376	0.9039	0.9027	0.8996	0.8813
$T_1 \times \text{MSE}_3$	0.9865	0.9427	0.9129	0.9023	0.8936	0.8846
$T_1 \times \text{MSE}_4$	0.9897	0.9338	0.9096	0.8947	0.9033	0.8905
$T_1^3 \times \text{MSE}_{\text{L.C.}}$	16.92	16.42	16.31	16.12	16.20	15.904
$T_1 \times \text{MSE}_{\hat{\sigma}^2}$	3.835	3.694	3.604	3.595	3.561	3.552