Inference for Factor Model Based Average Treatment Effects

Kathleen T. Li*
Marketing Department
The Wharton School, the University of Pennsylvania

January, 2018

Abstract

In this paper we consider the problem of estimating average treatment effects (ATE) with a single (or a few) treated unit(s), a large number of control units, and large pre and post-treatment sample sizes. Long panel data are quite common in marketing due to the prevalence of weekly data at the customer, store or company level and or even daily data from scanner or online transaction data. This is in contrast to many economic data at the quarterly or annual frequency. We estimate counterfactual outcomes with a factor model. To select the number of factors, we propose a modification of Bai and Ng's (2002) procedure to improve its finite sample performance. We establish the asymptotic distribution theory of the ATE estimator allowing for both stationary and non-stationary data. Simulations confirm our theoretical analysis, and an empirical application examines the effect of opening a showroom by WarbyParker.com on its online sales.

Key words: Average treatment effects; factor model; asymptotic distribution, stationary and non-stationary data.

JEL number: C1.

*I would like to thank Christophe Van den Bulte and Eric T. Bradlow for helpful comments that greatly improved the paper. Correspondence regarding this manuscript can be addressed to Kathleen T. Li, katli@wharton.upenn.edu, The Wharton School, 700 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104.

1 Introduction

Estimation of average treatment effects (ATE) is important because it is commonly encountered in almost all areas of science: economics, management science, marketing, political science, sociology, medical science, and others. The early literature on estimating ATE focused mainly on the effects of education or training programs on labor market outcomes (Ashenfelter 1978, Ashenfelter and Card 1985). In recent years, marketing and management scientists increasingly have applied various ATE estimation methodologies to a wide range of issues. These include offline bookstore openings on sales at Amazon (Forman, Ghose and Goldfarb 2009), offline TV advertising on online chatter (Tirunillai and Tellis 2017) and on online shopping (Liaukonyte, Teixeira and Wilbur 2015), online multidimensional ratings on consumption (Chen, Hong and Liu 2017), online book reviews on driving sales (Chevalier and Mayzlin 2006), and offline stores on online sales (Wang and Goldfarb 2017).

Most of this work uses the difference-in-differences (DID) method to estimate ATE. DID is best suited for estimating ATE when both treated and control individuals are random draws from a common population. But even when treated individuals are not randomly selected, DID still may be used to estimate ATE. For example, when the average outcomes of the treated and the control groups follow a 'parallel line', the DID method can be used to accurately estimate ATE.

Much of the empirical data involves a large number of treated and control individuals over a short time horizon. Then, the DID methodology, together with some proper matching methods, is perhaps the best available approach for estimating ATE. That is because the 'parallel line' assumption is likely to hold for properly matched pairs within a short panel. However, when there are few (possibly only one) treated unit(s), and many control units in a long panel (a panel with long time-series data), alternative methods may be more suitable for estimating ATE because without random assignment of the treated units, the 'parallel line' assumption is likely to be violated in a long panel.

The synthetic control method (SCM) proposed by Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmeller (2010) is another popular and powerful approach for estimating ATE. It is designed for use with a few treated units (or a single one) and a small or moderate number of control units in a long panel. However, when the number of control units is large, this method may lead to large estimation variance (because of the need to estimate a large number of parameters) and to imprecise ATE estimation. The least squares approach suggested by Hsiao, Ching and Wan (2012, HCW) also can be used to estimate ATE. However, similar to the synthetic control method, having a large number of control units can lead the HCW method to over-fit the in-sample data and generate imprecise out-of-sample predictions. If the number of control units is larger than the pre-treatment time period, then the HCW method becomes invalid.[1]

In this paper, we estimate counterfactual outcomes based on a factor model also known as the generalized synthetic control method. We argue in Section 2 that a factor model structure is ideal for estimating ATE of a single treated unit with a large number of control units and a long time-series panel.

Long panels are quite common in marketing due to the prevalence of daily or weekly data at the customer, store or company level such as sales or scanner data. This is in contrast to many economic data at the quaterly or annual frequency. Most existing inference methods focus on panels with short post-treatment period. We provide inference theory for long panel which should be particularly useful to marketing and management scientists.

While Gobillon and Magnac (2016) and Chan and Kwok (2016) consider the case of large number of treated and control units, we focus on the case with one treated unit and a large number of control units. In addition, Chan and Kwok (2016) only consider the case that treatment effects are time invariant, whereas we consider a more general case that allows for time varying treatment effects.

Xu (2017) has a similar set up to ours, but he does not provide distribution theory for the factor-model-based method, which is also referred to as the generalized synthetic control method (GSC).[2] Xu proposes using a bootstrap method to conduct inferences, which relies on the assumption that the variances of the idiosyncratic errors are the same for the treated and control groups, However, this assumption is likely to be violated, especially when treated units are not randomly assigned such as in quasi-experimental settings.

Recently, Chernozhukov, Wüthrich and Zhu (2017) propose a general inference procedure covering different ATE estimators, including Difference-in-Differences (DID), synthetic control, and a factor-model-based method. They consider two situations: (i) Under the assumption that the idiosyncratic error term satisfies an exchangeability condition (e.g., iid), they propose a permutation inference method that achieves exact finite sample size. (ii) If the data exhibit dynamic and serial correlation, they propose an inference procedure

[1]When the number of control units is large. Chernozhukov, Wüthrich and Zhu (2017), and Li and Bell (2017) show that the LASSO method can be used to select control units.

[2]Xu (2017) refers to the factor-model-based method as the 'generalized synthetic control method' (GSC). We also use these two terminologies interchangeably in this paper.

that achieves approximate uniform size control under the condition that the pre-treatment sample is larger and the post-treatment sample is small. Exchangeability assumption is strong and may not be plausible in many applications; also, for many marketing data, the post-treatment sample period may not be negligibly small compared with pre-treatment sample size, rendering inference methods, based on small-post-treatment-sample-size, invalid to this type of data.

For panels with a single treated unit and a short post-treatment period, existing inference methods either assume that idiosyncratic errors have the same variance for the treated and control units (Ferman and Pinto 2017); or require the treated units's error variances to be the same during the pre and post-treatment periods (Chernozhukov et al. 2017). Equal variance assumption can be violated in practice because the treated and control units are usually heterogeneous without random assignment of treated units. Also, a treatment may not only affect the mean value of an outcome variable, but also affects its variance so that the treated unit's idiosyncratic error variance can be substantially different during pre and post-treatment periods.

In this paper, we derive the asymptotic distribution theory in order to facilitate inference for our factor-model-based ATE estimator that is ideal for settings with a large number of control units, and large pre- and post-treatment sample sizes. Our distribution theory fills a gap in the literature and provides a simple inference method for the generalized synthetic control ATE estimator with long panels under quite general (regularity) conditions. Specifically, we allow for both stationary and non-stationary data, idiosyncratic errors can be weakly dependent both cross sectionally and over time (e.g., Bai 2003, 2004). Moreover, we allow for the variances of the idiosyncratic errors from the treated and control groups to be different, and we also allow for the treated unit's idiosyncratic errors to have different variances during the pre- and post-treatment periods. Because our inference method is based on normal distribution theory, an additional advantage is that it is computationally much more efficient than bootstrap or permutation based inference procedures. Simulations reported in Section 5 confirm our theoretical result.

The paper is organized as follows. In Section 2 we review several popular ATE estimation methods. Then we discuss our set up and estimation method. In Section 3 we propose a modification to Bai and Ng's (2002) method of estimating the number of factors in a factor model (for the stationary data case). Section 4 provides distribution theory of the factor-model-based ATE estimator. In Section 5, we use simulations to examine the finite sample performances of our modified method in selecting the number of factors, and the estimated confidence intervals based on our distribution theory. Section 6 reports on an empirical application where we estimate the ATE of opening a showroom in Boston on Boston online sales by WarbyParker.com. Section 7 concludes and Appendix A contains proofs of the main theoretical results.

2 Different ATE estimation methods

First, we describe the general problem and notation for average treatment effect (ATE) estimation. Then, we introduce several methods for estimating ATE and describe their relative advantages and disadvantages. Let y_{it}^1 and y_{it}^0 denote the outcomes of unit i in period t with and without treatment, respectively. The treatment intervention effect for the i^{th} unit at time t is defined as

$$\Delta_{it} = y_{it}^1 - y_{it}^0. \tag{2.1}$$

However, for the same unit i, we do not simultaneously observe y_{it}^0 and y_{it}^1. Thus, the observed data takes the form $y_{it} = d_{it} y_{it}^1 + (1 - d_{it}) y_{it}^0$, where $d_{it} = 1$ if the i^{th} unit is under the treatment at time t, and $d_{it} = 0$ otherwise. That is, we observe y_{it}^1 or y_{it}^0, depending on whether unit i receives a treatment at time t or not.

In estimating ATE, the difficulty lies in how to obtain the counterfactual outcome y_{it}^0 when the i^{th} unit receives a treatment, because we only observe y_{it}^1 for post-treatment period t for a treated unit. We focus on the case where only one unit receives a treatment at time $T_1 + 1$ and the other N_{co} control units do not receive any treatment throughout the sample period.[3] Without loss of generality, we assume that the first unit receives a treatment at time $T_1 + 1$. We want to estimate the post-treatment period average treatment effects for the treated unit. Specifically, we want to estimate $\Delta_1 = E(y_{1t}^1 - y_{1t}^0)$. Let $\hat{y}_{1t}^0$ be a generic estimator of y_{1t}^0. Then, a sample analogue of Δ_1 is given by

$$\hat{\Delta}_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}_{1t}^0),$$

where $T_2 = T - T_1$ is the number of post-treatment time periods. Here we would like to emphasize that since there is only one unit (the first unit) that receives a treatment, the ATE estimator is obtained by averaging over the post-treatment periods (time series

[3] If more than one unit receives treatment, possibly at different times, we can first estimate ATE (over post-treatment period) for each treated unit, then average these units' ATE to obtain a total ATE, an average over both treated units and their post-treatment periods.

averaging) for the treated unit. This differs from the usual Difference-in-Differences (DID) method in which one often has a larger number of units receiving treatments and the average is usually done over many treatment units (cross sectional averaging).

Because we only have one unit that receives a treatment, consistent estimation of Δ_1 usually requires that both the number of pre-treatment periods, T_1, and the number of post-treatment period, T_2 are large, while the number of control units, N_{co}, can be either large or small. In the following subsections, we discuss different estimation methods that are designed for estimating ATE for a single treated unit, and compare their relative advantages and disadvantages, in particular their performance when the total number of time periods, T, and the number of control units, N_{co}, are large.

2.1 The synthetic control method

The synthetic control method (SCM) proposed by Abadie and Gardeazabal (2003), and Abadie, Diamond and Hainmeller (2010), uses a weighted average of control units to approximate the counterfactual outcome of the treated unit in the absence of treatment. The weights are restricted to be non-negative and to sum to one. The synthetic control method includes the Difference-in-Differences method as a special case when each control unit is assigned a weight of $1/N_{co}$. The performance of the synthetic control method relies crucially on the assumption that the weighted average of the control units' sample path is parallel to the treated unit's sample path in the absence of treatment. Doudchenko and Imbens (2016) suggest that when the parallel line assumption is violated, one can remove the restriction that weights sum to one. This modification effectively relaxes the original 'parallel line' assumption to a weaker condition that the weighted average (with weights summing to one) of the control unit times a positive constant (a scale factor) is parallel to the sample path of the treated unit in the absence of treatment (so that the new weights may not sum to one). This modified synthetic control method (MSCM) contains the synthetic control method as a special case when the scale factor equals one. The performance of the modified version requires large a T_1 and T_2 but that N_{co} is not too large. When the number of control units, N_{co}, is large, it may be better not to use all of the control units when estimating ATE using the synthetic control method (or the modified version) because a large number of explanatory variables in a regression model can lead to overfitting in-sample and imprecise out-of-sample predictions. In this case, one can use the best subset selection method proposed by Doudchenko and Imbens (2016) or the least absolute shrinkage and selection operator (LASSO) method suggested by Chernozhukov, Wüthrich and Zhu (2017), and Li and Bell (2017) to select control units in applications. Recent work on synthetic control methods include Chernozhukov, Wüthrich and Zhu (2017), Ferman and Pinto (2017), Firpo and Possebom (2017), Hahn and Shi (2016), Li (2017) and Xu (2017) who provide inference methods or asymptotic distribution theory of the synthetic control (or modified syntheic control) ATE estimator.

2.2 The least squares method

Hsiao, Ching and Wan (2012, HCW) propose estimating ATE based on a least squares approach. They motivate their estimation method via a factor model data generating process. The common factors are the main forces driving all the outcome variables, y_{it}, to exhibit some co-movement over time. Specifically, suppose that the outcome variables for the control units are generated by the following factor model

$$y_{it}^0 = \lambda_i' F_t + e_{it}, \qquad i = 2, ..., N; \quad t = 1, ..., T, \tag{2.2}$$

where λ_i is a r-dimensional vector of factor loadings, F_t is a r-dimensional vector of (unobservable) common factors and e_{it} is a zero mean idiosyncratic error. Note that the first element of F_t can be a constant of 1 so that the model contains an individual specific intercept term.

Similarly, the treated unit's pre-treatment period outcome is generated by $y_{1t} = \lambda_1' F_t + e_{1t}$ for $t = 1, ..., T_1$. At time $T_1 + 1$, the first unit receives a policy intervention or treatment. Therefore, the treated unit outcome for the post-treatment periods is

$$y_{1t} = y_{1t}^1 = \lambda_1' F_t + \Delta_{1t} + e_{1t} \qquad \text{for } t = T_1 + 1, ..., T, \tag{2.3}$$

where Δ_{1t} is the treatment effect to the first unit at time t.

HCW suggest a method for estimating the counterfactual outcome y_{1t}^0 without estimating unobserved factors and factor loadings. They propose first regressing the treated unit outcome, y_{1t}, on control units' outcomes, $x_t = (1, y_{2t}, ..., y_{Nt})'$, using pre-treatment period data to estimate the coefficients β in the following regression model

$$y_{1t} = x_t' \beta + u_{1t} \qquad t = 1, ..., T_1, \tag{2.4}$$

Let $\hat{\beta}_{OLS}$ denote the resulting estimator of β. Then, the counterfactual outcome, y_{1t}^0, can be predicted by $\hat{y}_{1t}^0 = x_t' \hat{\beta}_{OLS}$ for $t = T_1 + 1, ..., T$.

Effectively, HCW's method uses the control units' outcome variables to replace F_t in the regression analysis. The rationale for this approach is that outcome variables are all correlated via common factors. Under the identifying assumption that the correlation between outcomes of the treated and the control units remains stable in the absence of treatment, HCW's method can be used to estimate a counterfactual outcome and the ATE. However, one problem with this approach is that when the number of control units, N_{co}, is large (e.g. $N_{co} > T_1$), the least squares estimator of β does not exist. Hence, the HCW method is not well suited when the number of control units, N_{co}, is large. Even when the number of control units is smaller than the pre-treatment sample size ($N_{co} \leq T_1$), a large number of regressors usually leads to large estimation variation and may lead to inaccurate estimation results. HCW propose using some model selection criteria, such as AICC or BIC, to select a subset of control units for estimating the counterfactual outcomes, y_{1t}^0. AICC or BIC procedures can be computationally extremely costly when the number of control units, N_{co}, is large. Chernozhukov, Wüthrich and Zhu (2017), and Li and Bell (2017) show that the LASSO method is computationally more efficient than the AICC or BIC method in selecting control units.

2.3 A factor model approach

Unlike HCW, Gobillon and Magnac (2016), Chan and Kwok (2016) and Xu (2017) suggest estimating ATE by directly estimating a factor model. Suppose that in the absence of treatment, outcomes are generated by the following factor model

$$y_{it}^0 = \lambda_i' F_t + e_{it}, \tag{2.5}$$

where F_t is a $r \times 1$ vector of unobservable common factors, λ_i is a $r \times 1$ vector of factor loading coefficients, and e_{it} is an idiosyncratic error term that has a zero mean and finite fourth moment. For the control units ($i = 2, ..., N$), this outcome equation holds for all time periods, $t = 1, ..., T$ whereas for the treated unit ($i = 1$), this outcome equation only holds for the pre-treatment time periods ($t = 1, ..., T_1$).

If F_t were observable, then one could estimate λ_1 by the least-squares method using the pre-treatment data, and predict the counterfactual outcome as in HCW. However, in practice, F_t is not observable. A popular approach for estimating factors is principal components analysis. Gobillon and Magnac (2016) and Xu (2017) suggest using control units' full data (y_{it} for $i = 2, ..., N$ and $t = 1, ..., T$) to estimate the common factors (principal components).

Usually the number of factors (r) is small. For example, in our empirical application the number of factors is one. However, accurately estimating the number of factors (r)

requires both a large number of time periods (T) and a large number of control units (N_{co}). In addition, to accurately estimate the factors, F_t, we need a large number of controls units. When the number of control units is large, the HCW's least squares method involves estimation of the high dimensional model (2.4) (which has N_{co} coefficients to be estimated). In contrast, estimating ATE using a factor model approach only requires estimation of a low dimensional model (e.g., (2.5)) with r unknown coefficients. With a given sample size T_1, a low dimensional model can be estimated much more accurately than a high dimensional model. Hence, using a factor model approach to estimate ATE is ideal in a setting with a large number of control units and a long time series.

Following Gobillon and Magnac (2016) and Xu (2017), we consider estimating ATE based on a factor model. However, first we need to determine the number of factors. We propose a modification to Bai and Ng's (2002) model selection criteria (a small sample correction): the modified method performs better in determining the number of factors for small to moderate sample sizes. Also, filling a gap in the literature, we derive the asymptotic distribution of the generalized synthetic control estimator under weak conditions. For example, we do not require that the idiosyncratic errors from treated and control groups have the same variance and also do not require that the treated unit have the same variance in the pre and posttreatment periods , which makes our inference theory applicable to a wider range of empirical data applications, which previously could not be accurately analyzed using the factor-model-based generalized synthetic control method. The simulation results show that our inference theory leads to accurate estimates of finite sample confidence intervals. Finally, the generalized synthetic control method is also computationally more efficient, especially when the number of control units is large, than some existing estimation methods such as HCW's panel data approach which relies on using AICC or BIC to select control units.

3 Determining the number of factors

In this section, we consider the problem of determining the number of factors (r) with stationary data. Determining the number of factors precisely is important in accurately predicting counterfactual outcomes, and thus the ATE, using a factor model. Xu (2017) proposes using the leave-one-out cross-validation method to estimate the number of factors. However, Shao (1993) shows that in a regression model, the leave-one-out cross-validation method tends to select a number of factors larger than the true number of factors even when the sample size is large. Bai and Ng (2002) propose several criteria functions that can be used to consistently estimate the number of factors. For the stationary data case, one popular approach for determining the number of factors is Bai and Ng's (2002) PC_{p_1} criterion. This criterion selects the number of factors by choosing the value of k that minimizes

$$PC_{p_1}(k) = \frac{1}{N_{co}T}\sum_{i=2}^{N}\sum_{t=1}^{T}(y_{it} - \hat{\lambda}_i'\hat{F}_{Kt})^2 + k\,\hat{\sigma}^2\left(\frac{N_{co}+T}{N_{co}T}\right)\ln\left(\frac{N_{co}+T}{N_{co}T}\right) \tag{2.6}$$

where $k \in \{0, 1, ..., k_{max}\}$, k_{max} is a pre-specific constant satisfying $0 \leq r \leq k_{max}$, $N_{co} = N - 1$ is the number of control units, $\hat{\sigma}^2$ is an estimator of σ^2, and σ^2 is the variance of the idiosyncratic error in the factor model. As shown in Bai and Ng (2002), the PC_{p_1} criterion can consistently estimate the true number of factors when the number of control units, N_{co}, and the total number of time periods, T, are large. However, for small to moderate sample sizes, we find that PC_{p_1} tends to select a value of $\hat{k}$ (the estimated number of factors) larger than r (the true number of factors). This means that the penalty term in PC_{p_1} is not large enough for small to moderate sample sizes. Therefore, we suggest modifying the PC_{p_1} criterion by multiplying its penalty term by a factor

$$c_{N,T,m_N,m_T} = (N_{co} + m_N)(T + m_T)/(N_{co}T), \tag{2.7}$$

where m_N and m_T are bounded non-negative integers so that $c_{N,T,m_N,m_T} \geq 1$ for all N_{co}, T and that $c_{N,T,m_N,m_T} \to 1$ as N_{co}, $T \to \infty$. The simulations reported in Section 5 show that the choice of m_N and m_T such that $N_{co} + m_N \geq 60$ and $T + m_T \geq 60$ perform quite well. Hence, we recommend using the following m_N and m_T in practice

$$m_N = \max\{0, 60 - N_{co}\} \qquad \text{and} \qquad m_T = \max\{0, 60 - T\}.$$

Note that m_N (m_T) is non-negative, non-increasing in N_{co} (T) and equals zero for $N_{co} \geq 60$ ($T \geq 60$). For example, when $N_{co} = 30$ and $T = 30$, this correction factor $c_{N,T,m_N,m_T} = (30+30)(30+30)/(30)^2 = 4$. While $c_{N,T,m_N,m_T} = 1$ if $\min\{N_{co}, T\} \geq 60$.

Therefore, we propose choosing the number of factors as the value of $k \in \{0, 1, ..., k_{max}\}$ that minimizes the following modified version of Bai and Ng's (2002) PC_{p_1} criterion:

$$PC_{p_1,m_N,m_T}(k) = \frac{1}{N_{co}T}\sum_{i=2}^{N}\sum_{t=1}^{T}(y_{it} - \hat{\lambda}_i'\hat{F}_{Kt})^2 + k\,\hat{\sigma}^2 c_{N,T,m_N,m_T}\left(\frac{N_{co}+T}{N_{co}T}\right)\ln\left(\frac{N_{co}+T}{N_{co}T}\right). \tag{2.8}$$

By construction, the modification factor $c_{N,T,m_N,m_T} \geq 1$ and as N_{co}, $T \to \infty$, $c_{N,T,m_N,m_T} \to 1$ so we have the nice result that the asymptotic property of PC_{p_1,m_N,m_T} is the same as that of PC_{p_1}. In fact, we have equivalence of the two criterion, $PC_{p_1,m_N,m_T} \equiv PC_{p_1}$, if $\min\{N_{co}, T\} \geq 60$. Hence, the modified criterion PC_{p_1,m_N,m_T} has the same behavior as PC_{p_1} for large N_{co} and T. In other words, when the number of control units and number of time periods goes to infinity, both criterions will choose the true number of factors with probability 1. Mathematically, $\lim_{N_{co},T\to\infty} P(\hat{k}_{PC_{p_1,m_N,m_T}} = r) = 1$ follows directly from $\lim_{N_{co},T\to\infty} P(\hat{k}_{PC_{p_1}} = r) = 1$ by Bai and Ng (2002), where $\hat{k}_{PC_{p_1,m_N,m_T}}$ and $\hat{k}_{PC_1}$ are the selected number of factors by the modified and the original criteria PC_{p_1,m_N,m_T} and PC_{p_1}, respectively. We examine the performance of the modified criterion by simulations in Section 5.

4 Inference Theory for the ATE Estimator

Previously, the inference theory for the generalized synthetic control (GSC) ATE estimator based on a factor model approach was limited. We will this gap for both the stationary and non-stationary data cases in order to facilitate inference using the GSC estimator.

4.1 The Stationary Data Case

In the absence of treatment, the outcome variable y_{it}^0 is generated by a factor model:

$$y_{it}^0 = \lambda_i' F_t + e_{it}, \tag{4.1}$$

For the control units ($i = 2, ..., N$), this outcome equation holds for all time periods, $t = 1, ..., T$ whereas for the treated unit ($i = 1$), this outcome equation only holds for the pre-treatment time periods ($t = 1, ..., T_1$).

Let $\hat{k}$ be the estimated number of factors obtained by minimizing PC_{p_1,m_N,m_T} using the full control units' data (i.e., using y_{it} for $i = 2, ..., N$ and $t = 1, ..., T$). Let $\hat{F}_t$ be the $\hat{k} \times 1$ vector of estimated factors for $t = 1, ..., T$. We estimate the factor loading λ_1 based on the following regression model using the pre-treatment data

$$y_{1t} = \lambda_1' \hat{F}_t + \epsilon_{1t}, \qquad t = 1, ..., T_1 \tag{4.2}$$

where $\epsilon_{1t} = \lambda_1'(F_t - \hat{F}_t) + e_{1t}$.

Let $\hat{\lambda}_1$ be the OLS estimator of λ_1 based on (4.2), i.e.,

$$\hat{\lambda}_1 = \left[\sum_{t=1}^{T_1} \hat{F}_t \hat{F}_t' \right]^{-1} \sum_{t=1}^{T_1} \hat{F}_t y_{1t}. \tag{4.3}$$

Then we estimate the post treatment period counterfactual outcome y_{1t}^0 by $\hat{y}_{1t}^0 = \hat{\lambda}_1' \hat{F}_t$, for $t = T_1+1, ..., T$. The treatment effects are estimated by $\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^0$ for $t = T_1+1, ..., T$, and the average treatment effect is given by

$$\hat{\Delta}_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \hat{\Delta}_{1t} = \frac{1}{T_2} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}_{1t}^0), \tag{4.4}$$

where $T_2 = T - T_1$ is the post-treatment sample size. In Appendix A we show that

$$\begin{aligned} \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) &= \Big[\sum_{s=1}^{T_1} e_{1s} F_s'\Big]\Big[\sum_{s=1}^{T_1} F_s F_s'\Big]^{-1}\Big[\frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} F_t\Big] + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\ &= A_1 + A_2 + o_p(1), \end{aligned} \tag{4.5}$$

where $v_{1t} = \Delta_{1t} - E(\Delta_{1t}) + e_{1t}$ is a zero mean stationary process. Equation (4.5) shows that the ATE consists of two parts, A_1 and A_2, where A_1 captures the estimation error in the pre-treatment period and A_2 captures the post-treatment error. From the proofs presented at Appendix A, we know that A_1 is from estimation error of $\hat{\lambda}_1 - \lambda_1$, while A_2 is from average of v_{1t} over the post-treatment period. It can be shown that under the regularity conditions discussed in Appendix A, $A_l \stackrel{d}{\to} N(0, \Omega_l)$ for $l = 1, 2$, and that A_1 and A_2 are asymptotically independent with each other. Hence, we have the following result.

Theorem 4.1 *Under the assumptions given in Appendix A.1 (mainly Bai's (2003) regularity conditions) and letting $\Delta_1 = E(\Delta_{1t})$, we have*

$$\frac{\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)}{\sqrt{\hat{\Omega}}} \stackrel{d}{\to} N(0, 1),$$

where $\hat{\Omega}$ is a consistent estimator of Ω (see Appendix A for a specific definition of $\hat{\Omega}$), $\Omega = \Omega_1 + \Omega_2$, $\Omega_1 = \phi\Sigma$, $\phi = \lim_{T_1,T_2\to\infty} T_2/T_1$, $\Sigma = \eta'\Psi^{-1}\eta$ with $\eta = E(F_t)$, $\Psi = [E(F_tF_t')]^{-1}V[E(F_tF_t')]^{-1}$, $V = \lim_{T_1\to\infty} T_1^{-1} \sum_{t=1}^{T_1} \sum_{s=1}^{T_1} E(e_{1t}e_{1s}F_tF_s')$, $\Omega_2 = \lim_{T_2\to\infty} T_2^{-1} \sum_{t=T_1+1}^{T} \sum_{s=T_1+1}^{T}$ and $v_{1t} = \Delta_{1t} - E(\Delta_{1t}) + e_{1t}$.

The proof of Theorem 4.1 is given in Appendix A.

Remark 4.1: *If F_t were observable, then one could replace $\hat{F}_t$ by F_t in estimating the ATE, Δ_1. Call this the 'infeasible ATE estimator'. Then by comparing the results of Theorem 4.1 and Theorem 3.2 of Li and Bell (2017), one can see that Theorem 4.1 claims that the asymptotic distribution of our feasible ATE estimator and the 'infeasible ATE estimator' that uses the true F_t in regression model (2.5) have the same asymptotic distribution. That* *is, asymptotically, the estimated factor $\hat{F}_t$ is as good as the true F_t as far as the asymptotic distribution of $\hat{\Delta}_1$ (the ATE estimator) is concerned.*

Remark 4.2: *When the number of pre-treatment time periods, T_1, is much larger than the number of post-treatment time periods, T_2 (i.e.$T_2/T_1 = o(1)$), we have $\Omega_1 = 0$ because $\phi = 0$. The asymptotic variance reduces to Ω_2. This result is quite intuitive. When T_1 is much larger than T_2, the factor loading λ_1 can be accurately estimated and the estimation error in $\hat{\lambda}_1 - \lambda_1$ is negligible. In other words, A_2 becomes the dominating term and A_1 is asymptotically negligible compared with A_2.*

Remark 4.3: *Our Theorem 4.1 allows for the idiosyncratic errors from the treated and control group to have different variances. It also allows for the treated unit's idiosyncratic errors to have different variances for the pre- and post-treatment periods.*

In Appendix A, we propose consistently estimating Ω_1 and Ω_2 by $\hat{\Omega}_1$ and $\hat{\Omega}_2$, respectively. A consistent estimator of Ω is given by $\hat{\Omega} = \hat{\Omega}_1 + \hat{\Omega}_2$. Then, for $\alpha \in (0, 1)$, Theorem 4.1 implies that

$$P[c_{\alpha/2} \leq \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)/\sqrt{\hat{\Omega}} \leq c_{1-\alpha/2}] \to 1 - \alpha, \tag{4.6}$$

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 $(1 - \alpha)$ confidence interval (CI) of Δ_1 is given by

$$[\hat{\Delta}_1 - c_{1-\alpha/2}\sqrt{\hat{\Omega}}/\sqrt{T_2},\ \ \hat{\Delta}_1 - c_{\alpha/2}\sqrt{\hat{\Omega}}/\sqrt{T_2}]. \tag{4.7}$$

For example, for $\alpha = 0.05$, a 95% CI for Δ_1 is given by

$$[\hat{\Delta}_1 - 1.96\sqrt{\hat{\Omega}}/\sqrt{T_2},\ \ \hat{\Delta}_1 + 1.96\sqrt{\hat{\Omega}}/\sqrt{T_2}]$$

because $c_{0.975} = 1.96$ and $c_{0.025} = -1.96$.

To our knowledge, we are the first to provide asymptotic theory for the factor model based ATE estimator, also known as the generalized synthetic control estimator. Xu (2017) does not provide an asymptotic theory for the generalized synthetic control estimator, but rather proposes using a bootstrap method to conduct inferences. Xu suggests drawing bootstrap sample for the treated unit (in the absence of treatment) from the control group's data. Hence, his method requires that the idiosyncratic error term e_{jt} have the same variance for the treated and control units. When the idiosyncratic errors from the treated and control units are heterogeneous and have different variances, Xu's method yields biased CI estimation results, while our inference method based on (4.7) works well whether the idiosyncratic errors from the treated and control groups have the same variance or not.

Chernozhukov, Wüthrich and Zhu (2017) propose a using permutation procedure to conduct inferences under quite general conditions, but they consider the cases of (i) fixed T_1 and T_2, and (ii) large T_1 and fixed T_2, while we consider large T_1 and T_2 and we allow for T_2/T_1 converges to a non-negative constant. Hence, while we consider large T_1 and T_2 scenario, we allow T_2 to have the same order as T_1, or has an order smaller than T_1. Hence, our results complement that of Chernozhukov, Wüthrich and Zhu (2017).

4.2 The Non-Stationary Data Case

The model and the notation is the same as described in the previous subsection except that now F_t is non-stationary rather than stationary. To be specific, we consider the case that F_t follows a drift-less unit root process: $F_t = F_{t-1} + u_t$, where u_t is weakly dependent stationary process with zero mean and finite fourth moment. Now equation (4.1) becomes a co-integration model because e_{1t} is a zero mean stationary process.

Let Y be the $T \times N_{co}$ data matrix of the control units' outcomes. As in Bai (2004), for a fixed value of k, we estimate factors and factor loadings by solving the following minimization problem

$$V(k) = \min_{\Lambda^k, F^k} \frac{1}{N_{co}T} \sum_{i=2}^{N} \sum_{t=1}^{T} (Y_{it} - \lambda_i^{k\prime} F_t^k)^2 \tag{4.8}$$

subject to the normalization $F^k F^{k\prime}/T^2 = I_k$ or $\Lambda^{k\prime}\Lambda^k/N_{co} = I_k$. The estimator of $F_{T\times k}$ is denoted by $\tilde{F}^k$, which is T times the eigenvectors corresponding to the k largest eigenvalues of the $T \times T$ matrix YY'. The estimator of the factor loading matrix $(\Lambda^k)_{k\times N_{co}}$ can be obtained by $\tilde{\Lambda}^{k\prime} = (\tilde{F}^{k\prime}\tilde{F}^k)^{-1}\tilde{F}^{k\prime}Y = \tilde{F}^{k\prime}Y/T^2$ because $F^kF^{k\prime}/T^2 = I_k$.

Another way of obtaining estimators for F and Λ is to first obtain $\bar{\Lambda}^k$ as $\sqrt{N_{co}}$ times the eigenvectors corresponding to the k largest eigenvalues of the $N_{co} \times N_{co}$ matrix $Y'Y$ and then obtain $\bar{F}^k = Y\bar{\Lambda}^k(\bar{\Lambda}^{k\prime}\bar{\Lambda}^k)^{-1} = Y\bar{\Lambda}^k/N_{co}$ because $\Lambda^{k\prime}\Lambda^k/N_{co} = I_k$. Bai (2004) further defines a rescaled version of $(\bar{F}^k, \bar{\Lambda}^k)$ as follows

$$\hat{F}^k = \bar{F}^k(\bar{F}^{k\prime}\bar{F}^k/T^2)^{1/2} \text{ and } \hat{\Lambda}^k = \bar{\Lambda}^k(\bar{F}^{k\prime}\bar{F}^k/T^2)^{-1/2}.$$

For a fixed $k \geq 1$, there exists H_l^k $(l = 1, 2)$ matrix with $rank(H_l) = \min\{k, r\}$ such that $\hat{F}_t^k$ estimates $(H_1^k)'F_t$ and $\tilde{F}_t^k$ estimates $(H_2^k)'F_t$. We will suppress the superscript k for notational simplicity, i.e., we will write $\hat{F}_t = \hat{F}_t^k$ and $H_l = H_l^k$ for $l = 1, 2$.

For determining the number of factors, we use the method suggested by Bai (2004, page

145) and select the number of factors by minimizing:

$$IPC_1(k) = V(k) + k\hat{\sigma}^2\alpha_T \left(\frac{N+T}{NT}\right) \log\left(\frac{NT}{N+T}\right), \tag{4.9}$$

where $\alpha_T = T/[4\log\log(T)]$ and $\hat{\sigma}^2$ is an estimator of the variance of the idiosyncratic error for the control group. Let $\hat{k}$ denote the number of factors selected by miniming $IPC_1(k)$, Bai (2004) proves that $Pr(\hat{k} = r) \to 1$ as $T, N_{co} \to \infty$.

The estimator of the factor loading, $\hat{\lambda}_1$, and the ATE estimator, $\hat{\Delta}_1$, are the same as given in (4.3) and (4.4). Moreover, we also have

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) = A_1 + A_2 + o_p(1), \tag{4.10}$$

where the definitions of A_1 and A_2 are the same as given in (4.5) except that now F_t is an unit root process. The following Theorem shows that standardized ATE estimator has an asymptotic standard normal distribution.

Theorem 4.2 *Under the assumptions given in Appendix A.2 (mainly Bai's (2004) regularity conditions) and the assumption that Δ_{1t} is a stationary process, we have*

$$\frac{\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)}{\sqrt{\hat{\Omega}}} \xrightarrow{d} N(0,1),$$

where $\Delta_1 = E(\Delta_{1t})$ and $\hat{\Omega}$ is the same as in Theorem 4.1 (which is defined in Appendix A).

Remark 4.4: *Comparing Theorem 4.1 and Theorem 4.2, we see that the asymptotic distributions of the ATE estimator are the same whether the data is stationary or non-stationary. The reason for this result follows similar logic to the estimation of a co-integration model, where we know that even when the co-integration coefficient estimator is not asymptotically normal, the t-statistic from a co-integration model can still be asymptotically normal under some conditions (Hayashi 2000, page 658). This greatly simplifies the inference procedure. In particular, we do not need to conduct a pre-test to determine whether or not the data is stationary, as the construction of confidence intervals is the same whether the data is stationary or non-stationary. The confidence intervals for Δ_1 is given in (4.7).*

Remark 4.5: *Similar to the stationary data case, when the number of pretreatment time periods, T_1, is much larger than the number of posttreatment time periods, T_2 (i.e., $\phi = \lim_{T_2,T_1\to\infty} T_2/T_1 = 0$), we show in supplementary Appendix B that A_1 is asymptotically* *negligible compared with A_2 (i.e. $A_1 = o_p(1)$ as shown Lemma B.6 and the non-stationary component F does not show up in the asymptotic distribution). Hence, the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$ is determined by $A_2 = T_2^{-1/2} \sum_{t=T_1+1}^{T} v_{1t}$, which converges to $N(0, \Omega_2)$ in distribution.*

Remark 4.6: *Note that Ω_1 defined in Theorem 4.1 is a non-negative constant, while it does not have a well defined meaning when F_t is a drift-less unit root process, and $\hat{\Omega}_1$ does not converge to a constant when F_t is non-stationary. Nevertheless, we still have $A_1/\sqrt{\hat{\Omega}_1} \xrightarrow{d} N(0,1)$ for the reason given in remark 4.4.*

5 Simulation Results

5.1 Selecting the number of factors: the modified Bai and Ng's (M-BN) criterion

In this subsection, we focus on the stationary data case and use simulations to examine the accuracy of using Xu's leave-one-out cross-validation method, Bai and Ng's (2002) PC_{p_1} criterion, and the modified criterion PC_{p_1,m_N,m_T} to select the number of factors. We show that the modified method overcomes the small sample bias problem of the original PC_{p_1} method. Following HCW (2012) and Du and Zhang (2015), we consider the following 3-factor model:

$$DGP1: \quad \begin{cases} F_{1t} = 0.8F_{1t-1} + v_{1t}, \\ F_{2t} = -0.68F_{1t-1} + v_{2t} + 0.8v_{2t-1}, \\ F_{3t} = v_{3t} + 0.9v_{3t-1} + 0.4v_{3t-2}, \end{cases}$$

where v_{it} is iid $N(0,1)$. The outcome variables (in the absence of treatment) are generated by $y_{it}^0 = \lambda_i' F_t + e_{it}$ for $i = 1, ..., N$; $t = 1, ..., T$, where e_{it} is iid $N(0, \sigma^2)$ and λ_i is iid $N(1,1)$. We choose the value of variance as $\sigma^2 = 2$, 1, 0.5 and 0.1 and we consider number of control units to be $N_{co} = 30$, 60, 120 ($N = N_{co} + 1$) and the number of time periods to be $T = 30$, 60, 120. The number of simulation replications is 5,000. We select the number of factors in the range of $\{0, 1, ..., 10\}$ by minimizing Bai and Ng's (2002) criterion PC_{p1} defined in (2.6), the modified criterion PC_{p_1,m_N,m_T} defined in (2.8), and the least-squares cross-validation method suggested by Xu (2017). Our simulation results are reported in Table 1.

The middle panel of Table 1 shows the number of estimated factors $\hat{k}$ using the PC_{p_1} criterion. From Table 1, we observe that when N_{co} and T are small, the PC_{p_1} criterion tends to select a number that is much larger than the true number of factors ($r = 3$). In fact, when $T = 30$ and $N = 30$, the PC_{p_1} method often selects the upper bound value $\hat{k} = 10$ (since the maximum allowed number of factors is $k_{max} = 10$). Even for $(N_{co}, T) = (60, 30)$ and $(N_{co}, T) = (30, 60)$, the PC_{p_1} method still selects a number much larger than three, the true number of factors. In contrast, the PC_{p1,m_N,m_T} criterion (whose results are in the right panel) performs better because it penalizes more for small sample cases and thus avoids selecting a large number of factors. It virtually always selects the correct number, three for all cases.

Note that when both N_{co} and T are large, Bai and Ng (2002) and the modified methods perform equally well as expected, because the two criteria become the same for large N_{co} and T. Also, the cross-validation method suggested by Xu (2017) tends to select a model with a number larger than the true value of three. This is true even for large values of N_{co} and T. Shao (1993) shows that the leave-one-out cross validation method is inconsistent in selecting the true model. Our simulations verify that the leave-one-out cross-validation method is also inconsistent in selecting the number of factors (see left panel).

Table 1: Selecting # of factors by LS-CV, PC_{p_1} and PC_{p1,m_N,m_T}

	LS-CV				PC_{p_1}				PC_{p1,m_N,m_T}			
σ^2	2	1	0.5	0.1	2	1	0.5	0.1	2	1	0.5	0.1
T	$N_{co} = 30$											
30	3.517	3.588	3.765	3.840	9.455	9.486	9.465	9.429	2.992	3.000	3.000	3.000
60	3.467	3.606	3.556	3.715	6.874	6.862	6.892	6.839	2.999	3.000	3.000	3.000
120	3.512	3.684	3.593	3.738	4.169	4.176	4.248	4.191	3.000	3.000	3.000	3.000
	$N_{co} = 60$											
30	3.502	3.582	3.763	3.801	6.854	6.897	6.852	6.869	3.000	3.001	3.000	3.000
60	3.480	3.587	3.564	3.672	3.187	3.198	3.198	3.206	3.000	3.000	3.000	3.000
120	3.568	3.632	3.639	3.656	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000
	$N_{co} = 120$											
30	3.511	3.613	3.666	3.744	4.226	4.155	4.187	4.201	3.000	3.000	3.000	3.000
60	3.453	3.594	3.646	3.827	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000
120	3.621	3.646	3.600	3.749	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000

5.2 Estimation of confidence intervals

In this section we use simulations to examine the coverage probabilities of confidence interval estimated using (4.7). We consider both cases where the idiosyncratic errors for the treated and control groups have the same and different variances in section 5.2.1 and section 5.2.2, respectively. Also, we consider both stationary and non-stationary data cases.

The key takeaway is that our inference method performs well (i.e. able to recover the coverage probabilities) in all cases allowing our inference method to be more widely applied compared to previous methods. In addition, because the simulation results are very similar for both the case where there the true ATE is a null effect, $\Delta_1 = 0$, and where the true ATE is non-zero, we only report the $\Delta_1 = 0$ case in this section for brevity.

5.2.1 The stationary data: equal variance case

The common factors are generated by DGP1. We choose $T_1 = 30$, 60, 120, $T_2 = 20$ and $N_{co} = 30$, 60, 120.

Table 2 reports estimated confidence intervals using the inference theory developed in Section 3 with the number of factors determined by our proposed modified Bai and Ng's criterion PC_{p_1,m_N,m_T}. The variance of e_{it} is the same for the treated and control units: $\sigma_{tr} = \sigma_{co}$ with $\sigma_{co}^2 \in \{2, 1, 0.5\}$. Table 2 reveals that our asymptotic theory works well with estimated confidence intervals close to their nominal values for all cases considered.

Table 2: Equal Variance Case Estimated CI (DGP1): $\sigma_{tr} = \sigma_{co}$, $T_2 = 20$

σ_{co}^2	2	1	0.5	2	1	0.5	2	1	0.5
T_1/N_{co}	$N_{co} = 30$			$N_{co} = 60$			$N_{co} = 120$		
	95% CI								
30	0.9660	0.9600	0.9650	0.9660	0.9600	0.9650	0.9330	0.9350	0.9440
60	0.9655	0.9670	0.9685	0.9655	0.9670	0.9685	0.9490	0.9340	0.9410
120	0.9595	0.9580	0.9555	0.9595	0.9580	0.9555	0.9595	0.9580	0.9555
	90% CI								
30	0.9140	0.9045	0.9135	0.9120	0.9045	0.9135	0.8845	0.8900	0.9015
60	0.9090	0.9145	0.9165	0.9090	0.9145	0.9165	0.8850	0.8915	0.8825
120	0.9110	0.9050	0.9090	0.9110	0.9050	0.9090	0.8935	0.8920	0.9015
	80% CI								
30	0.7975	0.7960	0.8060	0.8030	0.7980	0.8110	0.7975	0.7960	0.8085
60	0.8010	0.7995	0.8085	0.8130	0.7980	0.7975	0.8010	0.7995	0.8085
120	0.8180	0.7925	0.8135	0.8130	0.7980	0.7975	0.8010	0.7995	0.8085
	50% CI								
30	0.4830	0.4965	0.4975	0.4910	0.4785	0.4825	0.4830	0.4965	0.4975
60	0.4945	0.4885	0.4830	0.4820	0.5060	0.4960	0.4945	0.4885	0.4905
120	0.5105	0.4950	0.4935	0.4935	0.4880	0.5020	0.5105	0.4950	0.4935

5.2.2 The stationary data: unequal variance case

Now we consider the case where the variance of the idiosyncratic error e_{it} is different for the treated and control units. We consider both the case where the control group's error variance is larger than the treated unit's error variance (Table 3) and the reverse case where the treated unit's error variance is larger than that of the control group (Table 4). The number of factors is still determined by our modified Bai and Ng's criterion PC_{p_1,m_N,m_T}.

Table 3: Unequal Variance Case Estimated CI (DGP1): $\sigma_{tr} = 0.5\sigma_{co}$, $T_2 = 20$

σ^2_{co}	2	1	0.5	2	1	0.5	2	1	0.5
T_1/N_{co}	$N_{co} = 30$			$N_{co} = 60$			$N_{co} = 120$		
	95% CI								
30	0.9554	0.9539	0.9592	0.9564	0.9636	0.9622	0.9608	0.9626	0.9632
60	0.9636	0.9630	0.9632	0.9626	0.9610	0.9620	0.9596	0.9668	0.9612
120	0.9626	0.9630	0.9635	0.9592	0.9662	0.9690	0.9586	0.9614	0.9616
	90% CI								
30	0.8999	0.9011	0.9000	0.8996	0.9102	0.9144	0.9004	0.9110	0.9184
60	0.9090	0.9117	0.9127	0.9042	0.9088	0.9138	0.9018	0.9120	0.9030
120	0.9092	0.9120	0.9149	0.9018	0.9180	0.9118	0.9056	0.9114	0.9084
	80% CI								
30	0.7841	0.7839	0.7846	0.7850	0.8042	0.8142	0.7880	0.7988	0.8120
60	0.8003	0.8069	0.8035	0.7962	0.8024	0.8034	0.7924	0.8006	0.7980
120	0.8060	0.8064	0.8099	0.7868	0.8038	0.8068	0.7918	0.7932	0.8006
	50% CI								
30	0.4796	0.4855	0.4765	0.4764	0.4888	0.5024	0.4772	0.4910	0.4982
60	0.4886	0.4911	0.4890	0.4778	0.4804	0.4806	0.4782	0.4812	0.4936
120	0.5000	0.4927	0.4895	0.4796	0.4816	0.4962	0.4740	0.4828	0.4892

From Tables 3 and 4, we see that our inference theory is robust to different error variance values for the treated unit, σ^2_{tr}, and control group, σ^2_{co}. This is in contrast to Xu (2017) who proposes using a bootstrap method for inference for the generalized synthetic control estimator. Xu's approach requires that variance for treated and control group are equal ($\sigma^2_{tr} = \sigma^2_{co}$) because he draws a bootstrap sample for the treated units from the control group's data. It is easy to see that when the error variances are unequal (i.e. $\sigma^2_{tr} < (>)\sigma^2_{co}$), his treated unit from the bootstrap sample will have either a larger (smaller) variation than that of the treated unit from the original sample. This implies that the estimated confidence intervals using a bootstrap sample will be wider (narrower) than the true confidence intervals, resulting in over (under) coverage.

5.2.3 The non-stationary data

In this subsection, we examine the performance of our confidence interval estimates when common factors are non-stationary. Hence, the outcome variables y_{it} are also non-stationary.

Table 4: Unequal Variance Case Estimated CI (DGP1): $\sigma_{tr} = 2\sigma_{co}$, $T_2 = 20$

σ^2_{co}	2	1	0.5	2	1	0.5	2	1	0.5
T_1/N_{co}	$N_{co} = 30$			$N_{co} = 60$			$N_{co} = 120$		
	95% CI								
30	0.9579	0.9610	0.9541	0.9544	0.9700	0.9632	0.9580	0.9670	0.9686
60	0.9655	0.9645	0.9677	0.9578	0.9662	0.9636	0.9568	0.9672	0.9642
120	0.9624	0.9645	0.9640	0.9540	0.9658	0.9686	0.9582	0.9618	0.9656
	90% CI								
30	0.9033	0.9020	0.9019	0.8984	0.9158	0.9156	0.9050	0.9156	0.9138
60	0.9144	0.9140	0.9128	0.9038	0.9120	0.9094	0.9066	0.9190	0.9150
120	0.9139	0.9178	0.9143	0.8910	0.9112	0.9178	0.9014	0.9080	0.9128
	80% CI								
30	0.7954	0.7915	0.7875	0.7774	0.8154	0.8058	0.7934	0.8042	0.8040
60	0.8125	0.8026	0.8090	0.7980	0.8046	0.8022	0.8072	0.8164	0.8114
120	0.8105	0.8076	0.8093	0.7752	0.8000	0.7846	0.8024	0.8068	0.8144
	50% CI								
30	0.4821	0.4699	0.4763	0.4662	0.4862	0.4964	0.4876	0.4926	0.4920
60	0.4948	0.4873	0.4986	0.4748	0.5000	0.4888	0.4808	0.4924	0.5048
120	0.4894	0.4926	0.4877	0.4694	0.4842	0.5054	0.4784	0.4836	0.4914

The idiosyncratic errors e_{it} are stationary and are generated as before. We consider the following three factors, all of which are non-stationary:

$$DGP2: \quad \begin{cases} F_{4t} = F_{4,t-1} + \epsilon_{4t}, \\ F_{5t} = (0.2 + \xi_t)t + \epsilon_{5t}, \\ F_{6t} = \sqrt{t} + \epsilon_{6t} + 0.9\epsilon_{6,t-1} + 0.4\epsilon_{6,t-2}, \end{cases}$$

where ϵ_{it} is iid $N(0,1)$, ξ_t is iid Uniform[0, 1], and ϵ_{it} and ξ_s are independent with each other for all i, t and s. The above three factors are non-stationary. F_{4t} is a non-stationary (drift-less) unit root process, F_{5t} has a linear trend component with a random coefficient uniformly distributed between 0.2 to 1.2, and F_{6t} has a non-linear (square-root) trend and an MA(2) error structure.

The outcome variables y_{it} are generated the same way as before using $y^0_{it} = \lambda_i' F_t + e_{it}$ except the three stationary factors (DGP1) are replaced by the three non-stationary factors as described in DGP2. We consider two cases of generating the idiosyncratic errors e_{it}: (i) $e_{it} \sim \sigma N(0,1)$ as before and (ii) $e_{it} \sim \sigma$ uniform$[-\sqrt{3}, \sqrt{3}]$. Simulation results are almost identical for these two cases. Therefore, we only report case (ii) for brevity. The simulation results are presented in Table 5. The results are almost identical to the stationary data case. Hence, the results strongly support our theoretical analysis that our inference procedure is valid whether the data is stationary or non-stationary, and whether or not the idiosyncratic errors are normally distributed.

Table 5: Estimated CI for DGP2: uniformly distributed e_{it}, $\sigma_{tr} = \sigma_{co}$, $T_2 = 20$

σ^2	2	1	0.5	2	1	0.5	2	1	0.5
T_1/N_{co}	$N_{co} = 30$			$N_{co} = 60$			$N_{co} = 120$		
	95% CI								
30	0.9576	0.9536	0.9552	0.9544	0.9576	0.9560	0.9560	0.9608	0.9504
60	0.9628	0.9630	0.9658	0.9626	0.9642	0.9650	0.9652	0.9620	0.9634
120	0.9646	0.9618	0.9634	0.9608	0.9636	0.9612	0.9626	0.9658	0.9660
	90% CI								
30	0.9014	0.9040	0.9028	0.9008	0.9046	0.9010	0.8988	0.9000	0.8938
60	0.9140	0.9058	0.9200	0.9132	0.9120	0.9148	0.9214	0.9040	0.9108
120	0.9074	0.9186	0.9168	0.9096	0.9158	0.9078	0.9112	0.9224	0.9162
	80% CI								
30	0.7798	0.7866	0.7864	0.7836	0.7960	0.7936	0.7780	0.7902	0.7884
60	0.8056	0.7952	0.8142	0.8050	0.8052	0.8094	0.8088	0.8042	0.8000
120	0.7982	0.8048	0.8112	0.8048	0.8140	0.8058	0.8064	0.8120	0.8080
	50% CI								
30	0.4872	0.4794	0.4830	0.4800	0.4898	0.4748	0.4685	0.4828	0.4758
60	0.4884	0.4806	0.5008	0.4854	0.4840	0.4944	0.5058	0.4928	0.4880
120	0.4888	0.4934	0.5012	0.4968	0.5032	0.4970	0.5094	0.4892	0.5018

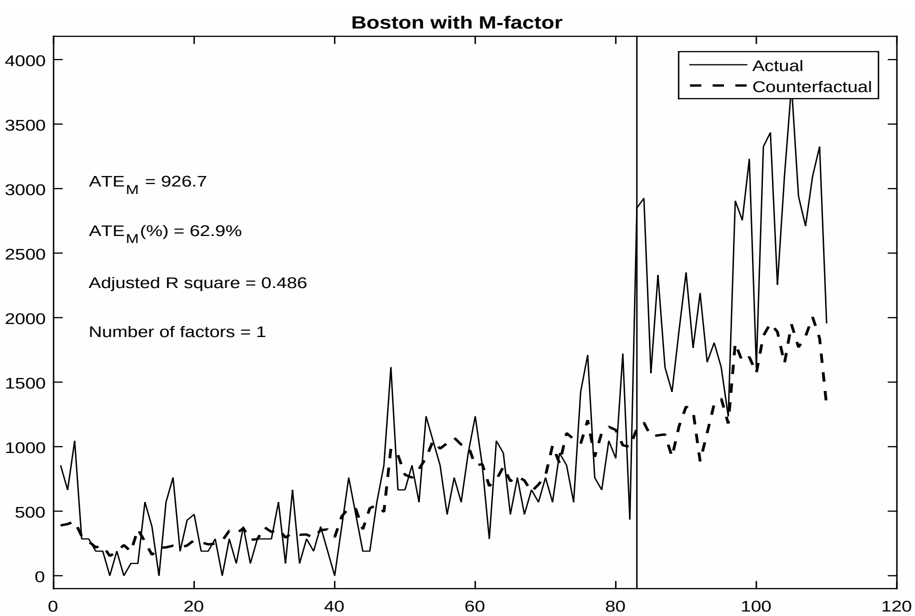
6 An Empirical Application

6.1 Estimating ATE of opening a showroom

We illustrate the usefulness of the factor-model-based inference with an application to an online-first e-tailer, WarbyParker.com. Starting completely online, WarbyParker.com disrupted the eyewear industry by offering the products at a lower price point due to cost advantages of vertical integration and online operations. However, as part of their growth strategy, WarbyParker.com opened physical showrooms in certain cities to allow customers to try their product before deciding to purchase or not. Specifically, we will examine the effect of a showroom opening in Boston on the total sales in Boston. The average treatment effect (ATE) is the average weekly change in total sales in Boston due to the openning of the showroom. The showroom opened in Boston on September 22, 2011. We use weekly sales data from February 2010 to March 2013, which is a total sample size of 110 weeks. We have 83 pre-treatment periods (before showroom openned) and 27 post-treatment periods (after showroom opened), i.e., $T_1 = 83$ and $T_2 = 27$. As the control group, we use the largest 30 U.S. cities that do not have showrooms during our sample period. First, we apply our modified Bai and Ng's (2002) criterion PC_{p1,m_N,m_T} to the control group data and select one factor.[4] The ATE estimation result, the in-sample fitted curve (the dashed curve for $t \leq 83$), and the predicted counterfactual curve (the dashed curve for $t > 83$), and the real data (the solid line) are drawn in Figure 1.

The time period when the showroom opens (treatment) is indicated by a vertical line ($T_1 = 83$). We observe that before the treatment occurs, the fitted curve traces the actual sales quite well, suggesting that our method gives a reasonably good in-sample fit. After the treatment, the two curves start to deviate from each other with the real sales being above the counterfactual sales most of the time. This indicates a positive treatment effect on sales from opening the showroom. Our results show that opening a showroom increases Boston's weekly sales by \$926.70, or a 62.9% increase in weekly sales.

Figure 1: Boston: GSC fitted curve, $N_{co} = 30$



6.2 Confidence interval estimation

We report estimated confidence intervals using (4.7). We compute the 80%, 90% and 95% confidence intervals for ATE Δ_1. They are given by $CI_{80\%} = [702.5, 1150.8]$, $CI_{90\%} =$

[4]The least squares cross validation method suggested by Xu also selects one factor for this empirical data.

$[639.0, 1214.4]$ and $CI_{95\%} = [583.9, 1269.5]$, respectively. All of these intervals lie far right of zero, suggesting that Boston's showroom opening has a significant positive effect on total weekly eyewear sales.

We also can conduct hypothesis tests for testing a specific positive ATE. That is, we can conduct a one-sided test for the null hypothesis H_1: $ATE = ATE_0$ against the alternative hypothesis H_1: $ATE > ATE_0$, where ATE_0 is a pre-specified positive constant. For example, if we conduct a 10% level one-sided test for $ATE_0 = 700$, we reject $ATE = 700$ and claim that $ATE > 700$ because the estimated 10^{th} percentile of ATE is 702.5 which is greater than 700. This can be very useful from a managerial perspective because WarbyParker.com knows the operating costs of a showroom and thus, may be interested in testing whether weekly sales revenue are above a certain cutoff to determine whether the keep the showroom open or not.

6.3 Comparison with other methods

For comparison, we also use the least squares method to estimate the ATE. This consists of the following steps: (i) Regress the treated unit outcome, y_{1t}, on control units' outcomes, $x_t = (1, y_{2t}, ..., y_{Nt})'$, using pre-treatment data to estimate β based on $y_{1t} = x_t'\beta + u_{1t}$. Let $\hat{\beta}_{OLS}$ denote the resulting estimator of β. We estimate ATE by $\hat{\Delta}_{1,OLS} = T_2^{-1} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}^0_{1t,OLS})$, where $\hat{y}_{1t,OLS} = x_t'\hat{\beta}_{OLS}$. The OLS-based method yields an estimated ATE of \$1041 increase in weekly sales or an 68.3% increase in weekly sales due to opening a showroom in Boston.

To examine which method gives a more accurate ATE estimate, we examine the out-of-sample prediction performances of the two methods. We select a value $T_0 < T_1$ and treat $T_0 + 1$ as a pseudo treatment time. We estimate the counterfactual outcome y^0_{1t} for $t = T_0 + 1, ..., T_1$. Specifically, let $\tilde{F}_t$ be the $r \times 1$ vector of estimated factor using the control group data for $t = 1, ..., T_1$. We estimate λ_1 by $\tilde{\lambda}_1 = (\tilde{F}'\tilde{F})^{-1}\tilde{F}'\tilde{y}_1$, where $\tilde{F}$ is the $T_0 \times r$ matrix of estimated factors and $\tilde{y}_1 = (y_{11}, ..., y_{1T_0})'$. We then estimate y^0_{1t} by $\hat{y}^0_{1t} = \tilde{\lambda}_1'\tilde{F}_t$ for $t = T_0 + 1, ..., T_1$. However, because there were no treatments during $T_0 + 1, ..., T_1$, we in fact observe $y^0_{1t} = y_{1t}$ for $t = T_0 + 1, ..., T_1$. Hence, we can compute the prediction mean squared error (PMSE) for our method by

$$PMSE_{GSC} = \frac{1}{T_1 - T_0} \sum_{t=T_0+1}^{T_1} (y_{1t} - \hat{y}^0_{1t})^2.$$

Similarly, we compute PMSE using Xu's (2017) method. That is, first we use the leave-one-out cross validation method to determine the number of factors using the pre-treatment sample (so the selected number of factors depends on T_0) and then we estimate PMSE as discussed above. We also use the least squares method as discussed in Section 2.2 to compute PMSE. We use $PMSE_{GSC,Xu}$ and $PMSE_{OLS}$ to denote the resulting PMSEs. We report PMSE ratio $PMSE_{OLS}/PMSE_{GSC}$ and $PMSE_{GSC,Xu}/PMSE_{GSC}$ where GSC is our method. The results are shown in Table 6.

Table 6: Out-of-sample Prediction MSE ratio

T_0	45	50	55	60	65	70	75	80
$\frac{PMSE_{OLS}}{PMSE_{GSC}}$	2.484	2.068	1.369	1.414	1.565	1.352	1.463	1.118
$\frac{PMSE_{GSC,Xu}}{PMSE_{GSC}}$	1.000	1.455	1.392	1.000	1.000	1.000	1.337	1.403
$\#factors$ by M-BN	1	1	1	1	1	1	1	1
$\#factors$ by Xu	1	8	6	1	1	1	6	6

From Table 6 we see that our proposed method gives smaller PMSE for all cases compared to the least-squares method. When compared to Xu's method (GSC, Xu), there are four cases where both our method and Xu's method select one factor and the PMSEs are the same for these four cases. For the remaining four cases, Xu's method selects more factors and gives a larger PMSE than our method. Thus, our proposed method performs well when compared with these competing methods.

7 Conclusion

In this paper we consider using a factor-model-based method, also known as the generalized synthetic control method, to estimate average treatment effects. This method is best suited for cases where there is only one (or a few) treated unit(s), a large number of control units, and large pre and post-treatment sample sizes (i.e., long panel). Long panel data are quite common in marketing due to the prevalence of daily and weekly data at the customer, store or company level.

Existing inference methods either assume that idiosyncratic errors have the same variance for the treated and control units or require that the treated units's error variances to be the same during the pre and post-treatment periods. However, equal variance assumption can be violated in practice because the treated and control units are usually heterogeneous without random assignment of treated units (i.e. any quasi-experimental setting). In addition, because treatment may affect ATE's variance in addition to its mean, the treated unit's idiosyncratic error variance can be substantially different during pre and post-treatment periods. Filling a gap in the literature, our inference for the factor model based ATE addresses both of these issues and provides previously unknown distribution theory properties. We establish asymptotic (normal) distribution theory of the ATE estimator that allows for idiosyncratic errors to have different variances for the treated and control groups, and during the pre- and post-treatment periods. In addition, the inference theory holds for both stationary and non-stationary data. To the best of our knowledge, we are the first to establish distribution theory for the generalized synthetic control method. We propose a modified Bai and Ng model selection criterion and show that it performs well in finite sample applications. Simulation results support our theoretical analysis and an empirical application that examines the effect of opening a showroom by WarbyParker.com on its average weekly sales demonstrates the usefulness of the factor-model-based method in estimating ATE. With the rise of quasi-experimental panel data with long panels and many control units, factor-model-based methods and the ability to conduct inference are powerful tools for marketing and social science researchers.

Appendix A: Proofs of Theorems 4.1 and 4.2

A.1: Proof of Theorem 4.1

Throughout this Appendix, we sometimes replace N_{co} by $N = N_{co} + 1 = O(N_{co})$ to simply notation, especially when we evaluate the order of a term because such a substitution will not affect our asymptotic analysis.

Since both the factors and factor loadings are unobservable. They are not identified without some identification conditions. In this subsection we use the same identification conditions as in Bai (2003). Let F denote the $T \times r$ matrix of common factors and Λ denote the $N \times r$ factor loading matrix. Following Bai (2003) we know that there exists an invertible matrix H such that $\hat{F}$ is an estimator of FH and $\hat{\Lambda}$ is an estimator of $\Lambda(H')^{-1}$. Because F and FH span the same space, a regression model using F or FH give the same result. We also adopt assumptions A to G in Bai (2003) so that we can borrow all needed results from Bai (2003). In addition we assume that that $(e_{1t}, F_t)_{t=1}^{T_1}$ and $(e_{1t}, F_t, \Delta_{1t})_{t=T_1+1}^{T}$ are weakly dependent ρ-mixing process with mixing coefficients satisfying $\rho(\tau) = \lambda^\tau$ for some $0 < \lambda < 1$ (see page 74 of Li and Bell (2017) for the definition of a ρ-mixing process). This enables us to use a result from Li and Bell (2017, Lemma A.1) to derive the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$. Also, we assume that $\lim_{T_2,N\to\infty} T_2/N^2 = 0$ and $\lim_{T_1,T_2\to\infty} T_2/T_1 = \phi$, where ϕ is a non-negative finite constant. We use a notation introduced in Bai (2003) that $\delta_{N,T} = \min\{\sqrt{N}, \sqrt{T}\}$ so that $\delta_{N,T}^{-2} = \max\left\{\frac{1}{N}, \frac{1}{T}\right\}$.

We first give an outline of the proof of Theorem 4.1. Let $\tilde{\Delta}_1$ be defined by replacing $\hat{F}_t$ by F_t in $\hat{\Delta}_1$, we will show that

$$\tilde{\Delta}_1 = \hat{\Delta}_1 + O_p(\delta_{N,T_2}^{-2}) = \hat{\Delta}_1 + O\left(\max\left\{\frac{1}{N}, \frac{1}{T_2}\right\}\right). \tag{A.1}$$

Hence, we have

$$\begin{aligned} \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) &= \sqrt{T_2}(\tilde{\Delta}_1 - \Delta_1) + \sqrt{T_2}(\hat{\Delta}_1 - \tilde{\Delta}_1) \\ &= \sqrt{T_2}(\tilde{\Delta}_1 - \Delta_1) + o_p(1) \end{aligned}$$

because as in Bai (2004), we assume that $T_2 = o(N^2)$. Thus, the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$ is the same as that of $\sqrt{T_2}(\tilde{\Delta}_1 - \Delta_1)$. The asymptotic of $\sqrt{T_2}(\tilde{\Delta}_1 - \Delta_1)$ follows from Theorem 3.2 of Li and Bell (2017). A detailed proof of Theorem 4.1 is given below.

By definition, $\hat{\lambda}_1 = \left[\sum_{t=1}^{T_1} \hat{F}_t \hat{F}_t'\right]^{-1} \left[\sum_{t=1}^{T_1} \hat{F}_t y_{1t}\right]$, and $\hat{y}_{1t}^0 = \hat{\lambda}_1' \hat{F}_t = \left[\sum_{s=1}^{T_1} y_{1s} \hat{F}_s'\right] \left[\sum_{s=1}^{T_1} \hat{F}_s \hat{F}_s'\right]^{-1} \hat{F}_t$ for $t = T_1 + 1, \ldots, T$. Then

$$\hat{\Delta}_{1t} \overset{def}{=} y_{1t} - \hat{y}_{1t}^0 = y_{1t} - \left[\sum_{s=1}^{T_1} y_{1s} \hat{F}_s'\right] \left[\sum_{s=1}^{T_1} \hat{F}_s \hat{F}_s'\right]^{-1} \hat{F}_t \qquad \text{for } t = T_1 + 1, \ldots, T. \tag{A.2}$$

By our model specification, for pre-treatment period,

$$y_{1s} = \lambda_1' F_s + e_{1s}, \qquad s = 1, 2, \ldots, T_1, \tag{A.3}$$

and for post-treatment period,

$$y_{1t} = F_t' \lambda_1 + \Delta_1 + v_{1t}, \qquad t = T_1 + 1, \ldots, T, \tag{A.4}$$

where $\Delta_1 = E(\Delta_{1t})$ and $v_{1t} = \Delta_{1t} - \Delta_1 + e_{1t}$.

Substituting (A.4) into (A.2), and using $y_{1s} = \lambda_1' F_s + e_{1s}$ for $s = 1, ..., T_1$, we get

$$\hat{\Delta}_{1t} = \Delta_1 + v_{1t} + \left\{F_t - \left[\sum_{s=1}^{T_1} F_s \hat{F}_s'\right] \left[\sum_{s=1}^{T_1} \hat{F}_s \hat{F}_s'\right]^{-1} \hat{F}_t\right\}' \lambda_1 - \left[\sum_{s=1}^{T_1} e_{1s} \hat{F}_s'\right] \left[\sum_{s=1}^{T_1} \hat{F}_s \hat{F}_s'\right]^{-1} \hat{F}_t. \tag{A.5}$$

Therefore, the ATE estimator $\hat{\Delta}_1 = T_2^{-1}\sum_{t=T_1+1}^{T}\hat{\Delta}_{1t}$ is given by

$$\begin{aligned}
\hat{\Delta}_1 &= \Delta_1 + \frac{1}{T_2}\sum_{t=T_1+1}^{T} v_{1t} + \underbrace{\frac{1}{T_2}\sum_{t=T_1+1}^{T}\Big\{F_t - \Big[\sum_{s=1}^{T_1} F_s\hat{F}_s'\Big]\Big[\sum_{s=1}^{T_1}\hat{F}_s\hat{F}_s'\Big]^{-1}\hat{F}_t\Big\}'\lambda_1}_{B_1} \\
&\quad \underbrace{-\Big[\sum_{s=1}^{T_1} e_{1s}\hat{F}_s'\Big]\Big[\sum_{s=1}^{T_1}\hat{F}_s\hat{F}_s'\Big]^{-1}\Big[\frac{1}{T_2}\sum_{t=T_1+1}^{T}\hat{F}_t\Big]}_{B_2} \\
&= \Delta_1 + \frac{1}{T_2}\sum_{t=T_1+1}^{T} v_{1t} + \Big[\sum_{s=1}^{T_1} e_{1s}F_s'\Big]\Big[\sum_{s=1}^{T_1} F_sF_s'\Big]^{-1}\Big[\frac{1}{T_2}\sum_{t=T_1+1}^{T} F_t\Big] + O_p(\delta_{N,T_2}^{-2})
\end{aligned} \tag{A.6}$$

where the last equality follows from Lemma A.3 and Lemma A.4 (for B_1 and B_2).

If we compare (A.5) with (A.6), we see that if one replaces $\hat{F}_t$ in (A.5) by F_t, one obtains the leading term of (A.6). Consequently, the asymptotic distribution of $\hat{\Delta}_1$ using $\hat{F}_t$ is the same as that using F_t. The reason that using $\hat{F}_t$ leads to the same asymptotic distribution as using F_t is that, although $\hat{F}_t - HF_t = O_p(N^{-1/2})$; by Lemma A.2 (ii) we know that $T^{-1}\sum_{t=1}^{T}(\hat{F}_t - HF_t) = O_p(\max\{N^{-1}, T^{-1}\}) = o_p(N^{-1/2})$ (because $N = o(T^2)$). In general, averaging over random variables reduce variance of the random variable and hence leads a smaller order quantity.

Hence, under the conditions that $\sqrt{T_2}/N \to 0$ and $T_2/T_1 \to \phi$, we have

$$\begin{aligned}
\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) &= -\sqrt{\phi}\frac{1}{\sqrt{T_1}}\sum_{s=1}^{T_1} e_{1s}F_s'\Big[\frac{1}{T_1}\sum_{s=1}^{T_1} F_sF_s'\Big]^{-1}\Big[\frac{1}{T_2}\sum_{t=T_1+1}^{T} F_t\Big] + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\
&= -\sqrt{\phi}\frac{1}{\sqrt{T_1}}\sum_{s=1}^{T_1} e_{1s}F_s'\Big[E(F_sF_s') + o_p(1)\Big]^{-1}\Big[E(F_t) + o_p(1)\Big] + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\
&= -\sqrt{\phi}\frac{1}{\sqrt{T_1}}\sum_{s=1}^{T_1} e_{1s}F_s'C + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\
&\equiv A_1 + A_2,
\end{aligned}$$

where $A_1 = -\sqrt{\phi}\frac{1}{\sqrt{T_1}}\sum_{s=1}^{T_1} e_{1s}F_s'C$, $C = \Big[E(F_sF_s')\Big]^{-1}\Big[E(F_t)\Big]$ is a $r \times 1$ vector of constants and $A_2 = \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t}$.

Under the assumption that $e_{1t}F_t$ and v_{1t} are zero mean weakly dependent processes so that central limit theorem hold for their partial sums, we have

$$A_1 \xrightarrow{d} N(0, \Omega_1), \qquad A_2 \xrightarrow{d} N(0, \Omega_2),$$

where Ω_1 and Ω_2 are defined in Theorem 4.1. Also, under the assumption that $(e_{1t}, F_t)_{t=1}^{T_1}$ and $(e_{1t}, F_t, \Delta_{1t})_{t=T_1+1}^{T}$ are ρ-mixing process with mixing coefficients satisfying $\rho(\tau) = \lambda^\tau$ for some $0 < \lambda < 1$. Then by Lemma A.1 of Li and Bell (2017) we know that $Cov(A_1, A_2) = o(1)$. Hence, we have that

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) = A_1 + A_2 + o_p(1) \xrightarrow{d} N(0, \Omega_1 + \Omega_2).$$

This completes the proof of Theorem 4.1.

Consistent estimators of Ω_1 and Ω_2

A consistent estimator of Ω_1 is given by

$$\hat{\Omega}_1 = (T_2/T_1)\hat{E}(F_t)'(\hat{V}/T_2)\hat{E}(F_t),$$

where $\hat{E}(F_t) = T_2^{-1}\sum_{t=T_1+1}^{T}\hat{F}_t$,

$$\hat{V} = [\hat{E}(F_tF_t')]^{-1}[T_1^{-1}\sum_{t=1}^{T_1}\sum_{s=1,|s-t|\le l}^{T_1}\hat{e}_{1t}\hat{e}_{1s}\hat{F}_t\hat{F}_s][\hat{E}(F_tF_t')]^{-1} \tag{A.7}$$

where $\hat{E}(F_tF_t') = T_1^{-1}\sum_{t=1}^{T_1}\hat{F}_t\hat{F}_t'$, $\hat{e}_{1t} = y_{1t} - \hat{\lambda}_1'\hat{F}_t$, l satisfies that $l \to \infty$ and $l/T_1 \to \infty$ as $T_1 \to \infty$, one can choose $l = O(T_1^{1/4})$ as suggested by Newey and West (1987).

A consistent estimator of Ω_2 is given by

$$\hat{\Omega}_2 = \frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1,|t-s|\le h}^{T}[\hat{\Delta}_{1t} - \hat{\Delta}_1][\hat{\Delta}_{1s} - \hat{\Delta}_1],$$

where $h \to \infty$ and $h/T_2 \to 0$ as $T_2 \to \infty$. For example, one can choose $h = O(T_2^{1/4})$.

When e_{1t} and Δ_{1t} are serially uncorrelated, $\hat{\Omega}_1$ and $\hat{\Omega}_2$ can be simplified. In $\hat{\Omega}_1$, one can replace $\hat{V}$ defined in (A.7) by $\hat{V} = T_1^{-1}\sum_{t=1}^{T_1}\hat{e}_{1t}^2\hat{F}_t\hat{F}_t'$, and $\hat{\Omega}_2$ is simplified to $\hat{\Omega}_2 = \frac{1}{T_2}\sum_{t=T_1+1}^{T}(\hat{\Delta}_{1t} - \hat{\Delta}_1)^2$.

Some useful lemmas

Below we present some Lemmas that are used in the proof of Theorem 4.1. Lemma A.1 summaries some useful results from Bai (2003).

Lemma A.1 *Let $\delta_{N,T} = \min\{\sqrt{N}, \sqrt{T}\}$. Then under the regularity conditions given in Bai (2003), we have*

(i) $\frac{1}{T_1}\sum_{t=1}^{T_1}||\hat{F}_t - H'F_t||^2 = O_p(\delta_{N,T_1}^{-2})$.

(ii) For all $i = 1, ..., N$, $\frac{1}{T_1}\sum_{t=1}^{T_1}(\hat{F}_t - H'F_t)e_{it} = O_p(\delta_{N,T_1}^{-2})$.

(iii) $\frac{1}{T_1}\sum_{s=1}^{T_1} F_s\hat{F}_s' = \frac{1}{T_1}\sum_{s=1}^{T_1} F_sF_s'H' + O_p(\delta_{N,T_1}^{-2})$.

(iv) $\frac{1}{T_1}\sum_{t=1}^{T_1}(\hat{F}_t - H'F_t)'\hat{F}_t = O_p(\delta_{N,T_1}^{-2})$.

Proof: See Bai's (2003) Lemma A.1, Lemma B.1, Lemma B.2 and Lemma B.3 for proofs of (i), (ii), (iii) and (iv), respectively.

Lemma A.2 *Under the regularity conditions as in Lemma A.1 we have*

(i) $\frac{1}{T_1}\sum_{s=1}^{T_1}\hat{F}_s\hat{F}_s' = \frac{1}{T_1}\sum_{s=1}^{T_1} HF_sF_s'H' + O_p(\delta_{N,T_1}^{-2})$.

(ii) $\frac{1}{T_2}\sum_{t=T_1+1}^{T}(F_t - H^{-1}\hat{F}_t) = O_p(\delta_{N,T_2}^{-2})$.

Proof of Lemma A.2 (i): By adding/subtracting terms, we have

$$\begin{aligned}\frac{1}{T_1}\sum_{s=1}^{T_1}\hat{F}_s\hat{F}_s' &= \frac{1}{T_1}\sum_{s=1}^{T_1}(\hat{F}_s - HF_s)(\hat{F}_s - HF_s)' + H\frac{1}{T_1}\sum_{s=1}^{T_1}F_s(\hat{F}_s - HF_s)' \\ &+ \left[\frac{1}{T_1}\sum_{s=1}^{T_1}(\hat{F}_s - HF_s)F_s'\right]H' + \frac{1}{T_1}\sum_{s=1}^{T_1}HF_sF_s'H'.\end{aligned} \tag{A.8}$$

The first term on the right-hand-side of (B.1) is $O_p(\delta_{N,T_1}^{-2})$ by Lemma A.1 (i). By noting that $H = O(1)$ and that H is time-invariant, we see that the second and the third terms on the right-hand-side of (B.1) are both $O_p(\delta_{N,T_1}^{-2})$ by Lemma A.1 (iii). Hence, we get

$$\frac{1}{T_1}\sum_{s=1}^{T_1}\hat{F}_s\hat{F}_s' = \frac{1}{T_1}\sum_{s=1}^{T_1}HF_sF_s'H' + O_p(\delta_{N,T_1}^{-2}).$$

This completes the proof of Lemma A.2 (i).

Proof of Lemma A.2 (ii):

Lemma B.2 of Bai (2003) establishes $T^{-1}\sum_{t=1}^{T}(\hat{F}_t - H'F_t)F_t' = O_p(\delta_{N,T}^{-2})$. Replacing the second F_t by 1, by using the exactly the same arguments as in the proof of Lemma B.2 of Bai (2003), one can show that $T^{-1}\sum_{t=1}^{T}(\hat{F}_t - H'F_t) = O_p(\delta_{N,T}^{-2})$, which implies that $T_2^{-1}\sum_{t=T_1+1}^{T}(\hat{F}_t - H'F_t) = O_p(\delta_{N,T_2}^{-2})$.

Lemma A.3 *Define $B_1 = \frac{1}{T_2}\sum_{t=T_1+1}^{T}\left\{F_t - \left[\sum_{s=1}^{T_1}F_s\hat{F}_s'\right]\left[\sum_{s=1}^{T_1}\hat{F}_s\hat{F}_s'\right]^{-1}\hat{F}_t\right\}$.*

Then $B_1 = O_p(\delta_{N,T_2}^{-2})$.

Proof: By Lemma A.2 (i) we know that

$$\frac{1}{T_1}\sum_{s=1}^{T_1}\hat{F}_s\hat{F}_s' \;=\; \frac{1}{T_1}\sum_{s=1}^{T_1}HF_sF_s'H' + O_p(\delta_{N,T_1}^{-2}). \tag{A.9}$$

Also, by Lemma A.1 (iii) we have that

$$\frac{1}{T_1}\sum_{s=1}^{T_1}F_s\hat{F}_s' = \frac{1}{T_1}\sum_{s=1}^{T_1}F_sF_s'H' + O_p(\delta_{N,T_1}^{-2}). \tag{A.10}$$

Substituting (B.9) and (B.10) into the expression of B_1, we get

$$B_1 \;=\; \frac{1}{T_2}\sum_{t=T_1+1}^{T}\left\{F_t - H^{-1}\hat{F}_t\right\} + O_p(\delta_{N,T_1}^{-2}) = O_p(\delta_{N,T_2}^{-2}), \tag{A.11}$$

where the last equality follows from $T_2^{-1}\sum_{t=T_1+1}^{T}\left(F_t - H^{-1}\hat{F}_t\right) = O_p(\delta_{N,T_2}^{-2})$ by Lemma A.2 (ii) and that $T_2 = O(T_1)$.

Lemma A.4 *Define* $B_2 = \left[\sum_{s=1}^{T_1} e_{1s}\hat{F}_s'\right]\left[\sum_{s=1}^{T_1}\hat{F}_s\hat{F}_s'\right]^{-1}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T}\hat{F}_t\right]$.

Then

$$B_2 = \left[\sum_{s=1}^{T_1} e_{1s}F_s'\right]\left[\sum_{s=1}^{T_1}F_sF_s'\right]^{-1}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T}F_t\right] + O_p(\delta_{N,T_2}^{-2}).$$

Proof: By adding/subtracting terms, we get

$$\begin{aligned}\frac{1}{T_1}\sum_{s=1}^{T_1}e_{1s}\hat{F}_s' \;&=\; \frac{1}{T_1}\sum_{s=1}^{T_1}e_{1s}(\hat{F}_s - HF_s)' + \frac{1}{T_1}\sum_{s=1}^{T_1}e_{1s}F_s'H' \\ &=\; \frac{1}{T_1}\sum_{s=1}^{T_1}e_{1s}F_s'H' + O_p(\delta_{N,T_1}^{-2}) \end{aligned} \tag{A.12}$$

by Lemma A.1 (ii).

Substituting (B.9) and (B.12) into B_2, also using the result of Lemma A.2 (ii) and noting that $T_2 = O(T_1)$, we obtain

$$B_2 = \frac{1}{T_1}\sum_{s=1}^{T_1}e_{1s}F_s'\left[\frac{1}{T_1}\sum_{s=1}^{T_1}F_sF_s'\right]^{-1}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T}F_t\right] + O_p(\delta_{N,T_2}^{-2}).$$

This completes the proof of Lemma A.4.

A.2: Proof of Theorem 4.2

We adopt assumptions A to H in Bai (2004) so that we can use all the results presented in Section 2 to Section 4 in Bai (2004). Let $\hat{k}$ denote the number of factors obtained by minimizing $IPC_1(k)$ defined in (4.9). Because $Pr(\hat{k} = r) \to 1$ as $T, N \to \infty$,[5] we can only consider the case of $\hat{k} = r$, that is, $\hat{F}_t$ is an $r \times 1$ vector. Also, we use the notation in Bai (2004) that $\delta_{2,NT} = \min\{\sqrt{N}, T\}$ so that $\delta_{2,NT}^{-1} = \max\{1/\sqrt{N}, 1/T\}$. We now start the proof of Theorem 4.2.

Because the ATE estimator is defined with the same algebraic expression whether the data is stationary or not, by the same derivations that lead to (A.6), for the non-stationary data case, the ATE estimator $\hat{\Delta}_1$ has the following expression.

$$\begin{aligned}\hat{\Delta}_1 &= \Delta_1 + \frac{1}{T_2}\sum_{t=T_1+1}^{T} v_{1t} + \lambda_1' \underbrace{\frac{1}{T_2}\sum_{t=T_1+1}^{T}\left\{F_t - \Big[\sum_{s=1}^{T_1} F_s\hat{F}_s'\Big]\Big[\sum_{s=1}^{T_1}\hat{F}_s\hat{F}_s'\Big]^{-1}\hat{F}_t\right\}}_{\mathcal{B}_1} \\ &\quad \underbrace{-\Big[\sum_{s=1}^{T_1} e_{1s}\hat{F}_s'\Big]\Big[\sum_{s=1}^{T_1}\hat{F}_s\hat{F}_s'\Big]^{-1}\Big[\frac{1}{T_2}\sum_{t=T_1+1}^{T}\hat{F}_t\Big]}_{\mathcal{B}_2} \\ &= \Delta_1 + \frac{1}{T_2}\sum_{t=T_1+1}^{T} v_{1t} - \Big[\sum_{s=1}^{T_1} e_{1s}F_s'\Big]\Big[\sum_{s=1}^{T_1}F_sF_s'\Big]^{-1}\Big[\frac{1}{T_2}\sum_{t=T_1+1}^{T}F_t\Big] + o_p(T_2^{-1/2}) \qquad \text{(A.13)}\end{aligned}$$

where the second equality follows from Lemma B.4 and Lemma B.5 (for $\mathcal{B}_1$ and $\mathcal{B}_2$) of the supplementary Appendix B.

From (A.13) we obtain

$$\begin{aligned}\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) &= -\frac{1}{\sqrt{T_2}}\left[\sum_{s=1}^{T_1} e_{1s}F_s'\right]\Big[\sum_{s=1}^{T_1}F_sF_s'\Big]^{-1}\Big[\sum_{s=T_1+1}^{T}F_s\Big] + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\ &\equiv A_1 + A_2 + o_p(1), \qquad \text{(A.14)}\end{aligned}$$

where the definitions of A_1 and A_2 should be apparent.

Let $C_T = \Big[\sum_{s=1}^{T_1} F_sF_s'\Big]^{-1}\Big[\sum_{t=T_1+1}^{T} F_t\Big]$. Then we have

$$A_1 = -\frac{1}{\sqrt{T_2}}\left[\sum_{s=1}^{T_1} e_{1s}F_s'\right] C_T.$$

Now we first make a strong assumption that $\{e_{1t}\}_{t=1}^T$ is iid $N(0, \sigma_e^2)$ and is independent with $\{u_t\}_{t=1}^T$ (recall that $F_t = F_{t-1} + u_t$), hence, independent with $\{F_t\}_{t=1}^T$. Then conditional on

[5] When discussing the order of an quantity, we often use N rather than N_{co} as they have the same order of magnitude.

the common factor F, we have

$$A_1|F \overset{d}{\sim} N(0, \Sigma_{1,T}), \tag{A.15}$$

where $\Sigma_{1,T} = \sigma_e^2 C_T' [\sum_{s=1}^{T_1} F_s F_s'] C_T / T_2$.

Now we make another strong assumptions that v_{1t} is normally distributed with zero mean and finite variance so that

$$A_2 \sim N(0, \Sigma_{2,T}), \tag{A.16}$$

where $\Sigma_{2,T} = T_2^{-1} \sum_{t=T_1+1}^{T} \sum_{s=T_1+1}^{T} E(v_{1t} v_{1s})$. Also, assume that $\{v_{1t}\}_{t=T_1+1}^{T}$ is independent with $\{F_t\}_{t=1}^{T}$, $\{e_{1t}\}_{t=1}^{T_1}$ and $\{e_{it}\}_{t=1}^{T}$ for $2 = 1, ..., N$. Then A_1 and A_2 are independent with each other. Hence, we have

$$\frac{A_1 + A_2}{\sqrt{\Sigma_{1,T} + \Sigma_{2,T}}} |F \overset{d}{\sim} N(0, 1). \tag{A.17}$$

Since the right-hand-side of (A.17) does not depend on F, we have, unconditionally,

$$\frac{A_1 + A_2}{\sqrt{\Sigma_{1,T} + \Sigma_{2,T}}} \overset{d}{\sim} N(0, 1). \tag{A.18}$$

The above strong assumptions can be relaxed. For example, by assuming that v_{1t} is a zero mean weakly dependent process such that a central limit theorem applies to its partial sum, then we have $A_2/\sqrt{\Sigma_{2,T}} \overset{d}{\to} N(0, 1)$. For A_1, Saikkonen (1991) and Stock and Watson (1993) show that the normality assumption on e_{1t} can be dropped by assuming proper (functional) central limit theorems hold to partial sums involving e_{1t} and u_s. Also, the independence assumption between e_{1t} and u_s can be relaxed to be uncorrelated for all t and s. Hayashi (2000, page 658) provides a summary of these results. The uncorrelated assumption can be further relaxed. When e_{1t} and u_s are correlated, but e_{1t} is uncorrelated with u_{t-s} for $|s| > p$, where p is a positive integer. Then one can add lead and lag values of $\Delta \hat{F}_{t-\tau} = \hat{F}_{t-\tau} - \hat{F}_{t-\tau-1}$, for $\tau = -p, -p+1, ..., p$, as additional regressors in the co-integration model to make the new idiosyncratic error, say $\tilde{e}_{1t}$, to be uncorrelated with u_s for all t and s, see Hamilton (1994, pages 608-610) for a detailed discussion on this. We assume that proper regularity conditions discussed in Stock and Watson (1993) hold so that the t-statistic for $\hat{\lambda}_1$ has an asymptotic standard normal distribution. This, together with that A_1 and A_2 are asymptotically independent with each other, leads to the following result.

$$\frac{A_1 + A_2}{\sqrt{\Sigma_{1,T} + \Sigma_{2,T}}} \overset{d}{\to} N(0, 1). \tag{A.19}$$

In applications we replace $\Sigma_{l,T}$ by $\hat{\Sigma}_{l,T}$, where for $l = 1, 2$, $\hat{\Sigma}_{l,T}$ can be obtained from $\Sigma_{l,T}$ with F_t, v_{1t} and σ_e^2 replaced by $\hat{F}_t$, $\hat{v}_{1t} = \hat{\Delta}_{1t} - \hat{\Delta}_1$ and $\hat{\sigma}_e^2 = T_1^{-1} \sum_{t=1}^{T_1} \hat{e}_{1t}^2$ with $\hat{e}_{1t} = y_{1t} - \hat{\lambda}_1' \hat{F}_1$ for $t = 1, ..., T_1$. It is easy to show that $\hat{\Sigma}_{l,T}/\Sigma_{l,T} = 1 + o_p(1)$ for $l = 1, 2$. Hence, we have

$$\frac{\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)}{\sqrt{\hat{\Sigma}_{1,T} + \hat{\Sigma}_{2,T}}} = \frac{A_1 + A_2}{\sqrt{\hat{\Sigma}_{1,T} + \hat{\Sigma}_{2,T}}} = \frac{A_1 + A_2}{\sqrt{\Sigma_{1,T} + \Sigma_{2,T}}} + o_p(1) \ \xrightarrow{d} \ N(0,1). \tag{A.20}$$

This completes the proof of Theorem 4.2.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of Californias tobacco control program. *Journal of the American statistical Association*, 105(490), 493-505.

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American economic review*, 93(1), 113-132.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, 47-57.

Ashenfelter, O. & Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67(4), 648-660.

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

Bai, J. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics*, 122(1), 137-183.

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1), 191-221.

Chan, M., & Kwok, S. S. M. (2016). Policy Evaluation with Interactive Fixed Effects (No. 2016-11). University of Sydney, School of Economics.

Chen, P. Y., Hong, Y., & Liu, Y. (2017). The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. Forthcoming in *Man-*

agement Science.

Chernozhukov, V., Wuthrich, K., & Zhu, Y. (2017). An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls. arXiv preprint arXiv:1712.09089.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3), 345-354.

Doudchenko, N., & Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis (No. w22791). National Bureau of Economic Research.

Du, Z., & Zhang, L. (2015). Home-purchase restriction, property tax and housing price in China: A counterfactual analysis. *Journal of Econometrics*, 188(2), 558-568.

Ferman, B., & Pinto, C. (2016). Revisiting the synthetic control estimator.

Firpo, S., & Possebom, V. (2017). Synthetic control method: Inference, sensitivity analysis and confidence sets.

Forman, C., Ghose, A., & Goldfarb, A. (2009). Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management science*, 55(1), 47-57.

Gobillon, L., & Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls.*Review of Economics and Statistics*, 98(3), 535-551.

Hahn, J., & Shi, R. (2017). Synthetic control and inference. *Econometrics*, 5(4), 52.

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2). Princeton: Princeton university press.

Hayashi, F. (2000). *Econometrics*. 2000. Princeton University Press. Section, 1, 60-69.

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

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

Li, K.T. (2017). Statistical inference for Synthetic Control Methods

Liaukonyte, J., Teixeira, T., & Wilbur, K. C. (2015). Television advertising and online shopping. *Marketing Science*, 34(3), 311-330.

Newey, W., & West, K. (1987). A simple positive semidefinite heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 55, 703 -708.

Saikkonen, P. (1991). Asymptotically efficient estimation of cointegration regressions. *Econometric Theory*, 7(1), 1-21.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422), 486-494.

Stock, J. H., & Watson, M. W. (1993). A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica*, 783-820.

Tirunillai, S., & Tellis, G. J. (2017). Does offline TV advertising affect online chatter? Quasi-experimental analysis using synthetic control. *Marketing Science*, 36(6), 862-878.

Wang, K., & Goldfarb, A. (2017). Can offline stores drive online sales?. *Journal of Marketing Research*, 54(5), 706-719.

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

Supplementary Appendix B: Some useful lemmas

In this appendix, we present some lemmas which are used in the proof of Theorem 4.2.

Lemma B.1 *Let $\delta_{2,NT} = \min\{\sqrt{N}, T\}$. Then under the regularity conditions given in Bai (2004), we have*

(i) $\frac{1}{T}\sum_{t=1}^{T} ||\hat{F}_t - H_1' F_t||^2 = O_p(\delta_{2,NT}^{-2})$;

(ii) $\frac{1}{T_1^2}\sum_{t=1}^{T_1} \hat{F}_t F_t' = \frac{1}{T_1^2}\sum_{t=1}^{T_1} H_1' F_t F_t' + O_p(T_1^{-1}\delta_{2,NT_1}^{-1})$;

(iii) $\frac{1}{T_1^2}\sum_{t=1}^{T_1}(\hat{F}_t - H_1' F_t)' \hat{F}_t = O_p(T_1^{-1}\delta_{2,NT_1}^{-1})$.

Note that all the results in Lemma B.1 hold true when $(\hat{F}_t, H_1)$ is replaced by $(\tilde{F}_t, H_2)$.
Proof of (i), see the proof of Lemma 1 in Bai (2004, equation (6) in page 143).
Proofs of (ii) and (iii), see Lemma B.4 (i) and Lemma B.4 (ii) of Bai (2004, page 171) for proofs of (ii) and (iii), respectively.

Lemma B.2 *Under the regularity conditions as in Lemma B.1 we have*

(i) $\frac{1}{T_1^2}\sum_{t=1}^{T_1} \hat{F}_t \hat{F}_t' = \frac{1}{T_1^2}\sum_{t=1}^{T_1} H_1' F_t F_t' H_1 + O_p(T_1^{-1}\delta_{2,NT_1}^{-1})$.

(ii) $\frac{1}{T_2}\sum_{t=T_1+1}^{T}(\hat{F}_t - H_1' F_t) = O_p((NT_2)^{-1/2})$.

Proof of Lemma B.2 (i): By adding/subtracting terms, we have

$$\begin{aligned}\frac{1}{T_1^2}\sum_{t=1}^{T_1}\hat{F}_t\hat{F}_t' &= \frac{1}{T_1^2}\sum_{s=1}^{T_1}(\hat{F}_t - H_1'F_t)(\hat{F}_t - H_1'F_t)' + H_1'\frac{1}{T_1^2}\sum_{t=1}^{T_1}F_t(\hat{F}_t - H_1'F_t)' \\ &\quad + \left[\frac{1}{T_1^2}\sum_{t=1}^{T_1}(\hat{F}_t - H_1'F_t)F_t'\right]H_1 + \frac{1}{T_1}\sum_{s=1}^{T_1}H_1'F_tF_t'H_1. \end{aligned} \tag{B.1}$$

The first term on the right-hand-side of the above equation is $O_p(T_1^{-1}\delta_{2,NT_1}^{-2})$ by Lemma B.1 (i). By noting that $H_1 = O(1)$ and H_1 is time-invariant, we see that the second and the third terms on the right-hand-side of the above equation are both $O_p(T_1^{-1}\delta_{2,NT_1}^{-1})$ by Lemma B.1 (iii). Hence, we get

$$\frac{1}{T_1^2}\sum_{t=1}^{T_1}\hat{F}_t\hat{F}_t' = \frac{1}{T_1^2}\sum_{s=1}^{T_1}H_1'F_tF_t'H_1 + O_p(T_1^{-1}\delta_{2,NT_1}^{-1}).$$

This completes the proof of Lemma B.2 (i).

Proof of Lemma B.2 (ii): Using the notation of (A.1) in Bai (2004, page 164), we have

$$\hat{F}_t - H_1' F_t = T^{-2}\sum_{s=1}^{T}\tilde{F}_s\gamma_N(s,t) + T^{-2}\sum_{s=1}^{T}\tilde{F}_s\zeta_{st} + T^{-2}\sum_{s=1}^{T}\tilde{F}_s\eta_{s,t} + T^{-2}\sum_{s=1}^{T}\tilde{F}_s\xi_{s,t}, \tag{B.2}$$

where $\gamma_N(s,t) = N_{co}^{-1}\sum_{i=2}^{N}E(e_{is}e_{it})$, $\zeta_{st} = N_{co}^{-1}\sum_{i=2}^{N}e_{is}e_{it} - \gamma_N(s,t)$, $\eta_{st} = F_s'\Lambda e_t/N_{co}$, $\xi_{st} = F_t'\Lambda' e_s/N_{co}$, $(e_t)_{N_{co}\times 1} = (e_{2t}, ..., e_{Nt})'$, Λ is the $r \times N_{co}$ factor loading matrix for the control group.

Hence

$$\begin{aligned} A_{1T} &\equiv \frac{1}{T_2}\sum_{t=T_1+1}^{T}(\hat{F}_t - H_1'F_t) \\ &= \frac{1}{T_2T^2}\sum_{t=T_1+1}^{T}\sum_{s=1}^{T}\tilde{F}_s\gamma_N(s,t) + \frac{1}{T_2T^2}\sum_{t=T_1+1}^{T}\sum_{s=1}^{T}\tilde{F}_s\zeta_{st} \\ &\quad + \frac{1}{T_2T^2}\sum_{t=T_1+1}^{T}\sum_{s=1}^{T}\tilde{F}_s\eta_{s,t} + \frac{1}{T_2T^2}\sum_{t=T_1+1}^{T}\sum_{s=1}^{T}\tilde{F}_s\xi_{s,t} \\ &= I + II + III + IV. \end{aligned}$$

By parts (a), (b) and (d) of Lemma B.2 in Bai (2004, page 167), we know that $I = O_p(T^{-3/2})$, $II = O_p((NT)^{-1/2})$ and $IV = O_p((NT)^{-1/2})$. For III, we have

$$\begin{aligned} \frac{1}{T^2}\sum_{s=1}^{T}\tilde{F}_s\eta_{st} &= H_2'\frac{1}{T^2}\sum_{s=1}^{T}F_s\eta_{st} + \frac{1}{T^2}\sum_{s=1}^{T}(\tilde{F}_s - H_2'F_s)\eta_{st} \\ &= H_2'\frac{1}{T^2}\sum_{s=1}^{T}F_s\eta_{st} + O_p((NT)^{-1/2}\delta_{2,NT}^{-1}), \end{aligned} \tag{B.3}$$

where $T^{-2}\sum_{s=1}^{T}(\tilde{F}_s - H_2'F_s)\eta_{st} = O_p((NT)^{-1/2}\delta_{2,NT}^{-1})$ was proved in Bai (2004, page 169, line 3).

Substituting (B.3) into III gives

$$\begin{aligned} III &= H_2'\frac{1}{T^2T_2}\sum_{t=T_1+1}^{T}\sum_{s=1}^{T}F_s\eta_{st} + o_p((NT)^{-1/2}) \\ &= H_2'\left(T^{-2}\sum_{s=1}^{T}F_sF_s'\right)\left(\frac{1}{NT_2}\sum_{t=T_1+1}^{T}\sum_{i=2}^{N}\lambda_i e_{it}\right) + o_p((NT)^{-1/2}) \\ &= O_p((NT_2)^{-1/2}) \end{aligned} \tag{B.4}$$

because $T^{-2}\sum_{s=1}^{T} F_s F_s' = O_p(1)$, and the second moment of $D \equiv \frac{1}{NT_2}\sum_{t=T_1+1}^{T}\sum_{i=2}^{N}\lambda_i e_{it}$ has an order $O((NT_2)^{-1})$. To see this. We evaluate the second moment of D:

$$\begin{aligned} E[||D||^2] &= E[tr(DD')] \\ &\leq \frac{1}{N^2T_2^2}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1}^{T}\sum_{i=2}^{N}\sum_{j=2}^{N}|tr[(\lambda_i\lambda_j')]|\;|E(e_{it}e_{js})| \\ &\leq \frac{C}{N^2T_2^2}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1}^{T}\sum_{i=2}^{N}\sum_{j=2}^{N}|\tau_{ij,ts}| \\ &= O((NT_2)^{-1}), \end{aligned} \tag{B.5}$$

because $|tr[(\lambda_i\lambda_j')]| \leq C$ and $\sum_{t=T_1+1}^{T}\sum_{s=T_1+1}^{T}\sum_{i=2}^{N}\sum_{j=2}^{N}|\tau_{ij,ts}| = O(NT_2)$, where $\tau_{ij,ts} = E(e_{it}e_{js})$. Now, (B.5) implies that $D = O_p((NT_2)^{-1/2})$.

Lemma B.3 *In additional to the assumptions made in Bai (2004), we also assume that* $\sum_{i=2}^{N}\sum_{j=2}^{N}\sum_{t=1}^{T_1}\sum_{s=1}^{T_1}|\tau_{0,ij,ts}| = O(NT_1)$, *where* $\tau_{0,ij,ts} = E(e_{1t}e_{it}e_{1s}e_{js})$. *Then we have* $A_{2T} = \frac{1}{T_1}\sum_{t=1}^{T_1}(\hat{F}_t - H_1F_t)e_{1t} = O_p((NT_1)^{-1/2})$.

Proof: Following the notation defined below (B.2), we have

$$\hat{F}_t - H_1'F_t = T^{-2}\sum_{s=1}^{T}\tilde{F}_s\gamma_N(s,t) + T^{-2}\sum_{s=1}^{T}\tilde{F}_s\zeta_{st} + T^{-2}\sum_{s=1}^{T}\tilde{F}_s\eta_{s,t} + T^{-2}\sum_{s=1}^{T}\tilde{F}_s\xi_{s,t},$$

where $\gamma_N(s,t) = N_{co}^{-1}\sum_{i=2}^{N}E(e_{is}e_{it})$, $\zeta_{st} = N_{co}^{-1}\sum_{i=2}^{N}e_{is}e_{it} - \gamma_N(s,t)$, $\eta_{st} = F_s'\Lambda e_t/N_{co}$, $\xi_{st} = F_t'\Lambda' e_s/N_{co}$.

Hence

$$\begin{aligned} A_{2T} &\equiv \frac{1}{T_1}\sum_{t=1}^{T_1}(\hat{F}_t - H_1'F_t)e_{1t} \\ &= \frac{1}{T_1T^2}\sum_{t=1}^{T_1}\sum_{s=1}^{T}\tilde{F}_s\gamma_N(s,t)e_{1t} + \frac{1}{T_1T^2}\sum_{t=1}^{T_1}\sum_{s=1}^{T}\tilde{F}_s\zeta_{st}e_{1t} \\ &\quad + \frac{1}{T_1T^2}\sum_{t=1}^{T_1}\sum_{s=1}^{T}\tilde{F}_s\eta_{s,t}e_{1t} + \frac{1}{T_1T^2}\sum_{t=1}^{T_1}\sum_{s=1}^{T}\tilde{F}_s\xi_{s,t}e_{1t} \\ &= I_1 + II_1 + III_1 + IV_1. \end{aligned}$$

Similar to the proof of Lemma B.2, one can show that $I_1 = O_p(I) = O_p(T^{-3/2})$, $II_1 = O_p(II) = O_p((NT)^{-1/2})$, $III_1 = O_p(III) = O_p((NT_1)^{-1/2})$ and $IV_1 = O_p(IV) =$

$O_p((NT)^{-1/2})$, where I, II, III and IV are defined in the proof of Lemma B.2. Below we will only prove for the term III_1.

For III_1, first note that

$$\begin{aligned} \frac{1}{T^2}\sum_{s=1}^{T}\tilde{F}_s\eta_{st}e_{1t} &= H_2'\frac{1}{T^2}\sum_{s=1}^{T}F_s\eta_{st}e_{1t}+\frac{1}{T^2}\sum_{s=1}^{T}(\tilde{F}_s-H_2'F_s)\eta_{st}e_{1t} \\ &= H_2'\frac{1}{T^2}\sum_{s=1}^{T}F_s\eta_{st}e_{1t}+O_p((NT)^{-1/2}\delta_{2,NT}^{-1}), \qquad \text{(B.6)} \end{aligned}$$

where $T^{-2}\sum_{s=1}^{T}(\tilde{F}_s-H_2'F_s)\eta_{st}e_{1t}=O_p((NT)^{-1/2}\delta_{2,NT}^{-1})$, which can be proved similar to Bai (2004, page 169, the third line).

Substituting (B.6) into III_1 gives

$$\begin{aligned} III_1 &= H_2'\frac{1}{T^2T_1}\sum_{t=1}^{T_1}\sum_{s=1}^{T}F_s\eta_{st}e_{1t}+o_p((NT)^{-1/2}) \\ &= H_2'\left(T^{-2}\sum_{s=1}^{T}F_sF_s'\right)\left(\frac{1}{NT_1}\sum_{t=1}^{T_1}\sum_{i=2}^{N}\lambda_ie_{it}e_{1t}\right)+o_p((NT)^{-1/2}) \\ &= O_p((NT_1)^{-1/2}) \qquad \text{(B.7)} \end{aligned}$$

because $T^{-2}\sum_{s=1}^{T}F_sF_s'=O_p(1)$ and $D_1\equiv\frac{1}{NT_1}\sum_{t=1}^{T_1}\sum_{i=2}^{N}\lambda_ie_{it}e_{1t}=O_p((NT_1)^{-1/2})$ since the second moment of D_1 has an order $O((NT_1)^{-1})$. To see this we evaluate the second moment of D_1:

$$\begin{aligned} E[||D_1||^2] &= E[tr(D_1D_1')] \\ &\leq \frac{1}{N^2T_1^2}\sum_{t=1}^{T_1}\sum_{s=1}^{T_1}\sum_{i=1}^{N}\sum_{j=1}^{N}|tr[(\lambda_i\lambda_j')]|\,|E(e_{it}e_{js}e_{1t}e_{1s})| \\ &\leq \frac{C}{N^2T_1^2}\sum_{t=1}^{T_1}\sum_{s=1}^{T_1}\sum_{i=2}^{N}\sum_{j=2}^{N}|\tau_{0,ij,ts}| \\ &= O((NT_1)^{-1}), \qquad \text{(B.8)} \end{aligned}$$

because $|tr[(\lambda_i\lambda_j')]|\leq C$ and $\sum_{t=1}^{T_1}\sum_{s=1}^{T_1}\sum_{i=2}^{N}\sum_{j=2}^{N}\tau_{0,ij,ts}=O(NT_1)$. Equation (B.8) implies that $D_1=O_p((NT_1)^{-1/2})$.

Lemma B.4 *Define* $\mathcal{B}_1=\frac{1}{T_2}\sum_{t=T_1+1}^{T}\left\{F_t-\left[\sum_{t=1}^{T_1}F_t\hat{F}_t'\right]\left[\sum_{t=1}^{T_1}\hat{F}_t\hat{F}_t'\right]^{-1}\hat{F}_t\right\}$.

Then $\mathcal{B}_1=o_p(T_2^{-1/2})$.

Proof: By Lemma B.2 (i) we know that

$$\frac{1}{T_1^2}\sum_{t=1}^{T_1}\hat{F}_t\hat{F}_t' = \frac{1}{T_1^2}\sum_{t=1}^{T_1}H_1'F_tF_t'H_1 + O_p(T_1^{-1}\delta_{2,NT_1}^{-1}). \tag{B.9}$$

Also, by Lemma B.1 (ii) we have that

$$\frac{1}{T_1^2}\sum_{t=1}^{T_1}F_t\hat{F}_t' = \frac{1}{T_1^2}\sum_{t=1}^{T_1}F_tF_t'H_1 + O_p(T_1^{-1}\delta_{2,NT_1}^{-1}). \tag{B.10}$$

Substituting (B.9) and (B.10) into the expression of $\mathcal{B}_1$, we get

$$\begin{aligned}\mathcal{B}_1 &= \frac{1}{T_2}\sum_{t=T_1+1}^{T}\Big\{F_t - \big[\,(H_1')^{-1} + O_p(T_1^{-1}\delta_{2,NT_1}^{-1})\,\big]\,\hat{F}_t\Big\} \\ &= O_p((NT_2)^{-1/2}) + O_p(T_1^{-1/2}\delta_{2,NT_2}^{-1}) = o_p(T_2^{-1/2}), \qquad \text{(B.11)}\end{aligned}$$

where the last equality follows from $T_2^{-1}\sum_{t=T_1+1}^{T}\Big(F_t - (H_1')^{-1}\hat{F}_t\Big) = (H_1')^{-1}T_2^{-1}\sum_{t=T_1+1}^{T}\Big(H_1'F_t - \hat{F}_t\Big) = O_p((NT_2)^{-1})$ by Lemma B.2 (ii), $T_2^{-1}\sum_{t=T_1+1}^{T}\hat{F}_t/\sqrt{T} = O_p(1)$ and $T_1/T = O(1)$.

Lemma B.5 *Define*

$$\begin{aligned}\mathcal{B}_2 &= \Big[\sum_{t=1}^{T_1}e_{1t}\hat{F}_t'\Big]\Big[\sum_{t=1}^{T_1}\hat{F}_t\hat{F}_t'\Big]^{-1}\Big[\frac{1}{T_2}\sum_{t=T_1+1}^{T}\hat{F}_t\Big] \\ &\equiv \frac{\sqrt{T}}{T_1}\Big[\frac{1}{T_1}\sum_{t=1}^{T_1}e_{1t}\hat{F}_t'\Big]\Big[\frac{1}{T_1^2}\sum_{t=1}^{T_1}\hat{F}_t\hat{F}_t'\Big]^{-1}\Big[\frac{1}{T_2\sqrt{T}}\sum_{t=T_1+1}^{T}\hat{F}_t\Big].\end{aligned}$$

Then

$$\mathcal{B}_2 = \frac{\sqrt{T}}{T_1}\Big[\frac{1}{T_1}\sum_{t=1}^{T_1}e_{1t}F_t'\Big]\Big[\frac{1}{T_1^2}\sum_{t=1}^{T_1}F_tF_t'\Big]^{-1}\Big[\frac{1}{T_2\sqrt{T}}\sum_{t=T_1+1}^{T}F_t\Big] + o_p(T_2^{-1/2}).$$

Proof: By adding/subtracting terms, we get

$$\begin{aligned}\frac{1}{T_1}\sum_{t=1}^{T_1}e_{1t}\hat{F}_t' &= \frac{1}{T_1}\sum_{t=1}^{T_1}e_{1t}(\hat{F}_t - H_1'F_t)' + \frac{1}{T_1}\sum_{t=1}^{T_1}e_{1t}F_t'H_1 \\ &= \frac{1}{T_1}\sum_{t=1}^{T_1}e_{1t}F_t'H_1 + O_p((NT_1)^{-1/2}) \qquad \text{(B.12)}\end{aligned}$$

by Lemma B.3.

By Lemma B.2 (ii), we know that

$$\frac{1}{T_2\sqrt{T}}\sum_{t=T_1+1}^{T}(\hat{F}_t - H_1'F_t) = O_p(T^{-1/2}(NT_2)^{-1/2}). \tag{B.13}$$

Substituting (B.9), (B.12) and (B.13) into $\mathcal{B}_2$, also using the result of Lemma B.2 (ii) and noting that $T_2 = O(T_1)$, we obtain

$$\mathcal{B}_2 = \frac{\sqrt{T}}{T_1}\frac{1}{T_1}\sum_{t=1}^{T_1} e_{1t}F_t'\left[\frac{1}{T_1^2}\sum_{t=1}^{T_1}F_tF_t'\right]^{-1}\left[\frac{1}{T_2\sqrt{T}}\sum_{t=T_1+1}^{T}F_t\right] + O_p(T_1^{-1}\delta_{2,NT_1}^{-1}).$$

This completes the proof of Lemma B.5.

Lemma B.6 *Let $\phi = \lim_{T_1,T_2\to\infty} T_2/T_1$, then we have (i) $A_1 = O_e(1)$ if $0 < \phi < \infty$, here $A_1 = O_e(1)$ means that $A_1 = O_p(1)$ and $A_1 \neq o_p(1)$; (ii) $A_1 = o_p(1)$ if $\phi = 0$.*

Proof of (i): We first consider the case that $\phi > 0$. Note $T/T_1 = (T_1+T_2)/T_1 = 1+T_1/T_2 \to 1+\phi$, and $(T_1+1)/T = T_1/T + 1/T \to (1+\phi)^{-1}$, we have

$$\begin{aligned}
A_1 &= -\sqrt{\frac{T_2}{T_1}}\left(\frac{T}{T_1}\right)^{3/2}\left[\frac{1}{T_1}\sum_{t=1}^{T_1}e_{1t}F_t'\right]\Big[\frac{1}{T_1^2}\sum_{t=1}^{T_1}F_tF_t'\Big]^{-1}\Big[\frac{T_1}{T_2T^{3/2}}\sum_{s=T_1+1}^{T}F_s\Big] && \text{(B.14)}\\
&= -\sqrt{\phi}(1+\phi)^{3/2}\int_0^1 B_u(r)dB_e(r)\Big[B_{2,u}+o_p(1)\Big]^{-1}\Big[B_{\phi,u}+o_p(1)\Big] + o_p(1) && \text{(B.15)}\\
&= -\sqrt{\phi}(1+\phi)^{3/2}\Big[\int_0^1 B_u(r)dB_e(r)\Big]B_{2,u}^{-1}B_{\phi,u} + o_p(1), && \text{(B.16)}
\end{aligned}$$

where $B_u(\cdot)$ and $B_e(\cdot)$ denote Browning motions generated by innovations u_t and e_{1t}, respectively, $B_{2,u} = \int_0^1 B_u(r)B_u(r)'dr$, $B_{\phi,u} = \phi^{-1}\int_{1/(1+\phi)}^1 B_u(r)dr$. Here, we assume that the following results hold: $T_1^{-2}\sum_{t=1}^{T_1}F_tF_t' \xrightarrow{d} \int_0^1 B_u(r)B_u(r)'dr$ and $T_1^{-1}\sum_{t=1}^{T_1}F_te_{1t} \xrightarrow{d} \int_0^1 B_u(r)dB_e(r)'dr$. It is obvious that $A_1 = O_e(1)$ since $\phi > 0$ is a finite positive constant.

Proof of (ii). We now consider the case $\phi = 0$. Let $\phi_T = T_2/T_1$. Then using the same arguments that lead to (B.14), we have

$$A_1 = -\sqrt{\phi_T}(1+\phi_T)^{3/2}\Big[\int_0^1 B_u(r)dB_e(r)\Big]B_{2,u}^{-1}B_{\phi_T,u} + o_p(1),$$

where $B_{\phi_T,u} = \frac{T_1}{T_2T^{3/2}}\sum_{s=T_1+1}^{T}F_s$. Because $\phi_T = o(1)$, it suffices to show that $B_{\phi_T,u} = O_p(1)$.

$$\begin{aligned}
B_{\phi_T,u} &= \frac{1}{\phi_T}\int_{1/(1+\phi_T)}^1 B_u(r)dr + o_p(1)\\
&= B_{0,\Phi_T,u} + o_p(1) = O_p(1),
\end{aligned}$$

because $B_{0,\phi_T,u} = \phi_T^{-1} \int_{1/(1+\phi_T)}^1 B_u(r)dr$ has zero mean and a bounded variance:

$$\begin{aligned}
Var(B_{0,\phi_T,u}) &= \phi_T^{-2} \int_{1/(1+\phi_T)}^1 \Big[\int_{1/(1+\phi_T)}^1 E[B_u(r)B_u(s)]dr \Big] ds \\
&= \sigma_u^2 \phi_T^{-2} \int_{1/(1+\phi_T)}^1 \Big[\int_{1/(1+\phi_T)}^1 \min\{r,s\} dr \Big] ds \\
&= \sigma_u^2 \phi_T^{-2} [\phi_T^2 + O(\phi_T^3)] \\
&= \sigma_u^2 + O(\phi_T) = O(1),
\end{aligned}$$

where we used the following calculation that for any $\delta \in (0,1)$, let $D_\delta = \int_\delta^1 \Big[\int_\delta^1 \min\{r,s\}dr \Big] ds$, then we have

$$\begin{aligned}
D_\delta &= \int_\delta^1 \Big[\int_\delta^s r dr + s \int_s^1 dr \Big] ds \\
&= \int_\delta^1 \Big[(1/2)[s^2 - \delta^2] + s(1-s) \Big] ds \\
&= (1/6)[1-\delta^3] - (1/2)\delta^2(1-\delta) + (1/2)[1-\delta^2] - (1/3)[1-\delta^3].
\end{aligned}$$

Replacing δ by $1/(1+\phi_T) = 1 - \phi_T + \phi_T^2 + O(\phi_T^3)$, we obtain

$$D_{1/(1+\phi_T)} = \phi_T^2 + O(\phi_T^3).$$