Causal Inference for Large Dimensional Non-Stationary Panels with Two-Way Endogenous Treatment and Latent Confounders[*]

Junting Duan[†] Markus Pelger[‡] Ruoxuan Xiong[§]

This draft: July, 2024
First draft: March, 2024

Abstract

This paper studies the imputation and inference for large-dimensional non-stationary panel data with missing observations. We propose a novel method, Within-Transform-PCA (wi-PCA), to estimate an approximate latent factor structure and non-stationary two-way fixed effects under general missing patterns. The missing patterns can depend on the latent factor model and two-way fixed effects. Our method combines a novel within-transformation for the estimation of two-way fixed effects with a PCA on within-transformed data. We provide entry-wise inferential theory for the values imputed with wi-PCA. The key application of wi-PCA is the estimation of counterfactuals on causal panels, where we allow for two-way endogenous treatment effects, time trends and general latent confounders. In an empirical study of the liberalization of marijuana, we show that wi-PCA yields more accurate estimates of treatment effects and more credible economic conclusions compared to its two special cases of conventional difference-in-differences and PCA.

Keywords: Factor Analysis, Principal Components, Fixed-Effects, Non-Stationary, Missing Entry, Large-Dimensional Panel Data, Causal Inference, Treatment Effect
JEL classification: C14, C38, C55

[*] We thank seminar and conference participants at Stanford, the Annual Meeting of the American Economic Association and the Stanford Causal Science Conference for helpful comments. Our analyses are partly based on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researchers and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

[†] Stanford University, Department of Management Science & Engineering, Email: duanjt@stanford.edu.

[‡] Stanford University, Department of Management Science & Engineering, and NBER, Email: mpelger@stanford.edu.

[§] Emory University, Department of Quantitative Theory and Methods, Email: ruoxuan.xiong@emory.edu.

1 Introduction

Panel data with large cross-section and time-series dimensions are prevalent in the social sciences and other disciplines. These large-dimensional panels often have missing observations. A growing literature is developing methods for imputing missing observations. A common approach is to assume an approximate latent factor structure. For these large panels, it is possible to leverage the large cross-section and time-series dimensions to obtain accurate imputation and develop the inferential theory for imputed values. A key application for the inference of the imputed values is causal inference, where the unobserved counterfactual outcomes can be modeled as missing values.

Most existing imputation methods based on factor models with entry-wise inferential theory are designed for stationary panels or require restrictive assumptions on the missing patterns.[1] However, in many applications, the panel data are non-stationary. For example, in the evaluation of the tobacco control program in Abadie, Diamond, and Hainmueller (2010), the tobacco consumption across states has a common declining trend. The conventional approach of differencing the time series to deal with common trends is generally not feasible with missing data. A commonly used solution is to control for general time fixed effects, which can capture a common time trend, seasonal fluctuations, or a common break in the mean. Moreover, missing patterns can be complex and depend on unit and time fixed effects. For example, due to a common break in the mean of a panel, specific units can have more missing observations. The inclusion of two-way fixed effects and the dependency of the missing pattern on fixed effects is particularly relevant for applications in causal inference. For example, the public perception of tobacco can vary for states and over time (modeled by two-way fixed effects), which affects both the tobacco consumption (outcome) and the adoption of tobacco control program (treatment).

In this paper, we study the imputation and inference for large panels with an approximate latent factor structure and general two-way fixed effects. Our novel approach is one of the most general models for data generation and for the missing pattern in the factor model literature. The time fixed effects are not restricted and can be non-stationary processes. The general missing pattern can depend on the loadings of the latent factors (i.e. latent confounders) and both dimensions of

[1] This includes, among others, Chen, Fan, Ma, and Yan (2019), Xia and Yuan (2021), Jin, Miao, and Su (2021), Bai and Ng (2021), Cahan, Bai, and Ng (2023), Xiong and Pelger (2023), Duan, Pelger, and Xiong (2023), and Chernozhukov, Hansen, Liao, and Zhu (2023).

the two-way fixed effects.

We develop a novel method, namely Within-Transform principal component analysis (wi-PCA), to estimate a latent factor model with two-way fixed effects from a (non-stationary) panel, and to impute the missing observations. Our wi-PCA is simple to implement. Importantly, it is applicable to many relevant observation patterns, including missing-at-random, block-missing, and mixed-frequency patterns, while allowing the missingness to depend on the model itself. The estimation follows two steps. In the first step, wi-PCA estimates unit and time fixed effects by carefully weighting the time-series and cross-sectional outcomes, respectively, where the weighting scheme depends on the observation pattern. The within-transformation is then applied to the data using the estimated fixed effects. In the second step, wi-PCA estimates the latent factor structure by applying a generalization of PCA to a re-weighted covariance matrix estimated from within-transformed data. Our wi-PCA imputes the missing observations with the plug-in estimator that combines the estimated two-way fixed effects and factor model. Separating the estimation of fixed effect from the latent factor component is crucial for three reasons: (1) it allows for non-stationary time fixed-effects, (2) the missingness can depend in a general way on the two-way fixed effects, and (3) it is generally more efficient than subsuming fixed effects by latent factors.

We show the consistency and asymptotic normality for the estimated factor model and imputed values from wi-PCA under general assumptions on the approximate factor model and observation patterns. Developing the inferential theory for wi-PCA is highly non-trivial for two reasons. First, it is challenging to estimate the fixed effects when the missing patterns depend on their values. Conventional estimators for fixed effects, such as simple averages of cross-sectional or time-series outcomes, are generally inconsistent in these cases. Second, the estimation error of the two-way fixed effects carries over to the estimation error of the latent factor model, and both estimation errors vary with the observation patterns. Hence, deriving the inferential theory requires a careful and comprehensive analysis of all the error terms.

A key application of our model and asymptotic theory is causal inference in panels. The unobserved counterfactual outcome can naturally be formulated as a missing observation problem. We derive test statistics for unit-specific treatment effects in a panel. Given the asymptotic normality of our estimator, we develop a bootstrap procedure to obtain a feasible variance estimator and confidence intervals for the estimated treatment effects. The generality of the missing patterns maps into general observational treatment patterns. Special cases of our estimator relate to commonly used methods in causal inference. With only the two-way fixed effects (and without the latent factor model), our estimator simplifies to a form of difference-in-difference estimator. With only a latent factor model, our estimator can be interpreted as a data-driven method to obtain synthetic controls for simultaneous, staggered, or other treatment adoption patterns, where latent factor loadings serve as unobserved confounders to construct weighted averages of similar units. The combination of both components and the more general treatment patterns pushes the boundary for inference in causal panels.

Our work contributes to three streams of literature: large dimensional factor modeling, missing data imputation for large panels, and causal inference. First, our work builds on the literature on large dimensional factor modeling. There has been substantial progress in this literature since the pioneering work by Bai and Ng (2002) for estimating the number of factors and by Bai (2003) for developing the inferential theory of the estimated latent factor model. The progress includes, but is not restricted to, the estimation of large covariance matrices (Fan, Liao, and Mincheva, 2013), the identification and estimation of more general models, such as models with additive and interactive effects and common slope coefficients (Bai, 2009), models with weak factors and factor loadings (Onatski, 2012; Lettau and Pelger, 2020; Bai and Ng, 2023), models with time-varying factor loadings (Su and Wang, 2017; Pelger and Xiong, 2021b; Urga and Wang, 2022), and the estimation of interpretable latent factors (Pelger and Xiong, 2021a).

Our work is the most closely related to the recent work on the estimation and inference of factor models from large panels with missing observations. The wi-PCA accommodates general observation patterns, complementing the methods developed by Chen, Fan, Ma, and Yan (2019), Xia and Yuan (2021), Jin, Miao, and Su (2021), and Chernozhukov, Hansen, Liao, and Zhu (2023) for missing-at-random,[2] Xu (2017), Bai and Ng (2021), Cahan, Bai, and Ng (2023), and Choi and Yuan (2023) for general block-missing patterns,[3] and Ng and Scanlan (2024) for mixed-frequency patterns.[4] Our wi-PCA builds on Xiong and Pelger (2023) and Duan, Pelger, and Xiong (2023) that

[2] Chen, Fan, Ma, and Yan (2019), Xia and Yuan (2021), and Jin, Miao, and Su (2021) assume a homogeneous missing probability, while Chernozhukov, Hansen, Liao, and Zhu (2023) allow for heterogeneous missing probabilities in either the cross-sectional or time dimension, but not both.

[3] General block-missing patterns require blocks of fully observed entries, which include the missing patterns implied by simultaneous and staggered treatment adoptions of causal panels.

[4] Chen, Fan, Ma, and Yan (2019), Athey, Bayati, Doudchenko, Imbens, and Khosravi (2021), Xia and Yuan (2021), Chernozhukov, Hansen, Liao, and Zhu (2023), and Choi and Yuan (2023) analyze the low-rank matrix estimator

already allow for general observation patterns, but makes two challenging generalizations. First, it allows for non-stationary two-way fixed effects in the outcomes, and, second, the missing patterns can depend on the time trends in addition to the cross-sectional (latent) characteristics of units. Hence, this is a strict generalization of the data generating process and the missing patterns. These two generalizations require new arguments to prove the theoretical results. However, these two generalizations make wi-PCA more broadly applicable, and are particularly important for causal inference in panels.

We conduct extensive simulations to demonstrate the superior performance of wi-PCA over benchmark approaches. We identify two cases, where an imputation that uses only a latent factor model, is particularly problematic. First, we show when the data generating process has non-stationary fixed effects, PCA on missing data can fail, while wi-PCA provides accurate out-of-sample imputed values. Hence, a latent factor model cannot compensate for fixed effects, and we need our more general model. Second, if the missingness itself depends on the fixed effects (even when those are stationary), wi-PCA also strongly outperforms PCA type approaches. Lastly, even for stationary panels when the fixed effects are subsumed by latent factors, it is more efficient to directly estimate the fixed effects.

In our empirical study, we use wi-PCA for an application in causal inference: the policy effect of legalizing marijuana on beer sales studied in Li and Sonnier (2023). We demonstrate that our general model is more accurate and can lead to fundamentally different economic conclusions compared to using special cases of it. Common approaches used in the literature are a difference-in-difference estimator (which is a special case of wi-PCA without the latent factor model) or a form of synthetic controls based only on a latent factor model (which is a special case of wi-PCA without the two-way fixed effects). We show that omitting the fixed effects or latent factors results in spurious significance of treatment effects as well as too large treatment effects. Our results illustrate the importance of our general estimator for estimating more credible policy effects and making more reliable policy recommendations.

from a convex optimization problem using a nuclear norm regularization (Mazumder, Hastie, and Tibshirani, 2010; Negahban and Wainwright, 2011, 2012). Athey, Bayati, Doudchenko, Imbens, and Khosravi (2021) also demonstrate that it improves the accuracy to estimate two-way fixed effects separately from a low rank component. Jin, Miao, and Su (2021) provide the inferential theory for the estimated factor model with the expectation-maximization (EM) algorithm (Stock and Watson, 2002; Bańbura and Modugno, 2014). Bai and Ng (2021) and Cahan, Bai, and Ng (2023) provide the inferential theory for the factor-based imputed values.

The rest of the paper is organized as follows. Section 2 introduces the model setup, while Section 3 proposes the wi-PCA estimator. Sections 4 formalizes the assumptions on the approximate factor model with two-way fixed effects, and provides the asymptotic results for our estimator. Section 5 shows a key application of our estimator to estimate the treatment effects in causal inference and provides feasible bootstrap estimators for the asymptotic variance. Section 6 discusses the extension of including observables in our model. Section 7 demonstrates in simulations the good performance of wi-PCA compared to the benchmarks and Section 8 revisits a case study in causal inference. Section 9 concludes the paper. The Internet Appendix collects the proofs and additional empirical and simulation results.

2 Model Setup

2.1 Model

Assume we partially observe a large dimensional panel $Y \in \mathbb{R}^{N \times T}$, where both N and T are large. We aim to impute the missing observations in Y and provide entrywise inferential theory for the imputed values. We work under the assumption that Y can be well approximated by two-way fixed effects and k common latent factors:[5]

$$Y_{it} = \mu + \alpha_i + \xi_t + \Lambda_i^\top F_t + \epsilon_{it}\,, \qquad i = 1, \cdots, N,\ t = 1, \cdots, T\,. \tag{1}$$

Here, Y_{it} denotes the outcome of unit i at time t, μ is the grand mean, α_i is the fixed effect of unit i, ξ_t is the time fixed effect at time t, F_t is a k-dimensional vector of latent factors at time t, Λ_i is a k-dimensional vector of unit i's loadings, and lastly ϵ_{it} is the unit i's idiosyncratic error at time t. We define the common component as the combination of the two-way fixed effects and the latent factor model:

$$C_{it} = \mu + \alpha_i + \xi_t + \Lambda_i^\top F_t\,.$$

The goal of our paper is to estimate C_{it} for all i and t, and to use the estimated C_{it} to impute Y_{it} when it is unobserved.

[5] We assume k is known. In practice, k can be determined by cross-validation arguments as discussed in more detail in Section 6.

Note that the fixed effects can be included as additional latent factors in the factor model. More specifically, the fixed effects $\mu + \alpha_i + \xi_t$ can be written as a two-factor model with $(1, \xi_t)$ as factors and $(\mu + \alpha_i, 1)$ as loadings. Combining $\mu + \alpha_i + \xi_t$ with $\Lambda_i^\top F_t$, we can reformulate model (1) as an approximate factor model with $k + 2$ common factors

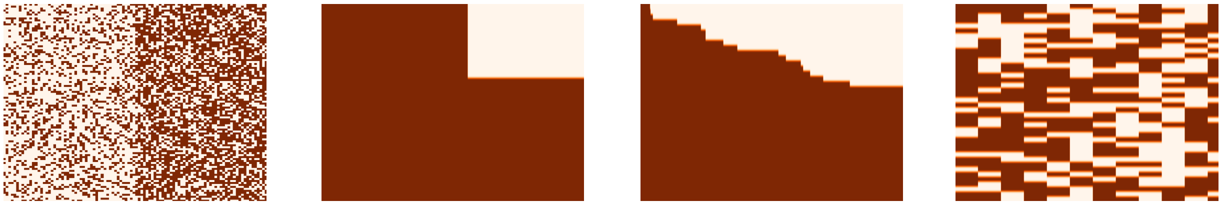
$$Y_{it} = \Lambda_i^{\mathrm{FE}^\top} F_t^{\mathrm{FE}} + \epsilon_{it} \,,$$

where $\Lambda_i^{\mathrm{FE}^\top} = (\mu + \alpha_i, 1, \Lambda_i^\top)$ and $F_t^{\mathrm{FE}^\top} = (1, \xi_t, F_t^\top)\,.$

However, separating the fixed-effects from the latent factor structure is important for three reasons. First, it allows for data with common non-stationary time patterns as the time-series fixed effects in our model can be almost arbitrary. These time patterns can be a common unit-root process, common seasonality or common structural breaks. Without separating the time-series fixed effects for non-stationarity data, the limit of the time-series average of $F_t^{\mathrm{FE}} F_t^{\mathrm{FE}^\top}$ does not exist for T going to infinity, invalidating the commonly made assumptions for consistent estimation of latent factor models. For fully observed data Y, certain non-stationarities can be removed by transformations, for example, by taking the first differences between consecutive periods to remove a common unit root process. However, with missing data, for example missing-at-random, such transformations might not be feasible. We address this problem by separating ξ_t from F_t, and using a different estimation approach that can allow for arbitrary non-stationary time trends in ξ_t.

Second, model (1) allows for broader empirical applications where the observation patterns (and treatment adoptions) depend on both (endogenous) cross-sectional and temporal information. For example, in our empirical study of the effect of legalization of marijuana in different U.S. states, the adoption might become more likely over time due to changes in the public opinion or federal law. At the same time, public awareness might be heterogeneous among states. Hence, the adoption pattern can depend endogenously on state and time fixed effects. However, when the model is written as an interactive latent structure $\Lambda_i^{\mathrm{FE}^\top} F_t^{\mathrm{FE}}$ and is estimated by applying a version of PCA, Xiong and Pelger (2023) show that for identification reasons, the observation patterns can only depend on either Λ_i^{FE} or F_t^{FE}, but not on both. By separating α_i and ξ_t from $\Lambda_i^\top F_t$ and treating them differently in the estimation procedure, we can allow for more general observation patterns that can depend on both α_i and ξ_t, and either Λ_i or F_t, and hence can be endogenous in both dimensions.

Figure 1: Examples of observation patterns



(a) Missing-at-random (b) Simultaneous adoption (c) Staggered adoption (d) Switchback

These figures show examples of important patterns of missing observations. The shaded entries indicate observed entries, whereas the unshaded entries indicate missing entries. Time is on the horizontal axis.

Lastly, estimating the fixed effects separately from the latent factors is more efficient. Even for stationary panels when fixed effects are subsumed by latent factors, the direct estimation is more precise as less parameters have to be estimated.

2.2 Observation Patterns

We allow for the most general assumptions on missing patterns in this literature. In order to provide some intuition, Figure 1 illustrates important examples. The first example represents missing-at-random, which is conceptually the simplest case. In this case, the probability of observing an entry does not depend on the outcome variable or whether other entries are observed. The second and third cases show observation patterns that commonly occur in causal inference applications. For both, the simultaneous adoption and staggered adoption settings, a unit stays treated for all periods after adopting the treatment, that is, missingness depends on prior missingness. The last pattern shows the observation pattern of the global control panel when the treatment can be turned on and off. This pattern occurs in "switchback experiments" that are commonly used in digital platforms to test changes to algorithms and products (Xiong, Chin, and Taylor, 2024).

The observation pattern in Y is captured by the random variables $W \in \{1, 0\}^{N \times T}$, where $W_{it} = 1$ denotes that Y_{it} is observed while $W_{it} = 0$ implies missing values. Assumption 1 formally states the assumptions on the missing pattern. The probability of observing entry Y_{it} can depend on a very general conditioning set denoted by I_{it}. We allow this conditioning set I_{it} to be $\{\alpha_1, \cdots, \alpha_N, \Lambda_1, \cdots, \Lambda_N, \xi_1, \cdots, \xi_t\}$, that is, the observation pattern can depend on all unit and time fixed effects and all loadings. A special case of the conditioning set is $\{\alpha_i, \Lambda_i, \xi_1, \cdots, \xi_t\}$, where the probability that Y_{it} is observed depends on unit i's unit fixed effect and factor loadings, as well as time fixed effects up to time t. As discussed in Xiong and Pelger (2023), the assumptions on the factors and loadings can be switched, and hence the observation pattern can depend on the factors instead of the loadings, but not on both. The prior literature only allows general dependencies on the endogeneity in either the unit or time dimension, but through the fixed effects, we allow for complex dependencies in both dimensions.

Assumption 1 (Observation Patterns). *The conditioning set I_{it} can be any subset (including the complete set) of $\{\alpha_1, \cdots, \alpha_N, \Lambda_1, \cdots, \Lambda_N, \xi_1, \cdots, \xi_t\}$.*[6]

1. *The observation pattern W is independent of the factors F and idiosyncratic errors ϵ.*
2. *There exists a positive constant η such that, for any i and t, the conditional observation probability $p_{it} := \mathbb{P}(W_{it} = 1 \mid I_{it}) \in [\eta, 1]$. For any W, there exists a positive constant q such that $N^{-1} \sum_{i=1}^{N} W_{it} \geq q$ and $T^{-1} \sum_{t=1}^{T} W_{it} \geq q$ for any i and t.*
3. *For any i and j satisfying $i \neq j$ and any t and s, W_{it} and W_{js} are independent conditional on I_{it} and I_{js}.*

Assumption 1.1 requires W to be independent of F and ϵ, but allows for arbitrary dependency on α, Λ and ξ. As we allow W to depend on both α and ξ, our setup is more general than prior literature (such as Xiong and Pelger (2023)) which allows for the dependence on either the endogenous cross-sectional or temporal information, but not both.

Assumption 1.2 assumes that each entry has a non-zero probability of being observed. The non-zero probability avoids the extrapolation problem when using information in the observed entries to impute missing entries. Essentially, the partially observed data needs to be sufficient to learn the underlying latent structure. The assumption implies that for N and T going to infinity, the fraction of observed entries for each time period t and each unit i are proportional to N and T, respectively. This assumption can be relaxed at the cost of more complex notation, but without changing the conceptual insights.

Assumption 1.3 allows for arbitrary time-series dependencies in the observation patterns, and hence includes complex staggered adoption as well as switchback designs. The assumption only assumes conditional independence of the observation pattern across units, but does not restrict

[6] The conditioning set I_{it} can also include observables, which we discuss in Section 6. Here we focus on the conceptual challenge of incorporating latent components within this set.

the time dependency for each unit. Specifically, whether Y_{it} is observed can depend on whether $Y_{i1}, \cdots, Y_{i,t-1}$ are observed. This, therefore, accommodates all the examples shown in Figure 1. The assumption of conditional independence across units is needed for the purpose of identification. This results in an asymmetry of the assumptions on observation patterns between the cross-sectional and time-series dimensions.

3 Within-Transform-PCA (wi-PCA)

In this section, we introduce our novel estimator, namely Within-Transform-PCA or, in short, wi-PCA, to estimate each term and the common component C_{it} in model (1). Our wi-PCA approach consists of two steps. First, we estimate the grand mean and two-way fixed effects with a within transformation. Second, as the name suggests, we estimate the latent loadings and factors by applying a modification of PCA to the "within transformed" data, where fixed effects are removed using the estimates from the first step. The estimation procedure in each step is carefully designed to account for the complex dependency between the observation pattern and the data generating process. This careful design is crucial for estimating C_{it} and imputing missing entries consistently and without bias.

3.1 Estimation of Grand Mean and Fixed Effects

In the first step, wi-PCA estimates the grand mean μ and two-way fixed effects α_i and ξ_t for all i and t. The estimator takes a weighted average of the observed entries in Y across different dimensions. The fundamental challenge of fixed effect estimation is how to properly weight the entries in these averages. The estimator takes the following form:

$$\begin{aligned} \tilde{\mu} &= \sum_{i=1}^{N} \sum_{t=1}^{T} M_{it}^{\mu} Y_{it} \,, \\ \tilde{\xi}_t &= \sum_{i=1}^{N} M_{it}^{\xi} Y_{it} - \tilde{\mu} \,, \\ \tilde{\alpha}_i &= \sum_{t=1}^{T} M_{it}^{\alpha} (Y_{it} - \tilde{\xi}_t) - \tilde{\mu} \,, \end{aligned} \tag{2}$$

where M_{it}^{μ}, M_{it}^{ξ} and M_{it}^{α} are the weights of unit i at time t in the estimation of μ, ξ_t, and α_i, respectively.

Note that the weighted averages for the time and unit fixed effects take a different and asymmetric form. Specifically, we have to remove ξ_t in the estimation of α_i, but not vice versa. The cancellation of ξ_t in $\tilde{\alpha}_i$ is necessary because we allow the observation patterns to be arbitrarily dependent in the time dimension, such as simultaneous or staggered adoption patterns. Without removing ξ_t, $\tilde{\alpha}_i$ could be biased. We show an example in Appendix A. In contrast, as the observation patterns in the cross-sectional dimension are conditionally independent, α_i does not have to be removed in $\tilde{\xi}_t$, as long as M_{it}^{ξ} is chosen appropriately. We will discuss this aspect in more detail below.

The most critical component of the estimators is the choice of weights M_{it}^{μ}, M_{it}^{ξ} and M_{it}^{α}. Our objective is to choose the weights such that $\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t$ is a consistent estimator of $\mu + \alpha_i + \xi_t$. As we are primarily interested in estimating the common components, obtaining the consistency for the sum of grand mean and two-way fixed effects is sufficient. When all entries in Y are observed, the natural choice is

$$M_{it}^{\mu} = \frac{1}{NT}, \qquad M_{it}^{\xi} = \frac{1}{N}, \qquad M_{it}^{\alpha} = \frac{1}{T}.$$

However, when Y has missing entries and the missing patterns depend on α_i and ξ_t themselves, the weights need to be chosen carefully. Naturally, the weights should satisfy two requirements. First, as we can only use the observed entries in the estimation, the weights for the missing entries have to be set to zero. Second, a necessary requirement for the weights is that the sum $\mu + \alpha_i + \xi_t$ can be consistently estimated. This imposes further conditions on the weights. We can identify these conditions by analyzing the estimation error of $\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t$:

$$\begin{aligned} &\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t - (\mu + \alpha_i + \xi_t) \\ =&\sum_{s=1}^{T} M_{is}^{\alpha} \left(\Lambda_i^\top F_s + \epsilon_{is} \right) - \sum_{s=1}^{T} M_{is}^{\alpha} \sum_{j=1}^{N} M_{js}^{\xi} \left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js} \right) + \sum_{j=1}^{N} M_{jt}^{\xi} \left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right). \end{aligned}$$

Note that the weight M_{it}^{μ} is canceled out in the estimation error of $\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t$. Therefore, as long as we are only concerned about the estimation of $\mu + \alpha_i + \xi_t$, there are no constraints on M_{it}^{μ} and the estimation of α_i and ξ_t subsumes the estimation of μ without further identification assumptions. We impose the normalization that the weights sum up to one, that is, $\sum_{i=1}^{N}\sum_{t=1}^{T} M_{it}^{\mu} = 1, \sum_{i=1}^{N} M_{it}^{\xi} = 1$, and $\sum_{t=1}^{T} M_{it}^{\alpha} = 1$. Without loss of generality, we use

$$M_{it}^{\mu} = \frac{W_{it}}{\sum_{s=1}^{T}\sum_{j=1}^{N} W_{js}} \, ,$$

that is, we estimate the grand mean μ by averaging Y_{it} over all the observed entries.

Next, we provide feasible choices of M_{it}^{α} and M_{it}^{ξ}, which ensure that the estimation error of $\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t$ is $o_p(1)$. First, we suggest to set M_{it}^{α} as

$$M_{it}^{\alpha} = \frac{W_{it}}{\sum_{s=1}^{T} W_{is}} \, . \tag{3}$$

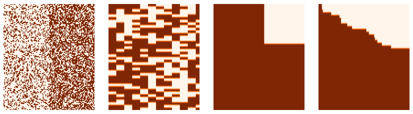
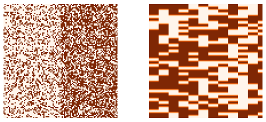
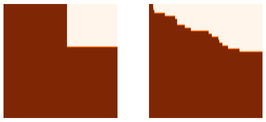
This choice of weights simply estimates the unit fixed effect α_i by averaging $Y_{it} - \tilde{\xi}_t - \tilde{\mu}$ over the time periods t where unit i is observed. Importantly, this average is taken after removing the time fixed effects. As F_t and ϵ_{it} can be normalized to have mean zero in the presence of ξ_t and are assumed to be independent of W, the first term in the decomposition is $o_p(1)$.

The most challenging problem is to specify the weights M_{it}^{ξ} in a way such that α_i and Λ_i can be averaged out. As W can depend on α_i and Λ_i in a complicated way, there does not exist a weight that works for all observation patterns. In other words, the weights M_{it}^{ξ} have to be specified depending on the missing pattern. We provide the weights M_{it}^{ξ} for three important cases, which are summarized in Table 1. These three cases altogether cover the examples of observation patterns shown in Figure 1 with known or unknown observation probabilities. In the first two cases, we leverage the structure of the known or unknown observation probabilities to construct weights M_{it}^{ξ}. In the third case, we leverage the structure in observation pattern (e.g., simultaneous or staggered adoption) to construct weights M_{it}^{ξ}. We provide a detailed discussion of the three cases in the following, and formalize the assumptions for the three cases in Assumption 4 in Section 4.1.

Case 1 – Known Observation Probability

When the observation probabilities $p_{it} = \mathbb{P}(W_{it} = 1 \mid I_{it})$ are known, for example, when Y is

Table 1: Summary of different cases and their corresponding weights M_{it}^{ξ}

	Description	Observation patterns	M_{it}^{ξ}
Case 1	Known p_{it}		$M_{it}^{\xi} = \left(\sum_{j=1}^{N} \frac{W_{jt}}{p_{jt}}\right)^{-1} \frac{W_{it}}{p_{it}}$
Case 2	Short-term dependency in missingness with factor structure		$M_{it}^{\xi} = \left(\sum_{j=1}^{N} \frac{W_{jt}}{W_{j,\cdot}}\right)^{-1} \frac{W_{it}}{W_{i,\cdot}}$
Case 3	Monotone missingness		$M_{it}^{\xi} = N_c^{-1} I(i \in \mathcal{N}_c)$

This table provides three important cases and our solutions for selecting the weights M_{it}^{ξ} for each of them. The first case assumes that the observation probabilities p_{it} are known, and the last two cases allow for unknown observation probabilities. The second case assumes that the probabilities p_{it} can be factorized into a one-factor model as $p_{it} = u_i v_t$ with $u_i, v_t \in [\eta, 1]$, and W_{it} can depend on W_{is} only when time t and s are sufficiently close (i.e., short-term dependency) for any i. The third case assume monotone observation patterns, and we denote by $\mathcal{N}_c$ and N_c the set and number of the fully observed units.

the experimental data in a design-based settings, we can set M_{it}^{ξ} as

$$M_{it}^{\xi} = \left(\sum_{j=1}^{N} \frac{W_{jt}}{p_{jt}}\right)^{-1} \frac{W_{it}}{p_{it}}. \tag{4}$$

The resulting estimator $\tilde{\xi}_t$ in Equation (2) is the Hajek estimator of ξ_t. We propose to use the Hajek estimator in this case because it is generally more efficient than the Horvitz-Thompson estimator (i.e., $M_{it}^{\xi} = N^{-1}W_{it}/p_{it}$). Intuitively, by weighting the observed entries with the inverse observation probability, we adjust for dependency in the observation pattern, and it can be treated similarly to missing-at-random.

Case 2 – Unknown Observation Probability: Short-Term Dependency in Missingness with Factor Structure

For most applications, the observation probabilities p_{it} are unknown. But if we can consistently estimate p_{it}, then we can still use the estimator in Equation (4) with p_{it} replaced by $\hat{p}_{it}$. Here we provide important examples when p_{it} can be consistently estimated.

The simplest example is missing-completely-at-random, that is each entry is observed with the same but unknown probability p, independent of whether other entries are observed and independent of Y. Then the fraction of observed entries $\hat{p} = (NT)^{-1} \sum_{j,s} W_{js}$ is a consistent estimator of p for all i and t. A more general example is that the unknown probability is time-dependent but the same for different units. Then $\hat{p}_t = N^{-1} \sum_j W_{jt}$ is a consistent estimator of p_t for all i. The symmetric example where the probability is unit-dependent but constant in time can be analyzed analogously.

The most challenging case is when the unknown probability varies with both units and time. If we do not impose any structure on p_{it}, then each p_{it} only has one observation W_{it}, and it is not possible to consistently estimate it. To make progress on this problem, we provide sufficient conditions below, under which p_{it} can be consistently estimated.

Specifically, the sufficient conditions consist of two assumptions, one directly on the observation probability p_{it} and another one on the time-series dependency of W. The first assumption is that p_{it} can be factorized into a one-factor model as $p_{it} = u_i v_t$ with $u_i, v_t \in [\eta, 1]$. The second assumption is that for any i and t, W_{it} can depend on W_{is} only when time t and s are sufficiently close (i.e., short-term dependency). Under these two conditions, we can estimate the observation probability up to a scaling constant using the time series average of W_{it}, and construct the weight M^{ξ} as

$$M^{\xi}_{it} = \left(\sum_{j=1}^{N} \frac{W_{jt}}{\bar{W}_{j,\cdot}} \right)^{-1} \frac{W_{it}}{\bar{W}_{i,\cdot}} \, . \tag{5}$$

Under mild assumptions specified in Assumption 4, we can show that this weight yields consistent $\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t$.

The one-factor model assumption in the observation probability seems to be a reasonable approximation for many practical problems. It has been imposed in a number of papers, such as Negahban and Wainwright (2012) and Chen and Bayati (2021), and has been shown to perform well in completing a low-rank matrix. In simulations, we also find that using multiple factors in the observation probability does not improve the imputation accuracy. We note that it is possible to generalize the one-factor model assumption. For example, we can allow for multiple factors at the expense of a more complicated notation. Alternatively, we can split the panel into sub-panels, where the observation probabilities for each sub-panel can be modeled by a different one-factor model. However, as mentioned before, the one-factor model in the observation probability already performs very well in many practical problems.

Case 3 – Unknown Observation Probability: Monotone Missingness

Simultaneous and staggered adoption are important treatment patterns in causal inference. These observation patterns have a monotone structure, that is, once a unit is treated it stays treated, which implies a long-term time dependency in missingness. The previous Case 2 assumes a short-term dependency in the missingness, which would not apply to these patterns. Here we propose weights M^{ξ}_{it} for the case of monotone observation patterns, that is, for patterns where either $W_{i1} \leq W_{i2} \leq \cdots \leq W_{iT}$ or $W_{i1} \geq W_{i2} \geq \cdots \geq W_{iT}$ holds for all i. Our solution can be applied for simultaneous or staggered treatment patterns.

In this case, the time fixed effects are weighted averages over the units that are observed for all times, which correspond to the fully observed block in simultaneous or staggered treatment patterns. Taking averages over the same cross-sectional units allows us to accommodate observation patterns that depend on the time fixed effects. Assumption 1.2 implies that there are at least qN units that are fully observed for all times. We denote by $\mathcal{N}_c$ and N_c the set and number of these units. We only assign non-zero weights to these fully observed units and simply take the cross-sectional average of these units for each time t, resulting in the following weight M^{ξ}_{it}:

$$M^{\xi}_{it} = \begin{cases} N_c^{-1} & \text{if unit } i \text{ is observed for all times } t = 1, \cdots, T \\ 0 & \text{otherwise}\,. \end{cases} \tag{6}$$

Under mild moment conditions stated in Assumption 4, we can show that with this weight $\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t$ is a consistent estimator of the fixed effect component.

The conventional difference-in-difference (DID) is a special case of our estimator. If we only estimate the fixed effect component without a latent factor structure, our estimator simplifies to

$$\begin{aligned} \tilde{C}_{it} =\ & \tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t = \frac{1}{N_c}\sum_{j \in \mathcal{N}_c} Y_{jt} + \left(\sum_{s=1}^{T} W_{is}\right)^{-1} \sum_{s=1}^{T} W_{is}\left(Y_{is} - \frac{1}{N_c}\sum_{j\in\mathcal{N}_c} Y_{js}\right) \\ =\ & \underbrace{\left(\sum_{s=1}^{T} W_{is}\right)^{-1}\sum_{s=1}^{T} W_{is}Y_{is}}_{\text{pretreatment average of unit } i} + \underbrace{\frac{1}{N_c}\sum_{j\in\mathcal{N}_c}\left(Y_{jt} - \left(\sum_{s=1}^{T} W_{is}\right)^{-1}\sum_{s=1}^{T} W_{is}Y_{js}\right)}_{\text{averaged difference among control units}}. \end{aligned}$$

This is equivalent to the DID estimator that uses the control units to estimate the time trend for each treated unit. In this sense, our wi-PCA is a strict generalization of the DID estimator by including latent factors in addition to the DID estimator.

3.2 Estimation of Latent Factor Model

After having estimated the fixed effect component, we estimate the latent factor structure in the second step. As suggested by the name of wi-PCA, we apply PCA to the within-transformed data to estimate the factors and loadings. As we have missing observations, we have to modify PCA to account for the missingness.

In more detail, we first transform the observed panel Y_{it} by subtracting the first-stage estimators, and obtain the within-transformed data $\dot{Y}_{it}$:

$$\dot{Y}_{it} = Y_{it} - \tilde{\mu} - \tilde{\alpha}_i - \tilde{\xi}_t \,.$$

This transformed data has an approximate factor structure

$$\dot{Y}_{it} = \Lambda_i^\top F_t + \dot{\epsilon}_{it}$$

with a new idiosyncratic error term $\dot{\epsilon}_{it} = \epsilon_{it} + (\mu + \alpha_i + \xi_t) - (\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t)$. Since $\dot{Y}_{it}$ shares the same latent factor model and observation pattern as Y_{it}, we can estimate F_t and Λ_i from $\dot{Y}_{it}$. The main challenge is that the new idiosyncratic errors $\dot{\epsilon}_{it}$ include the estimation error from the first step, which then carries over to the estimation of F_t and Λ_i. Hence, the usual assumptions of PCA-based factor model estimation do not directly apply, and we need to carefully account for the estimation error from the first step.

For the special case when $\dot{Y} \in \mathbb{R}^{N \times T}$ is fully observed, we can estimate the loadings Λ with PCA applied to the cross-sectional sample second-moment matrix $\dot{Y}\dot{Y}^\top/T$. The factors are estimated by regressing $\dot{Y}$ on the estimated loadings. The problem is more complicated for partially observed panels and requires to modify the PCA estimator.

For the general case when $\dot{Y}$ has missing observations, we adopt the all-purpose estimator proposed in Xiong and Pelger (2023). Specifically, we first estimate the cross-sectional sample second-moment matrix, denoted by $\tilde{\Sigma}$, from the partially observed panel. Essentially, we use the times when two units are observed together to estimate their covariance. We define $Q_{ij} := \{t : W_{it} = W_{jt} = 1\}$ to be the set of time periods when both units i and j are observed. Then, we estimate the (i, j)-th entry of $\tilde{\Sigma}$ using the time observations in Q_{ij}:

$$\tilde{\Sigma}_{ij} = \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} \dot{Y}_{it} \dot{Y}_{jt} \,.$$

Next, we apply PCA to $\tilde{\Sigma}$ to estimate Λ. Under the standard identification assumption $\tilde{\Lambda}^\top \tilde{\Lambda}/N = I_k$, the estimated loadings $\tilde{\Lambda}$ are $\sqrt{T}$ times the eigenvectors of the k largest eigenvalues of $\tilde{\Sigma}/N$. Lastly, we estimate the factors from a weighted regression on the loadings for every time period t, that is,

$$\tilde{F}_t = \left(\sum_{i=1}^{N} W_{it} \tilde{\Lambda}_i \tilde{\Lambda}_i^\top \right)^{-1} \left(\sum_{i=1}^{N} W_{it} \tilde{\Lambda}_i \dot{Y}_{it} \right) .$$

Finally, we estimate the common components of Y with the plug-in estimator

$$\tilde{C}_{it} = \tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t + \tilde{\Lambda}_i^\top \tilde{F}_t \,,$$

and use $\tilde{C}_{it}$ to impute the missing observations in Y.

4 Inferential Theory

In this section, we present the asymptotic results of our wi-PCA estimator. We first lay out the assumptions on the factor model and two-way fixed effects in Section 4.1. Sections 4.2 and 4.3 show the consistency and asymptotic normality results for wi-PCA.

4.1 Assumptions

We state the necessary assumptions on the approximate factor model and fixed effects for the consistency of our wi-PCA estimator. When the panel is fully observed, the assumptions on the approximate factor model in Assumption 2 are identical to those in Bai and Ng (2002) and Bai (2003). In the presence of missing observations, Assumption 2 is identical to the corresponding assumption in Xiong and Pelger (2023).

Assumption 2 (Factor Model)**.** *There exists a constant $M < \infty$ such that*

1. *Factors:* $\mathbb{E}[\|F_t\|^4] \leq M$ *for any* t. *There exists a positive definite* $k \times k$ *matrix* Σ_F *such that* $T^{-1}\sum_{t=1}^T F_t F_t^\top \xrightarrow{p} \Sigma_F$ *and* $\mathbb{E}\left\|\sqrt{T}(T^{-1}\sum_{t=1}^T F_t F_t^\top - \Sigma_F)\right\|^2 \leq M$. *Furthermore, for any set* Q_{ij}, $|Q_{ij}|^{-1}\sum_{t \in Q_{ij}} F_t F_t^\top \xrightarrow{p} \Sigma_F$ *and* $\mathbb{E}\left\|\sqrt{|Q_{ij}|}(|Q_{ij}|^{-1}\sum_{t\in Q_{ij}} F_t F_t^\top - \Sigma_F)\right\|^2 \leq M$.
2. *Loadings:* $\mathbb{E}[\|\Lambda_i\|^4] \leq M$ *for any* i. *There exist positive definite* $k \times k$ *matrices* Σ_Λ *and* $\Sigma_{\Lambda,t}$ *such that* $N^{-1}\sum_{i=1}^N \Lambda_i \Lambda_i^\top \xrightarrow{p} \Sigma_\Lambda$ *and* $N^{-1}\sum_{i=1}^N W_{it}\Lambda_i\Lambda_i^\top \xrightarrow{p} \Sigma_{\Lambda,t}$ *for any* t.
3. *Idiosyncratic errors:*

   (a) $\mathbb{E}[\epsilon_{it}] = 0, \mathbb{E}[\epsilon_{it}^8] \leq M$.

   (b) $\mathbb{E}\left[\epsilon_{is}\epsilon_{it}\right] = \gamma_{i,st}$ *with* $|\gamma_{i,st}| \leq \gamma_{st}$ *for some* γ_{st}, *and* $\sum_{s=1}^T \gamma_{st} \leq M$ *for all* t.

   (c) $\mathbb{E}\left[\epsilon_{it}\epsilon_{jt}\right] = \tau_{ij,t}$ *with* $|\tau_{ij,t}| \leq \tau_{ij}$ *for some* τ_{ij}, *and* $\sum_{j=1}^N \tau_{ij} \leq M$ *for all* i.

   (d) $\mathbb{E}\left[\epsilon_{it}\epsilon_{js}\right] = \tau_{ij,ts}$ *and* $\sum_{j=1}^N\sum_{s=1}^T |\tau_{ij,ts}| \leq M$ *for all* i *and* t.

   (e) *For all* i *and* j, $\mathbb{E}\Big[|Q_{ij}|^{-1/2}\sum_{t\in Q_{ij}}\left(\epsilon_{it}\epsilon_{jt} - \mathbb{E}\left[\epsilon_{it}\epsilon_{jt}\right]\right)\Big]^4 \leq M$.
4. *Dependence: Loadings are independent of factors and idiosyncratic errors. There is weak dependence between factor and errors:* $\mathbb{E}\left\||Q_{ij}|^{-1/2}\sum_{t\in Q_{ij}} F_t\epsilon_{it}\right\|^2 \leq M$ *for all* i *and* j.

Assumption 2.1 ensures that all the factors have nontrivial contributions to the variation in Y. Assumption 2.2 states that all the factors are strong, and the factor loadings of the observed units are systematic. Assumption 2.3 allows the idiosyncratic errors to be weakly correlated in both the cross-sectional and time series dimensions. This assumption implies bounded eigenvalues of the error covariance matrix. Assumption 2.4 allows the factors and errors to be weakly correlated, but we assume that the loadings are independent of the factors and errors. This is a necessary assumption on the observation pattern to be also independent of factors and errors. In summary, these are the standard assumptions in the literature on approximate factor models.

As our model includes the two-way fixed effects, we need additional assumptions to identify and consistently estimate the fixed effects and factor model.

Assumption 3 (Additional Assumptions on Fixed Effects and Factor Model)**.**

1. *Fixed effects: Fixed effects are independent of factors and errors, and* $\mathbb{E}[\alpha_i^4] \leq M$ *for any* i.
2. *Identification assumptions:* $\sum_{i=1}^N \alpha_i = 0$, $\sum_{t=1}^T \xi_t = 0$, $\sum_{i=1}^N \Lambda_i = 0$ *and* $\sum_{t=1}^T F_t = 0$.
3. *Additional moment conditions: (1)* $T^{-2}\sum_{s,t,y,z=1}^T |\mathbb{E}[F_{t,p}F_{s,q}F_{y,r}F_{z,h}]| \leq M$ *for any* p,q,r,h. *(2)* $\mathbb{E}[\epsilon_{is}\epsilon_{it}\epsilon_{iy}\epsilon_{iz}] = \gamma_{i,styz}$ *with* $|\gamma_{i,styz}| \leq \gamma_{styz}$ *for some* γ_{styz}, *and* $T^{-2}\sum_{s,t,y,z=1}^T \gamma_{styz} \leq M$. *(3)* $\sum_{s=1}^T \|\mathbb{E}[F_s\epsilon_{it}]\| \leq M$ *and* $\sum_{j=1}^N \|\mathbb{E}[F_t\epsilon_{it}\epsilon_{jt}]\| \leq M$ *for any* i,t.

Assumption 3 collects the additional assumptions arising from including fixed effects in the model. First, Assumption 3.1 ensures that the unit fixed effects have bounded fourth moments. Importantly, for time fixed effects, we do not impose restrictions on their dynamics and allow them to follow arbitrary non-stationary stochastic processes.[7] Assumption 3.2 imposes an identification condition. If we only want to draw inferences on the common component, we do not need this identification assumption. However, if we want to make inference separately on the fixed effect component, loadings and factors, we need to impose an identification assumption to uniquely separate the different elements. Our identification assumption assigns the non-zero mean of factors and loadings to the fixed effect structure. Assumption 3.3 imposes additional mild higher moment conditions. We assume weak time-series dependencies between the different factors and idiosyncratic error terms. This allows for a broad range of stochastic processes, for example, causal autoregressive processes.

Next, we formally state the assumptions for the three different cases of the observation pattern of the wi-PCA estimator that we have introduced in Section 3:

Assumption 4. *We assume one of the following cases holds:*

- *Case 1: The observation probability p_{it} is known.*
- *Case 2: The observation probability p_{it} can be factorized into a one-factor model as $p_{it} = u_i v_t$ with $u_i, v_t \in [\eta, 1]$, and the time series average $\bar{v} = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} v_t$ exists and $\bar{v} = T^{-1}\sum_{t=1}^{T} v_t + O(T^{-1/2})$. Furthermore, there exists a constant c such that $W_{it} \perp\!\!\!\perp W_{is} \mid I_{it} \cup I_{is}$ for any s and t satisfying $|s-t| > c$.*
- *Case 3: The observation pattern is monotone, i.e., either $W_{i1} \leq W_{i2} \leq \cdots \leq W_{iT}$ or $W_{i1} \geq W_{i2} \geq \cdots \geq W_{iT}$ holds for all i. Furthermore, $\mathbb{E}\left[\| |\mathcal{N}_c|^{-1} \sum_{i\in\mathcal{N}_c} \Lambda_i \|^4\right] \leq M/\delta_{N,T}^2$.*

Under Case 1 with known observation probability, the weight choice M_{it}^{ξ} results in the Hajek estimator of ξ_t, which is a widely used and efficient estimator in causal inference. Case 2 allows observation probabilities to be unknown, but assumes a one-factor structure in the observation

[7] Assumption 3.1 is stated assuming that fixed effects are realizations of random variables. Then we can state the assumptions about the moments or dependencies between different random variables in the model. We call these effects "fixed effects" as opposed to "random effects" for two reasons. First, we explicitly estimate the values of these effects. Second, we condition on the values of these effects when estimating the factor model.

probabilities. Under Case 2 the time series average of W_{it}

$$\bar{W}_{i,\cdot} = \frac{1}{T}\sum_{t=1}^{T} W_{it} = u_i\bar{v} + o_p(1) = \frac{p_{it}}{v_t}\cdot\bar{v} + o_p(1)$$

is a consistent estimator of $u_i\bar{v}$. Under a one-factor structure in the observation probability, the weight M^{ξ}_{it} also implies an estimator that is asymptotically the same as the Hajek estimator of ξ_t:

$$\begin{aligned} M^{\xi}_{it} &= \left(\sum_{i=1}^{N}\frac{W_{it}}{\bar{W}_{i,\cdot}}\right)^{-1}\frac{W_{it}}{\bar{W}_{i,\cdot}} \\ &= \left(\sum_{i=1}^{N}\frac{W_{it}}{p_{it}\cdot\bar{v}/v_t + o_p(1)}\right)^{-1}\frac{W_{it}}{p_{it}\cdot\bar{v}/v_t + o_p(1)} = \left(\sum_{i=1}^{N}\frac{W_{it}}{p_{it} + o_p(1)}\right)^{-1}\frac{W_{it}}{p_{it} + o_p(1)}. \end{aligned}$$

Case 3 violates the assumptions in both Cases 1 and 2. We therefore use a conceptually different estimator that leverages the structure in the observation pattern – this estimator averages over the control units that are fully observed for all time periods to estimate ξ_t. The additional moment condition in Case 3 implies that whether a unit adopts the treatment only weakly depends on the loadings. However, this condition allows the treatment adoption time to depend on the loadings, given that a unit adopts the treatment.

4.2 Consistency Results

Theorem 1 shows the consistency of wi-PCA under the observation pattern specified in Assumptions 1 and 4, and the general factor model specified in Assumptions 2 and 3.

Theorem 1 (Consistency)**.** *Suppose Assumptions 1, 2, 3 and 4 hold. We define $\delta_{N,T} = \min(N,T)$ and specify the weights for $\tilde{\alpha}_i$ and $\tilde{\xi}_t$ appropriately based on Case 1, Case 2 or Case 3. Then, as $N, T \to \infty$, wi-PCA consistently estimates the sum of grand mean and fixed effects and the common component:*

$$\begin{aligned} \sqrt{\delta_{N,T}}\big((\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t) - (\mu + \alpha_i + \xi_t)\big) &= O_p(1), \\ \sqrt{\delta_{N,T}}\big(\tilde{C}_{it} - C_{it}\big) &= O_p(1). \end{aligned}$$

Theorem 1 states that both $\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t$ and $\tilde{C}_{it}$ are consistent with a convergence rate of

$\min(\sqrt{N}, \sqrt{T})$. The consistency of $\tilde{C}_{it}$ implies the consistency of the individual factors $\tilde{F}_t$ and loading $\tilde{\Lambda}_i$ up to a rotation matrix. For brevity, we do not include the results for the individual factors and loadings, because the main object of interest is the common component for causal inference and data imputation applications. Next we present the asymptotic distribution of the estimated common components.

4.3 Asymptotic Normality

Showing the asymptotic distribution of wi-PCA is much more challenging than for the existing estimators in the literature (such as Bai (2003) for fully observed panels, and Jin, Miao, and Su (2021); Bai and Ng (2021); Cahan, Bai, and Ng (2023); Xiong and Pelger (2023); Duan, Pelger, and Xiong (2023) for partially observed panels). The main reason is that the first-stage estimation error of the grand mean and two-way fixed effects carries over to the second-stage estimation of the factor structure. Therefore, we consider a simplified model that conveys the main conceptual insights of the general factor model, but reduces the complexity due to including fixed effects in our estimator. We state two assumptions on the factor model, which altogether specifies the simplified factor model considered in this paper.

Assumption 5 (Simplified Factor Model with Two-Way Fixed Effects)**.**

1. *Fixed effects:* $\alpha_i \overset{i.i.d.}{\sim} (0, \sigma_\alpha^2)$, *and* $\mathbb{E}[\alpha_i^4] \leq M$ *for any* i.
2. *Factors:* $F_t \overset{i.i.d.}{\sim} (0, \Sigma_F)$, *and* $\mathbb{E}[\|F_t\|^4] \leq M$ *for any* t.
3. *Loadings:* $\Lambda_i \overset{i.i.d.}{\sim} (0, \Sigma_\Lambda)$, *and* $\mathbb{E}[\|\Lambda_i\|^4] \leq M$ *for any* i. *Furthermore, there exists a positive definite* $k \times k$ *matrix* $\Sigma_{\Lambda,t}$ *such that* $N^{-1} \sum_{i=1}^N W_{it} \Lambda_i \Lambda_i^\top \xrightarrow{p} \Sigma_{\Lambda,t}$ *for any* t.
4. *Idiosyncratic errors:* $\epsilon_{it} \overset{i.i.d.}{\sim} (0, \sigma_\epsilon^2)$ *and* $\mathbb{E}[\epsilon_{it}^8] \leq M$.
5. *Independence:* α, ξ, F, Λ *and* ϵ *are mutually independent.*

The simplified model is a special case of our general approximate factor model, that is, the simplified assumption implies the general Assumption 2. Assumption 5 assumes that the unit fixed effects, factors, loadings, and idiosyncratic errors are i.i.d and have bounded moments. In addition to Assumption 5, the asymptotic normality results require further assumptions on the observation pattern and its dependency on the factor model. We use $I_i = I_{i1} \cup \cdots \cup I_{iT}$ to denote the information set that impacts the full time-series of the observation pattern of unit i.

Assumption 6 (Additional Conditions for Central Limit Theorem)**.**

1. *For any t, we have the limits* $\text{plim}_{N\to\infty} N^{-1}\sum_{j=1}^{N} W_{jt}\Lambda_j$*,* $\text{plim}_{N\to\infty} N^{-1}\sum_{j=1}^{N} W_{jt}\bar{W}_{j,\cdot}^{-1}\Lambda_j$*, and* $\text{plim}_{N\to\infty} N^{-1}\sum_{j=1}^{N} W_{jt}\bar{W}_{j,\cdot}^{-1}\Lambda_j\Lambda_j^\top$ *exist.*
2. *For any t, $N^{-1}\sum_{j=1}^{N} p_{jt}^{-1} \xrightarrow{p} s_t$ as $N \to \infty$. Furthermore, $N^{-1}\sum_{j=1}^{N}\mathbb{E}[W_{jt}W_{js}/(p_{jt}p_{js}) \mid I_j]\alpha_j^2 \xrightarrow{p} s_{\alpha,st}$ and $N^{-1}\sum_{j=1}^{N}\mathbb{E}[W_{jt}W_{js}/(p_{jt}p_{js}) \mid I_j]\Lambda_j\Lambda_j^\top \xrightarrow{p} s_{\Lambda,st}$ for any s and t, and* $\text{plim}_{T\to\infty} T^{-2}\sum_{s,u=1}^{T} W_{is}W_{iu}s_{\alpha,su}$ *exists for any i.*
3. *For any i, j, h, l, $|Q_{ij}|/T \xrightarrow{p} q_{ij}$ and $|Q_{ij} \cap Q_{hl}|/T \xrightarrow{p} q_{ij,hl}$ as $T \to \infty$. Furthermore,* $\lim_{N\to\infty} N^{-2}\sum_{i,l=1}^{N} q_{ij,lj}$ *and* $\lim_{N\to\infty} N^{-4}\sum_{i,j,l,h=1}^{N} q_{ij,lh}/(q_{ij}q_{lh})$ *exist.*
4. *When Case 2 holds, $N^{-1}\sum_{j=1}^{N}\mathbb{E}\left[W_{jt}\bar{v}^2/(\bar{W}_{j,\cdot}^2 v_t^2) \mid I_j\right] \xrightarrow{p} \bar{s}_t$ and Assumption 6.2 holds with W_{jt}/p_{jt} replaced by $W_{jt}\bar{v}/(\bar{W}_{j,\cdot}v_t)$ and $s_{\alpha,st}, s_{\Lambda,st}$ replaced by $\bar{s}_{\alpha,st}, \bar{s}_{\Lambda,st}$ for any j, s, t. When Case 3 holds, $\sqrt{N_c}^{-1}\sum_{j\in\mathcal{N}_c}\Lambda_j \xrightarrow{d} \mathcal{N}(0, \Sigma_{\Lambda,c})$ for some $\Sigma_{\Lambda,c}$.*

Assumption 6.1 assumes that weighted averages of the loadings converge. Assumption 6.2 is necessary to show that the estimator $\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t$ is asymptotically normal, which is necessary for showing the asymptotic normality of the estimated factor model. Assumption 6.3 imposes assumptions on the proportion of time periods when multiple units are observed together. Lastly, Assumption 6.4 contains additional assumptions for Cases 2 and 3.

Below we present the asymptotic distribution of the estimated common components under all the aforementioned assumptions. To show this result, we first need to derive the asymptotic joint normal distributions of the estimators of the fixed effects, factors and loadings. Theorem 2 only presents the asymptotic results for the estimated common components, as the primary interest is the construction of confidence intervals for imputed values or draw inferences on treatment effects.

Theorem 2 (Asymptotic Normality)**.** *Suppose Assumptions 1, 4, 5 and 6 hold. We define $\delta_{N,T} = \min(N, T)$ and specify the weights for $\tilde{\alpha}_i$ and $\tilde{\xi}_t$ appropriately based on Case 1, Case 2 or Case 3. Then, as $N, T \to \infty$, the asymptotic distribution of the common components estimated by wi-PCA satisfies*

$$\sqrt{\delta_{N,T}} \cdot \sigma_{C,it}^{-1}\left(\tilde{C}_{it} - C_{it}\right) \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\sigma_{C,it}^2 = \delta_{N,T}/N \cdot \sigma_{C,it,1}^2 + \delta_{N,T}/T \cdot \sigma_{C,it,2}^2$ for some $\sigma_{C,it,1}^2$ and $\sigma_{C,it,2}^2$.

The inclusion of the grand mean and fixed effects substantially complicates the derivation and asymptotic variances $\sigma^2_{C,it,1}$ and $\sigma^2_{C,it,2}$, which contain more terms than the asymptotic variance for the pure factor model in Xiong and Pelger (2023). Specifically, we need to deal with the first-stage estimation error $\tilde{\Delta}_{it} = (\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t) - (\mu + \alpha_i + \xi_t)$ as well as two additional terms, which are respectively $\sqrt{T} N^{-1} \sum_{i=1}^{N} \Lambda_i \Lambda_i^\top |Q_{ij}|^{-1} \sum_{t \in Q_{ij}} F_t \tilde{\Delta}_{it}$ in the asymptotic distribution of $\tilde{\Lambda}_i$ and $N^{-1/2} \sum_{i=1}^{N} W_{it} \Lambda_i \tilde{\Delta}_{it}$ in the asymptotic distribution of $\tilde{F}_t$. In order to simplify the expression and focus on the main result, we do not explicitly state the form of $\sigma^2_{C,it,1}$ and $\sigma^2_{C,it,2}$. Instead we propose in the next section a practical bootstrap approach for the variance estimation of relevant quantities.

5 Application to Causal Inference

In this section, we discuss the application of our estimator to causal inference and provide a feasible variance estimator for the average treatment effects.

5.1 Estimation of Treatment Effects

A key application of our estimator is to estimate the treatment effects in causal inference. The fundamental problem of causal inference is that we observe either the control outcome or the treated outcome, but not both, for any unit at a specific time period. We can view the unobserved counterfactual control outcomes as missing values, which we impute with wi-PCA. By comparing the observed treated units with the imputed control units, we can estimate the treatment effects on the treated.

In the causal inference context, the outcome panel Y_{it} has both control and treated observations. We view the treated values as missing control observations, which results in the partially observed panel of control units $Y_{it}^{(\text{ct})}$. Correspondingly, the partially observed panel of treated units is denoted as $Y_{it}^{(\text{tr})}$. The treatment effect of unit i at time t is defined as

$$\tau_{it} = Y_{it}^{(\text{tr})} - Y_{it}^{(\text{ct})} .$$

Our focus is on the estimation and inference of the average treatment effect on a treated unit i

over time, denoted by τ_i , which is formally defined as

$$\tau_i := \frac{1}{T_{i,tr}} \sum_{t \in \mathcal{T}_{i,tr}} \left(Y_{it}^{(\mathrm{tr})} - Y_{it}^{(\mathrm{ct})} \right) = \frac{1}{T_{i,tr}} \sum_{t \in \mathcal{T}_{i,tr}} \tau_{it} \, ,$$

where $\mathcal{T}_{i,tr}$ and $T_{i,tr}$ denote the set and number of treated time periods of unit i, respectively. The average treatment effect on treated (ATT) is an important quantity that is often used in practice. Alternatively, we can also study the average treatment effects that are averaged over units, and over both time and units. This follows the same conceptual arguments as for τ_i and would be a straightforward extension of our framework.

We specify a model for the control panel to estimate the unobserved control units. We argue that the wi-PCA is particularly well suited for imputing the counterfactual control units. As discussed before, in the special case of only the fixed effect structure, but without the latent factor model, wi-PCA can simplify to a difference-in-difference estimator, which is widely used in causal inference. The latent factor model can be interpreted as a data-driven approach for constructing synthetic controls. Conceptually, synthetic controls are weighted averages of untreated units, where the weights depend on unit-specific features. Our approach does not require priori knowledge about which covariates describe if treated and control units are a good match. Instead, our latent loadings capture complex unit-specific information in a data-driven way. Our wi-PCA can be interpreted as combining a difference-in-difference estimator with a general synthetic control.

Specifically, under the assumption that the control panel $Y^{(\mathrm{ct})}$ follows the approximate factor model with two-way fixed effects in Equation (1), we use the common component $\tilde{C}_{it}$ estimated with wi-PCA to impute all values in $Y^{(\mathrm{ct})}$. This yields the following estimator for τ_i:

$$\hat{\tau}_i = \frac{1}{T_{i,tr}} \sum_{t \in \mathcal{T}_{i,tr}} (Y_{it}^{(\mathrm{tr})} - \tilde{C}_{it}) = \tau_i + \frac{1}{T_{i,tr}} \sum_{t \in \mathcal{T}_{i,tr}} (C_{it} - \tilde{C}_{it}) + \frac{1}{T_{i,tr}} \sum_{t \in \mathcal{T}_{i,tr}} \epsilon_{it} \, .$$

The estimator $\hat{\tau}_i$ is consistent if $\tilde{C}_{it}$ is consistent and the average over the idiosyncratic terms vanishes (which holds for $T_{i,tr}$ going to infinity). Moreover, if the assumptions of Theorem 2 are satisfied, then $\hat{\tau}_i$ is asymptotically normally distributed. In the next section, we provide a feasible estimator for the asymptotic variance of $\hat{\tau}_i$.

5.2 Feasible Variance Estimator for ATT

In this section, we show how to estimate the asymptotic variance and construct asymptotically valid confidence intervals for the average treatment effect on the treated τ_i using a resampling bootstrap approach. As discussed in the previous section, we can estimate the counterfactual control outcomes $\tilde{C}_{it}$ with wi-PCA and obtain the estimation error for τ_i as

$$\hat{\tau}_i - \tau_i = \frac{1}{T_{i,tr}} \sum_{t \in \mathcal{T}_{i,tr}} (C_{it} - \tilde{C}_{it}) + \frac{1}{T_{i,tr}} \sum_{t \in \mathcal{T}_{i,tr}} \epsilon_{it} \,.$$

Theorem 2 implies that the first term in the estimation error has an asymptotic normal distribution. Under mild assumptions and for $T_{i,tr} \to \infty$, the second term has mean zero and is jointly asymptotically normally distributed with the first term. Hence, the sum of the two parts is asymptotically normal, and the convergence rate is the smaller of the rate of each term.

Under the asymptotic normality, we obtain valid confidence interval (CI) for τ_i:

$$\mathbb{P}\left(\tau_i \in \left[\hat{\tau}_i - z_{1-\alpha/2}\sqrt{V_i}, \hat{\tau}_i + z_{1-\alpha/2}\sqrt{V_i}\,\right]\right) \overset{\text{asympt}}{\sim} 1 - \alpha\,,$$

where V_i is the asymptotic variance and z_α is the α-quantile of standard normal distribution.

We propose to estimate V_i with a resampling bootstrap procedure, which is easy to use despite the complex form of V_i. This procedure has three steps. First, we sample the idiosyncratic errors $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{iT})$ of the treated unit i for which we aim to draw inference. To obtain the sample distribution of these errors, we iteratively mask a control unit j with the same treatment pattern as the target treated unit i and apply wi-PCA to this new panel.[8] The difference between the estimated common components and observed outcomes of unit j serves as a residual time series in the sampling distribution. In the second step, we obtain estimates of τ_i from the bootstrap samples. We obtain a bootstrap sample by sampling $N-1$ units with replacement from all units besides the i-th one, and concatenating them with a "bootstrapped version" of unit i's time series. This "bootstrapped version" is obtained by adding unit i's estimated common components to a draw of ϵ_i from the sampling distribution obtained in step one. In the last step, we calculate the variance

[8] For brevity, we only consider the case where control units exist. If there is no control unit, we need to assume that the idiosyncratic errors are i.i.d. in the cross-sectional and time series dimension.

V_i using the estimates of τ_i from step two. This bootstrap provides consistent estimates of V_i under additional assumptions on the error terms. We provide a detailed description of this bootstrap procedure with the formal statements and assumptions in Appendix B.1.

The simulations in Appendix B.3 demonstrate the good finite sample performance of our bootstrap procedure and that it works well in general setups. We also show the benefits of sampling only the variance and leveraging the theoretical normal distribution, instead of sampling the complete distribution in step two. The procedure above assumes that ε_{jt} is i.i.d. in the cross-section. This assumption can be relaxed by using a block resampling procedure, where the sampling is based on blocks of correlated units. We provide the details of a block resampling procedure in Appendix B.2.

6 Discussions

6.1 Model with Observables

In empirical studies, researchers might observe additional covariates X_{it} which can capture variation in the outcome variables Y_{it}. Including these exogenous observables can reduce the variance of the imputed values and estimated treatment effects. In this section, we show how to extend wi-PCA to include additional exogenous observed covariates.

Formally, we can extend our model from Equation (1) by incorporating the observables $X_{it} \in \mathbb{R}^d$ as follows:

$$Y_{it} = \beta^\top X_{it} + \mu + \alpha_i + \xi_t + \Lambda_i^\top F_t + \epsilon_{it}\,,$$

with the coefficient $\beta \in \mathbb{R}^d$. The common component now becomes $C_{it} := \beta^\top X_{it} + \mu + \alpha_i + \xi_t + \Lambda_i^\top F_t\,$, and the goal is to estimate and draw inferences on the estimated common component. Note that C_{it} has three components: the fixed effects, latent factor structure and regression on observables. The estimator for wi-PCA with observables becomes a three-step approach with one step for each of the components. For the sake of brevity, we present the results for the case with known observation probability, and the arguments extend accordingly to the other cases.

The first step is the within transformation, which estimates the fixed effect structure and normalizes the observables. Note that with non-stationary fixed effects it is necessary to first apply the within transformation before running regressions or PCA. We apply the first step of the wi-PCA estimator to estimate the grand mean and fixed effects by calculating the weighted averages of observations in Y. Then, we subtract the first-stage estimators from Y_{it} and obtain

$$\dot{Y}_{it} = \beta^\top \dot{X}_{it} + \Lambda_i^\top F_t + \dot{\epsilon}_{it} \,,$$

where $\dot{Y}_{it} = Y_{it} - \tilde{\mu} - \tilde{\xi}_t - \tilde{\alpha}_i$ and $\dot{X}_{it} = X_{it} - \bar{X}_{\cdot,t} - \bar{X}_{i,\cdot} + \bar{X}$. The within transformation of observed covariates is needed for the purpose of identification, and does not pose any challenges as X is fully observed.

The second step estimates the coefficient β by running a regression

$$\tilde{\beta} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} \dot{X}_{it} \dot{X}_{it}^\top \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} \dot{X}_{it} \dot{Y}_{it} \,.$$

The residual from this regression $\ddot{Y}_{it} = \dot{Y}_{it} - \tilde{\beta}^\top \dot{X}_{it}$ has a latent factor structure, that is, $\ddot{Y}_{it} = \Lambda_i^\top F_t + \ddot{\epsilon}_{it}$. In the third step, we estimate the latent factor structure. By applying the PCA step of wi-PCA to $\ddot{Y}_{it}$, we obtain $\tilde{\Lambda}_i^\top \tilde{F}_t$. Aggregating all three estimations yields the estimate of the common components:

$$\tilde{C}_{it} = \tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i + \tilde{\beta}^\top (X_{it} - \bar{X}_{\cdot,t} - \bar{X}_{i,\cdot} + \bar{X}) + \tilde{\Lambda}_i^\top \tilde{F}_t.$$

We call this approach “wi-PCA with observables”. The following assumption and proposition formalize the asymptotic inference.

Assumption 7 (Observables)**.** *The observables X_{it} are independent of factors and idiosyncratic errors. Furthermore, for any i, t, $\mathbb{E}[\|X_{it}\|^{16}] \leq M$, $\mathbb{E}[\|T^{-1} \sum_{t=1}^{T} W_{it} \dot{X}_{it}\|^8] \leq M/T^4$, and the smallest eigenvalue of $(NT)^{-1} \sum_{i,t} W_{it} \dot{X}_{it} \dot{X}_{it}^\top$ is positive and bounded away from 0.*

Assumption 7 guarantees that the regression on X is well behaved. Proposition 1 shows the consistency and convergence rate.

Proposition 1. *Assume Assumptions 1, 2, 3, 7 and Case 1 in Assumption 4 hold. Furthermore, the eighth moments of the unit fixed effects, factors and loadings are bounded. Let $\delta_{N,T} = \min(N, T)$.*

Then, as $N, T \to \infty$, wi-PCA with observables consistently estimates the common component:

$$\sqrt{\delta_{N,T}}(\tilde{C}_{it} - C_{it}) = O_p(1).$$

Note that the regression on observables does not affects the overall convergence rate. Including the observables does not make the analysis more challenging and hence we omit it in the main model in equation (1). It is possible to generalize the regression to a non-parametric regression if its convergence rate is not slower than $\sqrt{\delta_{N,T}}$. It is also possible to change the order of the steps and estimate the latent factors before applying the regression of the residuals on observables X. However, this alternative order implies different assumptions, which for many applications seem to be more restrictive.

6.2 Number of Factors

We assume that the number of factors k is consistently estimated. Given a consistent estimator for the number of factors, we can treat k as known. The discussion in Xiong and Pelger (2023) about how to estimate the number of latent factors also applies to our estimator. After removing the fixed effects component, the latent factor structure is similar to Xiong and Pelger (2023), and we can build on the same insights. Fundamentally, the number of factors is a tuning parameter, and in practice, can be determined by cross-validation arguments. Specifically, we can mask entries in the panel and evaluate the accuracy of the imputed masked values with the actual observed entries for different number of factors. The model with the smallest validation imputation error is selected. However, the appropriate masking is a challenging problem, in particular for complex missing patterns that can depend on the factor structure. Conceptually, the correct masking would use the actual probability of missingness, which is generally unknown in observational studies. As a practical approach, in our empirical study we use a variety of realistic masking schemes and repeat the masking simulation many times. We find that our estimator is relatively robust to the number of factors once we include sufficiently many. Other estimators do not seem to share this robustness property.

7 Simulation

We demonstrate in comprehensive simulations the good performance of our wi-PCA estimator. For this purpose, we show that our estimator dominates natural benchmarks for a variety of relevant missing patterns and data generating processes. Specifically, we demonstrate that it is important to include the fixed effect structure and latent factor structure, and omitting one of them deteriorates the performance. We compare wi-PCA with the following three estimators:

- PCA: Special case of wi-PCA without estimating the fixed effect structure separately. This estimator corresponds to the PCA estimator using all observations proposed by Xiong and Pelger (2023).
- Block-PCA: PCA estimator that estimates factors only from fully observed blocks proposed by Xu (2017).[9]
- TWFE: Two-way fixed effects estimator, which is the special case of wi-PCA without latent factors.[10]

The PCA and TWFE estimators correspond to two special cases of wi-PCA with either only a latent structure or only the two-way fixed effects. The TWFE estimator corresponds to the conventional difference-in-difference (DID) estimator for monotone missing patterns. The Block-PCA serves as an alternative way to estimate a latent factor model without directly modeling the two-way fixed effects, and has been used for causal inference on panels.

We generate the data from a one-factor model with two-way fixed effects $Y_{it} = \mu + \alpha_i + \xi_t + \Lambda_i F_t + \epsilon_{it}$, where $\mu = 1$, $\alpha_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $F_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $\Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and $\epsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\epsilon^2)$. We consider both stationary and non-stationary time fixed effects ξ_t. The stationary time fixed effects follow $\xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, while non-stationary time fixed effects include a trend process and are generated as $\xi_t = 0.05t + \mathcal{N}(0,1)$ for any t. The main text uses the parameters $\sigma_\epsilon^2 = 4$, while the Internet Appendix shows that the results are robust to other choices.

We consider three different observation patterns for Y:

[9] We provide the details of this method in Appendix C.

[10] The TWFE estimator obtains the two-way fixed effects as

$$(\hat{\mu}, \hat{\alpha}, \hat{\xi}) = \arg\min_{\mu,\alpha,\xi} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(Y_{it} - \mu - \alpha_i - \xi_t)^2,$$

and the common components as $\hat{C}_{it} = \hat{\mu} + \hat{\alpha}_i + \hat{\xi}_t$.

1. Missing-at-random: Entries are missing independently at random with probability $p = 0.2$.
2. Simultaneous treatment adoption: 50% randomly selected units are completely missing after time $0.4 \cdot T$.
3. Staggered treatment adoption: Starting form $0.1 \cdot T$, units are selected to adopt the treatment with probability $0.1 \cdot I(|\alpha_i \xi_t| > 2.5)$. Once a unit adopts treatment, it stays treated without subsequent observations.

The observation patterns are illustrated in Table 2. The missing-at-random pattern is conceptually the simplest, and all entries are missing independently and exogenously. With simultaneous treatment adoption the entries are still missing exogenously but the missingness has dependence in the time series dimension. The simulated staggered treatment adoption pattern is the most complex as in addition to the monotone missingness we further allow endogenous missingness that depends on both two-way fixed-effects.

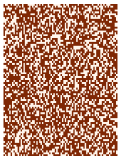
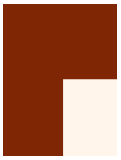
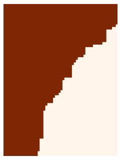
In Table 2, we compare the performance of different estimators in estimating the common components of Y on the observed, missing, and all entries and report for each method the relative mean squared error (relative MSE) defined as

$$\text{relative MSE}_{\mathcal{S}} = \frac{\sum_{(i,t)\in\mathcal{S}} (\tilde{C}_{it} - C_{it})^2}{\sum_{(i,t)\in\mathcal{S}} C_{it}^2},$$

where $\mathcal{S}$ denotes the set of either observed, missing or all entries in Y. We pay particular attention to the relative MSE for the missing entries in Y, which serves as the out-of-sample evaluation. This is important as models with more parameters are expected to approximate the data better in-sample, but due to overfitting the noise can perform worse out-of-sample. In the main text we show the results for one latent factor with wi-PCA and up to three latent factors for the pure factor models. Internet Appendix IA.A collects the results for other parameters of the data generating process. Note that for stationary data a pure latent factor model with three latent factors should theoretically be able to recover the same model as a one-factor model with two-way fixed effects.

Table 2 shows that wi-PCA dominates all other benchmarks in terms of in-sample and out-of-sample performance. We discuss the findings for each missing pattern. In the case of missing-at-random, both wi-PCA and PCA perform well for either stationary or non-stationary data, but wi-PCA is overall more efficient. A pure factor model with less than three factors suffers from an

Table 2: Relative MSE of common components for different estimators

			wi-PCA		PCA			Block-PCA			TWFE
			(FE+factor model)		(factor model only)			(factor model only)			(FE only)
	ξ_t	$\mathcal{S}$	known	unknown	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	
		obs	0.055	0.055	0.360	0.147	0.088	-	-	-	0.270
	S	**miss**	**0.056**	**0.056**	0.377	0.157	0.093	-	-	-	0.283
		all	0.055	0.055	0.364	0.149	0.089	-	-	-	0.273
		obs	0.036	0.036	0.292	0.145	0.059	-	-	-	0.177
	N	**miss**	**0.037**	**0.037**	0.305	0.157	0.063	-	-	-	0.186
		all	0.036	0.036	0.294	0.148	0.059	-	-	-	0.179
		obs	0.064	0.069	0.369	0.156	0.104	0.372	0.161	0.110	0.283
	S	**miss**	0.108	**0.105**	0.414	0.210	0.165	0.412	0.204	0.152	0.314
		all	0.077	0.079	0.383	0.172	0.122	0.384	0.174	0.122	0.292
		obs	0.048	0.051	0.376	0.209	0.091	0.359	0.186	0.080	0.207
	N	**miss**	0.055	**0.054**	0.594	0.573	0.527	0.476	0.322	0.154	0.162
		all	0.051	0.052	0.460	0.349	0.259	0.404	0.238	0.108	0.190
		obs	0.054	0.050	0.362	0.132	0.088	0.363	0.131	0.086	0.283
	S	**miss**	0.083	**0.079**	0.431	0.325	0.213	0.440	0.310	0.176	0.206
		all	0.058	0.055	0.371	0.159	0.106	0.373	0.157	0.100	0.271
		obs	0.056	0.045	0.301	0.178	0.078	0.287	0.126	0.083	0.216
	N	**miss**	0.097	**0.081**	0.911	0.759	0.602	1.135	1.065	0.494	0.175
		all	0.068	0.055	0.487	0.353	0.237	0.541	0.408	0.207	0.204

This table reports the relative MSE of different estimators for different setups. We compare the performance of wi-PCA, with known and unknown observation probability, with three benchmarks: PCA, block-PCA, and TWFE. The figures on the left show the observation patterns with shaded entries indicating observed entries and unshaded entries indicating missing entries. Each observation pattern corresponds to two rows with either stationary time fixed effects (S) or non-stationary time fixed effects (N). Bold numbers indicate the best out-of-sample relative model performance. We set $N = T = 100$ and run 200 simulations for each setup.

omitted variable bias, while the three-factor model is less efficient than directly modeling the two-way fixed effects. Using only the two-way fixed effects without a factor structure (TWFE) provides the worst results as it also suffers from an omitted variable bias. The block-PCA method is not applicable as there are no sufficiently large fully observed blocks.

In the case of simultaneous treatment adoption, wi-PCA continues to dominate the other methods. For data with stationary time fixed effects, all three PCA methods perform relatively well. However, for the non-stationary data, the PCA method has poor performance due to the correlation between the upward trend and missing pattern in the time series dimension. This correlation deteriorates the estimation of the covariance matrix in PCA, whereas block-PCA is unaffected because it only uses the fully observed block to estimate factors. This property of block-PCA, however, makes it less efficient than wi-PCA which leverages all the data to estimate factors. TWFE continues to suffer from omitting the latent factor structure. Note that with the upward trend in the time-fixed effects, the two-way fixed effect component explains more of the overall variation.

In the last setting of staggered treatment adoption, wi-PCA significantly outperforms the other methods for both stationary and non-stationary data. In this setting, we assume an endogenous observation pattern that depends on both cross-sectional and time series components of the model. This endogeneity cannot be accounted for by both PCA and block-PCA methods, leading to their poor performance. Their performance deteriorates even more with non-stationary data. In contrary, TWFE suffers less from non-stationarity and complex missingness, but obviously still omits the factor component. The performance gains of wi-PCA are impressive as it has two to eight times better out-of-sample accuracy.

We conclude that wi-PCA combines the advantages of the three benchmark methods for different settings and is more efficient than any of them. Moreover, wi-PCA with unknown or known observation probability perform quite similarly. The wi-PCA with unknown observation probability can have even smaller relative MSEs compared to its counterpart with known observation probability, as illustrated in the last two settings. Intuitively, the direct estimation can correct for variability in the sampling distribution of the observation pattern.

Our findings are robust to the parameters of the simulation. The Internet Appendix collects the results for different proportions of missing data and noise variances. Our wi-PCA dominates the benchmarks in all setups. We conclude that it is crucial to explicitly combine both the latent factor model and two-way fixed effects in a model.

8 Empirical Study: Liberalization of Marijuana Policy

We demonstrate the benefits of our wi-PCA estimator in an empirical application in causal inference: the liberalization of marijuana. This empirical application is originally studied in Li and Sonnier (2023), where they estimate the effect of legalizing recreational marijuana on retail sales for beer. The underlying question is whether marijuana represents a substitute or complement to alcohol.

We use the weekly NielsenIQ retail scanner beer sales data from the Kilts Center for Marketing. Our data consists of 208 weekly observations of beer sales revenue from January 01, 2017, to

December 26, 2020, aggregated at the state level for the U.S.[11]

Our cross-section consists of 45 states, of which 39 are control states and 6 are treated states that legalized the retail sale of marijuana at different times. In our study, we have to drop three states (Colorado, Washington, and Oregon) as they legalized recreational marijuana before 2017. Table 4 shows the treatment times of the 6 treated states. Different from Li and Sonnier (2023), our outcome variable is the weekly beer sales revenue normalized by the number of stores included in the Kilts Center dataset. This normalization is important as the number of stores changes substantially over time, which we discuss in more detail in Internet Appendix IA.B.

In this empirical study, we continue to compare wi-PCA with the three benchmark estimators introduced in Section 7: PCA estimator (with factor model only), block-PCA estimator (with factor model only), and TWFE estimator (with fixed effects only). The PCA and TWFE estimators correspond to two special cases of our model with only a factor model or fixed effects, and the block-PCA method is implemented as in Li and Sonnier (2023).

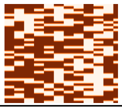
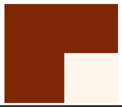
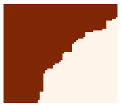
8.1 Synthetic Treatment Assignment

We start our analysis by comparing the performance of the four estimators. For this purpose we generate synthetic treatment assignments on the control panel that consists of the 39 control states. Specifically, we generate synthetic treatment patterns for the control panel and mask the corresponding control outcomes in the data as if they were treated. Then, we compare the estimation accuracy of different estimators on these synthetic treatments, which should have zero effects on the outcomes. This type of synthetic treatment is frequently considered in this literature, for example in Bertrand, Duflo, and Mullainathan (2004) and Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021).

This analysis allows us to answer the following questions. First, as this is an out-of-sample evaluation of the factor models, it allows us to compare the accuracy of the different models. The in-sample fit of a model naturally increases with the complexity of the model; however, due to potential overfitting, this is not the case out-of-sample. Second, it shows how the masking mechanism affects the estimation and accuracy of models. Third, this analysis can also serve as a validation to select

[11] Guha and Ng (2019) show the benefit of using factor models for the weekly scanner data and the presence of common seasonality in this type of data.

Table 3: Results for synthetic treatment patterns (unit: \$100/store)

		wi-PCA			PCA			Block-PCA			TWFE
		(FE+factor model)			(factor model only)			(factor model only)			(FE only)
		$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	
	RMSE_Y	2.21	**1.93**	2.01	2.53	2.64	2.95	-	-	-	2.54
	$\lvert\text{Bias}_{\text{ATT}}\rvert$	0.16	**0.14**	0.16	0.22	0.59	0.98	-	-	-	0.22
	RMSE_{ATT}	1.24	**1.11**	1.18	1.68	1.76	2.03	-	-	-	1.57
	RMSE_Y	2.42	2.09	**1.94**	2.37	4.61	5.45	2.31	2.25	2.13	2.45
	$\lvert\text{Bias}_{\text{ATT}}\rvert$	**0.26**	0.27	0.26	0.83	3.76	4.19	0.45	0.32	0.30	0.42
	RMSE_{ATT}	**0.99**	1.03	1.06	1.67	4.08	4.52	1.57	1.13	1.18	1.53
	RMSE_Y	4.30	**3.92**	4.10	5.00	6.77	7.46	4.88	7.16	5.93	4.60
	$\lvert\text{Bias}_{\text{ATT}}\rvert$	0.75	**0.47**	0.57	1.45	4.47	5.25	0.90	1.26	1.04	0.59
	RMSE_{ATT}	2.89	**2.77**	2.91	4.12	5.43	6.06	3.88	5.11	4.21	3.50

This table shows the results for three synthetic treatment patterns. The figures on the left illustrate the treatment patterns where the shaded entries indicate the observed control outcomes and the blank entries represent the missing observations. For each of the treatment pattern, we compare the RMSE of the imputation (RMSE_Y) and the absolute estimation bias and RMSE of the ATT ($|\text{Bias}_{\text{ATT}}|$ and RMSE_{ATT}) for different methods. Bold numbers indicate the best relative model performance. The three treatment patterns are generated as follows: (1) Uniformly random treatment pattern: for each unit, we set 20 time periods as a group and randomly assign treatment with probability 0.2 for each 20 time periods. (2) Simultaneous treatment adoption pattern: States are randomly selected into the treatment group with a probability of 0.5. Treated units simultaneously adopt the treatment at time $T_0 = 140$ and stay treated afterward. (3) Staggered treatment adoption pattern: States with unit fixed effects $\alpha_i > 0.5$ are randomly selected into the treatment group with a probability of 0.9. For treated units, they adopt the treatment with probability 0.1 in each time period with time fixed effects $\xi_t > 1$ and stay treated afterwards. We iterate 100 times in the simulations for each setup.

the number of latent factors. Tuning parameters like the number of factors can be selected based on which model provides the best out-of-sample performance for relevant masking schemes.

Table 3 summarizes the results for three treatment patterns, and provides a detailed description of how we generate the treatment patterns. The first pattern is a uniformly random pattern where we randomly assign treatments to each unit. The second pattern is a simultaneous treatment adoption pattern where units are randomly selected to adopt treatment at a given time. The last one is a staggered treatment adoption pattern where the treatments are endogenous and can depend on the two-way fixed effects. To generate endogenous treatment patterns, we first estimate the two-way fixed effects as the natural averages $\hat{\mu} = (NT)^{-1} \sum_{j,s} Y_{js}$, $\hat{\xi}_t = N^{-1} \sum_j Y_{jt} - \hat{\mu}$ and $\hat{\alpha}_i = T^{-1} \sum_s Y_{is} - \hat{\mu}$ for the fully observed control panel. Then, we determine the treatment mechanism based on these fixed effects.

We use the four estimators to impute the missing observations in the control panel and estimate the state-level ATT for each treated state. Table 3 compares the performance of different methods in estimating the out-of-sample counterfactual outcomes and treatment effects and reports for each method the root-mean-squared error (RMSE) of the imputation as well as the absolute bias and RMSE of the estimated ATT, which are respectively defined as

$$\mathrm{RMSE}_Y = \sqrt{\sum_{(i,t)\in\mathcal{S}_{\text{miss}}} (\tilde{C}_{it} - Y_{it})^2},$$

$$|\mathrm{Bias}_{\mathrm{ATT}}| = \left| \frac{1}{N_{tr}} \sum_{i\in\mathcal{N}_{tr}} (\hat{\tau}_i - \tau_i) \right|, \qquad \mathrm{RMSE}_{\mathrm{ATT}} = \sqrt{\frac{1}{N_{tr}} \sum_{i\in\mathcal{N}_{tr}} (\hat{\tau}_i - \tau_i)^2},$$

where $\mathcal{S}_{\text{miss}}$ denotes the set of missing entries in Y, $\mathcal{N}_{tr}$ and N_{tr} denote the set and number of treated units, and $\hat{\tau}_i$ denotes the ATT estimator for unit i.

For conciseness, Table 3 shows the results for only three latent factors, but all the results extend to up to 8 factors. Appendix D provides detailed robustness results for more factors. In fact, wi-PCA is remarkably stable and including “too many” factors has only a marginal effect. In contrast, other PCA methods deteriorate substantially for too many factors. We will revisit this important point in the next section.

The results demonstrate three main insights. First, wi-PCA dominates all benchmark estimators and is the most accurate estimator under all treatment assignments and error metrics. This means wi-PCA has the most precise estimates of first and second moments on the out-of-sample data. Second, wi-PCA is extremely stable for different number of factors, which is not the case for PCA and Block-PCA. Our findings also demonstrate the importance of estimating both the fixed effects and the factor model. Note that under stationarity assumptions PCA and Block-PCA could include the fixed effects as additional factors, but even with more latent factors their performance is substantially worse than separately estimating the fixed effects. On the other hand, there is a benefit of including at least one latent factor in addition to the fixed effects, that is, TWFE, which can be interpreted as the “zero-factor model” of wi-PCA, is never optimal.

Third, we demonstrate how the performance of models depends on the treatment assignment and confirm that wi-PCA outperforms for all the different treatment patterns. The case of missing uniformly completely at random can be viewed as the simplest case. In this setting, wi-PCA, PCA, and TWFE estimators exhibit comparable performance; however, wi-PCA excels in providing more precise imputations of counterfactual outcomes and more accurate estimations of treatment effects. Note that the block-PCA method is infeasible as there are no fully observed control units. In the case of simultaneous treatment adoption, wi-PCA and block-PCA perform well, while the PCA estimator performs poorly because of the dependency between the time series missingness. This result aligns with the simulation results with non-stationary fixed effects. Our wi-PCA significantly outperforms the other three methods for the staggered adoption pattern where the treatment assignment is endogenous and correlated with cross-sectional and temporal components.

Our findings are supported by our theory. In contrast to TWFE, wi-PCA incorporates interactive effects that better explain the correlation structure in the panel. In contrast to the PCA and block-PCA estimators, wi-PCA explicitly estimates and controls for the additive effects in the panel data. In contrast to all three benchmarks, the estimation of wi-PCA is valid under more general treatment assignments.

8.2 Estimation of ATT

In the second part, we estimate the treatment effects and their significance for the six actually treated states. We demonstrate how the economic conclusions are affected by the choice of estimation approach. Table 4 shows the estimated ATTs and corresponding standard errors estimated with the resampling bootstrap. We report the results for the wi-PCA, PCA, and block-PCA estimators for up to $k = 4$ latent factors.

There are three key observations. First, Table 4 shows that different estimators give substantially different point estimates of ATT. The point estimates of PCA and block-PCA can be more than 10 times larger than the ones for wi-PCA. Hence, the economic magnitude of the estimates seems to be largely exaggerated with PCA and block-PCA. Second, the standard errors of the benchmarks are generally much larger than with wi-PCA. In particular, for block-PCA, which uses the data inefficiently, the standard errors can explode for a larger number of factors and be twice as large as with wi-PCA. The different point estimates and standard errors can lead to different conclusions. Our wi-PCA finds that in none of the six states legalizing marijuana causes a change in beer sales. In contrast, for both PCA and block-PCA there are cases with statistically significant treatment effects.

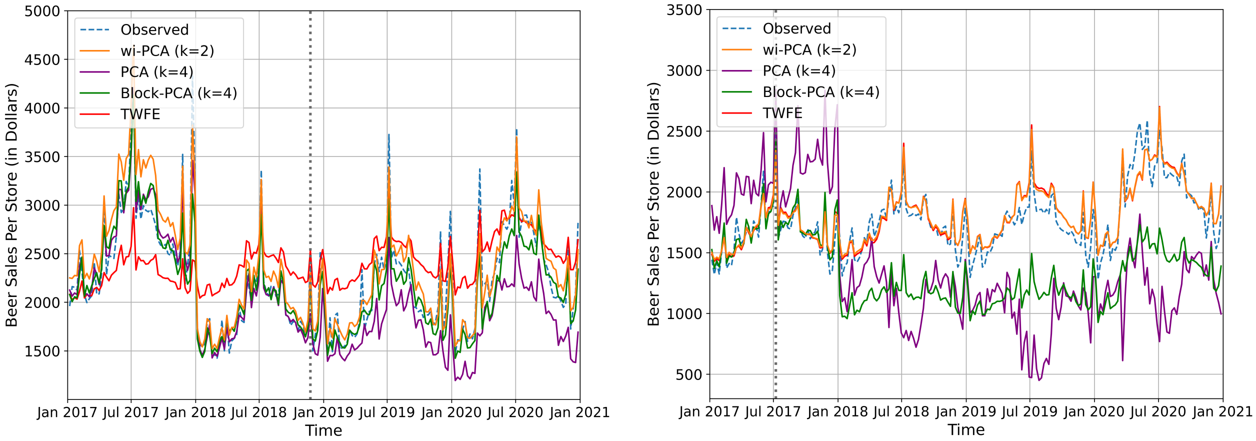
Third and most importantly, wi-PCA is extremely stable for different number of latent factors, while the instability of PCA and block-PCA makes them unreliable. To illustrate this point, we focus on the states of Massachusetts (MA) and Nevada (NV). For both states, the ATT of wi-PCA

Table 4: Average treatment effects on the actually treated states (unit: \$100/store)

		CA.		MI.		IL.		MA.		NV.		ME.	
	k	ATT	SE	ATT	SE	ATT	SE	ATT	SE	ATT	SE	ATT	SE
wi-PCA	1	0.52	3.78	0.43	1.17	0.32	1.21	-0.48	1.11	-0.47	3.18	-1.15	1.11
(FE+factor	2	0.57	3.64	0.44	1.17	0.31	1.22	-0.37	1.07	-0.47	3.02	-0.95	1.18
model)	3	0.55	3.68	0.52	1.12	0.09	1.12	-0.42	1.28	-0.48	3.05	-0.96	1.47
	4	0.54	3.69	1.00	1.14	0.08	1.08	-0.86	1.54	-0.47	3.02	-0.16	0.60
PCA	1	2.96	3.85	3.19	1.64	1.20	1.64	-1.29	2.36	3.02	3.10	-1.32	1.34
(factor model	2	7.44	3.61	2.74	1.15	2.27	1.21	3.63	1.49	5.35	3.01	-1.06	1.11
only)	3	7.65	3.62	6.56	2.44	5.39	2.37	4.18	2.35	5.49	3.04	-0.36	1.13
	4	8.01	3.60	6.66	2.31	5.55	2.27	4.28	2.36	4.94	3.03	0.32	1.08
Block-PCA	1	-0.30	4.26	1.65	1.52	-0.05	1.53	-4.41	2.50	-1.24	3.44	-1.32	1.19
(factor model	2	4.25	6.65	0.38	1.46	0.33	1.31	1.16	1.41	7.79	8.32	-1.18	1.09
only)	3	0.18	6.14	0.32	1.41	0.36	1.29	1.36	1.74	2.84	6.37	-0.91	1.31
	4	1.69	6.92	0.30	1.44	0.33	1.41	1.26	1.83	4.80	6.68	0.31	1.15
TWFE (FE only)		0.56	3.73	1.73	1.64	0.02	1.62	-2.94	2.29	-0.47	3.04	-1.28	1.19

This table reports the point estimations (ATT) and standard errors (SE) of ATT with different estimators. We show the results of wi-PCA, PCA, and block-PCA estimators with respectively $k = 1, 2, 3, 4$ latent factor(s). The standard errors are calculated by the resampling bootstrap method with $B = 1000$. The full name and treatment time of each treated state is CA: California (Jan. 1, 2018), MI: Michigan (Dec. 1, 2019), IL: Illinois (Jan. 1, 2020), MA: Massachusetts (Nov. 20, 2018), NV: Nevada (Jul. 1, 2017) and ME: Maine (Oct. 9, 2020).

Figure 2: Observed time series and estimated control time series of per-store beer sales



(a) Massachusetts

(b) Nevada

These figures show observed time series and estimated control time series of beer sales per store for Massachusetts and Nevada. We use wi-PCA estimator with $k = 2$ factors and PCA and block-PCA estimators with $k = 4$ factors. The grey vertical curve denotes the time of treatment.

is essentially identical for 1 to 4 factors. In contrast, for block-PCA and PCA the magnitude varies drastically and even the sign of the ATT can change for different number of factors. The choice of the unknown number of factors can result in drastically different economic conclusions for these benchmark estimators.

Figure 2 provides an explanation for our findings and shows the observed time series and estimated control time series of Massachusetts and Nevada. We use wi-PCA with $k = 2$ factors and PCA and block-PCA with $k = 4$ factors to have a fair comparison in terms of the degrees of freedom. We start with the right subfigure for Nevada. Before the treatment, the orange line representing the control outcomes with wi-PCA is close to the observed blue time-series. After the treatment the orange line continues to be close to the blue line, and therefore we do not find a significant treatment effect with wi-PCA. The fact that wi-PCA is very close to the control and treatment observations suggests that the same model can describe the full sample. In contrast, the purple line for PCA is far away from the blue line on the control and on the treatment data. This indicates that finding a treatment effect with PCA is due to a bad model fit. Similarly, the green line for block-PCA seems to deviate because it does not capture the variation in the time-series well. We now turn to the left plot for Massachusetts. While obtaining similar results for PCA, we want to draw attention to the red line of TWFE. A pure fixed effect model suffers from an omitted variable bias and neglects relevant time-series variation on the control and treatment data. Hence, it might imply spurious treatment effects due to this omitted variable bias.

Figure 2 suggests that the discrepancy of other methods seems to indicate either an omitted variable bias or overfitting of noise on the control data and hence a high variance on the treatment data. In summary, the economic conclusions can largely differ depending on the estimation approach, and the accuracy and generality of wi-PCA indicates that for this empirical study we do not find evidence for a significant treatment effect.

9 Conclusion

In this paper, we study the imputation and inference for large-dimensional non-stationary panels with general missing patterns. We propose a novel method, Within-Transform-PCA (wi-PCA), to estimate an approximate latent factor model and non-stationary two-way fixed effects. The general missing patterns can depend on both the latent factor model and two-way fixed effects. We show the consistency and provide entry-wise inference for the imputed values with wi-PCA. Including general fixed effects, which can affect missingness, is particularly important for applications in causal inference for panels. In an empirical study on the liberalization of marijuana, we illustrate that wi-PCA yields more accurate estimates of treatment effects and more credible economic conclusions compared to the benchmark estimators.

References

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

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

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2021): "Matrix completion methods for causal panel data models," *Journal of the American Statistical Association*, pp. 1–15.

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

——— (2009): "Panel data models with interactive fixed effects," *Econometrica*, 77(4), 1229–1279.

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

——— (2021): "Matrix completion, counterfactuals, and factor analysis of missing data," *Journal of the American Statistical Association*, 116(536), 1746–1763.

——— (2023): "Approximate factor models with weaker loadings," *Journal of Econometrics*, 235(2), 1893–1916.

Bańbura, M., and M. Modugno (2014): "Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data," *Journal of Applied Econometrics*, 29(1), 133–160.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust differences-in-differences estimates?," *The Quarterly Journal of Economics*, 119(1), 249–275.

Cahan, E., J. Bai, and S. Ng (2023): "Factor-based imputation of missing values and covariances in panel data of large dimensions," *Journal of Econometrics*, 233(1), 113–131.

Chen, W., and M. Bayati (2021): "Learning to Recommend Using Non-Uniform Data," *arXiv preprint arXiv:2110.11248.*

Chen, Y., J. Fan, C. Ma, and Y. Yan (2019): "Inference and uncertainty quantification for noisy matrix completion," *Proceedings of the National Academy of Sciences*, 116(46), 22931–22937.

Chernozhukov, V., C. Hansen, Y. Liao, and Y. Zhu (2023): "Inference for low-rank models," *The Annals of Statistics*, 51(3), 1309–1330.

Choi, J., and M. Yuan (2023): "Matrix Completion When Missing Is Not at Random and Its Applications in Causal Panel Data Models," *arXiv preprint arXiv:2308.02364.*

Duan, J., M. Pelger, and R. Xiong (2023): "Target PCA: Transfer learning large dimensional panel data," *Journal of Econometrics*, p. 105521.

Fan, J., Y. Liao, and M. Mincheva (2011): "High dimensional covariance matrix estimation in approximate factor models," *Annals of Statistics*, 39(6), 3320.

——— (2013): "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4), 603–680.

Guha, R., and S. Ng (2019): "A machine learning analysis of seasonal and cyclical sales in weekly scanner data," in *Big Data for 21st Century Economic Statistics*. University of Chicago Press.

Jin, S., K. Miao, and L. Su (2021): “On factor models with random missing: EM estimation, inference, and cross validation,” *Journal of Econometrics*, 222(1), 745–777.

Lettau, M., and M. Pelger (2020): “Estimating latent asset-pricing factors,” *Journal of Econometrics*, 218(1), 1–31.

Li, K. T., and G. P. Sonnier (2023): “Statistical Inference for the Factor Model Approach to Estimate Causal Effects in Quasi-Experimental Settings,” *Journal of Marketing Research*, 60(3), 449–472.

Mazumder, R., T. Hastie, and R. Tibshirani (2010): “Spectral regularization algorithms for learning large incomplete matrices,” *Journal of Machine Learning Research*, 11(Aug), 2287–2322.

Negahban, S., and M. J. Wainwright (2011): “Estimation of (near) low-rank matrices with noise and high-dimensional scaling,” *The Annals of Statistics*, pp. 1069–1097.

——— (2012): “Restricted strong convexity and weighted matrix completion: Optimal bounds with noise,” *Journal of Machine Learning Research*, 13(May), 1665–1697.

Ng, S., and S. Scanlan (2024): “Constructing high frequency economic indicators by imputation,” *The Econometrics Journal*, 27(1), C1–C30.

Onatski, A. (2012): “Asymptotics of the principal components estimator of large factor models with weakly influential factors,” *Journal of Econometrics*, 168(2), 244–258.

Pelger, M., and R. Xiong (2021a): “Interpretable sparse proximate factors for large dimensions,” *Journal of Business & Economic Statistics*, pp. 1–23.

——— (2021b): “State-varying factor models of large dimensions,” *Journal of Business & Economic Statistics*, pp. 1–50.

Stock, J. H., and M. W. Watson (2002): “Forecasting using principal components from a large number of predictors,” *Journal of the American Statistical Association*, 97(460), 1167–1179.

Su, L., and X. Wang (2017): “On time-varying factor models: Estimation and testing,” *Journal of Econometrics*, 198(1), 84–101.

Urga, G., and F. Wang (2022): “Estimation and inference for high dimensional factor model with regime switching,” *arXiv preprint arXiv:2205.12126*.

Xia, D., and M. Yuan (2021): “Statistical inferences of linear forms for noisy matrix completion,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(1), 58–77.

Xiong, R., A. Chin, and S. J. Taylor (2024): “Data-Driven Switchback Experiments: Theoretical Tradeoffs and Empirical Bayes Designs,” *arXiv preprint arXiv:2406.06768*.

Xiong, R., and M. Pelger (2023): “Large dimensional latent factor modeling with missing observations and applications to causal inference,” *Journal of Econometrics*, 233(1), 271–301.

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

A Inconsistency of Symmetric Fixed Effect Estimators: An Illustrative Example

In this subsection, we provide an illustrative example to demonstrate the importance of subtracting $\tilde{\xi}_t$ in the estimation of α_i . In this toy example, we consider a block-missing pattern, where each unit has 50% chance of being observed for all time periods or of being observed for only the first half of times.

For simplicity, we assume that[12]

$$\begin{aligned} &\xi_1 = \cdots = \xi_{T/2} = 1, \qquad \xi_{T/2+1} = \cdots = \xi_T = -1, \\ &\Lambda_i \overset{i.i.d.}{\sim} (0, \Sigma_\Lambda), \qquad F_t \overset{i.i.d.}{\sim} (0, \Sigma_F), \qquad \epsilon_{it} \overset{i.i.d.}{\sim} (0, \sigma_\varepsilon^2)\,. \end{aligned}$$

In this illustrative example, the observation probability equals $p_{it} = P(W_{it} = 1) = 1$ if $t \leq T/2$ (with $\xi_t = 1$) and $p_{it} = 1/2$ otherwise (with $\xi_t = -1$), that is, the observation pattern is correlated with the time fixed effects.

Our key point is to illustrate what goes wrong when we do not properly account for the time fixed effects. If we do not subtract $\tilde{\xi}_t$ from the estimator for α_i and instead use an estimator symmetric to the one of $\tilde{\xi}_t$ with inverse probability weighting, then the estimation error is

$$\begin{aligned} \tilde{\alpha}_i^{\text{sym}} - \alpha_i &= \left(\frac{1}{T} \sum_{t=1}^T \frac{W_{it}}{p_{it}} \right)^{-1} \frac{1}{T} \sum_{t=1}^T \frac{W_{it}}{p_{it}} Y_{it} - \tilde{\mu} - \alpha_i \\ &= \underbrace{(\mu - \tilde{\mu})}_{o_p(1)} + \left(\frac{1}{T} \sum_{t=1}^T \frac{W_{it}}{p_{it}} \right)^{-1} \left[\frac{1}{T} \sum_{t=1}^T \frac{W_{it}}{p_{it}} \xi_t + \underbrace{\frac{1}{T} \sum_{t=1}^T \frac{W_{it}}{p_{it}} \Lambda_i^\top F_t}_{o_p(1)} + \underbrace{\frac{1}{T} \sum_{t=1}^T \frac{W_{it}}{p_{it}} \epsilon_{it}}_{o_p(1)} \right]. \end{aligned}$$

Straightforward calculations yield

$$\left(\frac{1}{T} \sum_{t=1}^T \frac{W_{it}}{p_{it}} \right)^{-1} \frac{1}{T} \sum_{t=1}^T \frac{W_{it}}{p_{it}} \xi_t = \begin{cases} -1/3 & \text{if unit } i \text{ is fully observed,} \\ 1 & \text{otherwise,} \end{cases} \tag{7}$$

implying that $\tilde{\alpha}_i^{\text{sym}}$ is a biased estimator of α_i and the bias does not diminish as T grows.[13] This example illustrates why we propose to remove the bias by subtracting $\tilde{\xi}_t$ in the estimation of $\tilde{\alpha}_i^{\text{sym}}$.

[12] For simplicity, assume $T/2$ is an integer.

[13] Note that this issue does not occur in the estimation of ξ_t. The reason is that we can show that

$$\left(\frac{1}{N} \sum_{j=1}^N \frac{W_{jt}}{p_{jt}} \right)^{-1} \frac{1}{N} \sum_{i=1}^N \frac{W_{it}}{p_{it}} \alpha_i \xrightarrow{p} 0\,.$$

This follows from Assumption 1.3, stating that W_{it} and W_{jt} are conditionally independent for any $i \neq j$, and Assumption 3.2. In more detail, the proof relies on the key property that $\frac{1}{N^2} \text{Cov} \left(\sum_{i=1}^N \frac{W_{it}}{p_{it}} \alpha_i, \sum_{i=1}^N \frac{W_{it}}{p_{it}} \alpha_i \right) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \frac{1}{p_{it} p_{jt}} \mathbb{E}[\mathbb{E}[W_{it} \alpha_i W_{jt} \alpha_j \mid I_{it}, I_{jt}]] = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j + O(\frac{1}{N}) = O(\frac{1}{N})\,.$

After subtracting $\tilde{\xi}_t$, we do not need to use an inverse probability weighting to adjust for the selection bias because W_{it} does not depend on the remaining temporal information (that is F_t and ϵ_t). Therefore, we can use the equal weights M_{it}^{α} for all observed entries in Equation (3).

B Feasible Variance Estimators for ATT

In this section, we provide the detailed description of the resampling bootstrap method in Section 5.2, and discuss how to extend it to allow for cross-sectional dependency through a block resampling bootstrap method.

B.1 Resampling Bootstrap Method

In Section 5.2, we show that under the asymptotic normality, the valid confidence interval (CI) for τ_i is:

$$\mathbb{P}\left(\tau_i \in \left[\hat{\tau}_i - z_{1-\alpha/2}\sqrt{V_i}, \hat{\tau}_i + z_{1-\alpha/2}\sqrt{V_i}\,\right]\right) \overset{\text{asympt}}{\sim} 1-\alpha\,,$$

where V_i is the asymptotic variance and z_α is the α-quantile of standard normal distribution.

We propose to estimate V_i with a resampling bootstrap procedure, which is easy to use in spite of the very complex form of V_i. To implement this method, we assume that conditional on the factors and time fixed effects, the time series and its observation pattern of each unit are drawn independently from the same distribution P. To estimate the conditional variance $\text{Var}(\hat{\tau}_i \mid \xi, F)$, we need to resample the error for target unit i. This requires the assumption that the residual time series are homoscedastic across units.

We describe the resampling bootstrap method in the following. For brevity, we denote the observed time series in the control panel as $\tilde{Y}_i^{(\text{ct})} := (Y_{i1}^{(\text{ct})}, \cdots, Y_{iT}^{(\text{ct})}) \odot (W_{i1}, \cdots, W_{iT})$.

Resampling Bootstrap Algorithm:

1. Construct sampling distribution of the idiosyncratic errors $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{iT})$ of the treated unit i.

   Start a loop for $j \in \{\text{control units}\}$:

   - Mask unit j's observations as if it had the same treatment pattern as unit i, i.e. $\tilde{Y}_j^{(\text{ct})'} = Y_j^{(\text{ct})} \odot W_i$. Construct a new control panel $\tilde{Y}^{(\text{ct})'} = (\tilde{Y}_{[N]\backslash j}^{(\text{ct})}; \tilde{Y}_j^{(\text{ct})'})$.
   - Estimate $\tilde{C}'_{it}$ with wi-PCA. Calculate residual time series $\tilde{\epsilon}_j^{(i)} = Y_j^{(\text{ct})} - \tilde{C}'_j$.

   Collect $\tilde{\epsilon}^{(i)} := \{\tilde{\epsilon}_j^{(i)} : j \in \text{control units}\}$.
2. Obtain estimates of τ_i from bootstrap samples.

   Start a loop for $b \in \{1, \cdots, B\}$:

   - Randomly sample $N-1$ indices $i_1, \cdots, i_{N-1}$ from $[N] \setminus i$ with replacement and randomly sample $\tilde{\epsilon}_b^{(i)}$ from $\tilde{\epsilon}^{(i)}$. Construct a new control panel $\tilde{Y}_b^{(\text{ct})*} = (\tilde{Y}_{i_1}^{(\text{ct})}; \cdots; \tilde{Y}_{i_{N-1}}^{(\text{ct})}; \tilde{C}_i + \tilde{\epsilon}_b^{(i)})$ with the target unit i in the last row.

    - Estimate $\tilde{C}^*_{b,it}$ with wi-PCA. Calculate $\hat{\tau}^*_{i,b} = T^{-1}_{i,tr}\sum_{t\in\mathcal{T}_{i,tr}}(\tilde{C}_{it} + \tilde{\epsilon}^{(i)}_{b,t} - \tilde{C}^*_{b,Nt})$.[14]

3. Calculate the variance as $\mathrm{Var}_{\hat{P}}(\hat{\tau}^*_i) = B^{-1}\sum_{b=1}^{B}\left(\hat{\tau}^*_{i,b} - \bar{\tau}_i\right)^2$, where $\bar{\tau}_i = \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}^*_{i,b}$.

We show the good finite sample performance of the CI obtained from this procedure in Appendix B.3. We also show the benefits of leveraging the asymptotic normal distribution and only sampling the variance, instead of sampling the complete distribution. The bootstrap procedure assumes that α_j and Λ_j are independent in j and ε_{jt} is i.i.d. in j. This assumption can be relaxed by sampling blocks of correlated units. We provide the details of block resampling procedure in the next subsection.

B.2 Block Resampling Bootstrap Method

We discuss how to extend the resampling bootstrap algorithm in Section 5.2 to allow for cross-sectional dependency in the idiosyncratic errors if the the off-diagonal elements of the covariance matrix of the errors have a sparse structure. Specifically, we assume that the number of non-zero off-diagonal entries grows slowly with respect to N and T. Under this assumption, we can use a block resampling bootstrap method to estimate the variance, which treats correlated units as an entirety for the resampling. Intuitively, we apply a similar bootstrap algorithm as before, but apply it to uncorrelated blocks instead of individual units.

In some applications we might know these blocks, for example, based on geographical or industry clusters. Alternatively, we can identify blocks statistically with threshold techniques. Specifically, we first estimate the common components $\tilde{C}_{it}$ of the control panel $Y^{(\mathrm{ct})}$ and derive the idiosyncratic error $\tilde{\epsilon}_{it} = Y^{(\mathrm{ct})}_{it} - \tilde{C}_{it}$ for each entry. Based on this, we calculate the empirical error covariance matrix

$$\tilde{\Sigma}_{\epsilon} = \frac{1}{T}\sum_{t=1}^{T}(\tilde{\epsilon}_t - \bar{\epsilon})(\tilde{\epsilon}_t - \bar{\epsilon})^\top \in \mathbb{R}^{N\times N},$$

where $\tilde{\epsilon}_t = (\tilde{\epsilon}_{t1}, \cdots, \tilde{\epsilon}_{tN})^\top$, $\bar{\epsilon} = T^{-1}\sum_{t=1}^{T}\tilde{\epsilon}_t$. Then, we can apply a thresholding technique like in Fan, Liao, and Mincheva (2011) to $\tilde{\Sigma}_{\epsilon}$ to obtain a sparse error covariance matrix $\tilde{\Sigma}^{\mathcal{T}}_{\epsilon}$. We can rearrange the cross-sectional units of $\tilde{\Sigma}^{\mathcal{T}}_{\epsilon}$ to obtain a block-diagonal matrix.

Given the block structure, the block resampling bootstrap method works as follows:

Block Resampling Bootstrap Algorithm:

1. Get the block structure of panel $Y^{(\mathrm{ct})}$. Group the units into N_0 groups $g_1, \cdots, g_{N_0}$. Denote by $g^{(i)}$ the group containing the target treated unit i.
2. Construct sampling distribution of of the idiosyncratic errors of group $g^{(i)}$.
   Start a loop for $j \in \{$control groups $\cup$ groups with the same error structure as $g^{(i)}\}$:
    - Mask group j's observations as if it had the same treatment pattern as group $g^{(i)}$, i.e. $\tilde{Y}^{(\mathrm{ct})\prime}_{g_j} = Y^{(\mathrm{ct})}_{g_j} \odot W_{g^{(i)}}$. Construct a new control panel $\tilde{Y}^{(\mathrm{ct})\prime} = (\tilde{Y}^{(\mathrm{ct})}_{[N]\setminus g_j}; \tilde{Y}^{(\mathrm{ct})\prime}_{g_j})$.

[14] The treated outcome for unit i at time t can be drawn as $\tilde{C}_{it} + \tilde{\epsilon}^{(i)}_{b,t} + \tau_{it}$. However, since τ_{it} would not affect the variance of $\hat{\tau}_i$, we directly use $\tilde{C}_{it} + \tilde{\epsilon}^{(i)}_{b,t}$ $(\tau_{it} = 0)$ for the bootstrap treated outcome in estimating the variance.

- Estimate $\tilde{C}'$ from $\tilde{Y}^{(\mathrm{ct})'}$ with wi-PCA. Calculate the residual time series $\tilde{\epsilon}^{(i)}_{g_j} = Y^{(\mathrm{ct})}_{g_j} - \tilde{C}'_{g_j}$.

Collect $\tilde{\epsilon}^{(i)} := \{\tilde{\epsilon}^{(i)}_{g_j}\}$.

3. Obtain estimates of τ_i from bootstrap samples.
   Start a loop for $b \in \{1, \cdots, B\}$:

   - Randomly sample groups $g_{j_1}, \cdots, g_{j_d}$ from the $N_0 - 1$ groups with replacement, where d is the smallest number such that the number of sampled units $|g_{j_1}| + \cdots + |g_{j_d}| \geq N - |g^{(i)}|$. Then, randomly remove $|g_{j_1}| + \cdots + |g_{j_d}| - (N - |g^{(i)}|)$ units. Randomly sample $\tilde{\epsilon}^{(i)}_b$ from $\tilde{\epsilon}^{(i)}$. Combine the sampled units with $\tilde{C}_{g^{(i)}} + \tilde{\epsilon}^{(i)}_b$ to construct a new control panel $\tilde{Y}^{(\mathrm{ct})*}_b$.
   - Estimate $\tilde{C}^*_b$ from $\tilde{Y}^{(\mathrm{ct})*}_b$ with wi-PCA and calculate $\hat{\tau}^*_{i,b} = T^{-1}_{i,tr} \sum_{t \in \mathcal{T}_{i,tr}} (\tilde{C}_i + \tilde{\epsilon}^{(i)}_b - \tilde{C}^*_{b,(i)t})$, where (i) denotes the position of unit i.

4. Calculate the variance as

$$\mathrm{Var}_{\hat{P}}(\hat{\tau}^*_i) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\tau}^*_{i,b} - \bar{\tau}_i\right)^2, \quad \text{where } \bar{\tau}_i = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}^*_{i,b}.$$

$\hat{V}_i = \mathrm{Var}_{\hat{P}}(\hat{\tau}^*_i)$ is the estimate of $\mathrm{Var}(\hat{\tau}_i | \xi, F)$.

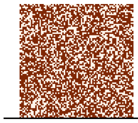
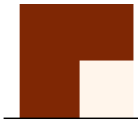
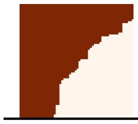
The resampling bootstrap algorithm proposed in Section 5.2 is a special case of this block resampling bootstrap algorithm with $N_0 = N$. The performance of the block resampling bootstrap algorithm increases in the number of groups N_0.

B.3 Bootstrapped Asymptotic Variance

In this section, we demonstrate the good finite sample performance of our bootstrap procedure from Section 5.2 for estimating the asymptotic variance and constructing asymptotically valid confidence intervals for the average treatment effect on the treated. In particular, we show that the bootstrapped variances can still be reliable estimates even if some of the assumptions are violated, and that leveraging the normal distribution of the estimator provides better coverage than estimating the complete distribution with the bootstrap.

We generate the control panel $Y^{(\mathrm{ct})}$ from the same one-factor model with two-way fixed effects specified as in Section 7, that is, $Y^{(\mathrm{ct})}_{it} = \mu + \alpha_i + \xi_t + \Lambda_i F_t + \epsilon_{it}$, where $\mu = 1$, $\alpha_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $\Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and $F_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. We consider both stationary and non-stationary time fixed effects ξ_t and both homoscedastic and heteroscedastic idiosyncratic errors. The stationary time fixed effects follow $\xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, while non-stationary time fixed effects include a trend and are generated as $\xi_t = 0.05t + \mathcal{N}(0,1)$ for any t. The homoscedastic errors follow $\epsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,4)$ for any i and t. For heteroscedastic errors, we let $\epsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,4)$ for the target unit i, while $\epsilon_{jt} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_{\epsilon,j})$ with $\sigma_{\epsilon,j} \overset{i.i.d.}{\sim} \mathrm{Uniform}(1,3)$ for units other than i. We assume the treated panel $Y^{(\mathrm{tr})} = Y^{(\mathrm{ct})}$, i.e. the treatment effect is zero for any unit i at any time period t. We consider the three treatment patterns in Section 7: treated-at-random, simultaneous treatment adoption, and staggered treatment adoption.

Table A.1: Coverage of constructed confidence intervals (CI)

		Homoscedastic Error			Heteroscedastic Error		
	ξ_t	95%	90%	80%	95%	90%	80%
	S	95.0%	89.8%	80.0%	94.4%	89.8%	80.5%
	N	95.0%	89.8%	80.0%	94.4%	89.8%	80.5%
	S	94.3%	89.2%	79.5%	94.1%	89.6%	80.6%
	N	94.3%	89.2%	79.5%	94.1%	89.6%	80.6%
	S	94.1%	89.2%	78.1%	94.2%	89.4%	79.7%
	N	94.1%	89.2%	78.1%	94.2%	89.4%	79.7%

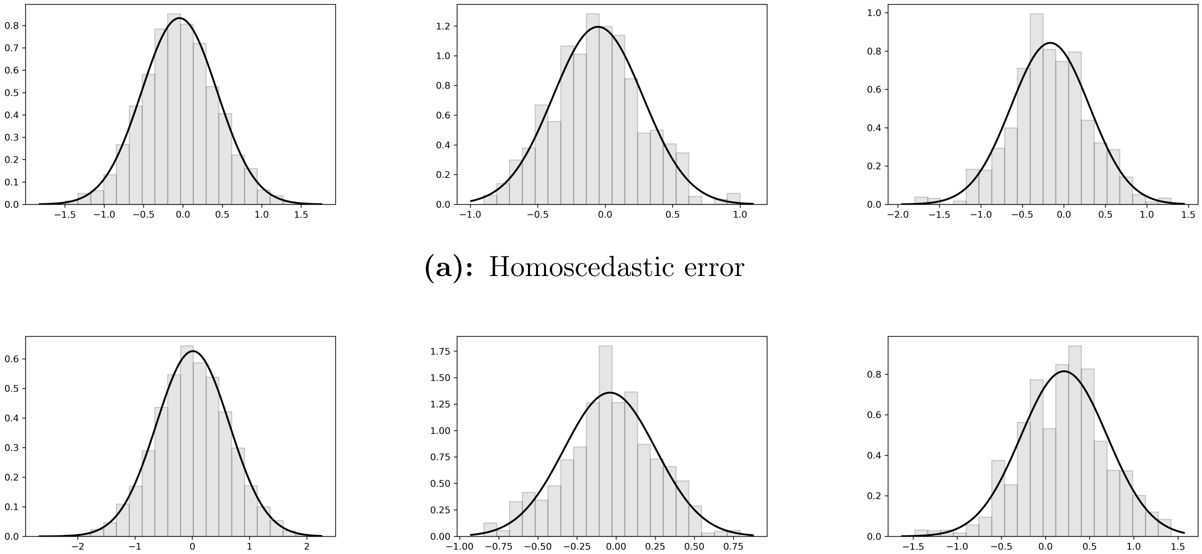
This table reports the coverage of CI for ATT using wi-PCA with unknown observation probability. The figures on the left show the observation patterns for the control panel with shaded entries indicating observed entries and unshaded entries indicating missing entries. Each observation pattern corresponds to two rows with either stationary time fixed effects (S) or non-stationary time fixed effects (N). We set $N = T = 100$, $B = 100$ and run 1000 simulations for each setup.

Table A.1 shows the good finite sample coverage of the confidence intervals for τ_i for all the three treatment patterns. In the setting with homoscedastic errors, our theoretical results guarantee that our constructed confidence intervals are asymptotically valid. However, for the case with heteroscedastic errors, the confidence intervals maintain good coverage, which demonstrates the robustness of the bootstrap procedure to minor deviations from the i.i.d assumption of the error time series.

Additionally, Figure A.1 shows the histograms of bootstrap ATTs for the three treatment patterns under both homoscedastic and heteroscedastic settings. In both settings, the bootstrap distributions closely resemble normal distributions.

We also show the benefits of sampling only the variance and leveraging the theoretical normal distribution, instead of sampling the complete distribution in the second step of the bootstrap procedure. Table A.2 compares the coverage of the confidence intervals constructed by two methods, where the first method ("bootstrapped variance with normal distribution") is our proposed bootstrap method that samples only the variance and uses the asymptotic normal distribution implied by our theory, while the second method ("bootstrapped distribution") uses the complete distribution of the bootstrap estimators $\hat{\tau}^*_{i,b}$ for $b = 1, \ldots, B$. In the second method, we construct confidence intervals as $[\hat{\tau}^*_{i,\alpha/2}, \hat{\tau}^*_{i,1-\alpha/2}]$, where $\hat{\tau}^*_{i,\alpha/2}$ and $\hat{\tau}^*_{i,\alpha/2}$ are respectively the $\alpha/2$ and $1 - \alpha/2$ quantiles of the bootstrap estimators $\hat{\tau}^*_{i,b}$. Table A.2 shows that the method that samples only the variance and leverages the theoretical normal distribution achieves better coverage than the method that uses the quantiles of the entire distribution.

Figure A.1: Histograms of bootstrap ATT's



(a): Homoscedastic error

(b): Heteroscedastic error

These figures show histograms of bootstrap ATT's of the three treatment patterns (from left to right): treated-at-random, simultaneous treatment adoption, and staggered treatment adoption. We set $B = 2000$ for each setup.

C Block-PCA Estimator

In this section, we briefly review the generalized synthetic control (GSC) estimator proposed in Xu (2017), which we refer to as "block-PCA". The GSC estimator with latent confounders also employs a latent factor model to describe the control outcomes, that is, it assumes the model

$$Y_{it} = \Lambda_i^\top F_t + \epsilon_{it} \tag{8}$$

for units $i = 1, \cdots, N$ and time periods $t = 1, \cdots, T$.

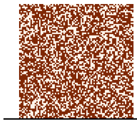
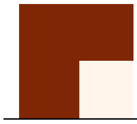
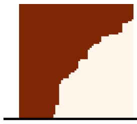
GSC estimator requires a block of fully observed control units. It estimates the model in three steps. First, it estimates the factor model using only the control units and obtain the latent factors $\tilde{F}$ and factor loadings for the control units $\tilde{\Lambda}_{co}$. In the second step, it estimates the factor loadings for each treated units by regressing the observed control outcomes on estimated factors $\tilde{F}$ using the pretreatment period. Finally, in the last step, it estimates the counterfactual control outcomes by combining the estimated factors and loadings, as $\tilde{C}_{it} = \tilde{\Lambda}_i^\top \tilde{F}_t$.

The GSC estimator with observed covariates X takes the form:

$$Y_{it} = \beta^\top X_{it} + \Lambda_i^\top F_t + \epsilon_{it}. \tag{9}$$

The coefficients $\tilde{\beta}$ on the observables are estimated in the first step, using only the fully observed

Table A.2: Comparison between confidence intervals (CI) constructed by different methods

		Bootstrapped Variance with Normal Distribution			Bootstrapped Distribution		
	ξ_t	95%	90%	80%	95%	90%	80%
	S	95.0%	89.8%	80.0%	93.4%	88.6%	78.8%
	N	95.0%	89.8%	80.0%	93.4%	88.6%	78.8%
	S	94.3%	89.2%	79.5%	92.4%	87.5%	80.4%
	N	94.3%	89.2%	79.5%	92.4%	87.5%	80.4%
	S	94.1%	89.2%	78.1%	92.6%	87.2%	76.4%
	N	94.1%	89.2%	78.1%	92.6%	87.2%	76.4%

This table reports the coverage of CI for ATT where we either bootstrap the variance and use the asymptotic normal distribution or bootstrap the full distribution. The figures on the left show the observation patterns for the control panel with shaded entries indicating observed entries and unshaded entries indicating missing entries. Each observation pattern corresponds to two rows with either stationary time fixed effects (S) or non-stationary time fixed effects (N). We consider the setting with homoscedastic errors. We set $N = T = 100$, $B = 100$ and run 1000 simulations for each setup.

control units.

D Empirical Study: Synthetic Treatment Assignment

In Section 8.1, we compare the performance of different estimators through synthetic treatment assignments on the control panel that consists of the 39 control states. Specifically, we generate synthetic treatment patterns for the control panel and mask the corresponding control outcomes in the data as if they were treated. Then, we compare the estimation accuracy of different estimators on these synthetic treatments.

Tables A.3-A.5 provide detailed robustness results with a larger number of factors for wi-PCA, PCA and block-PCA. For conciseness, Table 3 in the main text shows the results for only three latent factors, but all the results from the main text extend to up to 8 factors.

Table A.3: Results for the uniformly random treatment pattern

		RMSE_Y	$\lvert\text{Bias}_{\text{ATT}}\rvert$	RMSE_{ATT}
wi-PCA (FE+factor model)	k=1	2.207	0.156	1.236
	k=2	1.933	0.139	1.110
	k=3	2.012	0.158	1.179
	k=4	2.033	0.167	1.206
	k=5	2.026	0.173	1.214
	k=6	2.051	0.177	1.258
	k=7	2.061	0.177	1.291
	k=8	2.067	0.177	1.303
PCA (factor model only)	k=1	2.525	0.216	1.677
	k=2	2.642	0.586	1.755
	k=3	2.952	0.979	2.032
	k=4	3.302	1.357	2.368
	k=5	3.598	1.637	2.615
	k=6	3.764	1.806	2.782
	k=7	3.883	1.929	2.908
	k=8	3.978	2.022	3.002
Block-PCA (factor model only)	k=1	-	-	-
	k=2	-	-	-
	k=3	-	-	-
	k=4	-	-	-
	k=5	-	-	-
	k=6	-	-	-
	k=7	-	-	-
	k=8	-	-	-
TWFE (FE only)		2.576	0.218	1.604

This table reports the complete results for the uniformly random treatment pattern for a larger number of factors in wi-PCA, PCA, and block-PCA estimators. Note that the block-PCA method is infeasible here because we do not have control units with fully observed data. We simulate the masking 100 times and report the average results. The units are 100 dollars per store.

Table A.4: Results for the simultaneous treatment adoption pattern

		RMSE_Y	$\lvert\text{Bias}_{\text{ATT}}\rvert$	RMSE_{ATT}
wi-PCA (FE+factor model)	k=1	2.424	0.255	0.995
	k=2	2.090	0.275	1.029
	k=3	1.942	0.262	1.056
	k=4	1.909	0.284	1.084
	k=5	1.955	0.299	1.138
	k=6	2.006	0.299	1.166
	k=7	2.025	0.301	1.177
	k=8	2.039	0.304	1.184
PCA (factor model only)	k=1	2.374	0.825	1.672
	k=2	4.605	3.760	4.076
	k=3	5.450	4.190	4.523
	k=4	5.123	4.131	4.459
	k=5	5.038	4.138	4.467
	k=6	5.009	4.147	4.478
	k=7	5.005	4.152	4.483
	k=8	5.012	4.151	4.482
Block-PCA (factor model only)	k=1	2.307	0.450	1.568
	k=2	2.246	0.318	1.128
	k=3	2.125	0.304	1.176
	k=4	2.085	0.284	1.218
	k=5	2.037	0.289	1.228
	k=6	1.968	0.282	1.209
	k=7	1.940	0.272	1.208
	k=8	1.924	0.286	1.204
TWFE (FE only)		2.445	0.415	1.528

This table reports the complete results for the simultaneous treatment adoption pattern for a larger number of factors in wi-PCA, PCA, and block-PCA estimators. We simulate the masking 100 times and report the average results. The units are 100 dollars per store.

Table A.5: Results for the staggered treatment adoption pattern

		RMSE_Y	$\lvert\text{Bias}_{\text{ATT}}\rvert$	RMSE_{ATT}
wi-PCA (FE+factor model)	k=1	4.299	0.751	2.893
	k=2	3.923	0.475	2.770
	k=3	4.097	0.574	2.911
	k=4	4.185	0.642	2.985
	k=5	4.241	0.692	3.068
	k=6	4.270	0.733	3.102
	k=7	4.289	0.735	3.114
	k=8	4.278	0.726	3.096
PCA (factor model only)	k=1	5.004	1.445	4.119
	k=2	6.766	4.472	5.430
	k=3	7.463	5.246	6.063
	k=4	7.715	5.437	6.242
	k=5	7.794	5.558	6.369
	k=6	7.900	5.647	6.467
	k=7	7.985	5.719	6.539
	k=8	8.016	5.764	6.580
Block-PCA (factor model only)	k=1	4.877	0.902	3.884
	k=2	7.163	1.256	5.114
	k=3	5.934	1.035	4.212
	k=4	6.237	1.015	4.456
	k=5	6.243	1.009	4.434
	k=6	6.359	1.008	4.521
	k=7	6.617	1.039	4.731
	k=8	6.725	1.034	4.793
TWFE (FE only)		4.603	0.593	3.498

This table reports the complete results for the staggered treatment adoption pattern for a larger number of factors in wi-PCA, PCA, and block-PCA estimators. We simulate the masking 100 times and report the average results. The units are 100 dollars per store.

Internet Appendix to
Causal Inference for Large Dimensional Non-Stationary Panels with Two-Way Endogenous Treatment and Latent Confounders

Junting Duan[*] Markus Pelger[†] Ruoxuan Xiong[‡]

July 2, 2024

Abstract

This Internet Appendix collects all the supplementary simulation results, empirical results, and the detailed proofs for all the theoretical statements in the main text.

[*] Stanford University, Department of Management Science & Engineering, Email: duanjt@stanford.edu.

[†] Stanford University, Department of Management Science & Engineering, and NBER, Email: mpelger@stanford.edu.

[‡] Emory University, Department of Quantitative Theory and Methods, Email: ruoxuan.xiong@emory.edu.

Contents

IA.A Simulation

IA.A.1 Comparison with Benchmark Estimators

Table IA.1: Relative MSE of common components for missing at random with stationary time fixed effects

			wi-PCA		PCA			Block-PCA			TWFE
			(FE+factor model)		(factor model only)			(factor model only)			(FE only)
Parameters		$\mathcal{S}$	known	unknown	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	
$p = 0.4$	$\sigma_\epsilon = 1$	obs	0.041	0.041	0.352	0.138	0.075	-	-	-	0.252
		miss	**0.050**	**0.050**	0.386	0.167	0.101	-	-	-	0.278
		all	0.046	0.047	0.372	0.155	0.091	-	-	-	0.268
	$\sigma_\epsilon = 2$	obs	0.121	0.122	0.390	0.221	0.214	-	-	-	0.290
		miss	**0.133**	**0.133**	0.425	0.254	0.253	-	-	-	0.319
		all	0.128	0.129	0.411	0.241	0.238	-	-	-	0.308
	$\sigma_\epsilon = 3$	obs	0.273	0.274	0.464	0.388	0.495	-	-	-	0.352
		miss	**0.288**	0.289	0.503	0.428	0.551	-	-	-	0.383
		all	0.282	0.283	0.487	0.412	0.529	-	-	-	0.370
$p = 0.6$	$\sigma_\epsilon = 1$	obs	0.024	0.024	0.347	0.119	0.042	-	-	-	0.252
		miss	**0.028**	**0.028**	0.369	0.134	0.052	-	-	-	0.270
		all	0.026	0.026	0.355	0.125	0.046	-	-	-	0.259
	$\sigma_\epsilon = 2$	obs	0.076	0.077	0.370	0.172	0.128	-	-	-	0.276
		miss	**0.080**	0.081	0.393	0.188	0.139	-	-	-	0.295
		all	0.078	0.078	0.379	0.178	0.132	-	-	-	0.283
	$\sigma_\epsilon = 3$	obs	0.172	0.172	0.419	0.272	0.308	-	-	-	0.319
		miss	**0.177**	**0.177**	0.443	0.291	0.325	-	-	-	0.337
		all	0.174	0.174	0.429	0.279	0.315	-	-	-	0.326
$p = 0.8$	$\sigma_\epsilon = 1$	obs	0.015	0.015	0.336	0.107	0.026	-	-	-	0.245
		miss	**0.017**	**0.017**	0.352	0.117	0.029	-	-	-	0.257
		all	0.016	0.016	0.339	0.109	0.026	-	-	-	0.247
	$\sigma_\epsilon = 2$	obs	0.055	0.055	0.360	0.147	0.088	-	-	-	0.270
		miss	**0.056**	**0.056**	0.377	0.157	0.093	-	-	-	0.283
		all	0.055	0.055	0.364	0.149	0.089	-	-	-	0.273
	$\sigma_\epsilon = 3$	obs	0.124	0.124	0.398	0.219	0.213	-	-	-	0.303
		miss	**0.126**	**0.126**	0.413	0.230	0.219	-	-	-	0.315
		all	0.124	0.124	0.401	0.221	0.214	-	-	-	0.305

This table reports the relative MSE of different estimators for the missing at random pattern with stationary time fixed effects. We compare the performance of wi-PCA, with known and unknown observation probability, with three benchmarks: PCA, block-PCA, and TWFE. We generate data from a one-factor model with two-way fixed effects: $Y_{it} = \mu + \alpha_i + \xi_t + \Lambda_i F_t + \epsilon_{it}$, where $\mu = 1, \alpha_i, \xi_t, F_t, \Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and errors $\epsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ with different σ_ϵ. The entries of Y are uniformly missing at random with observation probability p. Bold numbers indicate the best out-of-sample relative model performance. We set $N = T = 100$ and run 200 simulations for each setup.

Table IA.2: Relative MSE of common components for missing at random with non-stationary time fixed effects

			wi-PCA		PCA			Block-PCA			TWFE
			(FE+factor model)		(factor model only)			(factor model only)			(FE only)
Parameters		$\mathcal{S}$	known	unknown	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	
$p=0.4$	$\sigma_\epsilon=1$	obs	0.027	0.027	0.293	0.148	0.065	-	-	-	0.168
		miss	**0.033**	**0.033**	0.322	0.182	0.090	-	-	-	0.185
		all	0.030	0.030	0.311	0.168	0.080	-	-	-	0.178
	$\sigma_\epsilon=2$	obs	0.080	0.080	0.318	0.203	0.151	-	-	-	0.191
		miss	**0.087**	**0.087**	0.347	0.240	0.184	-	-	-	0.209
		all	0.084	0.084	0.336	0.225	0.171	-	-	-	0.201
	$\sigma_\epsilon=3$	obs	0.179	0.180	0.361	0.307	0.325	-	-	-	0.222
		miss	**0.190**	**0.190**	0.393	0.350	0.372	-	-	-	0.246
		all	0.186	0.186	0.380	0.332	0.353	-	-	-	0.236
$p=0.6$	$\sigma_\epsilon=1$	obs	0.016	0.016	0.284	0.129	0.034	-	-	-	0.166
		miss	**0.018**	**0.018**	0.303	0.147	0.043	-	-	-	0.178
		all	0.017	0.017	0.292	0.136	0.038	-	-	-	0.171
	$\sigma_\epsilon=2$	obs	0.050	0.050	0.299	0.163	0.088	-	-	-	0.181
		miss	**0.053**	**0.053**	0.317	0.182	0.098	-	-	-	0.194
		all	0.051	0.051	0.306	0.171	0.092	-	-	-	0.186
	$\sigma_\epsilon=3$	obs	0.112	0.113	0.328	0.226	0.192	-	-	-	0.202
		miss	**0.115**	**0.115**	0.345	0.246	0.204	-	-	-	0.214
		all	0.113	0.114	0.335	0.234	0.197	-	-	-	0.207
$p=0.8$	$\sigma_\epsilon=1$	obs	0.010	0.010	0.277	0.117	0.019	-	-	-	0.162
		miss	**0.011**	**0.011**	0.290	0.129	0.023	-	-	-	0.170
		all	0.010	0.010	0.279	0.119	0.020	-	-	-	0.163
	$\sigma_\epsilon=2$	obs	0.036	0.036	0.292	0.145	0.059	-	-	-	0.177
		miss	**0.037**	**0.037**	0.305	0.157	0.063	-	-	-	0.186
		all	0.036	0.036	0.294	0.148	0.059	-	-	-	0.179
	$\sigma_\epsilon=3$	obs	0.082	0.082	0.316	0.193	0.133	-	-	-	0.192
		miss	**0.083**	**0.083**	0.328	0.205	0.137	-	-	-	0.205
		all	0.082	0.082	0.318	0.196	0.134	-	-	-	0.194

This table reports the relative MSE of different estimators for the missing at random pattern with non-stationary time fixed effects. We compare the performance of wi-PCA, with known and unknown observation probability, with three benchmarks: PCA, block-PCA, and TWFE. We generate data from a one-factor model with two-way fixed effects: $Y_{it} = \mu + \alpha_i + \xi_t + \Lambda_i F_t + \epsilon_{it}$, where $\mu = 1, \alpha_i, F_t, \Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,the time fixed effects $\xi_t = 0.05t + \mathcal{N}(0,1)$, and errors $\epsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ with different σ_ϵ. The entries of Y are uniformly missing at random with observation probability p. Bold numbers indicate the best out-of-sample relative model performance. We set $N = T = 100$ and run 200 simulations for each setup.

Table IA.3: Relative MSE of common components for simultaneous treatment adoption pattern with stationary time fixed effects

			wi-PCA		PCA			Block-PCA			TWFE
			(FE+factor model)		(factor model only)			(factor model only)			(FE only)
Parameters		$\mathcal{S}$	known	unknown	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	
$c=40\%$	$\sigma_\epsilon=1$	obs	0.017	0.017	0.339	0.106	0.025	0.340	0.108	0.026	0.250
		miss	0.036	**0.031**	0.374	0.137	0.050	0.373	0.131	0.038	0.274
		all	0.023	0.021	0.349	0.115	0.032	0.350	0.114	0.030	0.257
	$\sigma_\epsilon=2$	obs	0.064	0.069	0.369	0.156	0.104	0.372	0.161	0.110	0.283
		miss	0.108	**0.105**	0.414	0.210	0.165	0.412	0.204	0.152	0.314
		all	0.077	0.079	0.383	0.172	0.122	0.384	0.174	0.122	0.292
	$\sigma_\epsilon=3$	obs	0.149	0.161	0.412	0.253	0.256	0.418	0.261	0.265	0.321
		miss	**0.232**	0.233	0.477	0.344	0.377	0.469	0.327	0.335	0.368
		all	0.173	0.182	0.431	0.279	0.292	0.433	0.280	0.285	0.334
$c=50\%$	$\sigma_\epsilon=1$	obs	0.015	0.016	0.344	0.106	0.023	0.347	0.108	0.026	0.255
		miss	0.029	**0.025**	0.367	0.128	0.041	0.367	0.124	0.033	0.268
		all	0.019	0.018	0.349	0.111	0.028	0.351	0.112	0.028	0.258
	$\sigma_\epsilon=2$	obs	0.059	0.064	0.366	0.151	0.093	0.373	0.161	0.106	0.280
		miss	0.093	**0.090**	0.406	0.197	0.144	0.404	0.192	0.132	0.302
		all	0.067	0.070	0.376	0.162	0.106	0.380	0.168	0.112	0.285
	$\sigma_\epsilon=3$	obs	0.137	0.152	0.400	0.234	0.234	0.413	0.256	0.259	0.319
		miss	**0.205**	**0.205**	0.456	0.311	0.334	0.452	0.296	0.289	0.354
		all	0.154	0.165	0.414	0.253	0.258	0.423	0.266	0.266	0.327
$c=60\%$	$\sigma_\epsilon=1$	obs	0.014	0.015	0.340	0.104	0.021	0.344	0.108	0.025	0.252
		miss	0.027	**0.022**	0.369	0.127	0.036	0.369	0.126	0.031	0.263
		all	0.016	0.016	0.345	0.108	0.024	0.348	0.111	0.026	0.254
	$\sigma_\epsilon=2$	obs	0.054	0.061	0.359	0.145	0.086	0.368	0.159	0.103	0.277
		miss	0.084	**0.082**	0.397	0.183	0.129	0.396	0.179	0.118	0.300
		all	0.060	0.065	0.366	0.152	0.094	0.373	0.162	0.106	0.281
	$\sigma_\epsilon=3$	obs	0.125	0.140	0.399	0.220	0.209	0.415	0.248	0.247	0.321
		miss	0.190	**0.189**	0.456	0.294	0.301	0.451	0.279	0.264	0.353
		all	0.137	0.149	0.410	0.234	0.227	0.422	0.253	0.250	0.327

This table reports the relative MSE of different estimators for the simultaneous treatment adoption pattern with stationary time fixed effects. We compare the performance of wi-PCA, with known and unknown observation probability, with three benchmarks: PCA, block-PCA, and TWFE. We generate data from a one-factor model with two-way fixed effects: $Y_{it} = \mu + \alpha_i + \xi_t + \Lambda_i F_t + \epsilon_{it}$, where $\mu = 1, \alpha_i, \xi_t, F_t, \Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and errors $\epsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ with different σ_ϵ. The entries of Y are fully observed in the first cT time periods with different p. Starting from $t = cT + 1$, units adopt the treatment with probability 0.5. The units selected in the treatment group do not have observation in the last $(1-c)T$ time periods. Bold numbers indicate the best out-of-sample relative model performance. We set $N = T = 100$ and run 200 simulations for each setup.

Table IA.4: Relative MSE of common components for simultaneous treatment adoption pattern with non-stationary time fixed effects

			wi-PCA		PCA			Block-PCA			TWFE
			(FE+factor model)		(factor model only)			(factor model only)			(FE only)
Parameters		$\mathcal{S}$	known	unknown	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	
$c=40\%$	$\sigma_\epsilon=1$	obs	0.013	0.013	0.355	0.180	0.041	0.334	0.147	0.019	0.185
		miss	0.019	**0.016**	0.579	0.554	0.480	0.416	0.253	0.038	0.142
		all	0.015	0.014	0.440	0.323	0.208	0.365	0.187	0.027	0.169
	$\sigma_\epsilon=2$	obs	0.048	0.051	0.376	0.209	0.091	0.359	0.186	0.080	0.207
		miss	0.055	**0.054**	0.594	0.573	0.527	0.476	0.322	0.154	0.162
		all	0.051	0.052	0.460	0.349	0.259	0.404	0.238	0.108	0.190
	$\sigma_\epsilon=3$	obs	0.111	0.120	0.399	0.266	0.183	0.387	0.254	0.189	0.228
		miss	**0.119**	0.120	0.607	0.599	0.586	0.549	0.437	0.324	0.191
		all	0.114	0.119	0.478	0.392	0.336	0.449	0.324	0.240	0.213
$c=50\%$	$\sigma_\epsilon=1$	obs	0.012	0.012	0.368	0.179	0.046	0.347	0.152	0.019	0.192
		miss	0.014	**0.011**	0.496	0.453	0.402	0.278	0.143	0.027	0.126
		all	0.012	0.012	0.413	0.276	0.172	0.323	0.149	0.022	0.168
	$\sigma_\epsilon=2$	obs	0.044	0.049	0.382	0.209	0.092	0.366	0.191	0.079	0.207
		miss	0.043	**0.042**	0.503	0.468	0.433	0.317	0.202	0.104	0.141
		all	0.044	0.046	0.425	0.300	0.211	0.349	0.195	0.088	0.183
	$\sigma_\epsilon=3$	obs	0.104	0.116	0.410	0.271	0.180	0.402	0.266	0.189	0.230
		miss	**0.095**	**0.095**	0.509	0.508	0.493	0.363	0.285	0.230	0.165
		all	0.101	0.108	0.445	0.354	0.290	0.388	0.273	0.203	0.207
$c=60\%$	$\sigma_\epsilon=1$	obs	0.011	0.011	0.372	0.180	0.043	0.352	0.159	0.019	0.192
		miss	0.011	**0.009**	0.381	0.330	0.287	0.176	0.082	0.019	0.111
		all	0.011	0.011	0.375	0.227	0.119	0.297	0.135	0.019	0.166
	$\sigma_\epsilon=2$	obs	0.041	0.046	0.379	0.205	0.089	0.365	0.193	0.078	0.205
		miss	0.035	**0.034**	0.387	0.347	0.324	0.216	0.125	0.076	0.126
		all	0.039	0.042	0.382	0.249	0.163	0.318	0.171	0.077	0.180
	$\sigma_\epsilon=3$	obs	0.095	0.107	0.409	0.259	0.170	0.402	0.260	0.182	0.231
		miss	**0.078**	**0.078**	0.408	0.384	0.384	0.269	0.203	0.170	0.148
		all	0.090	0.098	0.408	0.298	0.237	0.360	0.242	0.178	0.205

This table reports the relative MSE of different estimators for the simultaneous treatment adoption pattern with non-stationary time fixed effects. We compare the performance of wi-PCA, with known and unknown observation probability, with three benchmarks: PCA, block-PCA, and TWFE. We generate data from a one-factor model with two-way fixed effects: $Y_{it} = \mu + \alpha_i + \xi_t + \Lambda_i F_t + \epsilon_{it}$, where $\mu = 1, \alpha_i, F_t, \Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, the time fixed effects $\xi_t = 0.05t + \mathcal{N}(0,1)$, and errors $\epsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\epsilon^2)$ with different σ_ϵ. The entries of Y are fully observed in the first cT time periods with different p. Starting from $t = cT + 1$, units adopt the treatment with probability 0.5. The units selected in the treatment group do not have observation in the last $(1-c)T$ time periods. Bold numbers indicate the best out-of-sample relative model performance. We set $N = T = 100$ and run 200 simulations for each setup.

Table IA.5: Relative MSE of common components for staggered treatment adoption pattern with stationary time fixed effects

			wi-PCA		PCA			Block-PCA			TWFE
			(FE+factor model)		(factor model only)			(factor model only)			(FE only)
Parameters		$\mathcal{S}$	known	unknown	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	
$p = 0.5$	$\sigma_\epsilon = 1$	obs	0.035	0.017	0.345	0.080	0.032	0.343	0.078	0.026	0.282
		miss	0.063	**0.039**	0.423	0.291	0.122	0.432	0.268	0.071	0.200
		all	0.044	0.023	0.368	0.141	0.059	0.370	0.133	0.040	0.258
	$\sigma_\epsilon = 2$	obs	0.090	0.066	0.374	0.136	0.119	0.369	0.126	0.115	0.308
		miss	0.135	**0.110**	0.455	0.363	0.292	0.471	0.355	0.287	0.226
		all	0.104	0.080	0.398	0.205	0.172	0.400	0.196	0.168	0.283
	$\sigma_\epsilon = 3$	obs	0.206	0.161	0.424	0.247	0.274	0.412	0.215	0.289	0.346
		miss	0.307	**0.252**	0.541	0.529	0.569	0.548	0.514	0.621	0.290
		all	0.236	0.188	0.457	0.328	0.360	0.452	0.304	0.389	0.330
$p = 0.7$	$\sigma_\epsilon = 1$	obs	0.026	0.015	0.345	0.084	0.029	0.346	0.083	0.024	0.279
		miss	0.047	**0.032**	0.404	0.279	0.111	0.414	0.265	0.061	0.181
		all	0.031	0.019	0.359	0.132	0.049	0.362	0.127	0.033	0.256
	$\sigma_\epsilon = 2$	obs	0.080	0.061	0.363	0.134	0.111	0.361	0.129	0.105	0.294
		miss	0.120	**0.099**	0.439	0.356	0.264	0.447	0.339	0.233	0.213
		all	0.090	0.071	0.381	0.189	0.149	0.383	0.181	0.138	0.274
	$\sigma_\epsilon = 3$	obs	0.178	0.144	0.411	0.232	0.255	0.406	0.211	0.270	0.334
		miss	0.257	**0.219**	0.512	0.499	0.531	0.517	0.472	0.534	0.261
		all	0.198	0.163	0.435	0.297	0.322	0.434	0.276	0.337	0.317
$p = 0.9$	$\sigma_\epsilon = 1$	obs	0.015	0.013	0.343	0.092	0.022	0.344	0.092	0.020	0.264
		miss	0.034	**0.030**	0.395	0.264	0.094	0.407	0.251	0.044	0.179
		all	0.018	0.015	0.349	0.116	0.032	0.352	0.114	0.023	0.252
	$\sigma_\epsilon = 2$	obs	0.054	0.050	0.362	0.132	0.088	0.363	0.131	0.086	0.283
		miss	0.083	**0.079**	0.431	0.325	0.213	0.440	0.310	0.176	0.206
		all	0.058	0.055	0.371	0.159	0.106	0.373	0.157	0.100	0.271
	$\sigma_\epsilon = 3$	obs	0.123	0.119	0.401	0.206	0.218	0.402	0.201	0.215	0.318
		miss	0.186	**0.183**	0.482	0.447	0.487	0.489	0.426	0.407	0.244
		all	0.132	0.128	0.412	0.239	0.255	0.414	0.233	0.241	0.308

This table reports the relative MSE of different estimators for the staggered treatment adoption pattern with stationary time fixed effects. We compare the performance of wi-PCA, with known and unknown observation probability, with three benchmarks: PCA, block-PCA, and TWFE. We generate data from a one-factor model with two-way fixed effects: $Y_{it} = \mu + \alpha_i + \xi_t + \Lambda_i F_t + \epsilon_{it}$, where $\mu = 1, \alpha_i, \xi_t, F_t, \Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, and errors $\epsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ with different σ_ϵ. The entries of Y are fully observed in the first $0.1 \cdot T$ time periods. Starting from $t = 0.1 \cdot T + 1$, entries are missing with probability. Specifically, if Y_{it} is missing, the observations for that unit in the following time periods are also missing, i.e. if $W_{it} = 0$, then $W_{it'} = 0$ for $t' \geq t$. If Y_{it} is observed, the observation probability of $Y_{i,t+1}$ is p when $|\alpha_i \xi_t| > 2.5$ and 1 otherwise. Bold numbers indicate the best out-of-sample relative model performance. We set $N = T = 100$ and run 200 simulations for each setup.

Table IA.6: Relative MSE of common components for staggered treatment adoption pattern with non-stationary time fixed effects

			wi-PCA		PCA			Block-PCA			TWFE
			(FE+factor model)		(factor model only)			(factor model only)			(FE only)
Parameters		$\mathcal{S}$	known	unknown	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	
$p=0.5$	$\sigma_\epsilon=1$	obs	0.034	0.016	0.260	0.101	0.034	0.241	0.057	0.026	0.217
		miss	0.066	**0.042**	0.961	0.818	0.584	1.099	0.985	0.255	0.177
		all	0.049	0.028	0.586	0.433	0.290	0.637	0.487	0.133	0.199
	$\sigma_\epsilon=2$	obs	0.100	0.066	0.295	0.153	0.104	0.265	0.101	0.111	0.258
		miss	0.158	**0.120**	1.037	0.868	0.685	1.156	1.081	0.868	0.232
		all	0.127	0.091	0.640	0.485	0.374	0.679	0.555	0.466	0.246
	$\sigma_\epsilon=3$	obs	0.229	0.169	0.342	0.255	0.245	0.291	0.177	0.255	0.325
		miss	0.346	**0.268**	1.131	0.976	0.873	1.257	1.313	1.468	0.333
		all	0.283	0.215	0.713	0.595	0.541	0.745	0.711	0.828	0.328
$p=0.7$	$\sigma_\epsilon=1$	obs	0.030	0.015	0.264	0.111	0.033	0.245	0.065	0.025	0.209
		miss	0.056	**0.036**	0.981	0.816	0.578	1.118	1.016	0.212	0.165
		all	0.041	0.023	0.562	0.404	0.260	0.608	0.460	0.103	0.191
	$\sigma_\epsilon=2$	obs	0.093	0.059	0.290	0.162	0.095	0.268	0.107	0.105	0.250
		miss	0.146	**0.108**	1.006	0.844	0.660	1.157	1.118	0.790	0.220
		all	0.115	0.079	0.592	0.449	0.334	0.643	0.533	0.396	0.238
	$\sigma_\epsilon=3$	obs	0.206	0.151	0.339	0.251	0.221	0.301	0.181	0.247	0.320
		miss	0.319	**0.247**	1.147	0.952	0.839	1.297	1.324	1.389	0.316
		all	0.254	0.191	0.683	0.548	0.484	0.724	0.665	0.733	0.318
$p=0.9$	$\sigma_\epsilon=1$	obs	0.018	0.011	0.283	0.141	0.030	0.269	0.091	0.019	0.194
		miss	0.039	**0.029**	0.898	0.734	0.544	1.119	1.012	0.129	0.144
		all	0.024	0.016	0.469	0.320	0.186	0.523	0.366	0.053	0.179
	$\sigma_\epsilon=2$	obs	0.056	0.045	0.301	0.178	0.078	0.287	0.126	0.083	0.216
		miss	0.097	**0.081**	0.911	0.759	0.602	1.135	1.065	0.494	0.175
		all	0.068	0.055	0.487	0.353	0.237	0.541	0.408	0.207	0.204
	$\sigma_\epsilon=3$	obs	0.125	0.109	0.328	0.232	0.170	0.312	0.182	0.202	0.246
		miss	0.202	0.185	0.982	0.840	0.741	1.198	1.190	1.028	0.217
		all	0.147	0.132	0.524	0.413	0.341	0.576	0.483	0.450	0.237

This table reports the relative MSE of different estimators for the staggered treatment adoption pattern with non-stationary time fixed effects. We compare the performance of wi-PCA, with known and unknown observation probability, with three benchmarks: PCA, block-PCA, and TWFE. We generate data from a one-factor model with two-way fixed effects: $Y_{it} = \mu + \alpha_i + \xi_t + \Lambda_i F_t + \epsilon_{it}$, where $\mu = 1, \alpha_i, F_t, \Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, the time fixed effects $\xi_t = 0.05t + \mathcal{N}(0,1)$, and errors $\epsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\epsilon^2)$ with different σ_ϵ. The entries of Y are fully observed in the first $0.1 \cdot T$ time periods. Starting from $t = 0.1 \cdot T + 1$, entries are missing with probability. Specifically, if Y_{it} is missing, the observations for that unit in the following time periods are also missing, i.e. if $W_{it} = 0$, then $W_{it'} = 0$ for $t' \geq t$. If Y_{it} is observed, the observation probability of $Y_{i,t+1}$ is p when $|\alpha_i \xi_t| > 2.5$ and 1 otherwise. Bold numbers indicate the best out-of-sample relative model performance. We set $N = T = 100$ and run 200 simulations for each setup.

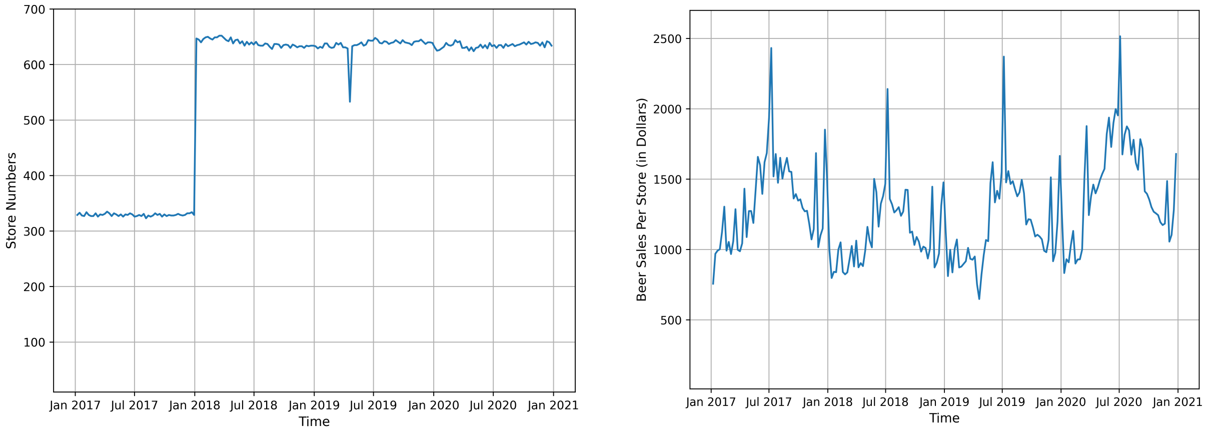
IA.B Empirical Study

IA.B.1 Data for Empirical Study

We use the weekly NielsenIQ retail scanner beer sales data from the Kilts Center for Marketing. Our data consists of 208 weekly observations of beer sales revenue from January 01, 2017, to December 26, 2020, which is aggregated at the state level for the U.S. Our cross-section consists of 45 states, of which 39 are control states and 6 are treated states that legalized the retail sale of marijuana at different time points during this period.

We explain why it important to normalize the weekly beer sales revenue by the number of stores in each state. The coverage of stores in the Kilts Center dataset fluctuates over time, in particular at the end of the sample. Figure IA.1(a) shows an example of the varying number of stores over time in Connecticut covered by Kilts Center dataset. Notably, there is a substantial increase in the number of stores in 2018. Therefore, a straightforward aggregation of sales revenue from the archived stores would yield misleading results. To address this issue, we normalize the total beer sales revenue by the number of stores in each state at each time period. As demonstrated in Figure IA.1(b), this normalization procedure helps to mitigate the impact of changes in store coverage and provides a more accurate representation of the beer sales data.

Figure IA.1: Number of stores and per-store beer sales revenue of Connecticut



(a) Store number

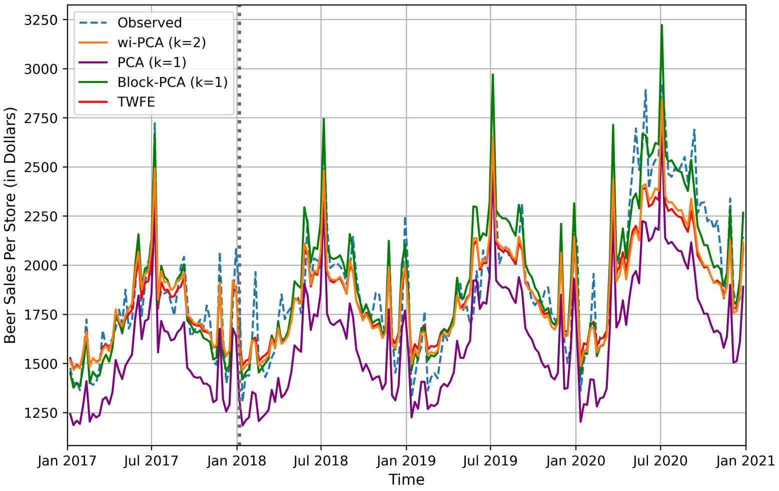
(b) Beer sales per store

IA.B.2 Observed and Estimated Time Series

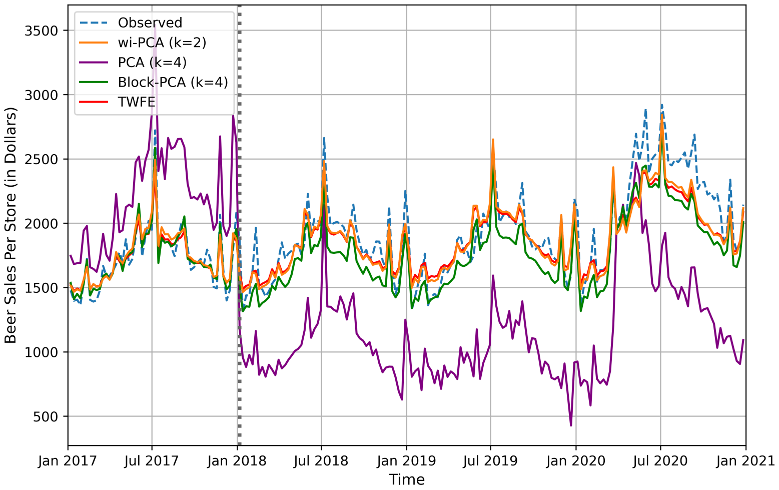
We present the observed time series and estimated control time series for all six treated states. We use wi-PCA, PCA, block-PCA and TWFW to estimate the common components, which serve as the control outcomes. We use wi-PCA with $k = 2$ factors and PCA and block-PCA with $k = 4$ factors to have a fair comparison in terms of the degrees of freedom. We also show the results for $k = 1$ factor for PCA and block-PCA.

We observe that the results from the main text for Massachusetts and Nevada generalize to the other treated states. We see that using more factors for PCA and block-PCA can lead to overfitting on the in-sample data, while using too few factors leads to an omitted variable bias.

Figure IA.2: California



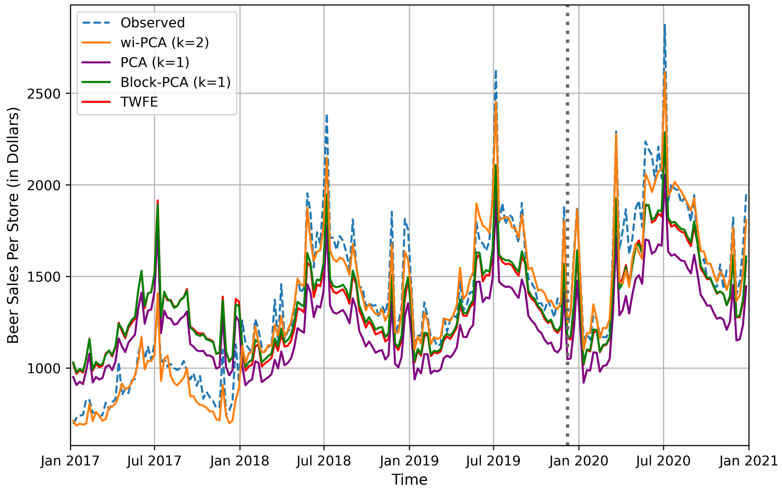
(a) k=1



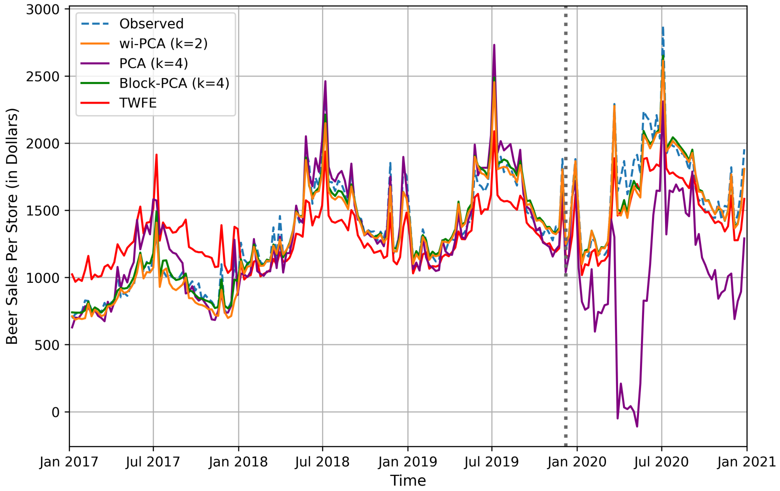
(b) k=4

These figures show observed outcomes and estimated control outcomes of beer sales per store for California. The figure above uses PCA and block-PCA estimators with $k = 1$ factor, while the figure below uses PCA and block-PCA with $k = 4$ factors. The grey vertical curve denotes the time of treatment.

Figure IA.3: Michigan



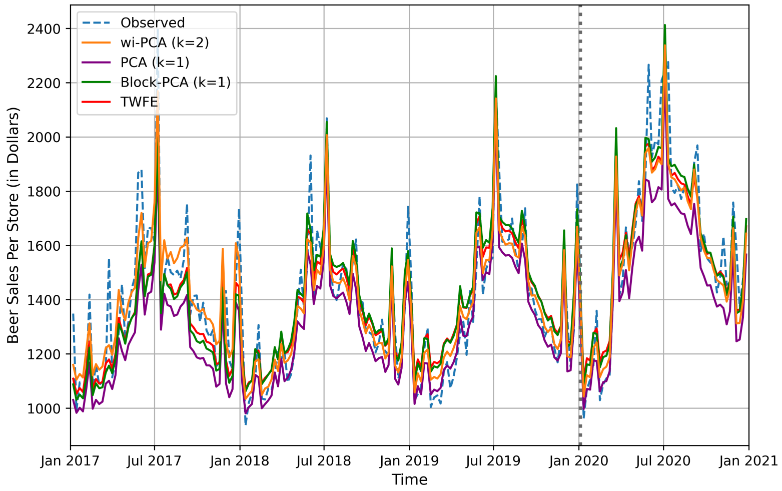
(a) k=1



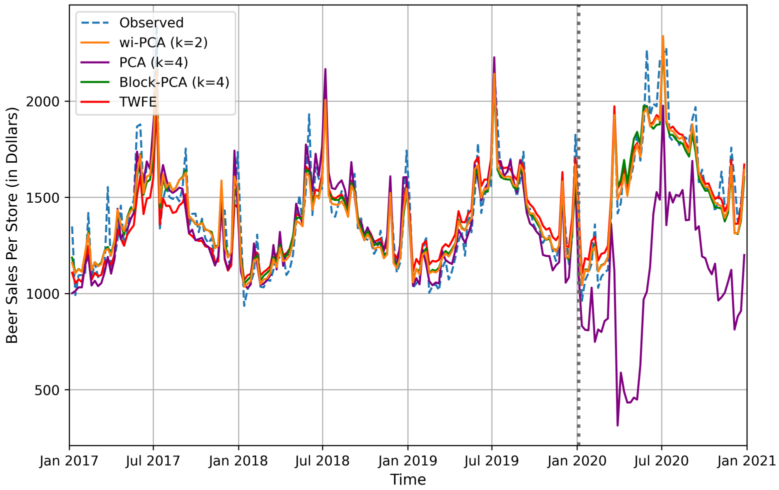
(b) k=4

These figures show observed outcomes and estimated control outcomes of beer sales per store for Michigan. The figure above uses PCA and block-PCA estimators with $k = 1$ factor, while the figure below uses PCA and block-PCA with $k = 4$ factors. The grey vertical curve denotes the time of treatment.

Figure IA.4: Illinois



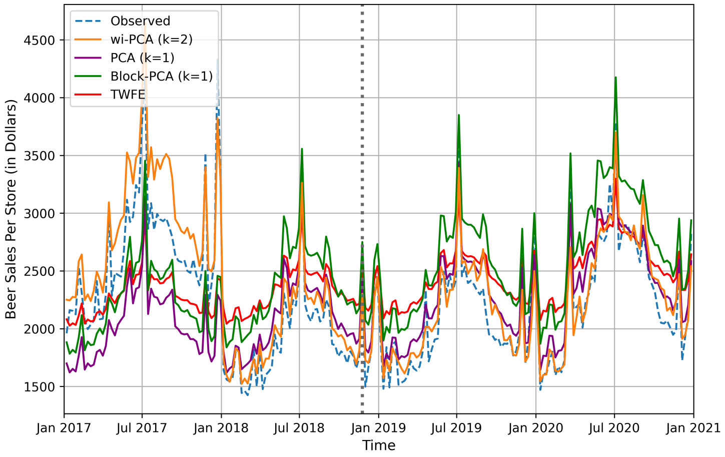
(a) k=1



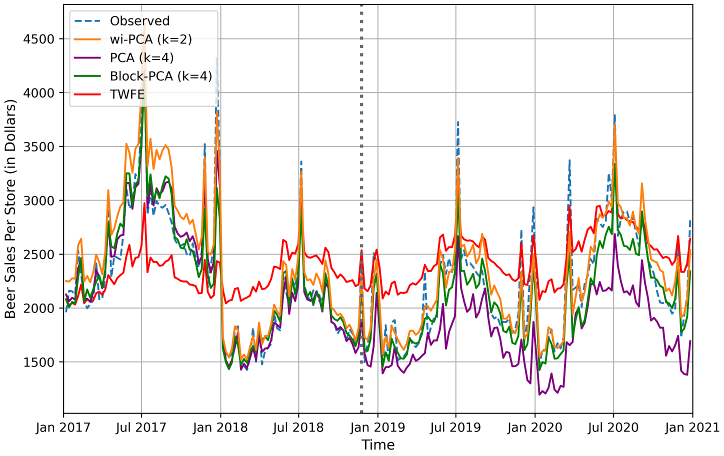
(b) k=4

These figures show observed outcomes and estimated control outcomes of beer sales per store for Illinois. The figure above uses PCA and block-PCA estimators with $k = 1$ factor, while the figure below uses PCA and block-PCA with $k = 4$ factors. The grey vertical curve denotes the time of treatment.

Figure IA.5: Massachusetts



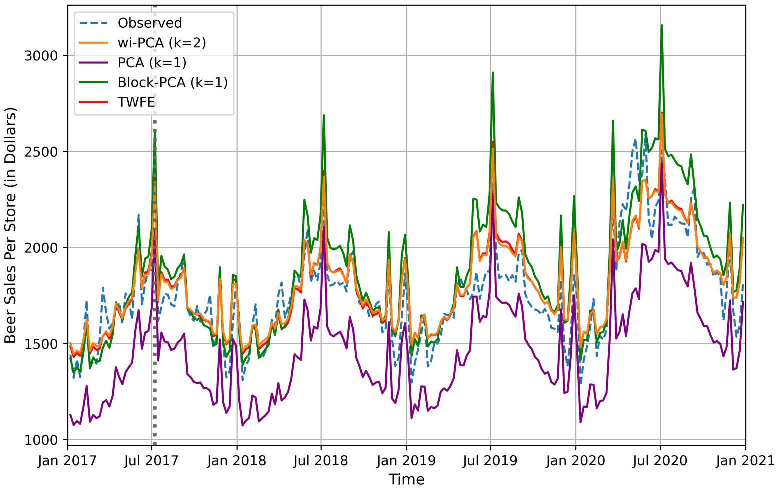
(a) k=1



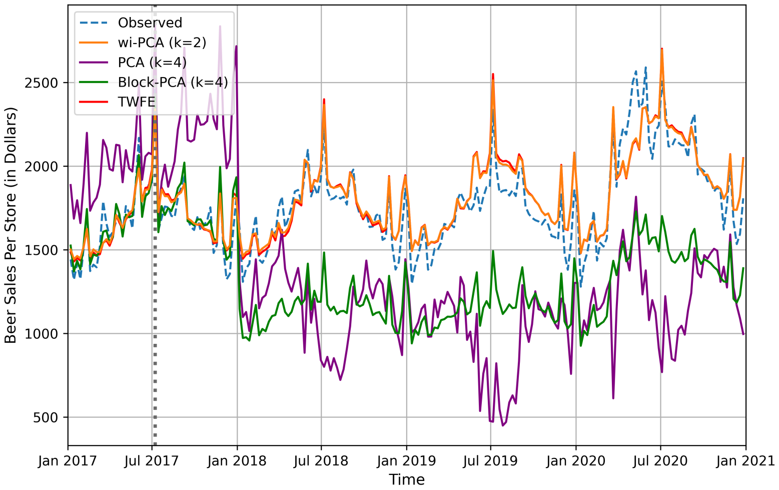
(b) k=4

These figures show observed outcomes and estimated control outcomes of beer sales per store for Massachusetts. The figure above uses PCA and block-PCA estimators with $k = 1$ factor, while the figure below uses PCA and block-PCA with $k = 4$ factors. The grey vertical curve denotes the time of treatment.

Figure IA.6: Nevada



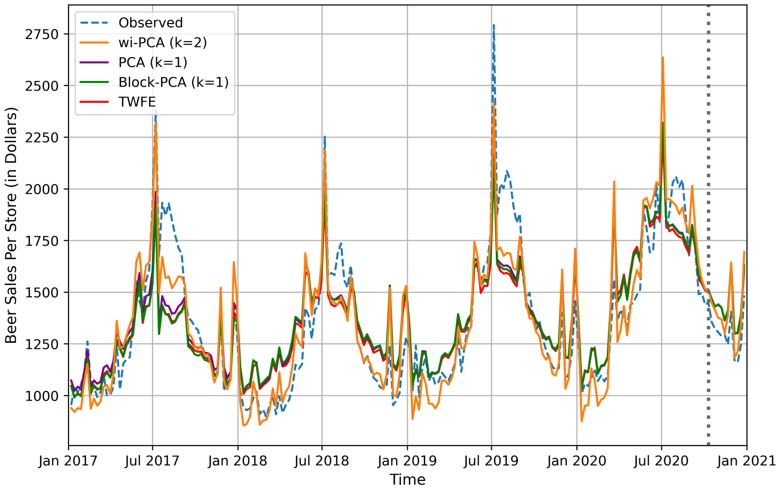
(a) k=1



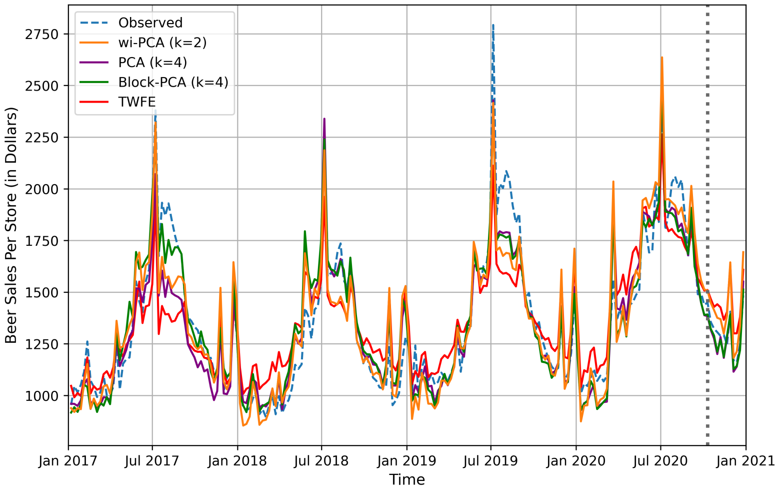
(b) k=4

These figures show observed outcomes and estimated control outcomes of beer sales per store for Nevada. The figure above uses PCA and block-PCA estimators with $k = 1$ factor, while the figure below uses PCA and block-PCA with $k = 4$ factors. The grey vertical curve denotes the time of treatment.

Figure IA.7: Maine



(a) k=1



(b) k=4

These figures show observed outcomes and estimated control outcomes of beer sales per store for Maine. The figure above uses PCA and block-PCA estimators with $k = 1$ factor, while the figure below uses PCA and block-PCA with $k = 4$ factors. The grey vertical curve denotes the time of treatment.

IA.C Proofs

To simplify the exposition we use the notation $D_t = \frac{1}{N}\sum_{i=1}^{N} W_{it}/p_{it}$, $\bar{W}_{i,\cdot} = \frac{1}{T}\sum_{t=1}^{T} W_{it}$, and the information set $I_i = I_{i1} \cup \cdots \cup I_{iT}$ for any unit i , and $I = I_1 \cup \cdots \cup I_N$.

IA.C.1 Proof of Theorem 1: Consistency

IA.C.1.1 Consistency for the sum of grand mean and fixed effects

Lemma 1. *Let $\delta_{N,T} = \min(N,T)$. Suppose Assumptions 1, 2, 3 and Case 1 in Assumption 4 hold. When $N, T \to \infty$, wi-PCA with weights* (4) *consistently estimate the sum of grand mean and fixed effects:*

$$\mathbb{E}\left[\Big((\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t) - (\mu + \alpha_i + \xi_t)\Big)^4\right] \le \frac{M}{\delta_{N,T}} ,$$

which implies that $\sqrt{\delta_{N,T}}\Big((\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t) - (\mu + \alpha_i + \xi_t)\Big) = O_p(1)$.

Proof. Plugging the expression of Y_{it} into $\tilde{\xi}_t$, we have

$$\begin{aligned}\tilde{\xi}_t &= D_t^{-1} \cdot \frac{1}{N}\sum_{i=1}^{N} \frac{W_{it}}{p_{it}}\Big(\mu + \alpha_i + \xi_t + \Lambda_i^\top F_t + \epsilon_{it}\Big) - \tilde{\mu} \\ &= \xi_t + (\mu - \tilde{\mu}) + D_t^{-1} \cdot \frac{1}{N}\sum_{i=1}^{N} \frac{W_{it}}{p_{it}}\Big(\alpha_i + \Lambda_i^\top F_t + \epsilon_{it}\Big) .\end{aligned}$$

Then, plugging Y_{it} and $\tilde{\xi}_t$ into $\tilde{\alpha}_i$ yields

$$\begin{aligned}\tilde{\alpha}_i &= \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T}\sum_{t=1}^{T} W_{it}\Big(\mu + \alpha_i + \xi_t + \Lambda_i^\top F_t + \epsilon_{it} - \tilde{\xi}_t\Big) - \tilde{\mu} \\ &= \alpha_i + (\mu - \tilde{\mu}) + \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T}\sum_{t=1}^{T} W_{it}\Big(\xi_t - \tilde{\xi}_t + \Lambda_i^\top F_t + \epsilon_{it}\Big) \\ &= \alpha_i + \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T}\sum_{t=1}^{T} W_{it}\Big(\Lambda_i^\top F_t + \epsilon_{it}\Big) - \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T} W_{it}\frac{W_{jt}}{p_{jt}} D_t^{-1}\Big(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt}\Big) .\end{aligned}$$

We combine these two terms and derive the estimation error of $\tilde{\mu}+\tilde{\alpha}_i+\tilde{\xi}_t$:

$$\begin{aligned}\tilde{\Delta}_{it} =&\ \tilde{\mu}+\tilde{\alpha}_i+\tilde{\xi}_t-(\mu+\alpha_i+\xi_t)\\ =&\ \underbrace{D_t^{-1}\cdot\frac{1}{N}\sum_{j=1}^{N}\frac{W_{jt}}{p_{jt}}\left(\alpha_j+\Lambda_j^\top F_t+\epsilon_{jt}\right)}_{\tilde{\Delta}_{it,1}}+\underbrace{\bar{W}_{i,\cdot}^{-1}\cdot\frac{1}{T}\sum_{s=1}^{T}W_{is}\left(\Lambda_i^\top F_s+\epsilon_{is}\right)}_{\tilde{\Delta}_{it,2}}\\ &\underbrace{-\bar{W}_{i,\cdot}^{-1}\cdot\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}W_{is}\frac{W_{js}}{p_{js}}D_s^{-1}\left(\alpha_j+\Lambda_j^\top F_s+\epsilon_{js}\right)}_{\tilde{\Delta}_{it,3}}.\end{aligned}$$

In the next steps, we derive bounds for the fourth moments of the three terms $\tilde{\Delta}_{it,1}$, $\tilde{\Delta}_{it,2}$ and $\tilde{\Delta}_{it,3}$, respectively. To simplify the exposition, we use the notation $X_{it}=W_{it}/p_{it}$.

First, we consider the term $\tilde{\Delta}_{it,1}=D_t^{-1}\frac{1}{N}\sum_{j=1}^{N}X_{jt}(\alpha_j+\Lambda_j^\top F_t+\epsilon_{jt})$. By Assumption 1.2, we have $q\leq\bar{W}_{i,\cdot}\leq 1$ and $q\leq D_t\leq 1/\eta$. We denote $\lambda_{it}=\alpha_i+\Lambda_i^\top F_t+\epsilon_{it}$. Since $X_{it}\perp\!\!\!\perp\lambda_{js}|I$ for any i,j,s,t, it holds that

$$\begin{aligned}\mathbb{E}[\tilde{\Delta}_{it,1}^4]&\leq q^{-4}\cdot\mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N}X_{jt}\lambda_{jt}\right)^4\right]\\ &=q^{-4}\cdot\frac{1}{N^4}\sum_{i,j,h,l=1}^{N}\mathbb{E}\big[\mathbb{E}[X_{it}X_{jt}X_{ht}X_{lt}|I]\cdot\mathbb{E}\left[\lambda_{it}\lambda_{jt}\lambda_{ht}\lambda_{lt}|I\right]\big].\end{aligned}$$

By Assumption 1.3, $W_{jt}\perp\!\!\!\perp W_{it}|I$ for any $i\neq j$. When i,j,h,l are distinct, we have that $\mathbb{E}[X_{it}X_{jt}X_{ht}X_{lt}|I]=1$ and $\mathbb{E}[\lambda_{it}\lambda_{jt}\lambda_{ht}\lambda_{lt}]=\mathbb{E}[(\alpha_i+\Lambda_i^\top F_t)(\alpha_j+\Lambda_j^\top F_t)(\alpha_h+\Lambda_h^\top F_t)(\alpha_l+\Lambda_l^\top F_t)]$; when i,h,l are distinct, $\mathbb{E}[X_{it}^2X_{ht}X_{lt}|I]=\mathbb{E}[X_{it}^2|I]$ and $\mathbb{E}[\lambda_{it}^2\lambda_{ht}\lambda_{lt}|I]=\mathbb{E}[(\alpha_i+\Lambda_i^\top F_t+\epsilon_{it})^2(\alpha_h+\Lambda_h^\top F_t)(\alpha_l+\Lambda_l^\top F_t)|I]$. Furthermore, since $\sum_{i=1}^{N}\alpha_i=0$ and $\sum_{i=1}^{N}\Lambda_i=0$, it holds that

$$\begin{aligned}\mathbb{E}[\tilde{\Delta}_{it,1}^4]\leq&\ q^{-4}\cdot\frac{1}{N^4}\sum_{i,j,h,l=1}^{N}\mathbb{E}\left[(\alpha_i+\Lambda_i^\top F_t)(\alpha_j+\Lambda_j^\top F_t)(\alpha_h+\Lambda_h^\top F_t)(\alpha_l+\Lambda_l^\top F_t)\right]\\ &-q^{-4}\cdot\frac{6}{N^4}\sum_{i,h,l=1}^{N}\mathbb{E}\left[(\alpha_i+\Lambda_i^\top F_t)^2(\alpha_h+\Lambda_h^\top F_t)(\alpha_l+\Lambda_l^\top F_t)\right]\\ &+q^{-4}\cdot\frac{6}{N^4}\sum_{i,h,l=1}^{N}\mathbb{E}\left[X_{it}^2(\alpha_i+\Lambda_i^\top F_t+\epsilon_{it})^2(\alpha_h+\Lambda_h^\top F_t)(\alpha_l+\Lambda_l^\top F_t)\right]+\frac{M}{N^2}\leq\frac{M}{N^2}.\end{aligned}$$

Next, we consider the term $\tilde{\Delta}_{it,2} = \bar{W}_{i,\cdot}^{-1}\frac{1}{T}\sum_{s=1}^{T} W_{is}(\Lambda_i^\top F_s + \epsilon_{is})$. We have

$$\mathbb{E}[\tilde{\Delta}_{it,2}^4] \le 8q^{-4} \cdot \frac{1}{T^4}\sum_{s,t,y,z=1}^{T} \mathbb{E}\left[W_{is}W_{it}W_{iy}W_{iz}\Lambda_i^\top F_t \Lambda_i^\top F_s \Lambda_i^\top F_y \Lambda_i^\top F_z\right]$$
$$+ 8q^{-4} \cdot \frac{1}{T^4}\sum_{s,t,y,z=1}^{T} \mathbb{E}\left[W_{it}W_{is}W_{iy}W_{iz}\right] \cdot \mathbb{E}\left[\epsilon_{it}\epsilon_{is}\epsilon_{iy}\epsilon_{iz}\right].$$

Since factors are independent of the missing patterns and loadings, and the loadings have bounded fourth moments, the first term on the RHS is bounded by

$$\begin{aligned}
&\frac{1}{T^4}\sum_{s,t,y,z=1}^{T} \mathbb{E}\left[W_{it}W_{is}W_{iy}W_{iz}\Lambda_i^\top F_t \Lambda_i^\top F_s \Lambda_i^\top F_y \Lambda_i^\top F_z\right] \\
=& \frac{1}{T^4}\sum_{s,t,y,z=1}^{T}\sum_{p,q,r,h=1}^{k} \mathbb{E}\left[W_{it}W_{is}W_{iy}W_{iz}\Lambda_{i,p}\Lambda_{i,q}\Lambda_{i,r}\Lambda_{i,h}\right] \cdot \mathbb{E}\left[F_{t,p}F_{s,q}F_{y,r}F_{z,h}\right] \\
\le& \frac{M}{T^4}\sum_{s,t,y,z=1}^{T}\sum_{p,q,r,h=1}^{k} \left|\mathbb{E}\left[F_{t,p}F_{s,q}F_{y,r}F_{z,h}\right]\right| \le \frac{M}{T^2},
\end{aligned}$$

where the last inequality holds by Assumption 3.3. Furthermore, we have that

$$\frac{1}{T^4}\sum_{s,t,y,z=1}^{T} \mathbb{E}\left[W_{it}W_{is}W_{iy}W_{iz}\right]\mathbb{E}\left[\epsilon_{it}\epsilon_{is}\epsilon_{iy}\epsilon_{iz}\right] \le \frac{1}{T^4}\sum_{s,t,y,z=1}^{T} \left|\mathbb{E}\left[\epsilon_{it}\epsilon_{is}\epsilon_{iy}\epsilon_{iz}\right]\right| \le M/T^2.$$

Combining these two terms gives us $\mathbb{E}[\tilde{\Delta}_{it,2}^4] \le M/T^2$.

Finally, for $\tilde{\Delta}_{it,3}$, we can write it as

$$\tilde{\Delta}_{it,3} = \bar{W}_{i,\cdot}^{-1} \cdot \underbrace{\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} W_{is}X_{js}\left(D_s^{-1}-1\right)\left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right)}_{\omega_1}$$
$$+ \bar{W}_{i,\cdot}^{-1} \cdot \underbrace{\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} W_{is}X_{js}\left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right)}_{\omega_2}.$$

For any t, we observe that

$$\mathbb{E}\left[(D_t^{-1}-1)^4|I\right] = \mathbb{E}\left[\left.\frac{(1-D_t)^4}{D_t^4}\right| I\right] \le q^{-4} \cdot \mathbb{E}\left[(1-D_t)^4|I\right] \le \frac{M}{N^2}.$$

Therefore, the fourth moment of ω_1 can be bounded by

$$\begin{aligned}\mathbb{E}\left[\omega_1^4\right] &\leq \frac{M}{NT}\sum_{j=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left[\mathbb{E}\left[(D_s^{-1}-1)^4|I\right]\cdot\mathbb{E}\left[(\alpha_j+\Lambda_j^\top F_s+\epsilon_{js})^4|I\right]\right]\\ &\leq \frac{M}{N^3T}\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbb{E}\left[(\alpha_j+\Lambda_j^\top F_s+\epsilon_{js})^4\right]\leq\frac{M}{N^2}.\end{aligned}$$

Additionally, following the proof of $\tilde{\Delta}_{it,1}$, we can show that

$$\mathbb{E}\left[\omega_2^4\right]\leq\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T}W_{is}^4\cdot\frac{1}{T}\sum_{s=1}^{T}\left(\frac{1}{N}\sum_{j=1}^{N}X_{js}\lambda_{js}\right)^4\right]\leq\frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N}X_{js}\lambda_{js}\right)^4\right]\leq\frac{M}{N^2}.$$

Therefore, $\mathbb{E}[\tilde{\Delta}_{it,3}^4]\leq M/N^2$ and we complete our proof. □

Lemma 2. *Let $\delta_{N,T}=\min(N,T)$. Suppose Assumptions 1, 2 and 3 and Case 2 in Assumption 4 hold. When $N,T\to\infty$, wi-PCA with weights* (5) *consistently estimates the sum of grand mean and fixed effects:*

$$\mathbb{E}\left[\Big((\tilde{\mu}+\tilde{\alpha}_i+\tilde{\xi}_t)-(\mu+\alpha_i+\xi_t)\Big)^4\right]\leq\frac{M}{\delta_{N,T}},$$

which implies that $\sqrt{\delta_{N,T}}\Big((\tilde{\mu}+\tilde{\alpha}_i+\tilde{\xi}_t)-(\mu+\alpha_i+\xi_t)\Big)=O_p(1)$.

Proof. Similar to Lemma 1, we decompose the estimation error $\tilde{\Delta}_{it}=(\tilde{\mu}+\tilde{\alpha}_i+\tilde{\xi}_t)-(\mu+\alpha_i+\xi_t)$ into three components

$$\begin{aligned}\tilde{\Delta}_{it}=&\underbrace{\tilde{D}_t^{-1}\cdot\frac{1}{N}\sum_{j=1}^{N}\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot}v_t}\Big(\alpha_j+\Lambda_j^\top F_t+\epsilon_{jt}\Big)}_{\tilde{\Delta}_{it,1}}+\underbrace{\bar{W}_{i,\cdot}^{-1}\cdot\frac{1}{T}\sum_{s=1}^{T}W_{is}\Big(\Lambda_i^\top F_s+\epsilon_{is}\Big)}_{\tilde{\Delta}_{it,2}}\\ &\underbrace{-\bar{W}_{i,\cdot}^{-1}\cdot\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}W_{is}\frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot}v_s}\tilde{D}_s^{-1}\Big(\alpha_j+\Lambda_j^\top F_s+\epsilon_{js}\Big)}_{\tilde{\Delta}_{it,3}},\end{aligned}$$

where $\tilde{D}_t=\frac{1}{N}\sum_{i=1}^{N}W_{it}\bar{v}/(\bar{W}_{i,\cdot}v_t)$. Based on Lemma 1, $\tilde{\Delta}_{it,2}$ has bounded fourth moment. Next, we bound the other two terms $\tilde{\Delta}_{it,1}$ and $\tilde{\Delta}_{it,3}$.

Consider the first term $\tilde{\Delta}_{it,1}$. We decompose it as follows

$$\tilde{\Delta}_{it,1} = \tilde{D}_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \left(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot} v_t} - \frac{W_{jt}}{p_{jt}} \right) \left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right)$$
$$+ \tilde{D}_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \frac{W_{jt}}{p_{jt}} \left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right).$$

Since $W_{it} \perp\!\!\!\perp W_{is} | I_i$ for any s and t satisfying $|s-t| > c$ and $\bar{v} = \frac{1}{T}\sum_{t=1}^{T} v_t + O(\frac{1}{\sqrt{T}})$, it is easy to see that $\mathbb{E}\left[(\bar{W}_{j,\cdot} - u_j \bar{v})^4 | I\right] \leq M/T^2$, and thus,

$$\mathbb{E}\left[\left(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot} v_t} - \frac{W_{jt}}{p_{jt}} \right)^4 \middle| I \right] \leq \frac{1}{\eta^4 q^4} \cdot \mathbb{E}\left[(\bar{W}_{j,\cdot} - u_j \bar{v})^4 | I\right] \leq \frac{M}{T^2}.$$

Moreover, by Assumptions 1, $\tilde{D}_t$ satisfies $\eta q \leq \tilde{D}_t \leq 1/\eta q$. As a result, the first part of $\tilde{\Delta}_{it,1}$ can be bounded by

$$\begin{aligned}
&\mathbb{E}\left[\left(\tilde{D}_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \left(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot} v_t} - \frac{W_{jt}}{p_{jt}} \right) \left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right) \right)^4 \right] \\
&\leq M \cdot \mathbb{E}\left[\frac{1}{N} \sum_{j=1}^{N} \mathbb{E}\left[\left(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot} v_t} - \frac{W_{jt}}{p_{jt}} \right)^4 \middle| I \right] \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}\left[\left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right)^4 \middle| I \right] \right] \\
&\leq \frac{M}{T^2 N} \sum_{j=1}^{N} \mathbb{E}\left[\left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right)^4 \right] \\
&\leq \frac{M}{T^2}.
\end{aligned}$$

According to Lemma 1, we can bound the fourth moment of the second part of $\tilde{\Delta}_{it,1}$ by M/N^2. Therefore, we get that $\mathbb{E}[\tilde{\Delta}_{it,1}^4] \leq M/\delta_{N,T}^2$.

The last term $\tilde{\Delta}_{it,3}$ can be decomposed as

$$\tilde{\Delta}_{it,3} = \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \left(\frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot} v_s} - \frac{W_{js}}{p_{js}} \right) \tilde{D}_s^{-1} \left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js} \right)$$
$$+ \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \frac{W_{js}}{p_{js}} \left(\tilde{D}_s^{-1} - D_s^{-1} \right) \left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js} \right)$$
$$+ \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \frac{W_{js}}{p_{js}} D_s^{-1} \left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js} \right).$$

Note that the last part on the RHS is the term $\tilde{\Delta}_{it,3}$ in Lemma 1, and thus its fourth moment can be bounded by M/N^2. We can bound the first part of $\tilde{\Delta}_{it,3}$ by

$$\begin{aligned}
&\mathbb{E}\left[\left(\bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} W_{is}\left(\frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot}v_s} - \frac{W_{js}}{p_{js}}\right)\tilde{D}_s^{-1}\left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right)\right)^4\right] \\
&\leq M \cdot \mathbb{E}\left[\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbb{E}\left[\left(\frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot}v_s} - \frac{W_{js}}{p_{js}}\right)^4\middle| I\right] \cdot \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbb{E}\left[\left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right)^4\middle| I\right]\right] \\
&\leq \frac{M}{T^2}\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbb{E}\left[\left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right)^4\right] \\
&\leq \frac{M}{T^2}.
\end{aligned}$$

Observe that $\mathbb{E}[(\tilde{D}_s^{-1} - D_s^{-1})^4|I] \leq M \cdot \mathbb{E}[(\tilde{D}_s - D_s)^4|I] \leq M/T^2$. By similar arguments, we can bound the fourth moment of the second part of $\tilde{\Delta}_{it,3}$ by M/T^2. This completes our proof. □

Lemma 3. *Let $\delta_{N,T} = \min(N,T)$. Suppose Assumptions 1, 2 and 3 and Case 3 in Assumption 4 hold. When $N, T \to \infty$, wi-PCA with weights* (6) *consistently estimates the sum of grand mean and fixed effects:*

$$\mathbb{E}\left[\left((\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t) - (\mu + \alpha_i + \xi_t)\right)^4\right] \leq \frac{M}{\delta_{N,T}},$$

which implies that $\sqrt{\delta_{N,T}}\left((\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t) - (\mu + \alpha_i + \xi_t)\right) = O_p(1)$.

Proof. Since the observation pattern is monotone and $\frac{1}{N}\sum_{i=1}^{N} W_{it} \geq q$ for any t, there are at least qN fully observed units. We denote by $\mathcal{N}_c$ and N_c the set and number of these units. The time fixed effect estimator $\tilde{\xi}_t$ can be decomposed as

$$\tilde{\xi}_t = \xi_t + (\mu - \tilde{\mu}) + \frac{1}{N_c}\sum_{i \in \mathcal{N}_c}\left(\alpha_i + \Lambda_i^\top F_t + \epsilon_{it}\right).$$

We plug $\tilde{\xi}_t$ into $\tilde{\alpha}_i$ to obtain

$$\tilde{\alpha}_i = \alpha_i - \frac{1}{N_c}\sum_{j \in \mathcal{N}_c}\alpha_j + \left(\frac{1}{T}\sum_{t=1}^{T} W_{it}\right)^{-1}\frac{1}{T}\sum_{t=1}^{T} W_{it}\left[\Lambda_i^\top F_t + \epsilon_{it} - \frac{1}{N_c}\sum_{j \in \mathcal{N}_c}\left(\Lambda_j^\top F_t + \epsilon_{jt}\right)\right].$$

Combining these terms, we get the first-stage estimation error

$$\tilde{\Delta}_{it} = \underbrace{\frac{1}{N_c}\sum_{j\in\mathcal{N}_c}\left(\Lambda_j^\top F_t + \epsilon_{jt}\right)}_{\tilde{\Delta}_{it,1}} + \underbrace{\bar{W}_{i,\cdot}^{-1}\cdot\frac{1}{T}\sum_{s=1}^{T} W_{is}\left(\Lambda_i^\top F_s + \epsilon_{is}\right)}_{\tilde{\Delta}_{it,2}}$$

$$\underbrace{-\bar{W}_{i,\cdot}^{-1}\cdot\frac{1}{T}\sum_{s=1}^{T} W_{is}\frac{1}{N_c}\sum_{j\in\mathcal{N}_c}\left(\Lambda_j^\top F_s + \epsilon_{js}\right)}_{\tilde{\Delta}_{it,3}}.$$

According to the proof of Lemma 1, $\mathbb{E}[\tilde{\Delta}_{it,2}^4] \le M/\delta_{N,T}^2$. Since $\mathbb{E}\left\|\frac{1}{N_c}\sum_{i\in\mathcal{N}_c}\Lambda_i\right\|^4 \le M/\delta_{N,T}^2$, it is easy to see that $\mathbb{E}[\tilde{\Delta}_{it,1}^4] \le M/\delta_{N,T}^2$ and $\mathbb{E}[\tilde{\Delta}_{it,3}^4] \le M/\delta_{N,T}^2$. □

IA.C.1.2 Consistency for the common components

The transformed $\dot{Y}$ has an approximate factor structure

$$\dot{Y}_{it} = \Lambda_i^\top F_t + \dot{\epsilon}_{it}$$

with a new idiosyncratic error term $\dot{\epsilon}_{it} = \epsilon_{it} + (\mu + \alpha_i + \xi_t) - (\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t)$. We denote $\tilde{\Delta}_{it} = (\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t) - (\mu + \alpha_i + \xi_t)$.

We first estimate the cross-sectional second-moment matrix $\tilde{\Sigma}$ with $\tilde{\Sigma}_{ij} = \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\dot{Y}_{it}\dot{Y}_{jt}$. Then, we apply PCA to $\tilde{\Sigma}$ to estimate loadings. Specifically, we solve $\frac{1}{N}\tilde{\Sigma}\tilde{\Lambda} = \tilde{\Lambda}\tilde{D}$, where $\tilde{D}$ is the diagonal matrix consisting of the k largest eigenvalues of $\tilde{\Sigma}/N$. Therefore, $\tilde{\Lambda}_i$ can be expanded as

$$\tilde{\Lambda}_i = \tilde{D}^{-1}\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\Lambda_j^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_tF_t^\top\Lambda_i + \tilde{D}^{-1}\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\Lambda_i^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_t\dot{\epsilon}_{jt}$$

$$+\tilde{D}^{-1}\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\Lambda_j^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_t\dot{\epsilon}_{it} + \tilde{D}^{-1}\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\dot{\epsilon}_{it}\dot{\epsilon}_{jt}\,.$$

To simplify notation, we define three auxiliary terms

$$\eta_{ij} = \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\Lambda_i^\top F_t\dot{\epsilon}_{jt} = \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\Lambda_i^\top F_t(\epsilon_{jt} - \tilde{\Delta}_{jt}),$$

$$\xi_{ij} = \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\Lambda_j^\top F_t\dot{\epsilon}_{it} = \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\Lambda_j^\top F_t(\epsilon_{it} - \tilde{\Delta}_{it}),$$

$$\gamma_{ij} = \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\dot{\epsilon}_{it}\dot{\epsilon}_{jt} = \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}(\epsilon_{it} - \tilde{\Delta}_{it})(\epsilon_{jt} - \tilde{\Delta}_{jt})\,.$$

We let $H_i = \tilde{D}^{-1}\frac{1}{N}\sum_{j=1}^N \tilde{\Lambda}_j \Lambda_j^\top \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t F_t^\top$ and consider a unified rotation matrix H for all i as $H = (NT)^{-1}\tilde{D}^{-1}\tilde{\Lambda}^\top \Lambda F^\top F$. This yields the decomposition

$$\tilde{\Lambda}_i = H\Lambda_i + (H_i - H)\Lambda_i + \tilde{D}^{-1}\left(\frac{1}{N}\sum_{j=1}^N \tilde{\Lambda}_j \eta_{ij} + \frac{1}{N}\sum_{j=1}^N \tilde{\Lambda}_j \xi_{ij} + \frac{1}{N}\sum_{j=1}^N \tilde{\Lambda}_j \gamma_{ij}\right).$$

Lemma 4. *Let $\delta_{N,T} = \min(N,T)$. Suppose Assumptions 1, 2, 3 hold and the first-stage estimation error $\mathbb{E}[\tilde{\Delta}_{it}^4] \leq M/\delta_{N,T}^2$ for any i and t. Then, it holds that*

1. *For any i and j, $\mathbb{E}[\eta_{ij}^2] \leq M/\delta_{N,T}$ and $\mathbb{E}[\xi_{ij}^2] \leq M/\delta_{N,T}$.*
2. *For any j, $\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\gamma_{ij}^2] \leq M/\delta_{N,T}$.*

Proof. 1. For η_{ij}, we have

$$\mathbb{E}[\eta_{ij}^2] \leq 2\cdot\mathbb{E}\left[\left(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \Lambda_i^\top F_t \epsilon_{jt}\right)^2\right] + 2\cdot\mathbb{E}\left[\left(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \Lambda_i^\top F_t \tilde{\Delta}_{jt}\right)^2\right].$$

According to Lemma 1 in Xiong and Pelger (2023), $\mathbb{E}\left[(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \Lambda_i^\top F_t \epsilon_{jt})^2\right] \leq M/T$. For any i, j, t and s, by the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned}\mathbb{E}\left|\Lambda_i^\top F_t \tilde{\Delta}_{jt} \Lambda_i^\top F_s \tilde{\Delta}_{js}\right| &\leq \left(\mathbb{E}\left[(\Lambda_i^\top F_t \Lambda_i^\top F_s)^2\right]\cdot\mathbb{E}\left[(\tilde{\Delta}_{jt}\tilde{\Delta}_{js})^2\right]\right)^{1/2} \\ &\leq M\cdot\left(\mathbb{E}[\tilde{\Delta}_{jt}^4]\cdot\mathbb{E}[\tilde{\Delta}_{js}^4]\right)^{1/4} \leq \frac{M}{\delta_{N,T}}.\end{aligned}$$

Therefore, the second term $\mathbb{E}\left[(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \Lambda_i^\top F_t \tilde{\Delta}_{jt})^2\right] \leq M/\delta_{N,T}$, and thus, $\mathbb{E}[\eta_{ij}^2] \leq M/\delta_{N,T}$. By symmetry, $\mathbb{E}[\xi_{ij}^2] \leq M/\delta_{N,T}$.

2. Similar to part 1, we just need to bound the three terms $\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it}\epsilon_{jt})^2\right]$, $\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it}\tilde{\Delta}_{jt})^2\right]$ and $\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \tilde{\Delta}_{it}\tilde{\Delta}_{jt})^2\right]$.

The first term can be decomposed as

$$\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it}\epsilon_{jt} = \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \mathbb{E}[\epsilon_{it}\epsilon_{jt}] + \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \left(\epsilon_{it}\epsilon_{jt} - \mathbb{E}[\epsilon_{it}\epsilon_{jt}]\right).$$

By Assumption 2.3,

$$\frac{1}{N}\sum_{i=1}^N\left(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \mathbb{E}[\epsilon_{it}\epsilon_{jt}]\right)^2 \leq \frac{1}{N}\sum_{i=1}^N \frac{1}{|Q_{ij}|^2}\sum_{s,t\in Q_{ij}} \tau_{ij}^2 \leq \frac{M}{N}.$$

Additionally, we have $\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}(\epsilon_{it}\epsilon_{jt} - \mathbb{E}[\epsilon_{it}\epsilon_{jt}]))^2\right] \leq M/T$. Therefore, the first term can be bounded by $M/\delta_{N,T}$.

Since $\mathbb{E}[\tilde{\Delta}_{it}^4] \leq M/\delta_{N,T}^2$, the second term can be bounded by

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\epsilon_{it}\tilde{\Delta}_{jt}\right)^2\right] \leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\epsilon_{it}^2\tilde{\Delta}_{jt}^2\right] \leq \frac{M}{\delta_{N,T}}\,.$$

Similarly, the last term satisfies $\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\tilde{\Delta}_{it}\tilde{\Delta}_{jt})^2\right] \leq M/\delta_{N,T}^2\,.$ □

Lemma 5. *Suppose Assumptions 1, 2, 3 hold and the first-stage estimation error* $\mathbb{E}[\tilde{\Delta}_{it}^4] \leq M/\delta_{N,T}^2$ *for any* i *and* t*. Then, as* $N, T \to \infty$,

1. $\frac{1}{N^2}\tilde{\Lambda}^\top\left((\dot{Y}\odot W)(\dot{Y}\odot W)^\top \odot \left[\frac{1}{|Q_{ij}|}\right]\right)\tilde{\Lambda} = \tilde{D} \xrightarrow{p} D$;
2. $\frac{1}{N^2}\tilde{\Lambda}^\top\left((\Lambda F^\top\odot W)(\Lambda F^\top\odot W)^\top \odot \left[\frac{1}{|Q_{ij}|}\right]\right)\tilde{\Lambda} \xrightarrow{p} D$;
3. $\frac{1}{N^2}\tilde{\Lambda}^\top\left(\Lambda\frac{F^\top F}{T}\Lambda^\top\right)\tilde{\Lambda} \xrightarrow{p} D$;

where $D = diag(d_1, \cdots, d_k)$ *is the diagonal matrix consisting of the eigenvalues of* $\Sigma_\Lambda\Sigma_F$.

Proof. Define $\Gamma = \{\lambda \in \mathbb{R}^{N\times 1} : \lambda^\top\lambda = N\}$, and let

$$\begin{aligned} R(\lambda) &= \frac{1}{N^2}\lambda^\top\left((\dot{Y}\odot W)(\dot{Y}\odot W)^\top \odot \left[\frac{1}{|Q_{ij}|}\right]\right)\lambda, \\ \tilde{R}(\lambda) &= \frac{1}{N^2}\lambda^\top\left((\Lambda F^\top\odot W)(\Lambda F^\top\odot W)^\top \odot \left[\frac{1}{|Q_{ij}|}\right]\right)\lambda, \\ R^*(\lambda) &= \frac{1}{N^2}\lambda^\top\left(\Lambda\frac{F^\top F}{T}\Lambda^\top\right)\lambda. \end{aligned}$$

We have the decomposition

$$R(\lambda) - R^*(\lambda) = R(\lambda) - \tilde{R}(\lambda) + \tilde{R}(\lambda) - R^*(\lambda).$$

For $R(\lambda) - \tilde{R}(\lambda)$, we have

$$\begin{aligned} \sup_{\lambda\in\Gamma}|R(\lambda) - \tilde{R}(\lambda)| \leq & \sup_{\lambda\in\Gamma}\frac{1}{N^2}\left|\lambda^\top\left((\dot{\epsilon}\odot W)(\dot{\epsilon}^\top\odot W^\top)\odot\left[\frac{1}{|Q_{ij}|}\right]\right)\lambda\right| \\ & + \sup_{\lambda\in\Gamma}\frac{2}{N^2}\left|\lambda^\top\left((\dot{\epsilon}\odot W)(F\Lambda^\top\odot W^\top)\odot\left[\frac{1}{|Q_{ij}|}\right]\right)\lambda\right|. \end{aligned}$$

The first part on the RHS satisfies

$$\begin{aligned}\sup_{\lambda\in\Gamma}\frac{1}{N^2}\left|\lambda^\top\left((\dot{\epsilon}\odot W)(\dot{\epsilon}^\top\odot W^\top)\odot\left[\frac{1}{|Q_{ij}|}\right]\right)\lambda\right| &= \sup_{\lambda\in\Gamma}\frac{1}{N^2}\left|\sum_{i,j=1}^{N}\lambda_i\lambda_j\gamma_{ij}\right| \\ &\le \sup_{\lambda\in\Gamma}\left(\frac{1}{N^2}\sum_{i,j=1}^{N}\lambda_i^2\lambda_j^2\right)^{1/2}\left(\frac{1}{N^2}\sum_{i,j=1}^{N}\gamma_{ij}^2\right)^{1/2} \\ &= o_p(1),\end{aligned}$$

where the last equality follows from Lemma 4. Similarly,

$$\sup_{\lambda\in\Gamma}\frac{2}{N^2}\left|\lambda^\top\left((\dot{\epsilon}\odot W)(F\Lambda^\top\odot W^\top)\odot\left[\frac{1}{|Q_{ij}|}\right]\right)\lambda\right| = \sup_{\lambda\in\Gamma}\frac{2}{N^2}\left|\sum_{i,j=1}^{N}\lambda_i\lambda_j\xi_{ij}\right| = o_p(1).$$

As a result, $\sup_{\lambda\in\Gamma}|R(\lambda)-\tilde{R}(\lambda)| \xrightarrow{p} 0$. In the rest of this proof, we follow similar steps as in Lemma 4 in Xiong and Pelger (2023) and sequentially show that

1. $\sup_{\lambda\in\Gamma}|\tilde{R}(\lambda)-R^*(\lambda)| \xrightarrow{p} 0$, $\sup_{\lambda\in\Gamma}|R(\lambda)-R^*(\lambda)| \xrightarrow{p} 0$
2. $|\sup_{\lambda\in\Gamma}R(\lambda)-\sup_{\lambda\in\Gamma}R^*(\lambda)| \xrightarrow{p} 0$
3. $\sup_{\lambda\in\Gamma}R^*(\lambda) \xrightarrow{p} d_1$, $\sup_{\lambda\in\Gamma}R(\lambda) \xrightarrow{p} d_1$
4. Let $\tilde{\Lambda}_1 = \arg\sup_{\lambda\in\Gamma}R(\lambda)$; then $\tilde{R}(\tilde{\Lambda}_1) \xrightarrow{p} d_1$, $R^*(\tilde{\Lambda}_1) \xrightarrow{p} d_1$
5. $R(\tilde{\Lambda}_i) \xrightarrow{p} d_i$, $\tilde{R}(\tilde{\Lambda}_i) \xrightarrow{p} d_i$, and $R^*(\tilde{\Lambda}_i) \xrightarrow{p} d_i$ for $i = 1,\cdots,k$.

□

Lemma 6. *Suppose Assumptions 1, 2, 3 hold and the first-stage estimation error* $\mathbb{E}[\tilde{\Delta}_{it}^4] \le M/\delta_{N,T}^2$ *for any* i *and* t*. Then, as* $N,T\to\infty$*,*

1. $\frac{1}{N}\tilde{\Lambda}^\top\Lambda \xrightarrow{p} Q = D^{1/2}\Upsilon\Sigma_F^{-1/2}$*, where the diagonal entries of* D *are eigenvalues of* $\Sigma_\Lambda\Sigma_F$*, and* Υ *is the corresponding eigenvector matrix such that* $\Upsilon^\top\Upsilon = I$*.*
2. $H \xrightarrow{p} (Q^\top)^{-1}$*, where* $H = \frac{1}{NT}\tilde{D}^{-1}\tilde{\Lambda}^\top\Lambda F^\top F$*.*

Proof. We left-multiply $\frac{1}{N}\tilde{\Sigma}\tilde{\Lambda} = \tilde{\Lambda}\tilde{D}$ on both sides by $\frac{1}{N}(\frac{1}{T}F^\top F)^{1/2}\Lambda^\top$ and obtain

$$\left(\frac{1}{T}F^\top F\right)^{1/2}\frac{1}{N^2}\Lambda^\top\tilde{\Sigma}\tilde{\Lambda} = \left(\frac{1}{T}F^\top F\right)^{1/2}\frac{1}{N}\Lambda^\top\tilde{\Lambda}\tilde{D}.$$

This can be written as

$$\left(\frac{1}{T}F^\top F\right)^{1/2}\frac{1}{N}\Lambda^\top\Lambda\cdot\frac{1}{T}F^\top F\cdot\frac{1}{N}\Lambda^\top\tilde{\Lambda} + d_{N,T} = \left(\frac{1}{T}F^\top F\right)^{1/2}\frac{1}{N}\Lambda^\top\tilde{\Lambda}\tilde{D}$$

with

$$d_{N,T} = \left(\frac{1}{T}F^\top F\right)^{1/2} \frac{1}{N}\Lambda^\top \left(\tilde{\Sigma} - \Lambda \frac{1}{T}F^\top F \Lambda^\top\right) \frac{1}{N}\tilde{\Lambda}.$$

For any i and j, $(\tilde{\Sigma} - \Lambda\frac{1}{T}F^\top F\Lambda^\top)_{ij}$ can be decomposed as $\Lambda_i^\top \Delta_{F,ij}\Lambda_j + \eta_{ij} + \xi_{ij} + \gamma_{ij}$, where $\Delta_{F,ij} \coloneqq \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t F_t^\top - \frac{1}{T}F^\top F$. Therefore, it holds that

$$\begin{aligned}
&\left\| \frac{1}{N}\Lambda^\top \left(\tilde{\Sigma} - \Lambda \frac{1}{T}F^\top F \Lambda^\top\right) \frac{1}{N}\tilde{\Lambda} \right\|^2 \\
= &\sum_{p,q=1}^{k} \frac{1}{N^2}\sum_{i,j=1}^{N} \Lambda_{ip}\tilde{\Lambda}_{jq}\left[\Lambda_i^\top \Delta_{F,ij}\Lambda_j + \eta_{ij} + \xi_{ij} + \gamma_{ij}\right] \\
\leq &\sum_{p,q=1}^{k} \left(\frac{1}{N^2}\sum_{i,j=1}^{N} \Lambda_{ip}^2\tilde{\Lambda}_{jq}^2\right)^{1/2} \left[\frac{1}{N^2}\sum_{i,j=1}^{N}(\Lambda^\top \Delta_{F,ij}\Lambda_j)^2 + \frac{1}{N^2}\sum_{i,j=1}^{N}\eta_{ij}^2 + \frac{1}{N^2}\sum_{i,j=1}^{N}\xi_{ij}^2 + \frac{1}{N^2}\sum_{i,j=1}^{N}\gamma_{ij}^2\right].
\end{aligned}$$

By Lemma 4, $d_{N,T} = O_p(1/\sqrt{\delta_{N,T}})$. We follow the same steps as Proposition 1 in Bai (2003) and show that

$$\frac{1}{N}\tilde{\Lambda}^\top \Lambda \xrightarrow{p} Q = D^{1/2}\Upsilon\Sigma_F^{-1/2}.$$

Furthermore, it holds that

$$H = \frac{1}{NT}\tilde{D}^{-1}\tilde{\Lambda}^\top \Lambda F^\top F \xrightarrow{p} D^{-1}D^{1/2}\Upsilon\Sigma_F^{-1/2}\Sigma_F = D^{-1/2}\Upsilon\Sigma_F^{1/2} = (Q^\top)^{-1}.$$

□

Lemma 7. *Let $\delta_{N,T} = \min(N,T)$. Suppose Assumptions 1, 4, 2 and 3 hold. For any t and $r = 1,\cdots,k$, we have*

$$\left|\mathbb{E}\left[\frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}W_{jt}\Lambda_{i,r}\Lambda_{j,r}\epsilon_{it}\tilde{\Delta}_{jt}\right]\right| \leq \frac{M}{\delta_{N,T}}.$$

Proof. We only show the proof for Case 1 with known observation probability, and the proof for the other two cases are similar. Based on Lemma 1, $\tilde{\Delta}_{it}$ can be decomposed as $\tilde{\Delta}_{it} = \tilde{\Delta}_{it,1} + \tilde{\Delta}_{it,2} - \tilde{\Delta}_{it,3}$. Since both of the idiosyncratic errors and factors are independent of the loadings, fixed effects and observation patterns, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}W_{jt}\Lambda_{i,r}\Lambda_{j,r}\epsilon_{it}\tilde{\Delta}_{jt,3}\right]\right| \\
=& \left|\mathbb{E}\left[\frac{1}{N^3}\sum_{i,j,l=1}^{N}\frac{1}{T}\sum_{s=1}^{T}\bar{W}_{j,\cdot}^{-1}W_{it}W_{jt}W_{js}X_{ls}D_s^{-1}\Lambda_{i,r}\Lambda_{j,r}\epsilon_{it}(\alpha_l+\Lambda_l^{\top}F_s+\epsilon_{ls})\right]\right| \\
\leq& \frac{1}{N^3}\sum_{i,j,l=1}^{N}\frac{1}{T}\sum_{s=1}^{T}\left|\mathbb{E}\left[\bar{W}_{j,\cdot}^{-1}W_{it}W_{jt}W_{js}X_{ls}D_s^{-1}\Lambda_{i,r}\Lambda_{j,r}\alpha_l\right]\cdot\mathbb{E}[\epsilon_{it}]\right| \\
&+\frac{1}{N^3}\sum_{i,j,l=1}^{N}\frac{1}{T}\sum_{s=1}^{T}\sum_{p=1}^{k}\left|\mathbb{E}\left[\bar{W}_{j,\cdot}^{-1}W_{it}W_{jt}W_{js}X_{ls}D_s^{-1}\Lambda_{i,r}\Lambda_{j,r}\Lambda_{l,p}\right]\cdot\mathbb{E}[F_{s,p}\epsilon_{it}]\right| \\
&+\frac{1}{N^3}\sum_{i,j,l=1}^{N}\frac{1}{T}\sum_{s=1}^{T}\left|\mathbb{E}\left[\bar{W}_{j,\cdot}^{-1}W_{it}W_{jt}W_{js}X_{ls}D_s^{-1}\Lambda_{i,r}\Lambda_{j,r}\right]\cdot\mathbb{E}[\epsilon_{it}\epsilon_{ls}]\right| \\
\leq& \frac{M}{NT}\sum_{i=1}^{N}\sum_{s=1}^{T}\sum_{p=1}^{k}|\mathbb{E}[F_{s,p}\epsilon_{it}]|+\frac{M}{N^2T}\sum_{i,l=1}^{N}\sum_{s=1}^{T}|E[\epsilon_{it}\epsilon_{ls}]| \\
\leq& \frac{M}{T}.
\end{aligned}$$

By similar arguments, we can show the bounds $\left|\mathbb{E}[\frac{1}{N^2}\sum_{i,j=1}^{N}W_{it}W_{jt}\Lambda_{i,r}\Lambda_{j,r}\epsilon_{it}\tilde{\Delta}_{jt,1}]\right|\leq M/\delta_{N,T}$ and $\left|\mathbb{E}[\frac{1}{N^2}\sum_{i,j=1}^{N}W_{it}W_{jt}\Lambda_{i,r}\Lambda_{j,r}\epsilon_{it}\tilde{\Delta}_{jt,2}]\right|\leq M/\delta_{N,T}$. □

Proof of Theorem 1:

Proof. Step 1 – Consistency of loadings

Observe that for any unit i,

$$\left\|\tilde{\Lambda}_i-H\Lambda_i\right\|^2\leq\left\|\tilde{\Lambda}_i-H_i\Lambda_i\right\|^2+\|(H_i-H)\Lambda_i\|^2.$$

We can decompose the first term on the RHS as

$$\tilde{\Lambda}_i-H_i\Lambda_i=\tilde{D}^{-1}\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\eta_{ij}+\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\xi_{ij}+\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\gamma_{ij}\right).$$

By Lemma 4 and Lemma 5, this implies that

$$\begin{aligned}\left\|\tilde{\Lambda}_i - H_i\Lambda_i\right\|^2 &\leq 3\left\|\tilde{D}^{-1}\right\|^2 \cdot \left(\left\|\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\eta_{ij}\right\|^2 + \left\|\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\xi_{ij}\right\|^2 + \left\|\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\gamma_{ij}\right\|^2\right)\\ &\leq 3\left\|\tilde{D}^{-1}\right\|^2 \cdot \frac{1}{N}\sum_{j=1}^{N}\left\|\tilde{\Lambda}_j\right\|^2 \cdot \left(\frac{1}{N}\sum_{j=1}^{N}\eta_{ij}^2 + \frac{1}{N}\sum_{j=1}^{N}\xi_{ij}^2 + \frac{1}{N}\sum_{j=1}^{N}\gamma_{ij}^2\right)\\ &= O_p(\frac{1}{\delta_{N,T}}).\end{aligned}$$

The second term $\|(H_i - H)\Lambda_i\|^2$ can be bounded by

$$\begin{aligned}\|(H_i - H)\Lambda_i\|^2 &= \left\|\tilde{D}^{-1}\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\Lambda_j^\top\left(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_tF_t^\top - \frac{1}{T}\sum_{t=1}^{T}F_tF_t^\top\right)\Lambda_i\right\|^2\\ &\leq \left\|\tilde{D}^{-1}\right\|^2\|\Lambda_i\|^2 \cdot \frac{1}{N}\sum_{j=1}^{N}\left\|\tilde{\Lambda}_j\right\|^2 \cdot \frac{1}{N}\sum_{j=1}^{N}\|\Lambda_j\|^2\|\Delta_{F,ij}\|^2,\end{aligned}$$

where $\Delta_{F,ij} = \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_tF_t^\top - \frac{1}{T}\sum_{t=1}^{T}F_tF_t^\top$. Following Assumption 2, we have

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\|\Lambda_j\|^2\|\Delta_{F,ij}\|^2\right] = \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\|\Lambda_j\|^2 \cdot \mathbb{E}\|\Delta_{F,ij}\|^2 \leq \frac{M}{T}.$$

Therefore, $\|(H_i - H)\Lambda_i\|^2 = O_p(\frac{1}{T})$, and thus,

$$\left\|\tilde{\Lambda}_i - H\Lambda_i\right\|^2 = O_p(\frac{1}{\delta_{N,T}}) + O_p(\frac{1}{T}) = O_p(\frac{1}{\delta_{N,T}}).$$

Step 2 – Consistency of factors

We derive the estimated factors $\tilde{F}_t$ by regressing the observed $\dot{Y}_{it}$ on $\tilde{\Lambda}_i$:

$$\tilde{F}_t = \left(\frac{1}{N}\sum_{i=1}^{N}W_{it}\tilde{\Lambda}_i\tilde{\Lambda}_i^\top\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}W_{it}\tilde{\Lambda}_i\dot{Y}_{it}\right), \quad t = 1, \cdots, T.$$

We define an auxiliary $\tilde{F}_t^*$ as

$$\tilde{F}_t^* := \left(\frac{1}{N}\sum_{i=1}^{N}W_{it}H\Lambda_i\Lambda_i^\top H^\top\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}W_{it}\tilde{\Lambda}_i\dot{Y}_{it}\right).$$

We have the decomposition

$$\left\|\tilde{F}_t - (H^\top)^{-1}F_t\right\|^2 \leq \left\|\tilde{F}_t - \tilde{F}_t^*\right\|^2 + \left\|\tilde{F}_t^* - (H^\top)^{-1}F_t\right\|^2.$$

We first analyze the second term $\|\tilde{F}_t^* - (H^\top)^{-1}F_t\|^2$. Observe that $\tilde{F}_t^*$ can be expanded as

$$\begin{aligned}\tilde{F}_t^* = (H^\top)^{-1}F_t + (H^\top)^{-1}\hat{\Sigma}_{\Lambda,t}^{-1} \cdot \frac{1}{N}\sum_{i=1}^N W_{it}\Lambda_i\dot{\epsilon}_{it} \\ + (H^\top)^{-1}\hat{\Sigma}_{\Lambda,t}^{-1}H^{-1} \cdot \frac{1}{N}\sum_{i=1}^N W_{it}(\tilde{\Lambda}_i - H\Lambda_i)(\Lambda_i^\top F_t + \dot{\epsilon}_{it}),\end{aligned}$$

where $\hat{\Sigma}_{\Lambda,t} \coloneqq \frac{1}{N}\sum_{i=1}^N W_{it}\Lambda_i\Lambda_i^\top$. By Assumption 2.2, it holds that $\hat{\Sigma}_{\Lambda,t} \xrightarrow{p} \Sigma_{\Lambda,t}$ for some positive definite matrix $\Sigma_{\Lambda,t}$. Thus, it follows that

$$\begin{aligned}\left\|\tilde{F}_t^* - (H^\top)^{-1}F_t\right\|^2 \leq 2\underbrace{\left\|(H^\top)^{-1}\hat{\Sigma}_{\Lambda,t}^{-1}\right\|}_{O_p(1)} \cdot \left\|\frac{1}{N}\sum_{i=1}^N W_{it}\Lambda_i\dot{\epsilon}_{it}\right\|^2 \\ + 2\underbrace{\left\|(H^\top)^{-1}\hat{\Sigma}_{\Lambda,t}^{-1}H^{-1}\right\|}_{O_p(1)} \cdot \left\|\frac{1}{N}\sum_{i=1}^N W_{it}(\tilde{\Lambda}_i - H\Lambda_i)(\Lambda_i^\top F_t + \dot{\epsilon}_{it})\right\|^2.\end{aligned}$$

For the first part on the RHS, we have

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N W_{it}\Lambda_i\dot{\epsilon}_{it}\right\|^2 &= \sum_{r=1}^k \mathbb{E}\left[\frac{1}{N^2}\sum_{i,j=1}^N W_{it}W_{jt}\Lambda_{i,r}\Lambda_{j,r}\dot{\epsilon}_{it}\dot{\epsilon}_{jt}\right] \\ &= \sum_{r=1}^k \mathbb{E}\left[\frac{1}{N^2}\sum_{i,j=1}^N W_{it}W_{jt}\Lambda_{i,r}\Lambda_{j,r}\epsilon_{it}\epsilon_{jt}\right] - 2\sum_{r=1}^k \mathbb{E}\left[\frac{1}{N^2}\sum_{i,j=1}^N W_{it}W_{jt}\Lambda_{i,r}\Lambda_{j,r}\epsilon_{it}\tilde{\Delta}_{jt}\right] \\ &\quad + \sum_{r=1}^k \mathbb{E}\left[\frac{1}{N^2}\sum_{i,j=1}^N W_{it}W_{jt}\Lambda_{i,r}\Lambda_{j,r}\tilde{\Delta}_{it}\tilde{\Delta}_{jt}\right].\end{aligned}$$

By Assumption 2.3, $\mathbb{E}\left[\frac{1}{N^2}\sum_{i,j=1}^N W_{it}W_{jt}\Lambda_{i,r}\Lambda_{j,r}\epsilon_{it}\epsilon_{jt}\right] = \frac{1}{N^2}\sum_{i,j=1}^N \mathbb{E}\left[W_{it}W_{jt}\Lambda_{i,r}\Lambda_{j,r}\right] \cdot \mathbb{E}\left[\epsilon_{it}\epsilon_{jt}\right] \leq M/N$. By Lemma 7, $\left|\mathbb{E}\left[\frac{1}{N^2}\sum_{i,j=1}^N W_{it}W_{jt}\Lambda_{i,r}\Lambda_{j,r}\epsilon_{it}\tilde{\Delta}_{jt}\right]\right| \leq M/\delta_{N,T}$. For the last term, there is

$$\mathbb{E}\left[\frac{1}{N^2}\sum_{i,j=1}^N W_{it}W_{jt}\Lambda_{i,r}\Lambda_{j,r}\tilde{\Delta}_{it}\tilde{\Delta}_{jt}\right] \leq \frac{1}{N^2}\sum_{i,j=1}^N \left(\mathbb{E}[\Lambda_{i,r}^4] \cdot \mathbb{E}[\tilde{\Delta}_{it}^4]\right)^{1/2} \leq \frac{M}{\delta_{N,T}}.$$

Therefore, we get that $\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N W_{it}\Lambda_i\dot{\epsilon}_{it}\right\|^2 \leq M/\delta_{N,T}$.

Moreover, based on the consistency results for the loadings, we have

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{i=1}^{N}W_{it}(\tilde{\Lambda}_i - H\Lambda_i)(\Lambda_i^\top F_t + \dot{\epsilon}_{it})\right\|^2\\ \leq&\frac{1}{N}\sum_{i=1}^{N}(\Lambda_i^\top F_t + \dot{\epsilon}_{it})^2 \cdot \frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\Lambda}_i - H\Lambda_i\right\|^2 = O_p(\frac{1}{\delta_{N,T}}).\end{aligned}$$

Thus, we have derived that $\|\tilde{F}_t^* - (H^\top)^{-1}F_t\|^2 = O_p(\frac{1}{\delta_{N,T}})$.

Next, we consider the term $\|\tilde{F}_t - \tilde{F}_t^*\|^2$. Observe that

$$\tilde{F}_t^* - \tilde{F}_t = \left(\frac{1}{N}\sum_{i=1}^{N}W_{it}\tilde{\Lambda}_i\tilde{\Lambda}_i^\top\right)^{-1}\underbrace{\left[\frac{1}{N}\sum_{i=1}^{N}W_{it}\tilde{\Lambda}_i\tilde{\Lambda}_i^\top - \frac{1}{N}\sum_{i=1}^{N}W_{it}H\Lambda_i\Lambda_i^\top H^\top\right]}_{\tilde{\Delta}_{\Lambda,t}}\tilde{F}_t^*.$$

We further decompose $\tilde{\Delta}_{\Lambda,t}$ as

$$\tilde{\Delta}_{\Lambda,t} = \frac{1}{N}\sum_{i=1}^{N}\left[W_{it}H\Lambda_i(\tilde{\Lambda}_i - H\Lambda_i)^\top + W_{it}(\tilde{\Lambda}_i - H\Lambda_i)(H\Lambda_i)^\top + W_{it}(\tilde{\Lambda}_i - H\Lambda_i)(\tilde{\Lambda}_i - H\Lambda_i)^\top\right].$$

Based on the consistency results for loadings, $\tilde{\Delta}_{\Lambda,t}$ can be bounded by

$$\begin{aligned}\|\tilde{\Delta}_{\Lambda,t}\|^2 \leq{}& 6\cdot\frac{1}{N}\sum_{i=1}^{N}\|H\Lambda_i\|^2\cdot\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\Lambda}_i - H\Lambda_i\right\|^2\\ &+3\cdot\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\Lambda}_i - H\Lambda_i\right\|^2\cdot\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\Lambda}_i - H\Lambda_i\right\|^2 = O_p(\frac{1}{\delta_{N,T}}).\end{aligned}$$

Therefore, according to Lemma 6, $\tilde{\Sigma}_{\Lambda,t} := \frac{1}{N}\sum_{i=1}^{N}W_{it}\tilde{\Lambda}_i\tilde{\Lambda}_i^\top \xrightarrow{p} (Q^\top)^{-1}\Sigma_{\Lambda,t}Q^{-1}$. Furthermore, since $\tilde{F}_t^* = O_p(1)$ and $\tilde{\Sigma}_{\Lambda,t} = O_p(1)$, it follows that

$$\begin{aligned}\left\|\tilde{F}_t - \tilde{F}_t^*\right\|^2 \leq{}& \left\|\tilde{\Sigma}_{\Lambda,t}\right\|^2\left\|\tilde{\Delta}_{\Lambda,t}\right\|^2\left\|\tilde{F}_t^*\right\|^2\\ &\leq 6\left\|\tilde{\Sigma}_{\Lambda,t}\right\|^2\left\|\tilde{F}_t^*\right\|^2\cdot\frac{1}{N}\sum_{i=1}^{N}\|H\Lambda_i\|^2\cdot\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\Lambda}_i - H\Lambda_i\right\|^2\\ &\quad+3\left\|\tilde{\Sigma}_{\Lambda,t}\right\|^2\left\|\tilde{F}_t^*\right\|^2\cdot\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\Lambda}_i - H\Lambda_i\right\|^2\cdot\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\Lambda}_i - H\Lambda_i\right\|^2\\ &= O_p(\frac{1}{\delta_{N,T}}).\end{aligned}$$

Therefore, it holds that

$$\left\|\tilde{F}_t-(H^\top)^{-1}F_t\right\|^2 \leq \left\|\tilde{F}_t-\tilde{F}_t^*\right\|^2+\left\|\tilde{F}_t^*-(H^\top)^{-1}F_t\right\|^2=O_p(\frac{1}{\delta_{N,T}}).$$

Step 3 – Consistency of common components:

We have

$$\begin{aligned}\tilde{C}_{it}-C_{it}&=(\tilde{\mu}+\tilde{\alpha}_i+\tilde{\xi}_t+\tilde{\Lambda}_i^\top\tilde{F}_t)-(\mu+\alpha_i+\xi_t+\Lambda_i^\top F_t)\\&=\tilde{\Delta}_{it}+\tilde{\Lambda}_i^\top\left(\tilde{F}_t-(H^\top)^{-1}F_t\right)+\left(\tilde{\Lambda}_i-H\Lambda_i\right)^\top(H^\top)^{-1}F_t.\end{aligned}$$

Based on the consistency results of loadings, factors and first-stage estimation, it holds that

$$\begin{aligned}(\tilde{C}_{it}-C_{it})^2=&\left(\tilde{\Delta}_{it}+\tilde{\Lambda}_i^\top\left(\tilde{F}_t-(H^\top)^{-1}F_t\right)+\left(\tilde{\Lambda}_i-H\Lambda_i\right)^\top(H^\top)^{-1}F_t\right)^2\\&\leq 3\tilde{\Delta}_{it}^2+3\left\|\tilde{\Lambda}_i\right\|^2\cdot\left\|\tilde{F}_t-(H^\top)^{-1}F_t\right\|^2+3\left\|\tilde{\Lambda}_i-H\Lambda_i\right\|^2\cdot\left\|(H^\top)^{-1}F_t\right\|^2\\&=O_p(\frac{1}{\delta_{N,T}}).\end{aligned}$$

□

IA.C.2 Proof of Theorem 2: Asymptotic Normality

First, we introduce the martingale central limit theorem that we will use in the main proof.

Theorem 1. *Let $\{S_{ni},\mathcal{F}_{ni},1\leq i\leq n,n\geq 1\}$ be a zero-mean, square-integrable martingale array with differences X_{ni}, and let η^2 be an a.s. finite r.v. Suppose that*

$$\begin{aligned}&\sum_i\mathbb{E}\left[X_{ni}^2\mid\mathcal{F}_{n,i-1}\right]\xrightarrow{p}\eta^2,\\&\sum_i\mathbb{E}\left[|X_{ni}|^{2+\delta}\mid\mathcal{F}_{n,i-1}\right]\xrightarrow{p}0\quad\text{for some }\delta>0,\end{aligned}$$

and the σ-fields are nested: $\mathcal{F}_{n,i}\subseteq\mathcal{F}_{n,i+1}$ for $1\leq i\leq n$. Then $S_{nn}=\sum_i X_{ni}\xrightarrow{d}Z$ stably, where the r.v. Z has characteristic function $\mathbb{E}[\exp(-\frac{1}{2}\eta^2t^2)]$.

Proof. It holds for any $\epsilon > 0$ that

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left[X_{ni}^2 \cdot I(|X_{ni}| > \epsilon) \mid \mathcal{F}_{n,i-1}\right] \\
\leq\ & \sum_{i=1}^{n} \mathbb{E}\left[\frac{|X_{ni}|^{2+\delta}}{\epsilon^\delta} \cdot I(|X_{ni}| > \epsilon) \mid \mathcal{F}_{n,i-1}\right] \\
\leq\ & \frac{1}{\epsilon^\delta} \sum_{i=1}^{n} \mathbb{E}\left[|X_{ni}|^{2+\delta} \mid \mathcal{F}_{n,i-1}\right].
\end{aligned}$$

Therefore, $\sum_i \mathbb{E}\left[|X_{ni}|^{2+\delta} \mid \mathcal{F}_{n,i-1}\right] \xrightarrow{p} 0$ implies that $\sum_{i=1}^n \mathbb{E}\left[X_{ni}^2 \cdot I(|X_{ni}| > \epsilon) \mid \mathcal{F}_{n,i-1}\right] \xrightarrow{p} 0$. By Corollary 3.1 in Hall and Heyde (2014), this completes our proof. □

IA.C.2.1 Asymptotic normality for the sum of grand mean and fixed effects

Lemma 8. *Suppose Assumptions 1, 5, 6 and Case 1 in Assumption 4 hold. When $N, T \to \infty$, the asymptotic distribution of $\tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i$ estimated by wi-PCA with weights (4) is:*

$$\sqrt{\delta_{N,T}} \cdot \sigma_{it}^{-1} \left((\tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i) - (\mu + \xi_t + \alpha_i)\right) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\sigma_{it}^2 = \frac{\delta_{N,T}}{N}\sigma_{it,1}^2 + \frac{\delta_{N,T}}{T}\sigma_{it,2}^2$ with some $\sigma_{it,1}$ and $\sigma_{it,2}$. Moreover,

$$\mathbb{E}\left[\left((\tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i) - (\mu + \xi_t + \alpha_i)\right)^4\right] \leq \frac{M}{\delta_{N,T}^2}.$$

Proof. We denote the re-weighted observation indicators by $X_{it} = W_{it}/p_{it}$. According to Lemma 1, the estimation error of $\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t$ is

$$\begin{aligned}
\tilde{\Delta}_{it} =\ & \underbrace{D_t^{-1} \cdot \frac{1}{N}\sum_{j=1}^{N} X_{jt}\left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt}\right)}_{\tilde{\Delta}_{it,1}} + \underbrace{\bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T}\sum_{s=1}^{T} W_{is}\left(\Lambda_i^\top F_s + \epsilon_{is}\right)}_{\tilde{\Delta}_{it,2}} \\
& \underbrace{- \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} W_{is} X_{js} D_s^{-1}\left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right)}_{\tilde{\Delta}_{it,3}}.
\end{aligned}$$

We decompose the first term $\tilde{\Delta}_{it,1}$ as

$$\tilde{\Delta}_{it,1} = D_t^{-1} \cdot \frac{1}{N}\sum_{j=1}^{N}(X_{jt} - 1)\left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt}\right) + D_t^{-1} \cdot \frac{1}{N}\sum_{j=1}^{N}\left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt}\right).$$

Similarly, we decompose the third term $\tilde{\Delta}_{it,3}$ as

$$\begin{aligned}\tilde{\Delta}_{it,3} = & \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} X_{js} \left(D_s^{-1} - 1\right) \left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right) \\ & + \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \left(X_{js} - 1\right) \left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right) \\ & + \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right).\end{aligned}$$

Since $\alpha_j, \Lambda_j, \epsilon_{js}$ are i.i.d. and $\mathbb{E}[(D_t^{-1} - 1)^4 \mid I] \leq M/N^2$, the first part on the RHS of $\tilde{\Delta}_{it,3}$ can be bounded by

$$\begin{aligned} & \mathbb{E}\left[\left(\bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} X_{js} \left(D_s^{-1} - 1\right) \left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right)\right)^4\right] \\ \leq & M \cdot \mathbb{E}\left[\frac{1}{T} \sum_{s=1}^{T} \mathbb{E}\left[\left(D_s^{-1} - 1\right)^4 \mid I\right] \cdot \frac{1}{T} \sum_{s=1}^{T} \mathbb{E}\left[\left(\frac{1}{N} \sum_{j=1}^{N} \left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right)\right)^4 \mid I\right]\right] \\ \leq & \frac{M}{N^2} \cdot \frac{1}{T} \sum_{s=1}^{T} \mathbb{E}\left[\left(\frac{1}{N} \sum_{j=1}^{N} \left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right)\right)^4\right] \\ \leq & \frac{M}{N^4}. \end{aligned}$$

For the second part of $\tilde{\Delta}_{it,3}$, we have

$$\begin{aligned} & \mathbb{E}\left[\left(\frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \left(X_{js} - 1\right) \left(\Lambda_j^\top F_s + \epsilon_{js}\right)\right)^4\right] \\ = \; & \frac{3}{N^4 T^4} \sum_{j,l,h,o=1}^{N} \sum_{s,u=1}^{T} \mathbb{E}[W_{is} W_{iu} (X_{js} - 1)(X_{ls} - 1)(X_{hu} - 1)(X_{ou} - 1)(\Lambda_j^\top F_s + \epsilon_{js}) \\ & (\Lambda_l^\top F_s + \epsilon_{ls})(\Lambda_h^\top F_u + \epsilon_{hu})(\Lambda_o^\top F_u + \epsilon_{ou})] \leq \frac{M}{N^2 T^2}, \end{aligned}$$

where the last inequality holds because $\mathbb{E}[W_{is} W_{iu} (X_{js} - 1)(X_{ls} - 1)(X_{hu} - 1)(X_{ou} - 1)|I] = 0$ when i, j, l, h, o are distinct or j, l, h, o take three different values other than i. For the third part in the decomposition of $\tilde{\Delta}_{it,3}$, we have $\frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \alpha_j = \bar{W}_{i,\cdot} \frac{1}{N} \sum_{j=1}^{N} \alpha_j$. Furthermore, it is easy to see that $\frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} (\Lambda_j^\top F_s + \epsilon_{js}) = O_p(\frac{1}{\delta_{N,T}})$ and its fourth moment can be bounded by M/T^2. Therefore, we conclude that

$$\tilde{\Delta}_{it,3} = \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \frac{1}{T} \sum_{s=1}^{T} W_{is} (X_{js} - 1) \alpha_j + \frac{1}{N} \sum_{j=1}^{N} \alpha_j + O_p(\frac{1}{\delta_{N,T}}).$$

Combining the three terms $\tilde{\Delta}_{it,1}$, $\tilde{\Delta}_{it,2}$ and $\tilde{\Delta}_{it,3}$, we can show that

$$\tilde{\Delta}_{it} = D_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} (X_{jt} - 1) \Big(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \Big) - \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T} \sum_{s=1}^{T} W_{is} \frac{1}{N} \sum_{j=1}^{N} (X_{js} - 1) \alpha_j$$

$$+ \bar{W}_{i,\cdot}^{-1} \cdot \Big(\frac{1}{T} \sum_{s=1}^{T} W_{is} \epsilon_{is} + \Lambda_i^\top \frac{1}{T} \sum_{s=1}^{T} W_{is} F_s \Big) + D_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \Big(\Lambda_j^\top F_t + \epsilon_{jt} \Big) + O_p(\frac{1}{\delta_{N,T}}).$$

This proves that $\mathbb{E}[\tilde{\Delta}_{it}^4] \leq M/\delta_{N,T}^2$.

We first analyze the asymptotic distribution of $\frac{1}{N}\sum_{j=1}^{N}(X_{jt}-1)\alpha_j$. Conditional on I, $(X_{jt}-1)\alpha_j$ and $(X_{lt}-1)\alpha_l$ are independent for any $j \neq l$. The expectation satisfies $\mathbb{E}[(X_{jt}-1)\alpha_j|I] = 0$. By Assumption 6.2, as $N \to \infty$, it holds that

$$\frac{1}{N} \sum_{j=1}^{N} \mathbb{E}\left[(X_{jt}-1)^2 \alpha_j^2 | I\right] = \frac{1}{N} \sum_{j=1}^{N} \left(\mathbb{E}\left[X_{jt}^2 \mid I\right] - 1 \right) \alpha_j^2 \xrightarrow{p} s_{\alpha,tt} - \sigma_\alpha^2.$$

Furthermore, we have

$$\sum_{j=1}^{N} \mathbb{E}\left[\left(\frac{1}{\sqrt{N}} (X_{jt}-1)\alpha_j \right)^4 \mid I \right] = \frac{1}{N^2} \sum_{j=1}^{N} \alpha_j^4 \cdot \mathbb{E}\left[(X_{jt}-1)^4 \mid I\right] \leq \frac{M}{N} \cdot \frac{1}{N} \sum_{j=1}^{N} \alpha_j^4,$$

where the inequality holds because $|X_{jt}|$ is bounded. By the law of large number, we have $\frac{1}{N}\sum_{j=1}^{N} \alpha_j^4 \xrightarrow{p} \mathbb{E}[\alpha_j^4] \leq M$. As a result, $\sum_{j=1}^{N} \mathbb{E}\left[(\frac{1}{\sqrt{N}}(X_{jt}-1)\alpha_j)^4 \mid I \right] \xrightarrow{p} 0$. According to Theorem 1, $\frac{1}{\sqrt{N}}\sum_{j=1}^{N}(X_{jt}-1)\alpha_j \xrightarrow{d} \mathcal{N}(0, s_{\alpha,tt} - \sigma_\alpha^2)$.

Similarly, we can prove the martingale CLT for the other two terms $\frac{1}{N}\sum_{j=1}^{N}(X_{jt}-1)\Lambda_j^\top F_t$ and $\frac{1}{N}\sum_{j=1}^{N}(X_{jt}-1)\epsilon_{jt}$. According to Assumption 6.2,

$$\frac{1}{N} \sum_{j=1}^{N} \mathbb{E}\left[(X_{jt}-1)^2 \Lambda_j \Lambda_j^\top \mid I \right] = \frac{1}{N} \sum_{j=1}^{N} \left(\mathbb{E}\left[X_{jt}^2 \mid I\right] - 1 \right) \Lambda_j \Lambda_j^\top \xrightarrow{p} s_{\Lambda,tt} - \Sigma_\Lambda,$$

and

$$\frac{1}{N} \sum_{j=1}^{N} \mathbb{E}\left[(X_{jt}-1)^2 \epsilon_{jt}^2 \mid I\right] = \sigma_\epsilon^2 \cdot \frac{1}{N} \sum_{j=1}^{N} \left(\mathbb{E}\left[X_{jt}^2 \mid I\right] - 1 \right) \xrightarrow{p} (s_t - 1)\sigma_\epsilon^2.$$

By similar arguments, we get stable convergence in law for $\frac{1}{\sqrt{N}}\sum_{j=1}^{N}(X_{jt}-1)\Lambda_j^\top F_t \xrightarrow{d} \mathcal{N}(0, F_t^\top (s_{\Lambda,tt} - \Sigma_\Lambda) F_t)$ and $\frac{1}{\sqrt{N}}\sum_{j=1}^{N}(X_{jt}-1)\epsilon_{jt} \xrightarrow{d} \mathcal{N}(0, \sigma_\epsilon^2(s_t - 1))$.

Next, we analyze the asymptotic behavior of $\frac{1}{T}\sum_{s=1}^{T} W_{is} \frac{1}{N}\sum_{j=1}^{N}(X_{js}-1)\alpha_j$. We denote $G_s = \frac{1}{N}\sum_{j=1}^{N}(X_{js}-1)\alpha_j$. Note that $\left[\sqrt{N}G_1, \cdots, \sqrt{N}G_T\right]$ is jointly asymptotically normal. For any

s, u, the asymptotic covariance between $\sqrt{N} G_s$ and $\sqrt{N} G_u$ is

$$\mathrm{ACov}(\sqrt{N} G_s, \sqrt{N} G_u) = \mathrm{plim}_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \alpha_j^2 \cdot (\mathbb{E}\left[X_{js} X_{ju} \mid I\right] - 1) = s_{\alpha,su} - \sigma_\alpha^2.$$

Furthermore, $\mathrm{plim}_{T\to\infty} \frac{1}{T^2} \sum_{s,u=1}^{T} W_{is} W_{iu} s_{\alpha,su}$ exists by Assumption 6. As a result, we have that $\frac{1}{T} \sum_{s=1}^{T} W_{is} \frac{1}{N} \sum_{j=1}^{N} (X_{js} - 1)\alpha_j$ is asymptotically normal with zero mean. Note that the above four terms are jointly asymptotically normal because their randomness comes from the cross-sectional missing pattern.

Next, we analyze the term $\frac{1}{T} \sum_{s=1}^{T} W_{is} \epsilon_{is}$. Conditional on $W_i = \{W_{is}, s \le T\}$, $W_{is}\epsilon_{is}$ and $W_{it}\epsilon_{it}$ are independent for any $s \neq t$. We have $\mathbb{E}[W_{is}\epsilon_{is} | W_i] = 0$, and by Assumption 6.1,

$$\frac{1}{T} \sum_{s=1}^{T} \mathbb{E}\left[W_{is}^2 \epsilon_{is}^2 \mid W_i\right] = \bar{W}_{i,\cdot} \sigma_\epsilon^2 \xrightarrow{p} q_{ii} \sigma_\epsilon^2.$$

Furthermore, $\frac{1}{T^2} \sum_{s=1}^{T} \mathbb{E}\left[W_{is} \epsilon_{is}^4 \mid W_i\right] \xrightarrow{p} 0$. Therefore, $\frac{1}{\sqrt{T}} \sum_{s=1}^{T} W_{is}\epsilon_{is} \xrightarrow{d} \mathcal{N}(0, q_{ii}\sigma_\epsilon^2)$. Similarly, we can prove that $\frac{1}{\sqrt{T}} \sum_{s=1}^{T} W_{is} \Lambda_i^\top F_s \xrightarrow{d} \mathcal{N}(0, q_{ii} \Lambda_i^\top \Sigma_F \Lambda_i)$. Note that these two terms are asymptotically independent because the randomness of $\frac{1}{\sqrt{T}} \sum_{s=1}^{T} W_{is}\epsilon_{is}$ comes from idiosyncratic errors while the randomness of $\frac{1}{\sqrt{T}} \sum_{s=1}^{T} W_{is} \Lambda_i^\top F_s$ comes from the factors.

Finally, for $\frac{1}{N} \sum_{j=1}^{N} (\Lambda_j^\top F_t + \epsilon_{jt})$, we have that $\frac{1}{\sqrt{N}} \sum_{j=1}^{N} (\Lambda_j^\top F_t + \epsilon_{jt}) \xrightarrow{d} \mathcal{N}(0, F_t^\top \Sigma_\Lambda F_t + \sigma_\epsilon^2)$. Since $D_t = \frac{1}{N} \sum_{i=1}^{N} X_{it} \xrightarrow{p} 1$ and $\bar{W}_{i,\cdot} \xrightarrow{p} q_{ii}$, using Slutsky's theorem we can complete our proof. □

Lemma 9. *Suppose Assumptions 1, 5, 6 and Case 2 in Assumption 4 hold. When $N, T \to \infty$ and $\sqrt{N}/T \to 0$, the asymptotic distribution of $\tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i$ estimated by wi-PCA with weights* (5) *is:*

$$\sqrt{\delta_{N,T}} \cdot \sigma_{it}^{-1} \left((\tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i) - (\mu + \xi_t + \alpha_i) \right) \xrightarrow{d} \mathcal{N}\left(0, 1\right),$$

where $\sigma_{it}^2 = \frac{\delta_{N,T}}{N} \sigma_{it,1}^2 + \frac{\delta_{N,T}}{T} \sigma_{it,2}^2$ with some $\sigma_{it,1}$ and $\sigma_{it,2}$. Moreover,

$$\mathbb{E}\left[\left((\tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i) - (\mu + \xi_t + \alpha_i) \right)^4 \right] \le \frac{M}{\delta_{N,T}^2}.$$

Proof. According to Lemma 2, the estimation error of $\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t$ can be expanded as

$$\tilde{\Delta}_{it} = \underbrace{\tilde{D}_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot} v_t} \left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right)}_{\tilde{\Delta}_{it,1}} + \underbrace{\bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T} \sum_{s=1}^{T} W_{is} \left(\Lambda_i^\top F_s + \epsilon_{is} \right)}_{\tilde{\Delta}_{it,2}}$$
$$\underbrace{- \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot} v_s} \tilde{D}_s^{-1} \left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js} \right)}_{\tilde{\Delta}_{it,3}},$$

where $\tilde{D}_t = \frac{1}{N} \sum_{i=1}^{N} W_{it}\bar{v} / (\bar{W}_{i,\cdot} v_t)$.

Let $\tilde{\Delta}_{it}^{\text{known}}$ denote the first-stage estimation error of the wi-PCA estimator with weights (4) in Lemma 8 for known observation probabilities. According to Lemma 8, we have

$$\tilde{\Delta}_{it}^{\text{known}} = \underbrace{D_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \frac{W_{jt}}{p_{jt}} \left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right)}_{\tilde{\Delta}_{it,1}^{\text{known}}} + \underbrace{\bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T} \sum_{s=1}^{T} W_{is} \left(\Lambda_i^\top F_s + \epsilon_{is} \right)}_{\tilde{\Delta}_{it,2}^{\text{known}}}$$
$$\underbrace{- \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \frac{W_{js}}{p_{js}} D_s^{-1} \left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js} \right)}_{\tilde{\Delta}_{it,3}^{\text{known}}}.$$

Observe that $\tilde{\Delta}_{it,2} = \tilde{\Delta}_{it,2}^{\text{known}}$, so

$$\tilde{\Delta}_{it} - \tilde{\Delta}_{it}^{\text{known}} = (\tilde{\Delta}_{it,1} - \tilde{\Delta}_{it,1}^{\text{known}}) - (\tilde{\Delta}_{it,3} - \tilde{\Delta}_{it,3}^{\text{known}}).$$

Next, we analyze the two differences separately.

First, the difference $\tilde{\Delta}_{it,1} - \tilde{\Delta}_{it,1}^{\text{known}}$ has the decomposition

$$\begin{aligned}
\tilde{\Delta}_{it,1} - \tilde{\Delta}_{it,1}^{\text{known}} &= \frac{1}{N} \sum_{j=1}^{N} \left(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot} v_t} \tilde{D}_t^{-1} - \frac{W_{jt}}{p_{jt}} D_t^{-1} \right) \left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right) \\
&= D_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \left(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot} v_t} - \frac{W_{jt}}{p_{jt}} \right) \left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right) \\
&\quad + \left(\tilde{D}_t^{-1} - D_t^{-1} \right) \cdot \frac{1}{N} \sum_{j=1}^{N} \left(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot} v_t} - \frac{W_{jt}}{p_{jt}} \right) \left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right) \\
&\quad + \left(\tilde{D}_t^{-1} - D_t^{-1} \right) \cdot \frac{1}{N} \sum_{j=1}^{N} \frac{W_{jt}}{p_{jt}} \left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right).
\end{aligned}$$

Since $\mathbb{E}[(u_i\bar{v} - \bar{W}_{i,\cdot})^8|I] \leq M/T^4$ for any i, it holds that

$$\mathbb{E}\left[\left(\tilde{D}_t - D_t\right)^8 \middle| I\right] \leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(\frac{W_{it}\bar{v}}{\bar{W}_{i,\cdot}v_t} - \frac{W_{it}}{p_{it}}\right)^8 \middle| I\right] \leq \frac{M}{T^4}.$$

The second term in the decomposition of $\tilde{\Delta}_{it,1} - \tilde{\Delta}_{it,1}^{\text{known}}$ can be bounded by

$$\begin{aligned}
&\mathbb{E}\left[\left(\left(\tilde{D}_t^{-1} - D_t^{-1}\right)\cdot\frac{1}{N}\sum_{j=1}^{N}\left(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot}v_t} - \frac{W_{jt}}{p_{jt}}\right)\left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt}\right)\right)^4\right] \\
&\leq \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\mathbb{E}\left[\left(\tilde{D}_t^{-1} - D_t^{-1}\right)^4\left(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot}v_t} - \frac{W_{jt}}{p_{jt}}\right)^4 \middle| I\right]\cdot\mathbb{E}\left[\left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt}\right)^4 \middle| I\right]\right] \\
&\leq \frac{M}{T^4}\cdot\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt}\right)^4\right] \\
&\leq \frac{M}{T^4}.
\end{aligned}$$

Similarly, the fourth moment of the third term in the decomposition of $\tilde{\Delta}_{it,1} - \tilde{\Delta}_{it,1}^{\text{known}}$ can be bounded by M/N^2T^2. As a result,

$$\tilde{\Delta}_{it,1} - \tilde{\Delta}_{it,1}^{\text{known}} = D_t^{-1}\cdot\frac{1}{N}\sum_{j=1}^{N}\left(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot}v_t} - \frac{W_{jt}}{p_{jt}}\right)\left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt}\right) + O_p(\frac{1}{\delta_{N,T}}).$$

Considering the difference between $\tilde{\Delta}_{it,3}$ and $\tilde{\Delta}_{it,3}^{\text{known}}$, we have

$$\tilde{\Delta}_{it,3} - \tilde{\Delta}_{it,3}^{\text{known}} = \bar{W}_{i,\cdot}^{-1}\cdot\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}W_{is}\left(\frac{W_{js}}{p_{js}}D_s^{-1} - \frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot}v_s}\tilde{D}_s^{-1}\right)\left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right).$$

Based on previous analysis, $\mathbb{E}\left[(\frac{W_{js}}{p_{js}}D_s^{-1} - \frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot}v_s}\tilde{D}_s^{-1})^4 \mid I\right] \leq M/T^2$. Furthermore, factors and errors are i.i.d. with mean zero and are independent of observation patterns. As a result,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}W_{is}\left(\frac{W_{js}}{p_{js}}D_s^{-1} - \frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot}v_s}\tilde{D}_s^{-1}\right)\left(\Lambda_j^\top F_s + \epsilon_{js}\right)\right)^4\right] \\
&\leq \frac{M}{NT^4}\sum_{j=1}^{N}\sum_{s,t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\left(\frac{W_{js}}{p_{js}}D_s^{-1} - \frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot}v_s}\tilde{D}_s^{-1}\right)^4 \mid I\right]\cdot\mathbb{E}\left[\left(\Lambda_j^\top F_s + \epsilon_{js}\right)^4 \mid I\right]\right] \\
&\leq \frac{M}{T^4}.
\end{aligned}$$

Moreover, we have the decomposition

$$
\begin{aligned}
&\bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \left(\frac{W_{js}}{p_{js}} D_s^{-1} - \frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot} v_s} \tilde{D}_s^{-1} \right) \alpha_j \\
&= \bar{W}_{i,\cdot}^{-1} \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \left(\frac{W_{js}}{p_{js}} - \frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot} v_s} \right) \left(\tilde{D}_s^{-1} - 1 \right) \alpha_j \\
&\quad + \bar{W}_{i,\cdot}^{-1} \frac{1}{T} \sum_{s=1}^{T} W_{is} \left(D_s^{-1} - \tilde{D}_s^{-1} \right) \cdot \frac{1}{N} \sum_{j=1}^{N} \frac{W_{js}}{p_{js}} \alpha_j + \bar{W}_{i,\cdot}^{-1} \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \left(\frac{W_{js}}{p_{js}} - \frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot} v_s} \right) \alpha_j \, .
\end{aligned}
$$

Similar to previous steps, we can bound the fourth moments of the first part and second part on the RHS by $M/\delta_{N,T}^4$. Therefore,

$$
\tilde{\Delta}_{it,3} - \tilde{\Delta}_{it,3}^{\text{known}} = \bar{W}_{i,\cdot}^{-1} \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \left(\frac{W_{js}}{p_{js}} - \frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot} v_s} \right) \alpha_j + O_p(\frac{1}{\delta_{N,T}}).
$$

Combining all the terms, we get that

$$
\begin{aligned}
\tilde{\Delta}_{it} - \tilde{\Delta}_{it}^{\text{known}} = &\bar{W}_{i,\cdot}^{-1} \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \left(\frac{W_{js}}{p_{js}} - \frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot} v_s} \right) \alpha_j \\
&+ D_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \left(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot} v_t} - \frac{W_{jt}}{p_{jt}} \right) \left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right) + O_p(\frac{1}{\delta_{N,T}}),
\end{aligned}
$$

and it is easy to show that $\mathbb{E}\left[(\tilde{\Delta}_{it} - \tilde{\Delta}_{it}^{\text{known}})^4 \right] \le M/\delta_{N,T}^2$. According to Lemma 8,

$$
\begin{aligned}
\tilde{\Delta}_{it} = &D_t^{-1} \frac{1}{N} \sum_{j=1}^{N} \left(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot} v_t} - 1 \right) \left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right) - \bar{W}_{i,\cdot}^{-1} \frac{1}{N} \sum_{j=1}^{N} \frac{1}{T} \sum_{s=1}^{T} W_{is} \left(\frac{W_{js}\bar{v}}{\bar{W}_{j,\cdot} v_s} - 1 \right) \alpha_j \\
&+ \bar{W}_{i,\cdot}^{-1} \cdot \left(\frac{1}{T} \sum_{s=1}^{T} W_{is}\epsilon_{is} + \Lambda_i^\top \frac{1}{T} \sum_{s=1}^{T} W_{is} F_s \right) + D_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \left(\Lambda_j^\top F_t + \epsilon_{jt} \right) + O_p(\frac{1}{\delta_{N,T}}).
\end{aligned}
$$

Observe that the only difference between $\tilde{\Delta}_{it}$ and $\tilde{\Delta}_{it}^{\text{known}}$ is that, the randomness of the first two terms of $\tilde{\Delta}_{it}$ comes from $W_{jt}\bar{v}/(\bar{W}_{j,\cdot} v_t) - 1$ instead of $W_{jt}/p_{jt} - 1$. Consider the conditional expectation of $W_{jt}\bar{v}/(\bar{W}_{j,\cdot} v_t) - 1$, we have

$$
\begin{aligned}
\mu_{jt} \coloneqq \mathbb{E}\left[\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot} v_t} - 1 \middle| I \right] &= \mathbb{E}\left[\frac{W_{jt}}{u_j v_t} + \frac{W_{jt}(u_j\bar{v} - \bar{W}_{j,\cdot})}{\bar{W}_{j,\cdot} u_j v_t} - 1 \middle| I \right] \\
&= \mathbb{E}\left[\frac{W_{jt}(u_j\bar{v} - \bar{W}_{j,\cdot})}{u_j^2 v_t \bar{v}} \middle| I \right] + \underbrace{\mathbb{E}\left[\frac{W_{jt}(u_j\bar{v} - \bar{W}_{j,\cdot})^2}{\bar{W}_{j,\cdot} u_j^2 v_t \bar{v}} \middle| I \right]}_{\le M/T} .
\end{aligned}
$$

By the short-term temporal dependency of the missingness, $\left|\frac{1}{T}\sum_{s=1}^{T}\mathbb{E}[W_{jt}W_{js}|I] - u_j^2 v_t \bar{v}\right| \leq M/T$ for any j and t, so the first part on the RHS can be bounded by

$$\left|\mathbb{E}\left[\frac{W_{jt}(u_j\bar{v} - \bar{W}_{j,\cdot})}{u_j^2 v_t \bar{v}}\middle| I\right]\right| = \left|1 - \frac{1}{u_j^2 v_t \bar{v}}\mathbb{E}\left[W_{jt}\bar{W}_{j,\cdot}|I\right]\right|$$
$$= \left|1 - \frac{1}{u_j^2 v_t \bar{v}} \cdot \frac{1}{T}\sum_{s=1}^{T}\mathbb{E}[W_{jt}W_{js}|I]\right| \leq \frac{M}{T}.$$

As a result, $|\mu_{jt}| = \left|\mathbb{E}[W_{jt}\bar{v}/(\bar{W}_{j,\cdot}v_t) - 1|I]\right| \leq M/T$ for any j and t.

We analyze the term $\frac{1}{N}\sum_{j=1}^{N}(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot}v_t} - 1)\alpha_j$. Conditional on I, $\left[(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot}v_t} - 1) - \mu_{jt}\right]\alpha_j$ and $\left[(\frac{W_{lt}\bar{v}}{\bar{W}_{l,\cdot}v_t} - 1) - \mu_{lt}\right]\alpha_l$ are independent for any $j \neq l$. The expectation $\mathbb{E}\left[((\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot}v_t} - 1) - \mu_{jt})\alpha_j \mid I\right] = 0$. By Assumption 6.4 and boundedness of μ_{jt},

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\left((\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot}v_t} - 1) - \mu_{jt}\right)^2 \alpha_j^2 \mid I\right]$$
$$= \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot}v_t})^2 \mid I\right]\alpha_j^2 - \frac{2}{N}\sum_{j=1}^{N}\mu_{jt}\alpha_j^2 - \frac{1}{N}\sum_{j=1}^{N}\mu_{jt}^2\alpha_j^2 - \frac{1}{N}\sum_{j=1}^{N}\alpha_j^2$$
$$\xrightarrow{p} \bar{s}_{\alpha,tt} - \sigma_\alpha^2.$$

Furthermore, we have

$$\sum_{j=1}^{N}\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\left((\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot}v_t} - 1) - \mu_{jt}\right)\alpha_j\right)^4 \mid I\right] \leq \frac{M}{N} \cdot \frac{1}{N}\sum_{j=1}^{N}\alpha_j^4.$$

By Theorem 1, we have $\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left[(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot}v_t} - 1) - \mu_{jt}\right]\alpha_j \xrightarrow{d} \mathcal{N}(0, \bar{s}_{\alpha,tt} - \sigma_\alpha^2)$. Furthermore, since $\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\mu_{jt}\alpha_j \xrightarrow{p} 0$, we get that $\frac{1}{\sqrt{N}}\sum_{j=1}^{N}(\frac{W_{jt}\bar{v}}{\bar{W}_{j,\cdot}v_t} - 1)\alpha_j \xrightarrow{d} \mathcal{N}(0, \bar{s}_{\alpha,tt} - \sigma_\alpha^2)$. By similar arguments, we can prove the asymptotic normality of other terms.

□

Lemma 10. *Suppose Assumptions 1, 5, 6 and Case 3 in Assumption 4 hold. When $N, T \to \infty$, the asymptotic distribution of $\tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i$ estimated by wi-PCA with weights (6) is:*

$$\sqrt{\delta_{N,T}} \cdot \sigma_{it}^{-1}\left((\tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i) - (\mu + \xi_t + \alpha_i)\right) \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\sigma_{it}^2 = \frac{\delta_{N,T}}{N}\sigma_{it,1}^2 + \frac{\delta_{N,T}}{T}\sigma_{it,2}^2$ with some $\sigma_{it,1}$ and $\sigma_{it,2}$. Moreover,

$$\mathbb{E}\left[\left((\tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i) - (\mu + \xi_t + \alpha_i)\right)^4\right] \leq \frac{M}{\delta_{N,T}^2}.$$

Proof. According to Lemma 3, $\tilde{\Delta}_{it}$ can be expanded by

$$\tilde{\Delta}_{it} = \underbrace{\frac{1}{N_c}\sum_{j\in\mathcal{N}_c}\left(\Lambda_j^\top F_t + \epsilon_{jt}\right)}_{\tilde{\Delta}_{it,1}} + \underbrace{\bar{W}_{i,\cdot}^{-1}\cdot\frac{1}{T}\sum_{s=1}^{T} W_{is}\left(\Lambda_i^\top F_s + \epsilon_{is}\right)}_{\tilde{\Delta}_{it,2}}$$
$$\underbrace{-\bar{W}_{i,\cdot}^{-1}\cdot\frac{1}{T}\sum_{s=1}^{T} W_{is}\frac{1}{N_c}\sum_{j\in\mathcal{N}_c}\left(\Lambda_j^\top F_s + \epsilon_{js}\right)}_{\tilde{\Delta}_{it,3}}.$$

We have $\frac{1}{\sqrt{N_c}}\sum_{j\in\mathcal{N}_c}\Lambda_j \xrightarrow{d} \mathcal{N}(0,\Sigma_{\Lambda,c})$. According to Lemma 8, $\tilde{\Delta}_{it,2}$ is asymptotically normal with zero mean and asymptotically independent with $\tilde{\Delta}_{it,1}$. For the last term $\tilde{\Delta}_{it,3}$, it is easy to show that $\frac{1}{T}\sum_{s=1}^{T} W_{is}\frac{1}{N_c}\sum_{j\in\mathcal{N}_c}(\Lambda_j^\top F_s + \epsilon_{js}) = O_p(\frac{1}{\delta_{N,T}})$ and $\mathbb{E}[\tilde{\Delta}_{it,3}^4] \le M/T^2$. We complete our proof. □

IA.C.2.2 Asymptotic normality for the common components

Lemma 11. *Let $\delta_{N,T} = \min(N,T)$. Suppose Assumptions 1, 5 hold and the first-stage estimation error satisfies $\mathbb{E}[\tilde{\Delta}_{it}^4] \le M/\delta_{N,T}^2$ for any i and t. Then, it holds that*

1. *For any i and j, $\mathbb{E}[\eta_{ij}^2] \le M/\delta_{N,T}$, $\mathbb{E}[\xi_{ij}^2] \le M/\delta_{N,T}$, and $\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\gamma_{ij}^2] \le M/\delta_{N,T}$.*
2. *When $N,T\to\infty$, $\frac{1}{N^2}\tilde{\Lambda}^\top\left((\dot{Y}\odot W)(\dot{Y}\odot W)^\top \odot \left[\frac{1}{|Q_{ij}|}\right]\right)\tilde{\Lambda} = \tilde{D} \xrightarrow{p} D$, where $D = diag(d_1,\cdots,d_k)$ is the diagonal matrix consisting of the eigenvalues of $\Sigma_\Lambda\Sigma_F$.*
3. *When $N,T\to\infty$, $\frac{1}{N}\tilde{\Lambda}^\top\Lambda \xrightarrow{p} Q = D^{1/2}\Upsilon\Sigma_F^{-1/2}$, where the diagonal entries of D are eigenvalues of $\Sigma_\Lambda\Sigma_F$, and Υ is the corresponding eigenvector matrix such that $\Upsilon^\top\Upsilon = I$. Moreover, $H \xrightarrow{p} (Q^\top)^{-1}$, where $H = \frac{1}{NT}\tilde{D}^{-1}\tilde{\Lambda}^\top\Lambda F^\top F$.*
4. *For any i, $\|\tilde{\Lambda}_i - H\Lambda_i\|^2 = O_p(\frac{1}{\delta_{N,T}})$. For any t, $\|\tilde{F}_t - (H^\top)^{-1}F_t\|^2 = O_p(\frac{1}{\delta_{N,T}})$.*

Proof. We can sequentially prove these statements following the proof of Lemmas 4, 5, 6 and Theorem 1. □

In the following, we prove the asymptotic normality of the common components estimated with wi-PCA with weights (4). We omit the proof for asymptotic normality of wi-PCA with weights (5) and (6), which is similar to the proof in the following.

Lemma 12. *Suppose Assumptions 1, 5 and Case 1 in Assumption 4 hold. We have*

1. *$\frac{1}{N}\sum_{i=1}^N \tilde{\Lambda}_i\eta_{ij} = O_p(\frac{1}{\sqrt{\delta_{N,T}}})$.*
2. *$\frac{1}{N}\sum_{i=1}^N \tilde{\Lambda}_i\xi_{ij} = O_p(\frac{1}{\delta_{N,T}})$.*
3. *$\frac{1}{N}\sum_{i=1}^N \tilde{\Lambda}_i\gamma_{ij} = O_p(\frac{1}{\delta_{N,T}})$.*

Proof. Observe that for any $\phi_{ij} = \eta_{ij}, \xi_{ij}$ and γ_{ij},

$$\frac{1}{N}\sum_{i=1}^{N}\tilde{\Lambda}_i\phi_{ij} = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{\Lambda}_i - H\Lambda_i\right)\phi_{ij} + H\cdot\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\phi_{ij}.$$

Based on Lemma 11, the first term on the RHS can be bounded by

$$\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{\Lambda}_i - H\Lambda_i\right)\phi_{ij}\right\|^2 \le \frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\Lambda}_i - H\Lambda_i\right\|^2\cdot\frac{1}{N}\sum_{i=1}^{N}\phi_{ij}^2 = O_p(\frac{1}{\delta_{N,T}^2}).$$

Because $H = O_p(1)$, we just need to bound the three terms $\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\eta_{ij}$, $\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\xi_{ij}$ and $\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\gamma_{ij}$ respectively in the following.

1. For $\frac{1}{N}\sum_{i=1}^{N}H\Lambda_i\eta_{ij}$, we have

$$\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\eta_{ij} = \frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_t\epsilon_{jt} - \frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_t\tilde{\Delta}_{jt}.$$

It is easy to see that the first part on the RHS

$$\left\|\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_t\epsilon_{jt}\right\|^2 \le \frac{1}{N}\sum_{i=1}^{N}\|\Lambda_i\|^4\cdot\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_t\epsilon_{jt}\right\|^2 = O_p(\frac{1}{T}).$$

According to Lemma 8,

$$\tilde{\Delta}_{it} = \underbrace{\bar{W}_{i,\cdot}^{-1}\left[\frac{1}{T}\sum_{s=1}^{T}W_{is}\left(\Lambda_i^\top F_s + \epsilon_{is}\right) - \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}W_{is}\frac{W_{js}}{p_{js}}D_s^{-1}\left(\alpha_j + \Lambda_j^\top F_s + \epsilon_{js}\right)\right]}_{\tilde{\Delta}_{i,1}}$$

$$+\underbrace{D_t^{-1}\cdot\frac{1}{N}\sum_{j=1}^{N}\frac{W_{jt}}{p_{jt}}\left(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt}\right)}_{\tilde{\Delta}_{t,2}}.$$

Plugging $\tilde{\Delta}_{it}$ into $\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\eta_{ij}$ gives us

$$\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\eta_{ij} = O_p(\frac{1}{\sqrt{T}}) - \underbrace{\tilde{\Delta}_{j,1}\cdot\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_t}_{O_p(\frac{1}{\delta_{N,T}})} - \frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_t\tilde{\Delta}_{t,2}$$

because $\tilde{\Delta}_{j,1} = O_p(\frac{1}{\sqrt{\delta_{N,T}}})$ and $\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_t = O_p(\frac{1}{\delta_{N,T}})$.

The last term $\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t\tilde{\Delta}_{t,2}$ can be further decomposed as

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t\tilde{\Delta}_{t,2} = \; & \underbrace{\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t\left(D_t^{-1}-1\right)\frac{1}{N}\sum_{l=1}^{N}\frac{W_{lt}}{p_{lt}}\Big(\alpha_l+\Lambda_l^\top F_t+\epsilon_{lt}\Big)}_{\omega_1} \\ & + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top \frac{1}{N}\sum_{l=1}^{N}\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t\left(\frac{W_{lt}}{p_{lt}}-1\right)\Big(\alpha_l+\Lambda_l^\top F_t+\epsilon_{lt}\Big)}_{\omega_2} \\ & + \underbrace{\frac{1}{N^2}\sum_{i,l=1}^{N}\Lambda_i\Lambda_i^\top \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t\Big(\alpha_l+\Lambda_l^\top F_t+\epsilon_{lt}\Big)}_{\omega_3}.\end{aligned}$$

Consider ω_1, we have

$$\|\omega_1\|^2 \le \frac{1}{N}\sum_{i=1}^{N}\|\Lambda_i\|^4\cdot\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\|F_t\|^2\left(D_t^{-1}-1\right)^2\left(\frac{1}{N}\sum_{l=1}^{N}\frac{W_{lt}}{p_{lt}}\Big(\alpha_l+\Lambda_l^\top F_t+\epsilon_{lt}\Big)\right)^2,$$

where

$$\begin{aligned}&\mathbb{E}\left[\|F_t\|^2\left(D_t^{-1}-1\right)^2\left(\frac{1}{N}\sum_{l=1}^{N}\frac{W_{lt}}{p_{lt}}\Big(\alpha_l+\Lambda_l^\top F_t+\epsilon_{lt}\Big)\right)^2\right] \\ \le\;& \left(\mathbb{E}\left[\|F_t\|^4\right]\cdot\mathbb{E}\left[\left(D_t^{-1}-1\right)^4\right]\cdot\mathbb{E}\left[\left(\frac{1}{N}\sum_{l=1}^{N}\frac{W_{lt}}{p_{lt}}\Big(\alpha_l+\Lambda_l^\top F_t+\epsilon_{lt}\Big)\right)^4\right]\right)^{1/2} \\ \le\;& \frac{M}{N^2}.\end{aligned}$$

Therefore, $\omega_1 = O_p(\frac{1}{N})$. For ω_2,

$$\begin{aligned}\|\omega_2\|^2 \le \frac{1}{N}\sum_{i=1}^{N}\|\Lambda_i\|^4\cdot\sum_{r=1}^{k}\frac{1}{N^3}\sum_{i,l,h=1}^{N}\frac{1}{|Q_{ij}|^2}\sum_{s,t\in Q_{ij}} F_{t,r}F_{s,r}\left(\frac{W_{lt}}{p_{lt}}-1\right)\left(\frac{W_{hs}}{p_{hs}}-1\right)\cdot \\ \Big(\alpha_l+\Lambda_l^\top F_t+\epsilon_{lt}\Big)\Big(\alpha_h+\Lambda_h^\top F_s+\epsilon_{hs}\Big) = O_p(\frac{1}{N}),\end{aligned}$$

where the last equality holds because $\mathbb{E}[F_{t,r}F_{s,r}(W_{lt}/p_{lt}-1)(W_{hs}/p_{hs}-1)(\alpha_l+\Lambda_l^\top F_t+\epsilon_{lt})(\alpha_h+\Lambda_h^\top F_s+\epsilon_{hs})] = 0$ when $l\neq h$. Furthermore, it is easy to show that

$$\omega_2 = \frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_tF_t^\top\frac{1}{N}\sum_{l=1}^{N}\left(\frac{W_{lt}}{p_{lt}}-1\right)\Lambda_l + O_p(\frac{1}{\delta_{N,T}}).$$

Similarly, we have

$$\omega_3 = \underbrace{\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_tF_t^\top \frac{1}{N}\sum_{l=1}^{N}\Lambda_l}_{O_p(\frac{1}{\sqrt{N}})} + O_p(\frac{1}{\delta_{N,T}}) = O_p(\frac{1}{\sqrt{\delta_{N,T}}}).$$

Combining these terms, we get that $\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\eta_{ij} = O_p(\frac{1}{\sqrt{\delta_{N,T}}})$.

2. For $\phi_{ij} = \xi_{ij}$, we have

$$\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\xi_{ij} = \left(\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t^\top\epsilon_{it} - \frac{1}{N}\sum_{i=1}^{N}\Lambda_i\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t^\top\tilde{\Delta}_{it}\right)\Lambda_j,$$

where it is easy to see that $\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t^\top\epsilon_{it} = O_p(\frac{1}{\delta_{N,T}})$.

Similar to the first part, we can decompose $\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t^\top\tilde{\Delta}_{it}$ as

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t^\top\tilde{\Delta}_{it}\\ =&\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\tilde{\Delta}_{i,1}\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t^\top + \frac{1}{N}\sum_{i=1}^{N}\Lambda_i\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t^\top\tilde{\Delta}_{t,2},\end{aligned}$$

where the first term on the RHS can be bounded by

$$\begin{aligned}\left\|\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\tilde{\Delta}_{i,1}\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t^\top\right\|^2 &\le \underbrace{\frac{1}{N}\sum_{i=1}^{N}\|\Lambda_i\|^2\tilde{\Delta}_{i,1}^2}_{O_p(\frac{1}{\delta_{N,T}})}\cdot\underbrace{\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t\right\|^2}_{O_p(\frac{1}{T})}\\ &= O_p(\frac{1}{T\delta_{N,T}}).\end{aligned}$$

Plugging the expression of $\tilde{\Delta}_{t,2}$ into the second term, we can further decompose it as

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t^\top\tilde{\Delta}_{t,2} = &\underbrace{\frac{1}{N^2}\sum_{i,l=1}^{N}\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}(D_t^{-1}-1)\frac{W_{lt}}{p_{lt}}\Lambda_iF_t^\top\left(\alpha_l + \Lambda_l^\top F_t + \epsilon_{lt}\right)}_{\omega_1}\\ &+\underbrace{\frac{1}{N^2}\sum_{i,l=1}^{N}\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\frac{W_{lt}}{p_{lt}}\Lambda_iF_t^\top\left(\alpha_l + \Lambda_l^\top F_t + \epsilon_{lt}\right)}_{\omega_2}.\end{aligned}$$

We bound ω_1 as

$$\begin{aligned}\|\omega_1\|^2 &\leq \frac{1}{N}\sum_{i=1}^N \|\Lambda_i\|^2 \cdot \frac{1}{N}\sum_{i=1}^N \left\| \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} (D_t^{-1}-1)F_t \frac{1}{N}\sum_{l=1}^N \frac{W_{lt}}{p_{lt}}\left(\alpha_l + \Lambda_l^\top F_t + \epsilon_{lt}\right)\right\|^2 \\ &\leq \underbrace{\frac{1}{N}\sum_{i=1}^N \|\Lambda_i\|^2}_{O_p(1)} \cdot \frac{1}{N}\sum_{i=1}^N \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} (D_t^{-1}-1)^2\|F_t\|^2 \left(\frac{1}{N}\sum_{l=1}^N \frac{W_{lt}}{p_{lt}}\left(\alpha_l + \Lambda_l^\top F_t + \epsilon_{lt}\right)\right)^2.\end{aligned}$$

According to part 1, $\mathbb{E}\left[(D_t^{-1}-1)^2\|F_t\|^2(\frac{1}{N}\sum_{l=1}^N \frac{W_{lt}}{p_{lt}}(\alpha_l + \Lambda_l^\top F_t + \epsilon_{lt}))^2\right] \leq M/N^2$. As a result, $\omega_1 = O_p(\frac{1}{N})$. For ω_2, we have

$$\begin{aligned}\|\omega_2\|^2 = &\sum_{r,o=1}^k \frac{1}{N^4}\sum_{i,l,h,q=1}^N \frac{1}{|Q_{ij}|}\frac{1}{|Q_{hj}|}\sum_{s\in Q_{ij}}\sum_{t\in Q_{hj}} \frac{W_{ls}}{p_{ls}}\frac{W_{qt}}{p_{qt}}\cdot \\ &\Lambda_{i,r}\Lambda_{h,r}F_{s,o}F_{t,o}(\alpha_l + \Lambda_l^\top F_s + \epsilon_{ls})(\alpha_q + \Lambda_q^\top F_t + \epsilon_{qt}) = O_p(\frac{1}{N^2}).\end{aligned}$$

As a result, $\omega_2 = O_p(\frac{1}{N})$ and $\frac{1}{N}\sum_{i=1}^N \tilde{\Lambda}_i\xi_{ij} = O_p(\frac{1}{\delta_{N,T}})$.

3. For $\phi_{ij} = \gamma_{ij}$,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \Lambda_i\gamma_{ij} = &\frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it}\epsilon_{jt} - \frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it}\tilde{\Delta}_{jt} \\ &- \frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \tilde{\Delta}_{it}\epsilon_{jt} + \frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \tilde{\Delta}_{it}\tilde{\Delta}_{jt}.\end{aligned}$$

It is easy to see that $\frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it}\epsilon_{jt} = O_p(\frac{1}{\delta_{N,T}})$. According to previous parts, $\tilde{\Delta}_{it} = \tilde{\Delta}_{i,1} + \tilde{\Delta}_{t,2}$, where $\mathbb{E}[\tilde{\Delta}_{i,1}^4] \leq M/\delta_{N,T}^2$ and $\mathbb{E}[\tilde{\Delta}_{t,2}^4] \leq M/\delta_{N,T}^2$. We have the decomposition

$$\frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it}\tilde{\Delta}_{jt} = \tilde{\Delta}_{j,1}\cdot \frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it} + \frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it}\tilde{\Delta}_{t,2}.$$

Since $\tilde{\Delta}_{j,1} = O_p(\frac{1}{\sqrt{\delta_{N,T}}})$ and

$$\left\|\frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it}\right\|^2 \leq \frac{1}{N}\sum_{i=1}^N \|\Lambda_i\|^2 \cdot \frac{1}{N}\sum_{i=1}^N \left(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it}\right)^2 = O_p(\frac{1}{T}),$$

we have $\tilde{\Delta}_{j,1} \cdot \frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it} = O_p(\frac{1}{\delta_{N,T}})$. Furthermore,

$$\frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it}\tilde{\Delta}_{t,2} = \underbrace{\frac{1}{N^2}\sum_{i,l=1}^N \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} (D_t^{-1}-1)\frac{W_{lt}}{p_{lt}}\Lambda_i\epsilon_{it}\left(\alpha_l + \Lambda_l^\top F_t + \epsilon_{lt}\right)}_{\omega_1}$$
$$+ \underbrace{\frac{1}{N^2}\sum_{i,l=1}^N \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \frac{W_{lt}}{p_{lt}}\Lambda_i\epsilon_{it}\left(\alpha_l + \Lambda_l^\top F_t + \epsilon_{lt}\right)}_{\omega_2}.$$

Similar to part 2, we have $\omega_1 = O_p(\frac{1}{N})$ and $\omega_2 = O_p(\frac{1}{N})$. Thus, $\frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{it}\tilde{\Delta}_{jt} = O_p(\frac{1}{\delta_{N,T}})$. We can also prove that $\frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \epsilon_{jt}\tilde{\Delta}_{it} = O_p(\frac{1}{\delta_{N,T}})$.

For $\frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \tilde{\Delta}_{it}\tilde{\Delta}_{jt}$, it holds that

$$\left\|\frac{1}{N}\sum_{i=1}^N \Lambda_i \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \tilde{\Delta}_{it}\tilde{\Delta}_{jt}\right\|^2 \leq \frac{1}{N}\sum_{i=1}^N \|\Lambda_i\|^2 \cdot \frac{1}{N}\sum_{i=1}^N \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \tilde{\Delta}_{it}^2\tilde{\Delta}_{jt}^2 = O_p(\frac{1}{\delta_{N,T}^2}).$$

Combining these terms, we get $\frac{1}{N}\sum_{i=1}^N \Lambda_i\gamma_{ij} = O_p(\frac{1}{\delta_{N,T}})$. We complete our proof. □

Lemma 13. *Suppose Assumptions 1, 5 and Case 1 in Assumption 4 hold. We have*

$$\tilde{\Lambda}_i - H\Lambda_i = \dot{D}^{-1}H\left(\omega_{\Lambda_i,1} - \omega_{\Lambda_i,2} + \omega_{\Lambda_i,3}\right) + O_p(\frac{1}{\delta_{N,T}}),$$

where

$$\omega_{\Lambda_i,1} = \frac{1}{N}\sum_{j=1}^N \Lambda_j\Lambda_j^\top \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t\epsilon_{it},$$
$$\omega_{\Lambda_i,2} = \frac{1}{N}\sum_{j=1}^N \Lambda_j\Lambda_j^\top \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_tF_t^\top \frac{1}{N}\sum_{l=1}^N \frac{W_{lt}}{p_{lt}}\Lambda_l,$$
$$\omega_{\Lambda_i,3} = \frac{1}{N}\sum_{j=1}^N \Lambda_j\Lambda_j^\top \Delta_{F,ij}\Lambda_i,$$

and $\Delta_{F,ij} = \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_tF_t^\top - \frac{1}{T}\sum_{t=1}^T F_tF_t^\top$.

Proof. For $i = 1, \cdots, N$, we have the decomposition

$$\tilde{\Lambda}_i - H\Lambda_i = \left(\tilde{\Lambda}_i - H_i\Lambda_i\right) + \left(H_i - H\right)\Lambda_i.$$

According to Lemma 12,

$$\begin{aligned}\tilde{\Lambda}_i - H_i\Lambda_i &= \dot{D}^{-1}\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\eta_{ij} + \frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\xi_{ij} + \frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\gamma_{ij}\right)\\ &= \dot{D}^{-1}H\Bigg[\underbrace{\frac{1}{N}\sum_{j=1}^{N}\Lambda_j\Lambda_j^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_t\epsilon_{it}}_{\omega_{\Lambda_i,1}} - \underbrace{\frac{1}{N}\sum_{j=1}^{N}\Lambda_j\Lambda_j^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_tF_t^\top\frac{1}{N}\sum_{l=1}^{N}\frac{W_{lt}}{p_{lt}}\Lambda_l}_{\omega_{\Lambda_i,2}}\Bigg]\\ &\quad + O_p(\frac{1}{\delta_{N,T}}).\end{aligned}$$

For $(H_i - H)\Lambda_i$, we have

$$\begin{aligned}H_i - H &= \dot{D}^{-1}\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\Lambda_j^\top\underbrace{\left(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_tF_t^\top - \frac{1}{T}\sum_{t=1}^{T}F_tF_t^\top\right)}_{\Delta_{F,ij}}\\ &= \dot{D}^{-1}\frac{1}{N}\sum_{j=1}^{N}\left(\tilde{\Lambda}_j - H\Lambda_j\right)\Lambda_j^\top\Delta_{F,ij} + \dot{D}^{-1}H\frac{1}{N}\sum_{j=1}^{N}\Lambda_j\Lambda_j^\top\Delta_{F,ij}\,.\end{aligned}$$

The first part $\frac{1}{N}\sum_{j=1}^{N}(\tilde{\Lambda}_j - H\Lambda_j)\Lambda_j^\top\Delta_{F,ij}$ can be bounded by

$$\left\|\frac{1}{N}\sum_{j=1}^{N}\left(\tilde{\Lambda}_j - H\Lambda_j\right)\Lambda_j^\top\Delta_{F,ij}\right\|^2 \le \underbrace{\frac{1}{N}\sum_{j=1}^{N}\left\|\tilde{\Lambda}_j - H\Lambda_j\right\|^2}_{O_p(\frac{1}{\delta_{N,T}})\text{ by Lemma 11}}\cdot\frac{1}{N}\sum_{j=1}^{N}\|\Lambda_j\|^2\|\Delta_{F,ij}\|^2 = O_p(\frac{1}{T\delta_{N,T}})\,,$$

where the last equality holds since $\mathbb{E}[\|\Delta_{F,ij}\|^2] \le M/T$. As a result,

$$(H_i - H)\Lambda_i = \dot{D}^{-1}H\underbrace{\frac{1}{N}\sum_{j=1}^{N}\Lambda_j\Lambda_j^\top\Delta_{F,ij}\Lambda_i}_{\omega_{\Lambda_i,3}} + O_p(\frac{1}{\delta_{N,T}}).$$

Combining the two terms, we have

$$\begin{aligned}\tilde{\Lambda}_i - H\Lambda_i &= \left(\tilde{\Lambda}_i - H_i\Lambda_i\right) + \left(H_i - H\right)\Lambda_i\\ &= \dot{D}^{-1}H\left(\omega_{\Lambda_i,1} - \omega_{\Lambda_i,2} + \omega_{\Lambda_i,3}\right) + O_p(\frac{1}{\delta_{N,T}}).\end{aligned}$$

□

Lemma 14. *Suppose Assumptions 1, 5 and Case 1 in Assumption 4 hold. We have*

1. $\frac{1}{N}\sum_{i=1}^N(\tilde{\Lambda}_i - H_i\Lambda_i)\Lambda_i^\top = O_p(\frac{1}{\delta_{N,T}})$.

2. $\frac{1}{N}\sum_{i=1}^N W_{it}(\tilde{\Lambda}_i - H_i\Lambda_i)\Lambda_i^\top = O_p(\frac{1}{\delta_{N,T}})$.

3. $\frac{1}{N}\sum_{i=1}^N W_{it}(\tilde{\Lambda}_i - H\Lambda_i)\dot{\epsilon}_{it} = O_p(\frac{1}{\delta_{N,T}})$.

Proof. 1. We have the decomposition

$$\frac{1}{N}\sum_{i=1}^N\left(\tilde{\Lambda}_i - H_i\Lambda_i\right)\Lambda_i^\top = \dot{D}^{-1}\left[\frac{1}{N^2}\sum_{i,j=1}^N\tilde{\Lambda}_j\Lambda_i^\top\eta_{ij} + \frac{1}{N^2}\sum_{i,j=1}^N\tilde{\Lambda}_j\Lambda_i^\top\xi_{ij} + \frac{1}{N^2}\sum_{i,j=1}^N\tilde{\Lambda}_j\Lambda_i^\top\gamma_{ij}\right].$$

Observe that for any $\phi_{ij} = \eta_{ij}, \xi_{ij}$ and γ_{ij},

$$\frac{1}{N^2}\sum_{i,j=1}^N\tilde{\Lambda}_j\Lambda_i^\top\phi_{ij} = \frac{1}{N^2}\sum_{i,j=1}^N\left(\tilde{\Lambda}_j - H\Lambda_j\right)\Lambda_i^\top\phi_{ij} + H\cdot\frac{1}{N^2}\sum_{i,j=1}^N\Lambda_j\Lambda_i^\top\phi_{ij}.$$

By Theorem 1 and Lemma 11, the first part on the RHS can be bounded by

$$\left\|\frac{1}{N^2}\sum_{i,j=1}^N\left(\tilde{\Lambda}_j - H\Lambda_j\right)\Lambda_i^\top\phi_{ij}\right\|^2 \leq \frac{1}{N}\sum_{j=1}^N\left\|\tilde{\Lambda}_j - H\Lambda_j\right\|^2\cdot\frac{1}{N}\sum_{i=1}^N\|\Lambda_i\|^2\cdot\frac{1}{N^2}\sum_{i,j=1}^N\phi_{ij}^2 = O_p(\frac{1}{\delta_{N,T}^2}).$$

We bound the three terms $\frac{1}{N^2}\sum_{i,j=1}^N\Lambda_j\Lambda_i^\top\eta_{ij}$, $\frac{1}{N^2}\sum_{i,j=1}^N\Lambda_j\Lambda_i^\top\xi_{ij}$ and $\frac{1}{N^2}\sum_{i,j=1}^N\Lambda_j\Lambda_i^\top\gamma_{ij}$ respectively in the following.

For $\frac{1}{N^2}\sum_{i,j=1}^N\Lambda_j\Lambda_i^\top\eta_{ij}$, it holds that

$$\frac{1}{N^2}\sum_{i,j=1}^N\Lambda_j\Lambda_i^\top\eta_{ij} = \underbrace{\frac{1}{N^2}\sum_{i,j=1}^N\Lambda_j\Lambda_i^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\Lambda_i^\top F_t\epsilon_{jt}}_{\omega_1} - \underbrace{\frac{1}{N^2}\sum_{i,j=1}^N\Lambda_j\Lambda_i^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\Lambda_i^\top F_t\tilde{\Delta}_{jt}}_{\omega_2}.$$

The first part ω_1 satisfies

$$\|\omega_1\|^2 = \sum_{p,q=1}^k\frac{1}{N^4}\sum_{i,j,h,l=1}^N\frac{1}{|Q_{ij}|}\frac{1}{|Q_{hl}|}\sum_{t\in Q_{ij}}\sum_{s\in Q_{hl}}\Lambda_{j,p}\Lambda_{i,q}\Lambda_{l,p}\Lambda_{h,q}\Lambda_i^\top F_t\Lambda_h^\top F_s\epsilon_{jt}\epsilon_{ls} = O_p(\frac{1}{NT}),$$

where the last equality holds since $\mathbb{E}[\Lambda_{j,p}\Lambda_{i,q}\Lambda_{l,p}\Lambda_{h,q}\Lambda_i^\top F_t\Lambda_h^\top F_s\epsilon_{jt}\epsilon_{ls}] = 0$ when $t \neq s$ or $j \neq l$. According to Lemma 12, $\tilde{\Delta}_{jt} = \tilde{\Delta}_{j,1} + \tilde{\Delta}_{t,2}$, so ω_2 can be further decomposed as

$$\omega_2 = \frac{1}{N^2}\sum_{i,j=1}^N\tilde{\Delta}_{j,1}\Lambda_j\Lambda_i^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\Lambda_i^\top F_t + \frac{1}{N^2}\sum_{i,j=1}^N\Lambda_j\Lambda_i^\top\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}\Lambda_i^\top F_t\tilde{\Delta}_{t,2}.$$

By Lemma 8,

$$\begin{aligned}
&\left\| \frac{1}{N^2} \sum_{i,j=1}^{N} \tilde{\Delta}_{j,1} \Lambda_j \Lambda_i^\top \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} \Lambda_i^\top F_t \right\|^2 \\
\leq & \underbrace{\frac{1}{N} \sum_{j=1}^{N} \tilde{\Delta}_{j,1}^2 \|\Lambda_j\|^2}_{O_p(\frac{1}{\delta_{N,T}})} \cdot \underbrace{\frac{1}{N} \sum_{i=1}^{N} \|\Lambda_i\|^4}_{O_p(1)} \cdot \underbrace{\frac{1}{N^2} \sum_{i,j=1}^{N} \left\| \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} F_t \right\|^2}_{O_p(\frac{1}{T})} \\
= & O_p(\frac{1}{\delta_{N,T}^2}).
\end{aligned}$$

We plug $\tilde{\Delta}_{t,2}$ into the second part of ω_2 and get

$$\begin{aligned}
\omega_2 = & \underbrace{\frac{1}{N^2} \sum_{i,j=1}^{N} \Lambda_j \Lambda_i^\top \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} \Lambda_i^\top F_t \left(D_t^{-1} - 1\right) \frac{1}{N} \sum_{l=1}^{N} \frac{W_{lt}}{p_{lt}} \left(\alpha_l + \Lambda_l^\top F_t + \epsilon_{lt}\right)}_{\omega_{2,1}} \\
& + \underbrace{\frac{1}{N^2} \sum_{i,j=1}^{N} \Lambda_j \Lambda_i^\top \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} \Lambda_i^\top F_t \frac{1}{N} \sum_{l=1}^{N} \frac{W_{lt}}{p_{lt}} \left(\alpha_l + \Lambda_l^\top F_t + \epsilon_{lt}\right)}_{\omega_{2,2}} + O_p(\frac{1}{\delta_{N,T}}).
\end{aligned}$$

The first part $\omega_{2,1}$ satisfies

$$\begin{aligned}
\|\omega_{2,1}\|^2 \leq & \frac{1}{N} \sum_{i=1}^{N} \|\Lambda_i\|^4 \cdot \frac{1}{N} \sum_{j=1}^{N} \|\Lambda_j\|^2 \cdot \\
& \frac{1}{N^2} \sum_{i,j=1}^{N} \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} \left\| F_t \left(D_t^{-1} - 1\right) \frac{1}{N} \sum_{l=1}^{N} \frac{W_{lt}}{p_{lt}} \left(\alpha_l + \Lambda_l^\top F_t + \epsilon_{lt}\right) \right\|^2 = O_p(\frac{1}{N^2}).
\end{aligned}$$

We can also show that $\omega_{2,2} = O_p(\frac{1}{\delta_{N,T}})$. Combining these terms, we get $\frac{1}{N^2} \sum_{i,j=1}^{N} \Lambda_j \Lambda_i^\top \eta_{ij} = O_p(\frac{1}{\delta_{N,T}})$. By symmetry, we have $\frac{1}{N^2} \sum_{i,j=1}^{N} \Lambda_j \Lambda_i^\top \xi_{ij} = O_p(\frac{1}{\delta_{N,T}})$.

For $\frac{1}{N^2} \sum_{i,j=1}^{N} \Lambda_j \Lambda_i^\top \gamma_{ij}$, we have

$$\begin{aligned}
\frac{1}{N^2} \sum_{i,j=1}^{N} \Lambda_j \Lambda_i^\top \gamma_{ij} = & \underbrace{\frac{1}{N^2} \sum_{i,j=1}^{N} \Lambda_j \Lambda_i^\top \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} \epsilon_{it} \epsilon_{jt}}_{\omega_1} - \underbrace{\frac{2}{N^2} \sum_{i,j=1}^{N} \Lambda_j \Lambda_i^\top \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} \epsilon_{it} \tilde{\Delta}_{jt}}_{\omega_2} \\
& + \underbrace{\frac{1}{N^2} \sum_{i,j=1}^{N} \Lambda_j \Lambda_i^\top \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} \tilde{\Delta}_{it} \tilde{\Delta}_{jt}}_{\omega_3}.
\end{aligned}$$

Since idiosyncratic errors are i.i.d. with zero mean and are independent with loadings,

$$\mathbb{E}\left[\|\omega_1\|^2\right] = \sum_{p,q=1}^{k} \frac{1}{N^4} \sum_{i,j,h,l=1}^{N} \frac{1}{|Q_{ij}|} \frac{1}{|Q_{hl}|} \sum_{t \in Q_{ij}} \sum_{s \in Q_{hl}} \mathbb{E}\left[\Lambda_{j,p}\Lambda_{i,q}\Lambda_{h,p}\Lambda_{l,q}\right] \cdot \mathbb{E}\left[\epsilon_{it}\epsilon_{jt}\epsilon_{hs}\epsilon_{ls}\right] \leq \frac{M}{N^2}.$$

Similar to $\frac{1}{N^2}\sum_{i,j=1}^{N} \Lambda_j \Lambda_i^\top \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \Lambda_i^\top F_t \tilde{\Delta}_{jt} = O_p(\frac{1}{\delta_{N,T}})$, we have $\omega_2 = O_p(\frac{1}{\delta_{N,T}})$. Since $\mathbb{E}[\tilde{\Delta}_{it}^4] \leq M/\delta_{N,T}^2$,

$$\|\omega_3\|^2 \leq \frac{1}{N}\sum_{i=1}^{N} \|\Lambda_i\|^2 \cdot \frac{1}{N}\sum_{j=1}^{N}\|\Lambda_j\|^2 \cdot \frac{1}{N^2}\sum_{i,j=1}^{N} \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} \tilde{\Delta}_{it}^2 \tilde{\Delta}_{jt}^2 = O_p(\frac{1}{\delta_{N,T}^2}).$$

Therefore, we have $\frac{1}{N^2}\sum_{i,j=1}^{N} \Lambda_j \Lambda_i^\top \gamma_{ij} = O_p(\frac{1}{\delta_{N,T}})$.

2. Similar to the first part, we have the decomposition

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} W_{it}\left(\tilde{\Lambda}_i - H_i\Lambda_i\right)\Lambda_i^\top \\
&= \dot{D}^{-1}\left[\frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}\tilde{\Lambda}_j\Lambda_i^\top \eta_{ij} + \frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}\tilde{\Lambda}_j\Lambda_i^\top \xi_{ij} + \frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}\tilde{\Lambda}_j\Lambda_i^\top \gamma_{ij}\right].
\end{aligned}$$

For any $\phi_{ij} = \eta_{ij}, \xi_{ij}$ and γ_{ij},

$$\frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}\tilde{\Lambda}_j\Lambda_i^\top \phi_{ij} = \frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}\left(\tilde{\Lambda}_j - H\Lambda_j\right)\Lambda_i^\top \phi_{ij} + H \cdot \frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}\Lambda_j\Lambda_i^\top \phi_{ij}.$$

Since $\frac{1}{N}\sum_{i=1}^{N} W_{it}\|\Lambda_i\|^2 \leq \frac{1}{N}\sum_{i=1}^{N}\|\Lambda_i\|^2$, the first part on the RHS can be bounded by

$$\begin{aligned}
\left\|\frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}\left(\tilde{\Lambda}_j - H\Lambda_j\right)\Lambda_i^\top \phi_{ij}\right\|^2 &\leq \frac{1}{N}\sum_{j=1}^{N}\left\|\tilde{\Lambda}_j - H\Lambda_j\right\|^2 \cdot \frac{1}{N}\sum_{i=1}^{N}\|\Lambda_i\|^2 \cdot \frac{1}{N^2}\sum_{i,j=1}^{N}\phi_{ij}^2 \\
&= O_p(\frac{1}{\delta_{N,T}^2}).
\end{aligned}$$

By similar arguments with the first part, we can show that $\frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}\Lambda_j\Lambda_i^\top \phi_{ij} = O_p(\frac{1}{\delta_{N,T}})$ for any $\phi_{ij} = \eta_{ij}, \xi_{ij}$ and γ_{ij}.

3. $\frac{1}{N}\sum_{i=1}^{N} W_{it}(\tilde{\Lambda}_i - H\Lambda_i)\dot{\epsilon}_{it}$ has the decomposition

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} W_{it}\left(\tilde{\Lambda}_i - H\Lambda_i\right)\dot{\epsilon}_{it} = & \underbrace{\frac{1}{N}\sum_{i=1}^{N} W_{it}\left(\tilde{\Lambda}_i - H_i\Lambda_i\right)\epsilon_{it}}_{\omega_1 = O_p(\frac{1}{\delta_{N,T}})} + \underbrace{\frac{1}{N}\sum_{i=1}^{N} W_{it}\left(H_i - H\right)\Lambda_i\epsilon_{it}}_{\omega_2} \\ & - \underbrace{\frac{1}{N}\sum_{i=1}^{N} W_{it}\left(\tilde{\Lambda}_i - H\Lambda_i\right)\tilde{\Delta}_{it}}_{\omega_3}.\end{aligned}$$

By definition, $H_i - H = \dot{D}^{-1}\frac{1}{N}\sum_{j=1}^{N}\tilde{\Lambda}_j\Lambda_j^\top\Delta_{F,ij}$, where $\Delta_{F,ij} = \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_tF_t^\top - \frac{1}{T}\sum_{t=1}^{T} F_tF_t^\top$. So ω_2 can be bounded by

$$\|\omega_2\|^2 \le \left\|\dot{D}^{-1}\right\|^2 \cdot \frac{1}{N}\sum_{j=1}^{N}\|\tilde{\Lambda}_j\|^2\|\Lambda_j\|^2 \cdot \frac{1}{N}\sum_{j=1}^{N}\left\|\frac{1}{N}\sum_{i=1}^{N} W_{it}\Delta_{F,ij}\Lambda_i\epsilon_{it}\right\|^2 = O_p(\frac{1}{NT})$$

since $\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N} W_{it}\Delta_{F,ij}\Lambda_i\epsilon_{it}\right\|^2 \le \mathbb{E}\left[\frac{1}{N^2}\sum_{i=1}^{N}\|\Delta_{F,ij}\|^2\|\Lambda_i\|^2\epsilon_{it}^2\right] \le M/NT$. Moreover,

$$\|\omega_3\|^2 \le \frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\Lambda}_i - H\Lambda_i\right\|^2 \cdot \frac{1}{N}\sum_{i=1}^{N}\tilde{\Delta}_{it}^2 = O_p(\frac{1}{\delta_{N,T}^2}).$$

We complete our proof. □

Lemma 15. *Suppose Assumptions 1, 5 and Case 1 in Assumption 4 hold, we have*

$$\tilde{F}_t - (H^\top)^{-1}F_t = \omega_{F_t,1} - \omega_{F_t,2} - \omega_{F_t,3} + O_p(\frac{1}{\delta_{N,T}}),$$

where

$$\begin{aligned}\omega_{F_t,1} &= (H^\top)^{-1}\hat{\Sigma}_{\Lambda,t}^{-1}\cdot\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\epsilon_{it},\\ \omega_{F_t,2} &= (H^\top)^{-1}\hat{\Sigma}_{\Lambda,t}^{-1}\cdot\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\tilde{\Delta}_{it},\\ \omega_{F_t,3} &= (H^\top)^{-1}\hat{\Sigma}_{\Lambda,t}^{-1}\boldsymbol{Z}_tH^\top(\dot{D}^{-1})^\top(H^\top)^{-1}F_t,\end{aligned}$$

$\hat{\Sigma}_{\Lambda,t} = \frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\Lambda_i^\top$, $\tilde{\Delta}_{it} = (\tilde{\mu}+\tilde{\alpha}_i+\tilde{\xi}_t)-(\mu+\alpha_i+\xi_t)$, *and* $\boldsymbol{Z}_t = \frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}\Lambda_j\Lambda_j^\top\Delta_{F,ij}\Lambda_i\Lambda_i^\top$.

Proof. We estimate factors by regressing the observed $\dot{Y}_{it}$ on $\tilde{\Lambda}_i$ as

$$\tilde{F}_t = \left(\frac{1}{N} \sum_{i=1}^{N} W_{it} \tilde{\Lambda}_i \tilde{\Lambda}_i^\top \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} W_{it} \tilde{\Lambda}_i \dot{Y}_{it} \right), \quad t = 1, \cdots, T.$$

As in the proof of Theorem 1, we define the auxiliary $\tilde{F}_t^*$ as

$$\tilde{F}_t^* \coloneqq \left(\frac{1}{N} \sum_{i=1}^{N} W_{it} H \Lambda_i \Lambda_i^\top H^\top \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} W_{it} \tilde{\Lambda}_i \dot{Y}_{it} \right).$$

We have the decomposition $\tilde{F}_t - (H^\top)^{-1} F_t = (\tilde{F}_t - \tilde{F}_t^*) + (\tilde{F}_t^* - (H^\top)^{-1} F_t)$.

We first analyze $\tilde{F}_t^*$, which has the decomposition

$$\begin{aligned} \tilde{F}_t^* = {} & (H^\top)^{-1} F_t + (H^\top)^{-1} \hat{\Sigma}_{\Lambda,t}^{-1} \cdot \left[\frac{1}{N} \sum_{i=1}^{N} W_{it} \Lambda_i \epsilon_{it} - \frac{1}{N} \sum_{i=1}^{N} W_{it} \Lambda_i \tilde{\Delta}_{it} \right] \\ & + (H^\top)^{-1} \hat{\Sigma}_{\Lambda,t}^{-1} H^{-1} \cdot \Bigg[\underbrace{\frac{1}{N} \sum_{i=1}^{N} W_{it} \left(\tilde{\Lambda}_i - H \Lambda_i \right) \dot{\epsilon}_{it}}_{O_p(\frac{1}{\delta_{N,T}}) \text{ by Lemma 14}} + \underbrace{\frac{1}{N} \sum_{i=1}^{N} W_{it} \left(\tilde{\Lambda}_i - H_i \Lambda_i \right) \Lambda_i^\top F_t}_{O_p(\frac{1}{\delta_{N,T}}) \text{ by Lemma 14}} \\ & + \frac{1}{N} \sum_{i=1}^{N} W_{it} \left(H_i - H \right) \Lambda_i \Lambda_i^\top F_t \Bigg], \end{aligned}$$

where $\hat{\Sigma}_{\Lambda,t} \coloneqq \frac{1}{N} \sum_{i=1}^{N} W_{it} \Lambda_i \Lambda_i^\top$. By Assumption 5, $\hat{\Sigma}_{\Lambda,t} \xrightarrow{p} \Sigma_{\Lambda,t}$. We further expand the last term $\frac{1}{N} \sum_{i=1}^{N} W_{it} \left(H_i - H \right) \Lambda_i \Lambda_i^\top F_t$ as

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} W_{it} \left(H_i - H \right) \Lambda_i \Lambda_i^\top F_t = {} & \dot{D}^{-1} \cdot \frac{1}{N^2} \sum_{i,j=1}^{N} W_{it} \left(\tilde{\Lambda}_j - H \Lambda_j \right) \Lambda_j^\top \Delta_{F,ij} \Lambda_i \Lambda_i^\top F_t \\ & + \dot{D}^{-1} H \cdot \frac{1}{N^2} \sum_{i,j=1}^{N} W_{it} \Lambda_j \Lambda_j^\top \Delta_{F,ij} \Lambda_i \Lambda_i^\top F_t, \end{aligned}$$

where $\Delta_{F,ij} = \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} F_t F_t^\top - \frac{1}{T} \sum_{t=1}^{T} F_t F_t^\top$. For $i \neq j$, $\mathbb{E}\left[\|\Lambda_j\|^2 \|\Lambda_i\|^4 \|\Delta_{F,ij}\|^2 \right] = \mathbb{E}[\|\Lambda_j\|^2] \cdot \mathbb{E}[\|\Lambda_i\|^4] \cdot \mathbb{E}[\|\Delta_{F,ij}\|^2] \leq M/T$, so the first part on the RHS can be bounded by

$$\begin{aligned} & \left\| \frac{1}{N^2} \sum_{i,j=1}^{N} W_{it} \left(\tilde{\Lambda}_j - H \Lambda_j \right) \Lambda_j^\top \Delta_{F,ij} \Lambda_i \Lambda_i^\top \right\|^2 \\ \leq {} & \frac{1}{N} \sum_{j=1}^{N} \left\| \tilde{\Lambda}_j - H \Lambda_j \right\|^2 \cdot \frac{1}{N^2} \sum_{i,j=1}^{N} \left(\|\Lambda_j\|^2 \|\Lambda_i\|^4 \|\Delta_{F,ij}\|^2 \right) = O_p(\frac{1}{\delta_{N,T}^2}). \end{aligned}$$

As a result,

$$\tilde{F}_t^* - (H^\top)^{-1}F_t = (H^\top)^{-1}\hat{\Sigma}_{\Lambda,t}^{-1} \cdot \frac{1}{N}\sum_{i=1}^N W_{it}\Lambda_i\epsilon_{it} - (H^\top)^{-1}\hat{\Sigma}_{\Lambda,t}^{-1} \cdot \frac{1}{N}\sum_{i=1}^N W_{it}\Lambda_i\tilde{\Delta}_{it}$$
$$+ (H^\top)^{-1}\hat{\Sigma}_{\Lambda,t}^{-1}H^{-1}\dot{D}^{-1}H \cdot \frac{1}{N^2}\sum_{i,j=1}^N W_{it}\Lambda_j\Lambda_j^\top\Delta_{F,ij}\Lambda_i\Lambda_i^\top F_t + O_p(\frac{1}{\delta_{N,T}}).$$

We let $\mathbf{Z}_t = \frac{1}{N^2}\sum_{i,j=1}^N W_{it}\Lambda_j\Lambda_j^\top\Delta_{F,ij}\Lambda_i\Lambda_i^\top$. By Proposition 3 in Xiong and Pelger (2023), $\sqrt{T}\cdot\mathbf{Z}_t$ is asymptotically normal with zero mean.

Next, we consider the difference between $\tilde{F}_t^*$ and $\tilde{F}_t$. We have

$$\tilde{F}_t^* - \tilde{F}_t = \left(\frac{1}{N}\sum_{i=1}^N W_{it}\tilde{\Lambda}_i\tilde{\Lambda}_i^\top\right)^{-1} \underbrace{\left[\frac{1}{N}\sum_{i=1}^N W_{it}\tilde{\Lambda}_i\tilde{\Lambda}_i^\top - \frac{1}{N}\sum_{i=1}^N W_{it}H\Lambda_i\Lambda_i^\top H^\top\right]}_{\tilde{\Delta}_{\Lambda,t}} \tilde{F}_t^*.$$

The difference $\tilde{\Delta}_{\Lambda,t}$ can be expanded as

$$\begin{aligned}\tilde{\Delta}_{\Lambda,t} &= \frac{1}{N}\sum_{i=1}^N \Big[W_{it}H\Lambda_i\Lambda_i^\top(H_i - H)^\top + W_{it}(H_i - H)\Lambda_i\Lambda_i^\top H^\top\Big] \\ &\quad + \underbrace{\frac{1}{N}\sum_{i=1}^N \Big[W_{it}H\Lambda_i(\tilde{\Lambda}_i - H_i\Lambda_i)^\top + W_{it}(\tilde{\Lambda}_i - H_i\Lambda_i)\Lambda_i^\top H^\top\Big]}_{O_p(\frac{1}{\delta_{N,T}}) \text{ by Lemma 14}} + \underbrace{\frac{1}{N}\sum_{i=1}^N W_{it}(\tilde{\Lambda}_i - H_i\Lambda_i)(\tilde{\Lambda}_i - H_i\Lambda_i)^\top}_{O_p(\frac{1}{\delta_{N,T}}) \text{ by Theorem 1}} \\ &= \frac{1}{N}\sum_{i=1}^N \Big[W_{it}H\Lambda_i\Lambda_i^\top(H_i - H)^\top + W_{it}(H_i - H)\Lambda_i\Lambda_i^\top H^\top\Big] + O_p(\frac{1}{\delta_{N,T}}).\end{aligned}$$

Plugging the expression of $H_i - H$ into the above equation, we get

$$\tilde{\Delta}_{\Lambda,t} = \dot{D}^{-1}H\mathbf{Z}_tH^\top + H\mathbf{Z}_tH^\top(\dot{D}^{-1})^\top + O_p(\frac{1}{\delta_{N,T}}).$$

$\mathbf{Z}_t = O_p(\frac{1}{\sqrt{T}})$ implies $\tilde{\Delta}_{\Lambda,t} \xrightarrow{p} 0$. Therefore,

$$\frac{1}{N}\sum_{i=1}^N W_{it}\tilde{\Lambda}_i\tilde{\Lambda}_i^\top = \frac{1}{N}\sum_{i=1}^N W_{it}H\Lambda_i\Lambda_i^\top H^\top + \tilde{\Delta}_{\Lambda,t} \xrightarrow{p} H\Sigma_{\Lambda,t}H^\top.$$

According to the first step, $\tilde{F}_t^* = (H^\top)^{-1}F_t + O_p(\frac{1}{\sqrt{\delta}})$. We infer that

$$\tilde{F}_t^* - \tilde{F}_t = (H^\top)^{-1}\hat{\Sigma}_{\Lambda,t}^{-1}H^{-1}\left(\dot{D}^{-1}H\mathbf{Z}_tH^\top + H\mathbf{Z}_tH^\top(\dot{D}^{-1})^\top\right)(H^\top)^{-1}F_t + O_p(\frac{1}{\delta_{N,T}}).$$

Combining these two steps, we finish our proof. □

Lemma 16. *Suppose Assumptions 1, 5 and Case 1 in Assumption 4 hold, we have*

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\tilde{\Delta}_{it} =& \frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i \frac{1}{N}\sum_{j=1}^{N}\left(\frac{W_{jt}}{p_{jt}}-1\right)\left(\alpha_j+\Lambda_j^\top F_t+\epsilon_{jt}\right)\\ &+\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i \frac{1}{N}\sum_{j=1}^{N}\left(\Lambda_j^\top F_t+\epsilon_{jt}\right)+\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\bar{W}_{i,\cdot}^{-1}\Lambda_i^\top \frac{1}{T}\sum_{s=1}^{T} W_{is}F_s\\ &-\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\bar{W}_{i,\cdot}^{-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{T}\sum_{s=1}^{T} W_{is}\left(\frac{W_{jt}}{p_{js}}-1\right)\alpha_j+O_p(\frac{1}{\delta_{N,T}}).\end{aligned}$$

Proof. We denote $X_{it}=W_{it}/p_{it}$. According to Lemma 8, $\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\tilde{\Delta}_{it}$ can be decomposed as

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\tilde{\Delta}_{it} =& \underbrace{D_t^{-1}\cdot\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\frac{1}{N}\sum_{j=1}^{N} X_{jt}\left(\alpha_j+\Lambda_j^\top F_t+\epsilon_{jt}\right)}_{\omega_1}\\ &+\underbrace{\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\bar{W}_{i,\cdot}^{-1}\frac{1}{T}\sum_{s=1}^{T} W_{is}\left(\Lambda_i^\top F_s+\epsilon_{is}\right)}_{\omega_2}\\ &-\underbrace{\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\bar{W}_{i,\cdot}^{-1}\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} W_{is}X_{js}D_s^{-1}\left(\alpha_j+\Lambda_j^\top F_s+\epsilon_{js}\right)}_{\omega_3}.\end{aligned}$$

We can further decompose ω_1 as

$$\begin{aligned}\omega_1 =& \left(D_t^{-1}-1\right)\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\frac{1}{N}\sum_{j=1}^{N} X_{jt}\left(\alpha_j+\Lambda_j^\top F_t+\epsilon_{jt}\right)\\ &+\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\frac{1}{N}\sum_{j=1}^{N} X_{jt}\left(\alpha_j+\Lambda_j^\top F_t+\epsilon_{jt}\right).\end{aligned}$$

We have shown that $\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\frac{1}{N}\sum_{j=1}^{N} X_{jt}\left(\alpha_j+\Lambda_j^\top F_t+\epsilon_{jt}\right)=O_p(\frac{1}{\sqrt{N}})$ and $\mathbb{E}[(D_t^{-1}-1)^2]\leq M/N$. Therefore, we have $\omega_1=\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\frac{1}{N}\sum_{j=1}^{N} X_{jt}(\alpha_j+\Lambda_j^\top F_t+\epsilon_{jt})+O_p(\frac{1}{N})$.

For ω_2, since ϵ_{it} are i.i.d. and independent of other terms,

$$\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i\bar{W}_{i,\cdot}^{-1}\frac{1}{T}\sum_{s=1}^{T} W_{is}\epsilon_{is}\right)^2\right]\leq\frac{M}{NT}.$$

Observe that ω_3 can be further decomposed as

$$\begin{aligned}\omega_3 = & \underbrace{\frac{1}{N^2}\sum_{i,j=1}^{N}\frac{1}{T}\sum_{s=1}^{T} W_{it}\Lambda_i\bar{W}_{i,\cdot}^{-1}W_{is}X_{js}\left(D_s^{-1}-1\right)\Big(\alpha_j+\Lambda_j^\top F_s+\epsilon_{js}\Big)}_{\omega_{3,1}} \\ & + \underbrace{\frac{1}{N^2}\sum_{i,j=1}^{N}\frac{1}{T}\sum_{s=1}^{T} W_{it}\Lambda_i\bar{W}_{i,\cdot}^{-1}W_{is}(X_{js}-1)\Big(\alpha_j+\Lambda_j^\top F_s+\epsilon_{js}\Big)}_{\omega_{3,2}} \\ & + \underbrace{\frac{1}{N^2}\sum_{i,j=1}^{N}\frac{1}{T}\sum_{s=1}^{T} W_{it}\Lambda_i\bar{W}_{i,\cdot}^{-1}W_{is}\Big(\alpha_j+\Lambda_j^\top F_s+\epsilon_{js}\Big)}_{\omega_{3,3}}.\end{aligned}$$

The first part $\omega_{3,1}$ can be bounded by

$$\|\omega_{3,1}\|^2 \le \underbrace{\frac{1}{T}\sum_{s=1}^{T}\left(D_s^{-1}-1\right)^2}_{O_p(\frac{1}{N})}\cdot\frac{1}{T}\sum_{s=1}^{T}\left\|\frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}W_{is}X_{js}\bar{W}_{i,\cdot}^{-1}\Lambda_i\Big(\alpha_j+\Lambda_j^\top F_s+\epsilon_{js}\Big)\right\|^2.$$

Since $\mathbb{E}\Big\|\frac{1}{N^2}\sum_{i,j=1}^{N} W_{it}W_{is}X_{js}\bar{W}_{i,\cdot}^{-1}\Lambda_i(\alpha_j+\Lambda_j^\top F_s+\epsilon_{js})\Big\|^2 \le M/N$, we have $\omega_{3,1}=O_p(\frac{1}{N})$. For $\omega_{3,2}$, it holds that

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{1}{N^2}\sum_{i,j=1}^{N}\frac{1}{T}\sum_{s=1}^{T} W_{it}\Lambda_i\bar{W}_{i,\cdot}^{-1}W_{is}(X_{js}-1)\Big(\Lambda_j^\top F_s+\epsilon_{js}\Big)\right\|^2\right] \\ = & \sum_{r=1}^{k}\frac{1}{N^4}\sum_{i,j,h,l=1}^{N}\frac{1}{T^2}\sum_{s,u=1}^{T}\mathbb{E}\Big[\mathbb{E}[W_{it}W_{ht}\bar{W}_{i,\cdot}^{-1}\bar{W}_{h,\cdot}^{-1}W_{is}W_{hu}(X_{js}-1)(X_{lu}-1)|I]\cdot \\ & \mathbb{E}[\Lambda_{i,r}\Lambda_{h,r}(\Lambda_j^\top F_s+\epsilon_{js})(\Lambda_l^\top F_u+\epsilon_{lu})|I]\Big] \le \frac{M}{NT}.\end{aligned}$$

This is because when $s\neq u$, $\mathbb{E}[\Lambda_{i,r}\Lambda_{h,r}(\Lambda_j^\top F_s+\epsilon_{js})(\Lambda_l^\top F_u+\epsilon_{lu})|I]=0$. Additionally, when i,j,h,l are distinct, $\mathbb{E}[W_{it}W_{ht}\bar{W}_{i,\cdot}^{-1}\bar{W}_{h,\cdot}^{-1}W_{is}W_{hu}(X_{js}-1)(X_{lu}-1)|I]=0$. For $\omega_{3,3}$, it is easy to see that $\frac{1}{N^2}\sum_{i,j=1}^{N}\frac{1}{T}\sum_{s=1}^{T}W_{it}\Lambda_i\bar{W}_{i,\cdot}^{-1}W_{is}\epsilon_{js}=O_p(\frac{1}{\delta_{N,T}})$. Furthermore,

$$\frac{1}{N^2}\sum_{i,j=1}^{N}\frac{1}{T}\sum_{s=1}^{T}W_{it}\Lambda_i\bar{W}_{i,\cdot}^{-1}W_{is}\Lambda_j^\top F_s = \underbrace{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{s=1}^{T}W_{it}\Lambda_i\bar{W}_{i,\cdot}^{-1}W_{is}F_s^\top}_{O_p(\frac{1}{\sqrt{T}})}\cdot\underbrace{\frac{1}{N}\sum_{j=1}^{N}\Lambda_j}_{O_p(\frac{1}{\sqrt{N}})} = O_p(\frac{1}{\delta_{N,T}}).$$

Therefore,

$$\omega_3 = \frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i \bar{W}_{i,\cdot}^{-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{T}\sum_{s=1}^{T} W_{is}(X_{js}-1)\alpha_j + \frac{1}{N}\sum_{i=1}^{N} W_{it}\Lambda_i \cdot \frac{1}{N}\sum_{j=1}^{N}\alpha_j + O_p(\frac{1}{\delta_{N,T}}).$$

Combining ω_1, ω_2 and ω_3, we get the results. □

Proof of Theorem 2:

Proof. The estimation error of the common component is

$$\begin{aligned}
\tilde{C}_{it} - C_{it} &= \left(\tilde{\mu} + \tilde{\alpha}_i + \tilde{\xi}_t + \tilde{\Lambda}_i^\top \tilde{F}_t\right) - \left(\mu + \alpha_i + \xi_t + \Lambda_i^\top F_t\right) \\
&= \Lambda_i^\top H^\top \left(\tilde{F}_t - (H^\top)^{-1} F_t\right) + F_t^\top H^{-1}\left(\tilde{\Lambda}_i - H\Lambda_i\right) + \tilde{\Delta}_{it} + O_p(\frac{1}{\delta_{N,T}}).
\end{aligned}$$

Following Lemmas 8, 13, 15, and 16, we can further decompose $\tilde{C}_{it} - C_{it}$ as

$$\begin{aligned}
\tilde{C}_{it} - C_{it} = \; & \underbrace{\Lambda_i^\top \hat{\Sigma}_{\Lambda,t}^{-1} \cdot \frac{1}{N}\sum_{j=1}^{N} W_{jt}\Lambda_j \epsilon_{jt}}_{\omega_{C_{it},1}} + \underbrace{F_t^\top H^{-1}\dot{D}^{-1} H \cdot \frac{1}{N}\sum_{j=1}^{N}\Lambda_j\Lambda_j^\top \frac{1}{|Q_{ij}|}\sum_{s\in Q_{ij}} F_s \epsilon_{is}}_{\omega_{C_{it},2}} \\
& + \underbrace{F_t^\top H^{-1}\dot{D}^{-1} H Z_i \Lambda_i - \Lambda_i^\top \hat{\Sigma}_{\Lambda,t}^{-1}\mathbf{Z}_t H^\top (\dot{D}^{-1})^\top (H^\top)^{-1} F_t}_{\omega_{C_{it},3}} \\
& + \underbrace{D_t^{-1}\frac{1}{N}\sum_{j=1}^{N}(X_{jt}-1)(\alpha_j + \Lambda_j^\top F_t + \epsilon_{jt}) - \bar{W}_{i,\cdot}^{-1}\frac{1}{T}\sum_{s=1}^{T} W_{is}\frac{1}{N}\sum_{j=1}^{N}(X_{js}-1)\alpha_j}_{\omega_{C_{it},4}} \\
& + \underbrace{\Lambda_i^\top \hat{\Sigma}_{\Lambda,t}^{-1}\frac{1}{N}\sum_{j=1}^{N} W_{jt}\Lambda_j \left[\bar{W}_{j,\cdot}^{-1}\frac{1}{T}\sum_{s=1}^{T} W_{js}\frac{1}{N}\sum_{l=1}^{N}(X_{ls}-1)\alpha_l - \frac{1}{N}\sum_{l=1}^{N}(X_{lt}-1)(\alpha_l + \Lambda_l^\top F_t + \epsilon_{lt})\right]}_{\omega_{C_{it},5}} \\
& - \underbrace{F_t^\top H^{-1}\dot{D}^{-1} H \frac{1}{N}\sum_{j=1}^{N}\Lambda_j\Lambda_j^\top \frac{1}{|Q_{ij}|}\sum_{s\in Q_{ij}} F_s F_s^\top \frac{1}{N}\sum_{l=1}^{N} X_{ls}\Lambda_l}_{\omega_{C_{it},6}} \\
& + \underbrace{\bar{W}_{i,\cdot}^{-1}\frac{1}{T}\sum_{s=1}^{T} W_{is}\epsilon_{is} + \bar{W}_{i,\cdot}^{-1}\Lambda_i^\top \frac{1}{T}\sum_{s=1}^{T} W_{is}F_s}_{\omega_{C_{it},7}} - \underbrace{\Lambda_i^\top \hat{\Sigma}_{\Lambda,t}^{-1}\frac{1}{N}\sum_{j=1}^{N} W_{jt}\bar{W}_{j,\cdot}^{-1}\Lambda_j\Lambda_j^\top \frac{1}{T}\sum_{s=1}^{T} W_{js}F_s}_{\omega_{C_{it},8}} \\
& - \underbrace{\Lambda_i^\top \hat{\Sigma}_{\Lambda,t}^{-1}\frac{1}{N}\sum_{j=1}^{N} W_{jt}\Lambda_j \frac{1}{N}\sum_{l=1}^{N}\left(\Lambda_l^\top F_t + \epsilon_{lt}\right) + D_t^{-1}\frac{1}{N}\sum_{j=1}^{N}\left(\Lambda_j^\top F_t + \epsilon_{jt}\right)}_{\omega_{C_{it},9}} + O_p(\frac{1}{\delta_{N,T}}),
\end{aligned}$$

where $\hat{\Sigma}_{\Lambda,t} = \frac{1}{N}\sum_{i=1}^N W_{it}\Lambda_i\Lambda_i^\top \xrightarrow{p} \Sigma_{\Lambda,t}$ by Assumption 5, $Z_i = \frac{1}{N}\sum_{j=1}^N \Lambda_j\Lambda_j^\top \Delta_{F,ij}$, $\mathbf{Z}_t = \frac{1}{N}\sum_{i=1}^N W_{it}Z_i\Lambda_i\Lambda_i^\top$, and $X_{it} = W_{it}/p_{it}$. The first three terms $\omega_{C_{it},1}$, $\omega_{C_{it},2}$ and $\omega_{C_{it},3}$ also appear in the asymptotic distribution of $\tilde{C}_{it} - C_{it}$ in the pure factor model in Xiong and Pelger (2023). According to Xiong and Pelger (2023), they are respectively asymptotically normal with zero mean. The other terms are brought by the first-stage estimation error of the fixed effects. We analyze them in the following.

Step 1 – The terms $\omega_{C_{it},4}$ and $\omega_{C_{it},5}$ are asymptotically normal

According to Lemma 8, $\omega_{C_{it},4}$ is asymptotically normal with zero mean.

For $\omega_{C_{it},5}$, by Assumption 6, $\text{plim}_{N\to\infty}\frac{1}{N}\sum_{j=1}^N W_{jt}\Lambda_j$ exists. Slutsky's theorem gives that $\frac{1}{N}\sum_{j=1}^N W_{jt}\Lambda_j\frac{1}{N}\sum_{l=1}^N (X_{lt}-1)(\alpha_l + \Lambda_l^\top F_t + \epsilon_{lt})$ is asymptotically normal with zero mean. Consider the term

$$\frac{1}{N}\sum_{j=1}^N W_{jt}\Lambda_j\bar{W}_{j,\cdot}^{-1}\frac{1}{T}\sum_{s=1}^T W_{js}\frac{1}{N}\sum_{l=1}^N (X_{ls}-1)\alpha_l.$$

We denote $G_j = \frac{1}{T}\sum_{s=1}^T W_{js}\frac{1}{N}\sum_{l=1}^N (X_{ls}-1)\alpha_l$. By Lemma 8, $[\sqrt{N}G_1, \sqrt{N}G_2, \cdots, \sqrt{N}G_N]$ is jointly asymptotically normal with asymptotic covariance $\sigma_{jl} = \text{ACov}(\sqrt{N}G_l, \sqrt{N}G_j)$. We denote $v_{jl} = \text{vec}(W_{jt}W_{lt}\bar{W}_{j,\cdot}^{-1}\bar{W}_{l,\cdot}^{-1}\Lambda_j\Lambda_l^\top)$ and $v_{jl,r}$ is the r-th entry in v_{jl}. We have

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{N^2}\sum_{j,l=1}^N \sigma_{jl}(v_{jl,r} - \mathbb{E}[v_{jl,r}])\right)^2\right]\\ =&\frac{1}{N^4}\sum_{i,j,h,l=1}^N \sigma_{ij}\sigma_{hl}\mathbb{E}\left[(v_{ij,r} - \mathbb{E}[v_{ij,r}])(v_{hl,r} - \mathbb{E}[v_{hl,r}])\right] = o(1),\end{aligned}$$

where the last equality holds because $\text{plim}_{N\to\infty}\frac{1}{N}\sum_{j=1}^N W_{jt}\bar{W}_{j,\cdot}^{-1}\Lambda_j$ exists in Assumption 6. Therefore, we have that $\text{plim}_{N\to\infty}\frac{1}{N^2}\sum_{i,j=1}^N W_{it}W_{jt}\bar{W}_{i,\cdot}^{-1}\bar{W}_{j,\cdot}^{-1}\sigma_{ij}\Lambda_i\Lambda_j^\top$ exists. This implies that $\frac{1}{N}\sum_{j=1}^N W_{jt}\Lambda_j\bar{W}_{j,\cdot}^{-1}\frac{1}{T}\sum_{s=}$ $1)\alpha_l$ is asymptotically normal with zero mean. Note that the randomness of both $\omega_{C_{it},4}$ and $\omega_{C_{it},5}$ come from the cross-sectional missingness, so they are jointly asymptotically normal.

Step 2 – The terms $\omega_{C_{it},6}$, $\omega_{C_{it},7}$, $\omega_{C_{it},8}$ and $\omega_{C_{it},9}$ are asymptotically normal

Based on Lemma 8 and due to the fact that $\text{plim}_{N\to\infty}\frac{1}{N}\sum_{j=1}^N W_{jt}\Lambda_j$ exists, $\omega_{C_{it},7}$ and $\omega_{C_{it},9}$ are asymptotically normal with zero mean. Consider $\omega_{C_{it},6}$, we have

$$\begin{aligned}&\frac{1}{N}\sum_{j=1}^N \Lambda_j\Lambda_j^\top\frac{1}{|Q_{ij}|}\sum_{s\in Q_{ij}} F_sF_s^\top\frac{1}{N}\sum_{l=1}^N X_{ls}\Lambda_l\\ =&\frac{1}{N}\sum_{j=1}^N \Lambda_j\Lambda_j^\top\frac{1}{|Q_{ij}|}\sum_{s\in Q_{ij}} F_sF_s^\top\frac{1}{N}\sum_{l=1}^N (X_{ls}-1)\Lambda_l + \frac{1}{N}\sum_{j=1}^N \Lambda_j\Lambda_j^\top\frac{1}{|Q_{ij}|}\sum_{s\in Q_{ij}} F_sF_s^\top\frac{1}{N}\sum_{l=1}^N \Lambda_l.\end{aligned}$$

It holds that $\frac{1}{N}\sum_{j=1}^N \Lambda_j\Lambda_j^\top\frac{1}{|Q_{ij}|}\sum_{s\in Q_{ij}} F_sF_s^\top \xrightarrow{p} \Sigma_\Lambda\Sigma_F$. By Slutsky's theorem, the second part in the decomposition of $\omega_{C_{it},6}$ is asymptotically normal with zero mean. For the first part, we denote $G_s = \frac{1}{N}\sum_{j=1}^N (X_{js}-1)\Lambda_j$. According to Lemma 8, $[\sqrt{N}G_1, \cdots, \sqrt{N}G_T]$ is jointly asymptotically normal with covariance $\mathrm{ACov}(\sqrt{N}G_s, \sqrt{N}G_u) = s_{\Lambda,st}$. By similar arguments with $\omega_{C_{it},5}$, the first part is asymptotically normal with zero mean. We can similarly show that $\omega_{C_{it},8}$ is asymptotically normal with zero mean.

Step 3 – Combining all the terms

We have proved that the nine terms $\omega_{C_{it},r}, r = 1, \cdots, 9$, which contribute to the asymptotic distribution of $\tilde{C}_{it} - C_{it}$ are all asymptotically normal with zero mean. Note that the randomness of $\omega_{C_{it},4}$ and $\omega_{C_{it},5}$ come from the cross-sectional average of the observation pattern. The randomness of $\omega_{C_{it},6}$ come from the cross-sectional average of the loadings. The randomness of $\omega_{C_{it},1}$, $\omega_{C_{it},2}$, $\omega_{C_{it},7}$ and $\omega_{C_{it},8}$ come from time series average of factors or idiosyncratic errors. The randomness of $\omega_{C_{it},9}$ comes from the cross-sectional average of the loadings and errors. We can show that $\frac{1}{\sqrt{T}}\sum_{t=1}^T W_{lt}F_t$ and $\sqrt{T}\mathrm{vec}(\frac{1}{T}\sum_{t=1}^T F_tF_t^\top - \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_tF_t^\top)$ are jointly asymptotically normal for any l, i, j. Therefore,

$$\sqrt{\delta_{N,T}}\sigma_{C,it}^{-1} \cdot \left(\tilde{C}_{it} - C_{it}\right) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\sigma_{C,it}^2 = \frac{\delta_{N,T}}{N}\sigma_{C,it,1}^2 + \frac{\delta_{N,T}}{T}\sigma_{C,it,2}^2$ with some $\sigma_{C,it,1}^2$ and $\sigma_{C,it,2}^2$.

Additionally, it is easy to show that

$$\sqrt{\delta_{N,T}}\bar{\sigma}_{C,it}^{-1} \cdot \frac{1}{N}\sum_{i=1}^N \left(\tilde{C}_{it} - C_{it}\right) \xrightarrow{d} \mathcal{N}(0,1)$$

with some $\bar{\sigma}_{C,it}^2 = \frac{\delta_{N,T}}{N}\bar{\sigma}_{C,it,1}^2 + \frac{\delta_{N,T}}{T}\bar{\sigma}_{C,it,2}^2$. □

IA.C.3 Proof of Proposition 1

Lemma 17. *Suppose the assumptions in Proposition 1 hold. Then, the error term $\dot{\epsilon}_{it}$ in the second step of wi-PCA with observables can be decomposed as $\dot{\epsilon}_{it} = \epsilon_{it} + \omega_{it}$, where ω_{it} satisfies $\mathbb{E}[\omega_{it}^8] \leq M/\delta_{N,T}^4$.*

Proof. By definition,

$$\dot{Y}_{it} = \beta^\top \dot{X}_{it} + \Lambda_i^\top F_t + \dot{\epsilon}_{it},$$

where $\dot{Y}_{it} = Y_{it} - \tilde{\mu} - \tilde{\xi}_t - \tilde{\alpha}_i$ and $\dot{X}_{it} = X_{it} - \bar{X}_{\cdot,t} - \bar{X}_{i,\cdot} + \bar{X}$. Rearranging the terms, we have

$$\begin{aligned} \dot{\epsilon}_{it} =& \ \dot{Y}_{it} - \beta^\top \dot{X}_{it} - \Lambda_i^\top F_t \\ =& \ \epsilon_{it} + \beta^\top\left(\bar{X}_{i,\cdot} + \bar{X}_{\cdot,t} - \bar{X}\right) - \left(\tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i - (\mu + \xi_t + \alpha_i)\right). \end{aligned}$$

We denote $\tilde{\Delta}_{it}^{\mathrm{old}}$ the first-stage estimation error without covariates X_{it}'s in Lemma 1. The first-stage estimation error with covariates can be written as

$$\begin{aligned}
&\tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i - (\mu + \xi_t + \alpha_i) \\
= \ & D_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \frac{W_{jt}}{p_{jt}} \left(\beta^\top X_{it} + \alpha_j + \Lambda_j^\top F_t + \epsilon_{jt} \right) + \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T} \sum_{s=1}^{T} W_{is} \left(\beta^\top X_{is} + \Lambda_i^\top F_s + \epsilon_{is} \right) \\
& - \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \frac{W_{js}}{p_{js}} D_s^{-1} \left(\beta^\top X_{js} + \alpha_j + \Lambda_j^\top F_s + \epsilon_{js} \right) \\
= \ & \tilde{\Delta}_{it}^{\text{old}} + D_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \frac{W_{jt}}{p_{jt}} \beta^\top X_{jt} + \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T} \sum_{s=1}^{T} W_{is} \beta^\top X_{is} \\
& - \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \frac{W_{js}}{p_{js}} D_s^{-1} \beta^\top X_{js} \, .
\end{aligned}$$

By similar arguments in Lemma 1, we can show that $\mathbb{E}[(\tilde{\Delta}_{it}^{\text{old}})^8] \leq M/\delta_{N,T}^4$. For the last term on the RHS, it holds that

$$\begin{aligned}
\frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \frac{W_{js}}{p_{js}} D_s^{-1} \beta^\top X_{js} = & \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \frac{W_{js}}{p_{js}} \left(D_s^{-1} - 1 \right) \beta^\top X_{js} \\
& + \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \left(\frac{W_{js}}{p_{js}} - 1 \right) \beta^\top X_{js} + \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} W_{is} \beta^\top X_{js}.
\end{aligned}$$

Since $\mathbb{E}[(D_s^{-1} - 1)^8] \leq M/N^4$ for any s and $\mathbb{E}[W_{js}/p_{js} - 1 | I] = 0$ for any j and s, it is easy to show that the eighth moments of the first term and second term on the RHS can be bounded by M/N^4. Therefore,

$$\begin{aligned}
&\tilde{\mu} + \tilde{\xi}_t + \tilde{\alpha}_i - (\mu + \xi_t + \alpha_i) \\
&= D_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \frac{W_{jt}}{p_{jt}} \beta^\top X_{jt} + \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T} \sum_{s=1}^{T} W_{is} \beta^\top X_{is} - \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T} \sum_{s=1}^{T} W_{is} \beta^\top \bar{X}_{\cdot,s} + \omega'_{it} \\
&= D_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \frac{W_{jt}}{p_{jt}} \beta^\top X_{jt} + \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T} \sum_{s=1}^{T} W_{is} \beta^\top (X_{is} - \bar{X}_{\cdot,s}) + \omega'_{it}
\end{aligned}$$

with some ω'_{it} satisfying $\mathbb{E}[\omega'^8_{it}] \leq M/\delta_{N,T}^4$.

Plugging this into $\dot{\epsilon}_{it}$, we get that

$$\dot{\epsilon}_{it} = \epsilon_{it} - D_t^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} \frac{W_{jt}}{p_{jt}} \beta^\top (X_{jt} - \bar{X}_{\cdot,t}) - \bar{W}_{i,\cdot}^{-1} \cdot \frac{1}{T} \sum_{s=1}^{T} W_{is} \beta^\top (X_{is} - \bar{X}_{\cdot,s} - \bar{X}_{i,\cdot} + \bar{X}) - \omega'_{it}.$$

We have $\mathbb{E}\left[(\frac{1}{N} \sum_{j=1}^{N} \frac{W_{jt}}{p_{jt}} \beta^\top (X_{jt} - \bar{X}_{\cdot,t}))^8 \right] \leq M/N^4$ because $\sum_{j=1}^{N} (X_{jt} - \bar{X}_{\cdot,t}) = 0$ and $\mathbb{E}[W_{jt}/p_{jt} | I] =$

1. According to Assumption 7, $\mathbb{E}\left[(\frac{1}{T}\sum_{s=1}^{T} W_{is}\beta^\top \dot{X}_{is})^8\right] \leq M/T^4$. We complete the proof. □

Lemma 18. *Suppose the assumptions in Proposition 1 hold. Then, the error term $\ddot{\epsilon}_{it}$ in the final step of wi-PCA with observables can be decomposed as $\ddot{\epsilon}_{it} = (\beta - \tilde{\beta})^\top \dot{X}_{it} + \dot{\epsilon}_{it} = \epsilon_{it} + \dot{\omega}_{it}$ with some $\dot{\omega}_{it}$ satisfying $\mathbb{E}[\dot{\omega}_{it}^4] \leq M/\delta_{N,T}^2$.*

Proof. We estimate the coefficient β as

$$\begin{aligned}\tilde{\beta} =& \left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\dot{X}_{it}\dot{X}_{it}^\top\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\dot{X}_{it}\dot{Y}_{it} \\ =& \ \beta + \left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\dot{X}_{it}\dot{X}_{it}^\top\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\dot{X}_{it}\left(\Lambda_i^\top F_t + \epsilon_{it} + \omega_{it}\right),\end{aligned}$$

where ω_{it} is defined in Lemma 17. Plugging this into $\ddot{\epsilon}_{it}$, we have

$$\ddot{\epsilon}_{it} = \epsilon_{it} + \underbrace{\omega_{it} - \dot{X}_{it}^\top \left(\sum_{j=1}^{N}\sum_{s=1}^{T} W_{js}\dot{X}_{js}\dot{X}_{js}^\top\right)^{-1} \sum_{j=1}^{N}\sum_{s=1}^{T} W_{js}\dot{X}_{js}\left(\Lambda_j^\top F_s + \epsilon_{js} + \omega_{js}\right)}_{\dot{\omega}_{it}}.$$

In the following, we prove that $\dot{\omega}_{it}$ has bounded fourth moment.

We denote the smallest eigenvalue of $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\dot{X}_{it}\dot{X}_{it}^\top$ to be $\lambda_{\min}$. By Assumption 7, there exists $q > 0$ such that $\lambda_{\min} \geq q > 0$. Therefore, it holds that

$$\begin{aligned}&\left\|\dot{X}_{it}^\top \left(\sum_{j=1}^{N}\sum_{s=1}^{T} W_{js}\dot{X}_{js}\dot{X}_{js}^\top\right)^{-1} \sum_{j=1}^{N}\sum_{s=1}^{T} W_{js}\dot{X}_{js}\left(\Lambda_j^\top F_s + \epsilon_{js} + \omega_{js}\right)\right\|^4 \\ \leq& \ q^{-4}\left\|\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} W_{js}\dot{X}_{js}\dot{X}_{it}^\top\left(\Lambda_j^\top F_s + \epsilon_{js} + \omega_{js}\right)\right\|^4 \\ =& \ q^{-4}\sum_{l,h=1}^{d}\left(\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} W_{js}\dot{X}_{js,l}\dot{X}_{it,h}\left(\Lambda_j^\top F_s + \epsilon_{js} + \omega_{js}\right)\right)^4.\end{aligned}$$

We can bound the fourth moment of $\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} W_{js}\dot{X}_{js,l}\dot{X}_{it,h}(\Lambda_j^\top F_s + \epsilon_{js})$ by M/T^2 since W and $\dot{X}$ are independent of F and ϵ. Furthermore, we can bound the fourth moment of $\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} W_{js}\dot{X}_{js,l}\dot{X}_{it,h}\omega$ by $M/\delta_{N,T}^2$ because of bounded moments of ω_{it} and $\dot{X}_{it}$. □

Following the proof of Theorem 1, we can prove that the proposed three-step method consistently estimates the loadings, factors, and common component.

References

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

HALL, P., AND C. C. HEYDE (2014): *Martingale limit theory and its application*. Academic Press.

XIONG, R., AND M. PELGER (2023): "Large dimensional latent factor modeling with missing observations and applications to causal inference," *Journal of Econometrics*, 233(1), 271–301.