Improved Inference for Interactive Fixed Effects Model with Cross Sectional Dependence

Zhenhao Gong *
University of Connecticut

Min Seong Kim †
University of Connecticut

September 17, 2021

Abstract

In this paper, we propose an improved inference procedure for the interactive fixed effects model in the presence of cross-sectional dependence and heteroskedasticity. It is well known in the literature that the LS estimator in this modelby Bai [2009] is asymptotically biased when the error term is cross-sectionally dependent, and we address this problem. Our procedure involves two parts: correcting the asymptotic bias of the LS estimator and employing the cross-sectional dependence robust covariance matrix estimator. We prove the validity of the proposed procedure in the asymptotic sense. Since our approach is based on the spatial HAC estimation, e.g., Conley (1999), Kelejian and Prucha (2007) and Kim and Sun (2011), we need a distance measure that characterizes the dependence structure. Such a distance may not be available in practice and we address this by considering a data-driven distance that does not rely on prior information. We also develop a bandwidth selection procedure based on a cluster wild bootstrap method. Monte Carlo simulations show our procedure work well in finite samples. As empirical illustrations, we apply the proposed method to make inference on the effects of divorce law reforms on the U.S. divorce rate, and the effects of clean water and sewerage interventions on the U.S. child mortality.

Keywords: Interactive fixed effects, Factor model, Bias correction, Robust inference, Data driven distance, Bandwidth selection, Time-series average spatial HAC estimator

*Address: 365 Fairfield Way, U-1063, Storrs, CT 06269, USA. Email: zhenhao.gong@uconn.edu
†Address: 365 Fairfield Way, U-1063, Storrs, CT 06269, USA. Email: min_seong.kim@uconn.edu

1 Introduction

Consider the following interactive fixed effects (IFE) model

$$Y_{it} = X_{it}'\beta_0 + u_{it}, \quad u_{it} = \lambda_i' F_t + \varepsilon_{it}, \quad i = 1, \cdots N;\ t = 1, \cdots, T, \tag{1}$$

where Y_{it} is an outcome variable; X_{it} is a $(p \times 1)$ vector of regressors; β_0 is a $(p \times 1)$ vector of unknown coefficients; u_{it} captures the unobserved heterogeneity, which is potentially correlated with X_{it}. We assume the factor-loading structure in u_{it}, where λ_i is a $(r \times 1)$ vector of factor loadings, F_t is a $(r \times 1)$ vector of common factors, and ε_{it} represents the idiosyncratic errors. The number of factors is r, and is assumed to be known. A crucial advantage of the interactive form is that it can be used to model the unobserved time-specific effects that impact the cross-sectional units heterogeneously. For example, in macroeconomics, unobserved common shocks would have heterogeneous impacts on different countries. The IFE model includes the standard additive fixed effects model as a special case but is significantly more flexible.

literature on IFE

This paper is about improving the inference for the IFE model with cross-sectionally dependent and heteroskedastic idiosyncratic errors. By employing our procedure, we can significantly improve the accuracy of inference on this model compared to the standard procedure by Bai [2009]. Our work is empirical relevant since the IFE model often require the idiosyncratic errors to be cross-sectionally correlated and heteroskedastic in practice. For example, for empirical studies using state-level data, while the cross-sectional correlations within a state can be captured by the factor structure, the cross-sectional correlations between states remain in the idiosyncratic errors. With large N and T, Bai [2009] proposed an IFE estimator of β_0 and shown that it is $\sqrt{NT}$ consistent and unbiased when the idiosyncratic errors are iid, but asymptotic bias exists in the presence of correlations and heteroskedasticities in both dimensions. This bias is caused by the estimation errors of F. Due to the growing dimension of F, we can only claim the spaces spanned by the estimator $\hat{F}$ and its true value F^0 are asymptotically the same. The terms left of the difference between the spaces spanned by $\hat{F}$ and F^0 contaminate

the estimation of β_0. This is known as incidental parameter problem (Neyman and Scott, 1948; Chamberlain, 1980; Nickell, 1981). Such bias is an important issue, and failure to control it can lead to misleading inference.

Our procedure involves two parts, correcting the asymptotic bias and estimating the covariance matrix of the IFE estimator. In the presence of serial correlation, we can estimate the bias by the truncated kernel method of Newey and West [1987]. But estimating the cross-sectional correlation bias is not straightforward, since the naturally ordered data are not available in the cross-sectional dimension. We focus on estimating it in this paper, assuming serial independent for simplicity. Due to the explicit form of the cross-sectional correlation bias has a long-run covariance structure, we introduce a time-series average spatial heteroskedasticity and autocorrelation (TA-SHAC) estimator to estimate it. Besides, we propose a TA-SHAC estimator to estimate the covariance matrix, which is robust under the cross-sectional correlation and heteroskedasticity.

Our work complements papers that provide bias correction and inference for the IFE model with cross-sectional correlation and heteroskedasticity. Based on our knowledge so far, there are two papers in the literature addressed this issue. Bai [2009] suggests the cross-sectional heteroskedasticity and autocorrelation (CS-HAC) estimator to estimate and correct the bias. The CS-HAC estimator uses a sub-sample in estimation, which is hard to implement into practice and different from our estimator since we use the whole sample. Bai and Liao [2017] develop a GLS method that focuses on the efficient estimation of β_0. By GLS transformation, their method can eliminate the bias caused cross-sectional correlation and is more efficient than the existing methods. But the GLS method has its own challenges. It is well know that the general GLS method has the side effect of invalid inference if the applied researcher did not model the heteroskedasticity correctly. We study the GLS method in our simulation and find that the inference of the GLS estimator may not practically reliable: confidence intervals do not generate the correct empirical coverage probabilities.

The TA-SHAC estimators we introduced have been discussed in a large literature. Conley [1996, 1999] is the first one introduced the spatial HAC estimation. Conley and Molinari

[2007] investigate the impact of distance measurement errors on the performance of parametric and nonparametric estimators. Kelejian and Prucha [2007] propose a spatial HAC estimator that models spatial dependence by a spatial weighting matrix. They assume that this weighting matrix is unknown and not parametrized. Kim and Sun [2011] generalize this estimator to apply to linear and nonlinear spatial models with moment conditions. They select the optimal bandwidth parameter based on the asymptotic truncated Mean Squared Error (MSE) criterion.

To drive the asymptotics and establish consistency, we decompose the difference of the TA-SHAC estimator and the true bias into three parts. The first part comes from the estimation errors of the parameters in the factor model. The second and third parts are the bias and variation of the infeasible estimator when we assume the model parameters are known. Due to the fact that the convergence rate of the variance is improved by taking the time-series average, the convergence rate of the estimation errors is no longer dominated by the variation and bias of the infeasible estimator. This is in contrast to the asymptotics of the standard spatial HAC estimators in the literature (e.g. Kelejian and Prucha [2007]; Kim and Sun [2011]; ect.), in which the optimal rate of convergence is achieved by balancing the bias and variation of the infeasible estimator. This result indicates that we cannot establish the bandwidth selection procedure based on the asymptotic MSE of our estimators.

There are two major challenges to implement our procedure in practice. The first challenge is how to choose a proper distance measure for the TA-SHAC estimators, which characterizes the dependence structure of data. In the literature, it is typically to find a relevant auxiliary variable as the distance, which captures the decaying pattern of dependence in the data (e.g. the transportation cost, Ligon and Conley [2001]; the geographic distance, Pinkse et al. [2002]; etc). But such auxiliary variable may not be available in some applications. For example, it would be impossible to find a variable that can be used as distance between the currency exchange rates in finance. We propose a data-driven distance measure that reflects the dependence structure of the data directly. This approach has been applied in many studies (e.g. Mantegna [1999]; Fernandez [2011]; Cui et al. [2020]; Kim [2021]; etc) and has a crucial advantage than the conventional distance, in which no prior information is required for implementation.

The second challenge is how to select the bandwidth parameters in our estimation procedure. This is particularly challenging in our setting, because we need to select two bandwidth parameters jointly in estimating the asymptotic bias and the covariance matrix. As we discuss before, we cannot establish the bandwidth selection procedure based on the asymptotic MSE. Instead, we propose a bootstrap-based bandwidth selection procedure. The problem is how to replicate the cross-sectional dependence of the data in one time period. As Gonçalves [2011], Vogelsang [2012], and Hidalgo and Schafgans [2017] suggested, we employ the cluster wild bootstrap approach, in which each cluster contains all cross-sectional units in one time period. By using a external variable that is common to all units in t, we can replicate the cross-sectional dependence of the original samples. Thus, the bootstrap process is expected to generate good approximations of the cross-sectional correlation bias and the corresponding covariance matrix. We show that the proposed bandwidth selection procedure performs well in the simulation with finite samples.

The remainder of the paper is as follows. Section 2 reviews the IFE estimator and corresponding asymptotics in Bai [2009]. Section 3 introduces the improved inference procedure for the IFE estimator. Section 4 provides the implementation procedure for our method. Section 5 discuss the GLS method and the possible extension of the proposed procedure to the dynamic IFE model. Section 6 presents the simulation design and results with finite samples. Section 7 applies our method to estimate the effects of divorce law reforms on the U.S. divorce rate, and the effects of clean water and sewerage interventions on U.S. child mortality. The last section concludes. All proofs are given in the Appendix.

2 IFE estimator and asymptotics

In this section, we review the IFE estimator and corresponding asymptotics in Bai [2009]. We can rewrite (1) as

$$Y_i = X_i\beta_0 + F\lambda_i + \varepsilon_i, \tag{2}$$

where $Y_i = (Y_{it}, \cdots, Y_{iT})'$, $X_i = (X_{it}, \cdots, X_{iT})$, $(T \times p)$, $F = (F_1, \cdots, F_T)'$, $(T \times r)$, and $\varepsilon_i = (\varepsilon_{it}, \cdots, \varepsilon_{iT})$. The IFE estimator to estimate β_0 is given by

$$\hat{\beta} = \arg\min_{\beta_0} \min_{F,\lambda_i} \sum_{i=1}^{N} (Y_i - X_i\beta_0 - F\lambda_i)'(Y_i - X_i\beta_0 - F\lambda_i), \tag{3}$$

subject to $\frac{1}{T}\sum_{t=1}^{T} F_t F_t' = I_r$ and $\sum_{i=1} \lambda_i\lambda_i'$ being diagonal. The two restriction conditions are used to identify factor loadings and factors in the factor structure. Concentrating out Λ, the least squares estimator for β_0 given $F = (F_1, F_2, \cdots, F_t)'$ is

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i' M_F X_i \right)^{-1} \sum_{i=1}^{N} X_i' M_F Y_i, \tag{4}$$

where $M_F = I_T - F(F'F)^{-1}F'$. Given β_0, the model reduces to a pure factor model, so we can estimate F by the principal components analysis (PCA). The estimator of F is equal to $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of the $T \times T$ matrix $\sum_{i=1}^{N}(Y_i - X_i\beta_0)(Y_i - X_i\beta_0)'$. Given $\hat{\beta}$ and $\hat{F}$, the estimator of Λ can be estimated by least square: $\hat{\Lambda} = (Y - X\hat{\beta})'\hat{F}/T$. Therefore, to estimate (β_0, F, Λ), we can start from a initial value of $\hat{\beta}_{in}$ and estimate $(\hat{F}_{in}, \hat{\Lambda}_{in})$ by PCA, and then simply iterated this process until convergence to get the estimators $(\hat{\beta}, \hat{F}, \hat{\Lambda})$.

We follow Bai [2003] and Bai [2009] to make the following assumptions. Throughout the paper, we define the Euclidean norm by $\|v\| = (v'v)^{1/2}$ for a vector v and the Frobenius norm by $\|A\|_F = (tr(A'A))^{1/2}$ for matrix A. We denote F^0 as the true parameter for F that satisfies Assumption A2 below and $\hat{F}$ as the estimator for a rotation of F^0.

Assumption A1. $E||X_{it}||^4 \leq M$ *and let* $\mathcal{F} = \{F : F'F/T = I\}$. *Define the* $p \times p$ *matrix*

$$H(F) = \frac{1}{NT}\sum_{i=1}^{N} X_i' M_F X_i - \frac{1}{T}\left[\frac{1}{N^2} \sum_{i=1}^{N}\sum_{k=1}^{N} X_i' M_F X_k a_{ik} \right], \tag{5}$$

where $a_{ik} = \lambda_i'(\Lambda'\Lambda/N)^{-1}\lambda_k$. *We assume* $\inf_{F\in\mathcal{F}} H(F) > 0$.

Assumption A2. *(i)* $E||F_t||^4 \leq M$ *and* $\frac{1}{T}\sum_{t=1}^{T} F_t F_t' \xrightarrow{p} \Sigma_F > 0$ *for some* $r \times r$ *matrix* Σ_F, *as*

$T \to \infty$*; (ii) We assume factor loading matrix* Λ *is non-random with its entries are bounded:* $||\lambda_i|| \leq C$ *for* $i = 1, \cdots, N$ *and* $\frac{1}{N}\Lambda'\Lambda \to Q_\Lambda$ *for some* $r \times r$ *positive definite matrix* Q_Λ*, as* $N \to \infty$*.*

Assumption A3.

(i) $E(\varepsilon_{it}) = 0$ *and* $E|\varepsilon_{it}|^8 \leq M$*;*

(ii) $E(\varepsilon_{it}\varepsilon_{ks}) = \sigma_{ik,ts}, |\sigma_{ik,ts}| < \bar{\sigma}_{ik}$ *for all* (i,k) *and* $|\sigma_{ik,ts}| < \tau_{ts}$ *for all* (t,s) *such that*

$$\frac{1}{N}\sum_{i,k=1}^{N}\bar{\sigma}_{ik} \leq M, \quad \frac{1}{T}\sum_{t,s=1}^{T}\tau_{ts} \leq M, \quad \frac{1}{NT}\sum_{i,k,t,s=1}|\sigma_{ik,ts}| \leq M.$$

(iii) For every (t,s)*,* $E\left|N^{-1/2}\sum_{i=1}^{N}\left[\varepsilon_{is}\varepsilon_{it} - E\left(\varepsilon_{is}\varepsilon_{it}\right)\right]\right|^4 \leq M$*.*

(iv) Moreover

$$T^{-2}N^{-1}\sum_{t,s,u,v}\sum_{i,k}\left|\text{cov}\left(\varepsilon_{it}\varepsilon_{is}, \varepsilon_{ku}\varepsilon_{kv}\right)\right| \leq M,$$
$$T^{-1}N^{-2}\sum_{t,s}\sum_{i,j,k,\ell}\left|\text{cov}\left(\varepsilon_{it}\varepsilon_{jt}, \varepsilon_{ks}\varepsilon_{\ell s}\right)\right| \leq M.$$

Assumption A4. ε_{it} *is independent of* X_{ks} *and* F_s *for all* i, t, k *and* s*.*

Assumption A5. *For some nonrandom positive definite matrix* H_Z*,*

$$\begin{aligned} &\text{plim}\,\tfrac{1}{NT}\textstyle\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\sigma_{ik,ts}Z_{it}Z_{ks}' = H_Z, \\ &\tfrac{1}{\sqrt{NT}}\textstyle\sum_{i=1}^{N}Z_i'\varepsilon_i \xrightarrow{d} N\left(0, H_Z\right), \end{aligned} \tag{6}$$

where $Z_i = M_{F^0}X_i - \frac{1}{N}\sum_{k=1}^{N}a_{ik}M_{F^0}X_k$ and a_{ik} defines in (5).

Assumption A1 indicates $H(F)$ is positive definite in the limit. It is used to identify β_0 and exclude the low-rank regressors (e.g. time-invariant and common regressors) in (2). Assumption A2 is a standard assumption for factor models. Under this assumption, the top r eigenvalues of the covariance matrix of Y diverge, while the rest of its eigenvalues are bounded as $N, T \to \infty$. It implies r factors and ensures the consistency of the PCA estimators for F and Λ in IFE model. Assumption A3 states the moment conditions and allows for weak serial and cross-sectional correlations and heteroskedasticities in the idiosyncratic errors. Assumption A4 rules out the dynamic panel data model for simplicity. Note that X_{it}, F_t, and ε_{it} are allowed to

be dynamic process; X_{it}, λ_i, and ε_{it} are allowed to be cross-sectionally correlated. Assumption A5 is a central limit theorem that is satisfied under various conditions.

Under 1-5 assumptions, for comparable N and T such that $T/N \to \rho > 0$, Bai [2009] shows that $\hat{\beta}$ is $\sqrt{NT}$ consistent and unbiased when the idiosyncratic errors are iid, but asymptotic bias appears in the presence of correlations and heteroskedasticities in both dimensions. This bias is caused by the estimation errors of F. Due to the growing dimension of F, we can only claim the spaces spanned by the estimator $\hat{F}$ and its true value F^0 are asymptotically the same. The terms left of the difference between the spaces spanned by $\hat{F}$ and F^0 contaminate the estimation of β_0. This is known as incidental parameter problem (Neyman and Scott, 1948; Chamberlain, 1980; Nickell, 1981). In the presence of serial correlation, we can estimate and correct the bias by the truncated kernel method of Newey and West [1987]. But estimating the cross-sectional correlation bias is not straightforward, since the naturally ordered data are not available in the cross-sectional dimension. We focus on estimating it in this paper.

Specifically, assume serial independence for simplicity, with $T/N \to \rho$, Bai [2009] shows that the distribution of the IFE estimator $\hat{\beta}$ is

$$\sqrt{NT}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(\rho^{1/2} B_0, H_0^{-1} H_Z H_0^{-1}\right), \tag{7}$$

where $H_0 = \text{plim} H(F^0)$ with $H(F^0)$ given in (5), and

$$H_Z = \text{plim} \frac{1}{NT} \sum_{i=1}^{N} \sum_{k=1}^{N} \sum_{t=1}^{T} E(\varepsilon_{it}\varepsilon_{kt}) Z_{it} Z_{kt}'. \tag{8}$$

The cross-sectional correlation bias B_0 is the probability limit of B_{NT} with

$$B_{NT} = -H\left(F^0\right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} w_i \lambda_k \left(\frac{1}{T} \sum_{t=1}^{T} E\varepsilon_{it}\varepsilon_{kt}\right), \tag{9}$$

where

$$w_i = \text{plim}\left[\frac{\left(X_i - V_i\right)' F^0}{T}\right] \left(\frac{F^{0\prime} F^0}{T}\right)^{-1} \left(\frac{\Lambda'\Lambda}{N}\right)^{-1} \text{ and } \quad V_i = \frac{1}{N} \sum_{k=1}^{N} a_{ik} X_k.$$

3 Improved inference procedure

In this section, we introduce our procedure to improve the inference of the IFE estimator $\hat{\beta}$. Our procedure involves two parts, estimating the cross-sectional bias B_{NT} and constructing a robust estimation for the covariance matrix H_Z.

3.1 Estimating the bias

To estimate B_{NT}, Bai [2009] suggests the CS-HAC estimator that allows for cross-sectional dependence and heteroskedasticity in ε_{it}. The estimator is defined as

$$\hat{B}_{CS} = -\hat{H}_0^{-1} \frac{1}{n_{sub}} \sum_{i=1}^{n_{sub}} \sum_{k=1}^{n_{sub}} \hat{w}_i \hat{\lambda}_k \left(\frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_{it} \hat{\varepsilon}_{kt} \right), \tag{10}$$

where n_{sub} is a sub-sample selected from the whole sample N such that $n_{sub}/\min\{N,T\} \to 0$, $\hat{H}_0$ and $\hat{w}_i$ are the estimators of H_0 and w_i with F^0, λ_i, and Λ replaced by $\hat{F}$, $\hat{\lambda}_i$, and $\hat{\Lambda}$. It can be shown that $\hat{B}_{CS}$ is consistent under the Assumption below.

Assumption B1. $E(\varepsilon_{it}\epsilon_{kt}) = \sigma_{ik}$ *for all i,k,t.*

This assumption implies that the covariance structure of $\{\varepsilon_{it}\}$ are time-invariant. The CS-HAC estimator is hard to implement into practice, however. Researchers need to select n_{sub} observations that can replicate the cross-sectional dependence structure of the whole sample. It can be shown that the performance of the CS-HAC estimator highly depends on this selection. If the n_{sub} units are randomly selected, then this estimator doesn't work at all since the selected observations do not maintain the cross-sectional dependence of the data. To the best of our knowledge, there is no practical guidance regarding this in the literature.

We propose a TA-SHAC estimator to estimate B_{NT}. Define

$$J_{NT} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} w_i \lambda_k \left(\frac{1}{T} \sum_{t=1}^{T} E\varepsilon_{it}\varepsilon_{kt} \right), \tag{11}$$

where w_i defines in (9). Then $B_{NT} = -H\left(F^0\right)^{-1} J_{NT}$. As $H\left(F^0\right)$ is easy to estimate by

replacing F^0, λ_i and Λ with their estimators, our focus is on consistent estimation of J_{NT}.

Define

$$J_t = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} w_i \lambda_k E(\varepsilon_{it}\varepsilon_{kt}). \tag{12}$$

We construct the TA-SHAC estimator of J_{NT} as

$$\hat{J}_{NT} = \frac{1}{T} \sum_{t=1}^{T} \hat{J}_t \text{ with } \hat{J}_t = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} K\left(\frac{d_{ik}}{d_n^{(1)}}\right) \hat{w}_i \hat{\lambda}_k \hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}, \tag{13}$$

where $K(\cdot)$ is a real-valued kernel function, $d_n^{(1)}$ is a bandwidth parameter, and d_{ik} is the distance measure between unit i and k that reflects the strength of cross sectional dependence. Note that $\hat{J}_t$ is a standard spatial HAC estimator in the literature (e.g., Kelejian and Prucha, 2007; Kim and Sun, 2011) and $\hat{J}_{NT}$ can be viewed as a time-series average of $\hat{J}_t, t = 1, \cdots, T$. Based on $\hat{J}_{NT}$, we can estimate B_{NT} by

$$\hat{B}_{NT} = -H(\hat{F})^{-1}\hat{J}_{NT}. \tag{14}$$

3.2 Estimating the covariance matrix

Recall the limit distribution of the IFE $\hat{\beta}$ in (7). We have already shown that the cross-sectional correlation bias B_{NT} can be estimated by $\hat{B}_{NT}$ in (14). To make valid inference, we also need to estimate the covariance matrix H_Z in (8), which is conventionally estimated as

$$\hat{H}_Z = \frac{1}{N} \sum_{i=1}^{N} \hat{\sigma}_i^2 \left(\frac{1}{T} \sum_{t=1}^{T} \hat{Z}_{it}\hat{Z}_{it}'\right), \tag{15}$$

where $\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_{it}^2$. This estimator does not take cross-sectional dependence into account and thus not valid when ε_{it} are cross-sectional correlated. Bai [2009] suggests that H_Z can be estimated by a consistent CS-HAC estimator, which is given by

$$\hat{H}_{CS} = \frac{1}{n_{sub}} \sum_{i=1}^{n_{sub}} \sum_{k=1}^{n_{sub}} \left(\frac{1}{T} \sum_{t=1}^{T} \hat{Z}_{it}\hat{Z}_{kt}'\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\right), \tag{16}$$

where $n_{sub}/\min\{N,T\} \to 0$. This estimator again is hard to implement into practice due to the sub-sample selection as we discussed before. In fact, instead of using the sub-sample n_{sub}, it can be shown that the estimator is also consistent if the whole sample size N is used in $\hat{H}_{CS}$ since $\frac{1}{N}\sum_{i=1}^{N}\hat{Z}_{it}\hat{\varepsilon}_{it} \neq 0$. That is, we can estimate H_Z directly by

$$\tilde{H}_{CS} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\hat{Z}_{it}\hat{Z}'_{kt}\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\right). \tag{17}$$

It can be shown that this estimator also works well in simulation.

We propose a TA-SHAC estimator to estimate H_Z, which is given by

$$\hat{H}_{NT} = \frac{1}{T}\sum_{t=1}^{T}\hat{H}_t \text{ with } \hat{H}_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\hat{Z}_{it}\hat{Z}'_{kt}\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}K_F\left(\frac{d_{ik}}{d_n^{(2)}}\right), \tag{18}$$

where $d_n^{(2)}$ is a bandwidth parameter, and $\hat{H}_t$ is a standard spatial HAC estimator. Inspired by Kim and Sun [2013], we employ the flat-top kernels $K_F(\cdot)$ to include the estimator $\tilde{H}_{CS}$ as a special case. The general form of flat-top kernels was introduced by Politis [2001] as

$$K_F = \left(K(\cdot) : K(x) = \begin{cases} 1 & \text{if } |x| \leq c_F \\ G(x) & \text{else} \end{cases}\right), \tag{19}$$

where $0 < c_F \leq 1$ and $G : |x| \in (c_F, 1] \to [0,1]$. A typical example of flat-top kernels is the trapezoidal kernel, in which $G(x) = \max\{(|x|-1)/(c_F-1), 0\}$. Thus, the rectangular kernel for the estimator $\tilde{H}_{CS}$ is an extreme case with $c_F = 1$.

The advantage of using flat-top kernels is that the corresponding estimators are higher-order accurate. The price is that the flat-top kernel estimators may not be positive semi-definiteness (psd), which is a highly desired property in the literature since Newey and West [1987]. Nevertheless, we can modify $\hat{H}_{NT}$ to be a psd estimator by the method suggested by Politis [2011]. That is, by applying eigen-decomposition to $\hat{H}_{NT}$, we have

$$\hat{H}_{NT} = U\Lambda U',$$

where Λ is a diagonal matrix that consist the eigenvalues of $\hat{H}_{NT}$ and U denotes the corresponding orthonormal eigenvectors. Then, the modified psd estimator can be defined as

$$\hat{H}_{NT}^{+} = U\Lambda^{+}U',$$

where $\Lambda^{+} = diag(\lambda_1^{+}, \cdots, \lambda_p^{+})$ with $\lambda_j^{+} = \max(\lambda_j, 0)$. It can be shown that $\hat{H}_{NT}^{+}$ maintains the higher-order accuracy property and has the same convergence rate as $\hat{H}_{NT}$. We apply this method in our simulation and empirical examples.

3.3 Asymptotic properties

This section establishes the consistency conditions of the TA-SHAC estimators $\hat{J}_{NT}$ and $\hat{H}_{NT}$. We start from the assumptions on the distance and kernel used in them.

Assumption B2. *(i)* $d_{ik} \geq 0, d_{ii} = 0$*, and* $d_{ik} = d_{ki}$*, (ii)* d_{ik} *is time invariant.*

This assumption implies that the TA-SHAC estimators does not require d_{ik} to satisfy the triangular inequality, $d_{ik} \leq d_{ij} + d_{jk}$, which is in contrast to the standard spatial HAC estimation in the literature (e.g., Conley, 1999; Kim and Sun, 2011). Data on economic distances usually contain measurement errors. Under certain conditions, we can generalize the results of this paper to the case when d_{ik} is contaminated by measurement errors. In this paper, however, we do not consider measurement errors for simplicity.

Assumption B3. *(i) The kernel* $K : \mathbb{R} \to [-1, 1]$ *satisfies* $K(0) = 1, K(x) = K(-x), K(x) = 0$ *for* $|x| \geq 1$*. (ii) For all* $x_1, x_2 \in R$ *there is a constant,* $c_L < 0$*, such that*

$$|K(x_1) - K(x_2)| \leq c_L |x_1 - x_2|.$$

Examples of kernels that satisfy this assumption are the Bartlett, Tukey-Hanning, and Parzen kernels. Next, we assume that ε_{it} has a linear representation

$$\varepsilon_{it} = \sum_{\ell=1}^{\infty} \gamma_{it,\ell} e_\ell, \tag{20}$$

where $\{\gamma_{it,\ell}\}$ are unknown constants and $\{e_\ell\}$ are iid innovations. This linear array process is commonly used to characterize spatial dependence in the literature (e.g. Kelejian and Prucha [2007]; Robinson [2011]; Kim and Sun [2011, 2013]; Pesaran and Tosetti [2011]; Kim [2021]), which includes the widely used spatial parametric models as special cases. By employing a linear array to establish the asymptotics, we avoid to introduce a mixing-type condition, which is difficult to justify in the cross-sectional dimension according to Bai and Ng [2006].

To establish the consistency of $\hat{J}_{NT}$, we define the infeasible version of $\hat{J}_{NT}$ as

$$\tilde{J}_{NT} = \frac{1}{T}\sum_{t=1}^{T} \tilde{J}_t \text{ with } \tilde{J}_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} K\left(\frac{d_{ik}}{d_n^{(1)}}\right) w_i \lambda_k \varepsilon_{it}\varepsilon_{kt}, \tag{21}$$

which is identical to $\hat{J}_{NT}$ but is based on the true value of w_i and λ_k. Using $\tilde{J}_{NT}$, the difference between $\hat{J}_{NT}$ and J_{NT} can be decomposed into three parts:

$$\hat{J}_{NT} - J_{NT} = (\hat{J}_{NT} - \tilde{J}_{NT}) + (\tilde{J}_{NT} - E\tilde{J}_{NT}) + (E\tilde{J}_{NT} - J_{NT}). \tag{22}$$

The first term is due to the effect of estimation errors in the factor model. The second and third terms represent the variance and bias of the infeasible estimator $\tilde{J}_{NT}$. The following assumptions are made to control the effect of estimation errors and characterize the variance and bias of $\tilde{J}_{NT}$.

Assumption B4. *$e_\ell \overset{iid}{\sim} (0,1)$ and $E(e_\ell^4) \leq \infty$, for all ℓ.*

Here we assume that $e_{\ell i}$ is independent of $e_{\ell k}$ for $i \neq k$. Under this assumption, J_{NT} in (11) can be expressed as

$$J_{NT} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} w_i \lambda_k \gamma_{it,\ell}\gamma_{kt,\ell}. \tag{23}$$

Assumption B5. *(i)* $\lim_{N,T\to\infty}\sum_{i=1}^{N}\sum_{t=1}^{T}|\gamma_{it,\ell}|<\infty$ *for all* ℓ*; (ii)* $\lim_{N,T\to\infty}\sum_{l=1}^{\infty}|\gamma_{it,\ell}|<\infty$ *for all* i *and* t*; (iii)* $\|w_i\|\leq C$ *for* $i=1,\cdots,N$.

This assumption requires the summation of the coefficients of the linear process in (20) to being finite, which corresponds to the weak dependence assumption of the idiosyncratic errors. Note that $|\gamma_{it,\ell}|$ can be interpreted as the absolute change of ε_{it} in response to one unit change in e_ℓ, so assumption B5 requires the aggregate response of ε_{it} to all innovations to be finite. We introduce this assumption to control the variance of $\tilde{J}_{NT}$.

Let

$$\ell_i=\sum_{k=1}^{N}1\left\{d_{ik}\leq d_n\right\}\text{ and }\bar{\ell}=\frac{1}{N}\sum_{i=1}^{N}\ell_i,$$

where ℓ_i is the number of pseudo-neighbors that unit i have within the bandwidth, and $\bar{\ell}$ is the average number of pseudo-neighbors. The number of pseudo-neighbors is increased with the bandwidth we choose.

Assumption B6. $\ell_i\leq c_\ell\bar{\ell}$ for all $i=1,\cdots,N$ with some constant c_ℓ.

This assumption allows different number of pseudo-neighbors for different units. It rules out the case that only a few units have many cross-sectional correlated units while others have none or very few.

The asymptotic bias of $\tilde{J}_{NT}$ are determined by the rate of decaying of the spatial dependence as well as the smoothness of kernel at zero. Let $q=\max\{q_0:K_{q_0}<\infty\}$ be the Parzen characteristic exponent of $K(x)$ with

$$K_{q_0}=\lim_{x\to 0}\frac{1-K(x)}{|x|^{q_0}}.\quad\text{ for }q_0\in[0,\infty)$$

Then, q is the largest value of q_0 for K_{q_0} to be finite, which reflects the smoothness of $K(x)$ at $x=0$.

Assumption B7. *There exists a finite constant* M *such that*

$$\lim_{N,T\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\|\Gamma_{ik,t}\|\,d_{ik}^{q}<M,\ \textit{with}\ \Gamma_{ik,t}=E\left(\varepsilon_{it}\varepsilon_{kt}\right).\tag{24}$$

This assumption characterizes the weak dependence between ε_{it} and ε_{kt} with respect to d_{ik}. The equation (24) implies that d_{ik} captures the decaying pattern of the dependence structure in the idiosyncratic errors, so that $\Gamma_{ik,t}$ decreases to zero fast enough as d_{ik} grows. This is a critical assumption for us to control the asymptotic bias of $\tilde{J}_{NT}$ caused by the truncation and downweight imposed by the kernel function. For example, when d_{ik} increases, the weight that $K(d_{ik}/d_n^{(1)})$ assigns to $\varepsilon_{it}\varepsilon_{kt}$ in (13) will decrease, which does not cause much bias under this assumption since $E(\varepsilon_{it}\varepsilon_{kt})$ also decreases.

The consistency of $\hat{J}_{NT}$ based on the decomposition in (22) is given in Theorem 1. The consistency of $\hat{H}_{NT}$ is given in Theorem 2. The proofs are all contained in Appendix.

Theorem 1. *Under the Assumptions A1-A4 and B2-B7, and* $d_n, l_n, N, T \to \infty$ *such that* $l_n/N, l_n/T \to \infty$ *and* $T/N \to \rho$*, we have* $\hat{J}_{NT} - J_{NT} = o_p(1)$.

Theorem 2. *Under the Assumptions A1-A4 and B1-B7, and* $d_n, l_n, N, T \to \infty$ *such that* $l_n/N, l_n/T \to \infty$ *and* $T/N \to \rho$*, we have* $\hat{H}_{NT} - H_Z = o_p(1)$.

Define

$$\hat{\beta}^\dagger = \hat{\beta} - \frac{1}{N}\hat{B}_{NT}.$$

Corollary 1. *Under the Assumptions of Theorem 1 and 2, then*

$$\frac{\sqrt{NT}(\hat{\beta}^\dagger - \beta_0)}{\sqrt{\hat{H}_0^{-1}\hat{H}_Z\hat{H}_0^{-1}}} \xrightarrow{d} N(0,1).$$

4 Implementation

As we introduced before, there are two major challenges for implementing our procedure in practice. The first challenge is how to choose an appropriate distance measure d_{ik} for the TA-SHAC estimators. In the literature, it is typically to find a relevant auxiliary variable as the distance, which captures the decaying pattern of dependence in the data. For example, Ligon and Conley [2001] use the transportation cost, Pinkse et al. [2002] use the geographic distance etc. But such auxiliary variable may not be available in some applications, such as the currency

exchange rates in finance. To address this issue, we propose a data-driven distance measure that reflects the dependence structure directly. Specifically, define

$$d_{ik}^{\mathrm{D}} = \frac{1}{|\rho_{ik}|} - 1,$$

where $\rho_{ik} = Corr(\varepsilon_{it}, \varepsilon_{kt})$. d_{ik}^{D} captures the degree of dependence by definition. Note that d_{ik}^{D} is unobservable and does not satisfy the triangular inequality, $d_{ik} \leq d_{ij} + d_{jk}$, but we can estimate it by its sample counterpart

$$\hat{d}_{ik}^{D} = \min\left\{1/|\hat{\rho}_{ik}|, 100\right\} - 1,$$

with $\hat{\rho}_{ik} = \sum_{t=1}^{T} \hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} / \sqrt{\sum_{t=1}^{T} \hat{\varepsilon}_{it}^2 \sum_{t=1}^{T} \hat{\varepsilon}_{kt}^2}$ and show that our estimators are still valid without the triangular inequality. This approach has been applied in many applications (e.g. Mantegna [1999]; Fernandez [2011]; Cui et al. [2020]; Kim [2021]) and has a crucial advantage than the conventional distance, in which no prior information is required for implementation. We apply this approach in the simulation and the empirical applications.

The second challenge is how to select the bandwidth parameters properly. This is particularly challenging in our setting, because we need to select two bandwidth parameters jointly in estimating the asymptotic bias and the covariance matrix. In addition, the conventional MSE optimal bandwidth parameter is not applicable as we discussed before. We propose a bootstrap-based bandwidth selection procedure. The idea of this procedure comes from Kim et al. [2017], in which they also need to select two smoothing parameters in their test procedure. But we can not apply their method directly, since the time-series dependence sample can be simply generated by a regular AR model. The problem in our setting is how to replicate the cross-sectional dependence sample in one time period. Gonçalves [2011], Vogelsang [2012], and Hidalgo and Schafgans [2017] suggest we can employ the cluster wild bootstrap approach, in which each cluster contains all cross-sectional units in one time period. By using a external variable that is common to all units in t, we can replicate the cross-sectional dependence of the original samples. Thus, the bootstrap process is expected to generate a good approximation of

the cross-sectional correlation bias and the corresponding covariance matrix.

Specifically, let $\mathcal{D}_{nM}^{(1)} = \{d_{n1}^{(1)}, \cdots, d_{nM}^{(1)}\}$ and $\mathcal{D}_{nS}^{(2)} = \{d_{n1}^{(2)}, \cdots, d_{nS}^{(2)}\}$ be the sets of reasonable bandwidth parameters $d_n^{(1)}$ and $d_n^{(2)}$ for a given sample size. The procedure involves the following steps.

Step 1: Estimate $\hat{\beta}$, $\hat{F}_t$, $\hat{\Lambda}$ by the iteration procedure used in Bai [2009] and the error terms by

$$\hat{\varepsilon}_t = Y_t - X_t\hat{\beta} - \hat{\Lambda}\hat{F}_t.$$

Step 2: Generate bootstrap sample Y_t^* based on

$$Y_t^* = X_t\hat{\beta} + \hat{\Lambda}\hat{F}_t + \varepsilon_t^*,$$
$$\varepsilon_t^* = \hat{\varepsilon}_t\xi_t \text{ with } \xi_t \overset{iid}{\sim} (0,1).$$

Step 3: Estimate the bootstrap version of $\hat{\beta}^*$, $\hat{F}_t^*$, $\hat{\Lambda}^*$, and $\hat{\varepsilon}_t^*$ as step 1. Construct the bootstrap version of the bias estimator $\hat{B}_{NT}^*\left(d_{nm}^{(1)}\right)$ with $d_{nm}^{(1)} \in \mathcal{D}_{nM}^{(1)}$ such that

$$\hat{B}_{NT}^*\left(d_{nm}^{(1)}\right) = -H\left(\hat{F}^*\right)^{-1}\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}K\left(\frac{d_{ik}}{d_{nm}^{(1)}}\right)\hat{w}_i^*\hat{\lambda}_k^*\hat{\varepsilon}_{it}^*\hat{\varepsilon}_{kt}^*\right],$$

where $H(\hat{F}^*)$ and $\hat{w}_i^*$ are the bootstrap version estimators of $H\left(F^0\right)$ and w_i with F^0, λ_i, and Λ replaced by $\hat{F}^*$, $\hat{\lambda}_i^*$, and $\hat{\Lambda}^*$.

Step 4: Estimate the bootstrap version of the covariance matrix estimator $\hat{H}_{NT}^*\left(d_{ns}^{(2)}\right)$ with $d_{ns}^{(2)} \in \mathcal{D}_{nS}^{(2)}$ such that

$$\hat{H}_{NT}^*\left(d_{ns}^{(2)}\right) = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\hat{Z}_{it}^*\hat{Z}_{kt}^{*\prime}\hat{\varepsilon}_{it}^*\hat{\varepsilon}_{kt}^*K_F\left(\frac{d_{ik}}{d_{ns}^{(2)}}\right)\right],$$

where $Z_i^* = M_{F^*}X_i - \frac{1}{N}\sum_{k=1}^{N}a_{ik}^*M_{F^*}X_k$, $M_{F^*} = I_T - F^*(F^{*\prime}F^*)^{-1}F^{*\prime}$ and $a_{ik}^* = \lambda_i^{*\prime}(\Lambda^{*\prime}\Lambda^*/N)^{-1}\lambda_k^*$.

Step 5: Generate $\mathcal{B}$ bootstrap samples and compute the bootstrap based t-test statistics

$$t_b^*(d_{nm}^{(1)}, d_{ns}^{(2)}) = \frac{\hat{\beta}^{\dagger *}}{se(\hat{\beta}^*)}, \text{ for } b = 1, 2, \cdots, \mathcal{B},$$

where $\hat{\beta}^{\dagger *}$ is the bias corrected estimator and $se(\hat{\beta}^*)$ is the standard error for $\hat{\beta}^*$:

$$\hat{\beta}^{\dagger *} = \hat{\beta}^* - \frac{1}{N} B_{NT}^* \left(d_{nm}^{(1)}\right) \text{ and } se(\hat{\beta}^*) = \sqrt{\frac{H(\hat{F}^*)^{-1} \hat{H}_{NT}^* \left(d_{ns}^{(2)}\right) H(\hat{F}^*)^{-1}}{NT}}.$$

Step 6: Repeat Step 2 to Step 5 for each $(d_{nm}^{(1)}, d_{ns}^{(2)}) \in \mathcal{D}_{nM}^{(1)} \bigotimes \mathcal{D}_{nS}^{(2)}$. Compute

$$\mathcal{V}(d_{nm}^{(1)}, d_{ns}^{(2)}) = \frac{1}{\mathcal{B}} \sum_{b=1}^{\mathcal{B}} 1(|t_b^*(d_{nm}^{(1)}, d_{ns}^{(2)})| > t^{\alpha/2}),$$

and select $\left(d_{nm}^{(1*)}, d_{ns}^{(2*)}\right)$ that solves

$$\max_{d_{nm}^{(1)} \in \mathcal{D}_{nM}^{(1)}, d_{nm}^{(2)} \in \mathcal{D}_{nM}^{(2)}} \mathcal{V}(d_{nm}^{(1)}, d_{ns}^{(2)}), \quad s.t.\ \mathcal{V}(d_{nm}^{(1)}, d_{ns}^{(2)}) \leq \alpha.$$

Note that we employ the cluster wild bootstrap to generate bootstrap sample Y_t^* in step 2, in which each cluster contains all cross-sectional units in one time period. The external random variable ξ_t replicates the cross-sectional dependence of original sample for all units in time period t. Hence, $\hat{B}_{NT}^* \left(d_{nm}^{(1)}\right)$ and $\hat{H}_{NT}^* \left(d_{ns}^{(2)}\right)$ are expected to be a good approximation to $\hat{B}_{NT} \left(d_{nm}^{(1)}\right)$ and $\hat{H}_{NT} \left(d_{ns}^{(2)}\right)$. We generate ξ_t from Rademacher Distribution in our simulation and empirical application.

Based on $\left(d_{nm}^{(1*)}, d_{ns}^{(2*)}\right)$, the proposed confidence interval for β_0 at a $100(1-\alpha)\%$ level is

$$CI(\beta_0) = \left[\hat{\beta}^\dagger - \mathfrak{q}_{\alpha/2} \sqrt{se(\hat{\beta})}, \hat{\beta}^\dagger + \mathfrak{q}_{1-\alpha/2} \sqrt{se(\hat{\beta})}\right],$$

where $\hat{\beta}^{\dagger}$ is the bias corrected estimator and $se(\hat{\beta})$ is the robust standard error for $\hat{\beta}$ such that

$$\hat{\beta}^{\dagger} = \hat{\beta} - \frac{1}{N}\hat{B}_{NT}\left(d_{nm}^{(1*)}\right) \text{ and } se(\hat{\beta}) = \sqrt{\frac{H(\hat{F})^{-1}\hat{H}_{NT}\left(d_{ns}^{(2*)}\right)H(\hat{F})^{-1}}{NT}}.$$

Thus, our bootstrap-based bandwidth selection procedure is designed to choose $\left(d_{nm}^{(1*)}, d_{ns}^{(2*)}\right)$ jointly that improves the inference of the IFE estimator $\hat{\beta}$ by two parts, correcting the asymptotic bias and estimating the covariance matrix.

Different bootstrap methods could be used in step 2 as long as they replicate the cross-sectional dependence in one time period. (e.g. the CSD bootstrap by Gonçalves and Perron [2020]). We may also consider a parametric bootstrap-based spatial regression model if the location of each unit is available. The theoretical properties of our cluster wild bootstrap procedure are not examined in this paper, and we leave it for future research.

5 Discussion and extension

An alternative way to address the cross-sectional correlation bias is the GLS method proposed by Bai and Liao [2017]. They focus on the efficient estimation of β_0. The corresponding GLS estimator is given by

$$\hat{\beta}(\hat{\Sigma}_{\varepsilon}^{-1}) = \arg\min_{\beta_0}\min_{F,\lambda_i}\sum_{i=1}^{N}(Y_i - X_i\beta_0 - F\lambda_i)'\hat{\Sigma}_{\varepsilon}^{-1}(Y_i - X_i\beta_0 - F\lambda_i), \tag{25}$$

where $\hat{\Sigma}_{\varepsilon}$ is a consistent estimator of the covariance matrix of ε_{it}, which is a high-dimensional matrix. By using $\hat{\Sigma}_{\varepsilon}^{-1}$ as the optimal weight matrix, they can eliminate the bias due to cross-sectional correlation and heteroskedasticity. To estimate Σ_{ε}, they assume it is a sparse covariance matrix and apply the thresholding method in Fan et al. [2013] by estimating the small entries to be zero directly. The estimation of Σ_{ε} requires the choice of tuning parameters for the thresholding value, which can be chosen through multifold cross-validation. See Bai et al. [2020] for the details.

By applying the GLS method, there is no need to correct the bias and use the robust covariance matrix, since the GLS transformation can take account both of them. Besides, it is more efficient than the existing methods. But the GLS method has its own challenges. It is well know that the general GLS method has the side effect of invalid inference if the applied researcher did not model the heteroskedasticity correctly. For example, Angrist and Pischke [2010] argue that "If the conditional variance model is a poor approximation or if the estimates of it are very noisy, weighted least squares (WLS) estimators may have worse finite-sample properties than unweighted estimators." We study the performance of the GLS estimator in our simulation and find that its inference may not practically reliable: confidence intervals do not generate the correct empirical coverage probabilities.

A potential extension of the proposed method is to improve the inference for the dynamic panel data model with interactive fixed effects. By allowing the predetermined regressors (e.g. lagged-dependent variables) in the IFE model, Moon and Weidner [2015] found two sources of asymptotic biases of the least squares (LS) estimator. The first type of bias is the same bias as Bai [2009] and the other type of bias arises from the predetermined regressors. In their bias correction procedure, they proposed consistent estimators of the biases under heteroskedasticity assuming no correlations in the idiosyncratic errors. But their estimators are not valid when the idiosyncratic errors are correlated in both dimensions. The bias caused by the time-series correlated errors and the predetermined regressors can be estimated by the truncated kernel method of Newey and West [1987]. The problem is how to choose the bandwidth parameters for the corresponding bias estimators. In the presence of cross-sectional correlation and heteroskedasticity, we can apply the proposed procedure to improve the inference of the LS estimator by estimating the asymptotic bias and the covariance matrix. We leave this for our future research.

6 Monte Carlo Simulation

In this section, we investigate the finite sample performance of the proposed procedure for correcting the bias and improving the inference of the IFE estimator $\hat{\beta}$. Follow Bai [2009], the data generating process (DGP) we consider is

$$Y_{it} = X_{it}\beta_0 + \lambda_i' F_t + \varepsilon_{it},$$

where the true value of $\beta_0 = 1$. The number of common factors $r = 2$, which is assumed to be known. The regressors and factors are generated according to

$$\begin{aligned}
&X_{it} = \mu + c\lambda_i' F_t + \iota'\lambda_i + \iota' F_t + \eta_{it}; \text{ with } \iota' = (1,1), \\
&F_{rt} = \rho F_{r,t-1} + \sqrt{1-\rho^2} v_{rt}, r = 1,2; \\
&\lambda_{ir}, \eta_{it}, v_{rt} \overset{iid}{\sim} N(0,1).
\end{aligned}$$

We set $c = \mu = 1$ and $\rho = 0.3$, so there is a weak serial correlation between factors. We generate the cross-sectional correlated data using a popular spatial MA model. The design is based on an $(L_N \times L_N)$ square integer lattice structure $(L_N = 14, 16)$, where unit i is located on a square grid of integers (i_1, i_2) such that

$$\varepsilon_t = (I_n + \theta M_1 + \theta^2 M_2) v_t, \quad t = 1, 2, \cdots, T$$

where $\varepsilon_t = (\varepsilon_{t1}, \cdots, \varepsilon_{tN})'$, $v_t = (v_{t1}, \cdots, v_{tN})'$ and v_{it} is i.i.d $N(0,1)$. $M_1 = [m_{1,ik}]_{i,k=1}^N$ and $M_2 = [m_{2,ik}]_{i,k=1}^N$ are $(N \times N)$ spatial weighting matrices such that

$$m_{1,ik} = \begin{cases} 1 & \text{if } d_{ik} = 1 \\ 0 & \text{if } d_{ik} \neq 1 \end{cases} \text{ and } m_{2,ik} = \begin{cases} 1 & \text{if } d_{ik} = \sqrt{2} \\ 0 & \text{if } d_{ik} \neq \sqrt{2} \end{cases},$$

where $d_{ik} = \max\{|i_1 - k_1|, |i_2 - k_2|\}$. Thus, units i and k are cross-sectional dependent if the distance between them is 1 or $\sqrt{2}$. The distance between two units is measured by Euclidean distance.

To construct the TA-SHAC estimators, we use the data-driven distance measure d_{ik}^D that is defined as

$$d_{ik}^{\mathrm{D}} = \frac{1}{|\rho_{ik}|} - 1,$$

where $\rho_{ik} = Corr(\varepsilon_{it}, \varepsilon_{kt})$. By definition, we can see that d_{ik}^D is a decrease function of the correlation and reflects the degree of dependence between unit i and k. While, as we discussed before, d_{ik}^D does not satisfy the triangular inequality so it is not a valid distance, we can show that our estimators are still valid. Also, d_{ik}^D is unobservable in practice, but we can use the sample counterpart as

$$\hat{d}_{ik}^D = \min\left\{1/|\hat{\rho}_{ik}|, 100\right\} - 1,$$

where $\hat{\rho}_{ik} = \sum_{t=1}^T \hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} / \sqrt{\sum_{t=1}^T \hat{\varepsilon}_{it}^2 \sum_{t=1}^T \hat{\varepsilon}_{kt}^2}$. Note that we don't need any prior information about the dependence structure to construct $\hat{d}_{ik}^D$, which is a critical advantage than the conventional distance. To select the bandwidth parameters $(d_n^{(1)}, d_n^{(2)})$, we apply the bootstrap-based bandwidth selection procedure in section 5. We choose bandwidth parameter sets to be $\mathcal{D}_{nM}^{(1)} = \mathcal{D}_{nS}^{(2)} = \{2:10\}$, so we have (9×9) different pairs of $(d_n^{(1)}, d_n^{(2)})$ for selection. In the simulation, we find that the almost all of the selected bandwidth $\left(d_{nm}^{(1*)}, d_{ns}^{(2*)}\right)$ in step 6 fall in the interior of the bandwidth parameter sets, so we believe $\mathcal{D}_{nM}^{(1)}$ and $\mathcal{D}_{nS}^{(2)}$ are reasonable sets for our selection. For the kernel function, we employ Parzen kernel for estimating the bias term $\hat{B}_{NT}\left(d_n^{(1*)}\right)$, and Bartlett kernel (a special case of flat-top kernel) for estimating the covariance matrix $\hat{H}_{NT}\left(d_n^{(2*)}\right)$.

In Table 1, we report the scaled biases and root mean square error (RMSE) for different estimators with 1000 repetitions. $B(\hat{\beta})$ is scaled bias that equals $\sqrt{NT}$ times the difference between the IFE estimator $\hat{\beta}$ and its true value β_0. With similar interpretations, $B(\hat{\beta}_{gls})$, $B(\hat{\beta}_{hac}^*)$, and $B(\tilde{\beta}_{hac}^*)$ are the scaled biases for the GLS estimator, TA-SHAC estimator with data-driven distance, and TA-SHAC estimator with true distance. RMSE is the corresponding $\sqrt{NT}$ times RMSE for each estimator. The reason why we scaled the bias and RMSE with $\sqrt{NT}$ is that the IFE estimator $\hat{\beta}$ is $\sqrt{NT}$ consistent, so the inference of $\hat{\beta}$ is affected by the $\sqrt{NT}$ scaled bias.

The results in Table 1 show that when there is no cross-sectional correlation in ε_{it} $(\theta = 0)$,

the scaled biases and corresponding RMSE for all of the estimators are similar. When there exists weak cross-sectional dependence in ε_{it} ($\theta = .4$), a few patterns emerge. First, the asymptotic biases of $\hat{\beta}$ are almost twice than the case without cross-sectional correlation. For example, when $T = 150$, $N = 144$ and $\theta = 0$, $B(\hat{\beta}) = 0.801$; when $\theta = .4$, $B(\hat{\beta}) = 1.642$. Second, when N is fixed, the asymptotic bias of $\hat{\beta}$ increases as T increasing. For example, when $T = 50$ and $N = 144$, $B(\hat{\beta}) = 1.579$; when $T = 150$ and $N = 144$, $B(\hat{\beta}) = 1.642$. This is consistent with the theory, since the asymptotic bias of $\hat{\beta}$ in (7) depends on $\rho = T/N$. Third, $\hat{\beta}^*_{hac}$ have similar performance $\tilde{\beta}^*_{hac}$ in terms of bias correction. For example, when $T = 150$, $N = 144$, $B(\hat{\beta}) = 1.642$; while $B(\tilde{\beta}^*_{hac}) = 1.367$ and $B(\hat{\beta}^*_{hac}) = 1.383$. This implies that the data-driven distance measure is valid for bias correction. Lastly, the GLS estimator $\hat{\beta}_{gls}$ performs the best in terms of reducing bias and RMSE. For example, when $T = 150$, $N = 144$, $B(\hat{\beta}_{gls}) = 0.617$; while $B(\tilde{\beta}^*_{hac}) = 1.367$ and $B(\hat{\beta}^*_{hac}) = 1.383$.

Table 2 presents the empirical coverage probabilities (EPCs) of the 95% confidence intervals for different estimators. $\hat{\beta}_{hac1}$, $\hat{\beta}_{hac2}$, and $\hat{\beta}^*_{hac}$ are the TA-SHAC estimators using the data-driven distance. For $\hat{\beta}_{hac1}$, we estimate covariance matrix only by $\hat{H}_{NT}$ in (18) without bias correction. For $\hat{\beta}_{hac2}$, we correct the bias only by $\hat{B}_{NT}$ in (14) with the conventional covariance matrix estimator in (15). They are used to compare which part (bias correction or robust covariance estimation) is more important for inproving the inference of $\hat{\beta}$. For $\hat{\beta}^*_{hac}$, we both correct the bias and estimate the covariance matrix by the proposed estimators. For comparison, we use the true distance measure d^T_{ik} in $\tilde{\beta}_{hac1}$, $\tilde{\beta}_{hac2}$, and $\tilde{\beta}^*_{hac}$ with similar interpretations.

From the results in Table 2, we can see that when there is no correlation in ε_{it} ($\theta = 0$), all EPCs of those estimators are close to the nominal coverage probability (0.95). However, when there exist weak cross-sectional correlation in ε_{it} ($\theta = .4$), the EPCs of $\hat{\beta}$ is not valid. For example, when $N = 144$, $T = 150$, the EPC of $\hat{\beta}$ decreases to 0.777. Also, the EPCs of $\hat{\beta}_{gls}$ is not robust when T is large. For example, when $N = 144$, $T = 150$, the EPC of $\hat{\beta}_{gls}$ decreases to 0.797. In contrast, $\hat{\beta}^*_{hac}$ and $\tilde{\beta}^*_{hac}$ perform well in the present of weak cross-sectional correlation in ε_{it} and robust to different combination of N and T. Also, they have better performance with larger N. For example, when $N = 144$, $T = 150$, the EPC for $\hat{\beta}^*_{hac}$

is 0.86; when $N = 200$, it increases to 0.911. Furthermore, the TA-SHAC estimators are able to improve the EPCs regardless of the true distance measure d_{ik}^T or data-driven distance measure d_{ik}^D is used, although the one with true distance measure performs slightly better in general. For example, when $N = 144$, $T = 150$, the EPC for $\tilde{\beta}_{hac}^*$ using the true distance is 0.878, while the EPC for $\hat{\beta}_{hac}^*$ using the data-driven distance is 0.86. This finding gives us an important implication from an empirical point of view. That is we can use our method with the data-driven distance measure, which can be directly obtained from time-series observations. Besides, in terms of improving the EPCs, the performance of $\hat{\beta}_{hac1}$ and $\hat{\beta}_{hac2}$ clearly show that bias correction is more important than using the robust covariance matrix. For example, when $N = 144$ and $T = 150$, $\hat{\beta}_{hac2}$ can improve the EPC of $\hat{\beta}$ from 0.777 to 0.854, while $\hat{\beta}_{hac1}$ can only improve the EPC of $\hat{\beta}$ from 0.777 to 0.806.

In conclusion, in the present of weak cross-sectional correlation ($\theta = .4$), the IFE estimator $\hat{\beta}$ is biased with invalid inference; although the GLS estimator $\hat{\beta}_{gls}$ has the best performs in terms of reducing the bias, its inference is not robust when T is large; the TA-SHAC estimators can both reduce the bias and provide robust inference with different combinations of N and T.

Table 1: Scaled bias and RMSE of different estimators

						TA-SHAC (d_{ik}^T)		TA-SHAC (d_{ik}^D)	
T	N	$B(\hat{\beta})$	RMSE	$B(\hat{\beta}_{gls})$	RMSE	$B(\tilde{\beta}^*_{hac})$	RMSE	$B(\hat{\beta}^*_{hac})$	RMSE
					$\theta = 0$				
50	144	0.848	1.072	0.867	1.098	0.849	1.073	0.841	1.056
100		0.832	1.061	0.843	1.071	0.833	1.062	0.805	1.011
150		0.801	1.026	0.808	1.042	0.802	1.027	0.856	1.058
200		0.792	0.989	0.804	1.003	0.793	0.990	0.785	1.004
50	196	0.790	1.017	0.818	1.044	0.790	1.018	0.828	1.022
100		0.864	1.075	0.871	1.082	0.864	1.075	0.801	1.008
150		0.800	1.015	0.815	1.036	0.799	1.015	0.810	1.015
200		0.787	0.988	0.799	1.002	0.787	0.987	0.784	1.004
					$\theta = .4$				
50	144	1.597	2.019	0.897	1.137	1.426	1.807	1.492	1.892
100		1.584	1.956	0.601	0.756	1.308	1.704	1.393	1.728
150		1.642	2.072	0.491	0.617	1.367	1.734	1.383	1.764
200		1.660	2.087	0.453	0.577	1.426	1.816	1.346	1.697
50	196	1.442	1.851	0.837	1.069	1.336	1.703	1.361	1.742
100		1.368	1.708	0.550	0.686	1.260	1.624	1.261	1.568
150		1.387	1.766	0.454	0.566	1.235	1.560	1.220	1.560
200		1.475	1.861	0.428	0.535	1.228	1.525	1.264	1.584

Note: $B(\hat{\beta})$ is scaled bias of $\hat{\beta}$ that equals the difference between the IFE estimator $\hat{\beta}$ in Bai [2009] and its true value β_0 multiplied by $\sqrt{NT}$. $B(\hat{\beta}_{gls})$ and $B(\hat{\beta}^*_{hac})$ are the scaled biases for the GLS estimator in Bai and Liao [2017] and the TA-SHAC estimator with similar interpretation. We use the true distance measure for $\tilde{\beta}^*_{hac}$. RMSE is the corresponding root mean square error multiplied by $\sqrt{NT}$ for each estimator.

Table 2: 95% empirical coverage rates of different estimators

				TA-SHAC (d_{ik}^T)			TA-SHAC (d_{ik}^D)		
T	N	$\hat{\beta}$	$\hat{\beta}_{gls}$	$\tilde{\beta}_{hac1}$	$\tilde{\beta}_{hac2}$	$\tilde{\beta}^*_{hac}$	$\hat{\beta}_{hac1}$	$\hat{\beta}_{hac2}$	$\hat{\beta}^*_{hac}$
					$\theta = 0$				
50	144	0.922	0.905	0.923	0.923	0.924	0.924	0.922	0.927
100		0.923	0.915	0.928	0.922	0.925	0.949	0.949	0.948
150		0.946	0.937	0.943	0.947	0.945	0.935	0.934	0.934
200		0.941	0.951	0.950	0.950	0.951	0.939	0.940	0.938
50	196	0.934	0.916	0.935	0.935	0.934	0.952	0.952	0.950
100		0.932	0.921	0.929	0.933	0.930	0.946	0.945	0.946
150		0.942	0.934	0.945	0.942	0.945	0.945	0.944	0.944
200		0.952	0.950	0.952	0.953	0.954	0.937	0.936	0.936
					$\theta = .4$				
50	144	0.771	0.969	0.829	0.824	0.864	0.817	0.796	0.849
100		0.800	0.902	0.834	0.851	0.867	0.821	0.843	0.879
150		0.777	0.797	0.806	0.854	0.878	0.796	0.841	0.860
200		0.754	0.734	0.772	0.821	0.854	0.786	0.853	0.879
50	196	0.809	0.972	0.846	0.837	0.877	0.843	0.835	0.868
100		0.855	0.911	0.866	0.881	0.898	0.874	0.885	0.902
150		0.842	0.784	0.876	0.890	0.908	0.857	0.880	0.894
200		0.823	0.678	0.872	0.902	0.921	0.847	0.896	0.911

Note: $\hat{\beta}$ is the IFE estimator in Bai [2009] and $\hat{\beta}_{gls}$ is the GLS estimator in Bai and Liao [2017]. $\hat{\beta}_{hac1}$, $\hat{\beta}_{hac2}$, and $\hat{\beta}^*_{hac}$ are TA-SHAC estimators using the data driven distance. For $\hat{\beta}_{hac1}$, we estimate covariance matrix only by TA-SHAC without bias correction. For $\hat{\beta}_{hac2}$, we correct the bias only by TA-SHAC and use the conventional covariance matrix estimator. We both correct the bias and estimate the covariance matrix by TA-SHAC for $\hat{\beta}^*_{hac}$. $\tilde{\beta}_{hac1}$, $\tilde{\beta}_{hac2}$, and $\tilde{\beta}^*_{hac}$ are TA-SHAC estimators using the true distance.

7 Empirical Application

In this section, we use two empirical examples to illustrate the application of the proposed method. The first example is the well-known problem of the U.S. divorce rate that was affected by divorce law reforms in 1970. The second example studies the effects of clean water and effective sewerage systems on child mortality in the U.S.

7.1 Effects of divorce law reforms

During and after the 1970s, most states in the U.S. shifted from a consent divorce regime to no-fault unilateral divorce laws. The new laws allowed people to seek a divorce without the consent of their spouse. Economists have great interest in analyzing the causal relationships between divorce rates and divorce law reforms. Earlier studies include Allen [1992], Peters [1986] suggested that divorce rates were unaffected by the change of the laws, which seems to conflict with the practitioners.

Alternative results are presented in Friedberg [1998]. After controlling for fixed state and year effects, she found that states' law reforms have contributed to about one-sixth of the rise in state-level divorce rates in the first eight years following reforms. However, it is still unclear about the longer effects of the states' law reforms. That is the effects of law reforms on the divorce rates after nine to fourteen years in most states. Wolfers [2006] studied a fixed effects panel data model as following

$$\begin{aligned} y_{st} &= T_{st} + v_s t + u_{st}, \\ u_{st} &= \delta_s + \alpha_t + \varepsilon_{st}, \end{aligned} \tag{26}$$

where y_{st} is the annual number of new divorces per thousand people in state s at time t, T_{it} is the treatment effect of divorce law reform, and $v_s t$ is the time trend. u_{it} captures the unobserved heterogeneity, in which δ_s and α_t are the state and the time fixed effects. ε_{st} is the idiosyncratic errors. The treatment effects T_{st} is

$$\begin{aligned} T_{st} =& \mathbf{1}_{T_s \le t \le T_s+1}\beta_1 + \mathbf{1}_{T_s+2 \le t \le T_s+3}\beta_2 \\ &+ \cdots + \mathbf{1}_{T_s+12 \le t \le T_s+13}\beta_7 + \mathbf{1}_{T_s+14 \le t}\beta_8, \end{aligned} \tag{27}$$

where $\mathbf{1}_A$ is an indicator variable taking value one if the logical condition A is true and T_s is the law reform year of state s. Based on this model, Wolfers [2006] found that the divorce rate rose sharply in the first eight years after the divorce laws reform and identified negative effects for the subsequent nine to fourteen years.

The robustness of Wolfers [2006] has been doubted, due to the additive structure in u_{st} is not flexible enough to capture factors varying across time and state. Since the state-level data he used consisting aggregates, the unobserved heterogeneity can be affected by a number of omitting social and cultural factors (e.g. the stigma of divorce; religious belief; etc.), which are evolving over time and we do not have data or appropriate proxy variables. Besides, cross-sectional correlations may exist in the idiosyncratic errors, since the state-level data includes all available cross-sectional units rather than random samples. To address this problem, Kim and Oka [2013] apply the IFE model for the study, which can effectively control the heterogeneity and cross-sectional correlations through a factor structure. In the model, u_{st} is expressed as

$$u_{st} = \lambda_s' F_t + \varepsilon_{st}. \tag{28}$$

The common factors F_t correspond to the principal components of u_{st}, which dominant the portion of divorce rates not explained by the included regressors. The loading vector λ_s stands for the heterogeneous effect of F_t to each state. If we let $\lambda_s = (1, \delta_s)'$ and $F_t = (\alpha_t, 1)'$, then u_{it} in (26) and (28) are the same. Hence, the state and time fixed effects can be regarded as a special case of interactive fixed effects.

To estimate the treatment effects $(\beta_1, \cdots, \beta_8)$ in (27), Kim and Oka [2013] adopted the estimation and bias correction procedure in Bai [2009], which does not take the cross-sectional correction in the idiosyncratic errors into account. Besides, they estimated the standard errors by the conventional estimator in (15), which also does not valid under cross-sectional correlated errors. Their results confirm the significant effects of the first eight years of law reform on the divorce rates, while the effects after eight years and beyond are insignificant. However, we argue that estimating the treatment effects without bias correction and robust covariance matrix may lead to incorrect confidence intervals and possibly incorrect conclusions.

To correct the bias and provide valid inference, we apply the proposed method to the model of Kim and Oka [2013]. We use the same data as in Wolfers [2006] and Kim and Oka [2013], which contains the divorces rates, state-level reform years, and binary regressors from 1956

to 1988 over 48 states. We choose the same number of factors as Kim and Oka [2013]. For the TA-SHAC estimator $\hat{\beta}^*_{hac}$, we employ the data-driven distance measure and choose the bandwidth parameters by using the bootstrap-based bandwidth selection procedure in section 5. In addition, we apply the GLS method proposed by Bai and Liao [2017] to the study for comparison.

In Table 3, we report the effects of divorce law reform from different estimators with the log of divorce rates as a dependent variable. The results show that both the TA-SHAC estimator $\hat{\beta}^*_{hac}$ and the GLS estimator $\hat{\beta}_{gls}$ produce smaller estimates than the IFE estimator $\hat{\beta}$ due to the bias correction. Such bias, even the size is small, can have a large impact on the inference of the IFE estimator as we discussed before. All of the estimators confirm that the law reforms significantly contribute to the increase of the divorce rates for the first six years after the reforms. However, both $\hat{\beta}^*_{hac}$ and $\hat{\beta}_{gls}$ show that the effects of the law reforms on the divorce rates for 7-8 years after the reforms are insignificant, which are different with $\hat{\beta}$. Furthermore, $\hat{\beta}_{gls}$ generates narrower confidence intervals than $\hat{\beta}^*_{hac}$ and $\hat{\beta}$, since it is more efficient than other estimators. But the confidence intervals generated by $\hat{\beta}_{gls}$ may not robust due to the low EPCs showed in our simulation. In contrast, $\hat{\beta}^*_{hac}$ has wider confidence intervals than the other estimators, which is valid and robust. Overall, the proposed method can correct the bias of the IFE estimator and provide robust inference for the estimates.

Table 3: Methods comparison in effects of divorce law reform: real data

	$\hat{\beta}$		$\hat{\beta}^*_{hac}$		$\hat{\beta}_{gls}$	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
First 2 years	0.0183**	[0.003, 0.034]	0.0175*	[-0.002, 0.037]	0.0138**	[0.000, 0.027]
3–4 years	0.0418***	[0.020, 0.064]	0.0402***	[0.016, 0.065]	0.0340***	[0.014, 0.054]
5–6 years	0.0322**	[0.004, 0.060]	0.0297**	[0.004, 0.057]	0.0249**	[0.000, 0.050]
7–8 years	0.0293*	[-0.005, 0.063]	0.0266	[-0.005, 0.061]	0.0152	[-0.015, 0.045]
9–10 years	0.0073	[-0.032, 0.047]	0.0039	[-0.034, 0.046]	-0.0061	[-0.040, 0.028]
11–12 years	0.0092	[-0.037, 0.051]	0.0055	[-0.038, 0.051]	-0.0078	[-0.044, 0.028]
13–14 years	0.0050	[-0.041, 0.051]	0.0008	[-0.048, 0.052]	-0.0092	[-0.048, 0.029]
15 years+	0.0306	[-0.020, 0.081]	0.0264	[-0.027, 0.084]	0.0093	[-0.033, 0.052]

Note: 95 % confidence intervals are reported. The number of factors $r = 10$.
* $p < .1$. ** $p < .05$. *** $p < .01$.

7.2 Effects of water and sewerage interventions

An important and interesting question in public health is the cause of the sharp decrease in the U.S. and Massachusetts infant mortality from 1870 to 1930. The association between public health interventions and infant mortality has been explored by an extensive literature. Among the interventions, some researchers have focused on the roles of purer water in early twentieth-century cities in the U.S., since clean water interventions make water safe for consumption and washing. One of the best identified research is Cutler and Miller [2005]. They studied the impact of water chlorination and filtration on the death rate from waterborne diseases across 13 U.S. cities. Their results suggest that improved water quality decreases 47 percent in log infant mortality from 1900 to 1936.

On the other hand, effective sewerage systems were installed across many U.S. metropolitan areas, which should also respond to the decline of child mortality. By removing excrement from drinking water sources and reducing human contact with feces, sewerage reduces the fecal-oral transmission of pathogens. So the combined effects of clean water and sewerage interventions on infant mortality have been an interesting question. Alsan and Goldin [2019] exploited the independent and combined effects of clean water and effective sewerage systems on under-5 mortality in Massachusetts, 1880-1920. They use the data from the states that pioneered the collection of U.S. vital statistics. Their data are annual and include 60 municipalities in Massachusetts. For empirical strategy, they employed a fixed effects panel data model, and identified the two interventions together account for approximately one-third of the decline in log child mortality during the 41 years. Specifically, they estimate

$$\begin{aligned} y_{it} &= \mu + \beta_1 W_{it} + \beta_2 S_{it} + \beta_3 (W * S)_{it} + \theta X_{it} + u_{it}, \\ u_{it} &= \delta_i + \alpha_t + \delta_i t + \varepsilon_{it}, \end{aligned} \tag{29}$$

where i is a municipality and t is the year; y_{it} is the log under-5 mortality rate; W_{it} and S_{it} are indicator variables that equal to one if a municipality had adopted the safe water and/or sewerage intervention by year t; X_{it} is a vector of time- and municipality-varying demographic

controls. u_{it} captures the unobserved heterogeneity, which includes municipality and time fixed effects and municipality-specific time trends $\delta_i t$. ε_{it} is the idiosyncratic errors. The standard errors are clustered at the municipality-level with 60 clusters in their analysis. Since they used the municipality-level data, the potential unobserved heterogeneities and cross-sectional correlation in the idiosyncratic errors may affect the results. To check the robustness of their results, we first apply the IFE model for the study. That is, we express u_{it} as

$$u_{it} = \lambda_i' F_t + \varepsilon_{it}, \tag{30}$$

where F_t is a vector of factors that dominant the portion of child mortality rates not explained by the included regressors, and the loading vector λ_i represents the heterogeneous effect of F_t to each municipality. Note that if we let $\lambda_i = (\delta_i, 1, \delta_i)'$ and $F_t = (1, \alpha_t, t)'$, then u_{it} in (29) and (30) are the same. Hence, we use three factors in the model to include the original model as a special case.

Then, we apply the proposed method to correct the bias and estimate the covariance matrix of the IFE estimator. We use the same data as in Alsan and Goldin [2019], which contains the under-5 mortality rate, municipality-level water and/or sewerage interventions years, and demographic control regressors from 1981 to 1920 over 60 municipalities. To construct a balanced panel, we drop the data of Westwood and interpolate the data of Wellesley in 1980 and 1981 with its data in 1982. We also interpolate the missing values of under-5 child mortality in Weston 1904 with its average value in 1903 and 1905, and 1917 with its the average value in 1916 and 1918. We employ the data-driven distance measure and the bootstrap-based bandwidth selection procedure for the TA-SHAC estimators in the estimation. In addition, we apply the GLS method for the study to compare with our method.

We report our results in Table 4. In Panel A, we use the standard fixed effects model as the original paper with the balanced panel data we constructed. The results of Panel A and the original paper are similar, so the results of the original paper are not sensitive to the data we adjusted. Panel B shows the results of the IFE model with the same adjusted data. Compare the

results in Panel A and Panel B, we can see that the independent and combined effects of clean water and effective sewerage system on under-5 mortality in Panel B are much smaller than Panel A. For example, the combination of sewerage and safe water treatments lowered under-5 mortality by 26.6 log points in Panel A, while it decreased to 13.9 log points in Panel B. This reason is that the IFE model can more effectively control the heterogeneities and cross-sectional correlation in the data than the standard fixed effects model, which leads to more conservative results.

Panel C presents the estimated results by the proposed method. Compare the results in Panel B and Panel C, we can see that the estimation effects in Panel C are smaller than Panel B due to the bias correction. Also, the estimates of the independent and combined effects of safe water and sewerage interventions change from statistically significant in Panel B to statistically insignificant in Panel C. Besides, the estimates in Panel C have wider confidence intervals, which is valid and robust as we showed in the simulation. The estimates by the GLS method in Panel D have smaller estimation effects and narrower confidence intervals than the other estimators. But our simulation shows that the corresponding confidence intervals may not reliable. Therefore, by applying the proposed method, we can both correct the bias and provide valid and robust inference for the estimates.

Table 4: Estimated effects of clean water and sewerage on child mortality

	Panel A. Standard Fixed Effects				
	(1)	(2)	(3)	(4)	(5)
Safe water	-0.127		-0.102		0.108
	[-0.280, 0.026]		[-0.252, 0.047]		[-0.043, 0.258]
Sewerage		-0.124***	-0.106**		-0.068
		[-0.214, -0.033]	[-0.194, -0.018]		[-0.156, 0.021]
Interaction				-0.239***	-0.307***
				[-0.395 -0.084]	[-0.509, -0.106]
	Panel B. Interactive Fixed Effects				
	(1)	(2)	(3)	(4)	(5)
Safe water	-0.060***		-0.051**		0.126***
	[-0.103, -0.017]		[-0.096, -0.006]		[0.055, 0.197]
Sewerage		-0.052***	-0.042**		-0.003
		[-0.092, -0.013]	[-0.085, 0.001]		[-0.045, 0.044]
Interaction				-0.151***	-0.262***
				[-0.198, -0.104]	[-0.346, -0.177]
	Panel C. TA-SHAC Estimation				
	(1)	(2)	(3)	(4)	(5)
Safe water	-0.056		-0.048		0.119**
	[-0.126, 0.012]		[-0.120, 0.022]		[0.013, 0.225]
Sewerage		-0.049*	-0.039		-0.003
		[-0.107, 0.009]	[-0.100, 0.022]		[-0.068, 0.062]
Interaction				-0.147***	-0.252***
				[-0.218, -0.076]	[-0.376, -0.128]
	Panel D. GLS Estimation				
	(1)	(2)	(3)	(4)	(5)
Safe water	-0.021		-0.020		0.116***
	[-0.074, 0.033]		[-0.075, 0.034]		[0.028, 0.205]
Sewerage		-0.024	-0.023		0.006
		[-0.071, 0.023]	[-0.072, 0.025]		[-0.044, 0.058]
Interaction				-0.100***	-0.205***
				[-0.159, -0.040]	[-0.310, -0.101]

Note: 95 % confidence intervals are reported. Interaction: interaction of safe water and sewerage. We use three number of factors, which includes the standard fixed effects model in the original paper as a special case.

* $p < .1$. ** $p < .05$. *** $p < .01$.

8 Conclusion

This paper studies the estimation and inference of the panel data model with interactive fixed effects. Under both large N and large T, Bai [2009] has shown that the IFE estimator is $\sqrt{NT}$ consistent, but asymptotic bias exists with correlations and heteroskedasticities in both dimensions. We propose an improved inference procedure for the IFE estimator in the presence of cross-sectional dependence and heteroskedasticity by two parts, correcting the asymptotic bias and estimating the covariance matrix. To implement our approach, we develop a data-driven distance that does not rely on prior information and bandwidth selection procedure based on the bootstrap method.

References

D. Allen. Marriage and divorce: comment. *American Economic Review*, 82:679–685, 1992.

M. Alsan and C. Goldin. Watersheds in Child Mortality: The Role of Effective Water and Sewerage Infrastructure, 1880–1920. *Journal of Political Economy*, 127(2):586–638, 2019.

J. D. Angrist and J.-S. Pischke. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics. *Discussion Paper Series No. 4800, IZA*, 2010.

J. Bai. Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71(1):135–171, 2003.

J. Bai. Panel Data Models With Interactive Fixed Effects. *Econometrica*, 77(4):1229–1279, 2009.

J. Bai and Y. Liao. Inferences in panel data with interactive effects using large covariance matrices. *Journal of Econometrics*, 200(1):59–78, 2017.

J. Bai and S. Ng. Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions. *Econometrica*, 74(4):1133–1150, 2006.

J. Bai, S. H. Choi, and Y. Liao. Standard errors for panel data models with unknown clusters. *Journal of Econometrics*, 2020.

G. Chamberlain. Analysis of Convariance With Qualitative Data. *Review of Economic Studies*, (1232):225–238, 1980.

T. Conley. Econometric modelling of cross sectional dependence. *Ph.D. Thesis. University of Chicago, Dept. of Economics*, 1996.

T. Conley. GMM estimation with cross sectional dependence. *Journal of Econometrics*, 92(1): 1–45, 1999.

T. G. Conley and F. Molinari. Spatial correlation robust inference with errors in location or distance. *Journal of Econometrics*, 140(1):76–96, 2007.

G. Cui, M. Norkute, V. Sarafidis, and T. Yamagata. Two-Stage Instrumental Variable Estimation of Linear Panel Data Models with Interactive Effects. *SSRN Electronic Journal*, 2020.

D. Cutler and G. Miller. The role of public health improvements in health advances: The twentieth-century United States. *Demography*, 42(1):1–22, 2005.

J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 2013.

V. Fernandez. Spatial linkages in international financial markets. *Quantitative Finance*, 11(2): 237–245, 2011.

L. Friedberg. Did Unilateral Divorce Raise Divorce Rates? Evidence from Panel Data. *American Economic Review*, 88:608–627, 1998.

S. Gonçalves. The Moving Blocks Bootstrap For Panel Linear Regression Models With Individual Fixed Effects. *Econometric Theory*, 27(5):1048–1082, 2011.

S. Gonçalves and B. Perron. Bootstrapping factor models with cross sectional dependence. *Journal of Econometrics*, 218(2):476–495, 2020.

J. Hidalgo and M. Schafgans. Inference and testing breaks in large dynamic panels with strong cross sectional dependence. *Journal of Econometrics*, 196(2):259–274, 2017.

H. H. Kelejian and I. R. Prucha. HAC estimation in a spatial framework. *Journal of Econometrics*, 140(1):131–154, 2007.

D. Kim and T. Oka. Divorce Law Reforms And Divorce Rates In The Usa: An Interactive Fixed-Effects Approach. *Journal of Applied Econometrics*, 29(2):231–245, 2013.

M. S. Kim. Robust Inference for Diffusion-Index Forecasts With Cross-Sectionally Dependent Data. *Journal of Business Economic Statistics*, pages 1–15, 2021.

M. S. Kim and Y. Sun. Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix. *Journal of Econometrics*, 160(2):349–371, 2011.

M. S. Kim and Y. Sun. Heteroskedasticity and spatiotemporal dependence robust inference for linear panel models with fixed effects. *Journal of Econometrics*, 177(1):85–108, 2013.

M. S. Kim, Y. Sun, and J. Yang. A fixed-bandwidth view of the pre-asymptotic inference for kernel smoothing with time series data. *Journal of Econometrics*, 197(2):298–322, 2017.

E. A. Ligon and T. G. Conley. Economic Distance and Cross-Country Spillovers. *SSRN Electronic Journal*, 2001.

R. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B*, 11(1):193–197, 1999.

H. R. Moon and M. Weidner. DYNAMIC LINEAR PANEL REGRESSION MODELS WITH INTERACTIVE FIXED EFFECTS. *Econometric Theory*, 33(1):158–195, 2015.

W. K. Newey and K. D. West. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703, 1987.

J. Neyman and E. L. Scott. Consistent Estimates Based on Partially Consistent Observations. *Econometrica*, 16(1):1, 1948.

S. Nickell. Biases in Dynamic Models with Fixed Effects. *Econometrica*, 49(6):1417, 1981.

M. H. Pesaran and E. Tosetti. Large panels with common factors and spatial correlation. *Journal of Econometrics*, 161(2):182–202, 2011.

H. Peters. Marriage and divorce: informational constraints and private contracting. *American Economic Review*, 76:437–454, 1986.

J. Pinkse, M. E. Slade, and C. Brett. Spatial Price Competition: A Semiparametric Approach. *Econometrica*, 70(3):1111–1153, 2002.

D. Politis. On nonparametric function estimation with infinite-order flat-top kernels. *Probability and Statistical Model with Applications*, pages 469–483, 2001.

D. N. Politis. Higher-Order Accurate, Positive Semi-definite Estimation of Large-Sample Covariance and Spectral Density Matrices. *Econometric Theory*, 27(4):703–744, 2011.

P. Robinson. Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics*, 165(1):5–19, 2011.

T. J. Vogelsang. Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects. *Journal of Econometrics*, 166(2):303–319, 2012.

J. Wolfers. Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results. *American Economic Review*, 96(5):1802–1820, 2006.

Appendix: Proofs

We use the following facts throughout: $T^{-1}\|X_i\|^2 = T^{-1}\sum_{t=1}^{T}\|X_{it}\|^2 = O_p(1)$ or $T^{-1/2}\|X_i\| = O_p(1)$. Averaging over i, $(TN)^{-1}\sum_{i=1}^{N}\|X_i\|^2 = O_p(1)$. Similarly, $T^{-1/2}\|F^0\| = O_p(1), T^{-1}\|\hat{F}\|^2 = r, T^{-1/2}\|\hat{F}\| = \sqrt{r}, T^{-1}\|X_i'F^0\| = O_p(1)$ and so forth. Throughout, we define $\delta_{NT} = \min[\sqrt{N}, \sqrt{T}]$ so that $\delta_{NT}^2 = \min[N,T]$. Note that $\hat{J}_{NT} - J_{NT} = o_p(1)$ holds if and only if $A'\hat{J}_{NT}A$ - $A'J_{NT}A$ for any $A \in \mathcal{R}^p$. Therefore, without loss of generality, we assume $\hat{J}_{NT}$ is a scalar, i.e., $p = 1$.

Proof of Theorem 1

(a) Asymptotic Bias:

$$E(\tilde{J}_{NT}) - J_{NT} = O\left(\frac{1}{d_n^q}\right).$$

Note that

$$\begin{aligned}
&E(\tilde{J}_{NT}) - J_{NT} \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} E(w_i\varepsilon_{it}\varepsilon_{kt}\lambda_k)K\left(\frac{d_{ik}}{d_n}\right) - \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} w_i\lambda_k\left(E\varepsilon_{it}\varepsilon_{kt}\right) \\
&= -\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} w_i\lambda_k E\left(\varepsilon_{it}\varepsilon_{kt}\right)\left[1 - K\left(\frac{d_{ik}}{d_n}\right)\right] \\
&\leq -\frac{1}{d_n^q}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\|w_i\|\|\lambda_k\|\|\Gamma_{ik,t}\|d_{ik}^q\right)\left[\frac{1 - K\left(\frac{d_{ik}}{d_n}\right)}{\left(\frac{d_{ik}}{d_n}\right)^q}\right] \\
&\leq -\frac{K_q}{d_n^q}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\|\Gamma_{ik,t}\|d_{ik}^q\right) + o(1) \\
&= O\left(\frac{1}{d_n^q}\right), \text{ as } N,T,d_n \to \infty,
\end{aligned}$$

where $w_i = \text{plim}\left[\frac{(X_i - V_i)'F^0}{T}\right]\left(\frac{F^{0\prime}F^0}{T}\right)^{-1}\left(\frac{\Lambda'\Lambda}{N}\right)^{-1}$ is a constant and we assume λ_k is deterministic instead of a random variable. We use the Assumption B7 in the last equation.

(b) Asymptotic Variance:

$$\tilde{J}_{NT} - E(\tilde{J}_{NT}) = O_p\left(\sqrt{\frac{\ell_N}{NT}}\right) = o_p(1).$$

We want to show that $\tilde{J}_{NT} - E(\tilde{J}_{NT}) = o_p(1)$. By definition, it is equivalent to show that for any $\Delta > 0$,

$$P(|\tilde{J}_{NT} - E(\tilde{J}_{NT})| > \Delta) \to 0.$$

By Chebyshev's inequality, we need to show that

$$\begin{aligned} &P(|\tilde{J}_{NT} - E(\tilde{J}_{NT})| > \Delta) \\ &\quad \leq \frac{1}{\Delta^2} E[\tilde{J}_{NT} - E(\tilde{J}_{NT})]^2 \to 0. \end{aligned}$$

We note that

$$\begin{aligned} &E[\tilde{J}_{NT} - E(\tilde{J}_{NT})]^2 \\ &= E\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} K\left(\frac{d_{ik}}{d_n}\right) w_i\lambda_k\varepsilon_{it}\varepsilon_{kt} - \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} K\left(\frac{d_{ik}}{d_n}\right) E(w_i\lambda_k\varepsilon_{it}\varepsilon_{kt})\right]^2 \\ &= E\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} K\left(\frac{d_{ik}}{d_n}\right) w_i\lambda_k\left(\varepsilon_{it}\varepsilon_{kt} - E\varepsilon_{it}\varepsilon_{kt}\right)\right]^2 \\ &= \frac{1}{N^2T^2}\sum_{i,k=1}^{N}\sum_{a,b=1}^{N}\sum_{s,t=1}^{T} K(\frac{d_{ik}}{d_n})K(\frac{d_{ab}}{d_n})(w_i\lambda_k)(w_a\lambda_b)E\Big[(\varepsilon_{it}\varepsilon_{kt} - E\varepsilon_{it}\varepsilon_{kt})(\varepsilon_{as}\varepsilon_{bs} - E\varepsilon_{as}\varepsilon_{bs})\Big] \\ &= \frac{1}{N^2T^2}\sum_{i,k=1}^{N}\sum_{a,b=1}^{N}\sum_{s,t=1}^{T} K(\frac{d_{ik}}{d_n})K(\frac{d_{ab}}{d_n})(w_i\lambda_k)(w_a\lambda_b) \\ &\quad \times E\Big[\varepsilon_{it}\varepsilon_{kt}\varepsilon_{as}\varepsilon_{bs} - \varepsilon_{it}\varepsilon_{kt}E(\varepsilon_{as}\varepsilon_{bs}) - \varepsilon_{as}\varepsilon_{bs}E(\varepsilon_{it}\varepsilon_{kt}) + E(\varepsilon_{it}\varepsilon_{kt})E(\varepsilon_{as}\varepsilon_{bs})\Big] \\ &= \frac{1}{N^2T^2}\sum_{i,k=1}^{N}\sum_{a,b=1}^{N}\sum_{s,t=1}^{T} K(\frac{d_{ik}}{d_n})K(\frac{d_{ab}}{d_n})(w_i\lambda_k)(w_a\lambda_b)\Big[E\varepsilon_{it}\varepsilon_{kt}\varepsilon_{as}\varepsilon_{bs} - E(\varepsilon_{it}\varepsilon_{kt})E(\varepsilon_{as}\varepsilon_{bs})\Big] \\ &= \frac{1}{N^2T^2}\sum_{i,k=1}^{N}\sum_{a,b=1}^{N}\sum_{s,t=1}^{T} K(\frac{d_{ik}}{d_n})K(\frac{d_{ab}}{d_n})(w_i\lambda_k)(w_a\lambda_b)\Big\{\Big[E\varepsilon_{it}\varepsilon_{kt}\varepsilon_{as}\varepsilon_{bs} - E(\varepsilon_{it}\varepsilon_{kt})E(\varepsilon_{as}\varepsilon_{bs}) \\ &\quad - E(\varepsilon_{it}\varepsilon_{as})E(\varepsilon_{bs}\varepsilon_{kt}) - E(\varepsilon_{it}\varepsilon_{bs})E(\varepsilon_{as}\varepsilon_{kt})\Big] + E(\varepsilon_{it}\varepsilon_{as})E(\varepsilon_{bs}\varepsilon_{kt}) + E(\varepsilon_{it}\varepsilon_{bs})E(\varepsilon_{as}\varepsilon_{kt})\Big\} \\ &= A_1 + A_2 + A_3. \end{aligned}$$

For A_1, we use the linear representation of ε_{it} to have

$$
\begin{aligned}
&E\varepsilon_{it}\varepsilon_{kt}\varepsilon_{as}\varepsilon_{bs} - E(\varepsilon_{it}\varepsilon_{kt})E(\varepsilon_{as}\varepsilon_{bs}) - E(\varepsilon_{it}\varepsilon_{as})E(\varepsilon_{bs}\varepsilon_{kt}) - E(\varepsilon_{it}\varepsilon_{bs})E(\varepsilon_{as}\varepsilon_{kt}) \\
&= \sum_{\ell=1}^{\infty} \gamma_{it,\ell}\gamma_{kt,\ell}\gamma_{as,\ell}\gamma_{bs,\ell}(Ee_\ell^4 - 3).
\end{aligned}
$$

Thus, under Assumption B4 and B5

$$
\begin{aligned}
NT|A_1| &\leq \frac{1}{NT}\sum_{i,k=1}^{N}\sum_{a,b=1}^{N}\sum_{s,t=1}^{T}\sum_{\ell=1}^{\infty} K(\frac{d_{ik}}{d_n})K(\frac{d_{ab}}{d_n})|(w_i\lambda_k)(w_a\lambda_b)||\gamma_{it,\ell}\gamma_{kt,\ell}\gamma_{as,\ell}\gamma_{bs,\ell}||Ee_\ell^4 - 3| \\
&\leq \frac{|M-3|}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} \underbrace{\left(\sum_{\ell=1}^{\infty}|\gamma_{it,\ell}|\right)}_{\leq M} \underbrace{\left(\sum_{k=1}^{N}|\gamma_{kt,\ell}|\right)}_{\leq M} \underbrace{\left(\sum_{s=1}^{T}\sum_{a=1}^{N}|\gamma_{as,\ell}|\right)}_{\leq M} \underbrace{\left(\sum_{b=1}^{N}|\gamma_{bs,\ell}|\right)}_{\leq M} \\
&= O(1).
\end{aligned}
$$

For A_2,

$$
\begin{aligned}
\frac{NT}{\ell_N}|A_2| &\leq \frac{1}{\ell_N NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{a=1}^{N}\sum_{s=1}^{T} \sum_{k\in\{d_{ik}\leq d_n\}} \sum_{b\in\{d_{ab}\leq d_n\}} |(w_i\lambda_k)(w_a\lambda_b)||E(\varepsilon_{it}\varepsilon_{as})||E(\varepsilon_{kt}\varepsilon_{bs})| \\
&\leq \frac{1}{\ell_N NT}\sum_{t=1}^{T}\sum_{i=1}^{N} \sum_{k\in\{d_{ik}\leq d_n\}} \left(\sum_{\ell=1}^{\infty}|\gamma_{it,\ell}|\right)\left(\sum_{a=1}^{N}\sum_{s=1}^{T}|\gamma_{as,\ell}|\right)\left(\sum_{f=1}^{\infty}|\gamma_{kt,f}|\right)\left(\sum_{b=1}^{N}|\gamma_{bs,f}|\right) \\
&= O(1).
\end{aligned}
$$

Using the same argument, we can have

$$
\frac{NT}{\ell_N}|A_3| = O(1).
$$

Combine the results above, we have

$$
E[\tilde{J}_{NT} - E(\tilde{J}_{NT})]^2 = O\left(\frac{1}{NT}\right) + O\left(\frac{\ell_N}{NT}\right),
$$

which implies

$$\tilde{J}_{NT} - E(\tilde{J}_{NT}) = O_p\left(\sqrt{\frac{\ell_N}{NT}}\right) = o_p(1).$$

(c) Estimation Error:

$$\hat{J}_{NT} - \tilde{J}_{NT} = o_p(1).$$

Note that

$$\begin{aligned}
\hat{J}_{NT} - \tilde{J}_{NT} &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}(\hat{w}_i\hat{\lambda}_k\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} - w_i\lambda_k\varepsilon_{it}\varepsilon_{kt})\right]K(\frac{d_{ik}}{d_n}) \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{(X_i - \hat{V}_i)'\hat{F}}{T}\left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\right)^{-1}\hat{\lambda}_k\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\right. \\
&\quad \left. - \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{(X_i - V_i)'F^0}{T}\left(\frac{F^{0\prime}F^0}{T}\right)^{-1}\left(\frac{\Lambda'\Lambda}{N}\right)^{-1}\lambda_k\varepsilon_{it}\varepsilon_{kt}\right]K(\frac{d_{ik}}{d_n}).
\end{aligned}$$

We shall prove

$$\begin{aligned}
B_1 = \frac{1}{T}\sum_{t=1}^{T}&\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{X_i'\hat{F}}{T}\left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\right)^{-1}\hat{\lambda}_k\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\right. \\
&\left. - \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{X_i'F^0}{T}\left(\frac{F^{0\prime}F^0}{T}\right)^{-1}\left(\frac{\Lambda'\Lambda}{N}\right)^{-1}\lambda_k\varepsilon_{it}\varepsilon_{kt}\right]K(\frac{d_{ik}}{d_n}) = o_p(1).
\end{aligned}$$

and

$$\begin{aligned}
B_2 = \frac{1}{T}\sum_{t=1}^{T}&\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{\hat{V}_i'\hat{F}}{T}\left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\right)^{-1}\hat{\lambda}_k\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\right. \\
&\left. - \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{V_i'F^0}{T}\left(\frac{F^{0\prime}F^0}{T}\right)^{-1}\left(\frac{\Lambda'\Lambda}{N}\right)^{-1}\lambda_k\varepsilon_{it}\varepsilon_{kt}\right]K(\frac{d_{ik}}{d_n}) = o_p(1).
\end{aligned}$$

Consider B_1. There are four items being estimated, namely F^0, $\Lambda'\Lambda/N$, λ_k, and $\varepsilon_{it}\varepsilon_{kt}$. Using the identity $\hat{a}\hat{b}\hat{c}\hat{d} - abcd = (\hat{a}-a)\hat{b}\hat{c}\hat{d} + a(\hat{b}-b)\hat{c}\hat{d} + ab(\hat{c}-$ c$)\,\hat{d} + abc(\hat{d}-d)$, the first

corresponding term is

$$\left\| \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{N} \frac{X_i' \left(\hat{F} - F^0 H \right)}{T} \left(\frac{\hat{\Lambda}' \hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k \hat{\varepsilon}_{it} \hat{\varepsilon}_{kt} \right\|$$
$$\leq \frac{\left\| \hat{F} - F^0 H \right\|}{\sqrt{T}} \left(\frac{1}{T} \sum_{t=1}^{T} \left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} X_i \hat{\varepsilon}_{it} \right\| \left\| \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left(\frac{\hat{\Lambda}' \hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k \hat{\varepsilon}_{kt} \right\| \right).$$

Since

$$\hat{\varepsilon}_{it} = \varepsilon_{it} + X_{it}(\hat{\beta} - \beta) + \left(\hat{F}_t - H' F_t^0 \right)' H^{-1} \lambda_i + \hat{F}_t' \left(\hat{\lambda}_i - H^{-1} \lambda_i \right),$$

we first look at

$$\left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} X_i \hat{\varepsilon}_{it} \right\|$$
$$= \left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} X_i \left(\varepsilon_{it} + X_{it}(\hat{\beta} - \beta) + \left(\hat{F}_t - H' F_t \right) H^{-1} \lambda_i + \hat{F}_t' \left(\hat{\lambda}_i - H^{-1} \lambda_i \right) \right) \right\|$$
$$\leq \left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} X_i \varepsilon_{it} \right\| + \left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} X_i X_{it} (\hat{\beta} - \beta) \right\|$$
$$+ \left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} X_i \left(\hat{F}_t - H' F_t^0 \right)' H^{-1} \lambda_i \right\| + \left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} X_i \hat{F}_t' \left(\hat{\lambda}_i - H^{-1} \lambda_i \right) \right\|$$
$$= B_{11} + B_{12} + B_{13} + B_{14}.$$

For B_{11},

$$\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{X_i}{\sqrt{T}} \varepsilon_{it} \right\| = \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} \left(\frac{\| X_i \|^2}{T} \right) \varepsilon_{it} \varepsilon_{kt} \right)^{1/2}$$
$$= O_p(1).$$

For B_{12}, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} X_i X_{it} (\hat{\beta} - \beta) \right\| \\
&\leq \sqrt{N} \left(\frac{1}{NT} \sum_{i=1}^{N} \|X_i\|^2 \right)^{1/2} \left(\frac{1}{N} \sum_{i=1}^{N} \|X_{it}\|^2 \right)^{1/2} \|\hat{\beta} - \beta\| \\
&= \sqrt{N} O_p(\|\hat{\beta} - \beta\|) = O_p(1).
\end{aligned}$$

For B_{13},

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} X_i \left(\hat{F}_t - H' F_t^0 \right)' H^{-1} \lambda_i \right\| \\
&\leq \sqrt{N} \left\| \hat{F}_t - H' F_t^0 \right\| \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\|X_i\|}{\sqrt{T}} \left\| H^{-1} \lambda_i \right\| \right) \\
&= \sqrt{N} \left\| \hat{F}_t - H' F_t^0 \right\| O_p(1).
\end{aligned}$$

For B_{14},

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} X_i \hat{F}_t' \left(\hat{\lambda}_i - H^{-1} \lambda_i \right) \right\| \\
&\leq \sqrt{N} \|\hat{F}_t\| \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\|X_i\|}{\sqrt{T}} \left\| \left(\hat{\lambda}_i - H^{-1} \lambda_i \right) \right\| \right) \\
&= \sqrt{N} \|\hat{F}_t\| \left(O_p(\|\hat{\beta} - \beta\|) + O_p(\delta_{NT}^{-1}) \right) = \|\hat{F}_t\| O_p(1).
\end{aligned}$$

For the last equality, we use the Lemma A.10 (ii) in Bai (2009) that

$$\frac{1}{N} \sum_{i=1}^{N} \left\| \hat{\lambda}_i - H^{-1} \lambda_i \right\| = O_p\left(\delta_{NT}^{-1} \right) + O_p(\|\hat{\beta} - \beta\|).$$

In summary, we have

$$\left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} X_i \hat{\varepsilon}_{it} \right\| \leq \sqrt{N} \left\| \hat{F}_t - H' F_t^0 \right\| O_p(1) + \|\hat{F}_t\| O_p(1).$$

We next consider

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k \hat{\varepsilon}_{kt} \right\| \\
&\leq \left\| \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k \varepsilon_{kt} \right\| + \left\| \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k X_{kt} (\hat{\beta} - \beta) \right\| \\
&\quad + \left\| \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k \left(\hat{F}_t - H' F_t^0 \right)' H^{-1} \lambda_k \right\| \\
&\quad + \left\| \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k \hat{F}_t' \left(\hat{\lambda}_k - H^{-1} \lambda_k \right) \right\| \\
&= C_{11} + C_{12} + C_{13} + C_{14}.
\end{aligned}$$

For C_{11},

$$\left\| \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k \varepsilon_{kt} \right\| = O_p(1)$$

For C_{12}, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k X_{kt} (\hat{\beta} - \beta) \right\| \\
&\leq \sqrt{N} \left\| \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \right\| \left(\frac{1}{N} \sum_{k=1}^{N} \|\hat{\lambda}_k\|^2 \right)^{1/2} \left(\frac{1}{N} \sum_{k=1}^{N} \|X_{kt}\|^2 \right)^{1/2} \|\hat{\beta} - \beta\| \\
&= \sqrt{N} O_p(\|\hat{\beta} - \beta\|) = O_p(1).
\end{aligned}$$

For C_{13},

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k \left(\hat{F}_t - H' F_t^0 \right)' H^{-1} \lambda_k \right\| \\
&\leq \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left\| \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \right\| \left\| \hat{\lambda}_k \right\| \left\| (\hat{F}_t - H' F_t^0) \right\| \left\| H^{-1} \lambda_k \right\| \\
&= \sqrt{N} \left\| (\hat{F}_t - H' F_t^0) \right\| O_p(1).
\end{aligned}$$

For C_{14},

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k \hat{F}_t' \left(\hat{\lambda}_k - H^{-1} \lambda_k \right) \right\| \\
&\leq \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left\| \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \right\| \left\| \hat{\lambda}_k \right\| \| \hat{F}_t \| \left\| \hat{\lambda}_k - H^{-1} \lambda_k \right\| \\
&= \| \hat{F}_t \| \sqrt{N} \left[O_p(\| \hat{\beta} - \beta \|) + O_p(\delta_{NT}^{-1}) \right] = \| \hat{F}_t \| O_p(1).
\end{aligned}$$

In summary, we have

$$\left\| \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k \hat{\varepsilon}_{kt} \right\| \leq \sqrt{N} \left\| (\hat{F}_t - H' F_t^0) \right\| O_p(1) + \| \hat{F}_t \| O_p(1).$$

Therefore, we the first corresponding term is

$$\begin{aligned}
&\left\| \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{N} \frac{X_i' \left(\hat{F} - F^0 H \right)}{T} \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k \hat{\varepsilon}_{it} \hat{\varepsilon}_{kt} \right\| \\
&\leq \frac{\left\| \hat{F} - F^0 H \right\|}{\sqrt{T}} \left(\frac{1}{T} \sum_{t=1}^{T} \left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} X_i \hat{\varepsilon}_{it} \right\| \left\| \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} \hat{\lambda}_k \hat{\varepsilon}_{kt} \right\| \right) \\
&\leq \frac{\left\| \hat{F} - F^0 H \right\|}{\sqrt{T}} \frac{1}{T} \sum_{t=1}^{T} \left(\sqrt{N} \left\| \hat{F}_t - H' F_t^0 \right\| O_p(1) + \| \hat{F}_t \| O_p(1) \right)^2 \\
&= O_p\left(\delta_{NT}^{-1} \right) + O_p(\| \hat{\beta} - \beta \|) = o_p(1).
\end{aligned}$$

For the last equality, we use the proposition A.1 (ii) in Bai (2009) that

$$\frac{1}{\sqrt{T}}\left\|\hat{F}-F^0H\right\| = O_p(\|\hat{\beta}-\beta\|)+O_p(\delta_{NT}^{-1}).$$

The second corresponding term is

$$\begin{aligned}
&\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{X_i'F^0}{T}\left[\left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\right)^{-1}-H'\left(\frac{\Lambda'\Lambda}{N}\right)^{-1}H\right]\hat{\lambda}_k\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\\
&=\frac{1}{T}\sum_{t=1}^{T}\left[\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{X_i'F^0}{T}\hat{\varepsilon}_{it}\right)\left[\left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\right)^{-1}-H'\left(\frac{\Lambda'\Lambda}{N}\right)^{-1}H\right]\left(\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\hat{\lambda}_k\hat{\varepsilon}_{kt}\right)\right],
\end{aligned}$$

where the term HH' arises in the interim and $HH'-(F^{0\prime}F^0/T)^{-1}=O_p(\delta_{NT}^{-1})$ by Lemma A.7 in Bai (2009). Let $Q=\left\|\left(\hat{\Lambda}'\hat{\Lambda}/N\right)^{-1}-H'\left(\Lambda'\Lambda/N\right)^{-1}H\right\|$ and $Q=O_p(\|\hat{\beta}-\beta\|)+O_p\left(\delta_{NT}^{-2}\right)=O_p(\delta_{NT}^{-1})$ by Lemma A.10 (iv) in Bai (2009). Then we have

$$\begin{aligned}
&\left\|\frac{1}{T}\sum_{t=1}^{T}\left[\left(\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\hat{\lambda}_k\hat{\varepsilon}_{kt}\right)\otimes\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{X_i'F^0}{T}\hat{\varepsilon}_{it}\right)\right]\mathrm{vec}(Q)\right\|\\
&\leq\left\|\frac{1}{T}\sum_{t=1}^{T}\left[\left(\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\hat{\lambda}_k\hat{\varepsilon}_{kt}\right)\otimes\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{X_i'F^0}{T}\hat{\varepsilon}_{it}\right)\right]\right\|\mathrm{vec}(Q)\\
&=O_p(\|\hat{\beta}-\beta\|)+O_p\left(\delta_{NT}^{-2}\right)=O_p(\delta_{NT}^{-1}).
\end{aligned}$$

since $\|X_i'F^0/T\|=O_p(1)$.

The third corresponding term is given by

$$\begin{aligned}
&\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{X_i'F^0}{T}\left(\frac{\Lambda'\Lambda}{N}\right)^{-1}\left(\hat{\lambda}_k-H^{-1}\lambda_k\right)\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\\
&=\frac{1}{T}\sum_{t=1}^{T}\left[\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{X_i'F^0}{T}\right)\left(\frac{\Lambda'\Lambda}{N}\right)^{-1}\hat{\varepsilon}_{it}\right)\left(\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\left(\hat{\lambda}_k-H^{-1}\lambda_k\right)\hat{\varepsilon}_{kt}\right)\right].
\end{aligned}$$

Let $A_i = \left(X_i' F^0/T\right)\left(\Lambda'\Lambda/N\right)^{-1}$. Then, we have

$$\left\| \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} A_i\hat{\varepsilon}_{it}\right)\left(\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\left(\hat{\lambda}_k - H^{-1}\lambda_k\right)\hat{\varepsilon}_{kt}\right)\right\| = o_p(1),$$

using the fact that $\|A_i\| = O_p(1)$ and

$$\begin{aligned}&\left\|\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\left(\hat{\lambda}_k - H^{-1}\lambda_k\right)\hat{\varepsilon}_{kt}\right\| \\ &\leq \left(\frac{1}{N}\sum_{k=1}^{N}\left\|\hat{\lambda}_k - H^{-1}\lambda_k\right\|^2\hat{\varepsilon}_{kt}^2\right)^{1/2} = o_p(1).\end{aligned}$$

It is easy to show that the last corresponding term is equal to $o_p(1)$ since

$$\frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} - \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{it}\varepsilon_{kt} = o_p(1).$$

In summary, B_1 is equal to $O_p\left(\delta_{NT}^{-1}\right) = o_p(1)$. Next, consider B_2. The only difference between B_1 and B_2 is X_i replaced by $\hat{V}_i$. Let $G_k = \left(F^{0\prime}F^0/T\right)^{-1}\left(\Lambda'\Lambda/N\right)^{-1}\lambda_k$. Then, $\|G_k\| = O_p(1)$. Thus it is sufficient to prove

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\frac{\left(\hat{V}_i - V_i\right)' F^0}{T}G_k\varepsilon_{it}\varepsilon_{kt} = o_p(1).$$

Since

$$\begin{aligned}&\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\frac{\left(\hat{V}_i - V_i\right)' F^0}{T}G_k\varepsilon_{it}\varepsilon_{kt}\right\| \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\left(\hat{V}_i - V_i\right)\varepsilon_{it}\right\|\left\|\frac{1}{\sqrt{N}}\sum_{k=1}^{N}G_k\varepsilon_{kt}\right\|\frac{\|F^0\|}{\sqrt{T}},\end{aligned}$$

and, $\hat{V}_i - V_i = \frac{1}{N}\sum_{k=1}^{N}(\hat{a}_{ik} - a_{ik})X_k$, where

$$\begin{aligned}\hat{a}_{ik} - a_{ik} = & \left(\hat{\lambda}_i - H^{-1}\lambda_i\right)' \left(\hat{\Lambda}'\hat{\Lambda}/N\right)^{-1} \hat{\lambda}_k \\ & + \lambda_i' H'^{-1} \left[\left(\hat{\Lambda}'\hat{\Lambda}/N\right)^{-1} - H' \left(\Lambda'\Lambda/N\right)^{-1} H\right] \hat{\lambda}_k \\ & + \lambda_i' \left(\Lambda'\Lambda/N\right)^{-1} H \left(\hat{\lambda}_k - H^{-1}\lambda_k\right).\end{aligned}$$

We have

$$\begin{aligned}& \left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \left(\hat{V}_i - V_i\right) \varepsilon_{it} \right\| \\ & \leq \left\| \frac{1}{N\sqrt{NT}} \sum_{i=1}^{N}\sum_{k=1}^{N} \left(\hat{\lambda}_i - H^{-1}\lambda_i\right)' \left(\hat{\Lambda}'\hat{\Lambda}/N\right)^{-1} \hat{\lambda}_k X_k \varepsilon_{it} \right\| \\ & \quad + \left\| \frac{1}{N\sqrt{NT}} \sum_{i=1}^{N}\sum_{k=1}^{N} \lambda_i' H'^{-1} \left[\left(\hat{\Lambda}'\hat{\Lambda}/N\right)^{-1} - H' \left(\Lambda'\Lambda/N\right)^{-1} H\right] \hat{\lambda}_k X_k \varepsilon_{it} \right\| \\ & \quad + \left\| \frac{1}{N\sqrt{NT}} \sum_{i=1}^{N}\sum_{k=1}^{N} \lambda_i' \left(\Lambda'\Lambda/N\right)^{-1} H \left(\hat{\lambda}_k - H^{-1}\lambda_k\right) X_k \varepsilon_{it} \right\| \\ & = D_1 + D_2 + D_3.\end{aligned}$$

We first consider,

$$\begin{aligned}& \left\| \frac{1}{N\sqrt{NT}} \sum_{i=1}^{N}\sum_{k=1}^{N} \left(\hat{\lambda}_i - H^{-1}\lambda_i\right)' \left(\hat{\Lambda}'\hat{\Lambda}/N\right)^{-1} \hat{\lambda}_k X_k \varepsilon_{it} \right\| \\ & \quad \leq \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\hat{\lambda}_i - H^{-1}\lambda_i\right) \varepsilon_{it} \right\| \left\| \hat{\Lambda}'\hat{\Lambda}/N \right\|^{-1} \left(\frac{1}{N} \sum_{k=1}^{N} \|\lambda_k\| \left\| \frac{X_k}{\sqrt{T}} \right\| \right) \\ & \quad = o_p(1).\end{aligned}$$

Next,

$$\begin{aligned}& \left\| \frac{1}{N\sqrt{NT}} \sum_{i=1}^{N}\sum_{k=1}^{N} \lambda_i' H'^{-1} \left[\left(\hat{\Lambda}'\hat{\Lambda}/N\right)^{-1} - H' \left(\Lambda'\Lambda/N\right)^{-1} H\right] \hat{\lambda}_k X_k \varepsilon_{it} \right\| \\ & \leq \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \lambda_i \varepsilon_{it} \right\| \left\| H^{-1} \right\| \left\| \left(\hat{\Lambda}'\hat{\Lambda}/N\right)^{-1} - H' \left(\Lambda'\Lambda/N\right)^{-1} H \right\| \left(\frac{1}{N} \sum_{k=1}^{N} \|\lambda_k\| \left\| \frac{X_k}{\sqrt{T}} \right\| \right) \\ & = O_p\left(\delta_{NT}^{-2}\right) + O_p(\|\hat{\beta} - \beta\|).\end{aligned}$$

Finally,

$$\begin{aligned}
&\left\| \frac{1}{N\sqrt{NT}} \sum_{i=1}^{N} \sum_{k=1}^{N} \lambda_i' \left(\Lambda'\Lambda/N\right)^{-1} H \left(\hat{\lambda}_k - H^{-1}\lambda_k\right) X_k \varepsilon_{it} \right\| \\
&\leq \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \lambda_i \varepsilon_{it} \right\| \left\| \left(\hat{\Lambda}'\hat{\Lambda}/N\right)^{-1} \right\| \|H\| \left(\frac{1}{N} \sum_{k=1}^{N} \left\| \hat{\lambda}_k - H^{-1}\lambda_k \right\| \left\| \frac{X_k}{\sqrt{T}} \right\| \right) \\
&= O_p\left(\delta_{NT}^{-1}\right) + O_p(\|\hat{\beta} - \beta\|).
\end{aligned}$$

In summary, we have

$$\left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \left(\hat{V}_i - V_i\right) \varepsilon_{it} \right\| = O_p\left(\delta_{NT}^{-1}\right) + O_p(\|\hat{\beta} - \beta\|).$$

Thus, B_2 is equal to $O_p\left(\delta_{NT}^{-1}\right) + O_p(\|\hat{\beta} - \beta\|) = o_p(1)$. Combining B_1 and B_2, we have $\hat{J}_{NT} - \tilde{J}_{NT} = o_p(1)$.

Proof of Theorem 2

Recall

$$H_Z = \operatorname{plim} \frac{1}{NT} \sum_{i=1}^{N} \sum_{k=1}^{N} \sum_{t=1}^{T} E(\varepsilon_{it}\varepsilon_{kt}) Z_{it} Z_{kt}'.$$

Define

$$H_{NT} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{k=1}^{N} \sum_{t=1}^{T} E(\varepsilon_{it}\varepsilon_{kt}) E(Z_{it} Z_{kt}').$$

then $H_Z = \operatorname{plim} H_{NT}$ and the TA-SHAC estimator for H_{NT} is given by

$$\hat{H}_{NT} = \frac{1}{T} \sum_{t=1}^{T} \hat{H}_t \text{ with } \hat{H}_t = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} \hat{Z}_{it} \hat{Z}_{kt}' \hat{\varepsilon}_{it} \hat{\varepsilon}_{kt} K\left(\frac{d_{ik}}{d_n}\right).$$

To establish the consistency of $\hat{H}_{NT}$, we define the infeasible version of $\hat{H}_{NT}$ as

$$\tilde{H}_{NT} = \frac{1}{T}\sum_{t=1}^{T}\tilde{H}_t \text{ with } \hat{H}_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} Z_{it}Z_{kt}'\varepsilon_{it}\varepsilon_{kt}K\left(\frac{d_{ik}}{d_n}\right),$$

which is identical to $\hat{H}_{NT}$ but is based on the true value of Z_{it} and ε_{it}. Using $\tilde{H}_{NT}$, the difference between $\hat{H}_{NT}$ and H_Z can be decomposed into three parts:

$$\hat{H}_{NT} - H_{NT} = (\hat{H}_{NT} - \tilde{H}_{NT}) + (\tilde{H}_{NT} - E\tilde{H}_{NT}) + (E\tilde{H}_{NT} - H_{NT}).$$

The first term is due to the effect of estimation errors in the factor model. The second and third terms represent the variance and bias of the infeasible estimator $\tilde{H}_{NT}$. Note that $\hat{H}_{NT} - H_Z = o_p(1)$ holds if and only if $A'\hat{H}_{NT}A$ - $A'H_ZA$ for any $A \in \mathcal{R}^p$. Therefore, without loss of generality, we assume $\hat{H}_Z$ is a scalar, i.e., $p = 1$.

(a) Asymptotic Bias:

$$E(\tilde{H}_{NT}) - H_{NT} = O\left(\frac{1}{d_n^q}\right).$$

Note that

$$\begin{aligned}
&E(\tilde{H}_{NT}) - H_Z \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} E(Z_{it}Z_{kt}\varepsilon_{it}\varepsilon_{kt})K\left(\frac{d_{ik}}{d_n}\right) - \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} E(Z_{it}Z_{kt})E(\varepsilon_{it}\varepsilon_{kt}) \\
&= -\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} E(Z_{it}Z_{kt})E\left(\varepsilon_{it}\varepsilon_{kt}\right)\left[1 - K\left(\frac{d_{ik}}{d_n}\right)\right] \\
&\leq -\frac{1}{d_n^q}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\|\Gamma_{ik,t}\|d_{ik}^q\right)\left[\frac{1 - K\left(\frac{d_{ik}}{d_n}\right)}{\left(\frac{d_{ik}}{d_n}\right)^q}\right] \\
&\leq -\frac{K_q}{d_n^q}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\|\Gamma_{ik,t}\|d_{ik}^q\right) + o(1) \\
&= O\left(\frac{1}{d_n^q}\right), \text{ as } N, T, d_n \to \infty.
\end{aligned}$$

(b) Asymptotic Variance:

$$\tilde{H}_{NT} - E(\tilde{H}_{NT}) = O_p\left(\sqrt{\frac{\ell_N}{NT}}\right) = o_p(1).$$

The proof is similar with $\tilde{J}_{NT} - E(\tilde{J}_{NT})$ we showed before.

(c) Estimation Error:

$$\hat{H}_{NT} - \tilde{H}_{NT} = o_p(1).$$

Note that

$$\begin{aligned}
\hat{H}_{NT} - \tilde{H}_{NT} &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}(\hat{Z}_{it}\hat{Z}_{kt}\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} - Z_{it}Z_{kt}\varepsilon_{it}\varepsilon_{kt})\right]K(\frac{d_{ik}}{d_n}) \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\hat{Z}_{it}\hat{Z}_{kt}(\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} - \varepsilon_{it}\varepsilon_{kt})K(\frac{d_{ik}}{d_n}) \\
&\quad + \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}(\hat{Z}_{it}\hat{Z}_{kt} - Z_{it}Z_{kt})\varepsilon_{it}\varepsilon_{kt}K(\frac{d_{ik}}{d_n}).
\end{aligned}$$

The first term is bounded by

$$\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|Z_{it}\|^4\right)^{1/2}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}(\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} - \varepsilon_{it}\varepsilon_{kt})^2\right)^{1/2},$$

so it is easy to show $\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}(\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} - \varepsilon_{it}\varepsilon_{kt})^2 = o_p(1)$. The second term is $o_p(1)$ that analyzed in Bai (2009). Thus $\hat{H}_{NT} - \tilde{H}_{NT} = o_p(1)$.