Improved Inference for Interactive Fixed Effects Model with Cross Sectional Dependence

Zhenhao Gong *

University of Connecticut

November 5, 2021

Abstract

In this paper, I 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 lease square (LS) estimator in this model by Bai [2009] is asymptotically biased when the error term is cross-sectionally dependent, and I address this problem. My procedure involves two parts: correcting the asymptotic bias of the LS estimator and employing the cross-sectional dependence robust covariance matrix estimator. I prove the validity of the proposed procedure in the asymptotic sense. Since my approach is based on the spatial HAC estimation, e.g., Conley (1999), Kelejian and Prucha (2007), and Kim and Sun (2011), I need a distance measure that characterizes the dependence structure. Such a distance may not be available in practice, and I address this by considering a data-driven distance that does not rely on prior information. I also consider a bandwidth selection procedure based on the cluster wild bootstrap method. Monte Carlo simulations show my procedure work well in finite samples. As empirical illustrations, I apply the proposed approach to study the effects of divorce law reforms on U.S. divorce rates (Wolfers [2006]) and the impacts of clean water and sewerage interventions on U.S. child mortality (Alsan and Goldin [2019]).

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

1 Introduction

It is crucial to properly allow for unobserved heterogeneity that may evolve over time and potentially correlate with regressors in panel regression analysis. For example, in macroeconomics, the heterogeneous impacts of unobserved common shocks on different countries. In applied microeconomics, the demographic structure and social and cultural factors that change over time and affect policy reforms and their corresponding effects. The typical way to address this issue is to use the standard fixed effects model, in which the individual- and time-specific effects enter the model additively. However, one concern of this approach is that it may not be flexible enough to capture the various patterns of heterogeneity. An alternative way to tackle this concern is to employ the interactive fixed effects (IFE) model, in which the individual- and time-specific effects enter the model interactively. The IFE model includes the additive fixed effects model as a special case but is significantly more flexible. Bai [2009] develops the LS estimator for the IFE model using the principal component analysis. This paper considers inference for the IFE model in the presence of cross-sectional dependence.

There has been growing attention on the IFE model in the literature, and the model can apply to various economics disciplines, such as applied microeconomics, asset pricing, and forecasting. In large N but fixed T panels, Holtz-Eakin et al. [1988] explore a quasi-differencing approach, which uses appropriate lagged variables as instruments to estimate the quasi-differenced version of the IFE model; Ahn et al. [2001] propose a generalized method of moments (GMM) estimator and show it is more efficient than the LS estimator. In large N and large T panel models, Pesaran [2006] investigates the common correlated effects (CEE) estimator that controls time-specific effects by using cross-sectional averages of the dependent and independent variables; Bai [2009] develops the LS estimator using the principal component analysis and establish the asymptotics. Moon and Weidner [2017] also consider the LS estimator for the IFE model in the dynamic panel model context.

In this paper, I propose an inference procedure that improves upon Bai [2009]'s method in the presence of cross-sectional dependence and heteroskedasticity. More specifically, I de-

velop a bias correction of the LS estimator and employ the cross-sectional dependence robust covariance matrix estimator. My work is empirically relevant since the cross-sectional correlation is common in panel data. For example, when one uses state-level data, while nationwide cross-sectional correlation can be captured by the factor structure, local correlations among neighbor states may remain in the idiosyncratic errors. Under the large N and T asymptotics, Bai [2009] shows that the LS estimator is asymptotically biased when the idiosyncratic errors are heteroskedastic or correlated in both dimensions. This is an incidental parameters problem (Neyman and Scott, 1948; Nickell, 1981) in the IFE model, which is crucial because failure to control it can lead to a misleading inference. I assume there is no serial dependence in order to focus on issues caused by cross-sectional dependence.

My procedure involves two parts: correcting the incidental parameters bias due to cross-sectional dependence and employing the cross-sectional dependence robust covariance matrix estimator. Using the fact that the bias forms a cross-sectional version of the long-run covariance structure, I develop the estimator of this bias based on the time-series average of spatial heteroskedasticity and autocorrelation (TA-SHAC) estimators, which is first proposed by Conley [1996, 1999]. This approach is further studied by Kelejian and Prucha [2007] and Kim and Sun [2011]. I also propose an estimator for the covariance matrix of the model parameters using the spatial HAC estimation method.

My work complements the IFE model literature. Regarding the inference of this model under cross-sectional dependence, Bai [2009] proposes a partial sample approach. However, this approach may be tough to implement in practice because the partial sample should be selected to replicate the dependence structure of the whole cross-sample, which is often infeasible. Another approach that yields valid inference in this setting is the GLS method by Bai and Liao [2017]. Their GLS transformation eliminates cross-sectional correlation, so the estimator becomes asymptotically centered at the actual value. While the GLS approach is attractive because it is efficient and incidental parameters bias-free, its inference is not as stable as our procedure in finite samples. My simulation shows that the GLS inference often produces substantial size distortions.

There are two challenges to implementing my procedure and I employ simple approaches to overcome those challenges. The first challenge is how to choose a proper distance measure to construct the TA-SHAC estimators. Since my bias estimator and covariance matrix estimator are constructed based on the spatial HAC estimation method, I need a distance measure that characterizes the dependence structure of data. A typical way in the literature is to find an auxiliary variable that captures the decaying pattern of dependence in data (e.g., the transportation cost in Ligon and Conley [2001]; the geographic distance in Pinkse et al. [2002]). However, such a variable may not be available in some applications. To address this issue, I propose a data-driven distance that reflects the cross-sectional dependence structure directly. See, for example, Mantegna [1999], Fernandez [2011], Cui et al. [2020] and Kim [2021]. An advantage of this approach is that I do not need prior information about the dependence structure for implementation.

The second challenge is how to select the bandwidth parameters. This is particularly challenging in my setting because I need to choose two bandwidth parameters jointly in estimating the asymptotic bias and the covariance matrix. I consider a bootstrap-based bandwidth selection procedure. To replicate the cross-sectional dependence, I follow Hidalgo and Schafgans [2017] and Kim [2021] to employ a cluster wild bootstrap approach, in which each cluster contains all cross-sectional units in one time period. Using the bootstrap, I choose the bandwidths that control the size properly in finite samples.

My procedure can be applied to the broad empirical literature in economics. I illustrate this with two empirical examples. The first one is the well-known problem of the U.S. divorce rates affected by divorce law reforms around the 1970s. Using the standard fixed-effects model with weighted least squares (WLS) estimation, Wolfers [2006] identifies the rise of divorce rates in the first eight years after the law reform. However, the robustness of Wolfers [2006]'s results is doubted in two regards. First, the model he uses may not be flexible enough to capture the factors varying over time and across states (e.g., the stigma of divorce; religious belief). This may lead to the observed large discrepancy between the ordinary least squares (OLS) and WLS estimates found by Droes and Lamoen [2010] and Lee and Solon [2011]. Second, the

idiosyncratic errors in his model are assumed to be cross-sectionally independent, which does not seem to be appropriate in practice. Kim and Oka [2013] employ the IFE model for the study. Their results confirm the findings of Wolfers [2006] and are robust to the weighting schemes. However, their bias correction procedure and standard error estimation do not take the cross-sectional dependence into account. I apply the proposed approach and provide inference results for this model. I find the IFE model with the proposed procedure yields smaller estimates with wider confidence intervals than Kim and Oka [2013]'s results.

The second example studies the effects of clean water and effective sewerage systems on U.S. child mortality. An essential question in public health is the cause of the sharp decrease in the U.S. and Massachusetts infant mortality from 1870 to 1930. Alsan and Goldin [2019] exploit the independent and combined effects of clean water and effective sewerage systems on under-5 mortality in Massachusetts, 1880-1920. For empirical strategy, they employ a standard fixed-effects model, which identifies the two interventions together account for approximately one-third of the decline in the log of child mortality during the 41 years. Since they use the municipality-level data, the potential unobserved time-varying heterogeneity and cross-sectional correlation in the idiosyncratic errors may affect the results. To check the robustness of their results, I employ the IFE model with the proposed inference procedure for the study. I find that the combined impacts of sewerage and safe water treatments on child mortality are significantly decreased by using the IFE model with the proposed procedure.

The remainder of the paper is as follows. Section 2 reviews the IFE model. Sections 3 and 4 introduces my method and its implementation procedure. Section 5 presents the simulation results. Section 6 applies my method to estimate and make inferences on the effects of divorce law reforms on the U.S. divorce rate and the impacts of clean water and sewerage interventions on U.S. child mortality. The last section concludes. All proofs are given in the Appendix.

2 Review of IFE model

In this section, I review the LS estimator and its asymptotics in Bai [2009]. I consider the following panel model with interactive fixed effects

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

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} is an error term. I assume a 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 error. The number of factors is r and is assumed to be known. X_{it} can be correlated with λ_i or F_t alone, or simultaneously correlated with both of them; ε_{it} is allowed to be weakly correlated in both dimensions. I can rewrite (1) as

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

where $Y_i = (Y_{i1}, \cdots, Y_{iT})'$, $X_i = (X_{i1}, \cdots, X_{iT})'$, $(T \times p)$, $F = (F_1, \cdots, F_T)'$ and $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{iT})'$. Let $\Lambda = (\lambda_1, \cdots, \lambda_N)'$. Given F and $\{\Lambda\}$, the LS estimator for β_0 is given by

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

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. These two restrictions are used to identify factors and loadings. Using the first order condition to concentrate out Λ, we have

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

where $M_F = I_T - F(F'F)^{-1}F'$. For the estimation of F, note that (2) reduces to a pure factor model given β, and we can use the principal components analysis (PCA). More specifically, the LS estimator of F given β is equal to $\sqrt{T}$ times the eigenvectors that are associated with

the r largest eigenvalues of $\sum_{i=1}^{N}(Y_i - X_i\beta)(Y_i - X_i\beta)'$.

Therefore, the final LS estimator $(\hat{\beta}, \hat{F})$ is obtained by solving the following equations simultaneously:

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i' M_{\hat{F}} X_i\right)^{-1} \sum_{i=1}^{N} X_i' M_{\hat{F}} Y_i, \tag{5}$$

and

$$\left[\frac{1}{NT}\sum_{i=1}^{N}(Y_i - X_i\hat{\beta})(Y_i - X_i\hat{\beta})'\right]\hat{F} = \hat{F}V_{NT}, \tag{6}$$

where V_{NT} is a diagonal matrix of the r largest eigenvalues of the matrix in the square bracket. Given $\hat{\beta}$ and $\hat{F}$, we have $\hat{\Lambda} = (Y - X\hat{\beta})'\hat{F}/T$.

Throughout the paper, I 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. I denote F^0 as the true parameter for F that satisfies Assumption A2 below.

Bai [2009] makes the following assumptions to establish the asympotics.

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

$$H(F) = \frac{1}{NT}\sum_{i=1}^{N} X_i' M_F X_i - \frac{1}{T}\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N} X_i' M_F X_k a_{ik}\right], \tag{7}$$

where $a_{ik} = \lambda_i'(\Lambda'\Lambda/N)^{-1}\lambda_k$. I 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) The factor loading matrix Λ is non-random. $||\lambda_i|| \leq C$ and $\frac{1}{N}\Lambda'\Lambda \to \Sigma_\Lambda$ for some $r \times r$ positive definite matrix Σ_Λ, as $N \to \infty$.*

Assumption A3.

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

(ii) $E(\varepsilon_{it}\varepsilon_{ks}) = 0$ for all (i,k) if $t \neq s$. $E(\varepsilon_{it}\varepsilon_{kt}) = \sigma_{ik,t}$, $|\sigma_{ik,t}| < \bar{\sigma}_{ik}$ for all (i,k) and t such that

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

The largest eigenvalue of $\Omega_i = E\varepsilon_i\varepsilon_i'$ is uniformly bounded in i and T.

(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

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

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

Assumption A5. *We have*

$$\begin{aligned} H_Z &= \operatorname{plim}\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\sigma_{ik,ts}Z_{it}Z_{ks}', \\ &\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}Z_i'\varepsilon_i \xrightarrow{d} N\left(0,H_Z\right), \end{aligned} \tag{8}$$

where $Z_i=M_{F^0}X_i-\frac{1}{N}\sum_{k=1}^{N}a_{ik}M_{F^0}X_k$.

Assumption A1 indicates $H(F)$ is positive definite and excludes 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 largest r eigenvalues of the covariance matrix of Y diverge, while the rest are bounded as $N,T\to\infty$. It ensures the consistency of the PCA estimators for F and Λ in the factor model. Assumption A3 states ε_{it} is serially uncorrelated but allows for weak cross-sectional correlation and heteroskedasticity. Assumption A4 rules out the dynamic panel data model. Assumption A5 states a central limit theorem holds for the moment process.

Under Assumptions 1-5, Bai [2009] shows that

$$\sqrt{NT}\left(\hat{\beta}-\beta_0\right)\xrightarrow{d}N\left(\rho^{1/2}B_0,H_0^{-1}H_ZH_0^{-1}\right) \tag{9}$$

as $T/N \to \rho$, where $H_0 = \text{plim} H(F^0)$ and $H(F)$ is defined in (7),

$$\begin{aligned} H_Z &= \text{plim}\, \frac{1}{NT} \sum_{i=1}^{N} \sum_{k=1}^{N} \sum_{t=1}^{T} E(\varepsilon_{it}\varepsilon_{kt}) Z_{it} Z_{kt}', \\ B_0 &= \text{plim}\, B_{NT} \text{ with } B_{NT} = -H \left(F^0\right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} w_i \lambda_k \left(\frac{1}{T} \sum_{t=1}^{T} E\varepsilon_{it}\varepsilon_{kt} \right), \end{aligned} \tag{10}$$

and

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

The formula of B_0 in (12) implies that $B_0 = 0$ and $\hat{\beta}$ is unbiased when cross-sectional correlation and heteroskedasticity are absent. However, if not, $\hat{\beta}$ is asymptotically biased, and inference without correcting this bias causes misleading statistical conclusions.

3 Inference on β

In this section, I propose an inference procedure for β, which is valid under cross-sectional correlation and heteroskedasticity. My approach involves two parts: correcting the bias of the $\hat{\beta}$ and employing the cross-sectional dependence robust covariance matrix estimator for H_Z.

3.1 Correcting the bias

As presented in (9), $\hat{\beta}$ is asymptotically biased in the presence of cross-sectional dependence and heteroskedasticity and it is necessary to estimate B_0 for valid inference. In this regard, Bai [2009] suggests the partial sum estimator, which is given by

$$\hat{B}_{CS} = -\hat{H}_0^{-1} \frac{1}{n_{sub}} \sum_{i=1}^{n_{sub}} \sum_{k=1}^{n_{sub}} \hat{w}_i \hat{\lambda}_k \left(\frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_{it} \hat{\varepsilon}_{kt} \right), \tag{11}$$

where n_{sub} is a sub-sample selected from the whole sample and $\hat{H}_0$ and $\hat{w}_i$ are the estimators of H_0 and w_i, respectively, defined as

$$\begin{aligned}\hat{H}_0 =& \frac{1}{NT}\sum_{i=1}^{N} X_i' M_{\hat{F}} X_i - \frac{1}{T}\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N} X_i' M_{\hat{F}} X_k \hat{a}_{ik}\right], \\ \hat{w}_i =& \left[\frac{\left(X_i - \hat{V}_i\right)'\hat{F}}{T}\right]\left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\right)^{-1},\end{aligned} \tag{12}$$

with $\hat{V}_i = N^{-1}\sum_{k=1}^{N}\hat{a}_{ik}X_k$ and $\hat{a}_{ik} = \hat{\lambda}_i'(\hat{\Lambda}'\hat{\Lambda}/N)^{-1}\hat{\lambda}_k$. Bai shows that $\hat{B}_{CS}$ converges to B_0 as $n_{sub}/\min\{N,T\} \to 0$. A critical issue to implement this approach is how to select partial observations to replicate the cross-sectional dependence structure of the whole sample. This may not be feasible in practice if the dependence structure is unknown. To the best of my knowledge, there is no practical guidance on this selection in the literature.

Let

$$J_{NT} = \frac{1}{T}\sum_{t=1}^{T} J_t \text{ where } J_t = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} w_i\lambda_k E(\varepsilon_{it}\varepsilon_{kt}). \tag{13}$$

Then we can write $B_{NT} = -H\left(F^0\right)^{-1} J_{NT}$. I propose an estimator of J_{NT} based on the spatial HAC estimation approach. Define

$$\hat{J}_{NT} = \frac{1}{T}\sum_{t=1}^{T}\hat{J}_t \text{ with } \hat{J}_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} K\left(\frac{d_{ik}}{d_n^{(1)}}\right)\hat{w}_i\hat{\lambda}_k\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}, \tag{14}$$

where $K(\cdot)$ is a real-valued kernel function, $d_n^{(1)}$ is a bandwidth parameter, and d_{ik} is a distance between units i and k that reflects the strength of their cross-sectional dependence. Note that $\hat{J}_t$ has a form of the spatial HAC estimator (e.g., Kelejian and Prucha, 2007; Kim and Sun, 2011) and $\hat{J}_{NT}$ can be viewed as its time series average. Based on $\hat{J}_{NT}$, the estimator of B_0 is

$$\hat{B}_{NT} = -H(\hat{F})^{-1}\hat{J}_{NT}, \tag{15}$$

and the bias corrected estimator of β is given by

$$\hat{\beta}^{\dagger} = \hat{\beta} - \frac{1}{N}\hat{B}_{NT}. \tag{16}$$

3.2 Robust covariance estimation

I also propose an estimator for the covariance matrix H_Z in (12), which is consistent in the presence of cross-sectional dependence and heteroskedasticity. H_Z is conventionally estimated with

$$\hat{H}_Z = \frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}_i^2\left(\frac{1}{T}\sum_{t=1}^{T}\hat{Z}_{it}\hat{Z}_{it}'\right), \tag{17}$$

where $\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_{it}^2$. However, $\hat{H}_Z$ is constructed based on the assumption of cross-sectional independence, so it is not valid when ε_{it} are cross-sectional correlated. In this regard, Bai [2009] proposes a partial sample method, which is given by

$$\hat{H}_{CS} = \frac{1}{n_{sub}}\sum_{i=1}^{n_{sub}}\sum_{k=1}^{n_{sub}}\left(\frac{1}{T}\sum_{t=1}^{T}\hat{Z}_{it}\hat{Z}_{kt}'\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\right), \tag{18}$$

and establishes its consistency as $n_{sub}/\min\{N,T\} \to 0$.

As discussed in Section 3.1, a practical issue in implementing a partial sample estimator is that it is hard to construct a partial sample to replicate the overall cross-sectional dependence structure. In fact, we find that we do not need to rely on a partial sample to estimate H_Z. That is, it can be estimated with

$$\tilde{H}_{CS} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\hat{Z}_{it}\hat{Z}_{kt}'\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\right). \tag{19}$$

Since a distance measure is available in our setting, I propose an estimator of H_Z using the spatial HAC estimation method. My estimator is given by

$$\hat{H}_{NT} = \frac{1}{T}\sum_{t=1}^{T}\hat{H}_t \text{ with } \hat{H}_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\hat{Z}_{it}\hat{Z}_{kt}'\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\mathfrak{K}_F\left(\frac{d_{ik}}{d_n^{(2)}}\right), \tag{20}$$

where $d_n^{(2)}$ is a bandwidth parameter. As my bias estimator in Section 3.1, $\hat{H}_{NT}$ has a form of time average of spatial HAC estimators $\{\hat{H}_t\}$. Note that the proposed estimator $\hat{H}_{NT}$ includes $\tilde{H}_{CS}$ as a special case by choosing $d_n^{(2)}$ large enough if $\mathfrak{K}_F(\cdot)$ is a rectangular kernel.

3.3 Asymptotic properties

This section establishes the consistency conditions for the TA-SHAC estimators $\hat{J}_{NT}$ and $\hat{H}_{NT}$. I start from the assumptions on the distance measure and kernels 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, I can generalize the results of this paper to the case when d_{ik} is contaminated by measurement errors. In this paper, however, I 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, I assume that ε_{it} has a linear representation

$$\varepsilon_{it} = \sum_{\ell=1}^{\infty} \gamma_{it,\ell} e_\ell, \tag{21}$$

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, I 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 $\hat{J}_{NT}$, I define the infeasible version of $\hat{J}_{NT}$ as

$$\tilde{J}_{NT} = \frac{1}{T}\sum_{t=1}^{T}\tilde{J}_t \text{ with } \tilde{J}_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} K\left(\frac{d_{ik}}{d_n^{(1)}}\right) w_i\lambda_k\varepsilon_{it}\varepsilon_{kt}, \tag{22}$$

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:

$$\hat{J}_{NT} - J_{NT} = (\hat{J}_{NT} - \tilde{J}_{NT}) + (\tilde{J}_{NT} - E\tilde{J}_{NT}) + (E\tilde{J}_{NT} - J_{NT}). \tag{23}$$

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 I assume that $e_{\ell i}$ is independent of $e_{\ell k}$ for $i \neq k$. Under this assumption, J_{NT} in (13) can be expressed as

$$J_{NT} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} w_i\lambda_k\gamma_{it,\ell}\gamma_{kt,\ell}. \tag{24}$$

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 (21) 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. I introduce this assumption to control the variance of $\tilde{J}_{NT}$.

Let

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

where ℓ_i is the number of pseudo-neighbors that unit i have within the bandwidth, and ℓ_n is the average number of pseudo-neighbors. The number of pseudo-neighbors is increased with the bandwidth I choose.

Assumption B6. $\ell_i \leq c_\ell \ell_n$ 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 terms have many cross-sectional correlated neighbours while others have none or very few.

The asymptotic bias of $\tilde{J}_{NT}$ is determined by the smoothness of the kernel at zero and the rate of decaying of the spatial dependence . 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*

$$\lim_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} \|\Gamma_{ik,t}\| \, d_{ik}^{q} < M, \text{ with } \Gamma_{ik,t} = E\left(\varepsilon_{it}\varepsilon_{kt}\right). \tag{25}$$

This assumption characterizes the weak dependence between ε_{it} and ε_{kt} with respect to d_{ik}. The equation (25) 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 $K(\cdot)$ assigns to $\varepsilon_{it}\varepsilon_{kt}$ in (14) 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 (23) 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* $d_n, \ell_n, N, T \to \infty$ *such that* $\ell_n/N, \ell_n/T \to 0$ *and* $T/N \to \rho$*, I have* $\hat{J}_{NT} - J_{NT} = o_p(1)$.

Theorem 2. *Under the Assumptions A1-A4 and B1-B7, and* $d_n, \ell_n, N, T \to \infty$ *such that* $\ell_n/N, \ell_n/T \to 0$ *and* $T/N \to \rho$*, I have* $\hat{H}_{NT} - H_Z = o_p(1)$.

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

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

where $\hat{\beta}^{\dagger}$ defines in (16) and $\hat{H}_{NT}$ defines in (20).

4 Implementation

As I introduced before, there are two major challenges for implementing my procedure in practice. The first challenge is how to choose a proper distance measure to construct the TA-SHAC estimators. Since my bias estimator and covariance matrix estimator are constructed by the spatial HAC estimation approach, I need a distance measure that characterizes the dependence structure of data. The literature typically finds a relevant auxiliary variable as the distance, which captures the decaying pattern of dependence in the data (e.g., the transportation cost in Ligon and Conley [2001]; the geographic distance in Pinkse et al. [2002]). However, such a variable may not be available in some applications. To address this issue, I propose a data-driven distance that reflects the cross-sectional 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 I can estimate

it by its sample counterpart

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

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 my 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. I 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 my setting because I need to choose two bandwidth parameters jointly in estimating the asymptotic bias and the covariance matrix. I consider 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 jointly in their test procedure. But I can not apply their method directly since they can generate the time-series dependence of the sample simply by a regular AR model. To replicate the cross-sectional dependence, I follow Hidalgo and Schafgans [2017] and Kim [2021] to employ a cluster wild bootstrap approach, in which each cluster contains all cross-sectional units in one time period. Using the bootstrap, I choose the bandwidths that control the size properly in finite samples.

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 $\hat{B}_{NT}^*(d_{nm}^{(1)})$ with $d_{nm}^{(1)} \in \mathcal{D}_{nM}^{(1)}$ such that

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

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 $\hat{H}_{NT}^*(d_{ns}^{(2)})$ 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 bootstrap version of the bias corrected estimator and $SE(\hat{\beta}^*)$ is the standard error for $\hat{\beta}^*$ such that

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

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 $(d_{nm}^{(1*)}, d_{ns}^{(2*)})$ 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 I 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 the sample in time period t. Hence, $\hat{B}_{NT}^*(d_{nm}^{(1)})$ and $\hat{H}_{NT}^*(d_{ns}^{(2)})$ are expected to be a good approximation to $\hat{B}_{NT}(d_{nm}^{(1)})$ and $\hat{H}_{NT}(d_{ns}^{(2)})$. I generate ξ_t from Rademacher Distribution in my simulation and empirical application.

Based on $(d_{nm}^{(1*)}, d_{ns}^{(2*)})$, 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

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

Thus, my bootstrap-based bandwidth selection procedure is designed to choose $(d_{nm}^{(1*)}, d_{ns}^{(2*)})$ jointly that improves the inference of the LS estimator $\hat{\beta}$ by 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 of the sample in one time period (e.g., the CSD bootstrap by Gonçalves and Perron [2020]). I may also consider a parametric bootstrap-based spatial regression model if the location of each unit is available. The theoretical properties of the cluster wild bootstrap method are not examined in this paper, and I leave it for future research.

5 Monte Carlo Simulation

In this section, I investigate the finite sample performance of the proposed procedure for correcting the bias and improving the inference of the LS estimator $\hat{\beta}$. Follow Bai [2009], the data generating process (DGP) I 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 is two and is assumed to be known. The regressors and factors are generated according to

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

I set $c = \mu = 1$ and $\rho = 0.3$, so there is a weak serial correlation between factors. I 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, I 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 d_{ik}^D does not satisfy the triangular inequality, so it is not a valid distance, I can show that my 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 I 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 $(d_n^{(1)}, d_n^{(2)})$, I apply the bootstrap-based bandwidth selection procedure in section 4. I choose bandwidth parameter sets to be $\mathcal{D}_{nM}^{(1)} = \mathcal{D}_{nS}^{(2)} = \{2 : 10\}$, so I have (9×9) different pairs of $(d_n^{(1)}, d_n^{(2)})$ for the selection. In the simulation, I find that almost all of the selected bandwidth $(d_{nm}^{(1*)}, d_{ns}^{(2*)})$ in step 6 fall in the interior of the bandwidth parameter sets, so I believe $\mathcal{D}_{nM}^{(1)}$ and $\mathcal{D}_{nS}^{(2)}$ are reasonable sets for the selection. For the kernel function, I employ Parzen kernel for estimating the bias $\hat{B}_{NT}(d_n^{(1*)})$, and Bartlett kernel for estimating the covariance matrix $\hat{H}_{NT}(d_n^{(2*)})$.

As I discussed before, another approach that yields valid inference in this setting is the GLS method by Bai and Liao [2017]. They focus on the efficient estimation of β_0. The corresponding GLS estimator is given by

$$\hat{\beta}(\hat{\Sigma}_\varepsilon^{-1}) = \arg\min_{\beta_0} \min_{F,\lambda_i} \sum_{i=1}^N (Y_i - X_i\beta_0 - F\lambda_i)'\hat{\Sigma}_\varepsilon^{-1}(Y_i - X_i\beta_0 - F\lambda_i), \tag{26}$$

where $\hat{\Sigma}_\varepsilon$ is a consistent estimator of the covariance matrix of ε_{it}, which is a high-dimensional matrix. To estimate Σ_ε, they assume it is a sparse covariance matrix and apply the thresholding method in Fan et al. [2013]. Their GLS transformation eliminates cross-sectional correlation,

so the estimator becomes asymptotically centered at the actual value. I apply the GLS method in my simulation for comparison. I choose the tuning parameter recommended by the authors for employing the thresholding method.

In Table 1, I 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 LS 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 using the data-driven distance measure (d^D_{ik}), and TA-SHAC estimator using the true distance measure (d^T_{ik}). The reason why I scaled the bias and RMSE with $\sqrt{NT}$ is that the LS 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 scaled bias 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 scaled 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 (9) depends on $\rho = T/N$. Third, the TA-SHAC estimator using the data-driven distance ($\hat{\beta}^*_{hac}$) have similar performance with the TA-SHAC estimator using the true distance ($\tilde{\beta}^*_{hac}$) 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 measure (d^D_{ik}). For $\hat{\beta}_{hac1}$, I estimate the covariance matrix only by the TA-SHAC estimator $\hat{H}_{NT}$ in (20) without bias correction. For $\hat{\beta}_{hac2}$, I correct the bias only by the

TA-SHAC estimator $\hat{B}_{NT}$ in (15) with the conventional covariance matrix estimator in (17). I use $\hat{\beta}_{hac1}$ and $\hat{\beta}_{hac2}$ to compare which part (bias correction or robust covariance estimation) in my procedure is more important for inproving the inference of $\hat{\beta}$. For $\hat{\beta}^*_{hac}$, I both correct the bias and estimate the covariance matrix by the proposed estimators. For comparison, I use the true distance (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} ($\theta = 0$), the EPCs for all of the 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 valid 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 when N is larger. 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 d^T_{ik} or d^D_{ik} is used, although the one with d^T_{ik} performs slightly better in general. For example, when $N = 144$, $T = 150$, the EPC for $\tilde{\beta}^*_{hac}$ is 0.878, while the EPC for $\hat{\beta}^*_{hac}$ is 0.86. This finding gives us an important implication from an empirical point of view. That is, we can apply my method with d^D_{ik}, 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 presence of weak cross-sectional correlation, the LS 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 stable when T is large. My simulation shows that the GLS inference often produces substantial size distortions. The proposed procedure can correct the bias and provide valid inference using the TA-SHAC estimators with data-driven distance measures in finite samples.

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
					$\theta = .3$				
50	144	1.680	2.112	1.344	1.704	1.536	1.968	1.620	2.040
100		1.152	1.452	0.732	0.912	1.116	1.416	1.092	1.368
150		1.190	1.484	0.647	0.808	1.073	1.367	1.102	1.382
200		1.256	1.578	0.611	0.764	1.171	1.459	1.137	1.442
50	196	1.198	1.495	0.980	1.228	1.119	1.406	1.168	1.455
100		1.092	1.372	0.686	0.854	1.106	1.372	1.036	1.302
150		1.063	1.337	0.583	0.737	1.115	1.389	1.012	1.252
200		1.069	1.366	0.554	0.693	1.089	1.346	1.010	1.267
					$\theta = .4$				
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 LS estimator $\hat{\beta}$ in Bai [2009] and its true value β_0 multiplied by $\sqrt{NT}$. $B(\hat{\beta}_{gls})$, $B(\hat{\beta}^*_{hac})$, and $B(\tilde{\beta}^*_{hac})$ are the scaled biases for the GLS estimator in Bai and Liao [2017], the TA-SHAC estimator using the data-driven distance measure (d_{ik}^D), and the TA-SHAC estimator using the true distance measure (d_{ik}^T). RMSE is the corresponding scaled root mean square error 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
					$\theta = .3$				
50	144	0.844	0.955	0.882	0.874	0.906	0.872	0.864	0.888
100		0.880	0.947	0.879	0.879	0.907	0.893	0.889	0.908
150		0.863	0.881	0.886	0.894	0.913	0.883	0.898	0.912
200		0.841	0.773	0.851	0.876	0.884	0.867	0.887	0.900
50	196	0.862	0.946	0.887	0.882	0.901	0.887	0.872	0.895
100		0.902	0.957	0.898	0.907	0.917	0.910	0.916	0.928
150		0.908	0.872	0.899	0.903	0.913	0.917	0.919	0.930
200		0.910	0.742	0.913	0.923	0.930	0.917	0.931	0.937
					$\theta = .4$				
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 LS 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 measure (d_{ik}^D). For $\hat{\beta}_{hac1}$, I estimate covariance matrix only without bias correction. For $\hat{\beta}_{hac2}$, I correct the bias only and use the conventional covariance matrix estimator in (17). I correct the bias and use the robust covariance matrix in our procedure for $\hat{\beta}^*_{hac}$. $\tilde{\beta}_{hac1}$, $\tilde{\beta}_{hac2}$, and $\tilde{\beta}^*_{hac}$ are TA-SHAC estimators using the true distance measure (d_{ik}^T) with similar interpretation.

6 Empirical Application

In this section, I use two empirical examples to illustrate the application of the proposed procedure. The first one is the well-known problem of the U.S. divorce rates affected by divorce law reforms around the 1970s. The second one studies the effects of clean water and effective sewerage systems on child mortality in the U.S.

6.1 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 are interested in analyzing the causal relationships between the rise of divorce rates and divorce law reforms. Earlier studies include Allen [1992] and Peters [1986]. Using the same cross-section data in 1979, Peters found that divorce rates were unaffected by the switch to the unilateral law, while Allen found a significant impact.

Alternative results are presented in Friedberg [1998]. After controlling for fixed state and year effects, as well as state-specific time trends in her specification, she found that states' law reforms have contributed to about one-sixth of the rise in state-level divorce rates since the late 1960s. Based on her results, she concluded that the effect of unilateral divorce on divorce behavior was permanent. In contrast, Wolfers [2006] found that the divorce rate rose sharply in the first eight years after the divorce laws reform, but that this rise was reversed for the subsequent nine to fourteen years. The model he studied was a standard fixed-effects panel data model as following

$$\begin{aligned} y_{st} &= T_{st} + f(v_s, t) + u_{st}, \\ u_{st} &= \delta_s + \alpha_t + \varepsilon_{st}, \end{aligned} \tag{27}$$

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 $f(v_s, t)$ is the time trend. For example, we have $f(v_s, t) = v_s t$ for the linear trend. u_{it} captures the unobserved heterogeneities, 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

$$\begin{aligned} T_{st} =& \mathbf{1}_{T_s \leq t \leq T_s+1}\beta_1 + \mathbf{1}_{T_s+2 \leq t \leq T_s+3}\beta_2 \\ &+ \cdots + \mathbf{1}_{T_s+12 \leq t \leq T_s+13}\beta_7 + \mathbf{1}_{T_s+14 \leq t}\beta_8, \end{aligned} \tag{28}$$

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.

The robustness of Wolfers [2006] has been doubted in two regards. The first one is the additive structure in u_{st} may not be flexible enough to capture factors varying across time and state. Since the state-level data he used consisting aggregates, the unobserved heterogeneities can be affected by many omitting social and cultural factors (e.g., the stigma of divorce; religious belief). Those factors are evolving, and we do not have data or appropriate proxy variables to capture them. Second, the idiosyncratic errors are assumed to be cross-sectional independent. However, cross-sectional correlations may exist in the error terms since the state-level data includes all available cross-sectional units rather than random samples. To address this issue, Kim and Oka [2013] applied 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

$$u_{st} = \lambda_s' F_t + \varepsilon_{st}. \tag{29}$$

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 (27) and (29) 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 (28), Kim and Oka [2013] adopted the estimation and bias correction procedure in Bai [2009], which does not take the cross-sectional dependence into account. Besides, they estimated the standard errors by the conventional estimator in (17), which also does not valid in the presence of cross-sectional correlation. Their

results confirmed the significant effects of the first eight years of law reforms on the rise of divorce rates, while the effects after eight years and beyond are insignificant. However, I argue that estimating the treatment effects without a proper bias correction procedure and robust covariance matrix estimation may lead to biased estimates and invalid inference.

To correct the cross-sectional correlation bias and provide valid inference, I apply the proposed method to the model of Kim and Oka [2013]. I use the same data as in Kim and Oka [2013], which contains the divorces rates, state-level reform years, and binary regressors from 1956 to 1988 over 48 states. I choose the same number of factors as Kim and Oka [2013]. For the TA-SHAC estimator $\hat{\beta}^*_{hac}$, I employ the data-driven distance measure and choose the bandwidth parameters by using the bootstrap-based bandwidth selection procedure in section 4. In addition, I apply the GLS method proposed by Bai and Liao [2017] to the study for comparison.

In Table 3, I 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 LS estimator $\hat{\beta}$ by taking the cross-sectional correlation bias into account. All three estimators confirm that the law reforms significantly contribute to the rise of the divorce rates for the first six years after the law 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, while $\hat{\beta}$ show that it is significant. They also identify negative effects for the 9-14 years after the law reforms, which is in line with that of Wolfers [2006]. 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 be reliable as I showed in my simulation. In contrast, $\hat{\beta}^*_{hac}$ has wider confidence intervals than the other estimators, which is valid and robust to the cross-sectional dependence. Overall, the proposed procedure can correct the bias for the LS estimator $\hat{\beta}$ and provides valid and 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]	0.0156*	[-0.003, 0.034]	0.0138**	[0.000, 0.027]
3–4 years	0.0418***	[0.020, 0.064]	0.0368***	[0.013, 0.060]	0.0340***	[0.014, 0.054]
5–6 years	0.0322**	[0.004, 0.060]	0.0255**	[-0.001, 0.052]	0.0249**	[0.000, 0.050]
7–8 years	0.0293*	[-0.005, 0.063]	0.0208	[-0.012, 0.054]	0.0152	[-0.015, 0.045]
9–10 years	0.0073	[-0.032, 0.047]	-0.0034	[-0.043, 0.036]	-0.0061	[-0.040, 0.028]
11–12 years	0.0092	[-0.037, 0.051]	-0.0026	[-0.047, 0.041]	-0.0078	[-0.044, 0.028]
13–14 years	0.0050	[-0.041, 0.051]	-0.0079	[-0.057, 0.041]	-0.0092	[-0.048, 0.029]
15 years+	0.0306	[-0.020, 0.081]	0.0170	[-0.038, 0.072]	0.0093	[-0.033, 0.052]

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

6.2 Effects of water and sewerage interventions

An essential question in public health is the cause of the sharp decrease in the U.S. and Massachusetts infant mortality from 1870 to 1930. An extensive literature has explored the association between public health interventions and infant mortality. 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 made 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 suggested that improved water quality decreased 47 percent in log infant mortality from 1900 to 1936.

On the other hand, many U.S. metropolitan areas installed effective sewerage systems during that time, which should also respond to child mortality decline. By removing excrement from drinking water sources and reducing human contact with feces, sewerage reduces the fecal-oral transmission of pathogens. 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. Their data are annual and include 60 municipalities in Massachusetts for a period that predates national mortality statistics. For empirical strategy, they employed

a standard fixed-effects panel data model, which identified the two interventions together account for approximately one-third of the decline in log child mortality during the 41 years. Specifically, they estimate

$$\begin{aligned} y_{it} &= \mu + \beta_1 W_{it} + \beta_2 S_{it} + \beta_3 (W * S)_{it} + \theta X_{it} + u_{it}, \\ u_{it} &= \delta_i + \alpha_t + \delta_i t + \varepsilon_{it}, \end{aligned} \tag{30}$$

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 sewerage interventions by year t, respectively; X_{it} is a vector of time- and municipality-varying demographic controls. u_{it} captures the unobserved heterogeneities, which includes municipality and time fixed-effects and municipality-specific time trends $\delta_i t$. ε_{it} is the idiosyncratic errors. They clustered the standard errors in their analysis at the municipality-level with 60 clusters. Since they used the municipality-level data, the potential unobserved time-varying heterogeneity and cross-sectional correlation in the idiosyncratic errors may affect the results. To check the robustness of their results, I first apply the IFE model for the study. That is, I re-express u_{it} in (30) as

$$u_{it} = \lambda_i' F_t + \varepsilon_{it}, \tag{31}$$

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 (30) and (31) are the same. Hence, I choose three factors in the IFE model to include the original model as a special case.

Then, I apply the proposed procedure to correct the bias and improve the inference of the LS estimates. I use the same data as in Alsan and Goldin [2019], which contains the under-5 mortality rate, municipality-level water, sewerage interventions years, and demographic control regressors from 1981 to 1920 over 60 municipalities. To construct a balanced panel, I drop the data of Westwood since there are many missing observations, and I interpolate the data of

Wellesley in 1980 and 1981 with its data in 1982. I 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 average value in 1916 and 1918. As a result, the interpolation creates a balanced panel that contains 59 municipalities and 41 years. I employ the data-driven distance measure and the bootstrap-based bandwidth selection procedure for the TA-SHAC estimators in the estimation. In addition, I apply the GLS method for the study to compare with our method.

I report my results in Table 4. In Panel A, I use the same model as the original paper with the balanced panel data I constructed. The results of Panel A and the original paper are similar, which implies that the results of the original paper are not sensitive to the data I adjusted. Panel B shows the results of the IFE model with the same adjusted data. Comparing Panel A and Panel B results, 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, while the combination of sewerage and safe water treatments lowered under-5 mortality by 26.6 log points in Panel A, it decreased to 13.9 log points in Panel B. The 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.

Panel C presents the estimated results by applied the proposed procedure. Comparing the results in Panel B and Panel C, we can see that the estimation effects in Panel C are smaller than in Panel B due to the bias correction. Also, some estimates of the independent or combined effects of safe water and sewerage interventions change from statistically significant in Panel B to statistically insignificant in Panel C. The estimates in Panel C have wider confidence intervals than Panel B, which are valid and robust to cross-sectional dependence. The estimated effects by the GLS method show in Panel D. The GLS estimator has smaller estimation effects and narrower confidence intervals than the other estimators. The confidence intervals generated by the GLS method may not be reliable, as I showed in my simulation before. In summary, by applying the proposed method, I can correct the bias and provide valid and robust inferences 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.

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

7 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] showed that the LS estimator is $\sqrt{NT}$ consistent, but asymptotic bias exists in the presence of correlations and heteroskedasticity in both dimensions. I propose an improved inference procedure for the LS estimator in the presence of cross-sectional dependence and heteroskedasticities. My procedure involves two parts: correcting the asymptotic bias of the LS estimator and employing the cross-sectional dependence robust covariance matrix estimator. To implement my procedure, I develop a data-driven distance that does not rely on prior information and a bandwidth selection procedure based on a cluster wild bootstrap method.

References

S. C. Ahn, Y. Hoon Lee, and P. Schmidt. GMM estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics*, 101(2):219–255, 2001.

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. 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.

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.

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.

I. Droes and R. Lamoen. Did unilateral divorce laws raise divorce rates? A reconciliation and new results: comment. . *Working Paper*, 2010.

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 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.

D. Holtz-Eakin, W. Newey, and H. S. Rosen. Estimating Vector Autoregressions with Panel Data. *Econometrica*, 56(6):1371, 1988.

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.

J. Y. Lee and G. Solon. The Fragility of Estimated Effects of Unilateral Divorce Laws on Divorce Rates. *The B.E. Journal of Economic Analysis Policy*, 11(1), 2011.

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, 2017.

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. Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure. *Econometrica*, 74(4):967–1012, 2006.

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.

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

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

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

as $\ell_n, N, T \to \infty$ such that $\ell_n/NT \to 0$.

(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 $\hat{a}\hat{b}\hat{c}\hat{d} - abcd = (\hat{a}-a)\hat{b}\hat{c}\hat{d} + a(\hat{b}-b)\hat{c}\hat{d} + ab(\hat{c}-\text{c})\,\hat{d} + abc(\hat{d}-d)$, 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

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

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