University of Connecticut

Department of Economics Working Paper Series

Improved Inference for Interactive Fixed Effects Model under Cross-Sectional Dependence

by

Zhenhao Gong
Shanxi University of Finance
and Economics

Min Seong Kim
University of Connecticut

Working Paper 2024-02
February 2024

365 Fairfield Way, Unit 1063
Storrs, CT 06269-1063
Phone: (860) 486-3022
Fax: (860) 486-4463
http://www.econ.uconn.edu/

Improved Inference for Interactive Fixed Effects Model under Cross-Sectional Dependence

Zhenhao Gong[1] and Min Seong Kim[2]

[1*]Department of Economics, Shanxi University of Finance and Economics.

[2]Department of Economics, University of Connecticut.

Abstract

This paper proposes an inference procedure for the interactive fixed effects model that is valid in the presence of cross-sectional dependence. When the error terms are cross-sectionally dependent, the Least Square (LS) estimator of this model is asymptotically biased and therefore the associated confidence interval tends to have a large coverage error. To address this, we propose a bias correction of the LS estimator and a cross-sectional dependence robust variance estimator to construct associated test statistics. The paper also discusses practical issues in implementing the proposed method, including the construction of distance that reflects the decaying pattern of cross-sectional dependence and the selection of the bandwidth parameters. Monte Carlo simulations show our procedure works well in finite samples. As empirical illustrations, we apply our procedure to study the effect of divorce law reforms on divorce rates and the impact of clean water and sewerage interventions on child mortality.

Keywords: Bandwidth selection, Bias correction, Robust inference, Spatial HAC method

JEL Classification: C12 , C13 , C15 , C31 , C33 , J12 , I18

1 Introduction

It is crucial in panel regression analysis to control for unobserved individual heterogeneity and time effects that are possibly correlated with regressors. A typical approach in this regard is to use the standard fixed effects model, in which the individual effects and time effects enter the model additively. An alternative approach is the interactive fixed effects (IFE) model, in which individual effects and time effects are interacted multiplicatively. Since the multiplicative structure captures the impact of individual heterogeneity and time effects in a more flexible way, the IFE model has attracted a great deal of attention in the literature. Holtz-Eakin et al. (1988) explore a quasi-differencing approach and Ahn and Horenstein (2001) propose the generalized method of moments (GMM) estimation in large N and fixed T panel models. The seminal work of Pesaran (2006) and Bai (2009) develop estimation and inference procedures of this model under the large N and large T asymptotics. Pesaran (2006) proposes the common correlated effects (CCE) estimator for heterogeneous panel models, and Bai (2009) develops the LS estimator using the principal component method. Moon and Weidner (2017) also consider the LS estimator for this model in the dynamic panel context. More recently, Hong et al. (2023) propose a two-step profile GMM estimation procedure to estimate the IFE model when the regressor are allowed to be correlated with the idiosyncratic error terms. Callaway and Karami (2023) considers identifying and estimating the Average Treatment Effect on the Treated (ATT) when untreated potential outcomes are generated by IFE model.

In this paper, we propose an inference procedure on the IFE model that is valid in the presence of cross-sectional dependence in the large N and large T asymptotic framework. As discussed by Bai (2009), the LS estimator of this model is asymptotically biased when the error terms are cross-sectionally and/or serially heteroskedastic and dependent. This is known as an incidental parameters problem (Neyman and Scott, 1948; Nickell, 1981), and inference without taking this problem into account may yield a misleading statistical conclusion. We develop an inference procedure that addresses this issue. For technical simplicity we assume there is no serial dependence and focus on issues caused by cross-sectional dependence. Our work is empirically relevant since local dependence may often remain even after introducing the factor structure. For example, when we use state level data, nationwide dependence can be captured by the factor structure but local correlations among neighbor states are likely to still exist in the model.

Our procedure consists of two parts: First, we correct the bias of the LS estimator. Using the fact that the bias has a form of the time-series average of cross-sectional long-run covariances, we develop a bias estimator based on the spatial HAC approach. Second, we propose a cross-sectional dependence robust variance estimator to construct associated test statistics. As the bias estimator, we employ the spatial HAC method to estimate the variance. Spatial HAC estimation is first proposed by Conley (1996, 1999), and is further studied by Kelejian and Prucha (2007) and Kim and Sun (2011).

Bai (2009) proposes a partial sample approach in the presence of cross-sectional dependence, which constructs the bias and variance estimators using partial samples. A practical issue of this method is that there is no practical guidance on the selection of partial sample that reflects the dependence structure of all cross-sectional units. Another approach that yields valid inference in this setting is the GLS method by Bai and Liao (2017). Their GLS transformation eliminates the cross-sectional correlations of idiosyncratic errors, which makes their estimator asymptotically centered at the true value. This approach is attractive in that it is efficient and free of incidental parameters bias. On the other hand, our simulation shows that the GLS inference tends to yield substantial size distortions. Our procedure is shown to produce more accurate inference results though it is not as efficient as the GLS. From this perspective, our procedure complements the existing methods and we make a contribution to the literature.

There are two practical issues in implementing the proposed method. The first one is how to construct a distance measure. Since our bias and variance estimators are constructed on the kernel based spatial HAC method, we need a distance measure that characterizes the dependence structure of the data. A typical approach in the literature is to use an auxiliary variable that captures the decaying pattern of dependence (e.g., the transportation cost in Ligon and Conley, 2002; the geographic distance in Pinkse et al., 2002). However, such a variable may not be available in some applications. To address this, we define a distance that reflects the cross-sectional dependence structure directly and propose its implementation using the information from the time dimension. See, e.g., Cui et al. (2021), Fernandez (2011), Kim (2021), and Mantegna (1999). An advantage of this method is that no prior information about the dependence structure is required. The second issue is the selection of the bandwidth parameters. This is particularly challenging in our setting because we need to choose two bandwidths jointly in estimating the bias and variance. We propose a selection method based on the cluster wild bootstrap, in which each cluster contains all units in one time period to replicate the cross-sectional dependence of the original data in bootstrap samples. We choose the bandwidths that maximize the bootstrap rejection probability under the null after controlling the size.

While we establish our methods under the assumption of serial independence for simplicity, serial dependence is a common feature of panel data and is another source of incidental parameters problem in this model. If researchers suspect the existence of serial dependence, we recommend they consider the bias correction method proposed by Bai (2009). His method is easy to use because it is based on the well-known Newey and West HAC estimation procedure. Regarding variance estimation, we can apply our variance estimation method after clustering by individual. We discuss the possible extension for an inference procedure that is robust to both serial and cross-sectional dependence and heteroskedasticity in a remark under Section 4.

We illustrate the application of our approach in 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. Wolfers (2006) uses the additive fixed effects model and finds divorce rates rose in the first eight years after the law reforms. However, the robustness of his results has been doubted because the model may not be flexible enough to capture unobserved heterogeneity varying over time and across states. Kim and Oka (2013) address this issue by employing the IFE model, but they do not take the cross-sectional dependence into account. We apply the proposed method to conduct inference for this model. The second application studies the effects of clean water and effective sewerage systems on U.S. child mortality. An important question in public health is the cause of the sharp decrease in the U.S. infant mortality between the late 19th century and the early 20th century. To answer this question, Alsan and Goldin (2019) employ the additive fixed effects model to estimate the independent and combined effects of clean water and effective sewerage systems on under-5 mortality in Massachusetts in 1880-1920. We employ the IFE model with the proposed inference procedure to examine the robustness of their results.

The remainder of the paper is as follows. Section 2 reviews the IFE model and the LS estimator. In Sections 3, we introduce the proposed inference method and study its asymptotics. Section 4 proposes our distance measure and selection rule of the bandwidth parameters, which are necessary to implement our method. Section 5 presents the simulation results. In Section 6, we apply the proposed method to the empirical applications introduced above. The last section concludes. All proofs and additional simulation results are provided in the Appendix.

2 Review of IFE model

We consider a linear panel model with interactive fixed effects

$$Y_{it} = X_{it}'\beta + u_{it}, \quad u_{it} = \lambda_i' F_t + \varepsilon_{it}, \quad i = 1, \cdots N; \ \ t = 1, \cdots, T, \tag{2.1}$$

where Y_{it} is an outcome variable, X_{it} is a $(p \times 1)$ vector of regressors, β is a vector of regression coefficients, and u_{it} is an error term. We assume a factor loading structure in u_{it}, where F_t is a $(r \times 1)$ vector of common factors, λ_i is a vector of factor loadings, and ε_{it} represents the idiosyncratic error. The number of factors r is assumed to be known. X_{it} is possibly correlated with λ_i and/or F_t.

The model can be rewritten as

$$Y_i = X_i\beta + F\lambda_i + \varepsilon_i, \tag{2.2}$$

where $Y_i = (Y_{i1}, \cdots, Y_{iT})'$, $X_i = (X_{i1}, \cdots, X_{iT})'$, $F = (F_1, \cdots, F_T)'$ and $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{iT})'$. Bai (2009) considers the LS estimator of this model, which is given by

$$(\hat{\beta}, \hat{F}, \hat{\Lambda}) = \arg\min_{\beta, F, \Lambda} \sum_{i=1}^{N} (Y_i - X_i\beta - F\lambda_i)'(Y_i - X_i\beta - F\lambda_i), \tag{2.3}$$

where $\Lambda = (\lambda_1, \cdots, \lambda_N)'$. Since F and Λ are multiplicative, they are not separately identifiable and the following normalization

$$\frac{1}{T}\sum_{t=1}^{T} F_t F_t' = I_r \text{ and } \sum_{i=1} \lambda_i \lambda_i' = \text{ diagonal} \tag{2.4}$$

is employed. (2.4) generates r^2 restrictions, which are sufficient to determine F and Λ uniquely.

Using the first order condition, we have

$$\lambda_i(\beta, F) = \frac{1}{T} F' (Y_i - X_i\beta) \text{ and } \Lambda(\beta, F) = \frac{1}{T} (Y - X\beta)' F. \tag{2.5}$$

Plugging (2.5) into (2.3), we can obtain the LS estimator for β given F

$$\hat{\beta}(F) = \Big(\sum_{i=1}^{N} X_i' M_F X_i\Big)^{-1} \sum_{i=1}^{N} X_i' M_F Y_i, \tag{2.6}$$

where $M_F = I - F(F'F)^{-1}F'$.

For the estimation of F, since (2.2) reduces to a pure factor model given β, we can use the principal components method. More specifically, the 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 LS estimators $(\hat{\beta}, \hat{F})$ are obtained by solving the following equations simultaneously:

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i' M_{\hat{F}} X_i\right)^{-1} \sum_{i=1}^{N} X_i' M_{\hat{F}} Y_i, \tag{2.7}$$

and

$$\left[\frac{1}{NT}\sum_{i=1}^{N}(Y_i - X_i\hat{\beta})(Y_i - X_i\hat{\beta})'\right]\hat{F} = \hat{F} V_{NT}, \tag{2.8}$$

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

We follow Bai (2009) to make the following assumptions to establish the asympotics.

Assumption 2.1 *$E\|X_{it}\|^4 \leq M$ and let $\mathcal{F} = \{F : F'F/T = I\}$. Define*

$$D(F) = \frac{1}{NT}\sum_{i=1}^{N} X_i' M_F X_i - \frac{1}{T}\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N} X_i' M_F X_k a_{ik}\right), \tag{2.9}$$

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

Assumption 2.2 *(a) $E\|F_t\|^4 \leq M$ 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 $T \to \infty$; (b) $E\|\lambda_i\|^4 \leq M$ and $\frac{1}{N}\Lambda'\Lambda \xrightarrow{p} \Sigma_\Lambda$ for some $r \times r$ positive definite matrix Σ_Λ, as $N \to \infty$.*

Assumption 2.3
(a) $E(\varepsilon_{it}) = 0$ and $E|\varepsilon_{it}|^8 \leq M$;
(b) $E(\varepsilon_{it}\varepsilon_{ks}) = 0$ for all (i,k) if $t \neq s$ and $E(\varepsilon_{it}\varepsilon_{kt}) = \sigma_{kk,t} = \sigma_{kk}$ for all $t = 1, ..., T$. $E(\varepsilon_{it}\varepsilon_{kt}) = \sigma_{ik,t}$, $|\sigma_{ik,t}| \leq \bar{\sigma}_{ik}$ for all $t = 1, ..., T$ such that $\frac{1}{N}\sum_{i,k=1}^{N} \bar{\sigma}_{ik} \leq M$, 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;
(c) For every (t,s), $E\left|N^{-1/2}\sum_{i=1}^{N}(\varepsilon_{it}\varepsilon_{is} - E(\varepsilon_{it}\varepsilon_{is}))\right|^4 \leq M$;
(d) Moreover

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

Assumption 2.4 *ε_{it} is independent of X_{ks}, λ_k, and F_s for all i, t, k and s.*

Assumption 2.5 *Let $Z_i = (Z_{i1}, \ldots, Z_{iT})' = M_{F^0}X_i - \frac{1}{N}\sum_{k=1}^{N} a_{ik} M_{F^0} X_k$, where F^0 is the true F. We have*

$$\frac{1}{\sqrt{NT}}\sum_{i=1}^{N} Z_i'\varepsilon_i \xrightarrow{d} N(0, \Omega), \tag{2.10}$$

where $\Omega = \lim \Omega_{NT}$ with $\Omega_{NT} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} \sigma_{ik,t} E(Z_{it}Z_{kt}')$.

Assumption 2.1 indicates $D(F)$ is positive definite and excludes the low-rank regressors (e.g., time-invariant and common regressors) in (2.2). Assumption 2.2 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 principal component estimators of F and Λ in the factor model. Assumption 2.3 assumes ε_{it} to be serially uncorrelated and homoskedasticity but allow it to exhibit weak cross-sectional correlation and heteroskedasticity. Assumption 2.4 rules out the dynamic panel data model. Assumption 2.5 states a central limit theorem holds for the moment process.

Under Assumptions 2.1-2.5, Bai (2009) shows that

$$\sqrt{NT}(\hat{\beta}-\beta) \xrightarrow{d} N\left(\rho^{1/2}B_0, D_0^{-1}\Omega D_0'^{-1}\right), \tag{2.11}$$

as $T/N \to \rho$, where $D_0 = \text{plim}D(F^0) = \text{plim}\frac{1}{NT}\sum_{i=1}^{N} Z_i'Z_i$,

$$B_0 = \text{plim}\, B_{NT} \text{ with } B_{NT} = -D_0^{-1}\underbrace{\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} w_i\lambda_k\left(\frac{1}{T}\sum_{t=1}^{T}\sigma_{ik,t}\right)}_{:=J_{NT}}, \tag{2.12}$$

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

The expression of B_0 in (2.12) implies that $\hat{\beta}$ is asymptotically centered at the true value only in the absence of cross-sectional correlation and heteroskedasticity. If not, $\hat{\beta}$ becomes asymptotically biased, and inference without correcting this bias leads to misleading conclusions.

3 Inference under cross-sectional dependence

In this section, we propose an inference procedure for β, which is valid under cross-sectional dependence. As we can see from (2.11), the LS estimator of the IFE model is asymptotically biased when $\{\varepsilon_{it}\}$ are cross-sectionally dependent, and the bias needs to be corrected for valid inference. Bai (2009) considers a partial sample approach to address this problem. His bias estimator for B_0 is given by

$$\hat{B}_{CS} = -\hat{D}^{-1}\frac{1}{n_{sub}}\sum_{i=1}^{n_{sub}}\sum_{k=1}^{n_{sub}}\hat{w}_i\hat{\lambda}_k\left(\frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\right), \tag{3.1}$$

where n_{sub} is the size of the partial sample. $\hat{D}$ and $\hat{w}_i$ are the estimators of D_0 and w_i defined as

$$\hat{D} = \frac{1}{NT}\sum_{i=1}^{N}\hat{Z}_i'\hat{Z}_i = \frac{1}{NT}\sum_{i=1}^{N} X_i'M_{\hat{F}}X_i - \frac{1}{T}\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N} X_i'M_{\hat{F}}X_k\hat{a}_{ik}\right)$$

$$\hat{w}_i = \left(\frac{(X_i-\hat{V}_i)'\hat{F}}{T}\right)\left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\right)^{-1},$$

where $\hat{Z}_i = M_{\hat{F}}X_i - \frac{1}{N}\sum_{k=1}^{N}\hat{a}_{ik}M_{\hat{F}}X_k$, $\hat{V}_i = N^{-1}\sum_{k=1}^{N}\hat{a}_{ik}X_k$ and $\hat{a}_{ik} = \hat{\lambda}_i'(\hat{\Lambda}'\hat{\Lambda}/N)^{-1}\hat{\lambda}_k$. Bai shows $\hat{B}_{CS}$ is consistent as $n_{sub}/\min\{N,T\} \to 0$, and constructs the bias corrected estimator for β based on $\hat{B}_{CS}$. While this method is valid in the asymptotic sense, a practical issue for its implementation is that there is no practical guidance on how to select the partial sample to replicate the cross-sectional dependence of all observations. If the partial sample fails to do so, the finite sample performance of $\hat{B}_{CS}$ would be poor. This is a practically important problem because the cross-sectional dependence structure is unknown and very complex in general.

We propose an alternative bias estimator using the spatial HAC estimation approach. Since $B_{NT} = -D_0^{-1} J_{NT}$ and D_0 is consistently estimated by $\hat{D}$ in (3), our interest is in

$$J_{NT} = \frac{1}{T} \sum_{t=1}^{T} J_t \text{ where } J_t = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} w_i \lambda_k E(\varepsilon_{it} \varepsilon_{kt}). \tag{3.2}$$

We propose estimating J_{NT} with

$$\hat{J} = \frac{1}{T} \sum_{t=1}^{T} \hat{J}_t \text{ where } \hat{J}_t = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} K_1 \left(\frac{d_{ik}}{d_{(1)}} \right) \hat{w}_i \hat{\lambda}_k \hat{\varepsilon}_{it} \hat{\varepsilon}_{kt}, \tag{3.3}$$

where $K_1(\cdot)$ is a real-valued kernel function, $d_{(1)}$ is the bandwidth parameter, and d_{ik} is distance between units i and k that reflects the strength of their cross-sectional dependence. More specifically, we assume d_{ik} to be small if ε_{it} and ε_{kt} are highly dependent and vice versa. The distance measure is assumed to satisfy $d_{ik} \geq 0$, $d_{ii} = 0$ and $d_{ik} = d_{ki}$, but we do not require the triangle inequality $d_{ik} \leq d_{ij} + d_{jk}$ to hold.

$\hat{J}_t$ has a form of the spatial HAC estimator (e.g., Conley, 1999; Kelejian and Prucha, 2007) and $\hat{J}$ is its time-series average. Based on $\hat{J}$, we construct the bias corrected estimator for β as follows:

$$\hat{\beta}^{\dagger} = \hat{\beta} - \frac{1}{N} \hat{B} \text{ where } \hat{B} = -\hat{D}^{-1} \hat{J}. \tag{3.4}$$

To examine the asymptotic properties of our bias correction, we first assume that ε_{it} has the following linear representation.

Assumption 3.1 *(a)* $\varepsilon_{it} = \sum_{l=1}^{\infty} \gamma_{il} e_{lt}$, *where* $e_{lt} \overset{iid}{\sim} (0,1)$, $E(e_{lt}^4) \leq \infty$, *and* γ_{il} *is an unknown constant; (b)* $\sum_{i=1}^{\infty} |\gamma_{il}| < M$ *for each* l, *and* $\sum_{l=1}^{\infty} |\gamma_{il}| < M$ *for each* i.

Assumption 3.1(a) states ε_{it} is an infinite weighted sum of i.i.d random variables. This linear representation is widely used in the literature to characterize spatial processes (e.g., Kelejian and Prucha, 2007; Robinson, 2011; Kim and Sun, 2011; Pesaran and Tosetti, 2011; Kim, 2021). It also includes the popular spatial parametric models such as spatial autoregressive (SAR) processes and spatial moving average (SMA) processes as special cases. Assumption 3.1(b) states the summability conditions for the coefficients of the linear representation, which restricts the strength of cross-sectional dependence. This condition enables us to control the variance of $\hat{B}$. An alternative to this linear representation and summability assumptions would be to introduce some mixing and stationary assumptions to obtain the fourth-order cumulant condition as in the time-series (e.g., Andrews, 1991), which, as pointed out by Bai and Ng (2006), is difficult to justify in cross-sectional models.

We introduce the following assumption on the kernel functions.

Assumption 3.2 *(a) The kernel* $K_a : \mathbb{R} \to [-1,1]$ *satisfies* $K_a(0) = 1, K_a(x) = K_a(-x), K_a(x) = 0$ *for* $|x| \geq 1$. *(b) For all* $x_1, x_2 \in \mathbb{R}$ *there is a constant,* $c_L < 0$, *such that*

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

Assumption 3.2 is standard and is satisfied by most of the kernels in the HAC literature including the Bartlett, Tukey-Hanning, and Parzen kernels.

We suppose that there exists $q_a \in [0, \infty)$ such that

$$K_a^{(q_a)} = \lim_{x \to 0} \frac{1 - K_a(x)}{|x|^{q_a}} < \infty.$$

The largest value of q_a is called Parzen characteristic exponent of $K_a(x)$ and reflects the smoothness of $K_a(x)$ at $x = 0$. The assumption below provides the condition that we use to control the bias.

Assumption 3.3 *There exists a finite constant* M *such that*

$$\lim_{N,T \to \infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{k=1}^{N} \sum_{t=1}^{T} |\sigma_{ik,t}| d_{ik}^{q_a} < M \tag{3.5}$$

for $q_a \in [0, \infty)$.

Assumption 3.3 requires d_{ik} to capture the decaying pattern of the cross-sectional dependence of $\{\varepsilon_{it}\}$. That is, $\sigma_{ik,t}$ is assumed to decrease to zero fast enough as d_{ik} grows so that the summability condition in (3.5) holds. This is a key condition that our distance measure should satisfy. We rely on this assumption to control the bias caused by the truncation and downweight imposed by the kernel function. When d_{ik} increases, the weight that $K_a(\cdot)$ assigns to $\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}$ in (3.3) becomes smaller, but this does not lead to much bias since $\sigma_{ik,t}$ also decreases to zero with d_{ik} under this assumption.

We define

$$\ell_{iN(a)} = \sum_{k=1}^{N} 1\{d_{ik} \leq d_{(a)}\} \text{ and } \bar{\ell}_{N(a)} = \frac{1}{N} \sum_{i=1}^{N} \ell_{iN(a)}, \tag{3.6}$$

where $\ell_{iN(a)}$ is the number of pseudo-neighbors that unit i has within $d_{(a)}$, and $\bar{\ell}_{N(a)}$ is the average number of pseudo-neighbors. It is obvious that $\bar{\ell}_{N(a)}$ is an increasing function of $d_{(a)}$. To control the variations of our HAC estimators, we should increase $d_{(a)}$ slowly as N and T grow so that $\bar{\ell}_{N(a)} / \min\{N, T\}$ converges to zero.

Assumption 3.4 $\ell_{iN(a)} \leq c_\ell \bar{\ell}_{N(a)}$ *for all* $i = 1, \cdots, N$ *with some constant* c_ℓ.

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

The following theorem establishes the consistency of our bias estimator.

Theorem 3.1 *Suppose that Assumptions 2.1-2.5 and 3.1-3.4 hold. If* $d_{(1)}, N, T \to \infty$ *such that* $\frac{\bar{\ell}_{N(1)}}{\min\{N,T\}} \to 0$ *and* $\frac{T}{N} \to \rho > 0$,

$$\hat{B} - B_{NT} \xrightarrow{p} 0. \tag{3.7}$$

The corollary below follows directly from the asymptotic normality of the LS estimator in (2.11) and Theorem 3.1.

Corollary 3.1 *Suppose that the assumptions and the rate conditions in Theorem 3.1 hold.*

$$\sqrt{NT}\big(\hat{\beta}^\dagger - \beta\big) \xrightarrow{d} N\big(0, D_0^{-1}\Omega D_0'^{-1}\big). \tag{3.8}$$

Corollary 3.1 indicates that our bias correction removes the bias successfully and $\hat{\beta}^\dagger$ is asymptotically centered at the true value. We also find that our bias correction does not contribute to the asymptotic variance. This result implies that we can use $\hat{\beta}^\dagger$ to make valid inference on β if we have a consistent estimator of Ω_{NT} in the presence of cross-sectional dependence.

Regarding the estimation of Ω_{NT}, the cross-sectional heteroskedasticity robust (HR) variance estimator is commonly used, which is given by

$$\hat{\Omega}_{HR} = \frac{1}{N}\sum_{i=1}^{N} \hat{\sigma}_i^2 \left(\frac{1}{T}\sum_{t=1}^{T} \hat{Z}_{it}\hat{Z}_{it}'\right) \text{ with } \hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_{it}^2. \tag{3.9}$$

However, it is obvious that $\hat{\Omega}_{HR}$ is not valid in our setting, because it does not allow for cross-sectional dependence of ε_{it}.

Bai (2009) addresses this issue by developing a partial sample variance estimator

$$\hat{\Omega}_{CS} = \frac{1}{n_{sub}}\sum_{i=1}^{n_{sub}}\sum_{k=1}^{n_{sub}} \left(\frac{1}{T}\sum_{t=1}^{T} \hat{Z}_{it}\hat{Z}_{kt}'\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\right), \tag{3.10}$$

and shows it is consistent as $n_{sub}/\min\{N,T\} \to 0$. As Bai's bias estimator $\hat{B}_{CS}$ in (3.1), a practical problem for this partial sample approach is that it is not clear how to choose the partial sample to replicate the overall cross-sectional dependence structure.

We propose an estimator of Ω_{NT} using the spatial HAC approach. The estimator is given by

$$\hat{\Omega} = \frac{1}{T}\sum_{t=1}^{T}\hat{\Omega}_t \text{ with } \hat{\Omega}_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}K_2\left(\frac{d_{ik}}{d_{(2)}}\right)\hat{Z}_{it}\hat{Z}_{kt}'\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}, \tag{3.11}$$

where $d_{(2)}$ is the bandwidth parameter. As our bias estimator, $\hat{\Omega}$ is a time average of spatial HAC estimators $\{\hat{\Omega}_t\}$. The following theorem establishes the consistency of our variance estimator.

Theorem 3.2 *Suppose that Assumptions 2.1-2.5 and 3.1-3.4 hold. If* $d_{(2)}, N, T \to \infty$ *such that* $\frac{\bar{\ell}_{N(2)}}{\min\{N,T\}} \to 0$ *and* $\frac{T}{N} \to \rho > 0$,

$$\hat{\Omega} - \Omega_{NT} \xrightarrow{p} 0. \tag{3.12}$$

Suppose that we employ the rectangular kernel for $K_2(\cdot)$ and select $d_{(2)}$ to be large such that $\bar{\ell}_{N(2)} = N$ and $K_2\left(\frac{d_{ik}}{d_{(2)}}\right) = 1$ for all i and k. In this case, $\hat{\Omega}$ reduces to the clustered variance estimator of which cluster is formed by each time period. The clustering-by-time approach is widely used in the panel model to construct a variance estimator that is robust to cross-sectional dependence. See, for example, Driscoll and Kraay (1998) and Vogelsang (2012), who use this approach in the linear panel model with individual fixed effects setup. It is interesting to note that our rate condition requires $\frac{\bar{\ell}_{N(2)}}{N} \to 0$ and this implies the clustering-by-time approach does not produce a consistent variance estimator in the IFE model. This rate condition is necessary to control for the effect of estimation errors in $\hat{F}$ and $\hat{\Lambda}$ on $\hat{\Omega}$ under the large N and T asymptotics. See the proof of Theorem 3.2 in the Appendix B for more details.

Suppose that we are interested in testing the following null hypothesis:

$$H_0 : R\beta = r_0, \tag{3.13}$$

where R is a $q \times p$ matrix and r_0 is a $q \times 1$ vector. Using our bias corrected estimator and variance estimator, we propose the following Wald statistic to test H_0:

$$W_{NT} = \sqrt{NT}\left(R\hat{\beta}^{\dagger} - r_0\right)'\left(R\hat{D}^{-1}\hat{\Omega}\hat{D}^{-1}R'\right)^{-1}\sqrt{NT}\left(R\hat{\beta}^{\dagger} - r_0\right).$$

In case $q = 1$, we can also consider the t-statistic, which is given by

$$t_{NT} = \frac{\sqrt{NT}\left(R\hat{\beta}^{\dagger} - r_0\right)}{\sqrt{R\hat{D}^{-1}\hat{\Omega}\hat{D}^{-1}R'}}.$$

The corollary below characterizes the limiting distribution of our test statistics under the null hypothesis.

Corollary 3.2 *Suppose that the assumptions and the rate conditions in Theorems 3.1 and 3.2 hold. If H_0 is true, we have*

$$W_{NT} \xrightarrow{d} \chi_q^2 \text{ and } t_{NT} \xrightarrow{d} N(0,1).$$

Corollary 3.2 follows directly from Corollary 3.1 and Theorem 3.2. It suggests we use the χ^2 distribution with the degree of freedom q and the standard normal distribution to obtain critical values.

4 Implementation

In this section, we discuss two major issues in implementing the proposed procedure. The first is the construction of a distance measure. Since we estimate the bias and variance based on the kernel based spatial HAC approach, we need a distance that captures the decaying pattern of cross-sectional dependence. Different variables have been used as distance depending on applications. e.g., the transportation cost in Ligon and Conley (2002) and the geographic distance in Pinkse et al. (2002). However, such a variable may not be available in some empirical applications. To address this issue, we consider the following distance

$$d_{ik}^{\mathrm{D}} = \frac{1}{|\rho_{ik}|} - 1, \text{ with } \rho_{ik} = Corr(\varepsilon_{it}, \varepsilon_{kt}),$$

which reflects the strength of dependence by definition.

Though d_{ik}^{D} is not observed, we can estimate it using the sample counterpart

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

with $\hat{\rho}_{ik} = \sum_{t=1}^{T} \hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} / \sqrt{\sum_{t=1}^{T} \hat{\varepsilon}_{it}^2 \sum_{t=1}^{T} \hat{\varepsilon}_{kt}^2}$. An advantage of using $\hat{d}_{ik}^{D}$ is that it is easy to calculate and no prior information about the dependence structure is required. We note that d_{ik}^{D} does not satisfy the triangle inequality, but the validity of our procedure does not require this inequality. Constructing distance based on the correlation coefficient has been considered in the literature. See, e.g., Mantegna (1999), Fernandez (2011), Cui et al. (2021), and Kim (2021). The performance of our procedure based on $\hat{d}_{ik}^{\mathrm{D}}$ is investigated in our simulations and empirical applications.

Another issue in implementing our method is how to select the bandwidth parameters. This is particularly challenging in our model because two bandwidth parameters should be chosen jointly for $\hat{J}$ and $\hat{\Omega}$. We address this issue by developing a selection method based on the cluster wild bootstrap. Our approach is similar to the one in Kim et al. (2017), who propose a simulation-based calibration approach for the choice of bandwidth parameters in the time-series kernel method. They use an autoregression type parametric model to approximate time-series dependence. However, their approach is not applicable to our model because cross-sectional dependence is much more complex and hard to parametrize. We employ a cluster wild bootstrap, in which each cluster contains all units in one time period, to replicate cross-sectional dependence without using a parametric model.

Let $\mathcal{D}_M^{(1)} = \{d_{1,(1)}, \cdots, d_{M,(1)}\}$ and $\mathcal{D}_S^{(2)} = \{d_{1,(2)}, \cdots, d_{S,(2)}\}$ be the sets of reasonable bandwidth values for the choice of $d_{(1)}$ and $d_{(2)}$. Our selection procedure follows the steps below.

STEP 1. Estimate $\hat{\beta}$, $\hat{F}_t$, $\hat{\Lambda}$ using the LS and obtain the residuals from

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

where $Y_t = (Y_{1t}, \cdots, Y_{Nt})'$, $X_t = (X_{1t}, \cdots, X_{Nt})'$ and $\varepsilon_t = (\varepsilon_{1t}, \cdots, \varepsilon_{Nt})'$.

STEP 2. Generate $\mathcal{B}$ bootstrap samples based on the following procedure

$$Y_{b,t}^* = X_t\hat{\beta} + \hat{\Lambda}\hat{F}_t + \varepsilon_{b,t}^*,$$

where $\varepsilon_{b,t}^* = \hat{\varepsilon}_t \xi_{b,t}$ with $\xi_{b,t} \overset{iid}{\sim} (0,1)$, for $b = 1, ..., \mathcal{B}$.

STEP 3. For each bootstrap sample, compute the bootstrap LS estimators $\hat{\beta}_b^*$, $\hat{F}_{b,t}^*$, $\hat{\Lambda}_b^*$, and $\hat{\varepsilon}_{b,t}^*$ and construct the bootstrap bias estimator with $d_{1,(1)}$ such that

$$\hat{B}_b^*(d_{1,(1)}) = -\hat{D}_b^{*-1}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}K_1\left(\frac{d_{ik}}{d_{1,(1)}}\right)\hat{w}_{b,i}^*\hat{\lambda}_{b,k}^*\hat{\varepsilon}_{b,it}^*\hat{\varepsilon}_{b,kt}^*\right),$$

where $\hat{D}_b^*$ and $\hat{w}_{b,i}^*$ are the bootstrap versions of $\hat{D}$ and $\hat{w}_i$ in (3) by replacing $\hat{F}$ and $\hat{\Lambda}$ with $\hat{F}_b^*$ and $\hat{\Lambda}_b^*$, respectively.

STEP 4. Construct the bootstrap variance estimator $\hat{\Omega}_b^*(d_{1,(2)})$ with $d_{1,(2)}$ such that

$$\hat{\Omega}_b^*(d_{1,(2)}) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\hat{Z}_{b,it}^*\hat{Z}_{b,kt}^{*\prime}\hat{\varepsilon}_{b,it}^*\hat{\varepsilon}_{b,kt}^*K_2\left(\frac{d_{ik}}{d_{1,(2)}}\right)\right),$$

where $\hat{Z}_{b,it}^*$ is the bootstrap version of $\hat{Z}_{it}$ in (3.9) by replacing $\hat{F}$ and $\hat{\Lambda}$ with $\hat{F}_b^*$ and $\hat{\Lambda}_b^*$, respectively.

STEP 5. Compute the bootstrap Wald statistic (or t-statistic) for each bootstrap sample

$$W_b^*(d_{1,(1)}, d_{1,(2)}) = \sqrt{NT}\left(R\left(\hat{\beta}_b^{\dagger *} - \hat{\beta}\right)\right)'\left(R\hat{\Gamma}_b^*R'\right)^{-1}\sqrt{NT}\left(R\left(\hat{\beta}_b^{\dagger *} - \hat{\beta}\right)\right),$$

where

$$\hat{\beta}_b^{\dagger *} = \hat{\beta}_b^* - \frac{1}{N}\hat{B}_b^*\big(d_{1,(1)}\big) \text{ and } \hat{\Gamma}_b^* = \hat{D}_b^{*-1}\hat{\Omega}_b^*\big(d_{1,(2)}\big)\hat{D}_b^{*-1}.$$

STEP 6. Let α denote the pre-selected significant level. Compute the bootstrap rejection probability

$$\mathcal{V}^*(d_{1,(1)}, d_{1,(2)}) = \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\big(W_b^*(d_{1,(1)}, d_{1,(2)}) > \chi^2_{1-\alpha}(q)\big),$$

where $1(\cdot)$ is the indicator function and $\chi^2_{1-\alpha}(q)$ is the $(1-\alpha)$ quantile of the χ^2 distribution with the degree of freedom q.

STEP 7. Repeat Step 2 to Step 6 for all combinations of $\big(d_{(1)}, d_{(2)}\big) \in \mathcal{D}_M^{(1)} \bigotimes \mathcal{D}_S^{(2)}$ and select $\big(\hat{d}_{(1)}, \hat{d}_{(2)}\big)$ that solve

$$\max_{(d_{(1)}, d_{(2)}) \in \mathcal{D}_M^{(1)} \bigotimes \mathcal{D}_S^{(2)}} \mathcal{V}^*\big(d_{(1)}, d_{(2)}\big), \quad s.t.\ \mathcal{V}^*\big(d_{(1)}, d_{(2)}\big) \leq \alpha.$$

If none of $d_{(1)}$ and $d_{(2)}$ satisfy the constraint $\mathcal{V}^*\big(d_{(1)}, d_{(2)}\big) \leq \alpha$, choose the one that minimizes $\mathcal{V}^*\big(d_{(1)}, d_{(2)}\big)$.

Step 2 describes our cluster wild bootstrap. Since the external random variable ξ_t is common for $\{\hat{\varepsilon}_{it}\}_{i=1}^n$, our bootstrap maintains the cross-sectional dependence of residuals, and $\hat{B}^*(d_{(1)})$ and $\hat{\Omega}^*(d_{(2)})$ are expected to be good approximations to $\hat{B}(d_{(1)})$ and $\hat{\Omega}(d_{(2)})$. We generate ξ_t from the Rademacher distribution in our simulations and empirical applications. Step 6 suggests we choose the bandwidths that maximize the bootstrap rejection probability after controlling it under the significance level α. If the rejection probability exceeds α with all $(d_{(1)}, d_{(2)}) \in \mathcal{D}_M^{(1)} \bigotimes \mathcal{D}_S^{(2)}$, we choose the ones that minimize the size distortion. $\mathcal{B}$ denotes the number of bootstrap replications and we set $\mathcal{B} = 300$ in our simulations and empirical applications.

Remark: In the presence of serial and cross-sectional dependence and heteroskedasticity, as Bai (2009) shows, there exists another asymptotic bias term C_0, which is the probability limit of C_{NT} with

$$C_{NT} = -D\left(F^0\right)^{-1}\frac{1}{NT}\sum_{i=1}^{N} X_i' M_{F^0}\Phi F^0\left(\frac{F^{0\prime}F^0}{T}\right)^{-1}\left(\frac{\Lambda'\Lambda}{N}\right)^{-1}\lambda_i$$

and $\Phi = \frac{1}{N}\sum_{k=1}^{N}\Phi_k$ with $\Phi_k = E\left(\varepsilon_k\varepsilon_k'\right)$.

To estimate C_{NT}, we need to construct consistent estimators for $X_i'\Phi_k F^0$ and $F^{0\prime}\Phi_k F^0$, and then take averages over i and k. These terms are standard expressions in the usual HAC literature. Let $Q_i = (X_i, F^0)$, then

$$\frac{1}{T} Q_i' \Phi_k Q_i = \begin{bmatrix} T^{-1} X_i' \Phi_k X_i & T^{-1} X_i' \Phi_k F^0 \\ T^{-1} F^{0\prime} \Phi_k X_i & T^{-1} F^{0\prime} \Phi_k F^0 \end{bmatrix}$$

contains $T^{-1} X_i' \Phi_k F^0$ and $T^{-1} F^{0\prime} \Phi_k F^0$ as sub-blocks, and the limit of $\frac{1}{T} Q_i' \Phi_k Q_i$ is the long-run variance of $T^{-1/2} \sum_{t=1}^{T} Q_{it} \varepsilon_{kt}$. Thus, for each given (i,k), we can construct a consistent estimator of $\frac{1}{T} Q_i' \Phi_k Q_i$ by the truncated kernel method of Newey and West (1987) based on the sequence of $\hat{Q}_{it} \hat{\varepsilon}_{kt} (t = 1, \cdots, T)$:

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \hat{Q}_{it} \hat{Q}_{is}' \hat{\varepsilon}_{kt} \hat{\varepsilon}_{ks} K_3 \left(\frac{d_{ts}}{d_{(3)}} \right),$$

where $K_3(\cdot)$ is a real-valued kernel function, $d_{(3)}$ is the bandwidth parameter, and $d_{ts} = |t - s|$ is time period between t and s that reflects the strength of serial dependence.

To make valid inference on β, we also need to construct a variance estimator that is robust to both types of dependencies. Assuming correlation and heteroskedasticity in both dimensions, the variance of $\hat{\beta}$ is given by $D_0^{-1} \Omega D_0^{-1}$ with

$$\Omega = \text{plim } \Omega_{NT} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{k=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} \sigma_{ik,ts} Z_{it} Z_{ks}',$$

where $\sigma_{ik,ts} = E(\varepsilon_{it} \varepsilon_{ks})$. For estimating it, we can apply the Panel HAC estimator proposed by Kim and Sun (2013):

$$\hat{\Omega}_{NT} = \frac{1}{NT} \sum_{i,k=1}^{N} \sum_{t,s=1}^{T} K_2 \left(\frac{d_{ik}}{d_{(2)}} \right) K_4 \left(\frac{d_{ts}}{d_{(4)}} \right) \hat{Z}_{it} \hat{Z}_{ks}' \hat{\varepsilon}_{it} \hat{\varepsilon}_{ks},$$

where $K_2(\cdot)$ and $K_4(\cdot)$ are real-valued kernel functions, $d_{(2)}$ and $d_{(4)}$ are the bandwidth parameters.

5 Monte Carlo Simulation

This section presents simulation evidence on the finite sample properties of the proposed procedure. We conduct inference on β in the following IFE model:

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

where we set $\beta = 0$ and

$$X_{it} = 1 + \lambda_i' F_t + \lambda_i + F_t + \eta_{it};$$
$$F_t = \rho F_{t-1} + \sqrt{1 - \rho^2} v_t; \quad \lambda_i, \eta_{it}, v_t \overset{iid}{\sim} N(0, 1),$$

The number of common factors is one and is assumed to be known. We let the common factors be serially weakly dependent by setting $\rho = 0.3$.

We employ the spatial MA process to model the cross-sectional dependence of the idiosyncratic errors. Assuming each unit to be located at (i_1, i_2) on an $(L_N \times L_N)$ square integer lattice, we generate $\varepsilon_t = (\varepsilon_{1t}, \cdots, \varepsilon_{Nt})'$ from the following process:

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

where $v_t = (v_{1t}, \cdots, v_{Nt})' \overset{iid}{\sim} N(0, I_N)$ and $M_1 = [m_{1,ik}]_{i,k=1}^N$ and $M_2 = [m_{2,ik}]_{i,k=1}^N$ are $(N \times N)$ spatial weighting matrices such that

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

with $d_{ik} = \sqrt{(i_1 - k_1)^2 + (i_2 - k_2)^2}$. Thus, ε_{it} and ε_{kt} are cross-sectional dependent if $d_{ik} \leq \sqrt{2}$. The number of replications is 3000.

We employ the distance measure $\hat{d}_{ik}^D$ proposed in Section 4 to construct $\hat{B}$ and $\hat{\Omega}$. We use our bootstrap method to select $d_{(1)}$ and $d_{(2)}$. The Parzen kernel and the rectangular kernel are used for $\hat{B}$ and $\hat{\Omega}$, respectively. One concern about $\hat{\Omega}$ is that it may not be positive semi-definite, which is often regarded as a desirable property in kernel HAC variance estimation noted by Newey and West (1987). However, we can achieve positive semi-definiteness with a simple modification. As $\hat{\Omega}$ is symmetric, $\hat{\Omega} = \hat{\Phi}\hat{\Xi}\hat{\Phi}'$, where $\hat{\Phi}$ is an orthogonal matrix and $\hat{\Xi} = \text{diag}(\hat{\nu}_1, ..., \hat{\nu}_N)$ is a diagonal matrix of the eigenvalues of $\hat{\Omega}$. Let $\hat{\Xi}^+ = \text{diag}(\hat{\nu}_1^+, ..., \hat{\nu}_N^+)$ where $\hat{\nu}_i^+ = \max\{\hat{\nu}_i, 0\}$. Then, we define

$$\hat{\Omega}^+ = \hat{\Phi}\hat{\Xi}^+\hat{\Phi}'.$$

As all the eigenvalues of $\hat{\Omega}^+$ is non-negative, it is positive semi-definite. We have the consistency of $\hat{\Omega}^+$ according to Theorem 4.1. of Politis (2011).

For comparative purposes, we also consider the GLS method proposed by Bai and Liao (2017) and the partial sample (CS) approach suggested by Bai (2009). Let $Y_t = (Y_{1t}, ..., Y_{Nt})'$ and $Y_t = (X_{1t}, ..., X_{Nt})'$. The GLS estimator is defined as

$$\hat{\beta}_{GLS} = \arg\min_{\beta} \min_{\Lambda, F_t} \sum_{t=1}^{T} (Y_t - X_t\beta - \Lambda F_t)'\hat{\Sigma}_{\varepsilon}^{-1}(Y_t - X_t\beta - \Lambda F_t), \qquad (5.1)$$

where $\hat{\Sigma}_{\varepsilon}$ is an estimator of Σ_{ε} which is the $(N \times N)$ covariance matrix of ε_t. Since Σ_{ε} is high-dimensional, it is estimated based on the thresholding method by Fan et al. (2013). Their GLS transformation is designed to eliminate cross-sectional correlation, so $\hat{\beta}_{GLS}$ does not suffer from the asymptotic bias. The variance of $\hat{\beta}_{GLS}$ is estimated by

$$\hat{\Gamma} = \frac{1}{NT} X'\hat{W}X, \tag{5.2}$$

where $X = (X_{11}, \cdots, X_{1T}, X_{21}, \cdots, X_{2T}, \cdots, X_{N1}, \cdots, X_{NT})'$, and

$$\hat{W} = \left(\hat{\Sigma}_{\varepsilon}^{-1} - \hat{\Sigma}_{\varepsilon}^{-1}\hat{\Lambda}(\hat{\Lambda}'\hat{\Sigma}_{\varepsilon}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{\varepsilon}^{-1}\right) \bigotimes M_{\hat{F}}.$$

For implementation of this method, we use the matlab code which is available on the authors' webpage.

For the CS approach, we construct $\hat{B}_{CS}$ in (3.1) by randomly selecting $n_{sub} = \min\{\sqrt{n}, \sqrt{T}\}$ consecutive units 20 times to obtain $\hat{B}_{CS}^{(1)}, \cdots, \hat{B}_{CS}^{(20)}$, and then take their average. Similarly, we construct the partial sample variance estimator $\hat{\Omega}_{CS}$ in (3.10) by selecting the sub-samples . The CS approach combines the bias-corrected estimator based on $\hat{B}_{CS}$ and the partial sample variance estimator based on $\hat{\Omega}_{CS}$.

Table 1 presents the empirical coverage probabilities (ECPs) of the 95% confidence intervals based on the LS estimator, the GLS estimator, the CS approach, and the proposed procedure. The HR variance estimator in (3.9) is used for the LS estimator. For our method, to investigate the effects of bias correction and robust variance estimation separately, we consider the confidence intervals based on (i) our bias corrected estimator with HR variance estimator, (ii) the LS estimator and our cross-sectional dependence robust variance estimator, and (iii) our bias corrected estimator and cross-sectional dependence robust variance estimator, which are denoted by SHAC_1, SHAC_2 and $\text{SHAC}^{\dagger}$, respectively. Thus, $\text{SHAC}^{\dagger}$ is the one that we propose to use in the paper.

The table reveals several important findings. Firstly, when there is no cross-sectional correlation in ε_{it} (i.e., $\theta = 0$), the LS method produces accurate confidence intervals. However, in the presence of cross-sectional correlation, the LS estimator-based confidence intervals exhibit substantially lower coverage rates than the nominal probability. For instance, when $(N, T) = (144, 150)$, the ECP based on the LS method with $\theta = 0$ is 0.943, but it decreases to 0.761 when $\theta = 0.5$. Secondly, the GLS method does not perform well in general, as it yields substantially lower coverage rates than the nominal level in all cases. Thirdly, the CS approach performs well when the strength of cross-sectional dependence is weak but does not work as effectively as the strength of cross-sectional dependence increases. For example, when $(N, T) = (196, 150)$ and $\theta = 0.3$, the ECP of the CS approach is 0.891. However, this ECP decreases to 0.818 when $\theta = 0.5$, indicating that the CS approach may not be robust enough to handle stronger cross-sectional dependence.

The proposed SHAC procedure demonstrates good performance in the presence of cross-sectional correlation, and its accuracy improves with increasing sample size. For instance, when $(N, T) = (144, 100)$ and $\theta = 0.5$, the ECP of $\text{SHAC}^{\dagger}$ is 0.905, which shows a substantial improvement compared to the LS method with an ECP of 0.812. The coverage rate of our method increases further to 0.917 when $(N,T) = (196,100)$. Furthermore, the results clearly indicate that the proposed procedure significantly improves the accuracy of inference for the LS estimator by correcting its bias. For instance, when $(N,T) = (196,100)$ and $\theta = 0.5$, SHAC_1 improves the ECP of LS from 0.811 to 0.901. The ECP of LS can be further enhanced by employing our cross-sectional dependence-robust variance estimator. For instance, when $(N,T) = (196,100)$ and $\theta = 0.5$, $\text{SHAC}^\dagger$ improves the ECP of SHAC_1 from 0.901 to 0.917. In addition, our proposed approach outperforms the CS approach in the presence of cross-sectional dependence. For example, when $(N,T) = (144,150)$ and $\theta = 0.5$, the ECP of our approach and the CS approach are 0.910 and 0.827, respectively, demonstrating the superiority of our SHAC procedure in handling cross-sectional dependence and improving inference accuracy.

We provide further simulation results in the Appendix A.

6 Empirical Application

In this section, we 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 rate affected by the divorce law reform 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

In the 1970s, about three quarters of states in the U.S. shifted from a consent divorce regime to no-fault unilateral divorce laws. The new laws allow people to seek a divorce without the consent of their spouse. An interesting empirical question in this regard is to understand the causal relationship between the divorce law reform and the divorce rate increase. Peters (1986) and Allen (1992) use the same cross-section data in 1979, but they reach different empirical results: the former concludes divorce rates are unaffected by the switch to the unilateral law, while the latter finds its significant impact.

Friedberg (1998) employs the standard fixed effects approach to investigate this relation. After controlling for state and year effects as well as state-specific time trends, she finds the state law reform has contributed to about one-sixth of the rise in the state level divorce rate since the late 1960s. Wolfers (2006) focuses on the longer effects by considering the following model

$$Y_{st} = DR_{st} + f(v_s, t) + u_{st}, \quad u_{st} = \delta_s + \alpha_t + \varepsilon_{st}, \tag{6.1}$$

where Y_{st} is the number of new divorces per thousand people in state s and year t, DR_{st} is the treatment effect of divorce law reform, and $f(v_s, t)$ is the state-specific time trend. For example, we have $f(v_s, t) = v_s t$ to represent the linear trend. u_{st} contains the state and time effects δ_s and α_t additively. The treatment effects DR_{st} is formulated as

$$DR_{st} = \beta_1 1(T_s \leq t \leq T_s + 1) + \beta_2 1(T_s + 2 \leq t \leq T_s + 3) \\ + \cdots + \beta_8 1(T_s + 14 \leq t), \tag{6.2}$$

where T_s is the law reform year of state s and $1(\cdot)$ is the indicator function.

An issue about the robustness of the LS estimation of this additive fixed effects model in Wolfers (2006) has been raised. The concern is that u_{st} consists of many missing social and cultural variables (e.g., the stigma of divorce, religious belief, and female participation in the workforce), and the additive fixed effects model may not be flexible enough to capture unobserved heterogeneity that may evolve over time and across states.

To tackle this, Kim and Oka (2013) propose using the IFE model with

$$u_{st} = \delta_s + \alpha_t + \lambda_s' F_t + \varepsilon_{st}, \tag{6.3}$$

which can effectively accommodate the remaining unobserved heterogeneity in u_{st}. They estimate the model using the LS method without taking the cross-sectional dependence into account. Bai and Liao (2017) re-estimate this model using their GLS method to improve the efficiency in the presence of cross-sectional dependence.

We employ the proposed method to make inference based on this IFE model. We set the number of factors equal to ten as Kim and Oka (2013) and Bai and Liao (2017). To construct our bias and variance estimators, we use the proposed distance measure and bootstrap method to select bandwidth parameters.

Table 2 reports the estimates of $(\beta_1, ..., \beta_8)$ and their standard errors. We consider the additive fixed effects model as well as the IFE model. For the latter, we estimate the model using the LS, the GLS, and our procedure. All the models include state-specific time linear trends. The table shows that all the estimates by the additive fixed effects model in the first column are positive but not significant, which are the same as Kim and Oka (2013). We can also see that all the estimation results based on the IFE model suggest the contribution of the law reform to the rise of the divorce rate for the first six years after the reform is significant. However, our method yields an insignificant estimate for the 7-8 years, which is different from the other two methods. The proposed method tends to produce smaller estimates with larger standard errors than the LS method by taking the cross-sectional correlation into account. The GLS tends to produce the smallest standard errors.

6.2 Effects of water and sewerage interventions

An essential question in public health is the cause of the sharp decrease in infant mortality between the late 19th century and the early 20th century. To answer this question, Cutler and Miller (2005) study the impact of water chlorination and filtration on the death rates from waterborne diseases across

13 U.S. cities. Their results suggest that improved water quality decreased 47 percent in log infant mortality in 1900-1936.

On the other hand, many U.S. metropolitan areas installed effective sewerage systems during that time, which may also have resulted in the child mortality decline. Alsan and Goldin (2019) study the independent and combined effects of clean water and effective sewerage systems on under-5 mortality in Massachusetts in 1880-1920. Their data are annual and include 60 municipalities in Massachusetts for the period that predates national mortality statistics. For empirical strategy, they employ the additive fixed effects panel model, which suggests that the two interventions together account for approximately one third of the decline in log child mortality during this period. Their regression model is formulated as follows:

$$\begin{aligned} Y_{it} &= \mu + \beta_1 W_{it} + \beta_2 S_{it} + \beta_3 (W * S)_{it} + \theta X_{it} + u_{it}, \\ u_{it} &= \alpha_i + f_t + \delta_i t + \varepsilon_{it}, \end{aligned} \tag{6.4}$$

where Y_{it} is the under-5 mortality rate for the municipality i in year t; W_{it} and S_{it} are indicator variables that equal to one if the municipality i had adopted the safe water and sewerage interventions by year t, respectively; X_{it} is a vector of time and municipality varying demographic controls including the log of population density, percentage of the foreign-born, percentage of males, and the percentage of females employed in manufacturing. u_{it} captures unobserved components including municipality and time fixed effects and municipality-specific time trend.

We introduce the IFE specification in u_{it} to accommodate various patterns of unobserved heterogeneity. u_{it} is modeled as

$$u_{it} = \lambda_i' F_t + \varepsilon_{it}, \tag{6.5}$$

where F_t is a vector of common factors that dominate the portion of infant mortality rates not explained by the regressors, and the loading vector λ_i represents the heterogeneous responses to F_t from the municipality i. Note that if we let $\lambda_i = (\alpha_i, 1, \delta_i)'$ and $F_t = (1, f_t, t)'$, then u_{it} in (6.5) reduces to the additive fixed effects specification with the time trend presented in (6.4). Hence, we choose three factors in the IFE model to include the original model by Alsan and Goldin (2019) as a special case.

We make inference on the IFE model based on (i) the LS method without accounting for cross-sectional dependence, (ii) the GLS method, and (iii) the proposed procedure. We use the same data set as Alsan and Goldin (2019), which contains the under-5 mortality rates, municipality-level water and sewerage interventions, and demographic control regressors from 1981 to 1920 over 60 municipalities. To construct a balanced panel, we drop the data of Westwood which contains many missing observations. We interpolate the missing observations of Wellesley in 1980 and 1981 using the observations in 1982, and the missing values of under-5 child mortality of Weston in 1904 by taking the average of the values in 1903 and 1905. The same way of interpolation for missing values in 1917 with the values in 1916 and 1918. As a result, we have a balanced panel that consists of 59 municipalities and 41 years.

The estimation results are summarized in Table 3. Panel A reports the estimates based on the additive fixed effects model with the linear trend as Alsan and Goldin (2019). Panels B-D present the estimates on the IFE model. Panel B is based on the LS method. Comparing the results from Panels A and B, we can see that the independent and combined effects of clean water and effective sewerage system on under-5 mortality in Panel B tend to be smaller than the ones from the additive fixed effects model in Panel A. For example, while the combination of sewerage and safe water treatments in column (5) decreases under-5 mortality by 13.530 percent in Panel A, it decreases to 11.595 percent in Panel B.

Panel C reports the coefficients based on the proposed procedure. Comparing between Panel B and Panel C, we see that the effect of safe water in column (5) becomes statistically insignificant at 10% level when we take cross-sectional dependence into account. This is consistent with our expectation since safe water is unlikely to raise the child mortality even without the effect sewage system. The GLS estimates appear in Panel D. The GLS yields smaller estimates and standard errors than the other methods.

7 Conclusion

This paper considers inference on the panel regression model with interactive fixed effects. Under large N and large T asymptotics, the LS estimator is asymptotically biased and the usual standard error is invalid when the idiosyncratic errors are cross-sectionally dependent. We propose an inference procedure that addresses this problem. We first develop a bias correction for the LS estimator and then propose a cross-sectional dependence robust variance estimator to construct the associated test statistics. We also propose a distance measure and bootstrap method to select bandwidth parameters which are necessary to implement our method.

Statements and Declarations

Not applicable.

References

Ahn, S. and A. Horenstein. 2001. Gmm estimation of linear panel data models with time varying individual effects. *Journal of Econometrics 101*(2): 219–255. https://doi.org/http://dx.doi.org/10.1016/s0304-4076(00)00083-x .

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

Alsan, M. and C. Goldin. 2019. Watersheds in child mortality: the role of effective water and sewerage infrastructure, 1880–1920. *Journal of Political Economy 127*(2): 586–638. https://doi.org/https://doi.org/10.1086/700766 .

Andrews, D. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*(3): 817–858. https://doi.org/https://doi.org/10.2307/2938229 .

Bai, J. 2009. Panel data models with interactive fixed Effects. *Econometrica 77*(4): 1229–1279. https://doi.org/https://doi.org/10.3982/ecta6135 .

Bai, J. and Y. Liao. 2017. Inferences in panel data with interactive effects using large covariance matrices. *Journal of Econometrics 200*(1): 59–78. https://doi.org/https://doi.org/10.1016/j.jeconom.2017.05.014 .

Bai, J. and S. Ng. 2006. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica 74*(4): 1133–1150. https://doi.org/https://doi.org/10.1111/j.1468-0262.2006.00696.x .

Callaway, B. and S. Karami. 2023. Treatment effects in interactive fixed effects models with a small number of time periods. *Journal of Econometrics 233*(1): 184–208. https://doi.org/10.1016/j.jeconom.2022.02.001 .

Conley, T. 1999. GMM estimation with cross sectional dependence. *Journal of Econometrics 92*(1): 1–45. https://doi.org/https://doi.org/10.1016/s0304-4076(98)00084-0 .

Cui, G., M. Norkutė, V. Sarafidis, and T. Yamagata. 2021. Two-stage instrumental variable estimation of linear panel data models with interactive effects. *The Econometrics Journal 25*(2): 340–361. https://doi.org/https://doi.org/10.1093/ectj/utab029 .

Cutler, D. and G. Miller. 2005. The role of public health improvements in health advances: The twentieth-century United States. *Demography 42*(1): 1–22. https://doi.org/https://doi.org/10.1353/dem.2005.0002 .

Driscoll, J. and A. Kraay. 1998. Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics 80*(4): 549–560. https://doi.org/https://doi.org/10.1162/003465398557825 .

Fan, J., Y. Liao, and M. Mincheva. 2013. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(4): 603–680. https://doi.org/https://doi.org/10.1111/rssb.12016 .

Fernandez, V. 2011. Spatial linkages in international financial markets. *Quantitative Finance 11*(2): 237–245. https://doi.org/https://doi.org/10.1080/14697680903127403 .

Friedberg, L. 1998. Did unilateral divorce raise divorce rates? evidence from panel data. *American Economic Review* 88: 608–627. https://doi.org/https://doi.org/10.3386/w6398 .

Holtz-Eakin, D., W. Newey, and H.S. Rosen. 1988. Estimating vector autoregressions with panel data. *Econometrica 56*(6): 1371. https://doi.org/https://doi.org/10.2307/1913103 .

Hong, S., L. Su, and T. Jiang. 2023. Profile GMM estimation of panel data models with interactive fixed effects. *Journal of Econometrics 235*(2): 927–948. https://doi.org/10.1016/j.jeconom.2022.07.010 .

Kelejian, H.H. and I.R. Prucha. 2007. HAC estimation in a spatial framework. *Journal of Econometrics 140*(1): 131–154. https://doi.org/https://doi.org/10.1016/j.jeconom.2006.09.005 .

Kim, D. and T. Oka. 2013. Divorce law reforms and divorce rates in the USA: An interactive fixed-effects approach. *Journal of Applied Econometrics 29*(2): 231–245. https://doi.org/https://doi.org/10.1002/jae.2310 .

Kim, M.S. 2021. Robust inference for diffusion-index forecasts with cross-sectionally dependent data. *Journal of Business and Economic Statistics*: 1–15. https://doi.org/https://doi.org/10.1080/07350015.2021.1906258 .

Kim, M.S. and Y. Sun. 2011. Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix. *Journal of Econometrics 160*(2): 349–371. https://doi.org/https://doi.org/10.1016/j.jeconom.2010.10.002 .

Kim, M.S., Y. Sun, and J. Yang. 2017. A fixed-bandwidth view of the pre-asymptotic inference for kernel smoothing with time series data. *Journal of Econometrics 197*(2): 298–322. https://doi.org/https://doi.org/10.1016/j.jeconom.2016.11.008 .

Ligon, E.A. and T.G. Conley. 2002. Economic distance and cross-Country spillovers. *Journal of Economic Growth* 7: 157–187. https://doi.org/https://doi.org/10.1023/a:1015676113101 .

Mantegna, R. 1999. Hierarchical structure in financial markets. *The European Physical Journal B 11*(1): 193–197. https://doi.org/https://doi.org/10.1007/s100510050929 .

Moon, H.R. and M. Weidner. 2017. Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory 33*(1): 158–195. https://doi.org/https://doi.org/10.1017/s0266466615000328 .

Newey, W. and K. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3): 703–708. https://doi.org/https://doi.org/10.2307/1913610 .

Neyman, J. and E.L. Scott. 1948. Consistent estimates based on partially consistent observations. *Econometrica 16*(1): 1. https://doi.org/https://doi.org/10.2307/1914288 .

Nickell, S. 1981. Biases in dynamic models with fixed effects. *Econometrica 49*(6): 1417. https://doi.org/https://doi.org/10.2307/1911408 .

Pesaran, M.H. 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4): 967–1012. https://doi.org/https://doi.org/10.1111/j.1468-0262.2006.00692.x .

Pesaran, M.H. and E. Tosetti. 2011. Large panels with common factors and spatial correlation. *Journal of Econometrics 161*(2): 182–202. https://doi.org/https://doi.org/10.1016/j.jeconom.2010.12.003 .

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

Pinkse, J., M.E. Slade, and C. Brett. 2002. Spatial price competition: a semiparametric approach. *Econometrica 70*(3): 1111–1153. https://doi.org/https://doi.org/10.1111/1468-0262.00320 .

Politis, D. 2011. Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices. *Econometric Theory 27*(4): 703–744. https://doi.org/https://doi.org/10.1017/S0266466610000484 .

Robinson, P. 2011. Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics 165*(1): 5–19. https://doi.org/https://doi.org/10.1016/j.jeconom.2011.05.002 .

Vogelsang, T. 2012. Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects. *Journal of Econometrics 166*(2): 303–319. https://doi.org/https://doi.org/10.1016/j.jeconom.2011.10.001 .

Wolfers, J. 2006. Did unilateral divorce laws raise divorce rates? a reconciliation and new results. *American Economic Review 96*(5): 1802–1820.

https://doi.org/http://dx.doi.org/10.1257/aer.96.5.1802 .

Table 1 Empirical Coverage Rates of 95% Confidence Interval.

T	N	LS	$SHAC_1$	$SHAC_2$	$SHAC^{\dagger}$	GLS	CS
				$\theta = 0$			
100	144	0.942	0.941	0.942	0.942	0.386	0.942
100	196	0.951	0.951	0.952	0.952	0.321	0.952
150	144	0.943	0.944	0.942	0.943	0.388	0.943
150	196	0.944	0.944	0.943	0.943	0.309	0.945
				$\theta = .3$			
100	144	0.881	0.927	0.882	0.928	0.549	0.913
100	196	0.884	0.932	0.886	0.934	0.477	0.918
150	144	0.823	0.908	0.826	0.908	0.525	0.878
150	196	0.840	0.917	0.845	0.919	0.426	0.891
				$\theta = .5$			
100	144	0.812	0.884	0.837	0.905	0.708	0.869
100	196	0.811	0.901	0.833	0.917	0.630	0.875
150	144	0.761	0.900	0.776	0.910	0.688	0.827
150	196	0.738	0.899	0.757	0.908	0.566	0.818

Note: LS is the LS estimator, GLS is the GLS estimator, and CS is the partial sample estimator. $SHAC_1$ is the bias corrected estimator with HR variance estimator. $SHAC_2$ is the LS estimator with cross-sectional dependence robust variance estimator. $SHAC^{\dagger}$ is the bias corrected estimator with cross-sectional dependence robust variance estimator.

Table 2 Effects of divorce law reform on divorce rate

	Additive		LS		$SHAC^{\dagger}(\hat{d}_{ij}^{D})$		GLS	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
First 2 years	0.095	0.096	0.108***	0.035	0.101**	0.044	0.088***	0.031
3–4 years	0.159	0.108	0.228***	0.050	0.213***	0.056	0.209***	0.043
5–6 years	0.091	0.121	0.193***	0.068	0.174***	0.066	0.171***	0.059
7–8 years	0.157	0.132	0.168*	0.088	0.138	0.091	0.167**	0.076
9–10 years	0.067	0.143	0.097	0.103	0.062	0.102	0.099	0.089
11–12 years	0.052	0.153	0.060	0.116	0.021	0.118	0.047	0.100
13–14 years	0.093	0.167	0.033	0.128	-0.005	0.134	0.042	0.111
15 years+	0.222	0.186	0.117	0.143	0.079	0.150	0.101	0.123

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

Table 3 Effects of clean water and sewerage on child mortality

	Panel A. Additive					Panel B. LS				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Safe water	-5.012		-3.694		5.555	-1.543		-1.841		5.580*
	(3.373)		(3.271)		(4.022)	(1.905)		(1.846)		(3.301)
Sewerage		-6.251***	-5.613**		-3.922		-3.987***	-3.934***		-2.935**
		(2.792)	(2.741)		(2.696)		(1.415)	(1.442)		(1.468)
Interaction				-10.201***	-13.530***				-7.347***	-11.595***
				(3.431)	(4.835)				(1.903)	(3.713)
	Panel C. SHAC†($\hat{d}_{ij}^{D}$)					Panel D. GLS				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Safe water	-1.512		-1.942		5.111	-1.278		-2.418		3.126
	(1.902)		(1.661)		(3.340)	(1.667)		(1.560)		(3.064)
Sewerage		-4.020**	-3.954**		-2.997*		-3.220**	-3.092**		-2.701**
		(1.651)	(1.695)		(1.651)		(1.263)	(1.291)		(1.307)
Interaction				-7.282***	-11.058***				-5.608***	-7.698**
				(1.887)	(3.854)				(1.631)	(3.378)

Note:
(i) Standard Errors are reported in the parentheses.
(ii) Interaction: interaction of safe water and sewerage.
(iii) * $p < .1$. ** $p < .05$. *** $p < .01$.

Appendix A Monte Carlo

In this section, we present additional simulation results on the finite sample performance of the proposed procedure. We also consider the LS estimator, the GLS estimator, and the CS estimator for the purpose of comparison. We examine their performance in terms of the bias and root mean square error (RMSE). The DGP is the same as the one used in Section 5.

The simulation results are summarized in Table A1. When there is no cross-sectional correlation ($\theta = 0$), all the estimators perform very well, exhibiting very small biases and root mean square errors (RMSEs). However, in the presence of cross-sectional dependence, several patterns emerge. Firstly, the bias and RMSE of the LS estimator increase as the strength of cross-sectional dependence grows. For example, when $(N, T) = (144, 150)$ and $\theta = 0.3$, the bias and RMSE of the LS estimator are -0.0077 and 0.0111, respectively. These increase to -0.0105 and 0.0166, respectively, when $\theta = 0.5$. Secondly, the bias and RMSE of the LS estimator decrease with an increase in the sample size. For instance, in the case of $(N, T) = (144, 100)$ and $\theta = 0.3$, the bias and RMSE of $\hat{\beta}$ are -0.0069 and 0.0120, respectively. These decrease to -0.0059 and 0.0102, respectively, with $(N, T) = (196, 100)$. Thirdly, the proposed bias-corrected estimator tends to produce smaller bias and RMSE than the LS estimator. For example, when $(N, T) = (196, 150)$ and $\theta = 0.5$, the bias and RMSE of $\hat{\beta}$ are -0.0102 and 0.0146, respectively, and the proposed method reduces the bias and RMSE to -0.0032 and 0.0109, respectively.

We also compare the performances of the proposed bias-corrected estimator with the CS estimator and the GLS estimator. The table shows that the CS estimator tends to produce smaller bias and RMSE than the LS estimator in the presence of cross-sectional dependence. In comparison, the proposed estimator and the GLS estimator outperform the CS estimator by yielding substantially smaller bias and RMSE. For instance, when $(N, T) = (144, 150)$ and $\theta = 0.5$, the bias and RMSE of the LS estimator are -0.0105 and 0.0166. The CS estimator decreases the bias and RMSE to -0.0079 and 0.0151, respectively, whereas $\hat{\beta}_{GLS}$ reduces the bias and RMSE to -0.0030 and 0.0123, respectively, and $\hat{\beta}^{\dagger}(\hat{d}_{ik}^{D})$ decreases the bias and RMSE to -0.0028 and 0.0132, respectively.

Table A1 Performances of estimated β; true $\beta = 0$.

		$\hat{\beta}$		$\hat{\beta}^{\dagger}(\hat{d}_{ik}^{D})$		$\hat{\beta}_{GLS}$		$\hat{\beta}_{CS}$	
T	N	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
					$\theta = 0$				
100	144	0.0006	0.0084	0.0006	0.0084	0.0065	0.0315	0.0006	0.0084
100	196	0.0003	0.0072	0.0003	0.0072	0.0069	0.0329	0.0003	0.0072
150	144	0.0003	0.0069	0.0003	0.0069	0.0036	0.0266	0.0003	0.0069
150	196	0.0003	0.0060	0.0003	0.0060	0.0043	0.0264	0.0003	0.0060
					$\theta = .3$				
100	144	-0.0069	0.0120	-0.0032	0.0105	0.0015	0.0254	-0.0050	0.0111
100	196	-0.0059	0.0102	-0.0028	0.0089	0.0027	0.0260	-0.0042	0.0094
150	144	-0.0077	0.0111	-0.0041	0.0090	-0.0011	0.0194	-0.0057	0.0098
150	196	-0.0062	0.0092	-0.0033	0.0076	0.0003	0.0186	-0.0046	0.0083
					$\theta = .5$				
100	144	-0.0061	0.0263	0.0017	0.0251	-0.0021	0.0163	-0.0036	0.0257
100	196	-0.0078	0.0179	-0.0007	0.0160	-0.0009	0.0164	-0.0053	0.0169
150	144	-0.0105	0.0166	-0.0028	0.0132	-0.0030	0.0123	-0.0079	0.0151
150	196	-0.0102	0.0146	-0.0032	0.0109	-0.0016	0.0116	-0.0078	0.0130

Note: $\hat{\beta}$ is the LS estimator, $\hat{\beta}_{GLS}$ is the GLS estimator, and $\hat{\beta}_{CS}$ is the partial sample estimator. $\hat{\beta}^{\dagger}(\hat{d}_{ik}^{D})$ is our bias corrected estimators using the proposed distance measure.

Appendix B Proofs

Proof of Theorem 3.1: We define $\delta_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$. Based on the asymptotic result of the LS estimator by Bai (2009), we can complete the proof of this theorem by showing the consistency of the bias estimator $\hat{B}$. i.e., $\hat{B} - B_{NT} = o_p(1)$.

Since $\hat{B} = -\hat{D}^{-1}\hat{J}$ and the consistency of $\hat{D}$ is established by Bai (2009), it is sufficient to prove the consistency of $\hat{J}$. Note that $\hat{J} - J_{NT} = o_p(1)$ if and only if $a'\hat{J}a - a'J_{NT}a = o_p(1)$ for any $a \in \mathbf{R}^p$. Therefore, we assume $\hat{J}$ is a scalar ($p = 1$) without loss of generality.

We set

$$\hat{J} - J_{NT} = \left(E(\tilde{J}) - J_{NT}\right) + \left(\tilde{J} - E(\tilde{J})\right) + \left(\hat{J} - \tilde{J}\right), \tag{B1}$$

where $\tilde{J} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} K_1\left(\frac{d_{ik}}{d_{(1)}}\right) w_i \lambda_k \varepsilon_{it}\varepsilon_{kt}$ is the infeasible estimator. We show each term in (B1) converges to zero in probability.

(a) $E(\tilde{J}) - J_{NT} = O\left(\frac{1}{d_{(1)}^{q_1}}\right)$

Note that we have

$$\begin{aligned}\Big\|E(\tilde J)-J_{NT}\Big\| &\le \frac{1}{d_{(1)}^{q_1}}\Big(\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\|w_i\|\|\lambda_k\|\|\sigma_{ik,t}\|d_{ik}^{q_1}\Big)\left|\frac{1-K_1\Big(\frac{d_{ik}}{d_{(1)}}\Big)}{\Big(\frac{d_{ik}}{d_{(1)}}\Big)^{q_1}}\right| \\ &\le \frac{K_1^{(q_1)}}{d_{(1)}^{q_1}}\Big(\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\|w_i\|\|\lambda_k\|\|\sigma_{ik,t}\|d_{ik}^{q_1}\Big)+o(1) \\ &= O\Big(\frac{1}{d_{(1)}^{q_1}}\Big),\end{aligned}$$

if $d_{(1)}\to\infty$ as $N,T\to\infty$.

(b) $\tilde J-E(\tilde J)=O_p\Big(\sqrt{\frac{\bar\ell_{N(1)}}{NT}}\Big)$

By Chebyshev's inequality, we have

$$P(|\tilde J-E(\tilde J)|>\Delta)\le\frac{1}{\Delta^2}E(\tilde J-E(\tilde J))^2.$$

Thus, it is sufficient to show that $E\Big(\tilde J-E(\tilde J)\Big)^2=o(1)$. We can write

$$\begin{aligned}E\Big(\tilde J-E(\tilde J)\Big)^2 &= \frac{1}{N^2T^2}\sum_{i,k=1}^{N}\sum_{a,b=1}^{N}\sum_{s,t=1}^{T}K_1\Big(\frac{d_{ik}}{d_{(1)}}\Big)K_1\Big(\frac{d_{ab}}{d_{(1)}}\Big)(w_i\lambda_k)(w_a\lambda_b) \\ &\times\big[\{E\varepsilon_{it}\varepsilon_{kt}\varepsilon_{as}\varepsilon_{bs}-E(\varepsilon_{it}\varepsilon_{kt})E(\varepsilon_{as}\varepsilon_{bs})-E(\varepsilon_{it}\varepsilon_{as})E(\varepsilon_{bs}\varepsilon_{kt}) \\ &\quad -E(\varepsilon_{it}\varepsilon_{bs})E(\varepsilon_{as}\varepsilon_{kt})\}+E(\varepsilon_{it}\varepsilon_{as})E(\varepsilon_{bs}\varepsilon_{kt})+E(\varepsilon_{it}\varepsilon_{bs})E(\varepsilon_{as}\varepsilon_{kt})\big] \\ &:= A_1+A_2+A_3.\end{aligned}$$

For A_1, we use the linear representation of ε_{it} in Assumption 3.1 to have

$$\begin{aligned}&E\varepsilon_{it}\varepsilon_{kt}\varepsilon_{as}\varepsilon_{bs}-E(\varepsilon_{it}\varepsilon_{kt})E(\varepsilon_{as}\varepsilon_{bs})-E(\varepsilon_{it}\varepsilon_{as})E(\varepsilon_{bs}\varepsilon_{kt})-E(\varepsilon_{it}\varepsilon_{bs})E(\varepsilon_{as}\varepsilon_{kt}) \\ &=\sum_{l=1}^{\infty}\gamma_{il}\gamma_{kl}\gamma_{al}\gamma_{bl}(Ee_{lt}^4-3).\end{aligned}$$

Under the moment condition for e_{lt} and the summability conditions in Assumption 3.1, we have

$$\|A_1\| \leq \frac{1}{N^2T^2}\sum_{i,k=1}^{N}\sum_{a,b=1}^{N}\sum_{t=1}^{T}\sum_{l=1}^{\infty}K_1\Big(\frac{d_{ik}}{d_{(1)}}\Big)K_1\Big(\frac{d_{ab}}{d_{(1)}}\Big)\|(w_i\lambda_k)(w_a\lambda_b)\|\,|\gamma_{il}\gamma_{kl}\gamma_{al}\gamma_{bl}||Ee_{lt}^4-3|$$

$$\leq O\Big(\frac{1}{NT}\Big)\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\underbrace{\Big(\sum_{l=1}^{\infty}\gamma_{il}\Big)}_{\leq M}\underbrace{\Big(\sum_{k=1}^{N}\gamma_{kl}\Big)}_{\leq M}\underbrace{\Big(\sum_{a=1}^{N}\gamma_{al}\Big)}_{\leq M}\underbrace{\Big(\sum_{b=1}^{N}\gamma_{bl}\Big)}_{\leq M}$$

$$= O\Big(\frac{1}{NT}\Big).$$

For A_2,

$$\|A_2\| \leq \frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{a=1}^{N}\sum_{k\in\{d_{ik}\leq d_{(1)}\}}\sum_{b\in\{d_{ab}\leq d_{(1)}\}}\|(w_i\lambda_k)(w_a\lambda_b)\||E(\varepsilon_{it}\varepsilon_{at})||E(\varepsilon_{kt}\varepsilon_{bt})|$$

$$\leq O\Big(\frac{\bar{\ell}_{N(1)}}{NT}\Big)\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{1}{\bar{\ell}_{N(1)}}\sum_{k\in\{d_{ik}\leq d_{(1)}\}}\Big(\sum_{l=1}^{\infty}\gamma_{il}\Big)\Big(\sum_{a=1}^{N}\gamma_{al}\Big)\Big(\sum_{f=1}^{\infty}\gamma_{kf}\Big)\Big(\sum_{b=1}^{N}\gamma_{bf}\Big)$$

$$= O\Big(\frac{\bar{\ell}_{N(1)}}{NT}\Big).$$

Using the same argument, it is easy to show $A_3 = O\Big(\frac{\bar{\ell}_{N(1)}}{NT}\Big)$. Combining all the results, we have

$$E\Big(\tilde{J}-E(\tilde{J})\Big)^2 = O\left(\frac{1}{NT}\right)+O\left(\frac{\bar{\ell}_{N(1)}}{NT}\right),$$

which implies

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

as $\bar{\ell}_{N(1)}/NT \to 0$.

(c) $\hat{J}-\tilde{J} = o_p(1)$

Since

$$\hat{J}-\tilde{J} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\left(\frac{(X_i-\hat{V}_i)'\hat{F}}{T}\left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\right)^{-1}\hat{\lambda}_k\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\right.$$
$$\left.-\frac{(X_i-V_i)'F^0}{T}\left(\frac{F^{0\prime}F^0}{T}\right)^{-1}\left(\frac{\Lambda'\Lambda}{N}\right)^{-1}\lambda_k\varepsilon_{it}\varepsilon_{kt}\right)K_1\Big(\frac{d_{ik}}{d_{(1)}}\Big),$$

we need to prove

$$C_1 = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\Bigg(\frac{X_i'\hat{F}}{T}\Big(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\Big)^{-1}\hat{\lambda}_k\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} - \frac{X_i'F^0}{T}\Big(\frac{F^{0\prime}F^0}{T}\Big)^{-1}\Big(\frac{\Lambda'\Lambda}{N}\Big)^{-1}\lambda_k\varepsilon_{it}\varepsilon_{kt}\Bigg)K_1\Big(\frac{d_{ik}}{d_{(1)}}\Big) = o_p(1),$$

and

$$C_2 = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\Bigg(\frac{\hat{V}_i'\hat{F}}{T}\Big(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\Big)^{-1}\hat{\lambda}_k\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} - \frac{V_i'F^0}{T}\Big(\frac{F^{0\prime}F^0}{T}\Big)^{-1}\Big(\frac{\Lambda'\Lambda}{N}\Big)^{-1}\lambda_k\varepsilon_{it}\varepsilon_{kt}\Bigg)K_1\Big(\frac{d_{ik}}{d_{(1)}}\Big) = o_p(1).$$

We prove $C_1 = o_p(1)$ using the identity $\hat{a}\hat{b}\hat{c}\hat{d} - abcd = (\hat{a}-a)\hat{b}\hat{c}\hat{d} + a(\hat{b}-b)\hat{c}\hat{d} + ab(\hat{c}-$ c) $\hat{d} + abc(\hat{d}-d)$. Let $H = (\Lambda'\Lambda/N)(F^{0\prime}\hat{F}/T)V_{NT}^{-1}$, where V_{NT} is defined in (2.8). For the first corresponding term,

$$\begin{aligned}
&\Bigg\|\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{X_i'(\hat{F}-F^0H)}{T}\hat{\varepsilon}_{it}\Big(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\Big)^{-1}\hat{\lambda}_k\hat{\varepsilon}_{kt}K_1\Big(\frac{d_{ik}}{d_{(1)}}\Big)\Bigg\| \\
&\leq \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{N}\Bigg\|\frac{X_i'\big(\hat{F}-F^0H\big)}{T}\hat{\varepsilon}_{it}\Bigg\|\Bigg\|\Big(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\Big)^{-1}\Bigg\|\Big\|\hat{\lambda}_k\hat{\varepsilon}_{kt}\Big\|\Big|K_1\Big(\frac{d_{ik}}{d_{(1)}}\Big)\Big| \\
&\leq \frac{\bar{\ell}_{N(1)}}{T}\sum_{t=1}^{T}\Bigg(\frac{1}{N}\sum_{i=1}^{N}\Bigg\|\frac{X_i'\big(\hat{F}-F^0H\big)}{T}\Bigg\|^2\Bigg)^{1/2}\Big(\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{it}^4\Big)^{1/4}\Bigg\|\Big(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\Big)^{-1}\Bigg\| \\
&\times\Bigg(\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1}{\bar{\ell}_{N(1)}}\sum_{k\in\{d_{ik}\leq d_{(1)}\}}\big\|\hat{\lambda}_k\hat{\varepsilon}_{kt}\big\|\Big)^4\Bigg)^{1/4} \\
&= O_p\Big(\frac{\bar{\ell}_{N(1)}}{\sqrt{NT}}\Big) + O_p\Big(\frac{\bar{\ell}_{N(1)}}{\delta_{NT}^2}\Big),
\end{aligned}$$

because we have

$$\frac{1}{N}\sum_{i=1}^{N}\Bigg\|\frac{X_i'\big(\hat{F}-F^0H\big)}{T}\Bigg\|^2 = O_p(\hat{\beta}-\beta) + O_p\Big(\frac{1}{\delta_{NT}^2}\Big).$$

due to Lemma A.3 of Bai (2009).

The second corresponding term is

$$\begin{aligned}
&\left\| \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{N} \hat{\varepsilon}_{it} \frac{X_i' F^0}{T} \left(\left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} - H' \left(\frac{\Lambda'\Lambda}{N} \right)^{-1} H \right) \hat{\lambda}_k \hat{\varepsilon}_{kt} K_1 \left(\frac{d_{ik}}{d_{(1)}} \right) \right\| \\
&\leq \frac{\bar{\ell}_{N(1)}}{T} \sum_{t=1}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \frac{X_i' F^0}{T} \hat{\varepsilon}_{it} \right\|^2 \right)^{1/2} \left(\frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{\bar{\ell}_{N(1)}} \sum_{k \in \{d_{ik} \leq d_{(1)}\}} \|\hat{\lambda}_k \hat{\varepsilon}_{kt}\| \right)^2 \right)^{1/2} \\
&\times \left\| \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} - H' \left(\frac{\Lambda'\Lambda}{N} \right)^{-1} H \right\| \\
&= O_p\left(\frac{\bar{\ell}_{N(1)}}{\sqrt{NT}} \right) + O_p\left(\frac{\bar{\ell}_{N(1)}}{\delta_{NT}^2} \right),
\end{aligned}$$

where we use

$$\left\| \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N} \right)^{-1} - H' \left(\frac{\Lambda'\Lambda}{N} \right)^{-1} H \right\| = O_p\left(\left\| \hat{\beta} - \beta \right\| \right) + O_p\left(\frac{1}{\delta_{NT}^2} \right).$$

under Lemma A.10 of Bai (2009).

Similarly,

$$\begin{aligned}
&\left\| \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{N} \hat{\varepsilon}_{it} \frac{X_i' F^0}{T} \left(\frac{\Lambda'\Lambda}{N} \right)^{-1} (\hat{\lambda}_k - H^{-1}\lambda_k) \hat{\varepsilon}_{kt} K_1 \left(\frac{d_{ik}}{d_{(1)}} \right) \right\| \\
&= O_p\left(\frac{\bar{\ell}_{N(1)}}{\sqrt{NT}} \right) + O_p\left(\frac{\bar{\ell}_{N(1)}}{\delta_{NT}^2} \right),
\end{aligned}$$

and

$$\begin{aligned}
&\left\| \frac{1}{N\sqrt{NT}} \sum_{i=1}^{N} \sum_{k=1}^{N} \lambda_i' \left(\Lambda'\Lambda/N \right)^{-1} H(\hat{\lambda}_k - H^{-1}\lambda_k) X_k \varepsilon_{it} \right\| \\
&= O_p\left(\frac{\bar{\ell}_{N(1)}}{\sqrt{NT}} \right) + O_p\left(\frac{\bar{\ell}_{N(1)}}{\delta_{NT}^2} \right).
\end{aligned}$$

Combining all the results, we have

$$C_1 = o_p(1),$$

if $N, T \to \infty$ such that $T/N \to \rho$ and $\bar{\ell}_{N(1)}/\delta_{NT}^2 \to 0$.

Note that we can use the same argument to prove $C_2 = o_p(1)$ if we replace $X_i'\hat{F}/T$ in C_1 with $\hat{V}_i'\hat{F}/T$. Therefore, we have

$$\hat{J} - \tilde{J} = o_p(1),$$

which completes the proof of Theorem 3.1. □

Proof of Theorem 3.2: As in the proof of Theorem 3.1, we assume Ω_{NT} is a scalar ($p = 1$) without loss of generality. We have

$$\hat{\Omega} - \Omega_{NT} = (E\tilde{\Omega} - \Omega_{NT}) + (\tilde{\Omega} - E\tilde{\Omega}) + (\hat{\Omega} - \tilde{\Omega}), \tag{B2}$$

where $\tilde{\Omega}$ is an infeasible estimator of Ω_{NT} which is defined as

$$\tilde{\Omega} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} Z_{it}Z_{kt}'\varepsilon_{it}\varepsilon_{kt}K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big).$$

Using the same procedure in the proof of Theorem 3.1, we can show that $E\tilde{\Omega} - \Omega_{NT} = o(1)$ and $\tilde{\Omega} - E\tilde{\Omega} = o_p(1)$ as $d_{(2)}, N, T \to \infty$ such that $\bar{\ell}_{N(2)}/NT \to 0$. Thus, we need to prove $\hat{\Omega} - \tilde{\Omega} = o_p(1)$ to complete the proof.

$$\begin{aligned}
\hat{\Omega} - \tilde{\Omega} &= \frac{1}{T}\sum_{t=1}^{T}\Bigg(\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}(\hat{Z}_{it}\hat{Z}_{kt}\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} - Z_{it}Z_{kt}\varepsilon_{it}\varepsilon_{kt})\Bigg)K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big)\\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\hat{Z}_{it}\hat{Z}_{kt}(\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} - \varepsilon_{it}\varepsilon_{kt})K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big)\\
&\quad + \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}(\hat{Z}_{it}\hat{Z}_{kt} - Z_{it}Z_{kt})\varepsilon_{it}\varepsilon_{kt}K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big)\\
&= E_1 + E_2.
\end{aligned}$$

Applying

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

to E_1, we have

$$\begin{aligned}
E_1 &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\hat{Z}_{it}\hat{Z}_{kt}\big[X_{it}X_{kt}(\hat{\beta} - \beta)^2 + (\hat{F}_t - H'F_t^0)'H^{-1}\lambda_i(\hat{F}_t - H'F_t^0)'H^{-1}\lambda_k\\
&\quad + \hat{F}_t'(\hat{\lambda}_i - H^{-1}\lambda_i)\hat{F}_t'(\hat{\lambda}_k - H^{-1}\lambda_k) + 2\varepsilon_{it}X_{kt}'(\hat{\beta} - \beta) + 2\varepsilon_{it}(\hat{F}_t - H'F_t^0)'H^{-1}\lambda_k\\
&\quad + 2\varepsilon_{it}\hat{F}_t'(\hat{\lambda}_k - H^{-1}\lambda_k) + 2X_{it}'(\hat{\beta} - \beta)(\hat{F}_t - H'F_t^0)'H^{-1}\lambda_k + 2X_{it}'(\hat{\beta} - \beta)\hat{F}_t(\hat{\lambda}_k - H^{-1}\lambda_k)\\
&\quad + (\hat{F}_t - H'F_t^0)'H^{-1}\lambda_i\hat{F}_t(\hat{\lambda}_k - H^{-1}\lambda_k)\big]K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big)\\
&= E_{11} + \cdots + E_{19}.
\end{aligned}$$

For E_{11},

$$E_{11} \le O_p\Big(\frac{1}{T}\Big)\frac{1}{T}\sum_{t=1}^{T}\Big(\frac{1}{N}\sum_{i=1}^{N}\hat{Z}_{it}X_{it}\Big)^2 = O_p\Big(\frac{1}{T}\Big).$$

For E_{12},

$$\|E_{12}\| \leq O_p(\bar{\ell}_{N(2)}) \left(\frac{1}{T} \sum_{t=1}^{T} \left\| \hat{F}_t - H' F_t^0 \right\|^2 \right) \times$$
$$\left(\frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \hat{Z}_{it} \lambda_i \right\| \frac{1}{\bar{\ell}_{N(2)}} \sum_{k \in \{d_{ik} \leq d_2\}} \left\| \hat{Z}_{kt} \lambda_k \right\| \right)^2 \right)^{1/2}$$
$$= O_p\Big(\frac{\bar{\ell}_{N(2)}}{\delta_{NT}^2} \Big).$$

because we have $\frac{1}{T} \sum_{t=1}^{T} \|\hat{F}_t - H' F_t^0\|^2 = O_p(\|\hat{\beta} - \beta\|^2) + O_p(\delta_{NT}^{-2})$ from Proposition A.1 in Bai (2009).

For E_{13},

$$\|E_{13}\| \leq O_p\left(\sqrt{N \bar{\ell}_{N(2)}}\right) \frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \hat{Z}_{it} \hat{F}_t' \right\|^4 \right)^{1/4} \left(\frac{1}{N} \sum_{i=1}^{N} \left\| (\hat{\lambda}_i - H^{-1} \lambda_i) \right\|^4 \right)^{1/4}$$
$$\times \left(\frac{1}{N \bar{\ell}_{N(2)}} \sum_{i=1}^{N} \sum_{k=1}^{N} \left\| \hat{Z}_{kt} \hat{F}_t' K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big) \right\|^2 \right)^{1/2} \left(\frac{1}{N} \sum_{k=1}^{N} \left\| \hat{\lambda}_k - H^{-1} \lambda_k \right\|^2 \right)^{1/2}$$
$$\leq O_p\left(\sqrt{N \bar{\ell}_{N(2)}}\right) \frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \hat{Z}_{it} \hat{F}_t' \right\|^4 \right)^{1/4}$$
$$\times \left(\frac{1}{N \bar{\ell}_{N(2)}} \sum_{i=1}^{N} \sum_{k=1}^{N} \left\| \hat{Z}_{kt} \hat{F}_t' K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big) \right\|^2 \right)^{1/2} \left(\frac{1}{N} \sum_{k=1}^{N} \left\| \hat{\lambda}_k - H^{-1} \lambda_k \right\|^2 \right)$$
$$= O_p\left(\frac{\sqrt{N \bar{\ell}_{N(2)}}}{\delta_{NT}^2} \right),$$

where we use $\frac{1}{N} \sum_{i=1}^{N} \left\| (\hat{\lambda}_i - H^{-1} \lambda_i) \right\|^2 = O_p(\|\hat{\beta} - \beta\|^2) + O_p(\delta_{NT}^{-2})$ due to Lemma A.10 in Bai (2009).

Using similar procedures we can show that $E_{14} = \cdots = E_{18} = o_p(1)$ if $N, T \to \infty$ such that $T/N \to \rho$ and $\bar{\ell}_{N(2)}/\delta_{NT}^2 \to 0$. Thus, we have

$$E_1 = o_p(1)$$

if $N, T \to \infty$ such that $T/N \to \rho$ and $\bar{\ell}_{N(2)}/\delta_{NT}^2 \to 0$.

For the proof of E_2, we let $G_{it} = Z_{it} \varepsilon_{it}$ and $\hat{G}_{it} = \hat{Z}_{it} \varepsilon_{it}$. Then, we can rewrite

$$E_2 = \frac{1}{NT} \sum_{i=1}^{N} \sum_{k=1}^{N} (\hat{G}_i' \hat{G}_k - G_i' G_k) K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big),$$

where $G_i = (G_{i1}, ..., G_{iT})'$ and $\hat{G}_i = (\hat{G}_{i1}, ..., \hat{G}_{iT})'$. Thus, we need to prove

$$E_{21} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N} U_i' \left(M_{\hat{F}} - M_{F^0}\right) U_k K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big) = o_p(1),$$

$$E_{22} = \frac{1}{N^2T}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{j=1}^{N} \left(U_i' M_{\hat{F}} U_j \hat{a}_{kj} - U_i' M_{F^0} U_j a_{kj}\right) K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big) = o_p(1),$$

$$E_{23} = \frac{1}{N^3T}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{j=1}^{N}\sum_{l=1}^{N} \left(U_j' M_{\hat{F}} U_l \hat{a}_{ij}\hat{a}_{kl} - U_j' M_{F^0} U_l a_{ij} a_{kl}\right) K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big) = o_p(1).$$

where $U_i = (U_{i1}, ..., U_{iT})'$ and $U_{it} = X_{it}\varepsilon_{it}$.
For E_{21},

$$\begin{aligned}\|E_{21}\| &\leq O_p(\sqrt{\bar{\ell}_{N(2)}}) \left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{\sqrt{T}}U_i'\right\|^2\right)^{1/2} \|P_{F^0} - P_{\hat{F}}\| \\ &\quad \times \left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{\bar{\ell}_{N(2)}}\sum_{k=1}^{N}\frac{1}{\sqrt{T}}U_k K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big)\right\|^2\right)^{1/2} \\ &= O_p\left(\frac{\sqrt{\bar{\ell}_{N(2)}}}{\delta_{NT}}\right).\end{aligned} \tag{B3}$$

For E_{22},

$$\begin{aligned}E_{22} &= \frac{1}{N^2T}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{j=1}^{N} \left[U_i' M_{\hat{F}} U_j \left(\hat{a}_{kj} - a_{kj}\right) + U_i' \left(M_{\hat{F}} - M_{F^0}\right) U_j a_{kj}\right] K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big) \\ &= E_{22}^{(1)} + E_{22}^{(2)},\end{aligned}$$

where we can show $E_{22}^{(2)} = O_p\left(\frac{\sqrt{\bar{\ell}_{N(2)}}}{\delta_{NT}}\right)$ using a similar argument in (B3). For $E_{22}^{(1)}$, we follow the proof of Proposition 2 in Bai (2009) to have

$$\begin{aligned}E_{22}^{(1)} &= \frac{1}{N^2T}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{j=1}^{N} U_i' M_{\hat{F}} U_j \left(\hat{\lambda}_j - H^{-1}\lambda_j\right)' \left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\right)^{-1} \hat{\lambda}_k K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big) \\ &\quad + \frac{1}{N^2T}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{j=1}^{N} U_i' M_{\hat{F}} U_j H^{-1}\lambda_j \left[\left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\right)^{-1} - H'\left(\frac{\Lambda'\Lambda}{N}\right)^{-1} H\right] \hat{\lambda}_k K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big) \\ &\quad + \frac{1}{N^2T}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{j=1}^{N} U_i' M_{\hat{F}} U_j H^{-1}\lambda_j H' \left(\frac{\Lambda'\Lambda}{N}\right)^{-1} H \left(\hat{\lambda}_k - H^{-1}\lambda_k\right) K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big) \\ &:= e_1 + e_2 + e_3.\end{aligned}$$

For e_1,

$$\|e_1\| \leq \sqrt{\bar{\ell}_{N(2)}} \left(\frac{1}{N} \sum_{k=1}^{N} \left\| \frac{1}{\sqrt{\bar{\ell}_{N(2)}}} \sum_{i=1}^{N} K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big) \frac{1}{\sqrt{T}} U_i' \right\|^2 \right)^{1/2} \frac{1}{N} \sum_{j=1}^{N} \left\| \frac{1}{\sqrt{T}} U_j \right\| \left\| \hat{\lambda}_j - H^{-1} \lambda_j \right\|$$

$$\times \left\| \left(\frac{\hat{\Lambda}' \hat{\Lambda}}{N} \right)^{-1} \right\| \left(\frac{1}{N} \sum_{k=1}^{N} \left\| \hat{\lambda}_k \right\|^2 \right)^{1/2}$$

$$= O_p \left(\frac{\sqrt{\bar{\ell}_{N(2)}}}{\delta_{NT}} \right).$$

For e_2,

$$\|e_2\| \leq \sqrt{\bar{\ell}_{N(2)}} \left(\sum_{j=1}^{N} \left\| \frac{1}{\sqrt{\bar{\ell}_{N(2)}}} \sum_{i=1}^{N} K_2\Big(\frac{d_{ik}}{d_{(2)}}\Big) \frac{1}{\sqrt{T}} U_i' \right\|^2 \right)^{1/2} \left(\frac{1}{N} \sum_{j=1}^{N} \left\| \frac{1}{\sqrt{T}} U_j \right\| \left\| H^{-1} \right\| \|\lambda_j\| \right)$$

$$\times \left\| \left(\frac{\hat{\Lambda}' \hat{\Lambda}}{N} \right)^{-1} - H' \left(\frac{\Lambda' \Lambda}{N} \right)' H \right\| \left(\frac{1}{N} \sum_{k=1}^{N} \left\| \hat{\lambda}_k \right\|^2 \right)^{1/2}$$

$$= O_p \left(\frac{\sqrt{\bar{\ell}_{N(2)}}}{\delta_{NT}^2} \right) + O_p \left(\sqrt{\bar{\ell}_{N(2)}} \left\| \hat{\beta} - \beta \right\| \right).$$

because $\left(\frac{\hat{\Lambda}'\hat{\Lambda}}{N}\right)^{-1} - H' \left(\frac{\Lambda'\Lambda}{N}\right)' H = O_p\left(\frac{1}{\delta_{NT}^2}\right) + O_p\left(\left\|\hat{\beta} - \beta\right\|\right)$ due to Lemma A.10 in Bai (2009). Using the same procedure for $\|e_1\| = O_p\left(\frac{\sqrt{\bar{\ell}_{N(2)}}}{\delta_{NT}}\right)$, we can show that $\|e_3\| = O_p\left(\frac{\sqrt{\bar{\ell}_{N(2)}}}{\delta_{NT}}\right)$. Therefore, $E_{22} = o_p(1)$ as $N, T \to \infty$ such that $T/N \to \rho$ and $\bar{\ell}_{N(2)}/\delta_{NT}^2 \to 0$..

The proof of $E_{23} = o_p(1)$ is similar to the proof of $E_{22} = o_p(1)$, so it is omitted. Thus, $\hat{\Omega} - \tilde{\Omega} = o_p(1)$, which completes the proof of Theorem 3.2. □