Improved Inference for Interactive Fixed Effects Model with Cross Sectional Dependence

Zhenhao Gong

University of Connecticut

Outline

- Consider

$$
\begin{aligned}
Y_{it} &= X_{it}'\beta_0 + u_{it}, \\
u_{it} &= \lambda_i' F_t + \varepsilon_{it}.
\end{aligned}
$$

  - $F_t (r_0 \times 1)$: common factors; $\lambda_i (r_0 \times 1)$: factor loadings; r_0: number of factors, and assumed to be known.

- Consider

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

  - $F_t(r_0 \times 1)$: common factors; $\lambda_i(r_0 \times 1)$: factor loadings; r_0: number of factors, and assumed to be known.
  - X_{it} is potentially correlated with λ_i or F_t alone, or both.

- Consider

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

  - $F_t (r_0 \times 1)$: common factors; $\lambda_i (r_0 \times 1)$: factor loadings; r_0: number of factors, and assumed to be known.
  - X_{it} is potentially correlated with λ_i or F_t alone, or both.
  - ε_{it} is allowed to be weakly correlated in both dimensions.

- Consider

$$\begin{aligned} Y_{it} &= X_{it}'\beta_0 + u_{it}, \\ u_{it} &= \lambda_i' F_t + \varepsilon_{it}. \end{aligned}$$

  - $F_t(r_0 \times 1)$: common factors; $\lambda_i(r_0 \times 1)$: factor loadings; r_0: number of factors, and assumed to be known.
  - X_{it} is potentially correlated with λ_i or F_t alone, or both.
  - ε_{it} is allowed to be weakly correlated in both dimensions.
  - Model the unobservable common time-varying effects to impact the cross-sectional units heterogeneously
    $\Rightarrow$ More flexible than the standard fixed effects model.

- Consider

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

- $F_t (r_0 \times 1)$: common factors; $\lambda_i (r_0 \times 1)$: factor loadings; r_0: number of factors, and assumed to be known.
- X_{it} is potentially correlated with λ_i or F_t alone, or both.
- ε_{it} is allowed to be weakly correlated in both dimensions.
- Model the unobservable common time-varying effects to impact the cross-sectional units heterogeneously
  $\Rightarrow$ More flexible than the standard fixed effects model.
- Incidental parameters problem in estimation
  $\Rightarrow$ Asymptotic bias and invaild inference.

Why Use IFE?

- The causal relationships between divorce law reforms and the rise of divorce rates in 1970s (e.g., Firedberg, 1998; Wolfers, 2006; Kim and Oka, 2013).

Why Use IFE?

- The causal relationships between divorce law reforms and the rise of divorce rates in 1970s (e.g., Firedberg, 1998; Wolfers, 2006; Kim and Oka, 2013).
- Wolfers (2006) studied the model as

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

• Not flexible to capture unobserved time-varying factors (e.g., the stigma of divorce; religious belief)
$\Rightarrow$ Large discrepancy between the OLS and WLS estimates.
• ε_{st} assumed to be cross-sectionally independent.

Why Use IFE?

- The causal relationships between divorce law reforms and the rise of divorce rates in 1970s (e.g., Firedberg, 1998; Wolfers, 2006; Kim and Oka, 2013).
- Wolfers (2006) studied the model as

$$y_{st} = T_{st} + f(v_s, t) + u_{st},$$
$$u_{st} = \delta_s + \alpha_t + \varepsilon_{st}.$$

  - Not flexible to capture unobserved time-varying factors (e.g., the stigma of divorce; religious belief)

  $\Rightarrow$ Large discrepancy between the OLS and WLS estimates.
  - ε_{st} assumed to be cross-sectionally independent.
- **The IFE model is robust to the weighing schemes and provide a natural solution for robust standard errors.**

- $(\hat{\beta}, \hat{F}, \hat{\Lambda})$ minimizes

$$\mathrm{SSR}(\beta_0, F, \Lambda) = \sum_{i=1}^{N} \left(Y_i - X_i\beta_0 - F\lambda_i\right)' \left(Y_i - X_i\beta_0 - F\lambda_i\right),$$

  subject to $F'F/T = I_r$ and $\Lambda'\Lambda$ being diagonal.

- Concentrating out Λ, the LS estimator for β_0 given F is:

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

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

- Given β_0, the model reduces to a pure factor model, so we can estimate F using PCA:

$$\left[\frac{1}{NT}\sum_{i=1}^{N}\left(Y_i - X_i\beta_0\right)\left(Y_i - X_i\beta_0\right)'\right]\hat{F} = \hat{F}V_{NT},$$

  where V_{NT} is a diagonal matrix that consists the r_0 largest eigenvalues of the matrix in the brackets and $\hat{F}$ is $\sqrt{T}$ times the corresponding eigenvectors.

- The solution $(\hat{\beta}, \hat{F})$ can be obtained by iteration until convergence. Given $(\hat{\beta}, \hat{F})$, we have $\hat{\Lambda} = T^{-1}(Y - X\hat{\beta})'\hat{F}$.

- Bai (2009) shows that as $N, T \to \infty$, under some regularity assumptions and if $T/N \to \rho > 0$,

$$\sqrt{NT}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(\rho^{1/2} B_0 + \rho^{-1/2} C_0, H_0^{-1} H_Z H_0^{-1}\right).$$

where B_0 and C_0 arise from cross-sectional and serial correlations and heteroskedasticities in ε_{it} $\Rightarrow$ invalid inference.

- Bai (2009) shows that as $N, T \to \infty$, under some regularity assumptions and if $T/N \to \rho > 0$,

$$\sqrt{NT}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(\rho^{1/2} B_0 + \rho^{-1/2} C_0, H_0^{-1} H_Z H_0^{-1}\right).$$

  where B_0 and C_0 arise from cross-sectional and serial correlations and heteroskedasticities in $\varepsilon_{it} \Rightarrow$ invalid inference.

- In the presence of serial correlation, we can correct the bias C_0 by the truncated kernel method of Newey and West (1987).

- Bai (2009) shows that as $N, T \to \infty$, under some regularity assumptions and if $T/N \to \rho > 0$,

$$\sqrt{NT}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(\rho^{1/2}B_0 + \rho^{-1/2}C_0, H_0^{-1}H_Z H_0^{-1}\right).$$

  where B_0 and C_0 arise from cross-sectional and serial correlations and heteroskedasticities in $\varepsilon_{it} \Rightarrow$ invalid inference.

- In the presence of serial correlation, we can correct the bias C_0 by the truncated kernel method of Newey and West (1987).
- **Goal:** Developing a valid inference procedure under cross-sectional correlation and heteroskedasticity, assuming no serial correlation ($C_0 = 0$).
  - Correct the asymptotic bias B_0.
  - Employ a robust estimation for H_Z.

- The asymptotic bias B_0 is the probability limit of B_{NT} with

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

where

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

$$w_i = \operatorname{plim} \left[\frac{\left(X_i - V_i\right)' F^0}{T} \right] \left(\frac{F^{0\prime} F^0}{T} \right)^{-1} \left(\frac{\Lambda' \Lambda}{N} \right)^{-1},$$

$$V_i = \frac{1}{N} \sum_{k=1}^{N} a_{ik} X_k, \text{ and } a_{ik} = \lambda_i' (\Lambda' \Lambda / N)^{-1} \lambda_k.$$

- Panel data models with interactive fixed effects:
    - Holtz-Eakin et al.(1988); Ahn et al.(2001); Pesaran (2006); Bai (2009); Moon and Weidner (2017); etc.
    - Empirical studies: Kim and Oka (2013); Gobillon and Magnac (2016); Totty (2017); etc.

- Panel data models with interactive fixed effects:
  - Holtz-Eakin et al.(1988); Ahn et al.(2001); Pesaran (2006); Bai (2009); Moon and Weidner (2017); etc.
  - Empirical studies: Kim and Oka (2013); Gobillon and Magnac (2016); Totty (2017); etc.
- Methods for the cross-sectional correlation bias:
  - Bai (2009): CS-HAC method.
  - Bai and Liao (2017): GLS method.

- Panel data models with interactive fixed effects:
  - Holtz-Eakin et al.(1988); Ahn et al.(2001); Pesaran (2006); Bai (2009); Moon and Weidner (2017); etc.
  - Empirical studies: Kim and Oka (2013); Gobillon and Magnac (2016); Totty (2017); etc.
- Methods for the cross-sectional correlation bias:
  - Bai (2009): CS-HAC method.
  - Bai and Liao (2017): GLS method.
- The spatial HAC method:
  - Conley (1996, 1999); Conley and Molinari (2007); Kelejian and Prucha (2007); Kim and Sun (2011, 2013); Bester et al.,(2017); Mueller and Watson (2021); etc.

1. The CS-HAC estimator (Bai, 2009):

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

where $H_0 = \text{plim} H(F)$; $\hat{H}_0$ and $\hat{w}_i$ are the estimators of H_0 and w_i with F, λ_i, and Λ replaced by $\hat{F}$, $\hat{\lambda}_i$, and $\hat{\Lambda}$.

- Consistent as $n_{sub}/\min\{N,T\} \to 0$.
- Hard to implement properly. Need to select n_{sub} to replicate the dependence structure of the whole sample.
- Performance highly depends on the sub-sample selection and there is no practical guidance to select.

2. The GLS estimator (Bai and Liao, 2017):

$$\hat{\beta}(\Sigma_{\varepsilon}^{-1}) = arg\min_{\beta} \sum_{t=1}^{T} (Y_t - X_t\beta - \Lambda F_t)\, \Sigma_{\varepsilon}^{-1}\, (Y_t - X_t\beta - \Lambda F_t)\,,$$

where $\Sigma_{\varepsilon} = cov(\varepsilon_t), (N \times N)$. They assume Σ_{ε} is sparse and $\{\varepsilon_t : t \geq 1\}$ is serial independent.

- Advantages:
  - More efficient than existing methods.
  - Incidental parameters bias-free.
- Practical issues:
  - Its inference is not stable in finite samples
    $\Rightarrow$ Our simulation shows that is often produces substantial size distortion.
  - Romanno and Wolf (2006); Angrist and Pischke (2010).

Outline

Our procedure **improves** the inference of $\hat{\beta}$ by **correcting the bias** and **employing a robust covariance estimation**.

1. **Correcting the bias**

- Recall

$$
\begin{aligned}
B_{NT} &= -H(F)^{-1} \underbrace{\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} w_i \lambda_k \left(\frac{1}{T}\sum_{t=1}^{T} E\varepsilon_{it}\varepsilon_{kt}\right)}_{=J_{NT}}, \\
&= -H(F)^{-1} J_{NT}.
\end{aligned}
$$

- $H(F)$ is easy to estimate, our focus is on consistent estimation of J_{NT}.

- We propose a TA-SHAC estimator to estimate J_{NT},

$$\hat{J}_{NT} = \frac{1}{T}\sum_{t=1}^{T} \underbrace{\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} K\left(\frac{d_{ik}}{d_n^{(1)}}\right)\hat{w}_i\hat{\lambda}_k\hat{\varepsilon}_{it}\hat{\varepsilon}_{kt}\right]}_{=\hat{J}_t}$$
$$= \frac{1}{T}\sum_{t=1}^{T}\hat{J}_t.$$

- $\hat{J}_t$ is a standard spatial HAC estimator in the literature and $\hat{J}_{NT}$ can be viewed as a time average of $\hat{J}_t, t = 1, \cdots, T$.
- $K(\cdot)$ is a real-valued kernel function. d_{ik} is the distance measure between i and k and $d_n^{(1)}$ is a bandwidth parameter.

- Based on $\hat{J}_{NT}$, we can estimate B_{NT} by

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

- The bias-corrected LS estimator can be defined as

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

2. **Robust covariance estimation**

- Recall under cross-sectional dependence

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

where $H_Z = \text{plim}\,\frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} E(\varepsilon_{it}\varepsilon_{kt})Z_{it}Z_{kt}'$ with $Z_i = M_{F_0}X_i - \frac{1}{N}\sum_{k=1}^{N} a_{ik}M_{F_0}X_k$.

2. **Robust covariance estimation**

- Recall under cross-sectional dependence

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

where $H_Z = \text{plim}\, \frac{1}{NT}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{t=1}^{T} E(\varepsilon_{it}\varepsilon_{kt})Z_{it}Z_{kt}'$ with $Z_i = M_{F_0}X_i - \frac{1}{N}\sum_{k=1}^{N} a_{ik}M_{F_0}X_k$.

- H_Z is conventional estimated as

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

where $\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_{it}^2$. Not valid under cross-sectional dependence.

- Bai (2009) suggests CS-HAC estimator

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

  - Hard to implement in practice.
  - Sensitive to the sub-sample selection.

- We find the CS-HAC estimator is still consistent if the whole sample is used,

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

- We propose a TA-SHAC estimator given by,

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

$$= \frac{1}{T}\sum_{t=1}^{T}\hat{H}_t.$$

- If $\mathfrak{K}_F$ is a rectangle kernel, then our estimator $\hat{H}_{NT}$ includes $\tilde{H}_{CS}$ as a special case.

- To establish the consistency of $\hat{J}_{NT}$, we introduce an infeasible estimator $\tilde{J}_{NT}$,

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

$$= \frac{1}{T}\sum_{t=1}^{T} \tilde{J}_t.$$

- We have

$$\hat{J}_{NT} - J_{NT} = \underbrace{(\hat{J}_{NT} - \tilde{J}_{NT})}_{\text{Estimation error}} + \underbrace{(\tilde{J}_{NT} - E\tilde{J}_{NT})}_{\text{Variation}} + \underbrace{(E\tilde{J}_{NT} - J_{NT})}_{\text{Bias}}.$$

- We assume that ε_{it} has a linear representation:

$$\varepsilon_{it} = \sum_{\ell=1}^{\infty} c_{it,\ell} e_{\ell},$$

where $\{c_{it,\ell}\}$ are unknown constants.

- We can rely on Andrews (1991), Kim and Sun (2011, 2013) to show that

$$\tilde{J}_{NT} - E\tilde{J}_{NT} = O_p\left(\sqrt{\frac{\ell_n}{NT}}\right) \text{ and } E\tilde{J}_{NT} - J_{NT} = O\left(\frac{1}{d_n^q}\right),$$

where

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

- Based on the arguments in Bai (2009), we can show that

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

- **Assumption 1.** *(i)* $d_{ik} \geq 0, d_{ii} = 0,$ *and* $d_{ik} = d_{ki},$ *(ii)* d_{ik} *is time invariant.*
- **Assumption 2.** *(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*

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

- **Assumption 3.** $e_\ell \overset{iid}{\sim} (0, 1)$ *and* $E(e_\ell^4) \leq \infty,$ *for all* $\ell.$

- **Assumption 4.** *(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* $i=1,\cdots,N$.
- **Assumption 5.** $\ell_i\leq c_\ell \ell_n$ for all $i=1,\cdots,N$ with some constant c_ℓ.
- **Assumption 6.** *There exists a finite constant* M *such that*

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

where $\Gamma_{ik,t}=E\left(\varepsilon_{it}\varepsilon_{kt}\right).$

Theorem 1. *Under the Assumptions in Bai (2009) and Assumption 1-6, with* $d_n, \ell_n, N, T \to \infty$ *such that* $\ell_n/N, \ell_n/T \to 0$ *and* $T/N \to \rho$*, we have* $\hat{J}_{NT} - J_{NT} = o_p(1)$.

Theorem 1. *Under the Assumptions in Bai (2009) and Assumption 1-6, with* $d_n, \ell_n, N, T \to \infty$ *such that* $\ell_n/N, \ell_n/T \to 0$ *and* $T/N \to \rho$*, we have* $\hat{J}_{NT} - J_{NT} = o_p(1)$.

Theorem 2. *Under the the Assumptions in Bai (2009) and Assumption 1-6, with* $d_n, \ell_n, N, T \to \infty$ *such that* $\ell_n/N, \ell_n/T \to 0$ *and* $T/N \to \rho$*, we have* $\hat{H}_{NT} - H_Z = o_p(1)$.

Theorem 1. *Under the Assumptions in Bai (2009) and Assumption 1-6, with* $d_n, \ell_n, N, T \to \infty$ *such that* $\ell_n/N, \ell_n/T \to 0$ *and* $T/N \to \rho$, *we have* $\hat{J}_{NT} - J_{NT} = o_p(1)$.

Theorem 2. *Under the the Assumptions in Bai (2009) and Assumption 1-6, with* $d_n, \ell_n, N, T \to \infty$ *such that* $\ell_n/N, \ell_n/T \to 0$ *and* $T/N \to \rho$, *we have* $\hat{H}_{NT} - H_Z = o_p(1)$.

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

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

Outline

There are two major challenges in implementing our method:

- **How to choose the distance measure?**
  - Transportation cost (Conley and Ligon, 2000)
  - Economics/geographic distance (Pinkse et al., 2002), etc.
    $\Rightarrow$ May not be available and appropriate.
- **How to select the bandwidths jointly?**
  - Fixed bandwidth (Kelejian and Prucha, 2007)
  - Asymptotic truncated MSE (Kim and Sun, 2011)
    $\Rightarrow$ May not applicable to our estimators.

- We define the distance that reflects the dependence structure directly,

$$d_{ik} = |1/\rho_{ik}| - 1.$$

where $\rho_{ik} = corr(\varepsilon_{it}, \varepsilon_{kt})$. d_{ik} is unobservable but we can use the sample counter part,

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

where $\hat{\rho}_{ik} = \sum_{t=1}^{T} \hat{\varepsilon}_{it}\hat{\varepsilon}_{kt} / \sqrt{\sum_{t=1}^{T} \hat{\varepsilon}_{it}^2 \sum_{t=1}^{T} \hat{\varepsilon}_{kt}^2}$.

  - Mantegna (1998), Fernandez (2011), Kim (2020), etc.
  - No need prior information for implementation.
  - Does not satisfy triangle inequality.

Bandwidth Selection Procedure

- Kim and Sun (2017) use a simulation-based choice in time-series kernel method to select two smoothing parameters in their test procedure.
- How to replicate cross-sectional dependence?
  - Silvia Goncalves (2011)
  - Timothy Vogelsang (2012)
  - Javier Hidalgo and Marcia Schafgans (2017)
- Bootstrap-based procedure: avoid to use the parameter model in time series case.

- We consider a bootstrap-based bandwidth selection procedure. 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 $d_n^{(1)}$ and $d_n^{(2)}$.
    1. Estimate $\hat{\beta}$, $\hat{F}_t$, $\hat{\Lambda}$, and $\hat{\varepsilon}_t = Y_t - X_t\hat{\beta} - \hat{\Lambda}\hat{F}_t$. (Bai, 2009)

- We consider a bootstrap-based bandwidth selection procedure. 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 $d_n^{(1)}$ and $d_n^{(2)}$.

  1. Estimate $\hat{\beta}$, $\hat{F}_t$, $\hat{\Lambda}$, and $\hat{\varepsilon}_t = Y_t - X_t\hat{\beta} - \hat{\Lambda}\hat{F}_t$. (Bai, 2009)
  2. Generate bootstrap sample Y_t^* based on

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

- We consider a bootstrap-based bandwidth selection procedure. 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 $d_n^{(1)}$ and $d_n^{(2)}$.
  1. Estimate $\hat{\beta}$, $\hat{F}_t$, $\hat{\Lambda}$, and $\hat{\varepsilon}_t = Y_t - X_t\hat{\beta} - \hat{\Lambda}\hat{F}_t$. (Bai, 2009)
  2. Generate bootstrap sample Y_t^* based on $Y_t^* = X_t\hat{\beta} + \hat{\Lambda}\hat{F}_t + \varepsilon_t^*$ with $\varepsilon_t^* = \hat{\varepsilon}_t\xi_t$, and $\xi_t \overset{iid}{\sim} (0, 1)$ (e.g. Rademacher distribution).
  3. Estimate $\hat{\beta}^*$, $\hat{F}_t^*$, $\hat{\Lambda}^*$, and $\hat{\varepsilon}_t^*$. Construct the bootstrap version of the bias estimator $\hat{B}_{NT}^*(d_{nm}^{(1)})$ with $d_{nm}^{(1)} \in \mathcal{D}_{nM}^{(1)}$.

- We consider a bootstrap-based bandwidth selection procedure. 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 $d_n^{(1)}$ and $d_n^{(2)}$.

  1. Estimate $\hat{\beta}$, $\hat{F}_t$, $\hat{\Lambda}$, and $\hat{\varepsilon}_t = Y_t - X_t\hat{\beta} - \hat{\Lambda}\hat{F}_t$. (Bai, 2009)
  2. Generate bootstrap sample Y_t^* based on $Y_t^* = X_t\hat{\beta} + \hat{\Lambda}\hat{F}_t + \varepsilon_t^*$ with $\varepsilon_t^* = \hat{\varepsilon}_t\xi_t$, and $\xi_t \overset{iid}{\sim} (0,1)$ (e.g. Rademacher distribution).
  3. Estimate $\hat{\beta}^*$, $\hat{F}_t^*$, $\hat{\Lambda}^*$, and $\hat{\varepsilon}_t^*$. Construct the bootstrap version of the bias estimator $\hat{B}_{NT}^*(d_{nm}^{(1)})$ with $d_{nm}^{(1)} \in \mathcal{D}_{nM}^{(1)}$.
  4. Estimate the bootstrap version of the covariance matrix estimator $\hat{H}_{NT}^*(d_{ns}^{(2)})$ with $d_{ns}^{(2)} \in \mathcal{D}_{nS}^{(2)}$.

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

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

$$\text{with} \quad \hat{\beta}^{\dagger *} = \hat{\beta}^* - \frac{1}{N} B_{NT}^* \left(d_{nm}^{(1)} \right) \quad \text{and}$$

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

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

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

- Consider the following DGP:

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

- We set $\beta = \mu = c = 1$ and $\rho = 0.3$. The number of common factors is two, and is assumed to be known.

- We generates cross-sectional dependent data using a popular spatial MA model.

$$\varepsilon_t = \left(I_N + \theta M_1 + \theta^2 M_2\right) v_t,$$
$$v_t = (v_{1t}, \ldots, v_{Nt})', v_{it} \overset{iid}{\sim} N(0,1).$$

- We generates cross-sectional dependent data using a popular spatial MA model.

$$\varepsilon_t = \left(I_N + \theta M_1 + \theta^2 M_2\right) v_t,$$
$$v_t = (v_{1t}, \ldots, v_{Nt})', v_{it} \overset{iid}{\sim} N(0,1).$$

- $M_1 = [m_{1,ik}]_{i,k=1}^{N}$ and $M_2 = [m_{2,ik}]_{i,k=1}^{N}$ are $(N \times N)$ spatial weight matrices such that

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

Table: 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
					$\theta = .4$				
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: Scaled bias equals the difference between each estimator and its true value scaled by $\sqrt{NT}$. RMSE is the corresponding root mean square error scaled by $\sqrt{NT}$. d_{ik}^T denotes the true distance. d_{ik}^D denotes the data driven distance measure.

Table: 95% empirical coverage rates of different estimators

				TA-SHAC (d_{ik}^{T})			TA-SHAC (d_{ik}^{D})		
T	N	$\hat{\beta}$	$\hat{\beta}_{gls}$	$\tilde{\beta}_{hac1}$	$\tilde{\beta}_{hac2}$	$\tilde{\beta}^{*}_{hac}$	$\hat{\beta}_{hac1}$	$\hat{\beta}_{hac2}$	$\hat{\beta}^{*}_{hac}$
					$\theta = .4$				
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: For $\hat{\beta}_{hac1}$, we estimate covariance matrix only by TA-SHAC without bias correction. For $\hat{\beta}_{hac2}$, we correct the bias only by TA-SHAC. We correct the bias and estimate the covariance matrix by TA-SHAC for $\hat{\beta}^{*}_{hac}$.

1. **Effects of divorce law reforms**

- **Background:** During and after 1970s, most of states in 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.
- **Research question:** the causal relationships between divorce law reforms and divorce rates.

- **Literature:**
  - Peters (1986) suggested that divorce rates were unaffected by the law reforms, while Allen (1992) found a significant impact.

- **Literature:**
  - Peters (1986) suggested that divorce rates were unaffected by the law reforms, while Allen (1992) found a significant impact.
  - After controlling for fixed state and year effects, as well as state-specific time trends, Firedberg (1998) found that states' law reforms have contributed to one-sixth of the rise and claimed the change was permanent.

- **Literature:**
  - Peters (1986) suggested that divorce rates were unaffected by the law reforms, while Allen (1992) found a significant impact.
  - After controlling for fixed state and year effects, as well as state-specific time trends, Firedberg (1998) found that states' law reforms have contributed to one-sixth of the rise and claimed the change was permanent.
  - Wolfers (2006) confirmed the rise of divorce rates in the first eight years after the law reform, but this rise was reversed for the subsequent nine to fourteen year.

- Specifically, Wolfers (2006) studied the model as

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

  where y_{st} is the annual divorce rates; $f(v_s, t)$ is the time trend; δ_s and α_t are the state and the time fixed effects.

- The treatment effects T_{st} is

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

  where T_s is the law reform year of state s.

- The robustness of Wolfers (2006) has been doubted since
  - the additive structure in u_{st} is not flexible to capture factors varying across time and state (e.g. the stigma of divorce; religious belief).
  - ε_{st} is assumed to be cross-sectionally independent.

- The robustness of Wolfers (2006) has been doubted since
    - the additive structure in u_{st} is not flexible to capture factors varying across time and state (e.g. the stigma of divorce; religious belief).
    - ε_{st} is assumed to be cross-sectionally independent.
- Kim and Oka (2013) applied the IFE model, which u_{st} is expressed as

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

where F_t is principle components of u_{it}, which dominant the portion of divorce rates not explained by the included regressors. λ_s stands for the heterogeneous effect of F_t to each state.

- Kim and Oka (2013) adopted the estimation and bias correction procedure in Bai (2009)
  - Not take the cross-sectional correlated errors into account.
  - Estimated the standard errors by the conventional estimator.
- Bai and Liao (2017) re-estimate the model of Kim and Oka (2013) using the GLS method.
- We apply the proposed procedure to correct the cross-sectional correlation bias and improve the inference of the estimates.

Table: Methods comparison in effects of divorce law reform

	$\hat{\beta}$		$\hat{\beta}^*_{hac}$		$\hat{\beta}_{gls}$	
	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 $r = 10$.
* $p < .1$. ** $p < .05$. *** $p < .01$.

2. **Effects of water and sewerage interventions**

- **Background:** From 1880 to 1920, when Boston authorities developed a sewerage and water district, infant mortality plummeted from around 1/5 to 1/16 white infants, and deaths of noninfants under 5 years decreased by a factor of seven in Massachusetts.
- **Research question:** the causal relationships between water and sewerage interventions and child mortality.

- **Literature:**
  - Cutler and Miller (2015) studied the impact of water chlorination and filtration on the death rate from waterborne diseases across 13 US cities. Their results suggest that improved water quality decreases 47 percent in log infant mortality from 1900 to 1936.

- **Literature:**
  - Cutler and Miller (2015) studied the impact of water chlorination and filtration on the death rate from waterborne diseases across 13 US cities. Their results suggest that improved water quality decreases 47 percent in log infant mortality from 1900 to 1936.
  - Alsan and Goldin (2019) exploited the independent and combined effects of clean water and effective sewerage systems on under-5 mortality in Massachusetts, 1880-1920. They identified the two interventions together account for approximately one-third of the decline in log child mortality during the 41 years.

- Specifically, Alsan and Goldin (2019) estimate

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

  - i is municipality and t is year; y_{it} is the log under-5 mortality rate. 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, municipality-specific time trends.
  - The standard errors are clustered at the municipality level with 60 clusters in their analysis.

- Since they used the municipality level data, the potential unobserved heterogeneities and cross-sectional correlation in the errors may affect the results.

- Since they used the municipality level data, the potential unobserved heterogeneities and cross-sectional correlation in the errors may affect the results.
- To check the robustness of their results, we first apply the IFE model with u_{it} expressed as

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

where F_t dominant the portion of child mortality rates not explained by the included regressors. λ_i stands for the heterogeneous effect of F_t to each municipality.

- Then, we apply the proposed procedure to correct the bias and provide valid inference for the interactive fixed effects model.

- Then, we apply the proposed procedure to correct the bias and provide valid inference for the interactive fixed effects model.
- Note that if we let $\lambda_i = (\delta_i, 1, \delta_i)'$ and $F_t = (1, \alpha_t, t)'$, then u_{it} in above equations are the same. Hence, we choice three factors in our model to include the original model as a special case.

- Then, we apply the proposed procedure to correct the bias and provide valid inference for the interactive fixed effects model.
- Note that if we let $\lambda_i = (\delta_i, 1, \delta_i)'$ and $F_t = (1, \alpha_t, t)'$, then u_{it} in above equations are the same. Hence, we choice three factors in our model to include the original model as a special case.
- Finally, we apply the GLS method for the study to compare with our method.

Table: Estimated effects of clean water and sewerage

	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				
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]

Note: 95 % confidence intervals are reported.
* $p < .1$. ** $p < .05$. *** $p < .01$.

Table: Estimated effects of clean water and sewerage

	(1)	(2)	(3)	(4)	(5)
	Panel C. TA-SHAC Estimation				
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				
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.
* $p < .1$. ** $p < .05$. *** $p < .01$.

Conclusion

1. We propose an improved inference procedure for the IFE model in the presence of cross-sectional dependence and heteroskedasticity.
2. We prove the validity of the proposed procedure in the asymptotic sense.
3. To implement our approach, we develop a data driven distance that does not rely on prior information and a bandwidth selection procedure based on a cluster wild bootstrap method.
4. We show that our procedure performs well in simulation with finite samples.