Feasible Generalized Least Squares for Panel Data with Cross-sectional and Serial Correlations

Jushan Bai* Columbia University

Sung Hoon Choi† Rutgers University

Yuan Liao‡ Rutgers University

October 11, 2020

Abstract

This paper considers generalized least squares (GLS) estimation for linear panel data models. By estimating the large error covariance matrix consistently, the proposed feasible GLS (FGLS) estimator is more efficient than the ordinary least squares (OLS) in the presence of heteroskedasticity, serial and cross-sectional correlations. To take into account the serial correlations, we employ the banding method. To take into account the cross-sectional correlations, we suggest to use the thresholding method. We establish the limiting distribution of the proposed estimator. A Monte Carlo study is considered. The proposed method is applied to an empirical application.

Keywords: Panel data, efficiency, thresholding, banding, cross-sectional correlation, serial correlation, heteroskedasticity

*Address: 420 West 118th St. MC 3308, New York, NY 10027, USA. E-mail: jb3064@columbia.edu.

†Address: 75 Hamilton St., New Brunswick, NJ 08901, USA. E-mail: shchoi@economics.rutgers.edu.

‡Address: 75 Hamilton St., New Brunswick, NJ 08901, USA. Email: yuan.liao@rutgers.edu.

1 Introduction

Heteroskedasticity, cross-sectional and serial correlations are important problems in the error terms of panel regression models. There are two approaches to deal with these problems. The first approach is to use the ordinary least squares (OLS) estimator but with a robust standard error that is robust to heteroskedasticity and correlations, for example, White (1980); Newey and West (1987); Liang and Zeger (1986); Arellano (1987); Driscoll and Kraay (1998); Hansen (2007a); Vogelsang (2012), among others. A widely used class of robust standard errors are clustered standard errors, for example, Petersen (2009), Wooldridge (2010) and Cameron and Miller (2015). Bai, Choi, and Liao (2019) proposed a robust standard error with unknown clusters. In an interesting paper by Abadie, Athey, Imbens, and Wooldridge (2017), they argued for caution in the application of clustered standard errors since they may give rise to conservative confidence intervals. The second approach is to use the generalized least squares estimator (GLS) that directly takes into account heteroskedasticity, and cross-sectional and serial correlations in the estimation. It is well known that GLS is more efficient than OLS.

This paper focuses on the second approach. For panel models, the underlying covariance matrix involves a large number of parameters. It is important to make GLS operational. We thus consider feasible generalized least squares (FGLS). Hansen (2007b) studied FGLS estimation that takes into account serial correlation and clustering problems in fixed effects panel and multilevel models. His approach requires the cluster structure to be known. This gives motivation to our paper. We assume the unknown cluster structure, and control heteroskedasticity, both serial and cross-sectional correlations by estimating the large error covariance matrix consistently. In cross-sectional setting, Romano and Wolf (2017) obtained asymptotically valid inference of the FGLS estimator, combined with heteroskedasticity-consistent standard errors without knowledge of the conditional heteroskedasticity functional form. Moreover, Miller and Startz (2018) adapted machine learning methods (i.e., support vector regression) to take into account the misspecified form of heteroskedasticity.

In this paper, we consider (i) balanced panel data, (ii) the case of large-N large-T, and (iii) both serial and cross-sectional correlations, but unknown structure of clusters. We introduce a modified FGLS estimator that eliminates the cross-sectional and serial correlation bias by proposing a high-dimensional error covariance matrix estimator. In addition, our proposed method is applicable when the knowledge of clusters is not available. Let u_t be an $N \times 1$ vector of regressor noises, whose definition is to be clear later. Following the idea of Bai and Liao (2017), in this paper, the FGLS involves estimating an $NT \times NT$ dimensional inverse covariance matrix Ω^{-1}, where

$$\Omega = (Eu_t u_s')$$

where each block $Eu_t u_s'$ is an $N \times N$ autocovariance matrix. Here parametric structures on

the serial or cross-sectional correlations are not imposed. By assuming weak dependences, we apply nonparametric methods to estimate the covariance matrix. To address the estimation of serial autocorrelations, we employ the idea of Newey-West truncation. This method, in the FGLS setting, is equivalent to "banding", previously proposed by Bickel and Levina (2008b) for estimating large covariance matrices. We apply it to banding out off-diagonal $N \times N$ blocks that are far from the diagonal block. In addition, to control for the cross-sectional correlation, we assume that each of the $N \times N$ block matrices are sparse, potentially resulting from the presence of cross-sectional correlations within clusters. We then estimate them by applying the thresholding approach of Bickel and Levina (2008a). We apply thresholding separately to the $N \times N$ blocks, which are formed by time lags $Eu_t u_{t-h}' : h = 0, 1, 2, ...$ This allows the cluster-membership to be potentially changing over-time. A contribution of this paper is the theoretical justification for estimating the large error covariance matrix.

For the FGLS, it is crucial for the asymptotic analysis to prove that the effect of estimating Ω is first-order negligible. In the usual low-dimensional settings that involve estimating optimal weight matrix, such as the optimal GMM estimations, it has been well known that consistency for the inverse covariance matrix estimator is sufficient for the first-order asymptotic theory, e.g., Hansen (1982), Newey (1990), Newey and McFadden (1994). However, it turns out that when the covariance matrix is of high-dimensions, not even the optimal convergence rate for estimating Ω^{-1} is sufficient. In fact, proving the first-order equivalence between the FGLS and the infeasible GLS (that uses the true Ω^{-1}) is a very challenging problem under the large N, large T setting. We provide a new theoretical argument to achieve this goal.

The banding and thresholding methods, which we employ in this paper, are two useful regularization methods. In the recent machine learning literature, these methods have been extensively exploited for estimating high-dimensional parameters. Moreover, in the econometric literature, nonparametric machine learning techniques have been verified to be powerful tools: Bai and Ng (2017); Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2016, 2017); Wager and Athey (2018), etc.

The rest of the paper is organized as follows. In Section 2, we describe the model and the large error covariance matrix estimator. Also we introduce the implementation of FGLS estimatior and its limiting distribution. Section 3 presents Monte Carlo studies evaluating the finite sample performance of the estimators. In Section 4, we apply our methods to study the US divorce rate problem. Conclusions are provided in Section 5. All proofs are given in Appendix A.

Throughout this paper, let $\nu_{\min}(A)$ and $\nu_{\max}(A)$ denote the minimum and maximum eigenvalues of matrix A respectively. Also we use $\|A\| = \sqrt{\nu_{\max}(A'A)}$, $\|A\|_1 = \max_i \sum_j |A_{ij}|$ and $\|A\|_F = \sqrt{tr(A'A)}$ as the operator norm, ℓ_1-norm and the Frobenius norm of a matrix A, respectively. Note that if A is a vector, $\|A\| = \|A\|_F$ is equal to the Euclidean norm.

2 Feasible Generalized Least Squares

We consider a linear model [1]

$$y_{it} = x_{it}'\beta + u_{it}. \tag{2.1}$$

The model (2.1) can be stacked and represented in full matrix notation as

$$Y = X\beta + U, \tag{2.2}$$

where $Y = (y_1', \cdots, y_T')'$ is the $NT \times 1$ vector of y_{it} with each y_t being an $N \times 1$ vector; $X = (x_1', \cdots, x_T')'$ is the $NT \times d$ matrix of x_{it} with each x_t being an $N \times d$; $U = (u_1', \cdots, u_T')'$ is the $NT \times 1$ vector of u_{it} with each u_t being an $N \times 1$ vector.

Let $\Omega = (Eu_t u_s')$ be an $NT \times NT$ matrix, consisting of many "blocks" matricies. The (t, s)th block is an $N \times N$ covariance matrix $Eu_t u_s'$. We consider the following (infeasible) GLS estimator of β:

$$\widetilde{\beta}_{GLS}^{inf} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y. \tag{2.3}$$

Note that Ω is a high-dimensional conditional covariance matrix, which is very difficult to estimate. We aim to achieve the following: (i) obtain a "good" estimator of Ω^{-1}, allowing an arbitrary form of weak dependence in u_{it}, and (ii) show that the effect of replacing Ω^{-1} by $\widehat{\Omega}^{-1}$ is asymptotically negligible.

We start with a population approximation for Ω in order to gain the intuitions. Then, we suggest the estimator for Ω that takes into account both serial and cross-sectional correlations.

2.1 Population approximation

We start with a "banding" approximation to control serial correlations. Recall that $\Omega = (Eu_t u_s')$, where the (t, s) block is $Eu_t u_s'$. By assuming serial stationarity and strong mixing condition, $Eu_t u_s'$ depends on (t, s) only through $h = t - s$. Specifically, with slight abuse of notation, we can write $\Omega_{t,s} = \Omega_h = Eu_t u_{t-h}'$. Note for $i \neq j$, it is possible that $Eu_{it}u_{j,t-h} \neq Eu_{i,t-h}u_{jt}$, so Ω_h is possibly non-symmetric for $h > 0$. On the other hand, Ω is symmetric due to $\Omega_{s,t} = \Omega_{t,s}'$. The diagonal blocks are the same, and all equal $\Omega_0 = Eu_t u_t'$, while magnitudes of the elements of the off-diagonal blocks $\Omega_h = Eu_t u_{t-h}'$ decay to zero as $|h| \to \infty$ under the weak serial dependence assumption.

In the Newey-West spirit, Ω can be approximated by $\Omega^{NW} = (\Omega_{t,s}^{NW})$, where each block can be written as $\Omega_{t,s}^{NW} = \Omega_h^{NW}$ for $h = t - s$. Here Ω_h^{NW} is an $N \times N$ block matrix, defined

[1] For technical simplicity we focus on a simple model where there are no fixed effects. It is straightforward to allow additive fixed effects $\alpha_i + \mu_t$ by applying the de-meaning first. The theories would be slightly more sophisticated, though such extensions are straightforward.

as:

$$\Omega_h^{NW} = \begin{cases} Eu_t u_{t-h}', & \text{if } |h| \leq L \\ 0, & \text{if } |h| > L, \end{cases}$$

for some pre-determined $L \to \infty$. For instance, as suggested by Newey and West (1994), we can set L equal to $4(T/100)^{(2/9)}$. Note that $\Omega_h^{NW} = \Omega_{-h}^{NW\prime}$. We regard $\Omega^{NW} = (\Omega_h^{NW})$ as the "population banding approximation".

Next, we focus on the $N \times N$ block matrix $\Omega_h = Eu_t u_{t-h}'$ to control cross-sectional correlations. Under the intuition that u_{it} is cross-sectional weakly dependent, we assume Ω_h is a sparse matrix, that is, $\Omega_{h,ij} = Eu_{it}u_{j,t-h}$ is "small" for "many" pairs (i,j). Then Ω_h can be approximated by a sparse matrix $\Omega_h^{BL} = (\Omega_{h,ij}^{BL})_{N\times N}$ (Bickel and Levina (2008a)), where

$$\Omega_{h,ij}^{BL} = \begin{cases} Eu_{it}u_{j,t-h}, & \text{if } |Eu_{it}u_{j,t-h}| > \tau_{ij} \\ 0, & \text{if } |Eu_{it}u_{j,t-h}| \leq \tau_{ij}, \end{cases}$$

for some pre-determined threshold $\tau_{ij} \to 0$. We regard Ω_h^{BL} as the "population sparse approximation".

In summary, we approximate Ω by an $NT \times NT$ matrix $(\widetilde{\Omega}_{t,s}^{NT})$, where each block $\widetilde{\Omega}_{t,s}^{NT}$ is an $N \times N$ matrix, defined as: for $h = t - s$,

$$\widetilde{\Omega}_{t,s}^{NT} := \begin{cases} \Omega_h^{BL}, & \text{if } |h| \leq L \\ 0, & \text{if } |h| > L. \end{cases}$$

Therefore, we use "banding" to control the serial correlation, and "sparsity" to control the cross-sectional correlation.

2.2 Implementation of Feasible GLS

2.2.1 The estimator of Ω and FGLS

Given the intuition of the population approximation, we construct the large covariance estimator as follows. First, we denote the OLS estimator of β by $\widehat{\beta}_{OLS}$ and the corresponding residuals by $\widehat{u}_{it} = y_{it} - x_{it}'\widehat{\beta}_{OLS}$.

Now we estimate the $N \times N$ block matrix $\Omega_h = Eu_t u_{t-h}'$. To do so, let

$$\widetilde{R}_{h,ij} = \begin{cases} \frac{1}{T}\sum_{t=h+1}^{T} \widehat{u}_{it}\widehat{u}_{j,t-h}, & \text{if } h \geq 0 \\ \frac{1}{T}\sum_{t=1}^{T+h} \widehat{u}_{it}\widehat{u}_{j,t-h}, & \text{if } h < 0 \end{cases}, \quad \text{and } \widetilde{\sigma}_{h,ij} = \begin{cases} \widetilde{R}_{h,ii}, & \text{if } i = j \\ s_{ij}(\widetilde{R}_{h,ij}), & \text{if } i \neq j, \end{cases}$$

where $s_{ij}(\cdot) : \mathbb{R} \to \mathbb{R}$ is a "soft-thresholding function" with an entry dependent threshold τ_{ij}

such that

$$s_{ij}(z) = \text{sgn}(z)(|z| - \tau_{ij})_+,$$

where $(x)_+ = x$ if $x \geq 0$, and zero otherwise. Here $\text{sgn}(\cdot)$ denotes the sign function, and other thresholding functions, e.g., hard thresholding, are possible. For the threshold value, we specify

$$\tau_{ij} = M\gamma_T\sqrt{|\widetilde{R}_{0,ii}|\,|\widetilde{R}_{0,jj}|},$$

for some pre-determined value $M > 0$, where $\gamma_T = \sqrt{\frac{\log(LN)}{T}}$ is such that $\max_{h\leq L}\max_{i,j\leq N}|\widetilde{R}_{h,ij} - Eu_{it}u_{i,t-h}| = O_P(\gamma_T)$. Note that here we use an entry dependent threshold τ_{ij}, which may vary across (i,j). Then define

$$\widetilde{\Omega}_h = (\widetilde{\sigma}_{h,ij})_{N\times N}. \tag{2.4}$$

Next, we define the (t,s)th block $\widehat{\Omega}_{t,s}$ as an $N \times N$ matrix: for $h = t - s$,

$$\widehat{\Omega}_{t,s} = \begin{cases} \omega(|h|, L)\widetilde{\Omega}_h, & \text{if } |h| \leq L \\ 0, & \text{if } |h| > L. \end{cases}$$

Here $\omega(h, L)$ is the kernel function (see Andrews (1991) and Newey and West (1994)). We let $\omega(h, L) = 1 - h/(L + 1)$ be the Bartlett kernel function, where L is the bandwidth. Our final estimator of Ω is an $NT \times NT$ matrix:

$$\widehat{\Omega} = (\widehat{\Omega}_{t,s}).$$

Here $\widehat{\Omega}$ is a nonparametric estimator, which does not require an assumed parametric structure on Ω.

Finally, given $\widehat{\Omega}$, we propose the feasible GLS (FGLS) estimator of β as

$$\widehat{\beta}_{FGLS} = [X'\widehat{\Omega}^{-1}X]^{-1}X'\widehat{\Omega}^{-1}Y.$$

Note that the above defined FGLS estimator leaves two quantities to be specified to applied researchers: (i) the constant $M > 0$ in the threshold value for τ_{ij}, and (ii) the Newey-West bandwidth L. We discuss the choice of these two quantities in Section 2.2.2 below.

Remark 2.1 (Universal thresholding)**.** We apply thresholding separately to the $N \times N$ blocks, $(\widetilde{\sigma}_{h,ij})_{N\times N}$, which are estimated lagged blocks for $Eu_tu_{t-h} : h = 0, 1, 2, ...$ This allows the cluster-membership to be potentially changing over-time, that is, the identities of zeros and nonzero elements of Eu_tu_{t-h} can change over h. If it is known that the cluster-membership (i.e., identities of nonzero elements) is time-invariant, then one would set $\widetilde{\sigma}_{h,ij} =$

0 if $\max_{h\leq L}|\widetilde{R}_{h,ij}|\leq\tau_{ij}$ for $i\neq j$. This potentially would increase the finite sample accuracy of identifying the cluster-membership.

2.2.2 Choice of tuning parameters

Our suggested covariance matrix estimator, $\widehat{\Omega}$, requires the choice of tuning parameters L and M, which are the bandwidth and the threshold constant respectively. We write $\widehat{\Omega}(M,L)=\widehat{\Omega}$, where the covariance estimator depends on M and L. First, to choose the bandwidth L, we suggest using $L^*=4(T/100)^{2/9}$, which is proposed by Newey and West (1994). For a small size of T, we also recommend $L\leq 3$.

As for the choice of the thresholding constant M, our recommended rule-of-thumb choice is any constant that is on the interval $[0.5,2]$. Based on our simulations in extensive studies with various values for N and T, we find that $M=1.8$ is a universally good choice.

Alternatively, M can also be chosen through multifold cross-validation. To discuss this procedure, let us randomly split the data P times. We divide the data into $P=\log(T)$ blocks $J_1,...,J_P$ with block length $T/\log(T)$ and take one of the P blocks as the validation set. At the pth split, we denote by $\widetilde{\Omega}_0^p$ the sample covariance matrix based on the validation set, defined by $\widetilde{\Omega}_0^p=|J_p|^{-1}\sum_{t\in J_p}\widehat{u}_t\widehat{u}_t'$. Let $\widetilde{\Omega}_0^{S,p}(M)$ be the thresholding estimator with threshold constant M using the training data set $\{\widehat{u}_t\}_{t\notin J_p}$. Finally, we choose the constant M^* by minimizing the cross-validation objective function

$$M^*=\arg\min_{c<M<\bar{C}}\frac{1}{P}\sum_{j=1}^{P}\|\widetilde{\Omega}_0^{S,p}(M)-\widetilde{\Omega}_0^p\|_F^2,$$

where $\bar{C}$ is a large constant such that $\widetilde{\Omega}_0^S(\bar{C})$ is a diagonal matrix, and can be fixed as, e.g., $\bar{C}=3$; c is a constant that guarantees the positive definiteness of $\widehat{\Omega}(M,L)$ for $M>c$: for each fixed L,

$$c=\inf[M>0:\lambda_{\min}\{\widehat{\Omega}(C,L)\}>0,\forall C>M].$$

Here $\widetilde{\Omega}_0^S(M)$ is the soft-thresholded estimator as defined in the equation (2.4). Then the resulting estimator of Ω is $\widehat{\Omega}(M^*,L^*)$. To determine this value, one can plot $\lambda_{\min}\{\widehat{\Omega}(C,L)$ as a function of C, fixing $L=L^*$, and visually determine c.

In summary, Table 1 summarizes the recommended quantities for implementing the proposed FGLS estimator.

2.2.3 Incorporating known clusters

Note that an advantage of the method proposed in this paper is that it does not assume known cluster information (i.e., the number of clusters and the membership of clusters). On the other hand, when clustering information is available, this method can be modified to

Table 1: Recommended choices for implemtations

quantities	$s_{ij}(z)$	$\omega(\|h\|, L)$	τ_{ij}	γ_T	L
choice	$\text{sgn}(z)(\|z\| - \tau_{ij})_+$	$1 - \frac{\|h\|}{L+1}$	$M\gamma_T\sqrt{\|\widetilde{R}_{0,ii}\|\,\|\widetilde{R}_{0,jj}\|}$	$\sqrt{\frac{\log(LN)}{T}}$	$4(T/100)^{2/9}$
quantities	M (rule-of-thumb)	P	$\bar{C}$	c	
choice	1.8	$\log T$	3	visually by plotting $\lambda_{\min}\{\widehat{\Omega}(C, L)$	

Here P, $\bar{C}$ and c are required constants choosing M using cross-validations.

take into account that information, and is particularly suitable when the number of clusters is small, and the size of each cluster is large.

For example, let $C_1, ..., C_G$ be disjoint subsets of $\{1, ..., N\}$, so that they are *known* clusters and that u_{it} and u_{js} are uncorrelated if i and j belong to different clusters for any (t, s). Then naturally we can re-arranged the $N \times N$ matrix $\Omega_h = Eu_t u'_{t-h}$ so that it can be decomposed into G disjoint blocks on the diagonal and off-diagonal blocks are zeros:

$$\Omega_h = \begin{pmatrix} \Omega_{h,1} & & \\ & \ddots & \\ & & \Omega_{h,G} \end{pmatrix}.$$

It is assumed that G is small while the size of each diagonal block matrix is large. Within the g th ($g \leq G$) diagonal block matrix, say $\Omega_{h,g}$, we apply thresholding to further reduce the dimensionality. So we estimate $\Omega_{h,g}$ by $\widetilde{\Omega}_{h,g} = (\widetilde{\sigma}_{h,g,ij})$, where

$$\widetilde{\sigma}_{h,g,ij} = \begin{cases} \widetilde{R}_{h,ii}, & \text{if } i = j, \text{ and } i, j \in C_g \\ s_{ij}(\widetilde{R}_{h,ij}), & \text{if } i \neq j, \text{ and } i, j \in C_g. \end{cases}$$

Putting these estimated diagonal blocks together, we obtain $\widetilde{\Omega}_h$, the estimated Ω_h.

The within-cluster thresholding then allows unknown correlations within each cluster. In contrast, the conventional clustered standard errors lose a lot of degrees of freedom when the size of cluster is too large (because each cluster is effectively treated as a "single observation"), resulting in conservative confidence intervals. See Cameron and Miller (2015) for more discussions.

Moreover, when the number of clusters is large, and the size of each cluster is small, then this is the usual setting of cluster standard errors. One does not need to apply thresholding, as the known clusters naturally form small diagonal blocks on Ω_h. Because the size of these blocks are small, sufficient degrees of freedom is kept and it is then straightforward to estimate Ω_h.

2.3 The effect of $\widehat{\Omega}^{-1} - \Omega^{-1}$

A key step of proving the asymptotic property for $\widehat{\beta}_{FGLS}$ is to show that it is asymptotically equivalent to $\widetilde{\beta}^{inf}_{GLS}$, that is:

$$\frac{1}{\sqrt{NT}} X'(\widehat{\Omega}^{-1} - \Omega^{-1})U = o_P(1). \tag{2.5}$$

In the usual low-dimensional settings that involve estimating optimal weight matrix, such as the optimal GMM estimations, it has been well known that consistency for the inverse covariance matrix estimator is sufficient for the first-order asymptotic theory, e.g., Hansen (1982), Newey (1990), Newey and McFadden (1994). It turns out, when the covariance matrix is of high-dimensions, not even the optimal convergence rate of $\|\widehat{\Omega} - \Omega\|$ is sufficient. In fact, proving equation (2.5) is a very challenging problem. In the general case when both cross-sectional and serial correlations are present, our strategy is to use a careful expansion for $\frac{1}{\sqrt{NT}} X'(\widehat{\Omega}^{-1} - \Omega^{-1})U$. We shall proceed in two steps:
Step 1: Show that $\frac{1}{\sqrt{NT}} X'(\widehat{\Omega}^{-1} - \Omega^{-1})U = \frac{1}{\sqrt{NT}} W'(\widehat{\Omega} - \Omega)\varepsilon + o_P(1)$, where $W = \Omega^{-1}X$, and $\varepsilon = \Omega^{-1}U$.
Step 2: Show that $\frac{1}{\sqrt{NT}} W'(\widehat{\Omega} - \Omega)\varepsilon = o_P(1)$.

Now we suppose $\omega(h, L) = 1, \Omega \approx \Omega^{NW}$ and let $A_{b_h} = \{(i,j) : |Eu_{it}u_{j,t-h}| \neq 0\}, A_{s_h} = \{(i,j) : |Eu_{it}u_{j,t-h}| = 0\}$. As for Step 2, we shall show,

$$\frac{1}{\sqrt{NT}} W'(\widehat{\Omega} - \Omega)\varepsilon \approx \frac{1}{\sqrt{NT}} \sum_{|h| \leq L} \sum_{i,j \in A_{b_h}} \sum_{t=h+1}^{T} w_{it}\varepsilon_{j,t-h} \frac{1}{T} \sum_{s=h+1}^{T} (u_{is}u_{j,s-h} - Eu_{it}u_{j,t-h}). \tag{2.6}$$

Here w_{it} is defined such that, we can write $W = (w_1', \cdots, w_T')'$ with w_t being an $N \times d$ matrix of w_{it}; ε_{it} is defined similarly. While proving (2.6) to be $o_P(1)$, in the presence of both serial and cross-sectional correlations, is very technically challenging. We thus directly assume it is $o_P(1)$ as a high-level condition (see Assumption 2.4 in Section 2.4 below). To appreciate the need of this high-level condition, let us consider a simple example as follows.

A simple example. To illustrate the key technical issue, consider a simple and ideal case where u_{it} is known, and independent across both i and t, but with cross-sectional heteroskedasticity. In this case, the covariance matrix of the $NT \times 1$ vector U is a diagonal matrix, with diagonal elements $\sigma_i^2 = Eu_{it}^2$:

$$\Omega = \begin{pmatrix} D & & \\ & \ddots & \\ & & D \end{pmatrix}, \text{ where } D = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_N^2 \end{pmatrix}.$$

Then a natural estimator for Ω is

$$\widehat{\Omega} = \begin{pmatrix} \widehat{D} & & \\ & \ddots & \\ & & \widehat{D} \end{pmatrix}, \text{ where } \widehat{D} = \begin{pmatrix} \widehat{\sigma}_1^2 & & \\ & \ddots & \\ & & \widehat{\sigma}_N^2 \end{pmatrix},$$

and $\widehat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T} u_{it}^2$, because u_{it} is known. Then the GLS becomes:

$$(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}x_{it}'\widehat{\sigma}_i^{-2})^{-1}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}y_{it}\widehat{\sigma}_i^{-2}.$$

A key step is to prove that the effect of estimating D is asymptotically negligible:

$$\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}u_{it}(\widehat{\sigma}_i^{-2} - \sigma_i^{-2}) = o_P(1). \tag{2.7}$$

It can be shown that the problem reduces to proving:

$$A \equiv \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}u_{it}\sigma_i^{-2}(\frac{1}{T}\sum_{s=1}^{T}(u_{is}^2 - Eu_{is}^2))\sigma_i^{-2} = o_P(1). \tag{2.8}$$

Under the simplified conditions of this example (u_{it} is independent across both i and t), it is straightforward to calculate var(A) and show that it converges to zero as $N, T \to \infty$ regardless of whether $N < T$ or not.

As for EA, straightforward calculations yield

$$EA = \frac{\sqrt{NT}}{T}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} E(x_{it}E(u_{it}^3|x_{it}))\sigma_i^{-4}.$$

Generally, if $u_{it}|x_{it}$ is non-Gaussian and asymmetric, $E(u_{it}^3|x_{it}) \neq 0$. Hence we require $N/T \to 0$ to have $EA \to 0$. Hence, to allow for non-Gaussian and asymmetric conditional distributions, in the GLS setting it turns out $N = o(T)$ is required.

We shall not explicitly impose $N = o(T)$ in this paper as a formal assumption, but instead impose Assumption 2.4. On one hand, when the distribution of u_{it} is symmetric, we *do not* require $N = o(T)$ because as is shown in the above example, $E(u_{it}^3|x_{it}) = 0$ is sufficient for $EA \to 0$, and is satisfied by symmetric distributions. On the other hand, when u_{it} is non-symmetric, Assumption 2.4 then implicitly requires $N = o(T)$. Note that $N = o(T)$ is a strong assumption in many microeconomic applications for panel data models. But as illustated in the above simple example, if $u_{it}|x_{it}$ is not symmetric, it is required for feasible GLS even if Ω is diagonal. One possible approach to weakening this assumption is to remove

the higher order bias from $\widehat{\Omega}$. Higher order debiasing is a complicated procedure in the presence of general weak dependences. This is left for future research.

2.4 Asymptotic results of FGLS

We impose the following conditions, regulating the sparsity and serial weak dependence.

Assumption 2.1. *(i) $\{u_t, x_t\}_{t\geq 1}$ is strictly stationary. In addition, each u_t has zero mean vector, and $\{u_t\}_{t\geq 1}$ and $\{x_t\}_{t\geq 1}$ are independent.*
(ii) There are constants $c_1, c_2 > 0$ such that $\lambda_{\min}(\Omega_h) > c_1$ and $\|\Omega_h\|_1 < c_2$ for each fixed h.
(iii) Exponential tail: There exist $r_1, r_2 > 0$ and $b_1, b_2 > 0$, and for any $s > 0, i \leq N$ and $l \leq d$,

$$P(|u_{it}| > s) \leq exp(-(s/b_1)^{r_1}), \quad P(|x_{it,l}| > s) \leq \exp(-(s/b_2)^{r_2}).$$

(iv) Strong mixing: There exist $\kappa \in (0,1)$ such that $r_1^{-1} + r_2^{-1} + \kappa^{-1} > 1$, and $C > 0$ such that for all $T > 0$,

$$\sup_{A\in\mathcal{F}^0_{-\infty}, B\in\mathcal{F}^\infty_T} |P(A)P(B) - P(AB)| < \exp(-CT^\kappa),$$

where $\mathcal{F}^0_{-\infty}$ and $\mathcal{F}^\infty_T$ denote the σ-algebras generated by $\{(x_t, u_t) : t \leq 0\}$ and $\{(x_t, u_t) : t \geq T\}$ respectively.

Condition (ii) requires that Ω_h be well conditioned. Condition (iii) ensures the Bernstein-type inequality for weakly dependent data, which requires the underlying distributions to be thin-tailed. Condition (iv) is the standard α-mixing condition, adapted to the large-N panel. In addition, we impose the following regularity conditions.

Assumption 2.2. *(i) There exists a constant $C > 0$ such that for all $i \leq N$ and $t \leq T$, $E\|x_{it}\|^4 < C$ and $Eu_{it}^4 < C$.*
(ii) Define $\xi_T(L) = \max_{t\leq T} \sum_{|h|>L} \|Eu_t u'_{t-h}\|$. Then $\xi_T(L) \to 0$.
(iii) Define $f_T(L) = \max_{t\leq T} \sum_{|h|\leq L} \|Eu_t u'_{t-h}(1 - \omega(|h|, L))\|$. Then $f_T(L) \to 0$.

Assumption 2.2 allows us to prove the convergence rate of the covariance matrix estimator. Condition (ii) is an extension of the standard weak serial dependence condition to the high-dimensional case in panel data literature. It allows us to employ banding or Newey-West trunction procedure. Condition (iii) is well satisfied by various kernel functions for the HAC-type estimator. For the Bartlett kernel, for example,

$$\max_{t\leq T} \sum_{|h|\leq L} \|Eu_t u'_{t-h}(1 - \omega(|h|, L))\| \leq \frac{1}{L} \max_{t\leq T} \sum_{|h|=0}^{\infty} \|Eu_t u'_{t-h}\||h|$$

converges to zero as $L \to \infty$ as long as $\max_{t\leq T} \sum_{|h|=0}^{\infty} \|Eu_t u'_{t-h}\||h| < \infty$.

In this paper, we assume Ω_h to be a sparse matrix for each h and impose similar conditions as those in Bickel and Levina (2008a) and Fan, Liao, and Mincheva (2013): write $\Omega_h = (\Omega_{h,ij})_{N\times N}$, where $\Omega_{h,ij} = Eu_{it}u_{j,t-h}$. For some $q \in [0,1)$, we define

$$m_N = \max_{|h|\le L}\max_{i\le N}\sum_{j=1}^{N}|\Omega_{h,ij}|^q,$$

as a measurement of the sparsity. We would require that m_N should be either fixed or grow slowly as $N \to \infty$. In particular, when $q = 0$, $m_N = \max_{|h|\le L}\max_{i\le N}\sum_{j=1}^{N} 1(\Omega_{h,ij} \neq 0)$, which corresponds to the exact sparsity case.

Let

$$\gamma_T = \sqrt{\log(LN)/T}.$$

Assumption 2.3. *For any $NT \times NT$ matrix M, we denote $(M)_{ts,ij}$ as the (i,j)th element of the (t,s)th block of the matrix M.*
(i) $\sum_{|h|>L}\|\Omega_h\|_1 = O(L^{-\alpha})$, *for a constant* $\alpha > 0$.
(ii) $\max_{i\le N,t\le T}\sum_{s=1}^{T}\sum_{j=1}^{N}|(\Omega^{-1})_{ts,ij}| = O(1)$.
(iii) There is $q \in [0,1)$ *such that* $Lm_N\gamma_T^{1-q} = o(1)$ *holds. In addition,*

$$\sqrt{T}L^2m_N^2\gamma_T^{3-2q} = o(1), \quad \text{and } \sqrt{NT}L^3m_N^3\gamma_T^{3-3q} = o(1).$$

(iv) $\sqrt{NT}(\xi_T(L) + f_T(L))^3 = o(1)$ *and* $L^{-\alpha}T\sqrt{NT}m_N\gamma_T^{1-q} = o(1)$.

Conditions (i)-(ii) require the weak cross-sectional correlations. Condition (iii) is about the sparsity assumptions on the growth of m_N, associated with q and the speed of L.

Remark 2.2. To understand Assumption 2.3, consider a simple case where $Eu_{it}u_{j,t-h}$ is nonzero for only finitely many pairs $i \neq j$. This corresponds to $q = 0$ and $m_N = O(1)$. Then condition (iii) requires

$$\sqrt{N}L^3\log^{3/2}(LT) = o(T).$$

In practice, the bandwidth L and $\log(LT)$ are both grow very slowly compared to N and T. So essentially this condition requires $N = o(T^2)$. In addition, condition (iv) assumes that the autocorrelations should decay sufficiently fast as $L \to \infty$. Suppose both $\zeta_T(L)$ and $f_T(L)$ decay in a polynomial rate of L (e.g., with order L^{-c_0}), then this condition require that the order of the polynomial, c_0, should be sufficiently large.

All the above conditions, Theorem A.1 in the appendix gives the convergence of $\|\widehat{\Omega} - \Omega\|$. It then leads to the following proposition.

Proposition 2.1. *Under the Assumption 2.1-2.2, for $q \in [0,1)$ and $\alpha > 0$ such that Assumption 2.3 holds,*

$$\sqrt{NT}(\widehat{\beta}_{FGLS} - \beta) = \Gamma^{-1}\left(\frac{1}{\sqrt{NT}}X'\Omega^{-1}U\right) + \Gamma^{-1}\left(\frac{1}{\sqrt{NT}}X'\Omega^{-1}(\widehat{\Omega} - \Omega)\Omega^{-1}U\right) + o_P(1),$$

where $\Gamma = E(X'\Omega^{-1}X/NT)$.

As we see from the above proposition, the effect $\widehat{\Omega} - \Omega$ also appears as an "weighted average" in the second term on the right-hand-side of the expansion. The negligibility of this term relies on the following high-level condition. We define $W = \Omega^{-1}X$ and $\varepsilon = \Omega^{-1}U$. Then, $W = (w_1', \cdots, w_T')'$ with w_t being an $N \times d$ matrix of w_{it}, and ε_{it} is defined similarly.

Assumption 2.4. *Let $A_{b_h} = \{(i,j) : |Eu_{it}u_{j,t-h}| \neq 0\}$. Then*

$$\left\| \frac{1}{\sqrt{NT}} \sum_{h=0}^{L} \sum_{i,j \in A_{b_h}} \mathbb{G}^1_{T,ij}(h)\mathbb{G}^2_{T,ij}(h) \right\| = o_P(1), \tag{2.9}$$

where $\mathbb{G}^1_{T,ij}(h) = \frac{1}{\sqrt{T}}\sum_{t=h+1}^{T}(u_{it}u_{j,t-h} - Eu_{it}u_{j,t-h})$ and $\mathbb{G}^2_{T,ij}(h) = \frac{1}{\sqrt{T}}\sum_{t=h+1}^{T} w_{it}\varepsilon_{j,t-h}$.

Remark 2.3. While it is difficult to verity the above high-level condition in the presence of either serial dependence or cross-sectional dependence or both, the intuition can be understood in the simple i.i.d. case. Suppose u_{it} is independent across both i and t. Then we can set $L = 0$ and this condition becomes

$$A \equiv \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \frac{1}{T} \sum_{t=1}^{T} (u_{it}^2 - Eu_{it}^2) \sum_{s=1}^{T} x_{is}u_{is}\sigma_i^{-4} = o_P(1),$$

which is (2.8) as we discussed in Section 2.3. As discussed therein, it is straightforward to see that $\text{var}(A) = o(1)$, proving $EA = o(1)$ requires *either* u_{it} has a symmetric distribution so that $Eu_{it}^3 = 0$, *or* $N = o(T)$ for asymmetric distributions. Similar conditions were required for high-dimensional GLS problems, for instance, by Bai and Liao (2017) in panel data with interactive effect estimations.

Then we have the following limiting distribution by using the result of Theorem A.1.

Theorem 2.1. *Suppose $\text{var}(U|X) = \text{var}(U) = \Omega$. Under the Assumptions 2.1-2.4, for $q \in [0,1)$ and $\alpha > 0$ such that Assumption 2.3 holds, as $N, T \to \infty$,*

$$\sqrt{NT}(\widehat{\beta}_{FGLS} - \beta) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1}),$$

where $\Gamma = E(X'\Omega^{-1}X/NT)$. The consistent estimator of Γ is $\widehat{\Gamma} = X'\widehat{\Omega}^{-1}X/NT$.

The asymptotic variance of the FGLS estimator is $\text{Avar}(\widehat{\beta}_{FGLS}) = \Gamma^{-1}/NT$, and an estimator of it is $(X'\widehat{\Omega}^{-1}X)^{-1}$. Asymptotic standard errors can be obtained in the usual fashion from the asymptotic variance estimates.

3 Monte Carlo evidence

3.1 DGP and methods

In this section we compare the proposed FGLS estimator with OLS estimator. We consider the fixed effect linear regression model, although this paper focuses on the simple linear model for technical simplicity. Hence the de-meaning procedure is applied first. The data generating process (DGP) used for the simulations is given by

$$y_{it} = \alpha_i + \mu_t + \beta_0 x_{it} + u_{it},$$

where the true $\beta_0 = 1$ and fixed effects α_i, μ_t are generated from $\mathcal{N}(0, 0.5)$. The DGP allows for serial and cross-sectional correlation in both x_{it} and u_{it}, which are generated by $(NT) \times (NT)$ covariance matrices, Ω_X and Ω_U, as follows: let $R_\eta = (R_{\eta,ij})$ denote an $N \times N$ block diagonal correlation matrix. We fix the number of clusters as $G = 25$. Hence, each diagonal block is a $N/G \times N/G$ matrix with the off-diagonal entries (i, j) in the same cluster, $R_{\eta,ij}$ for $i \neq j$, which are generated from i.i.d. Uniform$(0, \gamma)$. In this study, we set the level of cross-sectional correlation in each cluster as $\gamma = 0.3$, or 0.7. For the cross-sectional heteroskedasticity, let $D = \text{diag}\{d_i\}$, where $\{d_i\}_{i \leq N}$ are i.i.d. Uniform(1,m). Finally, we define the $N \times N$ covariance matrix of u_t as $\Sigma_u = DR_\eta D$. In this case, we report results when $m = \sqrt{5}$. For the covariance matrix of the regressor, we simply set $\Sigma_x = R_\eta$, which does not have heteroskedasticity.

Now we introduce i-dependent serial correlation for the regressor and the error as follows: first let $\sigma_{ii} = \rho_i$ if $i = j$ and $\sigma_{ij} = \rho_i \rho_j$ if $i \neq j$. Then we define the $(NT) \times (NT)$ covariance matrix, $\Omega_U = (\Omega_{t,s})$. The (t, s)th block is an $N \times N$ covariance matrix, given by $\Omega_{t,s} = (\Omega_{t,s}(i, j))$, where $\Omega_{t,s}(i, j) = \Sigma_{u,ij}\sigma_{ij}^{|t-s|}$. The large covariance matrix of the regressor, Ω_X, is generated similarly. The level of i-dependent ρ_i of the regressor and the error is generated from i.i.d. Uniform$(0, 0.6)$, seperately.

Note that the (t, s)th block covariance decays exponentially as $|t-s|$ increases. Finally we generate the $NT \times 1$ vectors $(u_1', ..., u_T')' = \Omega_U^{1/2}\zeta$, where ζ is an $NT \times 1$ vector, whose entries are generated from i.i.d. $\mathcal{N}(0, 5)$. Similarly, the regressor is generated by $(x_1', ..., x_T')' = \Omega_X^{1/2}\xi$, where ξ is an $NT \times 1$ vector, whose entries are generated from i.i.d. $\mathcal{N}(0, 1)$. Note that x_{it} is uncorrelated with u_{it}.

In this numerical study, we use sample sizes $N = 50, 100$ and $T = 50, 100, 150$, and the

simulation is replicated for one thousand times in all cases.[2] For each $\{N, T\}$ combination, we set the bandwidth $L = 3$ in all cases. The threshold constant, M, is obtained by the cross-validation method as suggested in Section 2.2.2. For instance, when $T = 100$, the number of folds to split is $\log(100) \approx 5$. In general, the cross-validation chooses M between 1.4 and 1.8. Interestingly, as the level of cross-section correlation increases, the cross-validation tends to choose smaller M, so that the number of non-thresholded elements increases. Hence it takes into account the strength of cross-sectional correlation. We use the Bartlett kernel for our FGLS estimator. Results are summarized in Tables 2-3.

3.2 Results

Tables 2-3 present the simulation results, where each table corresponds to a different level of cross-sectional correlation, $\gamma = \{0.3, 0.7\}$. In each table, the mean and standard deviation of the estimators are reported. FGLS(Diag) refers to the FGLS estimator using the diagonal covariance matrix, which only takes into account heteroskedasticity. RMSE is the ratio of the mean squared error of FGLS to that of OLS. The mean and standard deviation of the estimated standard errors for OLS and FGLS are also reported. The robust unknown clustered standard error, suggested by Bai, Choi, and Liao (2019), is used for OLS. For FGLS, we report the results of the standard error as introduced in Theorem 2.1. The difference between the standard deviation of the estimators and the mean of standard errors can be explained as the bias of estimated standard errors. In addition, we present null rejection probabilities for the 5% level tests using the traditional $\mathcal{N}(0, 1)$ critical value based on each standard errors.

According to Tables 2-3, we see that both methods are almost unbiased, while our proposed FGLS has indeed smaller standard deviation of $\widehat{\beta}$ than that of OLS and FGLS(Diag). In all cases, the RMSE of our proposed FGLS is significantly smaller than one. Hence the results confirm that the FGLS estimator is more efficient than the OLS and the FGLS(Diag) estimators in presence of heteroskedasticity, serial and cross-sectional correlations. Regarding the t-test, in Table 2, the rejection probabilities of FGLS and OLS are close to 0.05 when T is large, while those of FGLS(Diag) tend to over-reject. Since the FGLS(Diag) estimator does not take into account the serial and the cross-sectional correlations, its standard errors are underestimated. On the other hand, in Table 3, we find that the standard errors of all estimators are underestimated and the t-test rejection probabilities are much larger than 0.05, especially when T is relatively smaller than N (e.g., $N = 100$ and $T = 50$). This is due to the strong cross-sectional correlation within clusters. However, the rejection probabilities of FGLS and OLS are much smaller than those of FGLS(Diag). In summary, FGLS does

[2]The procedure of proposed estimators require use of an $NT \times NT$ matrix as discussed in Section 2.2. Indeed, when NT is large, the procedure appears to be computationally demanding. Hence, we focus on the small sample size in this study.

improve efficiency in terms of mean squared error; also we obtain unbiased standard error estimator and appropriate rejection rate as T increases.

4 Empirical study: Effects of divorce law reforms on divorce rates

In the literature, the cause of the sharp increase in the U.S. divorce rate in the 1960-1970s is an important research question. During 1970s, more than half of states in the U.S. liberalized the divorce system, and the effects of reforms on divorce rates have been investigated by many such as Allen (1992) and Peters (1986). With controls for state and year fixed effects, Friedberg (1998) suggested that state law reforms siginificantly increased divorce rates. Also, she assumed that unilateral divorce laws affected divorce rates permanently. However, divorce rates from 1975 have been subsequently decreasing according to empirical evidence. Therefore the question of whether law reforms also affect the divorce rate decrease has arisen. Wolfers (2006) revisited this question by using a treatment effect panel data model, and identified only temporal effects of reforms on divorce rates. In particular, he used dummy variables for the first two years after the reforms, 3-4 years, 5-6 years, and so on. More specifically, the following fixed effect panel data model was considered:

$$y_{it} = \alpha_i + \mu_t + \sum_{k=1}^{8} \beta_k X_{it,k} + \delta_i t + u_{it}, \tag{4.1}$$

where y_{it} is the divorce rate for state i and year t, α_i a state fixed effect, μ_t a time fixed effects, and $\delta_i t$ a linear time trend with unknown coefficient δ_i. X_{it} is a binary regressor which denotes the treatment effect $2k$ years after the reform. Wolfers (2006) suggested that "the divorce rate rose sharply following the adoption of unilateral divorce laws, but this rise was reversed within about a decade". He also concluded that "15 years after reform the divorce rate is lower as a result of the adoption of unilateral divorce, although it is hard to draw any strong conclusions about long-run effects".

Both Friedberg (1998) and Wolfers (2006) used a weighted model by muliplying all variables by the square root of state population. In addition, they used ordinary OLS standard error, which does not take into account heteroskedasticity, serial and cross-sectional correlations. However, standard errors might be biased when one disregards these correlations. Therefore, we re-estimated the model of Wolfers (2006) using the proposed FGLS method and OLS with the heteroskedastic standard errors of White (1980), the clustered standard error of Arellano (1987), and the robust standard error of Bai, Choi, and Liao (2019). Note that the model (4.1) is more complicated than the model we formally studied in this paper, by not only including fixed effect, but also linear time trends. While theoretical studies of models

with trends might be challenging in the high-dimensional GLS setting, it is straightforward to implement it in the same FGLS framework by applying a projection transformation to eliminate the time trend. Specifically, let $\ell = (1, 2, ..., T)'$ and $P_\ell = I_T - \ell(\ell'\ell)^{-1}\ell'$. We can define $\widetilde{Y}_i = P_\ell(y_{i1}, ..., y_{iT})'$, and $\widetilde{X}_i = P_\ell(X_{i1}, ..., X_{iT})'$, and define $\dot{\widetilde{y}}_{it}$ and $\dot{\widetilde{X}}_{it}$ accordingly from $\widetilde{y}_{it}$ and $\widetilde{X}_{it}$ that further remove the fixed effects.

The same dataset as in Wolfers (2006) is used, which includes the divorce rate, state-level reform years, binary regressors, and state population. Due to missing observations around divorce law reforms, we exclude Indiana, New Mexico and Louisiana. As a result, we obtain balanced panel data from 1956 to 1988 for 48 states. We fit the models both with and without linear time trend, and use OLS and FGLS in each model to estimate β. In the FGLS estimation, we set bandwidth $L = 3$ as proposed by Newey and West (1994) ($L = 4(T/100)^{2/9}$). The thresholding values are chosen by the cross-validation method as discussed in Section 2.2.2, more specifically, $M = 1.8$ and $M = 1.9$ for the model with and without linear time trends, respectively. The Bartlett kernel is used in the OLS robust standard error and FGLS estimation. The estimated $\beta_1, \cdots, \beta_8$ with and without linear time trend and standard errors are summarized in Table 4 below.

The OLS and FGLS estimates in both models are similar to each other. The results show that divorce rates rose soon after the law reform. However, within a decade, divorce rates had fallen over time. Interestingly, FGLS confirms the negative effects of the law reforms on the divorce rates, specifically, 11-15+ years after the reform in the model with state-specific linear time trends, and 9-15+ years after the reform in the model without state-specific linear time trends. In addition, the FGLS estimates for 1-6 and 1-4 years are positive and statistically significant in the models with and without linear time trends, respectively. For OLS, the coefficient estimates for 3-4 and 7-15+ are significant in the model without linear time trends based on se_{BCL}. In contrast, the OLS estimates are statistically significant only for 1-4 years when a linear time trend is added. According to the clustered standard error, se_{CX}, note that only 11-15+ are statistically significant in the model without trends.

According to OLS and FGLS estimation results with and without a linear time trend, we make the following conlcusion: in the first 8 years, the overall trend of divorce rate is increasing, but the law reform reduces the divorce rate after 3-4 years. However, 8 years after the reform, we observe that the law reform has a negative effect on divorce rate. Note that Wolfers (2006) de-emphasized the negative coefficient at the end of the periods, as these are not robust to inclusion of state-specific quadratic trends, which we did not employ in this paper. Overall, the results of FGLS estimates are consistent with Wolfers (2006).

5 Conclusions

In this paper, we propose a large covariance matrix estimator and a modified version of FGLS that takes into account both serial and cross-sectional correlations in linear panel models that are robust to heteroskedasticity, serial and cross-sectional correlations. We derive the asymptotic distribution of FGLS, which incorporates the estimated high-dimensional covariance matrix. From simulated experiments, we confirmed that our FGLS estimates are more efficient than OLS estimates.

Table 2: Performance of estimated β_0; true $\beta_0 = 1$; i-dependent serial correlation and weak cross-sectional correlation ($\gamma = 0.3$).

		OLS	FGLS		OLS	FGLS		OLS	FGLS	
N	T		Diag	Our		Diag	Our		Diag	Our
		mean($\widehat{\beta}$)			std($\widehat{\beta}$)			RMSE		
50	50	1.001	1.002	1.001	0.080	0.075	0.069	1.000	0.883	0.740
	100	0.999	0.999	0.999	0.061	0.055	0.050	1.000	0.823	0.680
	150	1.000	1.000	1.000	0.045	0.041	0.038	1.000	0.842	0.710
100	50	1.000	1.000	1.001	0.058	0.053	0.050	1.000	0.842	0.745
	100	0.999	0.998	0.998	0.041	0.037	0.034	1.000	0.793	0.690
	150	1.000	1.000	1.000	0.034	0.029	0.027	1.000	0.749	0.628
		mean(s.e.)			std(s.e.)			t-test rejection prob.		
50	50	0.079	0.066	0.065	0.004	0.002	0.002	0.054	0.084	0.068
	100	0.058	0.048	0.047	0.002	0.001	0.001	0.052	0.090	0.073
	150	0.047	0.039	0.039	0.001	0.001	0.001	0.047	0.068	0.043
100	50	0.058	0.048	0.046	0.002	0.001	0.001	0.058	0.083	0.069
	100	0.040	0.034	0.033	0.001	0.000	0.000	0.053	0.069	0.067
	150	0.032	0.027	0.026	0.001	0.000	0.000	0.069	0.077	0.059

Note: OLS and FGLS comparison. RMSE is the ratio of the mean squared error of FGLS to that of OLS. The t-test rejection prob. is t-test rejection rates for 5% level tests. Robust standard error suggested by Bai, Choi, and Liao (2019) is used for OLS. Reported results are based on 1000 replications. The threshold value, M, is chosen through the cross-validation method as discussed in Section 2.2.2. For the bandwidth, we set $L = 3$.

Table 3: Performance of estimated β_0; true $\beta_0 = 1$; i-dependent serial correlation and strong cross-sectional correlation ($\gamma = 0.7$).

		OLS	FGLS		OLS	FGLS		OLS	FGLS	
N	T		Diag	Our		Diag	Our		Diag	Our
		mean($\widehat{\beta}$)			std($\widehat{\beta}$)			RMSE		
50	50	1.000	1.001	1.000	0.084	0.079	0.072	1.000	0.883	0.744
	100	0.999	1.000	1.000	0.063	0.057	0.052	1.000	0.817	0.677
	150	1.002	1.003	1.003	0.051	0.047	0.042	1.000	0.856	0.685
100	50	1.001	1.000	1.001	0.070	0.063	0.059	1.000	0.810	0.711
	100	0.998	0.999	0.999	0.048	0.044	0.042	1.000	0.840	0.742
	150	1.000	1.000	1.000	0.038	0.033	0.030	1.000	0.777	0.617
		mean(s.e.)			std(s.e.)			t-test rejection prob.		
50	50	0.079	0.066	0.065	0.004	0.002	0.002	0.065	0.100	0.082
	100	0.059	0.048	0.048	0.002	0.001	0.001	0.068	0.104	0.065
	150	0.047	0.039	0.039	0.001	0.001	0.001	0.065	0.110	0.064
100	50	0.059	0.048	0.048	0.002	0.001	0.001	0.105	0.143	0.125
	100	0.040	0.034	0.034	0.001	0.000	0.000	0.097	0.139	0.094
	150	0.032	0.027	0.028	0.001	0.000	0.000	0.101	0.123	0.079

Note: See notes to Table 2.

Table 4: Empirical application: effects of divorce law refrom with state and year fixed effects: US state level data annual from 1956 to 1988, dependent variable is divorce rate per 1000 persons per year. OLS and FGLS estimates and standard errors (using state population weights).

Effects:	$\widehat{\beta}_{OLS}$	se_W	se_{CX}	se_{BCL}	$\widehat{\beta}_{FGLS}$	se_{FGLS}
Panel A: Without state-specific linear time trends						
1–2 years	0.256	0.140	0.189	0.148	0.133	0.046*
3–4 years	0.209	0.081*	0.159	0.089*	0.165	0.056*
5–6 years	0.126	0.073	0.168	0.069	0.100	0.059
7–8 years	0.105	0.070	0.165	0.040*	0.026	0.061
9–10 years	-0.122	0.060*	0.161	0.054*	-0.129	0.061*
11–12 years	-0.344	0.071*	0.173*	0.075*	-0.253	0.062*
13–14 years	-0.496	0.074*	0.188*	0.062*	-0.324	0.063*
15+ years	-0.508	0.089*	0.223*	0.077*	-0.325	0.067*
Panel B: With state-specific linear time trends						
1–2 years	0.286	0.152	0.206	0.140*	0.171	0.044*
3–4 years	0.254	0.099*	0.171	0.126*	0.220	0.058*
5–6 years	0.186	0.102	0.206	0.143	0.175	0.067*
7–8 years	0.177	0.109	0.230	0.146	0.097	0.075
9–10 years	-0.037	0.111	0.241	0.154	-0.073	0.082
11–12 years	-0.247	0.128	0.268	0.183	-0.240	0.089*
13–14 years	-0.386	0.137*	0.295	0.209	-0.329	0.098*
15+ years	-0.414	0.158*	0.337	0.243	-0.382	0.108*

Note: Standard errors with asterisks indicate significance at 5% level using $N(0,1)$ critical values. For OLS standard errors, se_W and se_{CX} refer to the heteroskedastic standard errors by White (1980) and the clustered standard errors by Arellano (1987), respectively; se_{BCL} is the robust standard error suggested by Bai, Choi, and Liao (2019). The threshold values for FGLS by the cross-validation are $M = 1.9$ and $M = 1.8$ for Panel A and B, respectively.

A Appendix

Throughout the proof, $\max_i$, $\max_t$, $\max_h$, $\max_{ij}$, and $\max_{it}$ denote $\max_{i\leq N}$, $\max_{t\leq T}$, $\max_{h\leq L}$, $\max_{i\leq N,j\leq N}$, and $\max_{i\leq N,t\leq T}$ respectively. In addition, for technical simplicity, we assume that there is no fixed effects so that we do not take the de-meaning procedure. Extending to the more complete estimators with de-meaning is straightforward, but should require more technical arguments to show that the effect from added dependences due to the de-meaning is negligible.

A.1 Proofs of Theorem A.1-2.1.

Lemma A.1. *Under the Assumptions 2.1-2.2, for* $\gamma_T = \sqrt{\frac{\log(LN)}{T}}$, *for* $h \geq 0$,

$$\max_{h\leq L}\max_{i,j\leq N}\left\|\frac{1}{T}\sum_{t=h+1}^{T} x_{it}u_{j,t-h}\right\| = O_P(\gamma_T).$$

Proof. Let $\gamma_{ij,t,h} = x_{it}u_{j,t-h}I_{\{h+1\leq t\leq T\}}$, where I_A is the indicator function of the set A. To simplify notation, we assume $d = \dim(x_{it}) = 1$. By Lemma A.2 of Fan, Liao, and Mincheva (2011) and Assumption 2.1 (iii), $\gamma_{ij,t,h}$ satisfies the exponential tail condition. We set $\alpha_T = \sqrt{\frac{\log(LN)}{T}}$ and $c_1^2 = 3c_2$ for $c_1, c_2 > 0$. Using Bernstein inequality for weakly dependent data in Merlevède, Peligrad, and Rio (2011) and the Bonferroni method, we have

$$P\left(\max_h\max_{ij}\left|\frac{1}{T}\sum_{t=1}^{T}\gamma_{ij,t,h}\right| > c_1\alpha_T\right) \leq LN^2\max_{h\leq L}\max_{ij}P\left(\left|\frac{1}{T}\sum_{t=1}^{T}\gamma_{ij,t,h}\right| > c_1\alpha_T\right) \to 0.$$

Then $\max_h\max_{ij}\|\frac{1}{T}\sum_{t=h+1}^{T}x_{it}u_{j,t-h}\| = O_P(\sqrt{\frac{\log(LN^2)\max_{ij,h}\frac{1}{T}\sum_{t=1}^{T}\mathrm{var}(\gamma_{ij,t,h})}{T}}) = O_P(\sqrt{\frac{\log(LN)}{T}})$. □

Lemma A.2. *Under the Assumptions 2.1-2.2, for* $\gamma_T = \sqrt{\frac{\log(LN)}{T}}$,
(i) $\max_{h\leq L}\max_{i,j\leq N}|\widetilde{R}_{h,ij} - \Omega_{h,ij}| = O_P(\gamma_T)$, *where* $\widetilde{R}_{h,ij} = \frac{1}{T}\sum_{t=h+1}^{T}\widehat{u}_{it}\widehat{u}_{j,t-h}$ *for* $h \geq 0$.
(ii) $\max_{|h|\leq L}\|\widetilde{\Omega}_h - \Omega_h\|_1 = O_P(m_N\gamma_T^{1-q})$.

Proof. (i) First, we write

$$\begin{aligned}\max_{h\leq L}\max_{i,j\leq N}\left|\frac{1}{T}\sum_{t=h+1}^{T}\widehat{u}_{it}\widehat{u}_{j,t-h} - Eu_{it}u_{j,t-h}\right| \leq{}& \max_h\max_{ij}\left|\frac{1}{T}\sum_{t=h+1}^{T}(\widehat{u}_{it}\widehat{u}_{j,t-h} - u_{it}u_{j,t-h})\right| \\ &+ \max_h\max_{ij}\left|\frac{1}{T}\sum_{t=h+1}^{T}u_{it}u_{j,t-h} - Eu_{it}u_{j,t-h}\right| \\ &+ \max_h\max_{ij}\frac{L}{T}|Eu_{it}u_{j,t-h}|\end{aligned}$$

$$\equiv a_1 + a_2 + a_3.$$

Then, a_1 is bounded by $a_{11} + a_{12} + a_{13}$, where

$$a_{11} \equiv \max_h \max_{ij} \left| \frac{1}{T} \sum_{t=h+1}^{T} (\widehat{u}_{it} - u_{it})(\widehat{u}_{j,t-h} - u_{j,t-h}) \right|$$
$$a_{12} \equiv \max_h \max_{ij} \left| \frac{1}{T} \sum_{t=h+1}^{T} (\widehat{u}_{it} - u_{it}) u_{j,t-h} \right|$$
$$a_{13} \equiv \max_h \max_{ij} \left| \frac{1}{T} \sum_{t=h+1}^{T} u_{it}(\widehat{u}_{j,t-h} - u_{j,t-h}) \right|.$$

First, by the Cauchy-Schwarz inequality,

$$\begin{aligned} a_{11} &= \max_h \max_{ij} \left| \frac{1}{T} \sum_{t=h+1}^{T} x'_{it}(\widehat{\beta} - \beta) x'_{j,t-h}(\widehat{\beta} - \beta) \right| \\ &\leq \|\widehat{\beta} - \beta\|^2 \max_h \max_{ij} \frac{1}{T} \sum_{t=h+1}^{T} \|x_{it}\| \|x_{j,t-h}\| \\ &\leq O_P\left(\frac{1}{NT}\right) \max_i \frac{1}{T} \sum_{t=1}^{T} \|x_{it}\|^2 = O_P\left(\frac{1}{NT}\right). \end{aligned}$$

Note that $\max_i \frac{1}{T} \sum_{t=1}^{T} \|x_{it}\|^2$ is bounded by the exponential tail condition and Bernstein's inequality using the same argument of Lemma 3.1 of Fan, Liao, and Mincheva (2011). Next, by Lemma A.1,

$$\begin{aligned} a_{12} &\leq \|\beta - \widehat{\beta}\| \max_h \max_{ij} \left\| \frac{1}{T} \sum_{t=h+1}^{T} x_{it} u_{j,t-h} \right\| \\ &\leq O_P(\frac{1}{\sqrt{NT}}) \max_h \max_{ij} \left\| \frac{1}{T} \sum_{t=h+1}^{T} x_{it} u_{j,t-h} \right\| = O_P\left(\frac{1}{T}\sqrt{\frac{\log(LN)}{N}}\right). \end{aligned}$$

Similarly, a_{13} is bounded using the same argument. Then, we have $a_1 = O_P(\frac{1}{T}\sqrt{\frac{\log(LN)}{N}})$.

Next, we let $Z_{h,ij,t} = u_{it}u_{j,t-h} - Eu_{it}u_{j,t-h}$, which satisfies the exponential tail condition by Assumption 2.1 and Lemma A.2 of Fan, Liao, and Mincheva (2011). Then a_2 can be written as $\max_h \max_{ij} |\frac{1}{T} \sum_t Z_{h,ij,t}|$. Set $\alpha_T = \sqrt{\frac{\log(LN)}{T}}$ and $c_1^2 = 3c_2$ for $c_1, c_2 > 0$. Then, using Bernstein's inequality in Merlevède, Peligrad, and Rio (2011) and the same argument

as in the proof of Lemma A.1,

$$P\left(\max_{h\leq L}\max_{ij}\left|\frac{1}{T}\sum_{t=1}^{T}Z_{h,ij,t}\right|>c_1\alpha_T\right)\leq LN^2\max_{h\leq L}\max_{ij}P\left(\left|\frac{1}{T}\sum_{t=1}^{T}Z_{h,ij,t}\right|>c_1\alpha_T\right)\to 0.$$

Hence, we have $a_2=O_P(\sqrt{\frac{\log(LN)}{T}})$. In addition, $a_3=O_P(\frac{L}{T})$, which can be proved easily. Together,

$$\max_{h}\max_{ij}\left|\frac{1}{T}\sum_{t=h+1}^{T}\widehat{u}_{it}\widehat{u}_{j,t-h}-Eu_{it}u_{j,t-h}\right|=O_P\left(\sqrt{\frac{\log(LN)}{T}}\right).$$

(ii) Following Theorem 5 of Fan, Liao, and Mincheva (2013), we then have $\max_{|h|\leq L}\|\widetilde{\Omega}_h-\Omega_h\|_1=O_P(m_N\gamma_T^{1-q})$, where $\widetilde{\Omega}_h$ is defined in (2.4). □

Lemma A.3. *For $h\leq L$ and $v\leq L$, let $Q_{imp}^{hv}=\sum_{t=1}^{T}w_{it}\sum_{j=1}^{N}(\Omega^{-1})_{t+h,m+v,jp}$. Then, under the Assumption 2.1-2.2,*

$$\max_{h,v\leq L}\max_{i,p\leq N}\left\|\frac{1}{NT}\sum_{q=1}^{N}\sum_{m=1}^{T}\varepsilon_{qm}Q_{imp}^{hv}\right\|=O_P\left(\sqrt{\frac{\log(LN)}{NT}}\right).$$

Proof. First, we define $W=\Omega^{-1}X=(w_1',\cdots,w_T')'$ $(NT\times d)$, and $\varepsilon=\Omega^{-1}U=(\varepsilon_1',\cdots,\varepsilon_T')'$ $(NT\times 1)$. Let w_{it}' and ε_{it} denote the ith row of w_t and the ith element of ε_t, respectively. For simplicity, we assume $d=1$. Let $\zeta_{ipq,mhv}=\varepsilon_{qm}Q_{imp}^{hv}$. Note that due to $\|\Omega^{-1}\|_1<\infty$, we know $E(Q_{imp}^{hv})^2<\infty$. Then $\max_{hv}\max_{ip}\frac{1}{NT}\sum_{q=1}^{T}\sum_{m=1}^{T}\mathrm{var}(\zeta_{ipq,mhv})$ is bounded. Set $\alpha_{NT}=\sqrt{\frac{\log(LN)}{NT}}$ and $c_1^2=3c_2$ for $c_1,c_2>0$. Then, using Bernstein's inequality and the same argument as in the proof of Lemma A.1,

$$P\left(\max_{h,v\leq L}\max_{ip}\left|\frac{1}{NT}\sum_{q=1}^{N}\sum_{m=1}^{T}\zeta_{ipq,mhv}\right|>c_1\alpha_{NT}\right)\leq L^2N^2\max_{h,v\leq L}\max_{ij}P\left(\left|\frac{1}{NT}\sum_{q=1}^{N}\sum_{m=1}^{T}\zeta_{ipq,mhv}\right|>c_1\alpha_{NT}\right)\to 0.$$

Therefore, we have $\max_{h,v\leq L}\max_{i,p\leq N}\|\frac{1}{NT}\sum_{q=1}^{N}\sum_{m=1}^{T}\varepsilon_{qm}Q_{imp}^{hv}\|=O_P(\sqrt{\frac{\log(LN)}{NT}})$. □

Lemma A.4. *Consider a symmetric block matrix $A=(A_{ij})\in\mathbb{R}^{dn\times dn}$ where $A_{ij}\in\mathbb{R}^{d\times d}$. Then*

$$\|A\|\leq\max_{i}\sum_{j=1}^{n}\|A_{ij}\|.$$

Proof. Suppose $\sigma(\cdot)$ is the spectrum of a matrix, which is the set of its eigenvalues. By

Gershgorin's Theorem for block matrices (see Salas (1999)), if we define

$$G_i \equiv \sigma(A_{ii}) \cup T_i,$$

where $T_i = \{\lambda \notin \sigma(A_{ii}) : \|(A_{ii} - \lambda I_d)^{-1}\|^{-1} \leq \sum_{j=1, j\neq i}^{n} \|A_{ij}\|\}$, then

$$\sigma(A) \subset \bigcup_{i=1}^{n} G_i.$$

Note that this theorem means the eigenvalue of A either equals $\sigma(A_{ii})$ or in that specific region.

Let $\lambda \in \cup_{i=1}^{n} G_i$. If $\lambda \in \sigma(A_{ii})$ for some i, then $|\lambda| \leq \|A_{ii}\| \leq \max_i \sum_{j=1}^{n} \|A_{ij}\|$. If $\lambda \notin \sigma(A_{ii})$ for all i, then we know $\lambda \in T_i$ for some i. Now we consider two cases: (i) $\|A_{ii}\| < |\lambda|$, and (ii) $\|A_{ii}\| \geq |\lambda|$, where i such that $\lambda \in T_i$. For the case of (i), note that if a matrix M is such that $\|M\| < 1$, then

$$\frac{1}{1 + \|M\|} \leq \|(I - M)^{-1}\| \leq \frac{1}{1 - \|M\|}. \tag{A.1}$$

Then we have

$$\begin{aligned} |\lambda| - \|A_{ii}\| &\leq |\lambda| \left(1 - \frac{\|A_{ii}\|}{|\lambda|}\right) \\ &\leq |\lambda| \left\| \left(I_d - \frac{A_{ii}}{|\lambda|}\right)^{-1} \right\|^{-1} = \|(|\lambda| I_d - A_{ii})^{-1}\|^{-1} \\ &\leq \sum_{j=1, j\neq i}^{n} \|A_{ij}\|. \end{aligned}$$

Therefore, we have $|\lambda| \leq \sum_{j=1}^{n} \|A_{ij}\| \leq \max_i \sum_{j=1}^{n} \|A_{ij}\|$. Note that we have the second inequality since $\frac{\|A_{ii}\|}{|\lambda|} < 1$ with the inequality (A.1). For part (ii), if $\|A_{ii}\| \geq |\lambda|$, then $|\lambda| \leq \sum_{j=1}^{n} \|A_{ij}\| \leq \max_i \sum_{j=1}^{n} \|A_{ij}\|$. Therefore, $|\lambda| \leq \max_i \sum_{j=1}^{n} \|A_{ij}\|$ for all $\lambda \in \cup_{i=1}^{n} G_i$.

Finally, since $\sigma(A) \subset \bigcup_{i=1}^{n} G_i$, we know that for all $\lambda \in \sigma(A)$, $|\lambda| \leq \max_i \sum_{j=1}^{n} \|A_{ij}\|$. Therefore, we have $\|A\| \leq \max_i \sum_{j=1}^{n} \|A_{ij}\|$. □

Theorem A.1. *Under the Assumptions 2.1-2.2, when* $\|\Omega^{-1}\|_1 = O(1)$*, for* $q \in [0, 1)$ *such that* $Lm_N \gamma_T^{1-q} = o(1)$*,*

$$\|\widehat{\Omega} - \Omega\| = O_P(Lm_N \gamma_T^{1-q} + \xi_T(L) + f_T(L)) = \|\widehat{\Omega}^{-1} - \Omega^{-1}\|.$$

Proof of Theorem A.1. For any $NT \times NT$ blocked matrix $M = (m_{t,s})$ where the $m_{t,s}$ is the (t, s)th block $N \times N$ matrix. In addition, for any $0 \leq L < T$, we define $B_L(M) =$

$[(m_{t,s})1(|t-s| \leq L)]$, which is an $NT \times NT$ matrix. Then we can write

$$\|\widehat{\Omega} - \Omega\| \leq \|B_L(\Omega) - \Omega\| + \|\widehat{\Omega} - B_L(\Omega)\|.$$

First, we assume that $\xi_T(L) = \max_t \sum_{|h|>L} \|Eu_t u'_{t-h}\| = o(1)$ in Assumption 2.2(ii). This implies that off-diagonal $N \times N$ blocks that are far from the diagonal block are negligible due to weak dependences. As for the first part, by Lemma A.4,

$$\|B_L(\Omega) - \Omega\| \leq \max_t \sum_{s:|s-t|>L} \|Eu_t u'_s\| = \xi_T(L) \to 0.$$

Next, note that $f_T(L) = \max_t \sum_{|h|\leq L} \|Eu_t u'_{t-h}(1 - \omega(|h|, L))\| = o(1)$ (see Assumption 2.2(iii)). Then by Lemmas A.2 and A.4, for $C < \infty$,

$$\begin{aligned}
\|\widehat{\Omega} - B_L(\Omega)\| &\leq \max_t \sum_{s:|t-s|\leq L} \|\widehat{\Omega}_{t,s} - Eu_t u'_s\| \\
&\leq L \max_{|h|\leq L} \|(\widetilde{\Omega}_h - \Omega_h)\omega(|h|, L)\| + \max_t \sum_{|h|\leq L} \|Eu_t u'_{t-h}(1 - \omega(|h|, L))\| \\
&\leq CL \max_{|h|\leq L} \|\widetilde{\Omega}_h - \Omega_h\| + f_T(L) \\
&= O_P(Lm_N \gamma_T^{1-q}) + f_T(L),
\end{aligned}$$

where $\widetilde{\Omega}_h$ is defined in (2.4). Therefore,

$$\|\widehat{\Omega} - \Omega\| = O_P(Lm_N \gamma_T^{1-q} + \xi_T(L) + f_T(L)).$$

We now show the second statement of Theorem A.1. By the triangular inequality, we have

$$\begin{aligned}
\|\widehat{\Omega}^{-1} - \Omega^{-1}\| &\leq \|(\widehat{\Omega}^{-1} - \Omega^{-1})(\widehat{\Omega} - \Omega)\Omega^{-1}\| + \|\Omega^{-1}(\widehat{\Omega} - \Omega)\Omega^{-1}\| \\
&\leq \|\widehat{\Omega}^{-1} - \Omega^{-1}\|\|\Omega^{-1}\|\|\widehat{\Omega} - \Omega\| + \|\Omega^{-1}\|^2\|\widehat{\Omega} - \Omega\| \\
&= O_P(Lm_N \gamma_T^{1-q} + \xi_T(L) + f_T(L))\|\widehat{\Omega}^{-1} - \Omega^{-1}\| + O_P(Lm_N \gamma_T^{1-q} + \xi_T(L) + f_T(L)).
\end{aligned}$$

Hence we have $(1 + o_P(1))\|\widehat{\Omega}^{-1} - \Omega^{-1}\| = O_P(Lm_N \gamma_T^{1-q} + \xi_T(L) + f_T(L))$, that implies the result. □

Proof of Proposition 2.1. First the left hand side of equation (2.5) can be extended as

$$\frac{1}{\sqrt{NT}} X'(\widehat{\Omega}^{-1} - \Omega^{-1})U = \frac{1}{\sqrt{NT}} X'\Omega^{-1}(\widehat{\Omega} - \Omega)\Omega^{-1}U$$

$$
\begin{aligned}
&+ \frac{1}{\sqrt{NT}} X'\Omega^{-1}(\widehat{\Omega}-\Omega)\Omega^{-1}(\widehat{\Omega}-\Omega)\Omega^{-1}U \\
&+ \frac{1}{\sqrt{NT}} X'(\widehat{\Omega}^{-1}-\Omega^{-1})(\widehat{\Omega}-\Omega)\Omega^{-1}(\widehat{\Omega}-\Omega)\Omega^{-1}U \\
&\equiv a+b+c.
\end{aligned}
$$

Now we shall show that $\frac{1}{\sqrt{NT}} X'(\widehat{\Omega}^{-1}-\Omega^{-1})U = \frac{1}{\sqrt{NT}} X'\Omega^{-1}(\widehat{\Omega}-\Omega)\Omega^{-1}U + o_P(1)$. First, we define $W = \Omega^{-1}X$ and $\varepsilon = \Omega^{-1}U$. Then, $W = (w_1', \cdots, w_T')'$ with w_t being an $N \times d$ matrix of w_{it}, and ε_{it} is defined similarly. For any $NT \times NT$ matrix M, we denote $(M)_{t,s}$ or $(M)_h$ as the (t,s)th block matrix for $h = t-s$. Moreover, we denote $(M)_{ts,ij}$ or $(M)_{h,ij}$ as the (i,j)th element of the (t,s)th block matrix. Under Assumption 2.3, we show that $b = o_P(1)$ as follows:

We write, for $h = t-s$ and $v = k-m$,

$$
\begin{aligned}
b &= \frac{1}{\sqrt{NT}} \sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{k=1}^{T}\sum_{m=1}^{T} w_t'(\widehat{\Omega}-\Omega)_{t,s}(\Omega^{-1})_{s,k}(\widehat{\Omega}-\Omega)_{k,m}\varepsilon_m \\
&= \frac{1}{\sqrt{NT}} \sum_{t=1}^{T}\sum_{|h|\le L}\sum_{k=1}^{T}\sum_{m=1}^{T} w_t'(\widehat{\Omega}-\Omega)_{h}(\Omega^{-1})_{t-h,k}(\widehat{\Omega}-\Omega)_{k,m}\varepsilon_m \\
&\quad - \frac{1}{\sqrt{NT}} \sum_{t=1}^{T}\sum_{|h|> L}\sum_{k=1}^{T}\sum_{m=1}^{T} w_t'\Omega_h(\Omega^{-1})_{t-h,k}(\widehat{\Omega}-\Omega)_{k,m}\varepsilon_m \\
&= \frac{1}{\sqrt{NT}} \sum_{t=1}^{T}\sum_{|h|\le L}\sum_{|v|\le L}\sum_{m=1}^{T} w_t'(\widehat{\Omega}-\Omega)_{h}(\Omega^{-1})_{t-h,m+v}(\widehat{\Omega}-\Omega)_{v}\varepsilon_m \\
&\quad - \frac{1}{\sqrt{NT}} \sum_{t=1}^{T}\sum_{|h|\le L}\sum_{|v|> L}\sum_{m=1}^{T} w_t'(\widehat{\Omega}-\Omega)_{h}(\Omega^{-1})_{t-h,m+v}\Omega_v\varepsilon_m \\
&\quad - \frac{1}{\sqrt{NT}} \sum_{t=1}^{T}\sum_{|h|> L}\sum_{k=1}^{T}\sum_{m=1}^{T} w_t'\Omega_h(\Omega^{-1})_{t-h,k}(\widehat{\Omega}-\Omega)_{k,m}\varepsilon_m \\
&\equiv b_1 + b_2 + b_3.
\end{aligned}
$$

First, define $Q_{imp}^{hv} = \sum_{t=1}^{T} w_{it} \sum_{j=1}^{N} (\Omega^{-1})_{t-h,m+v,jp}$ as in Lemma A.3. We have, by Lemmas A.2-A.3 and Assumption 2.3,

$$
\|b_1\| = \left\| \frac{1}{\sqrt{NT}} \sum_{|h|\le L}\sum_{|v|\le L}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N} (\widehat{\Omega}-\Omega)_{h,ij}(\widehat{\Omega}-\Omega)_{v,pq} \sum_{t=1}^{T}\sum_{m=1}^{T} w_{it}\varepsilon_{qm}(\Omega^{-1})_{t-h,m+v,jp} \right\|
$$

$$
\begin{aligned}
&\leq \frac{1}{\sqrt{NT}} \left[\max_{|h|\leq L} \max_{j} \sum_{i=1}^{N} \left| (\widehat{\Omega} - \Omega)_{h,ij} \right| \right] \left[\max_{|v|\leq L} \max_{q} \sum_{p=1}^{N} \left| (\widehat{\Omega} - \Omega)_{v,pq} \right| \right] \\
&\quad \times \max_{i} \max_{p} \left\| \sum_{|h|\leq L} \sum_{|v|\leq L} \sum_{j=1}^{N} \sum_{q=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{T} w_{it} \varepsilon_{qm} (\Omega^{-1})_{t-h,m+v,jp} \right\| \\
&\leq \frac{1}{\sqrt{NT}} \max_{|h|\leq L} \|\widehat{\Omega}_h - \Omega_h\|_1^2 \max_{i,p} \left\| \sum_{|h|\leq L} \sum_{|v|\leq L} \sum_{q=1}^{N} \sum_{m=1}^{T} \varepsilon_{qm} Q_{imp}^{hv} \right\| \\
&\leq O(L^2\sqrt{NT}) \max_{|h|\leq L} \|\widehat{\Omega}_h - \Omega_h\|_1^2 \max_{i,p,h,v} \left\| \frac{1}{NT} \sum_{q=1}^{N} \sum_{m=1}^{T} \varepsilon_{qm} Q_{imp}^{hv} \right\| \\
&= O_P(\sqrt{T} L^2 m_N^2 \gamma_T^{3-2q}) = o_P(1).
\end{aligned}
$$

In addition, by Lemma A.2 and Bernstein inequality,

$$
\begin{aligned}
\|b_2\| &= \left\| \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{|h|\leq L} \sum_{|v|>L} \sum_{m=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{p=1}^{N} \sum_{q=1}^{N} w_{it} (\widehat{\Omega} - \Omega)_{h,ij} (\Omega^{-1})_{t-h,m+v,jp} \Omega_{v,pq} \varepsilon_{qm} \right\| \\
&\leq \frac{L}{\sqrt{NT}} \left[\max_{|h|\leq L} \max_{j} \sum_{i=1}^{N} \left| (\widehat{\Omega} - \Omega)_{h,ij} \right| \right] \left[\max_{q} \sum_{p=1}^{N} \sum_{|v|>L} |\Omega_{v,pq}| \right] \\
&\quad \times \max_{i} \sum_{t=1}^{T} \|w_{it}\| \max_{m} \sum_{q=1}^{N} |\varepsilon_{qm}| \max_{t,h,v,p} \sum_{m=1}^{T} \sum_{j=1}^{N} |(\Omega^{-1})_{t-h,m+v,jp}| \\
&\leq O(L\sqrt{NT}) \max_{|h|\leq L} \|\widehat{\Omega}_h - \Omega_h\|_1 \sum_{|v|>L} \|\Omega_v\|_1 \\
&= O_P(L^{1-\alpha}\sqrt{NT} m_N \gamma_T^{1-q}) = o_P(1)
\end{aligned}
$$

and

$$
\begin{aligned}
\|b_3\| &= \left\| \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{m=1}^{T} \sum_{|h|>L} \sum_{k=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{p=1}^{N} \sum_{q=1}^{N} w_{it} \Omega_{h,ij} (\Omega^{-1})_{t-h,k,jp} (\widehat{\Omega} - \Omega)_{km,pq} \varepsilon_{qm} \right\| \\
&\leq \frac{T}{\sqrt{NT}} \left[\max_{|v|\leq L} \max_{q} \sum_{p=1}^{N} \left| (\widehat{\Omega} - \Omega)_{v,pq} \right| \right] \left[\max_{j} \sum_{i=1}^{N} \sum_{|h|>L} |\Omega_{h,ij}| \right] \\
&\quad \times \max_{i} \sum_{t=1}^{T} \|w_{it}\| \max_{m} \sum_{q=1}^{N} |\varepsilon_{qm}| \max_{t,h,v,p} \sum_{m=1}^{T} \sum_{j=1}^{N} |(\Omega^{-1})_{t-h,m+v,jp}| \\
&\leq O(T\sqrt{NT}) \max_{v} \|\widehat{\Omega}_v - \Omega_v\|_1 \sum_{|h|>L} \|\Omega_h\|_1 \\
&= O_P(L^{-\alpha} T \sqrt{NT} m_N \gamma_T^{1-q}) = o_P(1).
\end{aligned}
$$

Therefore, we have $\|b\| = o_P(1)$. Next, we define $\gamma^* = Lm_N\gamma_T^{1-q} + \xi_T(L) + f_T(L)$. By Theorem A.1, $\|\widehat{\Omega} - \Omega\| = O_P(\gamma^*) = \|\widehat{\Omega}^{-1} - \Omega^{-1}\|$. Then, under the Assumption 2.3(v), when $\|\Omega^{-1}\|_1 = O(1)$, we have

$$\begin{aligned}\|c\| &= \left\| \frac{1}{\sqrt{NT}} X'(\widehat{\Omega}^{-1} - \Omega^{-1})(\widehat{\Omega} - \Omega)\Omega^{-1}(\widehat{\Omega} - \Omega)\Omega^{-1}U \right\| \\ &\leq \|\widehat{\Omega}^{-1} - \Omega^{-1}\|\|\widehat{\Omega} - \Omega\|^2\sqrt{NT} = O_P(\sqrt{NT}\gamma^{*3}) = o_P(1).\end{aligned}$$

Therefore, we have $\frac{1}{\sqrt{NT}} X'(\widehat{\Omega}^{-1} - \Omega^{-1})U = \frac{1}{\sqrt{NT}} X'\Omega^{-1}(\widehat{\Omega} - \Omega)\Omega^{-1}U + o_P(1)$.

From Theorem A.1, it is easy to that $\frac{1}{NT} X'\widehat{\Omega}^{-1}X = \frac{1}{NT} X'\Omega^{-1}X + o_P(1)$. Also, by the weak law of large numbers, $(\frac{1}{NT} X'\Omega^{-1}X)^{-1} = \Gamma^{-1} + o_P(1)$, where $\Gamma = E(\frac{1}{NT} X'\Omega^{-1}X)$. Then

$$\begin{aligned}\sqrt{NT}(\widehat{\beta}_{FGLS} - \beta) &= \left(\frac{1}{NT} X'\Omega^{-1}X\right)^{-1} \left(\frac{1}{\sqrt{NT}} X'\widehat{\Omega}^{-1}U\right) + o_P(1) \\ &= \left(\frac{1}{NT} X'\Omega^{-1}X\right)^{-1} \left(\frac{1}{\sqrt{NT}} X'\Omega^{-1}U + \frac{1}{\sqrt{NT}} X'(\widehat{\Omega}^{-1} - \Omega^{-1})U\right) + o_P(1) \\ &= \Gamma^{-1} \left(\frac{1}{\sqrt{NT}} X'\Omega^{-1}U + \frac{1}{\sqrt{NT}} X'(\widehat{\Omega}^{-1} - \Omega^{-1})U\right) + o_P(1) \\ &= \Gamma^{-1} \left(\frac{1}{\sqrt{NT}} X'\Omega^{-1}U\right) + \Gamma^{-1} \left(\frac{1}{\sqrt{NT}} X'\Omega^{-1}(\widehat{\Omega} - \Omega)\Omega^{-1}U\right) + o_P(1). \ \square\end{aligned}$$

Proof of Theorem 2.1. It suffices to prove $\left\| \frac{1}{\sqrt{NT}} X'\Omega^{-1}(\widehat{\Omega} - \Omega)\Omega^{-1}U \right\| = o_P(1)$.

Let $A_{b_h} = \{(i,j) : |Eu_{it}u_{j,t-h}| \neq 0\}$. Also, let $W = \Omega^{-1}X$, and $\varepsilon = \Omega^{-1}U$. In addition, w_{it} and ε_{it} are defined as in the proof of Proposition 2.1. We define $\mathbb{G}^1_{T,ij}(h) = \frac{1}{\sqrt{T}} \sum_{t=h+1}^{T} (u_{it}u_{j,t-h} - Eu_{it}u_{j,t-h})$ and $\mathbb{G}^2_{T,ij}(h) = \frac{1}{\sqrt{T}} \sum_{t=h+1}^{T} w_{it}\varepsilon_{j,t-h}$. Then under the Assumption 2.4, there is $C > 0$ so that

$$\begin{aligned}\left\| \frac{1}{\sqrt{NT}} X'\Omega^{-1}(\widehat{\Omega} - \Omega)\Omega^{-1}U \right\| &= \left\| \frac{1}{\sqrt{NT}} W'(\widehat{\Omega} - \Omega)\varepsilon \right\| \\ &\leq \left\| \frac{C}{\sqrt{NT}} \sum_{h=0}^{L} \sum_{i,j \in A_{b_h}} \sum_{t=h+1}^{T} w_{it}\varepsilon_{j,t-h} \frac{1}{T} \sum_{s=h+1}^{T} (u_{is}u_{j,s-h} - Eu_{it}u_{j,t-h}) \right\| + o_P(1) \\ &= \left\| \frac{C}{\sqrt{NT}} \sum_{h=0}^{L} \sum_{i,j \in A_{b_h}} \mathbb{G}^1_{T,ij}(h)\mathbb{G}^2_{T,ij}(h) \right\| + o_P(1) = o_P(1).\end{aligned}$$

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017): "When should you adjust standard errors for clustering?," *National Bureau of Economic Research Working Paper No. 24003*.

Allen, D. W. (1992): "Marriage and divorce: Comment," *The American Economic Review*, 82(3), 679–685.

Andrews, D. W. (1991): "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica: Journal of the Econometric Society*, 59(3), 817–858.

Arellano, M. (1987): "Computing Robust Standard Errors for Within-groups Estimators," *Oxford bulletin of Economics and Statistics*, 49(4), 431–434.

Bai, J., S. H. Choi, and Y. Liao (2019): "Standard Errors for Panel Data Models with Unknown Clusters," *arXiv preprint arXiv:1910.07406*.

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

Bai, J., and S. Ng (2017): "Principal components and regularized estimation of factor models," *arXiv preprint arXiv:1708.08137*.

Bickel, P. J., and E. Levina (2008a): "Covariance regularization by thresholding," *The Annals of Statistics*, 36(6), 2577–2604.

——— (2008b): "Regularized estimation of large covariance matrices," *The Annals of Statistics*, 36(1), 199–227.

Cameron, A. C., and D. L. Miller (2015): "A practitioner's guide to cluster-robust inference," *Journal of Human Resources*, 50(2), 317–372.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017): "Double/debiased/neyman machine learning of treatment effects," *American Economic Review*, 107(5), 261–65.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. K. Newey (2016): "Double machine learning for treatment and causal parameters," Discussion paper, cemmap working paper, Centre for Microdata Methods and Practice.

Driscoll, J. C., and A. C. Kraay (1998): "Consistent covariance matrix estimation with spatially dependent panel data," *Review of Economics and Statistics*, 80(4), 549–560.

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

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

Friedberg, L. (1998): "Did unilateral divorce raise divorce rates? Evidence from panel data," *American Economic Review*, 88(3), 608–627.

Hansen, C. B. (2007a): "Asymptotic properties of a robust variance matrix estimator for panel data when T is large," *Journal of Econometrics*, 141(2), 597–620.

——— (2007b): "Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects," *Journal of Econometrics*, 140(2), 670–694.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the Econometric Society*, pp. 1029–1054.

Liang, K.-Y., and S. L. Zeger (1986): "Longitudinal data analysis using generalized linear models," *Biometrika*, 73(1), 13–22.

Merlevède, F., M. Peligrad, and E. Rio (2011): "A Bernstein type inequality and moderate deviations for weakly dependent sequences," *Probability Theory and Related Fields*, 151(3-4), 435–474.

Miller, S., and R. Startz (2018): "Feasible Generalized Least Squares Using Machine Learning," *Available at SSRN 2966194*.

Newey, W. K. (1990): "Efficient instrumental variables estimation of nonlinear models," *Econometrica: Journal of the Econometric Society*, pp. 809–837.

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Newey, W. K., and K. D. West (1987): "A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix," *Econometrica: Journal of the Econometric Society*, 55, 703–708.

——— (1994): "Automatic lag selection in covariance matrix estimation," *The Review of Economic Studies*, 61(4), 631–653.

Peters, H. E. (1986): "Marriage and divorce: Informational constraints and private contracting," *The American Economic Review*, 76(3), 437–454.

Petersen, M. A. (2009): "Estimating standard errors in finance panel data sets: Comparing approaches," *The Review of Financial Studies*, 22(1), 435–480.

Romano, J. P., and M. Wolf (2017): "Resurrecting weighted least squares," *Journal of Econometrics*, 197(1), 1–19.

Salas, H. N. (1999): "Gershgorin's theorem for matrices of operators," *Linear algebra and its applications*, 291(1-3), 15–36.

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

Wager, S., and S. Athey (2018): "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 113(523), 1228–1242.

White, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica: Journal of the Econometric Society*, pp. 817–838.

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

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*. MIT press.