DSSR

Discussion Paper No. 131

GMM Estimation of Spatial Autoregressive Models with Cluster Dependent Errors

Takaki Sato

October, 2022

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

GMM Estimation of Spatial Autoregressive Models with Cluster Dependent Errors

Takaki Sato *

Abstract

This study considers the generalized method of moment (GMM) estimation of spatial autoregressive (SAR) models with unknown cluster correlations among error terms. In the presence of cluster correlations within errors, nonlinear moment conditions suitable for independent errors lose their validity and GMM estimators obtained from the moment condition are inconsistent. In this paper, we propose the GMM estimator obtained from another nonlinear moment condition suitable for cluster dependent error terms and show its asymptotic properties. Because the asymptotic variance of the GMM estimator depends on the choice of the weight matrix for GMM estimation, we also discuss the optimal weight which minimizes the asymptotic variance, and introduce the feasible optimal GMM estimator based on the consistent estimator of the weight. Monte Carlo experiments indicate that the proposed GMM estimator has a small bias and root mean squared errors when error terms in SAR models have cluster correlation compared to two stage least squares estimators and GMM estimators for independent errors.

Keywords: Cluster dependence, GMM estimation, Spatial autoregressive models

JEL Codes: C13, C15, C21

1 Introduction

Recently, spatial econometrics has been applied in many fields of economics such as urban, environmental, international, and others because it allows us to analyze the spatial spillover effects of policies implemented in one region on neighboring regions. The most widely used spatial econometrics models are spatial autoregressive (SAR) models

*Advanced Institute for Yotta Informatics, Tohoku University, Sendai 980-8577, Japan. takaki.sato.a6@tohoku.ac.jp

proposed in Cliff and Ord (1973, 1981), which are linear regression models with spatial lag terms representing the spatial correlation of cross-sectional units. For accurate statistical inferences about the effects of policies in an empirical analysis, the variance and covariance structure of the error terms in SAR models needs to be modeled appropriately according to the dataset. In order to make the SAR model applicable to a wider range of data, various assumptions about error terms have been considered.

The simplest assumption is that the error terms are independent and identically distributed (i.i.d.), i.e., variances are homoskedastic. Anselin (1988) and Lee (2004) propose quasi-maximum likelihood (QML) estimation method for the SAR model in the error case. The likelihood function for the SAR model contains Jacobian terms and the computational cost of the Jacobian terms increase when sample size is large. As a less computationally expensive method, Kelejian and Prucha (1998, 1999) propose two stage least squares (2SLS) estimation and Lee (2007) proposes generalized method of moment (GMM) estimation for the homoskedastic case. Lee (2007) shows that the GMM estimator obtained from nonlinear moment conditions for i.i.d. errors in addition to linear moment conditions based on the orthogonality of exogenous regressors is asymptotically more efficient than the 2SLS estimator.

Because homoskedastic errors may be restrictive in empirical application, extensions to the case where the error terms are independently and not identically distributed, i.e., variances are heteroskedastic, is considered. Since QML estimator becomes inconsistency in the presence of heteroskedastic variance, Lin and Lee (2010) propose the robust GMM estimator against an unknown heteroskedasticity obtained form nonlinear moment conditions suitable for heteroskedstity with linear moment conditions. The consistency and asymptotic normality of the GMM estimator and the optimal choice of the weight matrix are discussed in the paper.

Modeling of cross sectional dependence between error terms i.e. ,non-diagonal elements of variance-covariance matrix, are also considered in spatial econometrics literature. The first one is SAR errors modeling where the variance-covariance matrix is defined using a spatial weight matrix which is predetermined before analysis based on the geographical information of the data. Kelejian and Prucha (2010) propose gneralize moment estimator and Lee and Liu (2010) and Wang et al. (2018) proposes GMM estimators. The second one is nonparametric heteroskedasticity-autocorrelation consistent estimator introduced in Conley (1999) and Kelejian and Prucha (2007). The variance-covariance matrix is modeled using kernel functions based on distance measures. However, both approaches cannot remove the correlation in the error terms if the choices of spatial weight matrix or distance measures are wrong. Therefore, modeling with as few assumptions as possible regarding the covariance matrices of the errors is desirable for more precise empirical analysis.

The current paper aims to propose GMM estimators for SAR models with unknown cluster dependence. In this analysis, we assume n observations are grouped into G known clusters and the cluster size of each cluster may vary across clusters, but is finite. On the other hand, we do not impose any assumptions about correlations within clusters. Following White (1984), Hansen (2007), and Hansen and Lee (2019), our asymptotic framework is asymptotic as n and G simultaneously diverge to infinity. Moulton (1986, 1990) illustrate an example that when there is cluster correlation in the error terms in a linear regression model, the standard error robust to heteroskedasticity without considering cluster correlation can be smaller than the actual standard error. In spatial data, regions within the same state tend to have similar values due to common cultural backgrounds and other factors, so cluster correlations are likely to exist. Therefore, considering cluster correlation in addition to heteroskedasticity is also essential for precise statistical inference in application of spatial econometrics models. Furthermore, the SAR model with cluster dependent errors is a generalization of the SAR model for independent errors because it is the same as the SAR model with heterogeneous variance when the cluster size of each cluster is all 1. Therefore, the proposed model can be used for the same kind of analysis as the SAR model for independent errors, and it allow us to conduct more accurate statistical inferences by taking into account cluster correlations among errors. To the best of our knowledge, there are no other papers dealing with cluster correlations regarding SAR models.

In the GMM estimation of SAR models, nonlinear moment conditions are used together with linear moment conditions to improve the estimation accuracy of spatial parameters which measure the strength of spatial interaction effect between observations. However, the nonlinear moment condition, which is suitable for independent errors loses its validity when there is cluster correlation in the error term, and the GMM estimator based on the moment condition becomes inconsistent. In this study, we propose another nonlinear moment condition that takes into account the cluster correlation in the error term, and show that the GMM estimator based on the moment condition is consistent and asymptotically normal even in the presence of cluster correlations. Since the asymptotic variance of the proposed GMM estimator is affected by the weight matrix for GMM estimation, the optimal weight matrix that minimizes the asymptotic variance and feasible optimal GMM estimator with the weight matrix are also discussed.

We conduct some Monte Carlo studies to investigate the finite sample performances of the proposed GMM estimator. We find that when errors have cluster correlations the GMM estimator for independent errors have large biases and root mean squared errors, but the proposed GMM estimator has a small bias and root mean squared errors. Using the nonlinear moment condition together with the linear moment condition will greatly reduce the RMSE of the spatial parameters and improve the estimation accuracy when the value of the regression coefficient is small.

The rest of paper is organized as follows. Sections 2 introduces the SAR model which have cluster dependent errors. We propose the GMM estimator for the model and derive its asymptotic properties in Section 3. Moreover, we introduce feasible optimal GMM estimator which obtains minimum asymptotic variances. The results of some

Monte Carlo simulations are reported in Section 4. Section 5 concludes the paper. The proofs of all theorems are given in the Appendix.

2 The Model

We consider the following spatial autoregressive (SAR) model:

$$\boldsymbol{Y}_n = \lambda \boldsymbol{W}_n \boldsymbol{Y}_n + \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \tag{1}$$

where $\boldsymbol{Y}_n$ is an $n \times 1$ vector of observed dependent variables, $\boldsymbol{X}_n$ is an $n \times k$ matrix of nonstochastic exogenous variables, $\boldsymbol{W}_n$ is an $n \times n$ spatial weight matrix which have zero diagonal elements and are predetermined by the spatial information of observations, $\boldsymbol{\varepsilon}_n$ is an $n \times 1$ vector of error terms. The parameter λ is a spatial correlation parameter which measure the strength of spatial interaction between observations, and $\boldsymbol{\beta}$ is the vector of usual regression coefficients. The parameter space of spatial parameter λ is discussed in (Elhorst (2014)) and it is usually taken to be $(-1, 1)$.

In this analysis, we assume observations $\boldsymbol{Y}_n = (y_1, \ldots, y_n)'$ are grouped into G known clusters, indexed $g = 1, \ldots, G$. The cluster size for each cluster, $n_g, g = 1, \ldots, G$ may vary across clusters, but is finite. Total number of observations is $n = \sum_{g=1}^{G} n_g$. For convenience, we introduce a cluster level representation. We denote observations as $y_{g,j}$ for $g = 1, \ldots, G$ and $j = 1, \ldots, n_g$. Thus, observations are also written by $\boldsymbol{Y}_n = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_G')'$ and $\boldsymbol{Y}_g = (Y_{g,1}, \ldots, Y_{g,n_g})'$. The same cluster-level representation is used for other vectors and matrices, such as $\boldsymbol{X}_n$ and $\boldsymbol{\varepsilon}_n$.

The error terms are independent across clusters, while dependent within each cluster. Thus,

$$E(\boldsymbol{\varepsilon}_{g_1} \boldsymbol{\varepsilon}_{g_2}') = \begin{cases} \boldsymbol{\Sigma}_{g_1 g_1} & (g_1 = g_2) \\ \boldsymbol{0}_{g_1 g_2} & (g_1 \neq g_2) \end{cases},$$

where $\boldsymbol{\Sigma}_{g_1 g_1}$ is the $n_{g_1} \times n_{g_1}$ covariance matrix of the error $\boldsymbol{\varepsilon}_{g_1}$, and $\boldsymbol{0}_{g_1 g_2}$ is the $n_{g_1} \times n_{g_2}$ matrix whose elements are all zeros. The covariance matrix of $\boldsymbol{\varepsilon}_n$, $\boldsymbol{\Sigma}_n$, is a block diagonal matrix and there may be non-zero non-diagonal elements in $\boldsymbol{\Sigma}_n$. Because we admit that cluster sizes vary from cluster to cluster, the size of $\boldsymbol{\Sigma}_{g_1 g_1}$ and $\boldsymbol{\Sigma}_{g_2 g_2}$ are not necessarily the same when $g_1 \neq g_2$. Moreover, We do not impose any special constraints on the covariance structure of $\boldsymbol{\Sigma}_{g_1 g_1}$, and thus $\boldsymbol{\Sigma}_{g_1 g_1}$ is not necessary to be identical to $\boldsymbol{\Sigma}_{g_2 g_2}$ even if the sizes of both matrices are the same.

We note that the SAR model for cluster dependent errors includes SAR models for independent errors as the special case $n_g = 1, g = 1, \ldots, G$. When the cluster sizes in all cluster are 1, the number of observations n and clusters G is identical, and then the variance matrix $\boldsymbol{\Sigma}_{g_1 g_1}$ is a scalar variance parameter and $\boldsymbol{\Sigma}_{g_1 g_1}$ may be different from $\boldsymbol{\Sigma}_{g_2 g_2}$ as discussed above. This framework is the same as the SAR model which have independent heteroscedastic error terms. Thus, the SAR model for cluster dependent errors are an extension of SAR models for independent errors.

3 Estimation

In this section, we discuss suitable nonlinear and linear moment conditions for the SAR model with cluster dependent errors, and introduce the GMM estimator and feasible optimal GMM estimator obtained from the moment conditions. Their asymptotic properties are also considered. Nonlinear moment conditions for the SAR model with independent errors are ineligible to derive consistent GMM estimators when error terms have cluster correlations. Thus, we propose another nonlinear condition for cluster depended error cases, and show that the GMM estimator obtained by using the nonlinear moment conditions in addition to linear moment conditions based on exogenous regressors are consistent and asymptotically normal. Because asymptotic distribution of the GMM estimator are depend on the choice of a weight matrix for GMM estimation, we introduce an optimal weight matrix which minimizes asymptotic variances of the GMM estimator and propose the feasible optimal GMM estimator with the weight matrix.

3.1 Moment Conditions for GMM estimation

For GMM estimation, let us consider nonlinear moment conditions suitable for cluster dependent errors because using nonlinear moment conditions improves finite sample performance of GMM estimators when the variation from the exogenous regressors is small compared to error terms (see Lee and Liu (2010)). We define two classes of constant $n \times n$ matrices. The class $\mathcal{P}_1$ is a class of matrices whose diagonal elements are zeros, and $\mathcal{P}_2$ is a class of matrices whose block diagonal matrices are $\mathbf{0}_{g_1 g_1}, g_1 = 1, \ldots, G$, from the top left. When $n_g = 1, g = 1, \ldots, G$, two classes $\mathcal{P}_1$ and $\mathcal{P}_2$ are identical.

As mentioned in Lin and Lee (2010), the nonlinear moment condition, $E(\boldsymbol{\varepsilon}_n' \boldsymbol{P}_n \boldsymbol{\varepsilon}_n) = 0$, holds when $\boldsymbol{P}_n \in \mathcal{P}_1$ and ε_is are independent because all of non-diagonal element of the variance matrix $\boldsymbol{\Sigma}_n = E(\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n')$ are zeros. However, there are non-zero non-diagonal elements in $\boldsymbol{\Sigma}_n$ when error terms have cluster correlations. Then, the moment condition with $\boldsymbol{P}_n \in \mathcal{P}_1$ lose its validity for the SAR model with cluster dependent errors, because $E(\boldsymbol{\varepsilon}_n' \boldsymbol{P}_n \boldsymbol{\varepsilon}_n) =$

$tr(\boldsymbol{P}_n \boldsymbol{\Sigma}_n) \neq 0$. On the other hands, the moment condition obtained from $\boldsymbol{P}_n \in \mathcal{P}_2$ retains its validity because the appropriate diagonal block in $\boldsymbol{P}_n$ are all zeros and thus $E(\boldsymbol{\varepsilon}_n' \boldsymbol{P}_n \boldsymbol{\varepsilon}_n) = 0$. Thus, we adopt $\mathcal{P}_2$ as the class of matrices for nonlinear moment conditions this paper.

linear moment conditions are obtained from the orthogonality of the exogenous regressors. Let $\boldsymbol{Q}_n$ be an $n \times k^*$ matrix of instruments variables where $k^* \geq k+1$. In this paper we use the linear independent columns in $\boldsymbol{X}_n, \boldsymbol{W}_n \boldsymbol{X}_n, \boldsymbol{W}_n^2 \boldsymbol{X}_n$, etc. We set $E(\boldsymbol{Q}_n' \boldsymbol{\varepsilon}_n)$ as the linear moment condition for GMM estimation.

By using $\boldsymbol{P}_{n,j} \in \mathcal{P}_2, j = 1, \ldots, m$, and $\boldsymbol{Q}_n$, the set of moment functions for the GMM estimation is given by

$$\boldsymbol{g}_n(\theta) = (\boldsymbol{\varepsilon}_n(\boldsymbol{\theta})' \boldsymbol{P}_{n,1}' \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}), \ldots, \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})' \boldsymbol{P}_{n,m}' \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}), \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})' \boldsymbol{Q}_n)', \tag{2}$$

where $\boldsymbol{\theta} = (\lambda, \boldsymbol{\beta}')'$ and $\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) = \boldsymbol{Y}_n - \lambda \boldsymbol{W}_n \boldsymbol{Y}_n - \boldsymbol{X}_n \boldsymbol{\beta}$ from the model (1). As mentioned above, two classes $\mathcal{P}_1$ and $\mathcal{P}_2$ are identical when there is no cluster correlation within error terms. Thus, the GMM estimator proposed below is the same as the GMM estimator in Lin and Lee (2010) when error terms are independent.

We discuss choices of $\boldsymbol{P}_n$ and $\boldsymbol{Q}_n$ for empirical analysis. We denote true parameters as $\boldsymbol{\theta}_0 = (\lambda_0, \boldsymbol{\beta}_0')'$, and define the $n \times n$ identity matrix as $\boldsymbol{I}_n$, $\boldsymbol{S}_n = (\boldsymbol{I}_n - \lambda_0 \boldsymbol{W}_n)$ and $\boldsymbol{G}_n = \boldsymbol{W}_n \boldsymbol{S}_n^{-1}$. For independent and homoskedastic variances cases, the best choice of $\boldsymbol{P}_n \in \mathcal{P}_1$ is $\boldsymbol{G}_n - Diag(\boldsymbol{G}_n)$, and $\boldsymbol{Q}_n$ is $(\boldsymbol{G}_n \boldsymbol{X}_n \boldsymbol{\beta}_0, \boldsymbol{X}_n)$, respectively. However, as Lin and Lee (2010) indicated, the best selection of $\boldsymbol{P}_n$ and $\boldsymbol{Q}_n$ are not available when error terms have unknown heteroskedastic variances and the same is true for unknown cluster correlations because the covariance matrix and the first order derivative of (2) defined as (4) and (5) below involve the unknown matrix $\boldsymbol{\Sigma}_n$. They suggest making the same choices in such cases as in the case of independent homoskedastic variance. Following the manner, it might be desirable strategy to apply $\boldsymbol{G}_n - \boldsymbol{G}_n^*$ where $\boldsymbol{G}_n^*$ is the block diagonal matrix whose diagonal blocks are the $n_g \times n_g$ matrix whose elements are correspond to the appropriate elements of $\boldsymbol{G}_n$ as $\boldsymbol{P}_n$, and $(\boldsymbol{G}_n \boldsymbol{X}_n \boldsymbol{\beta}_0, \boldsymbol{X}_n)$ as $\boldsymbol{Q}_n$.

3.2 GMM estimation

Following Lee (2007), Lin and Lee (2010), we adopt the following regularity assumptions for GMM estimation to discuss asymptotic properties of the GMM estimators proposed below. Some assumptions are modified to fit into cluster dependent error case. We define $\boldsymbol{P}_{n,j}^s = \boldsymbol{P}_{n,j} + \boldsymbol{P}_{n,j}', j = 1, \ldots, m$.

Assumption 1. The error terms $\boldsymbol{\varepsilon}_n = \boldsymbol{\Sigma}_n^{-\frac{1}{2}} \boldsymbol{v}_n$, where $\boldsymbol{v}_n = (v_1, \ldots, v_n)'$ is independent random variables with mean 0 and variance 1. Furthermore, $\sup_i E|v_i|^{4+\delta} < \infty$ for some $\delta > 0$.

Assumption 2. The elements of the $n \times K$ regressor matrix $\boldsymbol{X}_n$ are uniformly bounded constants, $\boldsymbol{X}_n$ has the full rank k, and $\lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}_n'\boldsymbol{X}_n$ exists and is nonsingular.

Assumption 3. Both row and column sums of the spatial weight matrices $\boldsymbol{W}_n$ and the matrix $\boldsymbol{S}_n^{-1}$ are uniformly bounded.

Assumption 4. The matrices $\boldsymbol{P}_{n,j}, j = 1, \ldots, m$ belongs to the class $\mathcal{P}_2$ and uniformly bounded in both row and column sums. All elements in $\boldsymbol{Q}_n$ are uniformly bounded.

Assumption 5. Either (a) $\lim_{n\to\infty} \frac{1}{n}\boldsymbol{Q}_n'(\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0, \boldsymbol{X}_n)$ has the full rank $(k+1)$ or (b) $\lim_{n\to\infty} \frac{1}{n}\boldsymbol{Q}_n'\boldsymbol{X}_n$ has the full rank k, $\lim_{n\to\infty} \frac{1}{n}tr(\boldsymbol{\Sigma}_n\boldsymbol{P}_{nj}^s\boldsymbol{G}_n) \neq 0$, for some $j = 1, \ldots m$, and $\lim_{n\to\infty} \frac{1}{n}(tr(\boldsymbol{\Sigma}_n\boldsymbol{P}_{n,1}^s\boldsymbol{G}_n), \ldots, tr(\boldsymbol{\Sigma}_n\boldsymbol{P}_{nm}^s\boldsymbol{G}_n))$ and $\lim_{n\to\infty} \frac{1}{n}(tr(\boldsymbol{\Sigma}_n\boldsymbol{G}_n'\boldsymbol{P}_{n,1}^s\boldsymbol{G}_n), \ldots, tr(\boldsymbol{\Sigma}_n\boldsymbol{G}_n'\boldsymbol{P}_{nm}^s\boldsymbol{G}_n))$are linearly independent.

Assumption 6. The set of true parameters $\boldsymbol{\theta}_0$ is in the interior of the parameter space $\boldsymbol{\Theta}$, which is a compact set in $\mathbb{R}^{k+1}$.

Assumption 7. The numbers of both observations and clusters go to infinity. Thus, $n \to \infty$ and $G \to \infty$. Moreover, the number of each cluster size are fixed.

In Assumptions 1, we assume error terms have cluster correlations which is represented by the block diagonal element of $\boldsymbol{\Sigma}_n$. Higher order moment condition of v_i is necessary to apply the linear quadratic central limit theorem in Kelejian and Prucha (2001). Assumptions 2-6 are almost the same as the assumptions in Lee (2007), Lin and Lee (2010), and detailed discussion on the assumptions is given in the papers. Assumption 7 states that our asymptotic framework is asymptotic as n and G simultaneously diverge to infinity. Because the cluster sizes are bounded, G diverges at the same rate as n. Boundedness of cluster sizes allow us to obtain the GMM estimators whose convergence rate are the same as GMM estimators for independendt errors even if errors have strong correlations within clusters.

First, we introduce the GMM estimator obtained from the moment condition (2) and their consistency and asymptotic normality. Let $\boldsymbol{a}_n$ be a matrix which is a full row rank greater than or equal to the number of parameters in $\boldsymbol{\theta}$. Then, the GMM criterion function $J(\boldsymbol{\theta})$ obtained from the weight matrix $\boldsymbol{a}_n'\boldsymbol{a}_n$, is given by

$$J(\boldsymbol{\theta}) = \boldsymbol{g}_n'(\boldsymbol{\theta})\boldsymbol{a}_n'\boldsymbol{a}_n\boldsymbol{g}_n(\boldsymbol{\theta}), \tag{3}$$

and the GMM estimator $\hat{\boldsymbol{\theta}}_n$ minimizes (3).

We denote the (g_1, g_2) block of $\boldsymbol{P}_{n,j}$ as $\boldsymbol{P}_{j,g_1g_2}$ and define $\boldsymbol{P}_{j,g_2g_1}^s = \boldsymbol{P}_{j,g_2g_1} + \boldsymbol{P}_{g_1g_2}'$. For the GMM estimator, we have the following theorem which is a generalization of Proposition 1 in Lin and Lee (2010) to the cluster dependent error case.

Theorem 1. Suppose that assumptions 1-7 hold, $\lim_{n\to\infty}\frac{1}{n}\boldsymbol{a}_n\boldsymbol{D}_n$ exists and has the full rank (k+1), and $\lim_{n\to\infty}\boldsymbol{a}_nE(\boldsymbol{g}_n(\boldsymbol{\theta}))=\boldsymbol{0}$ has a unique root at $\boldsymbol{\theta}_0$. Then, $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator and $\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\xrightarrow{d}N(0,\boldsymbol{\Gamma})$, where

$$\boldsymbol{\Gamma}=\lim_{n\to\infty}\frac{1}{n}(\boldsymbol{D}_n'\boldsymbol{a}_n'\boldsymbol{a}_n\boldsymbol{D}_n)^{-1}\boldsymbol{D}_n'\boldsymbol{a}_n'\boldsymbol{a}_n\boldsymbol{\Omega}_n\boldsymbol{a}_n'\boldsymbol{a}_n\boldsymbol{D}_n(\boldsymbol{D}_n'\boldsymbol{a}_n'\boldsymbol{a}_n\boldsymbol{D}_n)^{-1},$$

$$\begin{aligned}
\boldsymbol{\Omega}_n &= Var(\boldsymbol{g}_n(\boldsymbol{\theta}_0)),\\
&= \begin{pmatrix} E(\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,1}\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,1}\boldsymbol{\varepsilon}_n) & E(\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,1}\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,2}\boldsymbol{\varepsilon}_n) & \cdots & 0\\ E(\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,2}\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,1}\boldsymbol{\varepsilon}_n) & E(\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,2}\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,2}\boldsymbol{\varepsilon}_n) & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \boldsymbol{Q}_n'\boldsymbol{\Sigma}_n\boldsymbol{Q}_n \end{pmatrix},\\
&= \begin{pmatrix} \sum_{g_1=1}^{G}\sum_{g_2=1}^{G}tr(\boldsymbol{\Sigma}_{g_1g_1}\boldsymbol{P}_{1,g_1g_2}\boldsymbol{\Sigma}_{g_2g_2}\boldsymbol{P}_{1,g_2g_1}^s) & \cdots & 0\\ \sum_{g_1=1}^{G}\sum_{g_2=1}^{G}tr(\boldsymbol{\Sigma}_{g_1g_1}\boldsymbol{P}_{2,g_1g_2}\boldsymbol{\Sigma}_{g_2g_2}\boldsymbol{P}_{1,g_2g_1}^s) & \cdots & 0\\ \vdots & & \vdots\\ 0 & \cdots & \boldsymbol{Q}_n'\boldsymbol{\Sigma}_n\boldsymbol{Q}_n \end{pmatrix},
\end{aligned}\tag{4}$$

$$\boldsymbol{D}_n=-\frac{\partial E(\boldsymbol{g}_n(\boldsymbol{\theta}_0))}{\partial\boldsymbol{\theta}'}=\begin{pmatrix} tr(\boldsymbol{\Sigma}_n\boldsymbol{P}_{1n}^s\boldsymbol{G}_n) & 0\\ \vdots & \vdots\\ tr(\boldsymbol{\Sigma}_n\boldsymbol{P}_{mn}^s\boldsymbol{G}_n) & 0\\ \boldsymbol{Q}_n'\boldsymbol{G}_n\boldsymbol{X}_n\beta_0 & \boldsymbol{Q}_n'\boldsymbol{X}_n \end{pmatrix}.\tag{5}$$

The rate of convergence of the proposed GMM estimator $\hat{\boldsymbol{\theta}}_n$ is the same as when the error terms are independent. As discussed in Hansen (2007) and Hansen and Lee (2019), clustering can alter the rate of convergence of estimators between $G^{-1/2}$ and $n^{-1/2}$ depending on the size of clusters and the strength of correlation within the cluster. In this analysis, we make no assumptions about the strength of the cluster correlations within the errors, but we do assume a finite number of clusters. Because the effect of correlation within clusters is limited, it does not affect the rate of convergence of the GMM estimator and The rate of convergence of the proposed GMM estimator $\hat{\boldsymbol{\theta}}_n$ is $\sqrt{n}$.

Next, let us consider the feasible optimal GMM estimator which have asymptotically efficient variance matrix. Theorem 1 indicates that the asymptotic distribution of the GMM estimator $\hat{\boldsymbol{\theta}}_n$ depends on the choice of the weight matrix $\boldsymbol{a}_n'\boldsymbol{a}_n$. An optimal weight matrix which minimizes the asymptotic variance of the GMM estimator is Ω_n^{-1}. Thus, if we obtain the consistent estimator of the optimal weight, then we can conduct feasible optimal GMM estimation by using the consistent estimator of the optimal weight as the weight matrix $\boldsymbol{a}_n'\boldsymbol{a}_n$.

Compared to independent error cases, The terms $E(\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,i}\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,j}\boldsymbol{\varepsilon}_n), i, j = 1, \dots, m$ which are the elements in $\boldsymbol{\Omega}_n$ related to the nonlinear moment condition have more sums of elements because $\boldsymbol{\Sigma}_n$ have non-zero non-diagonal elements. The number of sums in its expectation is $(\sum_{g=1}^{G} n_g^2)^2$ and grater than n^2 for independent cases. However, the convergence of these terms can be obtained from the uniform boundedness properties of $\boldsymbol{P}_{n,j}, j = 1, \dots, m$.

Residuals $\hat{\boldsymbol{\varepsilon}}_n = (\hat{\boldsymbol{\varepsilon}}_1', \dots, \hat{\boldsymbol{\varepsilon}}_G')'$ of the model (1) with $\hat{\theta}_n$ is defined by $\hat{\boldsymbol{\varepsilon}}_n = \boldsymbol{Y}_n - \hat{\lambda}_n \boldsymbol{W}\boldsymbol{Y}_n - \boldsymbol{X}\hat{\boldsymbol{\beta}}_n$, and then the estimates of the j-th diagonal block of the estimates of covariance matrix of $\boldsymbol{\varepsilon}_n$, $\hat{\boldsymbol{\Sigma}}_n$, is $\hat{\boldsymbol{\Sigma}}_{jj} = \hat{\boldsymbol{\varepsilon}}_j\hat{\boldsymbol{\varepsilon}}_j'$. We define $\frac{1}{n}\hat{\boldsymbol{D}}_n$ and $\frac{1}{n}\hat{\boldsymbol{\Omega}}_n$ by replacing $\boldsymbol{\theta}_0$ and $\boldsymbol{\Sigma}_n$ in $\frac{1}{n}\boldsymbol{D}_n$ and $\frac{1}{n}\boldsymbol{\Omega}_n$ with $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\Sigma}}_n$, respectively. Then, $\frac{1}{n}\hat{\boldsymbol{D}}_n$ and $\frac{1}{n}\hat{\boldsymbol{\Omega}}_n$ are consistent estimators of $\frac{1}{n}\boldsymbol{D}_n$ and $\frac{1}{n}\boldsymbol{\Omega}_n$ from the following theorem.

Theorem 2. Under assumptions 1-7, $\frac{1}{n}\hat{\boldsymbol{\Omega}}_n - \frac{1}{n}\boldsymbol{\Omega}_n = o_p(1)$ and $\frac{1}{n}\hat{\boldsymbol{D}}_n - \frac{1}{n}\boldsymbol{D}_n = o_p(1)$.

By using the consistent estimator of the optimal weight $\frac{1}{n}\hat{\boldsymbol{\Omega}}_n$ as the weight matrix $\boldsymbol{a}_n'\boldsymbol{a}_n$, feasible optimal GMM estimator, $\hat{\boldsymbol{\theta}}_{o,n}$, is obtained from

$$\hat{\boldsymbol{\theta}}_{o,n} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\, \boldsymbol{g}_n'(\boldsymbol{\theta})\hat{\boldsymbol{\Omega}}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\theta}).$$

Asymptotic properties of feasible optimal GMM estimator $\hat{\boldsymbol{\theta}}_{o,n}$ are also given in the following theorem.

Theorem 3. Suppose that $\lim_{n\to\infty} \frac{1}{n}\boldsymbol{\Omega}_n$ exists and nonsingular, and $(\frac{1}{n}\hat{\boldsymbol{\Omega}}_n)^{-1} - (\frac{1}{n}\boldsymbol{\Omega}_n)^{-1} = o_p(1)$. Under assumptions 1-7,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{o,n} - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(0, \left(\lim_{n\to\infty} \frac{1}{n}\boldsymbol{D}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{D}_n\right)^{-1}\right).$$

Moreover, a consistent estimator for $\left(\lim_{n\to\infty} \frac{1}{n}\boldsymbol{D}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{D}_n\right)^{-1}$ is $\left(\lim_{n\to\infty} \frac{1}{n}\hat{\boldsymbol{D}}_n'\hat{\boldsymbol{\Omega}}_n^{-1}\hat{\boldsymbol{D}}_n\right)^{-1}$.

4 Monte Carlo Simulations

Monte Carlo simulations are carried out to investigate small sample properties of the proposed GMM estimator for the SAR model with cluster dependent errors. In this paper, we report only the case where error terms have cluster dependence because our proposed GMM estimator is the same as the GMM estimator for independent errors proposed in Lin and Lee (2010) when error terms are independent and small sample properties of the estimator in the case are already have reported in their paper.

Table 1: Biases and RMSEs under Designs V-D1 and V-D2. The set of number of clusters and cluster sizes are (i)$G = 200, n_g = 4$, or (ii)$G = 100, n_g = 8$. True parameters P-D1:$(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.6, 0.8, 0.2, 1.5)$.

				High Cluster Correlation		Low Cluster Correlation	
	G	ng		Bias	RMSE	Bias	RMSE
2SLS	200	4	λ	0.0078	0.1361	-0.0057	0.1108
			β_1	-0.0244	0.5110	0.0133	0.4145
			β_2	-0.0018	0.0509	0.0018	0.0507
			β_3	-0.0092	0.0834	-0.0016	0.0857
	100	8	λ	0.0027	0.2283	-0.0013	0.1162
			β_1	-0.0005	0.8148	0.0139	0.4362
			β_2	-0.0033	0.0468	-0.0037	0.0487
			β_3	-0.0149	0.0829	-0.0085	0.0863
GMM(hetero)	200	4	λ	0.1896	0.1914	0.0567	0.0653
			β_1	-0.6461	0.6702	-0.2015	0.2773
			β_2	-0.0066	0.0497	0.0009	0.0504
			β_3	-0.0444	0.0926	-0.0102	0.0858
	100	8	λ	0.2696	0.2705	0.0972	0.1019
			β_1	-0.9126	0.9257	-0.3251	0.3735
			β_2	-0.0106	0.0448	-0.0057	0.0482
			β_3	-0.0683	0.1005	-0.0238	0.0884
GMM(cluster)	200	4	λ	-0.0052	0.0474	-0.0052	0.0451
			β_1	0.0167	0.2434	0.0085	0.2260
			β_2	-0.0003	0.0517	0.0031	0.0510
			β_3	0.0014	0.0815	0.0048	0.0857
	100	8	λ	-0.0078	0.0674	-0.0064	0.0605
			β_1	0.0313	0.2973	0.0285	0.2678
			β_2	-0.0014	0.0484	-0.0028	0.0492
			β_3	-0.0001	0.0734	0.0000	0.0850

The data generating process is as follows. The number of observations is 800 for all experiments. The spatial weight matrix $\boldsymbol{W}_n$ made from a standard 8 connection settings, namely $w_{i,j} = 1/8, j = i-4, \ldots, i-1, i+1, \ldots, i+4$ and $w_{i,j} = 0$ otherwise. The set of number of clusters and cluster sizes can be (i)$G = 200, n_g = 4$, or (ii)$G = 100, n_g = 8$, where we assume that the cluster size is the same for all clusters, and one cluster is created for every n_g individuals in turn. That is, if $n_g = 4$, then the first cluster consists of first through fourth observations, and the second cluster consists of fifth through eighth observations. We consider three regressors $\boldsymbol{X}_{n,1}, \boldsymbol{X}_{n,2}$ and $\boldsymbol{X}_{n,3}$. The first regressor $\boldsymbol{X}_{n,1}$ is the constant term, and the i-the element of $\boldsymbol{X}_{n,2}$ and $\boldsymbol{X}_{n,3}$ are generated from $x_{2,i} \sim N(3,1)$ and $x_{3,i} \sim U(-1,2)$, respectively.

For the covariance matrix $\boldsymbol{\Sigma}_{gg}$ which is the g-th diagonal block in $\boldsymbol{\Sigma}_n$, we consider two designs V-D1 and V-D2. The diagonal elements of $\boldsymbol{\Sigma}_{gg}$ are generated from $U(1,3)$ in the both case. The all non-diagonal elements in the block diagonal elements are 0.9 in V-D1, and thus the design V-D1 emphasizes a strong cluster correlations. As a week cluster correlation case, non-diagonal elements in the block diagonal matrix are set to be 0.2 in V-D2.

Table 2: Biases and RMSEs under Designs V-D1 and V-D2. The set of number of clusters and cluster sizes are (i)$G = 200, n_g = 4$, or (ii)$G = 100, n_g = 8$. True parameters P-D2:$(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.6, 0.2, 0.2, 0.1)$.

				High Cluster Correlation		Low Cluster Correlation	
	G	ng		Bias	RMSE	Bias	RMSE
2SLS	200	4	λ	0.2121	1.2284	0.2273	1.1479
			β_1	-0.4099	2.5391	-0.4531	2.2613
			β_2	-0.0033	0.0514	-0.0020	0.0544
			β_3	0.0006	0.0810	-0.0054	0.0911
	100	8	λ	0.3643	0.6764	0.2360	0.6524
			β_1	-0.7135	1.3989	-0.4588	1.3514
			β_2	-0.0048	0.0460	-0.0046	0.0518
			β_3	0.0035	0.0758	-0.0055	0.0852
GMM(hetero)	200	4	λ	0.2247	0.2585	0.0979	0.1851
			β_1	-0.4368	0.5211	-0.1958	0.3895
			β_2	-0.0037	0.0469	-0.0005	0.0498
			β_3	0.0002	0.0792	-0.0026	0.0852
	100	8	λ	0.2888	0.3252	0.1539	0.2108
			β_1	-0.5483	0.6531	-0.2969	0.4320
			β_2	-0.0102	0.0451	-0.0032	0.0492
			β_3	0.0003	0.0738	-0.0050	0.0829
GMM(cluster)	200	4	λ	0.0751	0.2171	0.0528	0.2015
			β_1	-0.1479	0.4428	-0.1149	0.4191
			β_2	0.0006	0.0489	0.0027	0.0511
			β_3	0.0019	0.0830	-0.0013	0.0865
	100	8	λ	0.1381	0.2736	0.0914	0.2198
			β_1	-0.2671	0.5381	-0.1893	0.4347
			β_2	-0.0040	0.0462	0.0022	0.0518
			β_3	0.0031	0.0799	-0.0029	0.0860

For each of the variance designs, we consider the 4 parameter designs P-D1, P-D2, P-D3, and P-D4 following Lin and Lee (2010). The design P-D1 has $\boldsymbol{\theta}_0 = (\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.6, 0.8, 0.2, 1.5)$ and the design P-D2 has $\boldsymbol{\theta}_0 = (0.6, 0.2, 0.2, 0.1)$. Because the parameters for βs in P-D2 are smaller than in P-D1, the stochastic part of the model (1) in P-D2 becomes more dominant compared to the model for P-D1. As stated in Lee (2007) and Lee and Liu (2010), we expect that the GMM estimator based on both of nonlinear and linear moment condtions improve small sample performances compared to 2SLS which is based on only linear moment conditions. In addition for $\lambda_0 = 0.6$, we conducted simulations in the case $\lambda_0 = 0.2$ to check the finite sample performance of estimators when spatial interactions are weak.

Simulated data generating from above settings are estimated by the following three estimation methods. The first one is 2SLS estimation where we use $(\boldsymbol{W}_n\boldsymbol{X}_n, \boldsymbol{X}_n)$ as IV matrix. The second method is GMM estimation for independent errors (GMM(hetero)) proposed in Lin and Lee (2010). We set $\boldsymbol{G}_n - Diag(\boldsymbol{G}_n)$ as $\boldsymbol{P}_n$ and $(\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}, \boldsymbol{X}_n)$ as $\boldsymbol{Q}_n$ where the estimate obtained by the 2SLS are used to evaluate $\boldsymbol{G}_n$ and $\boldsymbol{\beta}_0$. By using the residual $\hat{\varepsilon}_i$s with

Table 3: Biases and RMSEs under Designs V-D1 and V-D2. The set of number of clusters and cluster sizes are (i)$G = 200, n_g = 4$, or (ii)$G = 100, n_g = 8$. True parameters P-D3:$(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.2, 0.8, 0.2, 1.5)$.

				High Cluster Correlation		Low Cluster Correlation	
	G	ng		Bias	RMSE	Bias	RMSE
2SLS	200	4	λ	-0.0120	0.2506	-0.0034	0.1673
			β_1	0.0301	0.4775	0.0070	0.3404
			β_2	-0.0025	0.0497	-0.0002	0.0537
			β_3	-0.0096	0.0882	-0.0059	0.0868
	100	8	λ	-0.0018	0.3439	0.0001	0.1786
			β_1	0.0063	0.5913	0.0065	0.3610
			β_2	-0.0020	0.0465	-0.0031	0.0493
			β_3	-0.0113	0.1010	-0.0088	0.0872
GMM(hetero)	200	4	λ	0.3223	0.3255	0.0969	0.1119
			β_1	-0.5519	0.5782	-0.1678	0.2602
			β_2	-0.0040	0.0484	-0.0006	0.0538
			β_3	-0.0210	0.0841	-0.0071	0.0863
	100	8	λ	0.4621	0.4667	0.1620	0.1703
			β_1	-0.7982	0.8164	-0.2744	0.3284
			β_2	-0.0040	0.0429	-0.0037	0.0494
			β_3	-0.0320	0.0857	-0.0128	0.0858
GMM(cluster)	200	4	λ	-0.0073	0.0772	-0.0029	0.0737
			β_1	0.0183	0.2201	0.0038	0.2181
			β_2	-0.0014	0.0507	0.0004	0.0544
			β_3	0.0012	0.0798	0.0002	0.0861
	100	8	λ	-0.0070	0.1102	-0.0077	0.0976
			β_1	0.0141	0.2657	0.0171	0.2350
			β_2	-0.0005	0.0480	-0.0022	0.0500
			β_3	0.0023	0.0793	-0.0011	0.0841

the estimates of 2SLS, the diagonal matrix $\hat{\boldsymbol{\Sigma}}_{n,hetero}$ where the diagonal element is $\hat{\sigma}_i^2 = \hat{\epsilon}_i^2$ are used as the weight matrix for GMM(hetero). The last one is the GMM estimator proposed in the current paper (GMM(cluster)). As discussed in section 3, we set $\boldsymbol{G}_n - \boldsymbol{G}_n^*$ as $\boldsymbol{P}_n$, and $(\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0, \boldsymbol{X}_n)$ as $\boldsymbol{Q}_n$ where the estimate obtained by the 2SLS are used to evaluate $\boldsymbol{G}_n$ and $\boldsymbol{\beta}_0$. The weight matrix for the GMM(cluster) is the block diagonal matrix whose diagonal blocks are $\hat{\boldsymbol{\Sigma}}_{jj} = \hat{\boldsymbol{\varepsilon}}_j\hat{\boldsymbol{\varepsilon}}_j'$ by using the residual $\hat{\boldsymbol{\varepsilon}}_g$ with the estimates of 2SLS.

For each case, we conducted 1000 times Monte Carlo replications. We report biases and root mean squared errors to investigate small sample properties of three estimators discussed above.

Table 1 summarizes the results under the designs V-D1 and V-D2 with P-D1. The case where regression coefficient are small and stochastic part in the model (1) are more dominant are reported in Table 2. Table 3 and 4 reports the results with p-D3 and P-D4 respectively, where $\lambda_0 = 0.2$ and spatial interactions are weak.

In terms of bias, GMM(hetero) has large biases when errors have cluster correlaiton. The bias is especially large when there are many individuals in a cluster or when the correlation is strong. This result is consistent with the

Table 4: Biases and RMSEs under Designs V-D1 and V-D2. The set of number of clusters and cluster sizes are (i)$G = 200, n_g = 4$, or (ii)$G = 100, n_g = 8$. True parameters P-D4:$(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.2, 0.2, 0.2, 0.1)$.

				High Cluster Correlation		Low Cluster Correlation	
	G	ng		Bias	RMSE	Bias	RMSE
2SLS	200	4	λ	0.3690	1.4287	0.2201	1.1481
			β_1	-0.3583	1.5427	-0.1957	1.2385
			β_2	-0.0024	0.0666	-0.0090	0.0622
			β_3	0.0013	0.0850	-0.0040	0.0907
	100	8	λ	0.5919	1.2203	0.2878	1.3058
			β_1	-0.5917	1.3222	-0.2692	1.4294
			β_2	0.0022	0.0530	-0.0048	0.0705
			β_3	0.0038	0.0804	-0.0009	0.0899
GMM(hetero)	200	4	λ	0.3824	0.4526	0.1467	0.2854
			β_1	-0.3759	0.4717	-0.1334	0.3077
			β_2	-0.0020	0.0483	-0.0050	0.0518
			β_3	0.0031	0.0817	-0.0024	0.0864
	100	8	λ	0.5389	0.5683	0.2194	0.3208
			β_1	-0.5285	0.5753	-0.2107	0.3446
			β_2	-0.0033	0.0426	-0.0023	0.0509
			β_3	0.0018	0.0757	-0.0001	0.0857
GMM(cluster)	200	4	λ	0.0958	0.3501	0.0586	0.2893
			β_1	-0.0955	0.3708	-0.0535	0.3063
			β_2	0.0003	0.0493	-0.0023	0.0524
			β_3	0.0037	0.0857	-0.0011	0.0885
	100	8	λ	0.2678	0.4999	0.0850	0.3460
			β_1	-0.2716	0.5137	-0.0914	0.3591
			β_2	0.0004	0.0448	0.0030	0.0540
			β_3	0.0056	0.0824	0.0020	0.0892

fact that when the error terms are correlated, the moment condition assuming independence is inappropriate, and the estimator is inconsistency. For 2SLS and GMM(cluster), the bias of 2SLS is not affected by the correlation within the clusters, and the results do not change when the correlation is strong or low. However, when the values of the regression coefficients are small in designs P-D2 and P-D4, the bias of the estimator of the spatial coefficients of 2SLS is much larger than that of GMM. In contrast to the 2SLS, which uses only linear moment conditions, the GMM(cluster) uses both linear and nonlinear moment conditions and it reduces the bias of the estiamtor.

In terms of RMSE, the RMSE of the spatial parameter λ and the intercept β_1 for all estimators are slightly larger when the correlation within the clusters is strong. Comparing GMM(hetero) and GMM(cluster), we can find that the RMSEs of GMM(cluster) are smaller in all cases. Because the GMM(cluster) uses the optimal weight matrix for cluster correlations among errors, this would improve efficiency in finite samples. When the values of the regression coefficients are small in designs P-D2 and P-D4, the RMSEs of the 2SLS estimator for the spatial parameter and the intercept is much larger than those of GMM(hetero) and GMM(cluster). As in the discussion of bias, the nonlinear moment conditions improve finite sample performances of the GMM estimators. For β_2 and β_3, RMSEs for them are almost the same in all three estimators, regardless of the strength of the cluster correlation or the size of the regression coefficient.

5 Conclusion

We introduce the SAR model which have cluster dependent error terms and propose the GMM estimator obtained from nonlinear moment conditions suitable for cluster dependent errors in addition to linear moment conditions. The proposed estimator has consistency and asymptotic normality under general assumptions. Since the asymptotic variance of the GMM estimator is affected by the weight matrix, we also introduced an asymptotically valid feasible optimal GMM estimator. Monte Carlo simulations show that the GMM estimator obtained from independence error assumptions has a bias when there is cluster correlation in the error term, but the proposed GMM estimator has a small bias and RMSE regardless of the strength of the correlation.

Future research includes the extension to spatial panel models and spatial dynamic panel models. In the panel model, the error terms may have serial correlation within the same observations in addition to the cluster correlations, which makes the correlation among error terms more complicated. However, by extending the model to panel models, it will be possible to analyze the spatio-temporal correlations of the variables and the regression coefficients considering the individual effects, which will enable more precise empirical analysis.

References

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Kluwer Academic Publishers, Dordrecht.

Cliff, A. and Ord, J. (1973). *Spatial Autocorrelation.* Pion, London.

Cliff, A. and Ord, J. (1981). *Spatial Process, Models and Applications.* Pion, London.

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

Elhorst, J. P. (2014). *Spatial econometrics from cross-sectional data to spatial panels.* Springer.

Hansen, B. E. and Lee, S. (2019). Asymptotic theory for clustered samples. *Journal of Econometrics*, 210(2):268–290.

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

Kelejian, H. H. and Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, 17(1):99–121.

Kelejian, H. H. and Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International economic review*, 40(2):509–533.

Kelejian, H. H. and Prucha, I. R. (2001). On the asymptotic distribution of the moran i test statistic with applications. *Journal of Econometrics*, 104(2):219–257.

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

Kelejian, H. H. and Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of econometrics*, 157(1):53–67.

Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72(6):1899–1925.

Lee, L.-f. (2007). Gmm and 2sls estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*, 137(2):489–514.

Lee, L.-f. and Liu, X. (2010). Efficient gmm estimation of high order spatial autoregressive models with autoregressive disturbances. *Econometric Theory*, 26(1):187–230.

Lin, X. and Lee, L.-f. (2010). Gmm estimation of spatial autoregressive models with unknown heteroskedasticity. *Journal of Econometrics*, 157(1):34–52.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of econometrics*, 32(3):385–397.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The review of Economics and Statistics*, pages 334–338.

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

Wang, W., Lee, L.-F., and Bao, Y. (2018). Gmm estimation of the spatial autoregressive model in a system of interrelated networks. *Regional Science and Urban Economics*, 69:167–198.

White, H. (1984). *Asymptotic Theory for Econometricians.* Academic Press.

Appendix

A Some useful lemmas

Lemma A.1. Suppose that $\boldsymbol{\varepsilon}_n$ satisfy assumption 1, and $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ are $n \times n$ matrices whose $n_{g_1} \times n_{g_1}$ diagonal blocks are $\mathbf{0}_{n_{g_1} n_{g_1}}$. Then, $E(\boldsymbol{\varepsilon}_n' A_n \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n' B_n \boldsymbol{\varepsilon}_n) = \sum_{g_1=1}^{G} \sum_{g_2=g_1}^{G} tr(\boldsymbol{\Sigma}_{g_1 g_1} \boldsymbol{A}_{g_1 g_2} \boldsymbol{\Sigma}_{g_2 g_2} \boldsymbol{B}^*_{g_2, g_1})$ where $\boldsymbol{B}^*_{g_2, g_1} = (\boldsymbol{B}_{g_2 g_1} + \boldsymbol{B}'_{g_1 g_2}))$.

Proof.

$$E(\boldsymbol{\varepsilon}_n' A_n \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n' B_n \boldsymbol{\varepsilon}_n) = \sum_{g_1=1}^{G} \sum_{g_2=1}^{G} \sum_{g_3=1}^{G} \sum_{g_4=1}^{G} E(\boldsymbol{\varepsilon}_{g_1}' \boldsymbol{A}_{g_1, g_2} \boldsymbol{\varepsilon}_{g_2} \boldsymbol{\varepsilon}_{g_3}' \boldsymbol{B}_{g_3, g_4} \boldsymbol{\varepsilon}_{g_4}).$$

Because $\boldsymbol{\varepsilon}_n = (\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_G')'$ are independent among clusters and the diagonal blocks of $\boldsymbol{A}$ and $\boldsymbol{B}$ are zeros, $E(\boldsymbol{\varepsilon}_{g_1}' \boldsymbol{A}_{g_1, g_2} \boldsymbol{\varepsilon}_{g_2} \boldsymbol{\varepsilon}_{g_3}' \boldsymbol{A}_{g_3, g_4} \boldsymbol{\varepsilon}_{g_4}) \neq 0$ only if $g_1 = g_3 \neq g_2 = g_4$ or $g_1 = g_4 \neq g_2 = g_3$. It follows that

$$\begin{aligned}
E(\boldsymbol{\varepsilon}_n' A_n \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n' B_n \boldsymbol{\varepsilon}_n) &= \sum_{g_1=1}^{G} \sum_{g_2 \neq g_1}^{G} \left\{ E(\boldsymbol{\varepsilon}_{g_1}' \boldsymbol{A}_{g_1, g_2} \boldsymbol{\varepsilon}_{g_2} \boldsymbol{\varepsilon}_{g_1}' \boldsymbol{B}_{g_1, g_2} \boldsymbol{\varepsilon}_{g_2}) + E(\boldsymbol{\varepsilon}_{g_1}' \boldsymbol{A}_{g_1, g_2} \boldsymbol{\varepsilon}_{g_2} \boldsymbol{\varepsilon}_{g_2}' \boldsymbol{B}_{g_2, g_1} \boldsymbol{\varepsilon}_{g_2}) \right\}, \\
&= \sum_{g_1=1}^{G} \sum_{g_2 \neq g_1}^{G} \left\{ E(\boldsymbol{\varepsilon}_{g_1}' \boldsymbol{A}_{g_1, g_2} \boldsymbol{\varepsilon}_{g_2} \boldsymbol{\varepsilon}_{g_2}' \boldsymbol{B}^*_{g_2, g_1} \boldsymbol{\varepsilon}_{g_2}) \right\} \\
&= \sum_{g_1=1}^{G} \sum_{g_2 \neq g_1}^{G} tr(\boldsymbol{\Sigma}_{g_1 g_1} \boldsymbol{A}_{g_1 g_2} \boldsymbol{\Sigma}_{g_2 g_2} \boldsymbol{B}^*_{g_2, g_1}), \\
&= \sum_{g_1=1}^{G} \sum_{g_2 = g_1}^{G} tr(\boldsymbol{\Sigma}_{g_1 g_1} \boldsymbol{A}_{g_1 g_2} \boldsymbol{\Sigma}_{g_2 g_2} \boldsymbol{B}^*_{g_2, g_1}),
\end{aligned}$$

where the last equality is the diagonal blocks in $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ are all zeros. □

Lemma A.2. Suppose that $\boldsymbol{\varepsilon}_n$ satisfy assumption 1, $\boldsymbol{A}_n$ is an $n \times n$ matrix whose row and column sums are uniformly bounded. Then, $E(\boldsymbol{\varepsilon}_n' \boldsymbol{A}_n \boldsymbol{\varepsilon}_n) = O(n), V(\boldsymbol{\varepsilon}_n' \boldsymbol{A}_n \boldsymbol{\varepsilon}_n) = O(n), \boldsymbol{\varepsilon}_n' \boldsymbol{A}_n \boldsymbol{\varepsilon}_n = O_p(n)$, and $\frac{1}{n} \boldsymbol{\varepsilon}_n' \boldsymbol{A}_n \boldsymbol{\varepsilon}_n - \frac{1}{n} E(\boldsymbol{\varepsilon}_n' \boldsymbol{A}_n \boldsymbol{\varepsilon}_n) = o_p(1)$.

Proof. From the assumption of $\boldsymbol{\varepsilon}_n$, $\boldsymbol{\varepsilon}_n = \boldsymbol{\Sigma}_n^{-\frac{1}{2}} \boldsymbol{v}_n$ and both row and column sums of $\boldsymbol{\Sigma}_n$ are uniformly bounded because the cluster size fo each cluster is finite. Thus, both row and column sums of $\boldsymbol{A}_n^* = \boldsymbol{\Sigma}_n^{-\frac{1}{2}} \boldsymbol{A}_n \boldsymbol{\Sigma}_n^{-\frac{1}{2}}$ are also uniformly bounded. By applying Lemma A.3 in Lin and Lee (2010) to $\boldsymbol{\varepsilon}_n' \boldsymbol{A}_n \boldsymbol{\varepsilon}_n = \boldsymbol{v}_n' \boldsymbol{A}_n^* \boldsymbol{v}_n$, the lemma is proofed. □

Lemma A.3. Suppose that $\boldsymbol{\varepsilon}_n$ satisfy assumption 1, $\boldsymbol{A}_n$ is an $n \times n$ matrix whose row and column sums are uniformly bounded, and $\boldsymbol{C}_n$ is an $n \times k$ matrix whose elements are uniformly bounded. Then, $\frac{1}{\sqrt{n}} \boldsymbol{C}_n' \boldsymbol{A}_n \boldsymbol{\varepsilon}_n =$

$O_p(1)$ and $\frac{1}{n} = \boldsymbol{C}_n' \boldsymbol{A}_n \boldsymbol{\varepsilon}_n = o_p(1)$. Moreover, if the limit $\frac{1}{n}\boldsymbol{C}_n' \boldsymbol{A}_n \boldsymbol{\Sigma}_n \boldsymbol{A}_n' \boldsymbol{C}_n$ exists and is positive definite, then $\frac{1}{\sqrt{n}}\boldsymbol{C}_n' \boldsymbol{A}_n \boldsymbol{\varepsilon}_n \xrightarrow{d} N(0, \lim_{n\to\infty} \frac{1}{n}\boldsymbol{C}_n' \boldsymbol{A}_n \boldsymbol{\Sigma}_n \boldsymbol{A}_n' \boldsymbol{C}_n)$

Proof. From the assumption of $\boldsymbol{\varepsilon}_n$, $\boldsymbol{\varepsilon}_n = \boldsymbol{\Sigma}_n^{-\frac{1}{2}} \boldsymbol{v}_n$ and both row and column sums of $\boldsymbol{\Sigma}_n$ are uniformly bounded because the cluster size fo each cluster is finite. Thus, both row and column sums of $\boldsymbol{A}_n^* = \boldsymbol{A}_n \boldsymbol{\Sigma}_n^{-\frac{1}{2}}$ are also uniformly bounded. By applying Lemma A.4 in Lin and Lee (2010) to $\boldsymbol{C}_n \boldsymbol{A}_n \boldsymbol{\varepsilon}_n = \boldsymbol{C}_n' \boldsymbol{A}_n^* \boldsymbol{v}_n$, the lemma is proofed. □

Lemma A.4. Suppose that $\boldsymbol{\varepsilon}_n$ satisfy assumption 1, $\boldsymbol{A}_n$ is an $n \times n$ matrix whose row and column sums are uniformly bounded, and $\boldsymbol{b}_n = (b_1, \ldots, b_n)'$ is an n dimensional vector such that $\sup_n \frac{1}{n} \sum_{i=1}^n |b_i|^{2+\eta} < \infty$ for some $\eta > 0$. Moreover, we define $Q_n = \boldsymbol{\varepsilon}_n' \boldsymbol{A}_n \boldsymbol{\varepsilon}_n + \boldsymbol{b}_n' \boldsymbol{\varepsilon}_n$, and $\mu_{Q_n} = E(Q_n)$ and $\sigma_{Q_n}^2 = V(Q_n)$, respectively. Then, $\frac{Q_n - u_{Q_n}}{\sigma_{Q_n}} \xrightarrow{d} N(0,1)$.

Proof. From the assumption of $\boldsymbol{\varepsilon}_n$, $\boldsymbol{\varepsilon}_n = \boldsymbol{\Sigma}_n^{-\frac{1}{2}} \boldsymbol{v}_n$ and both row and column sums of $\boldsymbol{\Sigma}_n$ are uniformly bounded because the cluster size fo each cluster is finite. Thus, both row and column sums of $\boldsymbol{A}_n^* = \boldsymbol{\Sigma}_n^{-\frac{1}{2}} \boldsymbol{A}_n \boldsymbol{\Sigma}_n^{-\frac{1}{2}}$ are also uniformly bounded, and $\boldsymbol{b}_n^{*'} = \boldsymbol{b}_n' \boldsymbol{\Sigma}_n^{-\frac{1}{2}}$ satisfies $\sup_n \frac{1}{n} \sum_{i=1}^n |b_i^*|^{2+\eta} < \infty$ for some $\eta > 0$. From Theorem 1 in Kelejian and Prucha (2001),

$$\frac{Q_n - u_{Q_n}}{\sigma_{Q_n}} = \frac{\boldsymbol{v}_n' \boldsymbol{A}_n^* \boldsymbol{v}_n + \boldsymbol{b}_n^{*'} \boldsymbol{\varepsilon}_n - u_{Q_n}}{\sigma_{Q_n}} \xrightarrow{d} N(0,1)$$

□

B Proofs of Theorems 1-3

Proof of Theorem 1

Because the SAR model for cluster dependent errors proposed in this paper is the same as the SAR model proposed in Lin and Lee (2010) except for the assumption of error terms. Therefore, by replacing Lemma A.3, A.4 and A.5 in Lin and Lee (2010) with Lemma A.2, A.3 and A.4 in this paper, Theorem 1 can be proved by the same argument as Proposition 1 in Lee (2007) and Proposition 1 in Lin and Lee (2010). The following discussion is almost the same as proof of Proposition 1 in Lee (2007) and Proposition 1 in Lin and Lee (2010).

Let $\boldsymbol{s}_n(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{a}_n \boldsymbol{g}_n(\boldsymbol{\theta})$, and $\boldsymbol{s}(\boldsymbol{\theta}) = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{a}_n E[\boldsymbol{g}_n(\boldsymbol{\theta})]$. Because the consistency of the proposed GMM estimator is shown by Theorem 2.1 of Newey and McFadden (1994), we check the following 4 conditions: (i) the parameter space $\boldsymbol{\Theta}$ is compact, (ii) $\boldsymbol{s}(\boldsymbol{\theta})'\boldsymbol{s}(\boldsymbol{\theta})$ is a continuous at $\boldsymbol{\theta}$, (iii) $\boldsymbol{s}_n(\boldsymbol{\theta})'\boldsymbol{s}_n(\boldsymbol{\theta})$ converges to $\boldsymbol{s}(\boldsymbol{\theta})'\boldsymbol{s}(\boldsymbol{\theta})$ in probability uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, (iv) $-\boldsymbol{s}(\boldsymbol{\theta})'\boldsymbol{s}(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\Theta}_0$. By the assumption of the parameter space, $\boldsymbol{\Theta}$ is compact. As showed in (7) below, $\boldsymbol{s}_n(\boldsymbol{\theta})'\boldsymbol{s}_n(\boldsymbol{\theta})$ is a polynomial function of $\boldsymbol{\theta}$. Thus, the condition (ii) holds.

First, let us consider the identification condition (iv). We note that $-\boldsymbol{s}(\boldsymbol{\theta})'\boldsymbol{s}(\boldsymbol{\theta}) \leq 0$ and $\boldsymbol{s}(\boldsymbol{\theta}) = \boldsymbol{0}$ has a unique root at $\boldsymbol{\theta}_0$ from the assumption. Thus, $-\boldsymbol{s}(\boldsymbol{\theta})'\boldsymbol{s}(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\Theta}_0$.

Next, we will show the uniform convergence condition (iii). It suffices to show that $\boldsymbol{s}_n(\boldsymbol{\theta})$ converges to $\boldsymbol{s}(\boldsymbol{\theta})$ in probability uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let $\boldsymbol{a}_n = (\boldsymbol{a}_{n1}, \ldots, \boldsymbol{a}_{nm}, \boldsymbol{a}_{nx})$ where $\boldsymbol{a}_{nj}$ is the j-th column of the matrix $\boldsymbol{A}_n$ and $\boldsymbol{a}_{nx}$ is the submatrix of $\boldsymbol{A}_n$. We denote the i-the element of $\boldsymbol{a}_n$ as $\boldsymbol{a}_{i,n} = (a_{i,n1}, \ldots, a_{i,nm}, \boldsymbol{a}_{i,nx})$ where $a_{i,nj}, j = 1, \ldots, m$ are scalar and $\boldsymbol{a}_{i,nx}$ is a sub vector. It suffices to show that the uniform convergence of $\frac{1}{n}\boldsymbol{a}_{i,n}\boldsymbol{g}_n(\boldsymbol{\theta})$ for each i.

From (2), $\boldsymbol{a}_{i,n}\boldsymbol{g}_n(\boldsymbol{\theta}) = \boldsymbol{\varepsilon}'_n(\boldsymbol{\theta})\big(\sum_{j=1}^m a_{i,nj}\boldsymbol{P}_{n,j}\big)\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) + \boldsymbol{a}_{i,nx}\boldsymbol{Q}'_n\boldsymbol{\varepsilon}_n(\boldsymbol{\theta})$. By the definition of the model (1), $\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) = \boldsymbol{d}_n(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}_n + (\lambda_0 - \lambda)\boldsymbol{G}_n\boldsymbol{\varepsilon}_n$ where $\boldsymbol{d}_n(\boldsymbol{\theta}) = (\lambda_0 - \lambda)\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0 + \boldsymbol{X}_n(\boldsymbol{\beta}_0 - \boldsymbol{\beta})$. Thus, $\boldsymbol{\varepsilon}'_n(\boldsymbol{\theta})\big(\sum_{j=1}^m a_{i,nj}\boldsymbol{P}_{n,j}\big)\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) = \boldsymbol{d}'_n(\boldsymbol{\theta})\big(\sum_{j=1}^m a_{i,nj}\boldsymbol{P}_{n,j}\big)\boldsymbol{d}_n(\boldsymbol{\theta}) + \boldsymbol{l}_n(\boldsymbol{\theta}) + \boldsymbol{q}_n(\boldsymbol{\theta})$ where $\boldsymbol{l}_n(\boldsymbol{\theta}) = \boldsymbol{d}'_n(\boldsymbol{\theta})\big(\sum_{j=1}^m a_{i,nj}\boldsymbol{P}^s_{n,j}\big)(\boldsymbol{\varepsilon}_n + (\lambda_0 - \lambda)\boldsymbol{G}_n\boldsymbol{\varepsilon}_n)$ and $\boldsymbol{q}_n(\boldsymbol{\theta}) = (\boldsymbol{\varepsilon}_n + (\lambda_0 - \lambda)\boldsymbol{G}_n\boldsymbol{\varepsilon}_n)'\big(\sum_{j=1}^m a_{i,nj}\boldsymbol{P}_{n,j}\big)(\boldsymbol{\varepsilon}_n + (\lambda_0 - \lambda)\boldsymbol{G}_n\boldsymbol{\varepsilon}_n)$. Furthermore, $\boldsymbol{a}_{i,nx}\boldsymbol{Q}'_n\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) = \boldsymbol{a}_{i,nx}\boldsymbol{Q}'_n(\boldsymbol{d}_n(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}_n + (\lambda_0 - \lambda)\boldsymbol{G}_n\boldsymbol{\varepsilon}_n)$.

The nonstochastic function $\frac{1}{n}\boldsymbol{a}_n E[\boldsymbol{g}_n(\boldsymbol{\theta})]$ is given by,

$$\frac{1}{n}\boldsymbol{a}_n E[\boldsymbol{g}_n(\boldsymbol{\theta})] = \frac{1}{n}\boldsymbol{d}'_n(\boldsymbol{\theta})\Big(\sum_{j=1}^m a_{i,nj}\boldsymbol{P}_{n,j}\Big)\boldsymbol{d}_n(\boldsymbol{\theta}) + \frac{1}{n}\sum_{j=1}^m a_{i,nj} tr(\boldsymbol{\Sigma}_n\boldsymbol{P}_{n,j}) + (\lambda_0 - \lambda)\frac{1}{n}\sum_{j=1}^m a_{i,nj} tr(\boldsymbol{\Sigma}_n\boldsymbol{G}'_n\boldsymbol{P}^s_{n,j}) \tag{6}$$

$$+ (\lambda_0 - \lambda)^2\frac{1}{n}\sum_{j=1}^m a_{i,nj} tr(\boldsymbol{\Sigma}_n\boldsymbol{G}'_n\boldsymbol{P}_{n,j}\boldsymbol{G}_n) + \frac{1}{n}\boldsymbol{a}_{i,nx}\boldsymbol{Q}'_n\boldsymbol{d}_n(\boldsymbol{\theta}). \tag{7}$$

From the definition,

$$\frac{1}{n}\boldsymbol{l}_n(\boldsymbol{\theta}) = (\lambda_0 - \lambda)\frac{1}{n}(\boldsymbol{X}_n\boldsymbol{\beta}_0)'\left(\sum_{j=1}^m a_{i,nj}\boldsymbol{G}_n\boldsymbol{P}^s_{n,j}\right)\boldsymbol{\varepsilon}_n + (\lambda_0 - \lambda)^2\frac{1}{n}(\boldsymbol{X}_n\boldsymbol{\beta}_0)'\left(\sum_{j=1}^m a_{i,nj}\boldsymbol{G}_n\boldsymbol{P}^s_{n,j}\boldsymbol{G}_n\right)\boldsymbol{\varepsilon}_n$$
$$+ (\boldsymbol{\beta}_0 - \boldsymbol{\beta})\frac{1}{n}\boldsymbol{X}'_n\left(\sum_{j=1}^m a_{i,nj}\boldsymbol{P}^s_{n,j}\right)\boldsymbol{\varepsilon}_n + (\boldsymbol{\beta}_0 - \boldsymbol{\beta})(\lambda_0 - \lambda)\frac{1}{n}\boldsymbol{X}'_n\left(\sum_{j=1}^m a_{i,nj}\boldsymbol{P}^s_{n,j}\boldsymbol{G}_n\right)\boldsymbol{\varepsilon}_n.$$

We note that both row and column sums of $\boldsymbol{G}_n\boldsymbol{P}^s_{n,j}$ and $\boldsymbol{G}_n\boldsymbol{P}^s_{n,j}\boldsymbol{G}_n$ are uniformly bounded because both row and column sums of $\boldsymbol{G}_n$ and $\boldsymbol{P}^s_{n,j}$ are uniformly bounded. From the boundedness of the parameter space and Lemma A.3, $\frac{1}{n}\boldsymbol{l}_n(\boldsymbol{\theta}) = o_p(1)$ uniformly in $\theta \in \Theta$.

By Lemma A.2,

$$
\begin{aligned}
\frac{1}{n}\boldsymbol{q}_n(\boldsymbol{\theta}) = & \frac{1}{n}\boldsymbol{\varepsilon}_n'\left(\sum_{j=1}^m a_{i,nj}\boldsymbol{P}_{n,j}\right)\boldsymbol{\varepsilon}_n + (\lambda_0-\lambda)\frac{1}{n}\boldsymbol{\varepsilon}_n'\left(\sum_{j=1}^m a_{i,nj}\boldsymbol{G}_n'\boldsymbol{P}_{n,j}^s\right)\boldsymbol{\varepsilon}_n \\
& + (\lambda_0-\lambda)^2\boldsymbol{\varepsilon}_n'\left(\sum_{j=1}^m a_{i,nj}\boldsymbol{G}_n'\boldsymbol{P}_{n,j}\boldsymbol{G}_n\right)\boldsymbol{\varepsilon}_n, \\
= & (\lambda_0-\lambda)\frac{1}{n}\sum_{j=1}^m a_{i,nj}tr(\boldsymbol{\Sigma}_n\boldsymbol{G}_n'\boldsymbol{P}_{n,j}^s) + (\lambda_0-\lambda)^2\frac{1}{n}\sum_{j=1}^m a_{i,nj}tr(\boldsymbol{\Sigma}_n\boldsymbol{G}_n'\boldsymbol{P}_{n,j}\boldsymbol{G}_n) + o_p(1),
\end{aligned}
$$

uniformly in $\boldsymbol{\theta}\in\boldsymbol{\Theta}$.

Consequently,

$$
\begin{aligned}
\frac{1}{n}\boldsymbol{\varepsilon}_n'(\boldsymbol{\theta})\left(\sum_{j=1}^m a_{i,nj}\boldsymbol{P}_{n,j}\right)\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) = & \boldsymbol{d}_n'(\boldsymbol{\theta})\left(\sum_{j=1}^m a_{i,nj}\boldsymbol{P}_{n,j}\right)\boldsymbol{d}_n(\boldsymbol{\theta}) + (\lambda_0-\lambda)\frac{1}{n}\sum_{j=1}^m a_{i,nj}tr(\boldsymbol{\Sigma}_n\boldsymbol{G}_n'\boldsymbol{P}_{n,j}^s) \\
& + (\lambda_0-\lambda)^2\frac{1}{n}\sum_{j=1}^m a_{i,nj}tr(\boldsymbol{\Sigma}_n\boldsymbol{G}_n'\boldsymbol{P}_{n,j}\boldsymbol{G}_n) + o_p(1),
\end{aligned}
$$

uniformly in $\boldsymbol{\theta}\in\boldsymbol{\Theta}$.

Similarly,

$$
\begin{aligned}
\frac{1}{n}\boldsymbol{a}_{i,nx}\boldsymbol{Q}_n'\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) &= \frac{1}{n}\boldsymbol{a}_{i,nx}\boldsymbol{Q}_n'\boldsymbol{d}_n(\boldsymbol{\theta}) + \frac{1}{n}\boldsymbol{a}_{i,nx}\boldsymbol{Q}_n'(\boldsymbol{\varepsilon}_n + (\lambda_0-\lambda)\boldsymbol{G}_n\boldsymbol{\varepsilon}_n), \\
&= \frac{1}{n}\boldsymbol{a}_{i,nx}\boldsymbol{Q}_n'\boldsymbol{d}_n(\boldsymbol{\theta}) + o_p(1),
\end{aligned}
$$

uniformly in $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ by Lemma A.3.

To sum up the above arguments, we obtain that $\boldsymbol{s}_n(\boldsymbol{\theta})$ converges to $\boldsymbol{s}(\boldsymbol{\theta})$ in probability uniformly in $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, and thus the consistency of the GMM estimator is shown.

Next, let us consider the asymptotic normality of the GMM estimator. We note that $\frac{\partial \boldsymbol{g}_n'(\hat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}\boldsymbol{a}_n'\boldsymbol{a}_n\boldsymbol{g}_n(\hat{\boldsymbol{\theta}}_n) = 0$ by the definition of $\hat{\boldsymbol{\theta}}_n$. By applying the Taylor's expansion to the above equation, we have

$$
\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\left(\frac{1}{n}\frac{\partial \boldsymbol{g}_n'(\hat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}\boldsymbol{a}_n'\boldsymbol{a}_n\frac{1}{n}\frac{\partial \boldsymbol{g}_n(\bar{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}\right)^{-1}\frac{1}{n}\frac{\partial \boldsymbol{g}_n'(\hat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}\boldsymbol{a}_n'\frac{1}{\sqrt{n}}\boldsymbol{a}_n\boldsymbol{g}_n(\boldsymbol{\theta}_0),
$$

where each element in $\bar{\boldsymbol{\theta}}_n$ is between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$.

First, we will show that $\frac{1}{n}\frac{\partial \boldsymbol{g}_n(\hat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}$ converges in probability to $-\frac{1}{n}\boldsymbol{D}_n$. By the model (1) and the moment condition (2), $\frac{\partial\boldsymbol{\varepsilon}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} = (\boldsymbol{W}_n\boldsymbol{Y}_n, \boldsymbol{X}_n)$ and $\frac{\partial \boldsymbol{g}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} = (\boldsymbol{P}_{n,1}^s\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}),\ldots,\boldsymbol{P}_{n,m}^s\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}),\boldsymbol{Q}_n)'(\boldsymbol{W}_n\boldsymbol{Y}_n,\boldsymbol{X}_n)$. By the model (1)

$\frac{1}{n}\boldsymbol{\varepsilon}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{W}_n\boldsymbol{Y}_n = \frac{1}{n}\boldsymbol{\varepsilon}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0 + \frac{1}{n}\boldsymbol{\varepsilon}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n\boldsymbol{\varepsilon}_n, j = 1,\ldots,m$. By Lemma A.3,

$$\begin{aligned}\frac{1}{n}\boldsymbol{\varepsilon}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0 &= \frac{1}{n}\boldsymbol{d}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0 + \frac{1}{n}\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0 + (\lambda_0-\lambda)\frac{1}{n}\boldsymbol{\varepsilon}_n'\boldsymbol{G}_n'\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0,\\ &= \frac{1}{n}\boldsymbol{d}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0 + o_p(1).\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{1}{n}\boldsymbol{\varepsilon}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n\boldsymbol{\varepsilon}_n &= \frac{1}{n}\boldsymbol{d}_n'(\boldsymbol{\theta})'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n\boldsymbol{\varepsilon}_n + \frac{1}{n}\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n\boldsymbol{\varepsilon}_n + (\lambda_0-\lambda)\frac{1}{n}\boldsymbol{\varepsilon}_n'\boldsymbol{G}_n'\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n\boldsymbol{\varepsilon}_n,\\ &= \frac{1}{n}tr(\boldsymbol{\Sigma}_n\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n) + (\lambda_0-\lambda)\frac{1}{n}tr(\boldsymbol{\Sigma}_n\boldsymbol{G}_n'\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n) + o_p(1),\end{aligned}$$

by Lemma A.2 and A.3.

Consequently,

$$\frac{1}{n}\boldsymbol{\varepsilon}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{W}_n\boldsymbol{Y}_n = \frac{1}{n}\boldsymbol{d}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0 + \frac{1}{n}tr(\boldsymbol{\Sigma}_n\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n) + (\lambda_0-\lambda)\frac{1}{n}tr(\boldsymbol{\Sigma}_n\boldsymbol{G}_n'\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n) + o_p(1).$$

We note that $\hat{\theta}_n$ is consistent estimator from the above discussion, and $\boldsymbol{d}_n(\boldsymbol{\theta}_0) = \boldsymbol{0}$. Thus, $\lim_{n\to\infty}\frac{1}{n}\boldsymbol{\varepsilon}_n'(\hat{\boldsymbol{\theta}})\boldsymbol{P}_{n,j}^s\boldsymbol{W}_n\boldsymbol{Y}_n = \frac{1}{n}tr(\boldsymbol{\Sigma}_n\boldsymbol{G}_n'\boldsymbol{P}_{n,j}^s\boldsymbol{G}_n)$.

For $\frac{1}{n}\boldsymbol{\varepsilon}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{X}_n$,

$$\begin{aligned}\frac{1}{n}\boldsymbol{\varepsilon}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{X}_n &= \frac{1}{n}\boldsymbol{d}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{X}_n + \frac{1}{n}\boldsymbol{\varepsilon}_n'\boldsymbol{P}_{n,j}^s\boldsymbol{X}_n + (\lambda_0-\lambda)\frac{1}{n}\boldsymbol{\varepsilon}_n'\boldsymbol{G}_n\boldsymbol{P}_{n,j}^s\boldsymbol{X}_n,\\ &= \frac{1}{n}\boldsymbol{d}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}^s\boldsymbol{X}_n + o_p(1),\end{aligned}$$

by Lemma A.3, and thus $\lim_{n\to\infty}\frac{1}{n}\boldsymbol{\varepsilon}_n'(\hat{\boldsymbol{\theta}})\boldsymbol{P}_{n,j}^s\boldsymbol{X}_n = o_p(1)$.

For $\frac{1}{n}\boldsymbol{Q}_n\boldsymbol{W}_n\boldsymbol{Y}_n$,

$$\frac{1}{n}\boldsymbol{Q}_n\boldsymbol{W}_n\boldsymbol{Y}_n = \frac{1}{n}\boldsymbol{Q}_n\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0 + \frac{1}{n}\boldsymbol{Q}_n\boldsymbol{G}_n\boldsymbol{\varepsilon}_n, = \frac{1}{n}\boldsymbol{Q}_n\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0 + o_p(1),$$

by Lemma A.3.

To sum up the above argument, we obtain that $\frac{1}{n}\frac{\partial \boldsymbol{g}_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}$ converges in probability to $-\frac{1}{n}\boldsymbol{D}_n$. Because $\bar{\boldsymbol{\theta}}_n$ is between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$, $\frac{1}{n}\frac{\partial \boldsymbol{g}_n(\bar{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}$ also converges in probability to $-\frac{1}{n}\boldsymbol{D}_n$.

Next, we will show that $\frac{1}{\sqrt{n}}\boldsymbol{a}_n\boldsymbol{g}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0}, \lim_{n\to\infty}\boldsymbol{a}_n\boldsymbol{\Omega}_n\boldsymbol{a}_n')$. Let n_a be the number of rows of $\boldsymbol{a}_n$. $\boldsymbol{c}$ is an $n_a\times 1$ vector of any constants with $\boldsymbol{c}'\boldsymbol{c} = 1$. By Lemma A.4, $\frac{1}{\sqrt{n}}\boldsymbol{c}'\boldsymbol{a}_n\boldsymbol{g}_n(\boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}[\boldsymbol{\varepsilon}_n'(\sum_{k=1}^{n_a}\sum_{j=1}^{m}c_k a_{k,nj}\boldsymbol{P}_{n,j})\boldsymbol{\varepsilon}_n +$

$\sum_{k=1}^{n_a} c_k \boldsymbol{a}_{k,nx} \boldsymbol{Q}_n \boldsymbol{\varepsilon}_n] \xrightarrow{d} N\big(0, \lim_{n\to\infty} \frac{1}{n} \boldsymbol{c}' \boldsymbol{a}_n' \boldsymbol{\Omega}_n \boldsymbol{a}_n \boldsymbol{c}\big)$. By the Cramér-Wold Device, $\frac{1}{\sqrt{n}} \boldsymbol{a}_n \boldsymbol{g}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \lim_{n\to\infty} \boldsymbol{a}_n \boldsymbol{\Omega}_n \boldsymbol{a}_n')$.

Gathering the above discussion, asymptotic normality of the GMM estimator is obtained by the Slutsky's theorem.

Proof of Theorem 2

A. The consistency of $\frac{1}{n}\hat{\boldsymbol{\Omega}}_n$: Let us consider that each elements in $\frac{1}{n}\hat{\boldsymbol{\Omega}}_n - \frac{1}{n}\boldsymbol{\Omega}_n$ converges to zero in probability.
(a): The consistency of some elements: One generic element in the matrix $\frac{1}{n}\boldsymbol{\Omega}_n$ is written by

$$E(\boldsymbol{\varepsilon}_n' \boldsymbol{P}_{n,a} \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n' \boldsymbol{P}_{n,b} \boldsymbol{\varepsilon}_n) = \frac{1}{n} \sum_{g_1=1}^{G} \sum_{g_2=1}^{G} \sum_{g_3=1}^{G} \sum_{g_4=1}^{G} E(\boldsymbol{\varepsilon}_{g_1}' \boldsymbol{P}_{a,g_1 g_2} \boldsymbol{\varepsilon}_{g_2} \boldsymbol{\varepsilon}_{g_3}' \boldsymbol{P}_{b,g_3 g_4} \boldsymbol{\varepsilon}_{g_4}), \tag{8}$$

Because error terms are independent across clusters and block diagonal elements in $\boldsymbol{P}_{n,a}$ and $\boldsymbol{P}_{n,b}$ are zeros, $E(\boldsymbol{\varepsilon}_{g_1}' \boldsymbol{P}_{a,g_1 g_2} \boldsymbol{\varepsilon}_{g_2} \boldsymbol{\varepsilon}_{g_3}' \boldsymbol{P}_{b,g_3 g_4} \boldsymbol{\varepsilon}_{g_4})$ will not vanish only when $g_1 = g_3 \neq g_2 = g_4$ and $g_1 = g_4 \neq g_2 = g_3$. The same argument holds when $g_1 = g_4 \neq g_2 = g_3$ as below, thus we show convergence in probability when $g_1 = g_3 \neq g_2 = g_4$.

$$\begin{aligned} E(\boldsymbol{\varepsilon}_n' \boldsymbol{P}_{a,n} \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n' \boldsymbol{P}_{b,n} \boldsymbol{\varepsilon}_n) &= \frac{1}{n} \sum_{g_1=1}^{G} \sum_{g_2=1}^{G} E(\boldsymbol{\varepsilon}_{g_1}' \boldsymbol{P}_{a,g_1,g_2} \boldsymbol{\varepsilon}_{g_2} \boldsymbol{\varepsilon}_{g_1}' \boldsymbol{P}_{b,g_1,g_2} \boldsymbol{\varepsilon}_{g_2}), \\ &= \frac{1}{n} \sum_{g_1=1}^{G} \sum_{g_2=1}^{G} \sum_{i_1=1}^{n_{g_1}} \sum_{i_2=1}^{n_{g_2}} \sum_{i_3=1}^{n_{g_1}} \sum_{i_4=1}^{n_{g_2}} p^a_{g_1 g_2, i_1 i_2} p^b_{g_1 g_2, i_3 i_4} E(\varepsilon_{g_1,i_1} \varepsilon_{g_1,i_3}) E(\varepsilon_{g_2,i_2} \varepsilon_{g_2,i_4}), \end{aligned}$$

where $p^a_{g_1 g_2, i_1 i_2}$ is the (i_1, i_2) element of $\boldsymbol{P}_{a,g_1,g_2}$ and $p^b_{g_1 g_2, i_3 i_4}$ is defined similarly.

First, we show that

$$\frac{1}{n} \sum_{g_1=1}^{G} \sum_{g_2=1}^{G} \sum_{i_1=1}^{n_{g_1}} \sum_{i_2=1}^{n_{g_2}} \sum_{i_3=1}^{n_{g_1}} \sum_{i_4=1}^{n_{g_2}} p^a_{g_1 g_2, i_1 i_2} p^b_{g_1 g_2, i_3 i_4} [\varepsilon_{g_1,i_1} \varepsilon_{g_1,i_3} \varepsilon_{g_2,i_2} \varepsilon_{g_2,i_4} - E(\varepsilon_{g_1,i_1} \varepsilon_{g_1,i_3}) E(\varepsilon_{g_2,i_2} \varepsilon_{g_2,i_4})] = o_p(1)$$

and after that we establish that this convergence holds when $\varepsilon_{g,i}$s are replaced by the residuals $\hat{\varepsilon}_{g,i}$

Note that

$$\begin{aligned} \varepsilon_{g_1,i_1} \varepsilon_{g_1,i_3} \varepsilon_{g_2,i_2} \varepsilon_{g_2,i_4} - E(\varepsilon_{g_1,i_1} \varepsilon_{g_1,i_3}) E(\varepsilon_{g_2,i_2} \varepsilon_{g_2,i_4}) =& [\varepsilon_{g_1,i_1} \varepsilon_{g_1,i_3} - E(\varepsilon_{g_1,i_1} \varepsilon_{g_1,i_3})][\varepsilon_{g_2,i_2} \varepsilon_{g_2,i_4} - E(\varepsilon_{g_2,i_2} \varepsilon_{g_2,i_4})] \\ &+ E(\varepsilon_{g_2,i_2} \varepsilon_{g_2,i_4})[\varepsilon_{g_1,i_1} \varepsilon_{g_1,i_3} - E(\varepsilon_{g_1,i_1} \varepsilon_{g_1,i_3})] \\ &+ E(\varepsilon_{g_1,i_1} \varepsilon_{g_1,i_3})[\varepsilon_{g_2,i_2} \varepsilon_{g_2,i_4} - E(\varepsilon_{g_2,i_2} \varepsilon_{g_2,i_4})], \end{aligned}$$

Thus,

$$\begin{aligned}
&\frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}[\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3}\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4} - E(\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3})E(\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4})],\\
&= H_1 + H_2 + H_3,
\end{aligned}$$

where $H_1 = \frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}[\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3} - E(\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3})][\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4} - E(\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4})]$, $H_2 = \frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}E(\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4})[\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3} - E(\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3})]$ and $H_3 = \frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}E(\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3})[\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4} - E(\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4})]$.

Because error terms are independent across clusters and block diagonal elements in $\boldsymbol{P}_{n,a}$ and $\boldsymbol{P}_{n,b}$ are zeros, $E(H_1) = 0$.

$$\begin{aligned}
H_1^2 =& \frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{g_3=1}^{G}\sum_{g_4=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}}\sum_{i_5=1}^{n_{g_3}}\sum_{i_6=1}^{n_{g_4}}\sum_{i_7=1}^{n_{g_3}}\sum_{i_8=1}^{n_{g_4}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}p^a_{g_1g_2,i_5i_6}p^b_{g_1g_2,i_7i_8}\\
&[\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3} - E(\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3})][\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4} - E(\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4})][\varepsilon_{g_1,i_5}\varepsilon_{g_1,i_7} - E(\varepsilon_{g_1,i_5}\varepsilon_{g_1,i_7})][\varepsilon_{g_2,i_6}\varepsilon_{g_2,i_8} - E(\varepsilon_{g_2,i_6}\varepsilon_{g_2,i_8})].
\end{aligned}$$

As error terms are independent across clusters and block diagonal elements in $\boldsymbol{P}_{n,a}$ and $\boldsymbol{P}_{n,b}$ are zeros, $E(H_1^2)$ will not vanish only when $g_1 = g_3 \neq g_2 = g_4$ and $g_1 = g_4 \neq g_2 = g_3$. The same argument holds when $g_1 = g_4 \neq g_2 = g_3$ as below, thus we show convergence in probability when $g_1 = g_3 \neq g_2 = g_4$.

Because $E|\varepsilon_{g,1}\varepsilon_{g,2}\varepsilon_{g,3}\varepsilon_{g,4}| \leq E(\varepsilon_{g,1}^4)^{\frac{1}{4}}E(\varepsilon_{g,2}^4)^{\frac{1}{4}}E(\varepsilon_{g,3}^4)^{\frac{1}{4}}E(\varepsilon_{g,4}^4)^{\frac{1}{4}} < c$ for some constant c and for all elements in $\boldsymbol{\varepsilon}_g$,

$$\begin{aligned}
E(H_1^2) =& \frac{1}{n^2}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}}\sum_{i_5=1}^{n_{g_1}}\sum_{i_6=1}^{n_{g_2}}\sum_{i_7=1}^{n_{g_1}}\sum_{i_8=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}p^a_{g_1g_2,i_5i_6}p^b_{g_1g_2,i_7i_8}\\
&[\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3} - E(\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3})][\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4} - E(\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4})][\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3} - E(\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3})][\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4} - E(\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4})],\\
&\leq \frac{c}{n^2}\sum_{g_1=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_5=1}^{n_{g_1}}\sum_{i_7=1}^{n_{g_1}}\left(\sum_{g_2=1}^{G}\sum_{i_2=1}^{n_{g_2}}\sum_{i_4=1}^{n_{g_2}}\sum_{i_6=1}^{n_{g_2}}\sum_{i_8=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}p^a_{g_1g_2,i_5i_6}p^b_{g_1g_2,i_7i_8}\right)\\
&\leq \frac{c}{n}\sum_{g_1=1}^{G} n_{g_1}^4,\\
&= o(1),
\end{aligned}$$

where the second inequality holds because the column sums of matrix $\boldsymbol{P}_{n,a}\boldsymbol{P}'_{n,b}\boldsymbol{P}_{n,a}\boldsymbol{P}'_{n,b}\boldsymbol{P}_{n,a}\boldsymbol{P}'_{n,b}\boldsymbol{P}_{n,a}\boldsymbol{P}'_{n,b}$ is uniformly bounded, and $\left(\sum_{g_2=1}^{G}\sum_{i_2=1}^{n_{g_2}}\sum_{i_4=1}^{n_{g_2}}\sum_{i_6=1}^{n_{g_2}}\sum_{i_8=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}p^a_{g_1g_2,i_5i_6}p^b_{g_1g_2,i_7i_8}\right)$ is a part of a column sum of $\boldsymbol{P}_{n,a}\boldsymbol{P}'_{n,b}\boldsymbol{P}_{n,a}\boldsymbol{P}'_{n,b}\boldsymbol{P}_{n,a}\boldsymbol{P}'_{n,b}\boldsymbol{P}_{n,a}\boldsymbol{P}'_{n,b}$. Thus, $H_1 = o_p(1)$ from the Chebyshev's inequality. The same argument holds for H_2 and H_3, and thus $H_2 = o_p(1)$ and $H_3 = o_p(1)$.

Consequently,

$$\frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}[\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3}\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4} - E(\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3})E(\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4})] = o_p(1). \quad (9)$$

Next, we will show that this convergence holds when $\varepsilon_{g,i}$s are replaced by the residuals $\hat{\varepsilon}_{g,i}$.

$$\begin{aligned}&\frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}[\hat{\varepsilon}_{g_1,i_1}\hat{\varepsilon}_{g_1,i_3}\hat{\varepsilon}_{g_2,i_2}\hat{\varepsilon}_{g_2,i_4} - \varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3}\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4}],\\ &= H_4 + H_5 + H_6,\end{aligned}$$

where $H_4 = \frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}(\hat{\varepsilon}_{g_1,i1}\hat{\varepsilon}_{g_1,i3}-\varepsilon_{g_1,i1}\varepsilon_{g_1,i3})(\hat{\varepsilon}_{g_2,i2}\hat{\varepsilon}_{g_2,i4}-\varepsilon_{g_2,i2}\varepsilon_{g_2,i4})$, $H_5 = \frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}\varepsilon_{g_2,i2}\varepsilon_{g_2,i4}(\hat{\varepsilon}_{g_1,i1}\hat{\varepsilon}_{g_1,i3} - \varepsilon_{g_1,i1}\varepsilon_{g_1,i3})$, and $H_6 = \frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}\varepsilon_{g_1,i1}\varepsilon_{g_1,i3}(\hat{\varepsilon}_{g_2,i2}\hat{\varepsilon}_{g_2,i4} - \varepsilon_{g_2,i2}\varepsilon_{g_2,i4})$.

From the model, we have

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}}_n &= \boldsymbol{S}_n(\hat{\lambda})\boldsymbol{Y}_n - \boldsymbol{X}\hat{\boldsymbol{\beta}},\\ &= \boldsymbol{\varepsilon}_n + (\lambda_0 - \hat{\lambda})\boldsymbol{G}_n\boldsymbol{\varepsilon}_n + \boldsymbol{X}_n(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + (\lambda_0 - \hat{\lambda})\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0.\end{aligned}$$

By using $e_{g_1,i}$ which is the (g_1,i)-th row in the $n\times n$ identity matrix, the i-the element in $\hat{\boldsymbol{\varepsilon}}_{g_1}$ is given by $\hat{\varepsilon}_{g_1,i} = \varepsilon_{g_1,i} + b_{g_1,i} + c_{g_1,i}$, where $b_{g_1,i} = (\lambda_0 - \hat{\lambda})e_{g_1,i}\boldsymbol{G}_n\boldsymbol{\varepsilon}_n$ and $c_{g_1,i} = e_{g_1,i}\boldsymbol{X}_n(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + (\lambda_0 - \hat{\lambda})e_{g_1,n}\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0$.

Let us consider H_4.

$$\begin{aligned}H_4 =& \frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}\\ &((\varepsilon_{g_1,i_1} + b_{g_1,i_1} + c_{g_1,i_1})(\varepsilon_{g_1,i_3} + b_{g_1,i_3} + c_{g_1,i_3}) - \varepsilon_{g_1,i1}\varepsilon_{g_1,i3})\\ &((\varepsilon_{g_2,i_2} + b_{g_2,i_2} + c_{g_2,i_2})(\varepsilon_{g_2,i_4} + b_{g_2,i_4} + c_{g_2,i_4}) - \varepsilon_{g_2,i2}\varepsilon_{g_2,i4})\end{aligned}$$

We pay attention to convergence in probability of terms with the higher orders in $\varepsilon_{g,i}$s. Convergences of the other terms are shown by the similar manner.

Let $\boldsymbol{G}_{(g_1,i_1),k_1}$ be the $((g_1,i_1),k_1)$ element of $\boldsymbol{G}_n$. The highest term with $\varepsilon_{g,i}$s is

$$\frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}b_{g_1,i_1}b_{g_1,i_3}b_{g_2,i_2}b_{g_2,i_4},$$
$$= (\lambda_0-\hat{\lambda})^4\frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}(e_{g_1,i_1}\boldsymbol{G}_n\boldsymbol{\varepsilon}_n)(e_{g_1,i_3}\boldsymbol{G}_n\boldsymbol{\varepsilon}_n)(e_{g_2,i_2}\boldsymbol{G}_n\boldsymbol{\varepsilon}_n)(e_{g_2,i_4}\boldsymbol{G}_n\boldsymbol{\varepsilon}_n),$$
$$= (\lambda_0-\hat{\lambda})^4K_n,$$

where $K_n = \frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}}\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}\sum_{k_3=1}^{n}\sum_{k_4=1}^{n} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3i_4}$ $\boldsymbol{G}_{(g_1,i_1),k_1}\boldsymbol{G}_{(g_2,i_2),k_2}\boldsymbol{G}_{(g_1,i_3),k_3}\boldsymbol{G}_{(g_2,i_4),k_4}\ \varepsilon_{k_1}\varepsilon_{k_2}\varepsilon_{k_3}\varepsilon_{k_4}$.

We note that $E|\varepsilon_{k_1}\varepsilon_{k_2}\varepsilon_{k_3}\varepsilon_{k_4}| \le E(\varepsilon^4_{k_1})^{\frac{1}{4}}E(\varepsilon^4_{k_2})^{\frac{1}{4}}E(\varepsilon^4_{k_3})^{\frac{1}{4}}E(\varepsilon^4_{k_4})^{\frac{1}{4}} < c$ for some constant c. By the uniform boundedness in row and column sums for $\boldsymbol{P}_a\boldsymbol{P}_b'$ and $\boldsymbol{G}$,

$$\begin{aligned}
E|K_n| &\le \frac{c}{n}\frac{1}{n}\sum_{g_1=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_3=1}^{n_{g_1}}\left(\sum_{g_2=1}^{G}\sum_{i_2=1}^{n_{g_2}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3,i_4}\right)\left(\sum_{k_1=1}^{n}\boldsymbol{G}_{(g_1,i_1),k_1}\right)\\
&\times\left(\sum_{k_2=1}^{n}\boldsymbol{G}_{(g_2,i_2),k_2}\right)\left(\sum_{k_3=1}^{n}\boldsymbol{G}_{(g_1,i_3),k_3}\right)\left(\sum_{k_4=1}^{n}\boldsymbol{G}_{(g_2,i_4),k_4}\right),\\
&\le c\frac{\sum_{g=1}^{G}n_g^2}{n},\\
&= O(1).
\end{aligned}$$

Thus, $K_n = O_p(1)$ by the Markov inequality. Convergence in probability of other terms in H_4 are shown similarly. We conclude that $H_4 = o_p(1)$. Moreover, $H_5 = o_p(1)$ and $H_6 = o_p(1)$ can be shown form the same argument. Consequently,

$$\frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3,i_4}[\hat{\varepsilon}_{g_1,i_1}\hat{\varepsilon}_{g_1,i_3}\hat{\varepsilon}_{g_2,i_2}\hat{\varepsilon}_{g_2,i_4}-\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3}\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4}] = o_p(1). \tag{10}$$

Combining (9) and (10), we have

$$\frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\sum_{i_1=1}^{n_{g_1}}\sum_{i_2=1}^{n_{g_2}}\sum_{i_3=1}^{n_{g_1}}\sum_{i_4=1}^{n_{g_2}} p^a_{g_1g_2,i_1i_2}p^b_{g_1g_2,i_3,i_4}[\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3}\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4}-E(\varepsilon_{g_1,i_1}\varepsilon_{g_1,i_3})E(\varepsilon_{g_2,i_2}\varepsilon_{g_2,i_4})],$$
$$= o_p(1)$$

(b) The consistency of the other elements: Let$q_{i,j}$ be the (i, j) element of the matrix $\boldsymbol{Q}_n$. The other elements in the matrix $\frac{1}{n}\boldsymbol{\Omega}_n$ are $\frac{1}{n}\sum_{g=1}^{G} Q'_{gg}\Sigma_{gg}Q_{gg}$ and the (i, j) element of the matrix is $\frac{1}{n}\sum_{g=1}^{G}\sum_{k_1=1}^{n}\sum_{k_2=1}^{n} q_{i,k_1}q_{k_2,j}E(\varepsilon_{g,k_1}\varepsilon_{g,k_2})$. This form is simpler than (8) and thus convergence in probability of the form can be shown with the same arguments in part (a) above.

In conclusion, we have shown that $\frac{1}{n}\hat{\boldsymbol{\Omega}}_n - \frac{1}{n}\boldsymbol{\Omega}_n = o_p(1)$.

B. The consistency of $\frac{1}{n}\hat{\boldsymbol{D}}_n$: Let $\boldsymbol{P}^*_{n1} = \boldsymbol{P}^s_1\boldsymbol{G}_n$. Note that the row and column sums of $\boldsymbol{P}^*_{n1}$ are uniformly bounded because the row and column sums of both $\boldsymbol{P}^s_1$ and $\boldsymbol{G}_n$ are uniformly bounded. One generic form for the elements of $\frac{1}{n}\boldsymbol{D}_n$ is $\frac{1}{n}E(\boldsymbol{\varepsilon}'_n\boldsymbol{P}^*_{n1}\boldsymbol{\varepsilon}_n) = \frac{1}{n}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\boldsymbol{\varepsilon}_{g_1}\boldsymbol{P}^*_{g_1g_2}\boldsymbol{\varepsilon}_{g_2}$. This form is simpler than (8) and thus convergence in probability of the form can be shown with the same arguments in part (a) above.

Together, Theorem 2 is proved.

Proof of Theorem 3

We have shown that $\frac{1}{n}\hat{\boldsymbol{\Omega}}_n - \frac{1}{n}\boldsymbol{\Omega}_n = o_p(1)$ and $\frac{1}{n}\hat{\boldsymbol{D}}_n - \frac{1}{n}\boldsymbol{D} = o_p(1)$ in Theorem 2. Thus, Theorem 3 is shown by the same argument as Proposition 2 in Lee (2007) and Proposition 3 in Lin and Lee (2010).

First, we will show that consistency of the feasible optimal GMM estimator $\hat{\boldsymbol{\theta}}_{o,n}$. Let $\boldsymbol{a}_{o,n} = (\frac{1}{n}\boldsymbol{\Omega}_n)^{-\frac{1}{2}}, \boldsymbol{a}_o = (\lim_{n\to\infty}\frac{1}{n}\boldsymbol{\Omega}_n)^{-\frac{1}{2}}$. Because $\hat{\boldsymbol{\Omega}}_n^{-1} = \boldsymbol{\Omega}_n^{-1} + (\hat{\boldsymbol{\Omega}}_n^{-1} - \boldsymbol{\Omega}_n^{-1})$,

$$\frac{1}{n}\boldsymbol{g}'_n(\boldsymbol{\theta})\hat{\boldsymbol{\Omega}}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{g}'_n(\boldsymbol{\theta})\boldsymbol{\Omega}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\theta}) + \frac{1}{n}\boldsymbol{g}'_n(\boldsymbol{\theta})(\hat{\boldsymbol{\Omega}}_n^{-1} - \boldsymbol{\Omega}_n^{-1})\boldsymbol{g}_n(\boldsymbol{\theta}).$$

Let $\boldsymbol{s}_o(\boldsymbol{\theta}) = \lim_{n\to\infty}\frac{1}{n}\boldsymbol{a}_oE[\boldsymbol{g}_n(\boldsymbol{\theta})]$. Because $\boldsymbol{a}_o$ is full rank (k+1) from the assumption, $\boldsymbol{s}_o(\boldsymbol{\theta}) = 0$ has a unique root at $\boldsymbol{\theta}_0$ from Assumption 5. Thus, $-\boldsymbol{s}_o(\boldsymbol{\theta})\boldsymbol{s}_o(\boldsymbol{\theta})$ which is the well defined limit of $\frac{1}{n}\boldsymbol{g}'_n(\boldsymbol{\theta})\boldsymbol{\Omega}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\theta})$ satisfies the identification uniqueness condition. Let $\tilde{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\,\boldsymbol{g}'_n(\boldsymbol{\theta})\boldsymbol{\Omega}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\theta})$. From Theorem 1, $\tilde{\boldsymbol{\theta}}$ is a consistent estimator. Furthermore, if $\frac{1}{n}\boldsymbol{g}'_n(\boldsymbol{\theta})(\hat{\boldsymbol{\Omega}}_n^{-1} - \boldsymbol{\Omega}_n^{-1})\boldsymbol{g}_n(\boldsymbol{\theta})$ convereges to zero uniformly in $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, then $\hat{\boldsymbol{\theta}}_{o,n}$ is also consistent estimator by Lemma A.6 in Lee (2007).

We will show that the uniform convergence of $\frac{1}{n}\boldsymbol{g}'_n(\boldsymbol{\theta})(\hat{\boldsymbol{\Omega}}_n^{-1} - \boldsymbol{\Omega}_n^{-1})\boldsymbol{g}_n(\boldsymbol{\theta})$. Let $\|\cdot\|$ be the Euclidean norm for vectors and matrices. By the submultiplicativity of the matrix norm,

$$\left\|\frac{1}{n}\boldsymbol{g}'_n(\boldsymbol{\theta})(\hat{\boldsymbol{\Omega}}_n^{-1} - \boldsymbol{\Omega}_n^{-1})\boldsymbol{g}_n(\boldsymbol{\theta})\right\| \le \left(\frac{1}{n}\|\boldsymbol{g}_n(\boldsymbol{\theta})\|\right)^2\left\|\left(\frac{1}{n}\hat{\boldsymbol{\Omega}}_n\right)^{-1} - \left(\frac{1}{n}\boldsymbol{\Omega}_n\right)^{-1}\right\|.$$

By the assumption, $\left\|\left(\frac{1}{n}\hat{\boldsymbol{\Omega}}_n\right)^{-1} - \left(\frac{1}{n}\boldsymbol{\Omega}_n\right)^{-1}\right\| = o_p(1)$.

The nonstochastic function $\frac{1}{n}E[\boldsymbol{g}_n(\boldsymbol{\theta})]$ is given by,

$$\frac{1}{n}E[\boldsymbol{g}_n(\boldsymbol{\theta})] =\frac{1}{n}\boldsymbol{d}_n'(\boldsymbol{\theta})\boldsymbol{P}_{n,j}\boldsymbol{d}_n(\boldsymbol{\theta}) + \frac{1}{n}\sum_{j=1}^{m} tr(\boldsymbol{\Sigma}_n\boldsymbol{P}_{n,j}) + (\lambda_0-\lambda)\frac{1}{n}\sum_{j=1}^{m} tr(\boldsymbol{\Sigma}_n\boldsymbol{G}_n'\boldsymbol{P}_{n,j}^s)$$
$$+ (\lambda_0-\lambda)^2\frac{1}{n}\sum_{j=1}^{m} tr(\boldsymbol{\Sigma}_n\boldsymbol{G}_n'\boldsymbol{P}_{n,j}\boldsymbol{G}_n) + \frac{1}{n}\boldsymbol{Q}_n'\boldsymbol{d}_n(\boldsymbol{\theta}).$$

As the parameter space is bounded, $\frac{1}{n}E[\boldsymbol{g}_n(\boldsymbol{\theta})] = O(1)$ uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Thus, $\frac{1}{n}\boldsymbol{g}_n(\boldsymbol{\theta}) = O_p(1)$ uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ because $\frac{1}{n}\boldsymbol{g}_n(\boldsymbol{\theta}) - \frac{1}{n}E[\boldsymbol{g}_n(\boldsymbol{\theta})] = o_p(1)$ uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ form the proof of Theorem 1. Consequently, $\frac{1}{n}\boldsymbol{g}_n'(\boldsymbol{\theta})(\hat{\boldsymbol{\Omega}}_n^{-1} - \boldsymbol{\Omega}_n^{-1})\boldsymbol{g}_n(\boldsymbol{\theta})$ convereges to zero uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and the consistency of the feasible optimal GMM estimator $\hat{\boldsymbol{\theta}}_{o,n}$ is shown.

Next, let us consider the asymptotic normality of $\hat{\boldsymbol{\theta}}_{o,n}$. By applying the Taylor's expansion to $\frac{\partial \boldsymbol{g}_n'(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\hat{\boldsymbol{\Omega}}_n^{-1}\boldsymbol{g}_n(\hat{\boldsymbol{\theta}}_n)$, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\left[\frac{1}{n}\frac{\partial \boldsymbol{g}_n'(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\left(\frac{1}{n}\hat{\boldsymbol{\Omega}}_n\right)^{-1}\frac{1}{n}\frac{\partial \boldsymbol{g}_n(\bar{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\right]^{-1}\frac{1}{n}\frac{\partial \boldsymbol{g}_n'(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\left(\frac{1}{n}\hat{\boldsymbol{\Omega}}_n\right)^{-1}\frac{1}{\sqrt{n}}\boldsymbol{g}_n(\boldsymbol{\theta}_0),$$

where each element in $\bar{\boldsymbol{\theta}}_n$ is between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$. From the same argument in the proof of Theorem 1, the asymptotic normality of the feasible optimal GMM estimator $\hat{\boldsymbol{\theta}}_{o,n}$ is shown.

Finally, $\left(\frac{1}{n}\hat{\boldsymbol{D}}_n'\hat{\boldsymbol{\Omega}}_n^{-1}\hat{\boldsymbol{D}}_n\right)^{-1} - \left(\frac{1}{n}\boldsymbol{D}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{D}_n\right)^{-1} = o_p(1)$ by by the continuous mapping theorem.