Available online at www.sciencedirect.com
ScienceDirect
Journal of Econometrics 137 (2007) 489–514
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

GMM and 2SLS estimation of mixed regressive, spatial autoregressive models

Lung-fei Lee

Department of Economics, 410 Arps Hall, Ohio State University, 1945 N. High Street, Columbus, OH 43210-1172, USA

Received 31 October 2001; accepted 6 October 2005
Available online 30 May 2006

Abstract

The GMM method and the classical 2SLS method are considered for the estimation of mixed regressive, spatial autoregressive models. These methods have computational advantage over the conventional maximum likelihood method. The proposed GMM estimators are shown to be consistent and asymptotically normal. Within certain classes of GMM estimators, best ones are derived. The proposed GMM estimators improve upon the 2SLS estimators and are applicable even if all regressors are irrelevant. A best GMM estimator may have the same limiting distribution as the ML estimator (with normal disturbances).

JEL classification: C13; C21; R15

Keywords: Spatial econometrics; Spatial autoregressive models; Regression; 2SLS; GMM; ML; Efficiency

1. Introduction

In this paper, we propose a general GMM framework for the estimation of mixed regressive, spatial autoregressive (MRSAR) models. The GMM estimation for those models can be computationally simpler than the maximum likelihood (ML) or quasi maximum likelihood (QML) methods in a general setting. The GMM estimator (GMME) may be asymptotically more efficient than the two-stage least squares (2SLS) estimator (2SLSE) and may be asymptotically efficient as the ML estimator (MLE).

E-mail address: lflee@econ.ohio-state.edu.

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2005.10.004

The 2SLS method has been noted for the estimation of the MRSAR model in Anselin (1988, 1990), Land and Deane (1992), Kelejian and Robinson (1993), Kelejian and Prucha (1997, 1998), and Lee (2003), among others. The instrumental variables (IVs) are generated from exogenous regressors and the spatial weights matrix of the model. However, there are some issues on the 2SLS approach. The main issue is that the proposed 2SLSE is inefficient relative to the MLE. For the estimation of a conventional linear simultaneous equation model, the 2SLSE is asymptotically efficient as the limited information MLE (see, e.g., Amemiya, 1985). This is not so for the estimation of the MRSAR model. The 2SLSE has been shown to be consistent and asymptotically normally distributed (Kelejian and Prucha, 1998). Lee (2003) discusses the best one (B2SLSE) within the class of IV estimators. By comparing the limiting variance matrices, the 2SLSE and B2SLSE are less efficient relative to the MLE (when the disturbances are normally distributed). A subtle issue is that the 2SLS method would not be consistent when all the exogenous regressors in the MRSAR model are really irrelevant. Furthermore, it is not possible to test the joint significance of all the exogenous regressors based on those IV estimators (Kelejian and Prucha, 1998). On the contrary, the MLE of a MRSAR is consistent and its limiting distribution can be used for testing the joint significance of regression coefficients. These show that the 2SLS approach is less satisfactory than the ML approach in some of its statistical properties. However, the 2SLS approach is computationally simpler than the ML approach and is distribution free.

For pure spatial autoregressive (SAR) processes, a method of moments (MOM) has been introduced in Kelejian and Prucha (2001). The MOM method is computationally simpler than the ML method. Their MOM estimator is consistent but is unlikely to be efficient relative to the MLE.[1] Recently, Lee (2001) extends the MOM estimation into a more general GMM estimation framework. Within that GMM framework, a best GMM estimator (BGMME) can have the same limiting distribution as the MLE or QML estimator.

In this paper, we consider the possible generalization of the MOM method for the MRSAR model. We suggest a combination of the moments in the 2SLS framework with moment functions originated for the estimation of pure SAR processes. We show that the resulting GMME can be asymptotically efficient relative to the 2SLSE and B2SLSE. The best GMME can be made available and it can be efficient as the MLE. With this GMM framework, one may also test the joint significance of all the possible exogenous regressors.

This paper is organized as follows. In Section 2, we consider the estimation of a MRSAR model. We discuss the moment functions that can be used in addition to the moments based on the orthogonality of exogenous regressors with the model disturbance. Consistency and asymptotic distribution of the GMME will be derived in Section 3. In Section 4, the GMME and the 2SLSE are compared. The best selection of moment functions and IVs will be discussed and its possible efficiency property is derived. All the proofs of the results are collected in the appendices. Section 5 provides some Monte Carlo results for the comparison of finite sample properties of estimators. Conclusions are drawn in Section 6.

[1]The asymptotic distribution of their MOM estimator has not been established in their article. They have provided Monte Carlo results to demonstrate that their MOM estimator is only slightly inefficient relative to the QML estimator under various distributions.

2. GMM estimation and identification of the MRSAR model

The MRSAR model differs from a pure SAR process in the presence of exogenous regressors X_n as explanatory variables in the model:

$$Y_n = \lambda W_n Y_n + X_n\beta + \varepsilon_n, \tag{2.1}$$

where X_n is a $n \times k$ dimensional matrix of nonstochastic exogenous variables, W_n is a spatial weights matrix of known constants with a zero diagonal, and the disturbances ε_{ni}, $i = 1, \ldots, n$, of the n-dimensional vector ε_n are i.i.d. $(0, \sigma^2)$. Specifically, we assume that

Assumption 1. The ε_{ni} are i.i.d. with zero mean, variance σ^2 and that a moment of order higher than the fourth exists.

Assumption 2. The elements of X_n are uniformly bounded constants, X_n has the full rank k, and $\lim_{n\to\infty}(1/n)X_n'X_n$ exists and is nonsingular.

Because statistics involving quadratic forms of ε_n will be present in the estimation, the existence of the fourth order moment of ε_{ni}'s will guarantee finite variances for the quadratic forms. The higher than the fourth moment condition in Assumption 1 is needed in order to apply a central limit theorem due to Kelejian and Prucha (2001). The nonstochastic X_n and its uniform boundedness conditions in Assumption 2 are for convenience. If the elements of X_n are stochastic and have unbounded ranges, conditions in Assumption 2 can be replaced by some finite moment conditions.

The W_nY_n in (2.1) is called a spatial lag and its coefficient is supposed to represent the spatial effect due to the influence of neighboring units on a single spatial unit. The main interest in estimation of the model is, in general, the parameters λ and β. In order to distinguish the true parameters from other possible values in the parameter space, we denote λ_0, β_0, and σ_0^2 as the true parameters which generate an observed sample. Let $\theta = (\lambda, \beta')'$ and $\theta_0 = (\lambda_0, \beta_0')'$. This model is supposed to be an equilibrium model. The structural equation (2.1) implies the reduced form equation that

$$Y_n = (I_n - \lambda_0 W_n)^{-1} X_n\beta_0 + (I_n - \lambda_0 W_n)^{-1}\varepsilon_n. \tag{2.2}$$

It follows that $W_nY_n = W_n(I_n - \lambda_0 W_n)^{-1}X_n\beta_0 + W_n(I_n - \lambda_0 W_n)^{-1}\varepsilon_n$ and W_nY_n is correlated with ε_n because, in general, $\mathrm{E}((W_n(I_n - \lambda_0 W_n)^{-1}\varepsilon_n)'\varepsilon_n) = \sigma_0^2 \operatorname{tr}(W_n(I_n - \lambda_0 W_n)^{-1}) \neq 0$. There are some regularity conditions on W_n and $(I_n - \lambda_0 W_n)^{-1}$ which will be needed in order that the spatial correlations between units can be manageable. The following Assumption 3 is originated in the works of Kelejian and Prucha, e.g., Kelejian and Prucha (1998). A sequence of square matrices $\{A_n\}$, where $A_n = [a_{n,ij}]$, is said to be uniformly bounded in row sums (column sums) in absolute value if the sequence of row sum matrix norm $\|A_n\|_\infty = \max_{i=1,\ldots,n}\sum_{j=1}^n |a_{n,ij}|$ (column sum matrix norm $\|A_n\|_1 = \max_{j=1,\ldots,n}\sum_{i=1}^n |a_{n,ij}|$) are bounded.[2]

Assumption 3. The spatial weights matrices $\{W_n\}$ and $\{(I_n - \lambda W_n)^{-1}\}$ at $\lambda = \lambda_0$ are uniformly bounded in both row and column sums in absolute value.

Note that we have imposed the uniform boundedness condition on $\{(I_n - \lambda W_n)^{-1}\}$ only at $\lambda = \lambda_0$. The stronger assumption that $\{(I_n - \lambda W_n)^{-1}\}$ is uniformly bounded in both row

[2] Properties of those matrix norms can be found in Horn and Johnson (1985, pp. 294–295).

and column sums in absolute value, uniformly in λ (in a compact parameter space of λ) is not imposed.[3]

With the normal distribution for ε_n, the unknown parameters λ, β and σ^2 can be estimated by the ML (or QML) method (Ord, 1975). The ML method involves the computation of the determinant $|(I_n - \lambda W_n)|$ of $(I_n - \lambda W_n)$, at each possible value of λ during an optimization search. For the case that W_n is row-normalized and the corresponding spatial matrix before row normalization is symmetric, the eigenvalues of W_n are all real. As $|(I_n - \lambda W_n)|$ is solely a function of λ and eigenvalues of W_n, Ord (1975) points out that computation of $|(I_n - \lambda W_n)|$ can be easily updated as the eigenvalues of W_n need to be computed only once. [4] With a general spatial weights matrix which does not have special properties, such as sparseness and the symmetry property in Ord (1975), the MLE would be difficult to be computed for sample with a large size. Recently, Smirnov and Anselin (2001) discuss the attractability of using a characteristic polynomial approach.

For a computational point of view, a 2SLS method remains the simplest. Kelejian and Prucha (1998) suggest the use of W_nX_n, $W_n^2X_n$, etc., together with X_n as IVs in a 2SLS for estimating θ. Lee (2003) shows that the B2SLS corresponds to use $W_n(I_n - \lambda_0 W_n)^{-1}X_n$ and X_n as IV matrices. These 2SLSE and B2SLSE are computationally simple and have closed form expressions. However, the 2SLS and B2SL methods have the limitation in that at least one nonconstant regressors in X_n must have significant coefficients in order that valid IVs can be generated from them. As the 2SLSE and B2SLSE are based on the existence of relevant nonconstant regressors, it is impossible to test the joint significance of X_n with those estimators.[5]

Even if valid IVs do exist, the 2SLSE may be inefficient relative to the MLE. By comparing the limiting variance matrices of the MLE and 2SLSE (e.g., in Anselin, 1988; Anselin and Bera, 1998), neither the 2SLSE nor the B2SLE have the same limiting distribution of the MLE. In this paper, we suggest to incorporate some other moment conditions in addition to those based on X_n in order to improve upon the efficiency of the 2SLSE.

Let Q_n be an $n \times k_x$ matrix of IVs constructed as functions of X_n and W_n in a 2SLS approach. Denote $\varepsilon_n(\theta) = (I_n - \lambda W_n)Y_n - X_n\beta$ for any possible value θ. The moment functions corresponding to the orthogonality conditions of Q_n and ε_n are $Q_n'\varepsilon_n(\theta)$. Let $\mathscr{P}_{1n}$ be the class of constant $n \times n$ matrices which have a zero trace. A subclass $\mathscr{P}_{2n}$ of $\mathscr{P}_{1n}$ consisting of matrices with a zero diagonal is also interesting. By selecting matrices $P_{1n}, \ldots, P_{mn}$ from $\mathscr{P}_{1n}$, we suggest the use of $(P_{jn}\varepsilon_n(\theta))'\varepsilon_n(\theta)$ in addition to $Q_n'\varepsilon_n(\theta)$ to form a set of moment functions. For analytical tractability, the matrices in $\mathscr{P}_{1n}$ are assumed to have the uniformly boundedness properties as W_n.

Assumption 4. The matrices P_{jn}'s from $\mathscr{P}_{1n}$ are uniformly bounded in both row and column sums in absolute value, and elements of Q_n are uniformly bounded.

[3]The latter stronger assumption is needed for the ML approach. For the GMM method that we propose, because the GMM function is a polynomial function of θ, which is relatively simpler function, our analysis does not require the stronger uniform boundedness assumption.

[4]However, Kelejian and Prucha (1999) have pointed out that Ord's method may suffer from numerically imprecise problems for large sample.

[5]When $\beta_0 = 0$, the model would be a pure spatial autoregressive process. For a spatial autoregressive model with only a nonzero intercept term but no other spatially varying regressors, W_nl_n will not be an useful instrument for W_nY_n when W_n is row normalized. When W_n is row-normalized, $W_nl_n = l_n$ where l_n is the vector of ones.

With the selected matrices P_{jn}'s and IV matrices Q_n, the set of moment functions forms a vector

$$g_n(\theta) = (P_{1n}\varepsilon_n(\theta), \ldots, P_{mn}\varepsilon_n(\theta), Q_n)'\varepsilon_n(\theta) = (\varepsilon_n'(\theta)P_{1n}\varepsilon_n(\theta), \ldots, \varepsilon_n'(\theta)P_{mn}\varepsilon_n(\theta), \varepsilon_n'(\theta)Q_n)' \tag{2.3}$$

for the GMM estimation. At θ_0, $g_n(\theta_0) = (\varepsilon_n'P_{1n}\varepsilon_n, \ldots, \varepsilon_n'P_{mn}\varepsilon_n, \varepsilon_n'Q_n)'$, which has a zero mean because $\mathrm{E}(Q_n'\varepsilon_n) = Q_n'\mathrm{E}(\varepsilon_n) = 0$ and $\mathrm{E}(\varepsilon_n'P_{jn}\varepsilon_n) = \sigma_0^2\,\mathrm{tr}(P_{jn}) = 0$ for $j = 1, \ldots, m$.[6] The intuition is as follows. The IV variables in Q_n, can be used as IV variables for W_nY_n and X_n in (2.1). The $P_{jn}\varepsilon_n$ is uncorrelated with ε_n. As $W_nY_n = W_n(I_n - \lambda_0W_n)^{-1}(X_n\beta_0 + \varepsilon_n)$, $P_{jn}\varepsilon_n$ can be used as an IV for W_nY_n as long as $P_{jn}\varepsilon_n$ and $W_n(I_n - \lambda_0W_n)^{-1}\varepsilon_n$ are correlated.

For any possible value θ,

$$\mathrm{E}(g_n(\theta)) = \begin{pmatrix} d_n'(\theta)P_{1n}d_n(\theta) + \sigma_0^2\,\mathrm{tr}((I_n - \lambda_0W_n)'^{-1}(I_n - \lambda W_n')P_{1n}(I_n - \lambda W_n)(I_n - \lambda_0W_n)^{-1}) \\ \vdots \\ d_n'(\theta)P_{mn}d_n(\theta) + \sigma_0^2\,\mathrm{tr}((I_n - \lambda_0W_n)'^{-1}(I_n - \lambda W_n')P_{mn}(I_n - \lambda W_n)(I_n - \lambda_0W_n)^{-1}) \\ Q_n'd_n(\theta) \end{pmatrix}, \tag{2.4}$$

where $d_n(\theta) = X_n(\beta_0 - \beta) + (\lambda_0 - \lambda)W_n(I_n - \lambda_0W_n)^{-1}X_n\beta_0$. For these moment functions to be useful, they have to identify the true parameter θ_0 of the model. In the GMM framework, the identification condition requires the unique solution of the limiting equations, $\lim_{n\to\infty}(1/n)\mathrm{E}(g_n(\theta)) = 0$ at θ_0 (Hansen, 1982). The moment equations corresponding to Q_n are $\lim_{n\to\infty}(1/n)Q_n'd_n(\theta) = \lim_{n\to\infty}(1/n)\,(Q_n'X_n, Q_n'W_n(I_n - \lambda_0W_n)^{-1} X_n\beta_0)((\beta_0 - \beta)', \lambda_0 - \lambda)' = 0$. They will have a unique solution at θ_0 if $(Q_n'X_n, Q_n'W_n(I_n - \lambda_0W_n)^{-1}X_n\beta_0)$ has a full column rank, i.e., rank $(k+1)$, for large enough n. This sufficient rank condition implies the necessary rank condition that $(X_n, W_n(I_n - \lambda_0W_n)^{-1}X_n\beta_0)$ has a full column rank $(k+1)$ and that Q_n has a rank at least $(k+1)$, for large enough n. The sufficient rank condition requires Q_n to be correlated with W_nY_n in the limit as n goes to infinity. This is so because $\mathrm{E}(Q_n'W_nY_n) = Q_n'W_n(I_n - \lambda_0W_n)^{-1}X_n\beta_0$. Under the sufficient rank condition, θ_0 can thus be identified via $\lim_{n\to\infty}(1/n)Q_n'd_n(\theta) = 0$.

The necessary full rank condition of $(X_n, W_n(I_n - \lambda_0W_n)^{-1}X_n\beta_0)$ for large n is possible only if the set consisting of $W_n(I_n - \lambda_0W_n)^{-1}X_n\beta_0$ and X_n is not asymptotically linearly dependent. This rank condition would not hold, in particular, if β_0 were zero. There are other cases when this dependence can occur (see, e.g., Kelejian and Prucha, 1998). As X_n has rank k, if $(X_n, W_n(I_n - \lambda_0W_n)^{-1}X_n\beta_0)$ does not have a full rank $(k+1)$, its rank will be k, and there will exist a constant vector c such that $W_n(I_n - \lambda_0W_n)^{-1}X_n\beta_0 = X_nc$. Then, $d_n(\theta) = X_n(\beta_0 - \beta + (\lambda_0 - \lambda)c)$ and $Q_n'd_n(\theta) = Q_n'X_n(\beta_0 - \beta + (\lambda_0 - \lambda)c)$. The corresponding moment equations $Q_n'd_n(\theta) = 0$ will have many solutions but the solutions are all described by the relation that $\beta = \beta_0 + (\lambda_0 - \lambda)c$ as long as $Q_n'X_n$ has a full rank k.

[6]The selection of the number of m is not an important issue for our GMM approach, because there exists a best P_n as discussed in a subsequent section. Furthermore, from our Monte Carlo study, the selection of W_n and $(W_n^2 - (\mathrm{tr}(W_n^2)/n)I_n)$ provides accurate approximations for the best one.

Under this scenario, β_0 can be identified only if λ_0 is identifiable. The identification of λ_0 will rely on the remaining quadratic moment equations $\lim_{n\to\infty}(1/n)\,\mathrm{tr}((I_n - \lambda_0 W_n)'^{-1}(I_n - \lambda W_n')P_{jn}(I_n - \lambda W_n)(I_n - \lambda_0 W_n)^{-1}) = 0$ for $j = 1, \ldots, m$. In this case, $Y_n = \lambda_0(W_n(I_n - \lambda_0 W_n)^{-1}X_n\beta_0) + X_n\beta_0 + (I_n - \lambda_0 W_n)^{-1}\varepsilon_n = X_n(\beta_0 + \lambda_0 c) + u_n$, where $u_n = (I_n - \lambda_0 W_n)^{-1}\varepsilon_n$. The relationship $u_n = \lambda_0 W_n u_n + \varepsilon_n$ is a SAR process. The identification of λ_0 thus comes from the SAR process u_n. One can see from Lee (2001) that the set of the limiting quadratic moment equations has a unique solution at λ_0 if $\lim_{n\to\infty}(1/n)[\mathrm{tr}(P_{1n}^s W_n(I_n - \lambda_0 W_n)^{-1}), \ldots, \mathrm{tr}(P_{mn}^s W_n(I_n - \lambda_0 W_n)^{-1})]'$ is linearly independent of $\lim_{n\to\infty}(1/n)[\mathrm{tr}((I_n - \lambda_0 W_n')^{-1} W_n' P_{1n} W_n(I_n - \lambda_0 W_n)^{-1}), \ \ldots, \mathrm{tr}((I_n - \lambda_0 W_n')^{-1} \ W_n' P_{mn} W_n(I_n - \lambda_0 W_n)^{-1})]'$, where $A^s = A + A'$ for any square matrix A. The following assumption summarizes some sufficient conditions for the identification of θ_0 from the moment $\lim_{n\to\infty}(1/n)\mathrm{E}(g_n(\theta)) = 0$.

Assumption 5. Either (i) $\lim_{n\to\infty}(1/n)Q_n'(W_n(I_n - \lambda_0 W_n)^{-1}X_n\beta_0, X_n)$ has the full rank $(k+1)$, or (ii) $\lim_{n\to\infty}(1/n)Q_n'X_n$ has the full rank k, $\lim_{n\to\infty}(1/n)\,\mathrm{tr}(P_{jn}^s W_n(I_n - \lambda_0 W_n)^{-1}) \neq 0$ for some j, and $\lim_{n\to\infty}(1/n)[\mathrm{tr}(P_{1n}^s W_n(I_n - \lambda_0 W_n)^{-1}), \ldots, \mathrm{tr}(P_{mn}^s W_n(I_n - \lambda_0 W_n)^{-1})]'$ is linearly independent of

$$\lim_{n\to\infty}(1/n)[\mathrm{tr}((I_n - \lambda_0 W_n')^{-1} W_n' P_{1n} W_n(I_n - \lambda_0 W_n)^{-1}),$$
$$\ldots, \mathrm{tr}((I_n - \lambda_0 W_n')^{-1} W_n' P_{mn} W_n(I_n - \lambda_0 W_n)^{-1})]'.$$

In terms of computation of the GMM estimator, because the moment functions in $g_n(\theta)$ are quadratic functions in λ and β, the GMM objective function will be of polynomial of order four. The derivation of the GMM estimator will involve the minimization of a polynomial function in θ. The computation of polynomial coefficients, which do not involve the unknown θ, need to be done once. The evaluation of the corresponding objective function will thus involve the multiplication of these polynomial coefficients with powers of θ's at different values of θ. The computation is more complicated than that of the 2SLS but shall be simpler than that of the ML approach.

The variance matrix of these moment functions involves variances and covariances of linear and quadratic forms of ε_n. For any square $n \times n$ matrix A, let $vec_D(A) = (a_{11}, \ldots, a_{nn})'$ denote the column vector formed with the diagonal elements of A. Then, $\mathrm{E}(Q_n'\varepsilon_n \cdot \varepsilon_n' P_n \varepsilon_n) = Q_n' \sum_{i=1}^n \sum_{j=1}^n p_{n,ij}\mathrm{E}(\varepsilon_n \varepsilon_{ni}\varepsilon_{nj}) = Q_n'\, vec_D(P_n)\mu_3$ and $\mathrm{E}(\varepsilon_n' P_{jn}\varepsilon_n \cdot \varepsilon_n' P_{ln}\varepsilon_n) = (\mu_4 - 3\sigma_0^4)\, vec_D'(P_{jn})\, vec_D(P_{ln}) + \sigma_0^4\, \mathrm{tr}(P_{jn}P_{ln}^s)$ by Lemma A.2, where $\mu_3 = \mathrm{E}(\varepsilon_{ni}^3)$ and $\mu_4 = \mathrm{E}(\varepsilon_{ni}^4)$. It follows that $var(g_n(\theta_0)) = \Omega_n$ where

$$\Omega_n = \begin{pmatrix} (\mu_4 - 3\sigma_0^4)\omega_{nm}'\omega_{nm} & \mu_3 \omega_{nm}' Q_n \\ \mu_3 Q_n' \omega_{nm} & 0 \end{pmatrix} + V_n, \tag{2.5}$$

with $\omega_{nm} = [vec_D(P_{1n}), \ldots, vec_D(P_{mn})]$ and

$$V_n = \sigma_0^4 \begin{pmatrix} \mathrm{tr}(P_{1n}P_{1n}^s) & \cdots & \mathrm{tr}(P_{1n}P_{mn}^s) & 0 \\ \vdots & & \vdots & \vdots \\ \mathrm{tr}(P_{mn}P_{1n}^s) & \cdots & \mathrm{tr}(P_{mn}P_{mn}^s) & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_0^2} Q_n' Q_n \end{pmatrix} = \sigma_0^4 \begin{pmatrix} \Delta_{mn} & 0 \\ 0 & \frac{1}{\sigma_0^2} Q_n' Q_n \end{pmatrix}, \tag{2.6}$$

where $\Delta_{mn} = [vec(P'_{1n}), \ldots, vec(P'_{mn})]'[vec(P^s_{1n}), \ldots, vec(P^s_{mn})]$, by using $\mathrm{tr}(AB) = vec(A')' vec(B)$ for any conformable matrices A and B. When ε_n is normally distributed, Ω_n is simplified to V_n because $\mu_3 = 0$ and $\mu_4 = 3\sigma_0^4$. If P_{jn}'s are from $\mathscr{P}_{2n}$, $\Omega_n = V_n$ also because $\omega_{nm} = 0$. In general, from (2.3), Ω_n is nonsingular if and only if both $(vec(P_{1n}), \ldots, vec(P_{mn}))$ and Q_n have full column ranks. This is so, because Ω_n would be singular if and only if the moments in $g_n(\theta_0)$ are functionally dependent, equivalently, if and only if $\sum_{j=1}^{m} a_j P_{jn} = 0$ and $Q_n b = 0$ for some constant vector $(a_1, \ldots, a_m, b') \neq 0$. As elements of P_{jn}'s and Q_n are uniformly bounded by Assumption 4, it is apparent that $(1/n)\omega'_{nm}Q_n$, $(1/n)\omega'_{nm}\omega_{nm}$ and $(1/n)Q'_nQ_n$ are of order O(1). Furthermore, because $P_{jn}P^s_{ln}$ is bounded in row or column sums in absolute value, $(1/n)\,\mathrm{tr}(P_{jn}P^s_{ln}) = \mathrm{O}(1)$. Consequently, $(1/n)\Omega_n = \mathrm{O}(1)$. It is thus meaningful to impose the following conventional regularity condition on the limit of $(1/n)\Omega_n$.

Assumption 6. The limit of $(1/n)\Omega_n$ exists and is a nonsingular matrix.[7]

The variance matrix Ω_n is needed to formulate the optimum GMME with $g_n(\theta)$.

3. Consistency and asymptotic distributions

The following proposition provides the asymptotic distribution of the GMME with a linear transformation of the moment equations, $a_n g_n(\theta)$, where a_n is a matrix with a full row rank greater than or equal to $(k+1)$. The a_n is assumed to converge to a constant full rank matrix a_0. This corresponds to the Hansen's GMM setting, which illustrates the optimal weighting issue.

Assumption 7. θ_0 is in the interior of the parameter space Θ, which is a compact subset of R^{k+1}.[8]

Proposition 1. *Under Assumptions* 1–5, *suppose that* P_{jn} *for* $j = 1, \ldots, m$, *are from* $\mathscr{P}_{1n}$ *and* Q_n *is a* $n \times k_x$ *IV matrix so that* $a_0 \lim_{n\to\infty}(1/n)\mathrm{E}(g_n(\theta)) = 0$ *has a unique root at* θ_0 *in* Θ. *Then, the GMME* $\hat{\theta}_n$ *derived from* $\min_{\theta\in\Theta} g'_n(\theta)a'_na_ng_n(\theta)$ *is a consistent estimator of* θ_0, *and* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, \Sigma)$, *where*

$$\Sigma = \lim_{n\to\infty} \left[\left(\frac{1}{n}D'_n\right)a'_na_n\left(\frac{1}{n}D_n\right)\right]^{-1}\left(\frac{1}{n}D'_n\right)a'_na_n\left(\frac{1}{n}\Omega_n\right)a'_na_n\left(\frac{1}{n}D_n\right) \times \left[\left(\frac{1}{n}D'_n\right)a'_na_n\left(\frac{1}{n}D_n\right)\right]^{-1} \tag{3.1}$$

[7]In this paper, for simplicity, we do not consider the large group interactions scenario as in Case (1991), where $\lim_{n\to\infty}(1/n)\Omega_n$ might be singular. It is possible to extend our analysis to cover that case but the analysis would become much algebraically complicated. For an analysis on the MLE, see Lee (2004).

[8]For a nonlinear extremum estimator, the parameter space would usually be assumed to be a compact set (Amemiya, 1985). The extremum estimate always exists when the objective function is continuous on a compact set. Furthermore, for the proof of consistency of the extremum estimator, the uniform convergence argument will usually require a compact parameter space (see the uniform convergence theorem in Amemiya). For our GMM approach, the moment functions are quadratic functions of θ. Because of its simple nonlinear structure, the uniform convergence argument of the sample objective function after proper normalization can be easily established as long as the parameter space is bounded. So the compact parameter space assumption may be relaxed to a bounded set as long as the minimum of the objective function exists in such a parameter space.

and

$$D_n = \begin{pmatrix} \sigma_0^2 \operatorname{tr}(P_{1n}^s W_n(I_n - \lambda_0 W_n)^{-1}) & \cdots & \sigma_0^2 \operatorname{tr}(P_{mn}^s W_n(I_n - \lambda_0 W_n)^{-1}) & (W_n(I_n - \lambda_0 W_n)^{-1} X_n \beta_0)' Q_n \\ 0 & \cdots & 0 & X_n' Q_n \end{pmatrix}', \tag{3.2}$$

under the assumption that $\lim_{n\to\infty}(1/n)a_n D_n$ *exists and has the full rank* $(k+1)$.

The rank conditions in Assumption 5 imply that $\lim_{n\to\infty}(1/n)g_n(\theta_0) = 0$ has a unique root at θ_0 and, hence, its corresponding gradient matrix $\lim_{n\to\infty}(1/n)D_n$ of (2.8) has rank $(k+1)$. In the presence of Assumption 5, the extra conditions that $\lim_{n\to\infty}(1/n)a_n \mathrm{E}(g_n(\theta)) = 0$ has a unique root at θ_0 and the limit of $(1/n)a_n D_n$ has full rank $(k+1)$, are simply to eliminate the bad choice of a sequence $\{a_n\}$ for the linear combination $a_n g_n(\theta)$, which may result in the loss of identification from the original g_n. If Assumption 5 were not satisfied, the condition that $a_0 \lim_{n\to\infty}(1/n)g_n(\theta) = 0$ has a unique root at θ_0 would not be satisfied.[9] The identification condition is specified for the limiting function. The weaker requirement that $a_n \mathrm{E}(g_n(\theta)) = 0$ has a unique root at θ_0 for large enough n is not sufficient for θ_0 to be identifiably unique, because the objective function might flatten out in the limit.

From Proposition 1, with the moment functions $g_n(\theta)$ in (2.3), the optimal choice of a weighting matrix $a_n' a_n$ is $((1/n)\Omega_n)^{-1}$ by the generalized Schwartz inequality. If P_{jn}'s are selected from the subclass $\mathscr{P}_{2n}$ or ε_{ni}'s are normally distributed, Ω_n in (2.5) will be reduced to the simpler matrix V_n in (2.6). These variance matrices can be used to form the optimal GMM objective function with $g_n(\theta)$. The σ^2, μ_3, and μ_4 can be consistently estimated by using estimated residuals of ε_n from an initial consistent estimate of θ_0.[10] The Ω_n can then be consistently estimated as $\hat{\Omega}_n$. The following proposition shows that the feasible optimum GMME (OGMME) with a consistently estimated $\hat{\Omega}_n$ has the same limiting distribution as that of the OGMME based on Ω_n. With the optimum GMM objective function, an overidentification test is available.

Proposition 2. *Under Assumptions* 1–6, *suppose that* $(\hat{\Omega}_n/n)^{-1} - (\Omega_n/n)^{-1} = o_P(1)$, *then the feasible OGMME* $\hat{\theta}_{o,n}$ *derived from* $\min_{\theta\in\Theta} g_n'(\theta)\hat{\Omega}_n^{-1} g_n(\theta)$ *based on* $g_n(\theta)$ *in* (2.3) *with* P_{jn}*'s from* $\mathscr{P}_{1n}$ *has the asymptotic distribution*

$$\sqrt{n}(\hat{\theta}_{o,n} - \theta_0) \xrightarrow{D} N\left(0, \left(\lim_{n\to\infty}(1/n)D_n'\Omega_n^{-1}D_n\right)^{-1}\right). \tag{3.3}$$

Furthermore,

$$g_n'(\hat{\theta}_n)\hat{\Omega}_n^{-1} g_n(\hat{\theta}_n) \xrightarrow{D} \chi^2((m+k_x) - (k+1)). \tag{3.4}$$

---

[9]When the set of $W_n(I_n - \lambda_0 W_n)^{-1} X_n \beta_0$ and X_n were linearly dependent, Proposition 1 would not cover the large group interaction case of Case (1991) because, in this situation, $(1/n)\operatorname{tr}(P_{jn}^s W_n(I_n - \lambda_0 W_n)^{-1})$ would vanish in the limit and Assumptions 5 and 6 needed be strengthened. The MLE $\hat{\lambda}_n$ of λ_0 in this situation is known to have a slower rate of convergence than that of the MLE $\hat{\beta}_n$ of β_0 (see, Lee, 2004).

[10]The detailed proof is straightforward but tedious and is omitted here.

The 2SLSE $\hat{\beta}_{2sl,n}$ would be inconsistent when the set of $W_n(I_n - \lambda_0 W_n)^{-1} X_n \beta_0$ and X_n is linearly dependent. An obvious example is $\beta_0 = 0$. Another example is a model where the only relevant variable in X_n is the intercept term l_n and W_n is row-normalized (Kelejian and Prucha, 1998). In that case, $W_n(I_n - \lambda_0 W_n)^{-1} X_n \beta_0 = (\beta_{01}/(1-\lambda_0))l_n$ because $(I_n - \lambda_0 W_n)^{-1} l_n = (1/(1-\lambda_0))l_n$ and $W_n l_n = l_n$. However, even when the set of $W_n(I_n - \lambda_0 W_n)^{-1} X_n \beta_0$ and X_n is linearly dependent, the GMM approach may still work because of the additional moment functions with P_{jn}'s. The asymptotic distribution of the GMME in (2.9) can be used to formulate a Wald statistic for testing the overall significance of all exogenous variables while that of the 2SLSE cannot.

4. Efficiency and the BGMME

The optimal GMME $\hat{\theta}_{o,n}$ can be compared with the 2SLSE. With Q_n as the IV matrix, the 2SLSE of θ_0 is

$$\hat{\theta}_{2sl,n} = [Z_n' Q_n (Q_n' Q_n)^{-1} Q_n' Z_n]^{-1} Z_n' Q_n (Q_n' Q_n)^{-1} Q_n' Y_n, \tag{4.1}$$

where $Z_n = (W_n Y_n, X_n)$. The asymptotic distribution of $\hat{\theta}_{2sl,n}$ is

$$\sqrt{n}(\hat{\theta}_{2sl,n} - \theta_0) \xrightarrow{D} N\Bigg(0, \sigma_0^2 \lim_{n\to\infty} \bigg\{ \frac{1}{n} (W_n(I_n - \lambda_0 W_n)^{-1} X_n \beta_0, X_n)' \\ \times Q_n (Q_n' Q_n)^{-1} Q_n' (W_n(I_n - \lambda_0 W_n)^{-1} X_n \beta_0, X_n) \bigg\}^{-1} \Bigg), \tag{4.2}$$

under the assumptions that $\lim_{n\to\infty}(1/n)Q_n' Q_n$ is nonsingular and $\lim_{n\to\infty}(1/n)Q_n(W_n(I_n - \lambda_0 W_n)^{-1} X_n \beta_0, X_n)$ has the full column rank $(k+1)$ (Kelejian and Prucha, 1998). Because the 2SLSE can be derived from $\min_\theta \varepsilon_n'(\theta) Q_n' (Q_n Q_n')^{-1} Q_n \varepsilon_n(\theta)$, the 2SLS approach is a special case of the GMM estimation in Proposition 1 with $a_n = (0, (Q_n' Q_n/n)^{-1/2})$ and $(1/n)a_n g_n(\theta) = (Q_n' Q_n/n)^{-1/2}(1/n)Q_n' \varepsilon_n(\theta)$. It follows from Proposition 2, $\hat{\theta}_{o,n}$ shall be efficient relative to $\hat{\theta}_{2sl,n}$.

Within the 2SLS framework, by the generalized Schwartz inequality applied to the asymptotic variance of $\hat{\theta}_{2ls,n}$ in (4.2), the best IV matrix Q_n will be $(W_n(I_n - \lambda_0 W_n)^{-1} X_n \beta_0, X_n)$. Using the best IV matrix for Q_n in the GMM framework, the resulting GMME shall be efficient relative to the B2SLE. There is a related question on whether $(W_n(I_n - \lambda_0 W_n)^{-1} X_n \beta_0, X_n)$ can be the best IV matrix in the class of matrices Q_n with given P_{jn}'s. The answer can be affirmative for cases where the moments $Q_n' \varepsilon_n$ do not interact with the moments $\varepsilon_n' P_{jn} \varepsilon_n$ via their correlations. The covariance of $Q_n' \varepsilon_n$ and $\varepsilon_n' P_{jn} \varepsilon_n$, $j = 1, \ldots, m$, is $\mu_3 Q_n' \omega_{nm}$, which can be zero when $\mu_3 = 0$ or $\omega_{nm} = 0$.

The remaining issue is on the best selection of P_{jn}'s. When the disturbance ε_n is normally distributed or P_{jn}'s are from $\mathscr{P}_{2n}$,

$$D_n' \Omega_n^{-1} D_n = \begin{pmatrix} C_{mn} \Delta_{mn}^{-1} C_{mn}' & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{\sigma_0^2} (W_n(I_n - \lambda_0 W_n)^{-1} X_n \beta_0, X_n)' \\ \times (W_n(I_n - \lambda_0 W_n)^{-1} X_n \beta_0, X_n), \tag{4.3}$$

where $C_{mn} = [\mathrm{tr}(P^s_{1n} W_n(I_n - \lambda_0 W_n)^{-1}), \ldots, \mathrm{tr}(P^s_{mn} W_n(I_n - \lambda_0 W_n)^{-1})]$. Note that, because $\mathrm{tr}(P_{jn}P^s_{ln}) = \frac{1}{2}\mathrm{tr}(P^s_{jn}P^s_{ln})$, Δ_{mn} can be rewritten as

$$\begin{aligned}\Delta_{mn} &= \frac{1}{2}\begin{pmatrix} \mathrm{tr}(P^s_{1n}P^s_{1n}) & \cdots & \mathrm{tr}(P^s_{1n}P^s_{mn}) \\ \vdots & & \vdots \\ \mathrm{tr}(P^s_{mn}P^s_{1n}) & \cdots & \mathrm{tr}(P^s_{mn}P^s_{mn}) \end{pmatrix} \\ &= \frac{1}{2}[vec(P^s_{1n}) \cdots vec(P^s_{mn})]'[vec(P^s_{1n}) \cdots vec(P^s_{mn})].\end{aligned}$$

(i) When P_{jn}'s are from $\mathscr{P}_{2n}$,

$$\begin{aligned}\mathrm{tr}(P^s_{jn} W_n(I_n - \lambda_0 W_n)^{-1}) &= \frac{1}{2}\,\mathrm{tr}(P^s_{jn}[W_n(I_n - \lambda_0 W_n)^{-1} - Diag(W_n(I_n - \lambda_0 W_n)^{-1})]^s) \\ &= \frac{1}{2}\,vec'([W_n(I_n - \lambda_0 W_n)^{-1} - Diag(W_n(I_n - \lambda_0 W_n)^{-1})]^s) \\ &\quad \times vec(P^s_{jn})\end{aligned}$$

for $j = 1, \ldots, m$, in C_{mn}, where $Diag(A)$ denotes the diagonal matrix formed by the diagonal elements of a square matrix A. Therefore, the generalized Schwartz inequality implies that

$$\begin{aligned}C_{mn}\Delta^{-1}_{mn}C'_{mn} &\leqslant \frac{1}{2}\,vec'([W_n(I_n - \lambda_0 W_n)^{-1} - Diag(W_n(I_n - \lambda_0 W_n)^{-1})]^s) \\ &\quad \times vec([W_n(I_n - \lambda_0 W_n)^{-1} - Diag(W_n(I_n - \lambda_0 W_n)^{-1})]^s) \\ &= \mathrm{tr}([W_n(I_n - \lambda_0 W_n)^{-1} - Diag(W_n(I_n - \lambda_0 W_n)^{-1})]^s W_n(I_n - \lambda_0 W_n)^{-1}).\end{aligned}$$

Thus, in the subclass $\mathscr{P}_{2n}$, $[W_n(I_n - \lambda_0 W_n)^{-1} - Diag(W_n(I_n - \lambda_0 W_n)^{-1})]$ and together with $[W_n(I_n - \lambda_0 W_n)^{-1}X_n\beta_0, X_n]$ provide the set of best IV functions.[11]

(ii) For the case where ε_n is $N(0, \sigma_0^2 I_n)$, because, for any $P_{jn} \in \mathscr{P}_{1n}$,

$$\begin{aligned}&\mathrm{tr}(P^s_{jn} W_n(I_n - \lambda_0 W_n)^{-1}) \\ &\quad = \frac{1}{2}\,vec'\left(\left[W_n(I_n - \lambda_0 W_n)^{-1} - \frac{\mathrm{tr}(W_n(I_n - \lambda_0 W_n)^{-1})}{n}\,I_n\right]^s\right) vec(P^s_{jn}),\end{aligned}$$

for $j = 1, \ldots, m$, the generalized Schwartz inequality implies that

$$C_{mn}\Delta^{-1}_{mn}C'_{mn} \leqslant \mathrm{tr}\left(\left[W_n(I_n - \lambda_0 W_n)^{-1} - \frac{\mathrm{tr}(W_n(I_n - \lambda_0 W_n)^{-1})}{n} I_n\right]^s W_n(I_n - \lambda_0 W_n)^{-1}\right).$$

[11]Note that the best selected matrix for the quadratic moment is a single matrix which is best relative to any finite number of P_{jn}. So any additional P_{jn} in addition to the best one will not play a role in (asymptotically) efficient estimation. This is so also for the best IV matrix Q for linear moments.

Hence, in the broader class $\mathscr{P}_{1n}$, $[W_n(I_n-\lambda_0 W_n)^{-1}-(\mathrm{tr}(W_n(I_n-\lambda_0 W_n)^{-1})/n)I_n]$ and $[W_n(I_n-\lambda_0 W_n)^{-1}X_n\beta_0, X_n]$ provide the best set of IV functions. For all those cases, in the event that the set of $W_n(I_n-\lambda_0 W_n)^{-1}X_n\beta_0$ and X_n is linearly dependent, $W_n(I_n-\lambda_0 W_n)^{-1}X_n\beta_0$ is redundant and the best IV matrix shall simply be X_n.[12]

In practice, with initial consistent estimates $\hat{\lambda}_n$, $\hat{\beta}_n$ of λ_0 and β_0, $W_n(I_n-\lambda_0 W_n)^{-1}$ can be estimated as $W_n(I_n-\hat{\lambda}_n W_n)^{-1}$, and $W_n(I_n-\lambda_0 W_n)^{-1}X_n\beta_0$ by $\hat{W}_n(I_n-\hat{\lambda}_n W_n)^{-1}X_n\hat{\beta}_n$. The corresponding variance matrix V_n of these best moment functions can be estimated as $\hat{V}_n$. The following proposition summarizes the results and shows that the feasible BGMME has the same limiting distribution as the BGMME.

Proposition 3. *Under Assumptions* 1–3, *suppose that* $\hat{\lambda}_n$ *is a* $\sqrt{n}$*-consistent estimate of* λ_0, $\hat{\beta}_n$ *is a consistent estimate of* β_0, *and* $\hat{\sigma}_n^2$ *is a consistent estimate of* σ_0^2.

Within the class of GMMEs derived with $\mathscr{P}_{2n}$, *the BGMME* $\hat{\theta}_{2b,n}$ *has the limiting distribution that* $\sqrt{n}(\hat{\theta}_{2b,n}-\theta_0)\xrightarrow{D} N(0,\Sigma_{2b}^{-1})$ *where*

$$\Sigma_{2b} = \lim_{n\to\infty}\frac{1}{n}\begin{pmatrix} \mathrm{tr}[(G_n-Diag(G_n))^s G_n]+\frac{1}{\sigma_0^2}(G_nX_n\beta_0)'(G_nX_n\beta_0) & \frac{1}{\sigma_0^2}(G_nX_n\beta_0)'X_n \\ \frac{1}{\sigma_0^2}X_n'(G_nX_n\beta_0) & \frac{1}{\sigma_0^2}X_n'X_n \end{pmatrix}, \quad (4.4)$$

with $G_n = W_n(I_n-\lambda_0 W_n)^{-1}$, *which is assumed to exist.*

In the event that $\varepsilon_n \sim N(0,\sigma_0^2 I_n)$, *within the broader class of GMME's derived with* $\mathscr{P}_{1n}$, *the BGMME* $\hat{\theta}_{1b,n}$ *has the limiting distribution that* $\sqrt{n}(\hat{\theta}_{1b,n}-\theta_0)\xrightarrow{D} N(0,\Sigma_{1b}^{-1})$ *where*

$$\Sigma_{1b} = \lim_{n\to\infty}\frac{1}{n}\begin{pmatrix} \mathrm{tr}[(G_n-\frac{\mathrm{tr}(G_n)}{n}I_n)^s G_n]+\frac{1}{\sigma_0^2}(G_nX_n\beta_0)'(G_nX_n\beta_0) & \frac{1}{\sigma_0^2}(G_nX_n\beta_0)'X_n \\ \frac{1}{\sigma_0^2}X_n'(G_nX_n\beta_0) & \frac{1}{\sigma_0^2}X_n'X_n \end{pmatrix}, \quad (4.5)$$

which is assumed to exist.

When ε_n is $N(0,\sigma_0^2 I_n)$, model (2.1) can be estimated by the ML method. The log likelihood function of the MRSAR model via its reduced form equation in (2.2) is

$$\ln L_n = -\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\sigma^2+\ln|(I_n-\lambda W_n)|-\frac{1}{2\sigma^2}[Y_n-(I_n-\lambda W_n)^{-1}X_n\beta]' \times(I_n-\lambda W_n')(I_n-\lambda W_n)[Y_n-(I_n-\lambda W_n)^{-1}X_n\beta]. \quad (4.6)$$

[12]The best IV vector $W_n(I_n-\lambda_0 W_n)^{-1}X_n\beta_0$ together with X_n will satisfy the identification Assumption 5(i) as long as they are not linearly dependent in the limit. In the event that they are, the model identification will follow from the best matrix $W_n(I_n-\lambda_0 W_n)^{-1}-(1/n)\,\mathrm{tr}(W_n(I_n-\lambda_0 W_n)^{-1})I_n$ or $W_n(I_n-\lambda_0 W_n)^{-1}-Diag(W_n(I_n-\lambda_0 W_n)^{-1})$ for the quadratic moment (Lee, 2001). With those best IV vectors and matrices, its variance matrix will be nonsingular and Assumption 6 will also be satisfied.

The asymptotic variance of the MLE $(\hat{\theta}_{ml,n}, \hat{\sigma}^2_{ml,n})$ is

$$\mathrm{AsyVar}(\hat{\theta}_{ml,n}, \hat{\sigma}^2_{ml,n}) = \begin{pmatrix} \mathrm{tr}(G_n^2) + \mathrm{tr}(G_n' G_n) + \frac{1}{\sigma_0^2}(G_n X_n \beta_0)'(G_n X_n \beta_0) & \frac{1}{\sigma_0^2}(X_n' G_n X_n \beta_0)' & \frac{\mathrm{tr}(G_n)}{\sigma_0^2} \\ \frac{1}{\sigma_0^2} X_n' G_n X_n \beta_0 & \frac{1}{\sigma_0^2} X_n' X_n & 0 \\ \frac{\mathrm{tr}(G_n)}{\sigma_0^2} & 0 & \frac{n}{2\sigma_0^4} \end{pmatrix}^{-1}$$

(see, e.g., Anselin and Bera, 1998, p. 256). From the inverse of a partitioned matrix, the asymptotic variance of the MLE $\hat{\theta}_{ml,n}$ is

$$\mathrm{AsyVar}(\hat{\theta}_{ml,n}) = \begin{pmatrix} \mathrm{tr}(G_n^2) + \mathrm{tr}(G_n' G_n) + \frac{1}{\sigma_0^2}(G_n X_n \beta_0)'(G_n X_n \beta_0) - \frac{2}{n}\mathrm{tr}^2(G_n) & \frac{1}{\sigma_0^2}(X_n' G_n X_n \beta_0)' \\ \frac{1}{\sigma_0^2} X_n' G_n X_n \beta_0 & \frac{1}{\sigma_0^2} X_n' X_n \end{pmatrix}^{-1}. \tag{4.7}$$

As $\mathrm{tr}(G_n^2) + \mathrm{tr}(G_n' G_n) - (2/n)\,\mathrm{tr}^2(G_n) = \mathrm{tr}((G_n - (\mathrm{tr}(G_n)/n) I_n)^s G_n)$, the GMME $\hat{\theta}_{1b,n}$ has the same limiting distribution as the MLE of θ_0 from Proposition 3.

There is an intuition on the best GMM approach compared with the ML one. The derivatives of the log likelihood in (4.6) are

$$\frac{\partial \ln L_n}{\partial \beta} = \frac{1}{\sigma^2} X_n' \varepsilon_n(\theta),$$

$$\frac{\partial \ln L_n}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\varepsilon_n'(\theta)\varepsilon_n(\theta),$$

and

$$\frac{\partial \ln L_n}{\partial \lambda} = -\mathrm{tr}(W_n(I_n - \lambda W_n)^{-1}) + \frac{1}{\sigma^2}[W_n(I_n - \lambda W_n)^{-1} X_n \beta]' \varepsilon_n(\theta) + \frac{1}{\sigma^2}\varepsilon_n'(\theta)[W_n(I_n - \lambda W_n)^{-1}]'\varepsilon_n(\theta).$$

The equation $\partial \ln L_n / \partial \sigma^2 = 0$ implies that the MLE is $\hat{\sigma}_n^2(\theta) = (1/n)\varepsilon_n'(\theta)\varepsilon_n(\theta)$ for a given value θ. By substituting $\hat{\sigma}_n^2(\theta)$ into the remaining likelihood equations, the MLE $\hat{\theta}_{ml,n}$ will be characterized by the moment equations: $X_n'\varepsilon_n(\theta) = 0$, and

$$[W_n(I_n - \lambda W_n)^{-1} X_n \beta]' \varepsilon_n(\theta) + \varepsilon_n'(\theta)\left[W_n(I_n - \lambda W_n)^{-1} - \frac{1}{n}\mathrm{tr}(W_n(I_n - \lambda W_n)^{-1})\right]\varepsilon_n(\theta) = 0.$$

The similarity of the best GMM moments and the above likelihood equations is revealing. The best GMM approach has the linear and quadratic moments of $\varepsilon_n(\theta)$ in its formation and uses consistently estimated matrices in its linear and quadratic forms.

5. Some Monte Carlo results

The model in the Monte Carlo study is specified as

$$Y_n = \lambda W_n Y_n + X_{n1}\beta_1 + X_{n2}\beta_2 + X_{n3}\beta_3 + \varepsilon_n,$$

where x_{i1}, x_{i2} and x_{i3} are three independently generated standard normal variables and are i.i.d. for all i, and ε_{ni}'s are i.i.d. $N(0, \sigma^2)$. When the sample size is $n = 49$, the spatial weights matrix W_n corresponds to the weights matrix for the study of crimes across 49 districts in Columbus, Ohio in Anselin (1988). For large sample sizes of $n = 245$ and 490, the corresponding spatial weights matrices are block diagonal matrices with the preceding 49×49 matrix as their diagonal blocks. These correspond to the pooling, respectively, of five and ten separate districts with similar neighboring structures in each district. The estimation methods considered are the

(1) 2SLS—the 2SLS method with IV's X_n, W_nX_n, and $W_n^2X_n$;
(2) GMM—a simple unweighted GMM approach using $Q_n = (X_n, W_nX_n, W_n^2X_n)$ for linear moments and W_n and $W_n^2 - (\mathrm{tr}(W_n^2)/n)I_n$ for quadratic moments (with an identity matrix as the distance matrix);
(3) OGMM—the optimum GMM approach using $Q_n = (X_n, W_nX_n, W_n^2X_n)$ for linear moments and W_n and $W_n^2 - (\mathrm{tr}(W_n^2)/n)I_n$ for quadratic moments (with the inverse of their (estimated) variance matrix as the distance matrix)[13];
(4) BGMM—the best optimum GMM approach by using X_n and $(I_n - \hat{\lambda}_n W_n)^{-1}X_n\hat{\beta}_n$ for the linear moments, and $W_n(I_n - \hat{\lambda}_n W_n)^{-1} - (1/n)\,\mathrm{tr}[W_n(I_n - \hat{\lambda}_n W_n)^{-1}]I_n$ for the quadratic moment, where $(\hat{\lambda}_n, \hat{\beta}_n)$ is an initial consistent estimate;
(5) ML—the ML approach.

The number of repetitions is 1,000 for each case in this Monte Carlo experiment. The regressors are randomly redrawn for each repetition. In each case, we report the mean 'Mean' and standard deviation 'SD' of the empirical distributions of the estimates. To facilitate the comparison of various estimators, their root mean square errors 'RMSE' are also reported. In all the cases of this study, the true λ_0 is set to 0.6. The smallest sample size is $n = 49$, and the moderate sample sizes are 245 and 490. The variance of the equation errors σ_0^2 is 2. The β coefficients are varied in the experiments.

Table 1 reports the results of the case where $\beta_{10} = -1.0$, $\beta_{20} = 0$ and $\beta_{30} = 1.0$. In this case, the corresponding variance ratio of $x\beta_0$ with the sum of variances of $x\beta_0$ and ε is 0.5. If one ignores the interaction term, this ratio would represent $R^2 = 0.5$ in a regression equation. The results indicate that the main differences of the various estimation approaches are on the estimation of the spatial effect λ. For the small sample size $N = 49$, the 2SLS is biased upward by 12.7% and it has also the largest SD compared with the various GMMEs and the MLE. The MLE is biased downward by 4% and the OGMME is biased upward by 6.8%. The GMME and BGMME are essentially unbiased. The BGMME is not better than the GMME and OGMME in terms of SD and RMSE with this small sample. Among the various GMM estimates, the OGMME is better in terms of SD

[13]The normality of the disturbances is assumed. The σ^2 in the variance matrix of the moments is estimated with the estimated residuals of the model equation with the simple GMM estimates as its coefficients. Note that the minimization of an objective function in various GMM approaches is performed globally without imposing a restricted parameter space, such as λ lies in $(-1, 1)$, in our study.

Table 1
2SLSE, GMME, and MLE

Method	λ	β_1	β_2	β_3
	Mean (SD)[RMSE]	Mean (SD)[RMSE]	Mean (SD)[RMSE]	Mean (SD)[RMSE]
$N = 49$				
2SLS	0.676 (0.177)[0.192]	−0.968 (0.218)[0.220]	0.001 (0.204)[0.204]	0.985 (0.213)[0.214]
GMM	0.600 (0.150)[0.150]	−0.982 (0.221)[0.222]	0.001 (0.206)[0.206]	0.998 (0.211)[0.211]
O-GMM	0.641 (0.134)[0.141]	−0.971 (0.221)[0.223]	0.001 (0.208)[0.208]	0.987 (0.212)[0.213]
BGMM	0.593 (0.161)[0.161]	−0.978 (0.219)[0.220]	0.001 (0.207)[0.207]	0.993 (0.209)[0.210]
MLE	0.575 (0.115)[0.118]	−0.988 (0.218)[0.218]	0.000 (0.205)[0.205]	1.002 (0.211)[0.211]
$N = 245$				
2SLS	0.612 (0.078)[0.079]	−0.993 (0.092)[0.093]	−0.002 (0.090)[0.090]	0.991 (0.094)[0.094]
GMM	0.600 (0.053)[0.053]	−0.997 (0.092)[0.092]	−0.002 (0.090)[0.090]	0.996 (0.093)[0.093]
O-GMM	0.606 (0.049)[0.049]	−0.994 (0.091)[0.092]	−0.003 (0.090)[0.090]	0.993 (0.093)[0.094]
BGMM	0.598 (0.048)[0.048]	−0.995 (0.091)[0.092]	−0.002 (0.090)[0.090]	0.994 (0.093)[0.093]
MLE	0.596 (0.047)[0.047]	−0.998 (0.091)[0.091]	−0.002 (0.090)[0.090]	0.997 (0.093)[0.093]
$N = 490$				
2SLS	0.608 (0.056)[0.056]	−0.994 (0.063)[0.064]	−0.001 (0.065)[0.065]	0.996 (0.066)[0.067]
GMM	0.600 (0.037)[0.037]	−0.997 (0.063)[0.063]	−0.001 (0.066)[0.066]	0.999 (0.066)[0.066]
O-GMM	0.604 (0.032)[0.033]	−0.995 (0.063)[0.063]	−0.001 (0.066)[0.066]	0.997 (0.066)[0.066]
BGMM	0.599 (0.032)[0.032]	−0.996 (0.063)[0.063]	−0.001 (0.066)[0.066]	0.998 (0.066)[0.066]
MLE	0.598 (0.032)[0.032]	−0.997 (0.063)[0.063]	−0.001 (0.066)[0.066]	0.999 (0.066)[0.066]

True parameters: $\lambda = 0.6$, $\beta_1 = -1.0$, $\beta_2 = 0$, and $\beta_3 = 1.0$.

and RMSE. The MLE has the smallest SD and RMSE among all these estimates. The estimates of β's of the various methods do not have much differences. The estimates of β's have small biases. When n increases to 245 or 490, the upward bias of the 2SLS estimate of λ is reduced. All the other estimates are unbiased. Even for the moderate sample sizes, the 2SLS estimates of λ have apparently larger SD and RMSEs than the corresponding various GMMEs and MLEs. The BGMME is slightly better than the OGMME, and, in turn, the OGMME is slightly efficient relative to the GMME. The BGMM is efficient as the MLE when $N = 490$.

In Table 2, the true parameters are $\beta_1 = -0.2$, $\beta_2 = 0$, and $\beta_3 = 0.2$. The variance of $x\beta_0$ is much smaller than the variance of ε. If one ignores the interaction term, the implied R^2 is about 0.04 in a regression equation. In this case, the λ may be relatively more difficult to be estimated by the 2SLS. The upward bias of the 2SLSE of λ can be large. The SD and RMSE of the 2SLS estimates are larger than those of the MLE and various GMMEs. The OGMME can be the best among the various GMM estimates when $N = 49$. The BGMME can be better than the other GMMEs with larger N. The BGMME estimates can be as efficient as the MLE with large $N = 490$. For the estimates of the β's, there are not much differences among the various estimates.

In summary, the 2SLSE of λ has larger biases and SDs than those of the various GMMEs.[14] The performance of the 2SLSE becomes worse when the value of R^2 becomes

[14] This may be so because our cases have only moderate or very small R^2 values due to the explanatory variables.

Table 2
2SLSE, GMME, and MLE

Method	λ	β_1	β_2	β_3
	Mean (SD)[RMSE]	Mean (SD)[RMSE]	Mean (SD)[RMSE]	Mean (SD)[RMSE]
$N = 49$				
2SLS	0.906 (0.316)[0.440]	−0.177 (0.215)[0.216]	0.001 (0.209)[0.209]	0.195 (0.213)[0.213]
GMM	0.597 (0.174)[0.174]	−0.191 (0.217)[0.218]	−0.001 (0.205)[0.205]	0.201 (0.213)[0.213]
O-GMM	0.688 (0.216)[0.233]	−0.186 (0.219)[0.219]	0.001 (0.211)[0.211]	0.199 (0.216)[0.216]
BGMM	0.605 (0.193)[0.193]	−0.187 (0.214)[0.214]	−0.002 (0.206)[0.206]	0.198 (0.211)[0.211]
MLE	0.566 (0.142)[0.146]	−0.190 (0.216)[0.216]	0.000 (0.205)[0.205]	0.203 (0.210)[0.210]
$N = 245$				
2SLS	0.795 (0.258)[0.323]	−0.189 (0.092)[0.092]	−0.003 (0.089)[0.089]	0.188 (0.093)[0.094]
GMM	0.600 (0.059)[0.059]	−0.198 (0.091)[0.091]	−0.002 (0.091)[0.091]	0.197 (0.093)[0.093]
O-GMM	0.613 (0.060)[0.061]	−0.197 (0.091)[0.091]	−0.003 (0.090)[0.090]	0.196 (0.093)[0.093]
BGMM	0.600 (0.058)[0.058]	−0.197 (0.090)[0.091]	−0.002 (0.090)[0.090]	0.196 (0.092)[0.092]
MLE	0.596 (0.057)[0.057]	−0.198 (0.091)[0.091]	−0.002 (0.090)[0.090]	0.197 (0.093)[0.093]
$N = 490$				
2SLS	0.747 (0.218)[0.263]	−0.190 (0.063)[0.064]	−0.002 (0.064)[0.064]	0.192 (0.066)[0.067]
GMM	0.600 (0.041)[0.041]	−0.198 (0.062)[0.062]	−0.001 (0.066)[0.066]	0.199 (0.065)[0.065]
O-GMM	0.606 (0.041)[0.041]	−0.197 (0.062)[0.062]	−0.001 (0.065)[0.065]	0.199 (0.065)[0.065]
BGMM	0.600 (0.040)[0.040]	−0.197 (0.062)[0.062]	−0.001 (0.065)[0.065]	0.199 (0.065)[0.065]
MLE	0.597 (0.040)[0.040]	−0.197 (0.062)[0.062]	−0.001 (0.065)[0.065]	0.199 (0.065)[0.065]

True parameters: $\lambda = 0.6$, $\beta_1 = -0.2$, $\beta_2 = 0$, and $\beta_3 = 0.2$.

small. The various GMMEs can substantially improve upon the 2SLSE. The OGMME and BGMME can be efficient as the MLE with large sample sizes.[15] The differences of the various estimators occur only for the estimation of λ but not for estimation of the β's.

6. Conclusion

In this paper, we consider the estimation of the MRSAR model. The 2SLS method has been suggested in the literature for the estimation of the MRSAR model. The 2SLS method can be applicable only if some of the spatially varying exogenous variables are really relevant. It is impossible in the 2SLS framework to test the overall significance of all the exogenous variables in the MRSAR model. It is known that the 2SLSE does not attain the same limiting distribution of the MLE (under normal disturbances) of the MRSAR model. This paper improves upon the 2SLS approach by introducing additional moment functions in the GMM framework. The resulted GMME can be efficient relative to the 2SLSE. It is possible to derive the best GMMEs within certain classes of GMMEs. One of the BGMMEs can attain the same limiting distribution of the MLE (under normal disturbances). Within the GMM estimation framework, it is possible to test the overall significance of all the exogenous variables in the model. The GMM approach may, in principle, be generalized for the estimation of MRSAR models with higher order spatial

[15] However, in some repetitions, the BGMME may be sensitive to initial consistent estimates in the construction of G_n. The results in Tables 1 and 2 are based, respectively, on 2SLSE and GMME as initial estimates.

lags and models with both spatial lags and/or SAR disturbances. However, many issues for the models with higher order moments have not been well understood, for example, the proper parameter space of the lag coefficients and its identification problem. These will be studied in another occasion.

Acknowledgements

I am grateful to the National Science Foundation, Economics Program for financial support for my research under the Grant numbers SES-0111380 and SES-0519204. I thank Mr. Ji Tao, who has worked as my RA, to conduct the Monte Carlo study in this paper. I appreciate having valuable comments and suggestions from Professor Christopher R. Bollinger, an anonymous referee and an associate editor of this journal.

Appendix A. Some useful lemmas

In this appendix, we list some lemmas which are useful for the proofs of the results in the text.

Lemma A.1. *Suppose that the sequences of n-dimensional column vectors* $\{z_{1n}\}$ *and* $\{z_{2n}\}$ *are uniformly bounded. If* $\{A_n\}$ *are uniformly bounded in either row or column sums in absolute value, then* $|z'_{1n}A_n z_{2n}| = \mathrm{O}(n)$.

Proof. This is trivial. □

Lemma A.2. *Suppose that* $\varepsilon_{n1}, \ldots, \varepsilon_{nn}$ *are i.i.d. random variables with zero mean, finite variance* σ^2 *and finite fourth moment* μ_4. *Then, for any two square* $n \times n$ *matrices A and B,*

$$\mathrm{E}(\varepsilon'_n A\varepsilon_n \cdot \varepsilon'_n B\varepsilon_n) = (\mu_4 - 3\sigma_0^4)\, vec'_D(A)\, vec_D(B) + \sigma_0^4[\mathrm{tr}(A)\,\mathrm{tr}(B) + \mathrm{tr}(AB^s)],$$

where $B^s = B + B'$.

Proof. This is a Lemma in Lee (2001). □

Lemma A.3. *Suppose that* $\{A_n\}$ *are uniformly bounded in both row and column sums in absolute value. The* $\varepsilon_{n1}, \ldots, \varepsilon_{nn}$ *are i.i.d. with zero mean and its fourth moment exists. Then,* $\mathrm{E}(\varepsilon'_n A_n \varepsilon_n) = \mathrm{O}(n)$, $\mathrm{var}(\varepsilon'_n A_n \varepsilon_n) = \mathrm{O}(n)$, $\varepsilon'_n A_n \varepsilon_n = \mathrm{O_P}(n)$, *and* $(1/n)\varepsilon'_n A_n \varepsilon_n - (1/n)\mathrm{E}(\varepsilon'_n A_n \varepsilon_n) = \mathrm{o_P}(1)$.

Proof. Lee (2001). □

Lemma A.4. *Suppose that* A_n *is a* $n \times n$ *matrix with its column sums being uniformly bounded in absolute value, elements of the* $n \times k$ *matrix* C_n *are uniformly bounded, and* $\varepsilon_{n1}, \ldots, \varepsilon_{nn}$ *are i.i.d. with zero mean and finite variance* σ^2. *Then,* $(1/\sqrt{n})C'_n A_n \varepsilon_n = \mathrm{O_P}(1)$ *and* $(1/n)C'_n A_n \varepsilon_n = \mathrm{o_p}(1)$. *Furthermore, if the limit of* $(1/n)C'_n A_n A'_n C_n$ *exists and is positive definite, then* $(1/\sqrt{n})C'_n A_n \varepsilon_n \xrightarrow{D} N(0, \sigma^2 \lim_{n\to\infty}(1/n)C'_n A_n A'_n C_n)$.

Proof. See Lee (2004). □

Lemma A.5. *Suppose that* $\{A_n\}$ *is a sequence of symmetric* $n \times n$ *matrices with row and column sums uniformly bounded in absolute value and* $b_n = (b_{n1}, \ldots, b_{nn})'$ *is a n-dimensional* *vector such that* $\sup_n (1/n)\sum_{i=1}^{n} |b_{ni}|^{2+\eta_1} < \infty$ *for some* $\eta_1 > 0$. *The* $\varepsilon_{n1}, \ldots, \varepsilon_{nn}$ *are i.i.d. random variables with zero mean and finite variance* σ^2, *and its moment* $\mathrm{E}(|\varepsilon|^{4+2\delta})$ *for some* $\delta > 0$ *exists. Let* $\sigma^2_{Q_n}$ *be the variance of* Q_n *where* $Q_n = \varepsilon_n' A_n \varepsilon_n + b_n' \varepsilon_n - \sigma^2 \operatorname{tr}(A_n)$. *Assume that the variance* $\sigma^2_{Q_n}$ *is bounded away from zero at the rate n. Then,* $(Q_n/\sigma_{Q_n}) \xrightarrow{\mathrm{D}} N(0,1)$.

Proof. See Kelejian and Prucha (2001). □

Lemma A.6. *Suppose that* $(1/n)(Q_n(\theta) - \bar{Q}_n(\theta))$ *converges in probability to zero uniformly in* $\theta \in \Theta$ *which is a convex set, and* $\{(1/n)\bar{Q}_n(\theta)\}$ *satisfies the identification uniqueness condition at* θ_0. *Let* $\hat{\theta}_n$ *and* $\hat{\theta}_n^*$ *be, respectively, the minimizers of* $Q_n(\theta)$ *and* $Q_n^*(\theta)$ *in* Θ. *If* $(1/n)(Q_n^*(\theta) - Q_n(\theta)) = o_P(1)$ *uniformly in* $\theta \in \Theta$, *then both* $\hat{\theta}_n$ *and* $\hat{\theta}_n^*$ *converge in probability to* θ_0.

In addition, suppose that $(1/n)(\partial^2 Q_n(\theta)/\partial\theta\partial\theta')$ *converges in probability to a well defined limiting matrix, uniformly in* $\theta \in \Theta$, *which is nonsingular at* θ_0, *and* $(1/\sqrt{n})(\partial Q_n(\theta_0)/\partial\theta) = O_P(1)$. *If* $(1/n)(\partial^2 Q_n^*(\theta)/\partial\theta\partial\theta' - \partial^2 Q_n(\theta)/\partial\theta\partial\theta') = o_P(1)$ *uniformly in* $\theta \in \Theta$ *and* $(1/\sqrt{n})(\partial Q_n^*(\theta_0)/\partial\theta - \partial Q_n(\theta_0)/\partial\theta) = o_P(1)$, *then* $\sqrt{n}(\hat{\theta}_n^* - \theta_0)$ *and* $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *have the same limiting distribution.*

Proof. The convergence of $\hat{\theta}_n$ to θ_0 follows from the uniform convergence of $(1/n)(Q_n(\theta) - \bar{Q}_n(\theta))$ to zero in probability and the uniqueness identification condition of $\{\bar{Q}_n(\theta)\}$ (White, 1994). As $(1/n)(Q_n^*(\theta) - \bar{Q}_n(\theta)) = (1/n)(Q_n^*(\theta) - Q_n(\theta)) + (1/n)(Q_n(\theta) - \bar{Q}_n(\theta)) = o_P(1)$ uniformly in $\theta \in \Theta$, the convergence of $\hat{\theta}_n^*$ to θ_0 in probability follows. For the limiting distribution, the Taylor expansion of $\partial Q_n^*(\theta)/\partial\theta$ at θ_0 implies that

$$
\begin{aligned}
\sqrt{n}(\hat{\theta}_n^* - \theta_0) &= -\left(\frac{1}{n}\frac{\partial^2 Q_n^*(\bar{\theta}_n)}{\partial\theta\partial\theta'}\right)^{-1} \frac{1}{\sqrt{n}}\frac{\partial Q_n^*(\theta_0)}{\partial\theta} \\
&= \left(-\frac{1}{n}\frac{\partial^2 Q_n(\bar{\theta}_n)}{\partial\theta\partial\theta'} + o_P(1)\right)^{-1} \left(\frac{1}{\sqrt{n}}\frac{\partial Q_n(\theta_0)}{\partial\theta} + o_P(1)\right) \\
&= \left(-\frac{1}{n}\frac{\partial^2 Q_n(\bar{\theta}_n)}{\partial\theta\partial\theta'}\right)^{-1} \frac{1}{\sqrt{n}}\frac{\partial Q_n(\theta_0)}{\partial\theta} + o_P(1).
\end{aligned}
$$

Thus, $\sqrt{n}(\hat{\theta}_n^* - \theta_0)$ and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ have the same limiting distribution. □

Lemma A.7. *Suppose that the elements of the* $n \times k$ *matrices* X_n *are uniformly bounded for all n; and* $\lim_{n\to\infty}(1/n)X_n'X_n$ *exists and this limiting matrix is nonsingular, then the projectors,* $X_n(X_n'X_n)^{-1}X_n'$ *and* $I_n - X_n(X_n'X_n)^{-1}X_n'$, *are uniformly bounded in both row and column sums in absolute value.*

Proof. See Lee (2004). □

Lemma A.8. *Suppose that* $\{\|W_n\|\}$ *and* $\{\|S_n^{-1}(\lambda_0)\|\}$, *where* $\|\cdot\|$ *is a matrix norm, are bounded. Then* $\{\|S_n(\lambda)^{-1}\|\}$, *where* $S_n(\lambda) = I_n - \lambda W_n$, *are uniformly bounded in a neighborhood of* λ_0.

Proof. See Lee (2004). □

In the following Lemmas and Appendix B, some simplified notations shall be used to minimize the presentation of mathematical terms. For any square $n \times n$ matrix A, we shall denote the adjusted matrix $(A - ((\operatorname{tr}(A)/n)I_n)$ or $(A - Diag(A))$ by A^d. Furthermore, with the spatial weights matrix W_n, denote $S_n(\lambda) = I_n - \lambda W_n$, $S_n = S_n(\lambda_0)$, $G_n(\lambda) = W_n(I_n - \lambda W_n)^{-1}$, and $G_n = G_n(\lambda_0)$.

Lemma A.9. *Suppose that z_{1n} and z_{2n} are n-dimensional column vectors of constants which elements are uniformly bounded, the $n \times n$ constant matrices A_n are uniformly bounded in the maximum column sum norm, and ε_{ni}'s in $\varepsilon_n = (\varepsilon_{n1}, \ldots, \varepsilon_{nn})$, are i.i.d. with zero mean and a finite variance σ^2. Let $\hat{\lambda}_n$ be an $\sqrt{n}$-consistent estimate of λ_0. Then, under Assumption* 3,

(1) $(1/n)z'_{1n}(G_n(\hat{\lambda}_n) - G_n)' z_{2n} = o_P(1)$, $(1/n)z'_{1n}(G_n(\hat{\lambda}_n) - G_n)^{d'} z_{2n} = o_P(1)$; *and*
(2) $(1/\sqrt{n})z'_{1n}(G_n(\hat{\lambda}_n) - G_n)' A_n\varepsilon_n = o_P(1)$, $(1/\sqrt{n})z'_{1n}(G_n(\hat{\lambda}_n) - G_n)^{d'} A_n\varepsilon_n = o_P(1)$.

Proof. As $S_n - S_n(\hat{\lambda}_n) = (\hat{\lambda}_n - \lambda_0)W_n$, it follows that $G_n(\hat{\lambda}_n) - G_n = W_n[S_n^{-1}(\hat{\lambda}_n) - S_n^{-1}] = W_n S_n^{-1}(\hat{\lambda}_n)[S_n - S_n(\hat{\lambda}_n)]S_n^{-1} = (\hat{\lambda}_n - \lambda_0)G_n(\hat{\lambda}_n)G_n$, and $(G_n(\hat{\lambda}_n) - G_n)^d = (\hat{\lambda}_n - \lambda_0)(G_n(\hat{\lambda}_n) G_n)^d$. A further expansion implies that $G_n(\hat{\lambda}_n) - G_n = (\hat{\lambda}_n - \lambda_0)G_n^2 + (\hat{\lambda}_n - \lambda_0)^2 G_n(\hat{\lambda}_n)G_n^2$ and $(G_n(\hat{\lambda}_n) - G_n)^d = (\hat{\lambda}_n - \lambda_0)G_n^{2d} + (\hat{\lambda}_n - \lambda_0)^2(G_n(\hat{\lambda}_n)G_n^2)^d$. We note that, under Assumption 3, because S_n^{-1} is uniformly bounded in both row and column sums in absolute value, $S_n^{-1}(\lambda)$ and, hence, $G_n(\lambda)$ must be uniformly bounded in both row and column sums in absolute value uniformly in λ in a small neighborhood of λ_0 by Lemma A.8. As $\hat{\lambda}_n$ is consistent, it follow that $G_n(\hat{\lambda}_n)$ is uniformly bounded in both row and column sums in absolute value with probability one. Therefore, Lemma A.1 implies that $(1/n)z'_{1n}G'_n G'_n(\hat{\lambda}_n)z_{2n} = O_P(1)$. Hence, $(1/n)z'_{1n}(G_n(\hat{\lambda}_n) - G_n)' z_{2n} = (\hat{\lambda}_n - \lambda_0)(1/n)z'_{1n}G'_n G'_n(\hat{\lambda}_n)z_{2n} = o_P(1)$ as $\hat{\lambda}_n - \lambda_0 = o_P(1)$. Similarly, $(1/n)z'_{1n}(G_n(\hat{\lambda}_n) - G_n)^{d'} z_{2n} = o_P(1)$. This proves (1).

For (2), by the further expansion of $G_n(\hat{\lambda}_n)$ around G_n,

$$\frac{1}{\sqrt{n}} z'_{1n}(G_n(\hat{\lambda}_n) - G_n)' A_n\varepsilon_n = \sqrt{n}(\hat{\lambda}_n - \lambda_0)\frac{1}{n} z'_{1n}G'^2_n A_n\varepsilon_n + (\hat{\lambda}_n - \lambda_0)^2 \times \frac{1}{\sqrt{n}} z'_{1n}G'^2_n G'_n(\hat{\lambda}_n)A_n\varepsilon_n.$$

Lemma A.4 implies that $(1/n)z'_{1n}G'^2_n A_n\varepsilon_n = o_P(1)$. Therefore, with the $\sqrt{n}$-consistent $\hat{\lambda}_n$, the first term on the right-hand side is $o_P(1)$. The remainder term is also $o_P(1)$. This is so as follows. Let $\|\cdot\|_1$ be the maximum column sum norm. Because the product of matrices uniformly bounded in column sums in absolute value is uniformly bounded in column sums in absolute value, $\|G'_n G'_n(\hat{\lambda}_n)A_n\|_1 \leqslant c_1$ for some constant c_1 for all n. As elements of z_{1n} are uniformly bounded, there exists a constant c_2 such that $\|z'_{1n}\|_1 \leqslant c_2$. It follows that

$$\left\|(\hat{\lambda}_n - \lambda_0)^2 \frac{1}{\sqrt{n}} z'_{1n}G'^2_n G'_n(\hat{\lambda}_n)A_n\varepsilon_n\right\|_1 \leqslant \left[\sqrt{n}(\hat{\lambda}_n - \lambda_0)\right]^2 \frac{1}{n^{3/2}} \|\varepsilon_n\|_1 \cdot \|z'_{1n}\|_1 \cdot \|G'^2_n G'_n(\hat{\lambda}_n)A_n\|_1 \leqslant \frac{c_1 c_2}{n^{\frac{1}{2}}} \left[\sqrt{n}(\hat{\lambda}_n - \lambda_0)\right]^2 \frac{1}{n}\sum_{i=1}^{n} |\varepsilon_{ni}|.$$

Result (2) follows because $\sqrt{n}(\lambda_n - \lambda_0) = o_P(1)$ and $(1/n)\sum_{i=1}^n |\varepsilon_{ni}| = O_P(1)$ by the strong law of large numbers. Similarly arguments are applicable to $(1/\sqrt{n})z'_{1n} (G_n(\hat{\lambda}_n) - G_n)^{d'} A_n\varepsilon_n$. $\square$

Lemma A.10. *Let A_n and B_n be $n \times n$ matrices, uniformly bounded in both row and column sums in absolute value. The ε_{ni}'s in $\varepsilon_n = (\varepsilon_{n1}, \ldots, \varepsilon_{nn})$ are i.i.d. with zero mean and its fourth moment exists. Suppose that $\hat{\lambda}_n$ is a $\sqrt{n}$-consistent estimator of λ_0. Then, under Assumption* 3,

(i) $(1/n)\varepsilon_n' A_n'(G_n(\hat{\lambda}_n) - G_n)^d B_n \varepsilon_n = o_P(1)$, *and*
(ii) $(1/\sqrt{n})\varepsilon_n'(G_n(\hat{\lambda}_n) - G_n)^d \varepsilon_n = o_P(1)$.

Proof. This is a case of Lemma A.9 in Lee (2001). □

Appendix B. Proofs

Proof of Proposition 1. For consistency, we first show that $(1/n)a_n g_n(\theta) - (1/n)a_n \mathrm{E}(g_n(\theta))$ will converge in probability uniformly in $\theta \in \Theta$ to zero. Let $a_n = (a_{n1}, \ldots, a_{nm}, a_{nx})$ where a_{nx} is a (row) subvector. Then $a_n g_n(\theta) = \varepsilon_n'(\theta)(\sum_{j=1}^m a_{nj} P_{jn})\varepsilon_n(\theta) + a_{nx} Q_n' \varepsilon_n(\theta)$. By expansion, $\varepsilon_n(\theta) = d_n(\theta) + \varepsilon_n + (\lambda_0 - \lambda) G_n \varepsilon_n$ where $d_n(\theta) = (\lambda_0 - \lambda) G_n X_n \beta_0 + X_n(\beta_0 - \beta)$. It follows that

$$\varepsilon_n'(\theta)\left(\sum_{j=1}^m a_{nj} P_{jn}\right)\varepsilon_n(\theta) = d_n'(\theta)\left(\sum_{j=1}^m a_{nj} P_{jn}\right) d_n(\theta) + l_n(\theta) + q_n(\theta),$$

where $l_n(\theta) = d_n'(\theta)(\sum_{j=1}^m a_{nj} P_{jn}^s)(\varepsilon_n + (\lambda_0 - \lambda) G_n \varepsilon_n)$ and $q_n(\theta) = (\varepsilon_n' + (\lambda_0 - \lambda)\varepsilon_n' G_n')(\sum_{j=1}^m a_{nj} P_{jn})(\varepsilon_n + (\lambda_0 - \lambda) G_n \varepsilon_n)$. The term $l_n(\theta)$ is linear in ε_n. By expansion,

$$\begin{aligned}\frac{1}{n} l_n(\theta) &= (\lambda_0 - \lambda)\frac{1}{n}(X_n\beta_0)' G_n'\left(\sum_{j=1}^m a_{nj} P_{jn}^s\right)\varepsilon_n + (\beta_0 - \beta)'\frac{1}{n} X_n'\left(\sum_{j=1}^m a_{nj} P_{jn}^s\right)\varepsilon_n \\ &\quad + (\lambda_0 - \lambda)^2 \frac{1}{n}(X_n\beta_0)' G_n'\left(\sum_{j=1}^m a_{nj} P_{jn}^s\right) G_n \varepsilon_n \\ &\quad + (\lambda_0 - \lambda)(\beta_0 - \beta)'\frac{1}{n} X_n'\left(\sum_{j=1}^m a_{nj} P_{jn}^s\right) G_n \varepsilon_n = o_P(1),\end{aligned}$$

by Lemma A.4, uniformly in $\theta \in \Theta$. The uniform convergence in probability follows because $l_n(\theta)$ is simply a quadratic function of λ and β and Θ is a bounded set. Similarly,

$$\begin{aligned}\frac{1}{n} q_n(\theta) &= \frac{1}{n}\varepsilon_n'\left(\sum_{j=1}^m a_{nj} P_{jn}\right)\varepsilon_n + (\lambda_0 - \lambda)\frac{1}{n}\varepsilon_n' G_n'\left(\sum_{j=1}^m a_{nj} P_{jn}^s\right)\varepsilon_n \\ &\quad + (\lambda_0 - \lambda)^2 \frac{1}{n}\varepsilon_n' G_n'\left(\sum_{j=1}^m a_{nj} P_{jn}\right) G_n \varepsilon_n \\ &= (\lambda_0 - \lambda)\frac{\sigma_0^2}{n}\sum_{j=1}^m a_{nj}\,\mathrm{tr}(G_n' P_{jn}^s) + (\lambda_0 - \lambda)^2 \frac{\sigma_0^2}{n}\sum_{j=1}^m a_{nj}\,\mathrm{tr}(G_n' P_{jn} G_n) + o_P(1),\end{aligned}$$

uniformly in $\theta \in \Theta$, by Lemma A.3 and $\mathrm{E}(\varepsilon_n' P_{jn} \varepsilon_n) = \sigma_0^2 \operatorname{tr}(P_{jn}) = 0$ for all $j = 1, \ldots, m$. Consequently,

$$\frac{1}{n} \varepsilon_n'(\theta) \left(\sum_{j=1}^m a_{nj} P_{jn} \right) \varepsilon_n(\theta) = \frac{1}{n} d_n'(\theta) \left(\sum_{j=1}^m a_{nj} P_{jn} \right) d_n(\theta) + (\lambda_0 - \lambda) \frac{\sigma_0^2}{n} \sum_{j=1}^m a_{nj} \operatorname{tr}(P_{jn}^s G_n) + (\lambda_0 - \lambda)^2 \frac{\sigma_0^2}{n} \sum_{j=1}^m a_{nj} \operatorname{tr}(G_n' P_{jn} G_n) + o_P(1),$$

uniformly in $\theta \in \Theta$. As $g_n(\theta)$ is a quadratic function of θ and Θ is bounded, $(1/n) a_n \mathrm{E}(g_n(\theta))$ is uniformly equicontinuous on Θ. The identification condition and the uniform equicontinuity of $(1/n) a_n \mathrm{E}(g_n(\theta))$ imply that the identification uniqueness condition for $(1/n^2) \mathrm{E}(g_n'(\theta) a_n' a_n \mathrm{E}(g_n(\theta))$ must be satisfied. The consistency of the GMME $\hat{\theta}_n$ follows from the uniform convergence and the identification uniqueness condition (White, 1994).

For the asymptotic distribution of $\hat{\theta}_n$, by the Taylor expansion of $(\partial g_n'(\hat{\theta}_n)/\partial\theta)\, a_n' a_n g_n(\hat{\theta}_n) = 0$ at θ_0,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\left[\frac{1}{n} \frac{\partial g_n'(\hat{\theta}_n)}{\partial\theta} a_n' a_n \frac{1}{n} \frac{\partial g_n(\bar{\theta}_n)}{\partial\theta'} \right]^{-1} \frac{1}{n} \frac{\partial g_n'(\hat{\theta}_n)}{\partial\theta} a_n' \frac{1}{\sqrt{n}} a_n g_n(\theta_0).$$

As $\partial\varepsilon_n(\theta)/\partial\theta' = -(W_n Y_n, X_n)$, it follows that $\partial g_n(\theta)/\partial\theta' = -(P_{1n}^s \varepsilon_n(\theta), \ldots, P_{mn}^s \varepsilon_n(\theta), Q_n)'(W_n Y_n, X_n)$. Explicitly, $(1/n)\varepsilon_n'(\theta) P_{jn}^s W_n Y_n = (1/n)\, \varepsilon_n'(\theta) P_{jn}^s G_n X_n \beta_0 + (1/n)\, \varepsilon_n'(\theta) P_{jn}^s G_n \varepsilon_n$. By Lemmas A.4 and A.3,

$$\begin{aligned} \frac{1}{n} \varepsilon_n'(\theta) P_{jn}^s G_n X_n \beta_0 &= \frac{1}{n} d_n'(\theta) P_{jn}^s G_n X_n \beta_0 + \frac{1}{n} \varepsilon_n' P_{jn}^s G_n X_n \beta_0 + (\lambda_0 - \lambda) \frac{1}{n} \varepsilon_n' G_n' P_{jn}^s G_n X_n \beta_0 \\ &= \frac{1}{n} d_n'(\theta) P_{jn}^s G_n X_n \beta_0 + o_P(1), \end{aligned}$$

and

$$\begin{aligned} \frac{1}{n} \varepsilon_n'(\theta) P_{jn}^s G_n \varepsilon_n &= \frac{1}{n} d_n'(\theta) P_{jn}^s G_n \varepsilon_n + \frac{1}{n} \varepsilon_n' P_{jn}^s G_n \varepsilon_n + \frac{1}{n} (\lambda_0 - \lambda) \varepsilon_n' G_n' P_{jn}^s G_n \varepsilon_n \\ &= \frac{\sigma_0^2}{n} \operatorname{tr}(P_{jn}^s G_n) + (\lambda_0 - \lambda) \frac{\sigma_0^2}{n} \operatorname{tr}(G_n' P_{jn}^s G_n) + o_P(1), \end{aligned}$$

uniformly in $\theta \in \Theta$. Hence,

$$\frac{1}{n} \varepsilon_n'(\theta) P_{jn}^s W_n Y_n = \frac{1}{n} d_n'(\theta) P_{jn}^s G_n X_n \beta_0 + \frac{\sigma_0^s}{n} \operatorname{tr}(P_{jn}^s G_n) + (\lambda_0 - \lambda) \frac{\sigma_0^2}{n} \operatorname{tr}(G_n' P_{jn}^s G_n) + o_P(1),$$

uniformly in $\theta \in \Theta$. At θ_0, $d_n(\theta_0) = 0$ and, hence, $(1/n)\varepsilon_n'(\theta_0) P_{jn}^s W_n Y_n = (\sigma_0^2/n) \operatorname{tr}(P_{jn}^s G_n) + o_P(1)$. At θ_0, $(1/n)\varepsilon_n'(\theta_0) P_{jn}^s X_n = o_P(1)$. Finally, $(1/n) Q_n' W_n Y_n = (1/n) Q_n' G_n X_n \beta_0 + (1/n)\, Q_n' G_n \varepsilon_n = (1/n) Q_n' G_n X_n \beta_0 + o_P(1)$. In conclusion, $(1/n)(\partial g_n(\tilde{\theta}_n)/\, \partial\theta) = -(1/n) D_n + o_P(1)$ with D_n in (3.2). On the other hand, Lemma A.5 implies that

$$\frac{1}{\sqrt{n}} a_n g_n(\theta_0) = \frac{1}{\sqrt{n}} \left[\varepsilon_n' \left(\sum_{j=1}^m a_{nj} P_{jn} \right) \varepsilon_n + a_{nx} Q_n' \varepsilon_n \right] \xrightarrow{D} N\left(0, \lim_{n\to\infty} \frac{1}{n} a_n \Omega_n a_n' \right).$$

The asymptotic distribution of $\sqrt{n}(\hat{\lambda}_n - \lambda_0)$ follows. $\square$

Proof of Proposition 2. The generalized Schwartz inequality implies that the optimal weighting matrix for $a_n'a_n$ in Proposition 1 is $((1/n)\Omega_n)^{-1}$. For consistency, consider

$$\frac{1}{n}\, g_n'(\theta)\hat{\Omega}_n^{-1} g_n(\theta) = \frac{1}{n}\, g_n'(\theta)\Omega_n^{-1} g_n(\theta) + \frac{1}{n}\, g_n'(\theta)(\hat{\Omega}_n^{-1} - \Omega_n^{-1}) g_n(\theta).$$

With $a_n = ((1/n)\Omega_n)^{-1/2}$ in Proposition 1, Assumption 6 implies that $a_0 = (\lim_{n\to\infty} (1/n)\Omega_n)^{-1/2}$ exits. Because a_0 is nonsingular, the identification condition of θ_0 corresponds to the unique root of $\lim_{n\to\infty} \mathrm{E}(1/n)\, g_n(\theta)) = 0$ at θ_0, which is satisfied by Assumption 5. Hence, the uniform convergence in probability of $(1/n)g_n'(\theta)\Omega_n^{-1} g_n(\theta)$ to a well defined limit uniformly in $\theta \in \Theta$ follows by a similar argument in the proof of Proposition 1. So it remains to show that $(1/n)g_n'(\theta)(\hat{\Omega}_n^{-1} - \Omega_n^{-1})g_n(\theta) = o_P(1)$ uniformly in $\theta \in \Theta$. Let $\|\cdot\|$ be the Euclidean norm for vectors and matrices. Then,

$$\left\|\frac{1}{n}\, g_n'(\theta)(\hat{\Omega}_n^{-1} - \Omega_n^{-1}) g_n(\theta)\right\| \leqslant \left(\frac{1}{n}\|g_n(\theta)\|\right)^2 \left\|\left(\frac{\hat{\Omega}_n}{n}\right)^{-1} - \left(\frac{\Omega_n}{n}\right)^{-1}\right\|.$$

To show that this is of probability order $o_P(1)$ uniformly in $\theta \in \Theta$, it is sufficient to show that $(1/n)\|g_n(\theta)\| = O_P(1)$ uniformly in $\theta \in \Theta$. From the proof of Proposition 1, $(1/n)[g_n(\theta) - \mathrm{E}(g_n(\theta))] = o_P(1)$ uniformly in $\theta \in \Theta$. On the other hand, as

$$\begin{aligned}\frac{1}{n}\, d_n'(\theta)P_{jn}d_n(\theta) &= (\lambda_0-\lambda)^2 \frac{1}{n}\,(X_n\beta_0)'G_n'P_{jn}G_nX_n\beta_0 + (\lambda_0-\lambda)\frac{1}{n}\,(X_n\beta_0)'G_n'P_{jn}^sX_n(\beta_0-\beta)\\ &\quad + (\beta_0-\beta)'\frac{1}{n}X_n'P_{jn}X_n(\beta_0-\beta) = O_P(1),\end{aligned}$$

uniformly in $\theta \in \Theta$ by Lemma A.1, it follows that

$$\begin{aligned}&\frac{1}{n}\,\mathrm{E}(\varepsilon_n'(\theta)P_{jn}\varepsilon_n(\theta))\\ &\quad = \frac{1}{n}\, d_n'(\theta)P_{jn}d_n(\theta) + (\lambda_0-\lambda)\sigma_0^2\,\frac{1}{n}\,\mathrm{tr}(P_{jn}^sG_n) + (\lambda_0-\lambda)^2\sigma_0^2\,\frac{1}{n}\,\mathrm{tr}(G_n'P_{jn}G_n) = O(1),\end{aligned}$$

uniformly in $\theta \in \Theta$. Similarly, $(1/n)\mathrm{E}(Q_n'\varepsilon_n(\theta)) = (1/n)Q_n'd_n(\theta) = (\lambda_0-\lambda)(1/n)\, Q_n'G_nX_n\beta_0 + (1/n)\, Q_n'X_n(\beta_0-\beta) = O(1)$ uniformly in $\theta \in \Theta$. These imply that $\|(1/n)\mathrm{E}(g_n(\theta))\| = O(1)$ uniformly in $\theta \in \Theta$. Consequently, by the Markov inequality, $(1/n)\|g_n(\theta)\| = O_P(1)$ uniformly in $\theta \in \Theta$. Therefore, $\|1/n\, g_n'(\theta)(\hat{\Omega}_n^{-1} - \Omega_n^{-1})g_n(\theta)\| = o_P(1)$, uniformly in $\theta \in \Theta$. The consistency of the feasible optimum GMME $\hat{\theta}_{o,n}$ follows. For the limiting distribution, as $(1/n)\partial g_n(\hat{\theta}_n)/\partial\theta = -D_n/n + o_P(1)$ from the proof of Proposition 1,

$$\begin{aligned}\sqrt{n}(\hat{\theta}_{o,n} - \theta_0) &= -\left[\frac{1}{n}\frac{\partial g_n'(\hat{\theta}_n)}{\partial\theta}\left(\frac{\hat{\Omega}_n}{n}\right)^{-1}\frac{1}{n}\frac{\partial g_n(\hat{\theta}_n)}{\partial\theta'}\right]^{-1}\frac{1}{n}\frac{\partial g_n'(\hat{\theta}_n)}{\partial\theta}\left(\frac{\hat{\Omega}_n}{n}\right)^{-1}\frac{1}{\sqrt{n}}\, g_n(\theta_0)\\ &= \left[\frac{D_n'}{n}\left(\frac{\Omega_n}{n}\right)^{-1}\frac{D_n}{n}\right]^{-1}\frac{D_n'}{n}\left(\frac{\Omega_n}{n}\right)^{-1}\frac{1}{\sqrt{n}}\, g_n(\theta_0) + o_P(1).\end{aligned}$$

The limiting distribution of $\sqrt{n}(\hat{\theta}_{on} - \theta_0)$ in (3.3) follows from this expansion.

For the overidentification test, by the Taylor expansion,

$$\frac{1}{\sqrt{n}}\, g_n(\hat{\theta}_{o,n}) = \frac{g_n(\theta_0)}{\sqrt{n}} + \frac{1}{n}\frac{\partial g_n(\bar{\theta}_n)}{\partial \theta'}\sqrt{n}(\hat{\theta}_{o,n}-\theta_0) = \frac{g_n(\theta_0)}{\sqrt{n}} - \frac{D_n}{n}\sqrt{n}(\hat{\theta}_{o,n}-\theta_0) + o_P(1)$$
$$= A_n \frac{g_n(\theta_0)}{\sqrt{n}} + o_P(1),$$

where $A_n = I_n - (D_n/n)[(D_n'/n)(\Omega_n/n)^{-1}D_n/n]^{-1}D_n'/n(\Omega_n/n)^{-1}$. Therefore,

$$g_n'(\hat{\theta}_{on})\hat{\Omega}_n^{-1}g_n(\hat{\theta}_{on}) = \frac{1}{\sqrt{n}}\, g_n'(\theta_0)A_n'\left(\frac{\Omega_n}{n}\right)^{-1} A_n \frac{1}{\sqrt{n}}\, g_n(\theta_0) + o_P(1)$$
$$= \frac{1}{\sqrt{n}}\, g_n'(\theta_0)\left(\frac{\Omega_n}{n}\right)^{-1/2}\left\{ I_n - \left(\frac{\Omega_n}{n}\right)^{-1/2}\frac{D_n}{n}\left[\frac{D_n'}{n}\left(\frac{\Omega_n}{n}\right)^{-1}\frac{D_n}{n}\right]^{-1}\right.$$
$$\left.\times\frac{D_n'}{n}\left(\frac{\Omega_n}{n}\right)^{-1/2}\right\}\left(\frac{\Omega_n}{n}\right)^{-1/2}\frac{1}{\sqrt{n}}\, g_n(\theta_0) + o_P(1)$$
$$\xrightarrow{D} \chi^2((m+k_x)-(k+1)),$$

because $(1/\sqrt{n})\, g_n(\theta_0) \xrightarrow{D} N(0, \lim_{n\to\infty}\Omega_n/n)$ as in the proof of Proposition 1. □

Proof of Proposition 3. For the feasible best GMM estimation with $\mathscr{P}_{1n}$, the vector of moment functions is $\hat{g}_{b,n}(\theta) = (\varepsilon_n'(\theta)(\hat{G}_n - (\mathrm{tr}(\hat{G}_n)/n)I_n)\varepsilon_n(\theta), \varepsilon_n'(\theta)\hat{G}_nX_n\hat{\beta}_n, \varepsilon_n'(\theta)X_n)'$, and the corresponding estimated V_n is

$$\hat{V}_n = \hat{\sigma}_n^4\begin{pmatrix} \mathrm{tr}\left(\left(\hat{G}_n - \frac{\mathrm{tr}(\hat{G}_n)}{n}I_n\right)^s\hat{G}_n\right) & 0 \\ 0 & \frac{1}{\hat{\sigma}_n^2}(\hat{G}_nX_n\hat{\beta}_n, X_n)'(\hat{G}_nX_n\hat{\beta}_n, X_n)\end{pmatrix}.$$

When $G_nX_n\beta_0$ and X_n are linearly dependent, the linear moment of $\hat{G}_nX_n\hat{\beta}_n$ would be redundant and should be dropped. The moment function $g_{b,n}(\theta)$ and the corresponding proper weighting function $\hat{V}_n$ should simply be $\hat{g}_{b,n}(\theta) = (\varepsilon_n'(\theta)(\hat{G}_n - (\mathrm{tr}(\hat{G}_n)/n)I_n)\varepsilon_n(\theta), \varepsilon_n'(\theta)X_n)'$, and

$$\hat{V}_n = \hat{\sigma}_n^4\begin{pmatrix} \mathrm{tr}\left(\left(\hat{G}_n - \frac{\mathrm{tr}(\hat{G}_n)}{n}I_n\right)^s\hat{G}_n\right) & 0 \\ 0 & \frac{1}{\hat{\sigma}_n^2}X_n'X_n\end{pmatrix}.$$

The feasible best GMME with $\mathscr{P}_{1n}$ will be derived from $\min_{\theta\in\Theta} g_{b,n}'(\theta)\hat{V}_n^{-1}g_{b,n}(\theta)$.

For the best GMM estimation with the subclass $\mathscr{P}_{2n}$, its moment functions $\hat{g}_{2b,n}$ and estimated variance matrix are those with $\hat{G}_n - Diag(\hat{G}_n)$ replacing $(\hat{G}_n - (\mathrm{tr}(\hat{G}_n)/n)I_n)$ in the preceding expressions.

We shall show that the objective functions $Q_n^*(\theta) = \hat{g}_{b,n}'(\theta)\hat{V}_n^{-1}\hat{g}_{b,n}(\theta)$ and $Q_n(\theta) = g_{b,n}'(\theta)V_n^{-1}g_{b,n}(\theta)$, where $g_{b,n}$ is the counter part of $\hat{g}_{b,n}$ with G_n and β_0 replacing, respectively by $\hat{G}_n$ and $\hat{\beta}_n$, will satisfy the conditions in Lemma A.6. If so, the GMME from the minimization of $Q_n^*(\theta)$ will have the same limiting distribution as that of the minimization of $Q_n(\theta)$. The difference of $Q_n^*(\theta)$ and $Q_n(\theta)$ and its derivatives involve the difference of

$\hat{g}_{b,n}(\theta)$ and $g_{b,n}(\theta)$ and their derivatives. Furthermore, one has to consider the difference of $\hat{V}_n$ and V_n.

First, consider $(1/n)(\hat{g}_{b,n}(\theta) - g_{b,n}(\theta))$. Explicitly,

$$\frac{1}{n}(\hat{g}_{b,n}(\theta) - g_{b,n}(\theta))' = \left(\frac{1}{n}\varepsilon_n'(\theta)(\hat{G}_n - G_n)^d\varepsilon_n(\theta), \frac{1}{n}(\hat{G}_n X_n\hat{\beta}_n - G_n X_n\beta_0)'\varepsilon_n(\theta), 0\right).$$

The $\varepsilon_n(\theta)$ is related to ε_n as $\varepsilon_n(\theta) = \varepsilon_n + (\lambda_0 - \lambda)G_n\varepsilon_n + d_n(\theta)$ where $d_n(\theta) = (\lambda_0 - \lambda)G_n X_n\beta_0 + X_n(\beta_0 - \beta)$. It follows that $(1/n)X_n'G_n'\varepsilon_n(\theta) = (1/n)X_n'G_n'\varepsilon_n + (\lambda_0 - \lambda)(1/n)X_n'G_n'G_n\varepsilon_n + (1/n)X_n'G_n'd_n(\theta) = (1/n)X_n'G_n'd_n(\theta) + o_P(1)$ by Lemma A.4, uniformly in $\theta \in \Theta$. On the other hand, Lemma A.1 implies that $(1/n)X_n'G_n'd_n(\theta) = O_P(1)$ uniformly in $\theta \in \Theta$. The uniformity follows because $d_n(\theta)$ is linear in λ and β. Hence $(1/n)X_n'G_n'\varepsilon_n(\theta) = O_P(1)$ uniformly in $\theta \in \Theta$. Similarly, Lemma A.9 implies $(1/n)X_n'(\hat{G}_n - G_n)'\varepsilon_n(\theta) = (1/n)X_n'(\hat{G}_n - G_n)'\varepsilon_n + (\lambda_0 - \lambda)(1/n)X_n'(\hat{G}_n - G_n)'G_n\varepsilon_n + (1/n)X_n'(\hat{G}_n - G_n)'d_n(\theta) = o_P(1)$ uniformly in $\theta \in \Theta$. It follows that $(1/n)(\hat{G}_n X_n\hat{\beta}_n - G_n X_n\beta_0)'\varepsilon_n(\theta) = (1/n)\hat{\beta}_n X_n'(\hat{G}_n - G_n)'\varepsilon_n(\theta) + (\hat{\beta}_n - \beta_0)'(1/n)X_n'G_n'\varepsilon_n(\theta) = o_P(1)$ because $\hat{\beta}_n - \beta_0 = o_P(1)$. Similarly, Lemmas A.9 and A.10 imply that $(1/n)\varepsilon_n'(\theta)(\hat{G}_n - G_n)^d\varepsilon_n(\theta) = o_P(1)$ uniformly in $\theta \in \Theta$. Hence, we conclude that $(1/n)(\hat{g}_{b,n}(\theta) - g_{b,n}(\theta)) = o_P(1)$ uniformly in $\theta \in \Theta$.

Consider the derivatives of $\hat{g}_{b,n}(\theta)$ and $g_{b,n}(\theta)$. As the second derivatives of $\varepsilon_n(\theta)$ with respect to θ are zero because $\varepsilon_n(\theta)$ is linear in θ, it follows that

$$\frac{\partial g_{b,n}(\theta)}{\partial\theta'} = \begin{pmatrix} \varepsilon_n'(\theta)G_n^{ds}\frac{\partial\varepsilon_n(\theta)}{\partial\theta'} \\ (G_n X_n\beta_0)'\frac{\partial\varepsilon_n(\theta)}{\partial\theta'} \\ X_n\frac{\partial\varepsilon_n(\theta)}{\partial\theta'} \end{pmatrix}, \quad \frac{\partial^2 g_{b,n}(\theta)}{\partial\theta\partial\theta'} = \begin{pmatrix} \frac{\partial\varepsilon_n'(\theta)}{\partial\theta}G_n^{ds}\frac{\partial\varepsilon_n(\theta)}{\partial\theta'} \\ 0 \\ 0 \end{pmatrix}.$$

The first order derivatives of $\varepsilon_n(\theta)$ is $\partial\varepsilon_n(\theta)/\partial\theta' = -(W_n Y_n, X_n)$. Because $W_n Y_n = G_n X_n\beta_0 + G_n\varepsilon_n$,

$$\begin{aligned}
&\frac{1}{n}(W_n Y_n)'(\hat{G}_n - G_n)^{ds}\varepsilon_n(\theta) \\
&\quad= \frac{1}{n}(G_n X_n\beta_0)'(\hat{G}_n - G_n)^{ds}d_n(\theta) + \frac{1}{n}(G_n X_n\beta_0)'(\hat{G}_n - G_n)^{ds}(\varepsilon_n + (\lambda_0 - \lambda)G_n\varepsilon_n) \\
&\qquad + d_n'(\theta)(\hat{G}_n - G_n)^{ds}G_n\varepsilon_n + \frac{1}{n}\varepsilon_n'G_n'(\hat{G}_n - G_n)^{ds}(\varepsilon_n + (\lambda_0 - \lambda)G_n\varepsilon_n) = o_P(1),
\end{aligned}$$

uniformly in $\theta \in \Theta$, and

$$\begin{aligned}
&\frac{1}{n}(W_n Y_n)'(\hat{G}_n - G_n)^{ds}W_n Y_n \\
&\quad= \frac{1}{n}(X_n\beta_0)'G_n'(\hat{G}_n - G_n)^{ds}G_n X_n\beta_0 + \frac{2}{n}(X_n\beta_0)'G_n'(\hat{G}_n - G_n)^{ds}G_n\varepsilon_n \\
&\qquad + \frac{1}{n}\varepsilon_n'G_n'(\hat{G}_n - G_n)^{ds}G_n\varepsilon_n \\
&\quad= o_P(1),
\end{aligned}$$

by Lemmas A.9 and A.10. Similarly, Lemmas A.9 and A.10 imply that $(1/n)X_n'(\hat{G}_n - G_n)^{ds}\varepsilon_n(\theta) = o_P(1)$, $(1/n)X_n'(\hat{G}_n - G_n)^{ds}W_n Y = o_P(1)$, and $(1/n)X_n'(\hat{G}_n - G_n)^{ds}X_n = o_p(1)$. Hence, we conclude that $(1/n)(\partial\hat{g}_{b,n}(\theta)/\partial\theta - \partial g_{b,n}(\theta)/\partial\theta) = o_P(1)$ and $(1/n)(\partial^2\hat{g}_{b,n}(\theta)/\partial\theta\partial\theta' - \partial^2 g_{b,n}(\theta)/\partial\theta\partial\theta') = o_P(1)$ uniformly in $\theta \in \Theta$.

For the cases under consideration, $\hat{V}_n$ and V_n are block diagonal matrices. Without loss of generality, consider the situation that $G_nX_n\beta_0$ and X_n are not linearly dependent for large n. Thus,

$$\hat{V}_n = \begin{pmatrix} \mathrm{tr}(\hat{G}_n^{ds}\hat{G}_n) & 0 \\ 0 & \hat{\sigma}_n^2\hat{A}_n \end{pmatrix}, \quad V_n = \begin{pmatrix} \mathrm{tr}(G_n^{ds}G_n) & 0 \\ 0 & \sigma_0^2 A_n \end{pmatrix},$$

where $A_n = (G_nX_n\beta_0, X_n)'(G_nX_n\beta_0, X_n)$ and $\hat{A}_n = (\hat{G}_nX_n\hat{\beta}_n, X_n)'(\hat{G}_nX_n\hat{\beta}_n, X_n)$. The difference of $\hat{G}_n$ and G_n is $(\hat{G}_n - G_n) = (\hat{\lambda}_n - \lambda_0)G_n^2 + (\hat{\lambda}_n - \lambda_0)^2\hat{G}_nG_n^2$. This implies that

$$\frac{1}{n}\mathrm{tr}(G_n^{ds}(\hat{G}_n - G_n)) = (\hat{\lambda}_n - \lambda_0)\frac{1}{n}\mathrm{tr}(G_n^{ds}G_n^2) + (\hat{\lambda}_n - \lambda_0)^2\frac{1}{n}\mathrm{tr}(G_n^{ds}\hat{G}_nG_n^2) = o_P(1),$$

because $(1/n)\,\mathrm{tr}(G_n^{ds}G_n^2) = O(1)$, $(1/n)\,\mathrm{tr}(G_n^{ds}\hat{G}_nG_n^2) = O_P(1)$ and $(\hat{\lambda}_n - \lambda_0) = o_P(1)$. Similarly,

$$\frac{1}{n}\mathrm{tr}[(\hat{G}_n - G_n)^{ds}\hat{G}_n] = (\hat{\lambda}_n - \lambda_0)\frac{1}{n}\mathrm{tr}(G_n^{2ds}\hat{G}_n) + (\hat{\lambda}_n - \lambda_0)^2\frac{1}{n}\mathrm{tr}[(\hat{G}_nG_n^2)^{ds}\hat{G}_n] = o_P(1).$$

Therefore, $(1/n)[\mathrm{tr}(\hat{G}_n^{ds}\hat{G}_n) - \mathrm{tr}(G_n^{ds}G_n)] = (1/n)\,\mathrm{tr}[(\hat{G}_n^{ds} - G_n^{ds})\hat{G}_n + G_n^{ds}(\hat{G}_n - G_n)] = o_P(1)$. Consider the remaining block matrix. Because $\hat{\sigma}_n^2$ is a consistent estimate of σ_0^2 and $(1/n)A_n = O(1)$ by Lemma A.1,

$$\frac{1}{n}(\hat{\sigma}_n^2\hat{A}_n - \sigma_0^2A_n) = \hat{\sigma}_n^2\frac{1}{n}(\hat{A}_n - A_n) + (\hat{\sigma}_n^2 - \sigma_0^2)\frac{1}{n}A_n = \hat{\sigma}_n^2\frac{1}{n}(\hat{A}_n - A_n) + o_P(1).$$

The difference $(1/n)(\hat{A}_n - A_n)$ is $o_p(1)$ because

$$\frac{1}{n}(\hat{G}_nX_n\hat{\beta}_n)'X_n - \frac{1}{n}(G_nX_n\beta_0)'X_n = \frac{1}{n}\hat{\beta}_n'X_n'(\hat{G}_n' - G_n')X_n + \frac{1}{n}(\hat{\beta}_n - \beta_0)'X_n'G_n'X_n = o_P(1)$$

and

$$\begin{aligned}
&\frac{1}{n}(\hat{G}_nX_n\hat{\beta}_n)'(\hat{G}_nX_n\hat{\beta}_n) - \frac{1}{n}(G_nX_n\beta_0)'(G_nX_n\beta_0) \\
&\quad = \hat{\beta}_n'\frac{1}{n}X_n'(\hat{G}_n - G_n)'\hat{G}_nX_n\hat{\beta}_n + \hat{\beta}_n'\frac{1}{n}X_n'G_n'(\hat{G}_n - G_n)X_n\hat{\beta}_n \\
&\qquad + (\hat{\beta}_n + \beta_0)'\frac{1}{n}X_n'G_n'G_nX_n(\hat{\beta}_n - \beta_0) = o_P(1),
\end{aligned}$$

by Lemmas A.1 and A.9. In conclusion, $(1/n)\hat{V}_n - (1/n)V_n = o_P(1)$. It follows that $((1/n)\hat{V}_n)^{-1} - ((1/n)V_n)^{-1} = o_P(1)$ by the continuous mapping theorem.

Furthermore, because $(1/n)(\hat{g}_{b,n}(\theta) - g_{b,n}(\theta)) = o_P(1)$, and $(1/n)[g_{b,n}(\theta) - \mathrm{E}(g_{b,n}(\theta))] = o_P(1)$ uniformly in $\theta \in \Theta$, and $\sup_{\theta\in\Theta} 1/n|\mathrm{E}(g_{b,n}(\theta))| = O(1)$ in the proof of Proposition 2, $(1/n)g_{b,n}(\theta)$ and $(1/n)\hat{g}_{b,n}(\theta)$ are stochastically bounded, uniformly in $\theta \in \Theta$. Similarly, $(1/n)(\partial g_{b,n}(\theta)/\partial\theta)$, $(1/n)(\partial\hat{g}_{b,n}(\theta)/\partial\theta)$, $(1/n)(\partial^2 g_{b,n}(\theta)/\partial\theta\partial\theta)$ and $(1/n)(\partial^2\hat{g}_{b,n}(\theta)/\partial\theta\partial\theta)$ are stochastically bounded, uniformly in $\theta \in \Theta$. With the uniform convergence in probability and uniformly stochastic boundedness properties, the difference of $Q_n^*(\theta)$ and $Q_n(\theta)$ can be investigated. By expansion,

$$\begin{aligned}\frac{1}{n}(Q_n^*(\theta)-Q_n(\theta)) &= \frac{1}{n}\hat{g}_{b,n}'(\theta)\hat{V}_n^{-1}(\hat{g}_{b,n}(\theta)-g_{b,n}(\theta)) \\ &\quad+\frac{1}{n}g_{b,n}'(\theta)(\hat{V}_n^{-1}-V_n^{-1})\hat{g}_{b,n}+\frac{1}{n}g_{b,n}'(\theta)V_n^{-1}(\hat{g}_{b,n}(\theta)-g_{b,n}(\theta)) \\ &= o_P(1),\end{aligned}$$

uniformly in $\theta \in \Theta$. Similarly, for each component θ_l of θ,

$$\begin{aligned}\frac{1}{n}\frac{\partial^2 Q_n^*(\theta)}{\partial\theta_l\partial\theta'}-\frac{1}{n}\frac{\partial^2 Q_n(\theta)}{\partial\theta_l\partial\theta'} &= \frac{2}{n}\left[\frac{\partial\hat{g}_{b,n}'(\theta)}{\partial\theta_l}\hat{V}_n^{-1}\frac{\partial\hat{g}_{b,n}(\theta)}{\partial\theta'}+\hat{g}_{b,n}'(\theta)\hat{V}_n^{-1}\frac{\partial^2\hat{g}_{b,n}(\theta)}{\partial\theta_l\partial\theta'}\right. \\ &\quad\left.-\left(\frac{\partial g_{b,n}(\theta)}{\partial\theta_l}V_n^{-1}\frac{\partial g_{b,n}(\theta)}{\partial\theta'}+g_{b,n}'(\theta)V_n^{-1}\frac{\partial^2 g_{b,n}(\theta)}{\partial\theta_l\partial\theta'}\right)\right] \\ &= o_P(1).\end{aligned}$$

Finally, because $((\partial\hat{g}_{b,n}'(\theta_0)/\partial\theta)\hat{V}_n^{-1}-(\partial g_{b,n}'(\theta_0)/\partial\theta)V_n^{-1})=o_P(1)$ as above, and $(1/\sqrt{n})g_{b,n}(\theta_0)=O_P(1)$ by the central limit theorems in Lemmas A.4 and A.5,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\left(\frac{\partial Q_n^*(\theta_0)}{\partial\theta}-\frac{\partial Q_n(\theta_0)}{\partial\theta}\right) \\ &\quad= 2\left\{\frac{\partial\hat{g}_{b,n}'(\theta_0)}{\partial\theta}\hat{V}_n^{-1}\frac{1}{\sqrt{n}}(\hat{g}_{b,n}(\theta_0)-g_{b,n}(\theta_0))\right. \\ &\quad\quad\left.+\left(\frac{\partial\hat{g}_{b,n}'(\theta_0)}{\partial\theta}\hat{V}_n^{-1}-\frac{\partial g_{b,n}'(\theta_0)}{\partial\theta}V_n^{-1}\right)\frac{1}{\sqrt{n}}g_{b,n}(\theta_0)\right\} \\ &\quad= 2\frac{\partial\hat{g}_{b,n}'(\theta_0)}{\partial\theta}\hat{V}_n^{-1}\frac{1}{\sqrt{n}}(\hat{g}_{b,n}(\theta_0)-g_{b,n}(\theta_0))+o_P(1).\end{aligned}$$

This difference will be of order $o_P(1)$ if $(1/\sqrt{n})(\hat{g}_{b,n}(\theta_0)-g_{b,n}(\theta_0))=o_P(1)$. Lemma A.10 implies that the component $(1/\sqrt{n})\varepsilon_n'(\hat{G}_n-G_n)^d\varepsilon_n=o_P(1)$. Lemmas A.4 and A.9 imply that $(1/\sqrt{n})[(\hat{G}_nX_n\hat{\beta}_n)'-(G_nX_n\beta_0)']\varepsilon_n=\hat{\beta}_n'(1/\sqrt{n})X_n'(\hat{G}_n-G_n)'\varepsilon_n+(\hat{\beta}_n-\beta_0)'(1/\sqrt{n})X_n'G_n'\varepsilon_n=o_P(1)$, and $(1/\sqrt{n})(\hat{\beta}_n'X_n'\varepsilon_n-\beta_0'X_n'\varepsilon_n)=(\hat{\beta}_n-\beta_0)'(1/\sqrt{n})X_n'\varepsilon_n=o_P(1)$. Hence $(1/\sqrt{n})(\hat{g}_{b,n}(\theta_0)-g_{b,n}(\theta_0))=o_P(1)$.

Finally, the results of the proposition follows from Lemma A.6 and the preceding propositions. □

References

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press, Cambridge, MA.

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer, Dordrecht.

Anselin, L., 1990. Some robust approaches to testing and estimation in spatial econometrics. Regional Science and Urban Economics 20, 141–163.

Anselin, L., Bera, A.K., 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. In: Ullah, A., Giles, D.E.A. (Eds.), Handbook of Applied Economic Statistics. Marcel Dekker, New York, pp. 237–289.

Case, A.C., 1991. Spatial patterns in household demand. Econometrica 59, 953–965.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.

Horn, R., Johnson, C., 1985. Matrix Analysis. Cambridge University Press, New York.

Kelejian, H.H., Prucha, I.R., 1997. Estimation of spatial regression models with autoregressive errors by two-stage least squares procedures: a serious problem. International Regional Science Review 20 (1 and 2), 103–111.

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

Kelejian, H.H., Prucha, I.R., 1999. A generalized moments estimator for the autoregressive parameter in a spatial model. International Economic Review 40, 509–533.

Kelejian, H.H., Prucha, I.R., 2001. On the asymptotic distribution of the Moran I test statistic with applications. Journal of Econometrics 104, 219–257.

Kelejian, H.H., Robinson, D.P., 1993. A suggested method of estimation for spatial interdependence models with autocorrelated errors, and an application to a county expenditure model. Papers in Regional Science 72, 297–312.

Land, K., Deane, G., 1992. On the large-sample estimation of regression models with spatial-or network-effects terms: a two stage least squares approach. In: Marsden, P. (Ed.), Sociological Methodology. Jossey Bass, San Francisco, pp. 221–248.

Lee, L.F., 2001. Generalized method of moments estimation of spatial autoregressive processes. Manuscript, 2001, Department of Economics, OSU, Columbus, OH.

Lee, L.F., 2003. Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. Econometric Reviews 22, 307–335.

Lee, L.F., 2004. Asymptotic distributions of quasi-maximum likelihood estimators for spatial econometric models. Econometrica 72, 1899–1926.

Ord, J., 1975. Estimation methods for models of spatial interaction. Journal of the American Statistical Association 70, 120–126.

Smirnov, O., Anselin, L., 2001. Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach. Computational Statistics and Data Analysis 35, 301–319.

White, H., 1994. Estimation, Inference and Specification Analysis. Cambridge University Press, New York.