A Narrow and Wide Replication of Mixed Regressive, Spatial Autoregressive Model

Mingyuan Jia
mjia002@m.scnu.edu.cn
School of Economics and Management, South China Normal University

Yunzhi Lu
yunzhi.lu@m.scnu.edu.cn
School of Economics and Management, South China Normal University

Summary: We replicate the mixed regressive, spatial autoregressive model (MRSAR) using various approaches considered in Lee (2007) [Journal of Econometrics, 2007]. Our replication generally supports the view in his paper, except for the one which should impose restricted parameter space on the minimization program. We also consider two extended replications from the view of empirical econometricians.

Keywords: Mixed regressive spatial autoregressive model, endogenous regressors, panel data
JEL Codes: C13, C21, R15

Introduction

The original MRSAR model is proposed in Lee (2007), which considers the following spatial model with exogenous regressors X_n as explanatory variables

$$Y_n = \lambda W_n Y_n + X_n \beta + \varepsilon_n, \tag{1}$$

where Y_n is an $n \times 1$ dimensional vector, X_n is an $n \times k$ dimensional matrix of non-stochastic exogenous variables, W_n is a spatial weights matrix of known constants with a zero diagonal, and the stochastic disturbance ε_{ni}, $i = 1, \ldots, n$, of the n-dimensional vector ε_n are *i.i.d.* $(0, \sigma^2)$.

The parameter of interest is the spatial interdependent parameter λ. From the point of view of an econometrician, W_nY_n is called spatial lag in microeconomic empirical studies. For example, spatial lag in corporate finance is usually interpreted as *peer effects* or *network effects* of corporate strategies from other companies, usually within the same industry or city (Leary & Roberts 2014; Grieser *et al.* 2021).

Under some regular assumptions[1], Lee (2007) argues that one can select some matrices P_{jn}'s whose traces are zero, and instrumental variable matrices Q_n which satisfies the orthogonality conditions $Q_n'\varepsilon_n = 0$ for the GMM approach. Hence, the corresponding set of moment functions for the GMM estimation can be described by a vector

$$g_n(\theta) = \left(\varepsilon_n' P_{1n} \varepsilon_n, \ldots, \varepsilon_n' P_{mn} \varepsilon_n, \varepsilon_n' Q_n\right)', \tag{2}$$

where θ is the collection of λ and β. Unlike traditional 2SLS estimation, the moment conditions in Eq.(2) could be divided into two parts, namely, quadratic and linear moments. The quadratic moments are expressed in quadratic forms, $\varepsilon_n' P_{1n} \varepsilon_n, \ldots, \varepsilon_n' P_{mn} \varepsilon_n$, while the linear moments are a function of X_n and W_n.

Lee (2007) considers different cases for selecting these moments, including unweighted GMM (UGMM), optimum GMM (OGMM), and best optimum GMM (BGMM). The UGMM uses an identity matrix I_n as distance matrix for the objective function in GMM approach. Once we estimated the residual from UGMM, it could be used to generate the covariance matrix Σ. The inverse of Σ is used as the distance matrix for OGMM. The BGMM is similar to OGMM, except that it considers the best choice of P_{jn}.

The remainder of this paper can be divided into two main parts. The first part reports our replication of Lee (2007) with and without restricted parameter space. The second part implements two extended replications to guide the empirical researches with endogenous regressors and unbalance panel specification.

The Basic Replication

In this section, we follow Lee (2007) and replicate the Monte Carlo simulations for models shown in Table 1. The stochastic disturbance is assumed to be normally distributed, with variance $\sigma_0^2 = 2$. Under this condition, the maximal likelihood estimation should be the most efficient

[1] Interested readers are referred to his original article. Here we mention one important testable assumption (called *rank condition*), which requires that the coefficient of exogenous variables, β's, should not be zero. If this assumption is violated, one could not identify λ and β except for some special cases. See the discussion started in the bottom of Page 493 of Lee (2007).

estimator for model (1). We consider two sets of true parameters[2] following Lee(2007). The first set (Set 1 hereafter) is $\lambda_0 = 0.6, \beta_{10} = -1, \beta_{20} = 0, \beta_{30} = 1$, where the variance of explanatory variables is equal to the variance of ε_n. The other set (Set 2 hereafter) is $\lambda_0 = 0.6, \beta_{10} = -0.2, \beta_{20} = 0, \beta_{30} = 0.2$, where the variance of explanatory variables is smaller than the variance of ε_n. The weights matrix across 49 districts in Columbus, Ohio is the same one used in Anselin (1988). Also, as pointed out in the Footnote 15 of Lee (2007), we use the estimators based on 2SLS and GMM as initial estimators for the two sets of true parameters respectively. We replicate his work using MATLAB.

[Insert Table 1 around Here]

It is worthwhile to mention that our program is constrained minimization with bounded λ in the interval [-1, 1]. In most of the spatial models, researchers are interested in the cases where λ is bounded between [-1, 1]. However, the 2SLS and MLE methods may not guarantee the estimated λ would lie within the interval. One of the advantages of GMM approach is it can impose restricted parameter space during the optimization process. In this section, we compare the differences between the constrained and unconstrained minimization, even though Lee (2007) suggests that one should use global minimization in GMM.

[Insert Table 2 around Here]

We show our main replication results in Table 2. Table 2 only summarizes the estimates of λ across two sets of parameters since it is the only one could be affected by different GMM estimators. Most of the estimates are comparable to Lee (2007)'s results. For example, we identify the upward biases of the 2SLS method is much more prominent under the specification of Set 2. Overall, the 2SLS method is not as efficient as GMM approaches as can be seen from standard deviation (SD) and root of mean squared error (RMSE). In addition, OGMM outperforms other GMM methods in small sample. Meanwhile, BGMM yields estimates almost as efficient as MLE when sample size is large ($N = 490$). In summary, our replication shows that the GMM approach proposed by Lee (2007) is valid and thus provides a feasible framework to estimate MRSAR models.

However, the most crucial finding is that the simulation results in Lee (2007) can only be replicated when λ is constrained between [-1, 1]. The estimates between constrained and unconstrained optimization are quite different as shown in Table 2. Two interesting facts are worth mentioning. First, when the true parameters are from Set 1, the estimates of GMM with and without constraints exhibit slight differences when the sample size is small, but tend to converge as sample size increases. Second, when the true parameters are from Set 2, it seems that the global minimization program tends to over-estimate λ, and, more importantly, generate larger SDs and RMSEs than constrained minimization. This bias cannot be eliminated by increasing the sample size. Therefore, we suggest that *one should impose a restricted parameter space on λ to yield more efficient estimates in the empirical studies*.

The Extended Replication

In this section, we consider several extended replications of Lee (2007) in the view of empirical practice. These exercises aim to provide some meaningful guidance for empirical researchers when the GMM approach is used to estimate MRSAR models.

[2] As is mentioned in the previous footnote, the rank condition implies that not all β's are zeros. Hence, the second set of parameters will make the 2SLS method hard to estimate λ, as is suggested in Lee (2007).

1. Endogenous Control Variables

The exogeneity of the instrumental variables is a major concern in empirical study. For instance, when constructing IV matrices using control variables in corporate finance, the estimator might be biased as many of these variables are endogenous. We investigate this issue in the first extended replication. Specifically, we consider an extreme case where all regressors X_n are correlated with stochastic disturbance ε_n, with correlation coefficient equals –0.5.

[Insert Table 3 around here]

Table 3 presents the estimates of λ in the first extension. For the sake of brevity, we omit the coefficients of control variables[3]. In general, the estimates of λ remain robust across classes of GMM estimators when we use endogenous variables to build IV matrices. When the sample size is large enough, the estimates of λ converge to its true value. However , It is also found that the upward biases of 2SLS are smaller than those in Table 2, due to the negative correlation coefficient. In addition, we notice that BGMM approach outperforms ML in terms of SD and RMSE in some cases. This supports the view of Lee (2007) that GMM approaches may be useful when the stochastic disturbance is no longer *i.i.d.* normal. As for the over-identification test, almost all GMM approaches could not reject the null hypothesis that instrument variables are exogenous at the 10% significance level. This may be partly due to the adoption of quadratic moment functions. Therefore, *even at the presence of endogenous control variables, the GMM approaches can still provide consistent estimates of* λ.

2. (Unbalanced) Panel Setting

The second replication extends the MRSAR model into panel data with individual fixed effects. Although there have been many theoretical papers investigating the spatial dynamic panel models (Lee & Yu 2010; Yang 2018; Jin *et al.* 2020), few of them pay attention to the unbalanced panel cases. Consider the following spatial panel model

$$Y_{it} = v_i + \lambda W_t Y_{it} + X_{it}\beta + \varepsilon_{it}, i = 1,\ldots,N; t = 1,\ldots,T \,. \tag{3}$$

One standard procedure to estimate Eq.(3) is to eliminate the fixed effects first, and then apply the estimation approaches developed in the cross-sectional model. This is straightforward when Eq.(3) is balanced panel where the weights matrix W_t is also time-invariant. Suppose $\tilde{Y}$ and $\tilde{X}$ are the demeaned variable of Y_{it} and X_{it}, where

$$\tilde{Y} = Y_{it} - \frac{1}{T}\sum_{t=1}^{T} Y_{it}, \tilde{X} = X_{it} - \frac{1}{T}\sum_{t=1}^{T} X_{it} \,. \tag{4}$$

One can rewrite Eq.(3) as

$$\tilde{Y} = \lambda W \tilde{Y} + \tilde{X}\beta + \tilde{\varepsilon} \,. \tag{5}$$

By doing so, we again come to a cross-sectional MRSAR model in Eq.(5).

When we want to eliminate the fixed effects in the panel data, we can either demean first before weighting or weight before demean. Therefore, one of the crucial questions is which one is better? If we use D_n to denote the demean matrix such that $D_n l_T = 0$, then "demean first" implies

$$D_n Y_{it} = D_n v_i + \lambda W_t D_n Y_{it} + D_n X_{it}\beta + D_n \varepsilon_{it} \,, \tag{6}$$

while "weighting first" implies

[3] Without surprise, the estimates of control variables' coefficient are downward biased.

$$D_n Y_{it} = D_n v_i + \lambda D_n W_t Y_{it} + D_n X_{it}\beta + D_n \varepsilon_{it}\,. \tag{7}$$

In applied econometrics, some scholars such as Grieser *et al.* (2021) support the first approach, while others like Fell & Haynie (2013) favor the second one[4]. Which approach to choose in the empirical studies does matter as we discuss below.

The sequence of demean and weighting will not affect the estimation result under the balanced panel setting. If we denote l_T as a T×1 vector of ones, and $J = I_N \otimes l_T$, where ⊗ denotes Kronecker product, then a demean matrix D_n under the balance panel setting is

$$D_n = I_{NT} - J\left(J'J\right)^{-1} J'\,.$$

It is straightforward to prove that $D_nW = WD_n$ for any feasible matrix W. However, the unbalanced panel setting cannot ensure the interchangeable property of D_n and W, which leads to different results when an empirical researcher tries to estimate Eq.(3) with the two approaches mentioned above.

We extend the Monte Carlo simulation to the unbalanced panel to investigate the efficiency of these two methods. Specifically, we assume that v_i is generated from a *i.i.d.* standard normal distribution. We randomly drop 30% observations[5] in each simulation to generate an unbalance panel, and re-calculate the weights matrix. Hence, the weights matrix becomes time-varying and block diagonal.

[Insert Table 4 around here]

The simulation results are reported in Table 4. We find that "demean first" and "weighting first" generate different estimates, though they are both close to the true values. Regarding the accuracy of different approaches, "weighting first" combined with OGMM and BGMM provides smaller RMSE than "demean first", regardless of sample size and parameter setting. In contrast, "demean first" combined with UGMM yields smaller RMSE instead. Additionally, it is difficult to find a consistent procedure for 2SLS to generate smaller RMSEs across different specifications. Since OGMM and BGMM are more efficient than UGMM, we believe that they combined with "weighting first" should be preferred when we estimate a panel MRSAR model.

In short, our last suggestion is that *one should use OGMM and BGMM combined with "weighting first" approach to eliminate individual fixed effects for panel MRSAR models*.

References

Anselin, L., 1988. Spatial econometrics: methods and models. Kluwer, Dordrecht.

Fell, H., Haynie, A.C.(2013), SPATIAL COMPETITION WITH CHANGING MARKET INSTITUTIONS, *Journal of Applied Econometrics*, 28: 702-719

Grieser, W., Hadlock, C., LeSage, J., Zekhnini, M.(2021), Network effects in corporate financial policies, *Journal of Financial Economics*

Jin, F., Lee, L., Yu, J.(2020), First difference estimation of spatial dynamic panel data models with fixed effects, *Economics Letters*, 189: 109010

Leary, M.T., Roberts, M.R.(2014), Do peer firms affect corporate financial policy? *The Journal of Finance*, 69: 139-178

[4] From the code shared by Fell & Haynie (2013), we find they calculate W_nY first, then left multiply W_nY by the demean matrix.

[5] We also try 10%, 20%, and fixed number of observations (*e.g.*, 80). The qualitative results are similar.

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

Lee, L., Yu, J.(2010), Estimation of spatial autoregressive panel data models with fixed effects, *Journal of Econometrics*, 154: 165-185

Yang, Z.(2018), Unified M-estimation of fixed-effects spatial dynamic models with short panels, *Journal of Econometrics*, 205: 423-447

Tables

Table 1. The summary of MRSAR specifications.

This table summarizes the five specifications in Lee (2007), including two-stage least squared (2SLS), unweighted GMM (UGMM), optimum GMM (OGMM), best optimum GMM (BGMM), and maximal likelihood approach (ML). P_{jn} and Q_n are matrices used in Eq.(2). 2SLS use standard IV's as a function of X_n and W_n. UGMM use IV's for linear moments and some additional zero-trace matrices for quadratic moments of Eq.(2). OGMM and BGMM use the inverse of estimated variance matrix Σ as distance matrix. N/A stands for "Not Applicable". $\hat{\lambda}_n$ and $\hat{\beta}_n$ are initial consistent initial estimates.

Approach	P_{jn}	Q_n	Distance matrix
2SLS	N/A	$X_n, W_n X_n, W_n^2 X_n$	N/A
UGMM	$W_n, W_n^2 - \frac{\mathrm{tr}\left(W_n^2\right)}{n} I_n$	$X_n, W_n X_n, W_n^2 X_n$	I_n
OGMM	$W_n, W_n^2 - \frac{\mathrm{tr}\left(W_n^2\right)}{n} I_n$	$X_n, W_n X_n, W_n^2 X_n$	Σ^{-1}
BGMM	$W_n\left(I_n - \hat{\lambda}_n W_n\right)^{-1} - \frac{1}{n}\mathrm{tr}\left(W_n\left(I_n - \hat{\lambda}_n W_n\right)^{-1}\right) I_n$	$\left(I_n - \hat{\lambda}_n W_n\right)^{-1} X_n \hat{\beta}_n$	Σ^{-1}
ML	N/A	N/A	N/A

Table 2. Replication of Lee (2007) with and without restricted parameter space of *λ*.

This table shows the extended replication of Monte Carlo simulation in the Table 1 and Table 2 of Lee (2007). The cross-sectional mixed regressive, spatial autoregressive (MRSAR) model is defined by Eq.(1), with sample size N varying from 49 to 490. The true values of parameters for Set 1 are $\lambda_0 = 0.6$, $\beta_{10} = -1$, $\beta_{20} = 0$, $\beta_{30} = 1$, while those for Set 2 are $\lambda_0 = 0.6$, $\beta_{10} = -0.2$, $\beta_{20} = 0$, $\beta_{30} = 0.2$. The columns with constraint impose a restricted parameter space $\lambda \in [-1, 1]$. The MRSAR model is estimated by three different methods indicated in the first column. They include unweighted GMM (UGMM), optimum GMM (OGMM), and best optimum GMM (BGMM). Each method is repeated 1000 times. MEAN, SD, and RMSE, are the average, the standard deviation and the root of mean squared error of estimated coefficients, respectively.

Method	Set 1						Set 2					
	With Constraint			Without Constraint			With Constraint			Without Constraint		
	MEAN	SD	RMSE	MEAN	SD	RMSE	MEAN	SD	RMSE	MEAN	SD	RMSE
$N = 49$												
UGMM	0.599	0.134	0.134	0.603	0.151	0.151	0.598	0.158	0.158	0.664	0.333	0.333
OGMM	0.643	0.128	0.135	0.645	0.136	0.143	0.676	0.171	0.188	0.720	0.274	0.300
BGMM	0.588	0.120	0.121	0.592	0.137	0.137	0.591	0.155	0.155	0.638	0.293	0.295
$N = 245$												
UGMM	0.597	0.052	0.052	0.597	0.052	0.052	0.596	0.059	0.059	0.665	0.283	0.283
OGMM	0.605	0.047	0.047	0.605	0.047	0.047	0.609	0.060	0.061	0.619	0.116	0.118
BGMM	0.597	0.046	0.046	0.597	0.046	0.046	0.597	0.058	0.058	0.629	0.195	0.197
$N = 490$												
UGMM	0.598	0.036	0.036	0.598	0.036	0.036	0.598	0.040	0.040	0.659	0.262	0.262
OGMM	0.602	0.034	0.034	0.602	0.034	0.034	0.604	0.040	0.040	0.607	0.066	0.066
BGMM	0.598	0.033	0.033	0.598	0.033	0.033	0.598	0.040	0.040	0.622	0.193	0.195

Table 3. Estimation with endogenous regressors.

This table shows the extended replication of Monte Carlo simulation in the Table 1 and Table 2 of Lee (2007). All regressors are set to be correlated with stochastic disturbance, whose correlation coefficient is –0.5. The cross-sectional mixed regressive, spatial autoregressive (MRSAR) model is defined by Eq.(1), with sample size N varying from 49 to 490. The table presents the estimated values of λ. The true values of parameters for Set 1 are $\lambda_0 = 0.6$, $\beta_{10} = -1$, $\beta_{20} = 0$, $\beta_{30} = 1$, while those for Set 2 are $\lambda_0 = 0.6$, $\beta_{10} = -0.2$, $\beta_{20} = 0$, $\beta_{30} = 0.2$. They include two-stage least squared (2SLS), unweighted GMM (UGMM), optimum GMM (OGMM), best optimum GMM (BGMM), and maximal likelihood approach (ML). Each method is repeated 1000 times. MEAN, SD, and RMSE, are the average, the standard deviation and the root of mean squared error of estimated coefficients, respectively. The column labeled $p < 0.10$ is the proportion of the p-values of Hansen's J smaller than 10% of the simulations.

Method	Set 1				Set 2			
	MEAN	SD	RMSE	$p < 0.10$	MEAN	SD	RMSE	$p < 0.10$
$N = 49$								
2SLS	0.621	0.093	0.096	0.094	0.641	0.132	0.138	0.096
UGMM	0.613	0.094	0.094	0.000	0.612	0.111	0.111	0.000
OGMM	0.620	0.084	0.086	0.000	0.632	0.111	0.115	0.000
BGMM	0.596	0.079	0.079	0.000	0.593	0.101	0.101	0.000
ML	0.584	0.079	0.080	N/A	0.576	0.101	0.104	N/A
$N = 245$								
2SLS	0.603	0.037	0.038	0.074	0.607	0.054	0.055	0.074
UGMM	0.601	0.034	0.034	0.000	0.601	0.041	0.041	0.000
OGMM	0.602	0.031	0.031	0.000	0.604	0.039	0.039	0.000
BGMM	0.598	0.031	0.031	0.000	0.598	0.038	0.038	0.000
ML	0.592	0.031	0.032	N/A	0.588	0.039	0.041	N/A
$N = 490$								
2SLS	0.601	0.026	0.026	0.072	0.602	0.037	0.037	0.073
UGMM	0.600	0.024	0.024	0.000	0.600	0.028	0.028	0.000
OGMM	0.600	0.022	0.022	0.000	0.600	0.027	0.027	0.000
BGMM	0.598	0.022	0.022	0.000	0.598	0.027	0.027	0.000
ML	0.592	0.022	0.024	N/A	0.589	0.027	0.030	N/A

Table 4. Estimation with unbalance spatial panel data.

This table shows the extended replication of Monte Carlo simulation in the Table 1 and Table 2 of Lee (2007). The panel mixed regressive, spatial autoregressive (MRSAR) model is defined by Eq.(3), with sample size N varying from 49 to 98. We randomly drop 30% observation to capture the feature of unbalance panel data. The true values of parameters for Set 1 are $\lambda_0 = 0.6$, $\beta_{10} = -1$, $\beta_{20} = 0$, $\beta_{30} = 1$, while those for Set 2 are $\lambda_0 = 0.6$, $\beta_{10} = -0.2$, $\beta_{20} = 0$, $\beta_{30} = 0.2$. The columns with "Demean First" are results from demeaning the Y first, then calculating WY, while those with "Weight First" are results from calculating WY first, then demeaning the variables, including WY. The MRSAR model is estimated by three different methods indicated in the first column. They include unweighted GMM (UGMM), optimum GMM (OGMM), and best optimum GMM (BGMM). Each method is repeated 1000 times. MEAN, SD, and RMSE, are the average, the standard deviation and the root of mean squared error of estimated coefficients, respectively.

Method	Set 1						Set 2					
	Demean First			Weight First			Demean First			Weight First		
	MEAN	SD	RMSE	MEAN	SD	RMSE	MEAN	SD	RMSE	MEAN	SD	RMSE
$N = 49$, $T = 10$												
2SLS	0.582	0.062	0.064	0.607	0.059	0.059	0.694	0.236	0.254	0.716	0.234	0.261
UGMM	0.565	0.040	0.040	0.599	0.041	0.041	0.569	0.044	0.044	0.598	0.047	0.047
OGMM	0.573	0.037	0.045	0.604	0.037	0.037	0.577	0.045	0.050	0.605	0.047	0.048
BGMM	0.570	0.036	0.047	0.599	0.037	0.037	0.572	0.044	0.052	0.599	0.046	0.046
$N = 98$, $T = 10$												
2SLS	0.578	0.047	0.052	0.603	0.047	0.047	0.644	0.196	0.201	0.671	0.207	0.219
UGMM	0.567	0.032	0.032	0.600	0.034	0.034	0.571	0.035	0.035	0.600	0.037	0.037
OGMM	0.572	0.031	0.041	0.602	0.032	0.032	0.576	0.035	0.042	0.603	0.037	0.037
BGMM	0.571	0.031	0.042	0.600	0.032	0.032	0.573	0.035	0.044	0.600	0.037	0.037