Journal of Econometrics 157 (2010) 34–52

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

GMM estimation of spatial autoregressive models with unknown heteroskedasticity

Xu Lin [a], Lung-fei Lee [b,*]

[a] Department of Economics, Wayne State University, Detroit, MI 48202, USA

[b] Department of Economics, The Ohio State University, Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Available online 31 October 2009

JEL classification:
C13
C15
C21

Keywords:
Spatial autoregression
Unknown heteroskedasticity
Robustness
Consistent covariance matrix
GMM

ABSTRACT

In the presence of heteroskedastic disturbances, the MLE for the SAR models without taking into account the heteroskedasticity is generally inconsistent. The 2SLS estimates can have large variances and biases for cases where regressors do not have strong effects. In contrast, GMM estimators obtained from certain moment conditions can be robust. Asymptotically valid inferences can be drawn with consistently estimated covariance matrices. Efficiency can be improved by constructing the optimal weighted estimation.

The approaches are applied to the study of county teenage pregnancy rates. The empirical results show a strong spatial convergence among county teenage pregnancy rates.

1. Introduction

Many economic processes, for example, housing decisions, technology adoption, unemployment, welfare participation, price decisions, etc., exhibit spatial patterns. Recently, spatial models that have a long history in regional science and geography have received substantial attention in various areas of economics, including urban, environmental, labor, developmental and others. However, the allowance of dependence between observations complicates the estimation procedure and calls for some specialized techniques.

The most popular spatial econometric model is the spatial autoregressive (SAR) model (e.g., (1) in Section 2). For a standard SAR model where the error terms are assumed to follow a normal distribution $N(0, \sigma^2)$, the most conventional estimation method is the maximum likelihood (ML). Since there is a Jacobian term, the determinant of the $S_n(\lambda)$ in the likelihood function,[1] the ML method entails significant computational complexities. Even though some simplification or approximation techniques have been suggested,[2] the computation involved may still be demanding, especially for large sample sizes and general spatial weights matrices. Another estimation procedure is the two stage least square (2SLS) for the mixed regressive, spatial autoregressive model (Kelejian and Prucha, 1998; Lee, 2003). The 2SLS estimator (2SLSE) has the virtue of computational simplicity but it is inefficient relative to the maximum likelihood estimator (MLE) since it focuses only on the deterministic part of the model, leaving the information contained in the (reduced form) error terms unexplored. Furthermore, it will be inconsistent when all the exogenous regressors are irrelevant. Kelejian and Prucha (1999) propose a Method of Moment (MOM) method for the regression model with spatial autoregressive disturbances based on correlations of sample observations. However, their estimator is inefficient as compared to the MLE. Lee (2001) generalizes the MOM method into a systematic generalized method of moments (GMM) procedure based on quadratic moment functions and shows the existence of the best GMM estimator (GMME), which can be asymptotically as efficient as the MLE. In Lee (2007a), a GMM procedure that combines both advantages of computational simplicity and efficiency is introduced for the estimation of the mixed regressive, spatial autoregressive model. It is shown that the GMME can be asymptotically more efficient than the 2SLSE and that the best GMME exists and it has the same limiting distribution as the MLE. The basic idea is to combine quadratic moments with the linear moments, where the latter are based on the orthogonality of the exogenous regressors with the model disturbances that generates the 2SLSE. All these ML, MOM and GMM

* Corresponding address: Department of Economics, The Ohio State University, 410 Arps Hall, 1945 N. High St., Columbus, OH 43210-1172, USA.

E-mail addresses: xulin@wayne.edu (X. Lin), lflee@econ.ohio-state.edu (L.-f. Lee).

[1] $S_n(\lambda) = I_n - \lambda W_n$, where W_n is the spatial weights matrix. Note that its dimension is $n \times n$, which is large for large sample sizes.

[2] See, for example, Ord (1975), Smirnov and Anselin (2001).

0304-4076/$ – see front matter
doi:10.1016/j.jeconom.2009.10.035

estimators are, however, designed for models with homoskedastic disturbances.

The homoskedastic assumption may be restrictive in practice. In certain applications, we would expect the variances of the error terms to be different. For instance, consider the analysis of the spatial dependence in the unemployment or crime rates of contiguous states in the United States. As a rate variable is a result of aggregation, heteroskedasticity may be present. In the presence of social interactions, the variance of the aggregated level data will be inflated, the extent depending on the strength and structure of the interactions. In a study of cross-city crime rates, Glaeser et al. (1996) show that the high variance of cross-city crime rates is largely caused by social interactions among individuals. Therefore, the presence of social interactions could complicate the variance structure of aggregated data, especially when social interaction patterns depend not only on the population size in the city, but also on the distribution and composition of the population. LeSage (1999) illustrates how the mean and variance of home selling prices change as we move across observations with different distances from the central business district. More discussions on spatial heteroskedasticity can be found in Anselin (1988).

In this paper, we consider the case when the error terms in the model are independent but with an unknown heteroskedasticity. If variances of the disturbances or the exact structure of heteroskedasticity are known, we may get rid of the heteroskedasticity by some appropriate transformations and then apply the conventional MLE or GMM techniques to the transformed model. However, one may not have accurate information about the nature of the heteroskedasticity in a model and may be unsure of the specific structural form of the variances. With an unknown heteroskedasticity, we would like to know the consequences for various estimators if the SAR model were estimated as if the disturbances were i.i.d. As will be shown without taking into account the heteroskedasticity, the MLE is generally inconsistent. In contrast, the GMME obtained from certain carefully designed moment conditions can be robust against an unknown heteroskedasticity. Furthermore, one may improve the efficiency by constructing optimal weighting for the GMM estimation even when the form of heteroskedasticity is unknown.

Section 2 discusses the possible inconsistency property of the MLE and derives its asymptotic bias for some special case. Robust GMM estimation under unknown heteroskedasticity is considered in Section 3. Its consistency and asymptotic distribution are derived. Section 4 considers the optimal weighting of the robust GMM estimation. Some extensive Monte Carlo studies illustrate possible degrees of bias for the various estimators in finite samples in Section 5. Section 6 presents specification tests on the testing of unknown heteroskedasticity, and some Monte Carlo results on levels of significance and powers of the Hausman-type and Lagrange Multiplier (LM) test statistics. An empirical application on county teenage pregnancy rates is provided in Section 7. Conclusions are drawn in Section 8. The technical details are given in the Appendix.

2. Inconsistency of the MLE in the presence of heteroskedastic disturbances

The model considered is the mixed regressive, spatial autoregressive model

$$Y_n = \lambda_0 W_n Y_n + X_n\beta_0 + \epsilon_n, \tag{1}$$

where X_n is an $n \times k$ matrix of nonstochastic exogenous variables, W_n is an $n \times n$ spatial weights matrix of known constants with zero diagonal elements, and the elements ϵ_{ni}'s of the n-dimensional vector ϵ_n are independent with a mean 0 and variances σ_{ni}^2, $i = 1, \ldots, n$. The spatial effect coefficient λ_0 measures the average influence of neighboring observations on Y_n, which usually lies between $(-1, 1)$ when W_n is row-normalized such that the sum of elements of each row is unity. For a general W_n which is not row-normalized, the λ_0 will usually be assumed to be in a parameter space which guarantees that the determinant of $(I_n - \lambda_0 W_n)$ is positive. There will be more discussion on the parameter space of λ_0 later on. The reduced form of the model is $Y_n = S_n^{-1}X_n\beta_0 + S_n^{-1}\epsilon_n$, where $S_n = I_n - \lambda_0 W_n$.

For the SAR model in (1), under the assumption of i.i.d. $N(0, \sigma_0^2)$ disturbances, the log likelihood for this standard model is

$$\ln L_n(\delta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + \ln|S_n(\lambda)| - \frac{1}{2\sigma^2}\epsilon_n'(\theta)\epsilon_n(\theta), \tag{2}$$

where $\delta = (\lambda, \beta', \sigma^2)$, $\theta = (\lambda, \beta')$, $S_n(\lambda) = I_n - \lambda W_n$, and $\epsilon_n(\theta) = S_n(\lambda)Y_n - X_n\beta$.

Given a λ, (1) becomes a regression equation of $S_n(\lambda)$ on X_n, and, the MLE of β is

$$\widehat{\beta}_n(\lambda) = (X_n'X_n)^{-1}X_n'S_n(\lambda)Y_n \tag{3}$$

and the MLE of σ^2 as $\widehat{\sigma}_n^2(\lambda) = \frac{1}{n}[S_n(\lambda)Y_n - X_n\widehat{\beta}_n(\lambda)]'[S_n(\lambda)Y_n - X_n\widehat{\beta}_n(\lambda)] = \frac{1}{n}Y_n'S_n'(\lambda)M_nS_n(\lambda)Y_n,$, where $M_n = I_n - X_n(X_n'X_n)^{-1}X_n'$.

Then, we can get the concentrated log likelihood function of λ, which is

$$\ln L_n(\lambda) = -\frac{n}{2}(\ln(2\pi) + 1) - \frac{n}{2}\ln\widehat{\sigma}_n^2(\lambda) + \ln|S_n(\lambda)|. \tag{4}$$

The first order condition for the concentrated log likelihood function is

$$\frac{\partial \ln L_n(\lambda)}{\partial\lambda} = \frac{1}{\widehat{\sigma}_n^2(\lambda)}Y_n'W_n'M_nS_n(\lambda)Y_n - \mathrm{tr}(W_nS_n^{-1}(\lambda)). \tag{5}$$

For consistency of the MLE $\widehat{\lambda}_n$, the necessary condition is $\mathrm{plim}_{n\to\infty}\frac{1}{n}\frac{\partial \ln L_n(\lambda_0)}{\partial\lambda} = 0$. However, with heteroskedastic disturbances, this condition may not be satisfied. Consequently, the consistency of the MLE is not guaranteed.

In the presence of heteroskedasticity, at the true parameter λ_0,

$$\begin{aligned}\widehat{\sigma}_n^2(\lambda_0) &= \frac{1}{n}[S_nY_n - X_n\widehat{\beta}_n(\lambda_0)]'[S_nY_n - X_n\widehat{\beta}_n(\lambda_0)] \\ &= \frac{1}{n}\epsilon_n'M_n\epsilon_n = \frac{1}{n}\sum_{i=1}^n\sigma_{ni}^2 + o_p(1).\end{aligned} \tag{6}$$

So, $\widehat{\sigma}_n^2(\lambda_0)$ and the average of σ_{ni}^2, $\overline{\sigma}^2$ are asymptotically equivalent.[3] Let $G_n = W_nS_n^{-1}$. Then, from Eqs. (5) and (6), we have, at λ_0,

$$\begin{aligned}\frac{1}{n}\frac{\partial \ln L_n(\lambda_0)}{\partial\lambda} &= \frac{1}{n}\left[\frac{1}{\widehat{\sigma}_n^2(\lambda_0)}Y_n'W_n'M_nS_nY_n - \mathrm{tr}(W_nS_n^{-1})\right] \\ &= \frac{\frac{1}{n}\epsilon_n'G_n'M_n\epsilon_n}{\frac{1}{n}\epsilon_n'M_n\epsilon_n} + \frac{\frac{1}{n}(X_n\beta_0)'G_n'M_n\epsilon_n}{\frac{1}{n}\epsilon_n'M_n\epsilon_n} - \frac{1}{n}\mathrm{tr}(G_n) \\ &= \frac{\sum_{i=1}^n G_{n,ii}\sigma_{ni}^2}{\sum_{i=1}^n\sigma_{ni}^2} - \overline{G}_n + o_p(1) \\ &= \frac{\frac{1}{n}\sum_{i=1}^n[G_{n,ii} - \overline{G}_n](\sigma_{ni}^2 - \overline{\sigma}^2)}{\overline{\sigma}^2} + o_p(1) \\ &= \frac{\mathrm{COV}(G_{n,ii}, \sigma_{ni}^2)}{\overline{\sigma}^2} + o_p(1),\end{aligned} \tag{7}$$

[3] The asymptotic arguments can follow from the law of large numbers in the Appendix. In this section, we do not provide the rigorous analysis in order to save space.

where $\overline{G}_n = \frac{1}{n}\mathrm{tr}(G_n) = \frac{1}{n}\sum_{i=1}^n G_{n,ii}$. Therefore, the limit of $\frac{1}{n}\frac{\partial \ln L_n(\lambda_0)}{\partial\lambda}$ will be zero if and only if the covariance between the diagonal elements of the matrix G_n, $G_{n,ii}$, $i = 1, \ldots, n$, and the individual variances σ_{ni}^2, $i = 1, \ldots, n$, is zero in the limit. In the heteroskedastic case, this condition will be satisfied if almost all the diagonal elements of the matrix G_n are equal.[4]

It is of interest to see when we would have constant diagonal elements in the G_n matrix for some special cases. Consider a "circular" world where the units are arranged on a circle such that the last unit y_n has neighbors y_1 and y_{n-1}, y_1 has neighbors y_2 and y_n, and so forth.[5] If we assign an equal weight to each neighbor of the same unit, the diagonal elements of the resulting G_n matrix will be constant. The units in a "circular" world can have more neighbors, as long as each unit has the same numbers of neighbors and with half of the neighbors lead and the rest lag, the diagonal elements of the G_n matrix will be the same. Another special case is that W_n is a block-diagonal matrix with an identical submatrix in the diagonal blocks and zeros elsewhere. This corresponds to the group interactions scenario where all the group sizes are equal, and each neighbor of the same unit is assigned equal weight. When these special spatial weights matrices are used, the MLE will still be consistent in the presence of unknown heteroskedasticity. However, for general spatial weights matrices, the consistency is not ensured.

Following the inconsistency of the MLE of λ_0, a consequence is the inconsistency of the MLE of β_0. Because from (3), we have

$$\widehat{\beta}_n(\widehat{\lambda}) = \beta_0 + (\lambda_0 - \widehat{\lambda})(X_n'X_n)^{-1}X_n'G_nX_n\beta_0 + o_p(1), \tag{8}$$

which will not converge to β_0 in the limit if $\widehat{\lambda}$ is not consistent.

Thus, besides the computational burden it entails, the MLE for the SAR model with an unknown heteroskedasticity is inconsistent as long as the diagonal elements of the matrix G_n are not all equal.

Because of the nonlinearity of λ in the concentrated log likelihood function, it is hard to make any general conclusion about the asymptotic bias of $\widehat{\lambda}$. For the asymptotic bias of $\widehat{\beta}_n(\widehat{\lambda})$ from (8), it is $(\lambda_0 - \widehat{\lambda})(X_n'X_n)^{-1}X_n'(G_nX_n\beta_0)$. Thus, given the bias of $\widehat{\lambda}$, the asymptotic bias of $\widehat{\beta}_n(\widehat{\lambda})$ is determined by the term $(X_n'X_n)^{-1}X_n'(G_nX_n\beta_0)$, which is the OLSE of the coefficient in the artificial regression of $G_nX_n\beta_0$ on X_n. Thus, given the bias of $\widehat{\lambda}$, the relative asymptotic bias of $\widehat{\beta}_n(\widehat{\lambda})$ depends on the properties of X_n and W_n. Consider a special case, which is often used in empirical social interaction studies. This is the case of group interactions, where W_n is assumed to be a block-diagonal matrix, and in each block, $W_r = \frac{1}{m_r-1}(l_{m_r}l_{m_r}' - I_{m_r})$, $r = 1, \ldots, R$, where R is the number of groups, m_r is the group size for group r, l_{m_r} is the m_r-dimensional vector of ones, and I_{m_r} is the m_r-dimensional identity matrix. Note that the group sizes are not all equal, and for the asymptotic properties, we let the number of groups R go to infinity while maintaining $\{m_r\}$ is bounded. This interaction pattern means that there are no cross group interactions and a unit is equally affected by all the other members in the same group. A group could be village or a class, etc. This group interaction setting has been studied by Case (1991), Lee (2004, 2007c), among others. Let's assume for all the groups, the x's are i.i.d. with mean μ and variance Σ_x for all observations. In particular, in group r, let $X_{(r)} = (l_{m_r}, z_{(r)})$, $\overline{X}_{(r)} = (1, \bar{z}_{(r)})$, $\mu = (1, \mu_z)$, and $\Sigma_x = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_z \end{pmatrix}$, where $z_{(r)} = (z_{1r}', \ldots, z_{m_r,r}')'$ is the matrix of regressors excluding the intercept term and $\bar{z}_{(r)} = \frac{1}{m_r}\sum_{i=1}^{m_r} z_{ir}$. Then after some calculations we can get the equation in Box I and $(X_n'X_n)^{-1} = \left[\sum_{r=1}^R \begin{pmatrix} m_r & \sum_{i=1}^{m_r} z_{ir} \\ \sum_{i=1}^{m_r} z_{ir}' & \sum_{i=1}^{m_r} z_{ir}'z_{ir} \end{pmatrix}\right]^{-1}$. Note that

$$\lim_{R\to\infty}\left\{E\left(\frac{1}{n}X_n'G_nX_n\right) - \left[\frac{\mu'\mu}{1-\lambda_0} + \frac{1}{1-\lambda_0}\frac{R}{n}\Sigma_x - \frac{1}{n}\sum_{r=1}^R\left(\frac{m_r-1}{m_r-1+\lambda_0}\right)\Sigma_x\right]\right\} = 0 \tag{9}$$

and $(E(\frac{1}{n}X_n'X_n))^{-1} = \begin{pmatrix} 1+\mu_z\Sigma_z^{-1}\mu_z' & -\mu_z\Sigma_z^{-1} \\ -\Sigma_z^{-1}\mu_z' & \Sigma_z^{-1} \end{pmatrix}$. Thus, we can get the equation in Box II.

Therefore, in this group interaction setting with randomly distributed x's, if all the elements in x except the constant term have a zero mean, i.e., $\mu_z = 0$, the relative asymptotic bias of the intercept β_{10} will be $\frac{1}{1-\lambda_0}$ times the bias of the MLE of λ_0. Also, except the intercept β_{10}, the MLE for all the other β_0's have the same magnitude of relative asymptotic bias, which is the term $(\frac{R}{n}\frac{1}{1-\lambda_0} - \frac{1}{n}\sum_{r=1}^R\frac{m_r-1}{m_r-1+\lambda_0})$ times the bias of the MLE of λ_0. As $(\frac{R}{n}\frac{1}{1-\lambda_0} - \frac{1}{n}\sum_{r=1}^R\frac{m_r-1}{m_r-1+\lambda_0})$ is less than $\frac{R}{n}\frac{1}{(1-\lambda_0)}$ and $\frac{n}{R}$ is the average group size, the relative asymptotic bias of the intercept will be larger than those of the other regression coefficients in β_0. In particular, if the average group size is moderately large, the biases of the coefficients of regressors (rather than the intercept term) can be small.

The preceding paragraph has considered the asymptotic bias of the MLE under heteroskedasticity. Likewise, the MOM estimator suggested by Kelejian and Prucha (1999) is not consistent in the presence of unknown heteroskedasticity since the moment conditions they proposed do not have a zero mean at the true parameters. The following section discusses the feature of GMM estimation and possible robust estimation.

3. GMM estimation against unknown heteroskedasticity

3.1. A brief overview

The consistency of the GMME in Lee (2001, 2007a) with P_n from $\mathcal{P}_{1n}$ which is a class of constant $n \times n$ matrices P_n with $\mathrm{tr}(P_n) = 0$; or $\mathcal{P}_{2n}$, a subclass of $\mathcal{P}_{1n}$ with $Diag(P_n) = 0$, is based on the fundamental moment property that $E(\epsilon_n'P_n\epsilon_n) = 0$. If the ϵ_{ni}'s have heteroskedastic variances, $E(\epsilon_n'P_n\epsilon_n) = \mathrm{tr}[P_nE(\epsilon_n\epsilon_n')]$ will not necessarily be zero if P_n is from $\mathcal{P}_{1n} \setminus \mathcal{P}_{2n}$. Consider the i^{th} component of $P_n\epsilon_n$, $\sum_{j=1}^n P_{n,ij}\epsilon_{nj}$, which is clearly correlated with the corresponding component ϵ_{ni} of ϵ_n if $P_{n,ii} \neq 0$. With homoskedastic disturbances, the correlations of $P_n\epsilon_n$ and ϵ_n can be canceled out as long as $\mathrm{tr}(P_n) = 0$. In the presence of heteroskedastic error terms, letting $\mathrm{tr}(P_n) = 0$ may not guarantee the correlations between each component of $P_n\epsilon_n$ and the corresponding components of ϵ_n are exactly canceled out. Therefore, when P_n is from $\mathcal{P}_{1n}$ but not $\mathcal{P}_{2n}$, $P_n\epsilon_n$ may be correlated with ϵ_n and thus loses its validity as an instrumental variable (IV) vector. In contrast, if P_n is from $\mathcal{P}_{2n}$, $E(\epsilon_n'P_n\epsilon_n) = 0$ is true since $\mathrm{tr}[P_nE(\epsilon_n\epsilon_n')] = \mathrm{tr}[Diag(P_n)E(\epsilon_n\epsilon_n')] = 0$. We successfully maintain the uncorrelation between $P_n\epsilon_n$ and ϵ_n by excluding each component of ϵ_n from the corresponding term of $P_n\epsilon_n$. Thus, in the presence of unknown heteroskedasticity, the GMM estimation for the SAR model will be based on $\mathcal{P}_{2n}$ but not $\mathcal{P}_{1n}$. Lee (2001) has noticed this possible robust property of using quadratic moments with the matrix P_n's from $\mathcal{P}_{2n}$ but has not provided any rigorous theory. This paper follows up on this observation and will provide a rigorous theory and investigate finite sample properties in Monte Carlo studies for the SAR model.

[4] It will be zero if ϵ_{ni}'s are i.i.d., since in that case $\sigma_{ni}^2 = \overline{\sigma}^2$, Eq. (7) will converge to zero regardless of the diagonal elements of the matrix G_n.

[5] Kelejian and Prucha (1999) use this type of weights matrix in their Monte Carlo study.

$$X_n'G_nX_n = \sum_{r=1}^{R}\begin{pmatrix} \frac{m_r}{1-\lambda_0} & \frac{m_r}{1-\lambda_0}\bar{z}_{(r)} \\ \frac{m_r}{1-\lambda_0}\bar{z}_{(r)}' & \frac{m_r}{1-\lambda_0}\bar{z}_{(r)}'\bar{z}_{(r)} - \frac{1}{m_r-1+\lambda_0}\sum_{i=1}^{m_r}(z_{ir}-\bar{z}_{(r)})'(z_{ir}-\bar{z}_{(r)}) \end{pmatrix}$$

Box I.

$$\lim_{R\to\infty}(E(X_n'X_n))^{-1}E(X_n'G_nX_n)$$

$$= \lim_{R\to\infty}\begin{pmatrix} \frac{1}{1-\lambda_0} & \left(\frac{1}{1-\lambda_0} - \frac{R}{n}\frac{1}{1-\lambda_0} + \frac{1}{n}\sum_{r=1}^{R}\frac{m_r-1}{m_r-1+\lambda_0}\right)\mu_z \\ 0 & \left(\frac{R}{n}\frac{1}{1-\lambda_0} - \frac{1}{n}\sum_{r=1}^{R}\frac{m_r-1}{m_r-1+\lambda_0}\right)I_z \end{pmatrix},$$

where I_z is the $(k-1)$-dimensional identity matrix.

Box II.

The MOM method suggested in Kelejian and Prucha (1999) uses essentially the two moments $\epsilon_n'W_n\epsilon_n$ and $\epsilon_n'(W_n'W_n - \frac{\mathrm{tr}(W_n'W_n)}{n}I_n)\epsilon_n$. While W_n has a zero diagonal and the moment $\epsilon_n'W_n\epsilon_n$ is robust against unknown heteroskedasticity, the other moment is not, as the diagonal of $[W_n'W_n - \frac{\mathrm{tr}(W_n'W_n)}{n}I_n]$ may not be zero. A robust version of this MOM method may replace the second moment function by $\epsilon_n'(W_n'W_n - \mathrm{Diag}(W_n'W_n))\epsilon_n$, where $\mathrm{Diag}(A)$ for a square matrix A denotes the diagonal matrix formed by the diagonal elements of A.[6]

3.2. Robust GMM estimation

To analyze rigorously the robust property of GMM estimation with $\mathcal{P}_{2n}$, we adopt most regularity assumptions for GMM estimation in Lee (2001, 2007a) with proper modifications to fit into the heteroskedasticity setting. Interested readers may refer to Lee (2001, 2007a) for detailed discussions on related assumptions for the i.i.d. disturbances case.[7]

Assumption 1. The ϵ_{ni}'s are independent $(0, \sigma_{ni}^2)$ with finite moments larger than the fourth order such that $E|\epsilon_{ni}|^{4+\eta}$ for some $\eta > 0$ are uniformly bounded for all n and i.

This assumption implies the uniform boundedness of the variances σ_{ni}^2, the third moments, $\mu_{ni,3}$ and the fourth moments $\mu_{ni,4}$ of ϵ_{ni} are also uniformly bounded for all n and i.

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

Assumption 3. The spatial weights matrices $\{W_n\}$ and the matrix $\{S_n^{-1}\}$ are uniformly bounded in absolute value in both row and column sums.

This uniform boundedness assumption limits the spatial dependences among the units to a tractable degree and is originated by Kelejian and Prucha (1999). It rules out the unit root case (in time series as a special case).

Let Q_n be an $n \times k^*$ matrix, where $k^* \geq k+1$, of IV's constructed from X_n and W_n, such as X_n, W_nX_n, $W_n^2X_n$, etc. The moment functions corresponding to the orthogonality conditions of X_n and ϵ_n are $Q_n'\epsilon_n(\theta)$. However, these linear moments reflect only the information in the deterministic part of W_nY_n, leaving those in the stochastic part unexplored. This can be seen from the reduced form of the model. If $\parallel \lambda W_n \parallel < 1$ where $\parallel\cdot\parallel$ is a matrix norm, we have $(I_n - \lambda W_n)^{-1} = I_n + \lambda W_n + \lambda^2 W_n^2 + \cdots$, and the reduced-form equation becomes

$$\begin{aligned} Y_n &= S_n^{-1}X_n\beta_0 + S_n^{-1}\epsilon_n \\ &= X_n\beta_0 + \lambda_0 W_nX_n\beta_0 + \lambda_0^2 W_n^2X_n\beta_0 + \cdots + S_n^{-1}\epsilon_n. \end{aligned} \tag{10}$$

It is obvious from (10) that forming IV vectors from functions of W_n and X_n focuses only on the information in the nonstochastic part $E(W_nY_n|X_n)$ of W_nY_n. Lee (2007a) suggests the use of the moment conditions $(P_{jn}\epsilon_n(\theta))'\epsilon_n(\theta)$ in addition to $Q_n'\epsilon_n(\theta)$. These additional moments capture the correlations across the spatial units. They serve as the IV for $G_n\epsilon_n$, the other component of W_nY_n.[8] The matrices in $\mathcal{P}_{2n}$ (more generally, $\mathcal{P}_{1n}$) are assumed to have a similar uniform boundedness property as in W_n and S_n^{-1}.

Assumption 4. The matrices P_{jn}'s with $\mathrm{Diag}(P_{jn}) = 0$ are uniformly bounded in both row and column sums, and elements of Q_n are uniformly bounded.

The set of moment functions for the GMM estimation is as follows

$$\begin{aligned} g_n(\theta) &= (P_{1n}\epsilon_n(\theta), \ldots, P_{mn}\epsilon_n(\theta), Q_n)'\epsilon_n(\theta) \\ &= (\epsilon_n'(\theta)P_{1n}\epsilon_n(\theta), \ldots, \epsilon_n'(\theta)P_{mn}\epsilon_n(\theta), \epsilon_n'(\theta)Q_n)'. \end{aligned} \tag{11}$$

Denote $\mathrm{Var}(g_n(\theta)) = \Omega_n$ and, for any square matrix A_n, $A_n^s = A_n + A_n'$ is the sum of A_n and its transpose. Let $\Sigma_n = \mathrm{Diag}\{\sigma_{n1}^2, \ldots, \sigma_{nn}^2\}$, where $\sigma_{ni}^2 = E(\epsilon_{ni}^2)$, $i = 1, \ldots, n$.

Assumption 5. Either (a) $\lim_{n\to\infty}\frac{1}{n}Q_n'(G_nX_n\beta_0, X_n)$ has the full rank $(k+1)$, or

(b) $\lim_{n\to\infty}\frac{1}{n}Q_n'X_n$ has the full rank k, $\lim_{n\to\infty}\frac{1}{n}\mathrm{tr}(\Sigma_nG_n^sP_{jn}) \neq 0$ for some j, and $\lim_{n\to\infty}\frac{1}{n}(\mathrm{tr}(\Sigma_nG_n^sP_{1n}), \ldots, \mathrm{tr}(\Sigma_nG_n^sP_{mn}))'$ and $\lim_{n\to\infty}\frac{1}{n}(\mathrm{tr}(\Sigma_nG_n'P_{1n}G_n), \ldots, \mathrm{tr}(\Sigma_nG_n'P_{mn}G_n))'$ are linearly independent.

6 After the completion of this paper, we realize that Kelejian and Prucha (2010) has extended their approach in Kelejian and Prucha (1999) to cover the estimation of the SAR model with spatial SAR process with unknown heteroskedaticity. Their approach for the SAR disturbance process has used the two moments $\widehat{\epsilon}_n'W_n\widehat{\epsilon}_n$ and $\widehat{\epsilon}_n'(W_n'W_n - \mathrm{Diag}(W_n'W_n))\widehat{\epsilon}_n$, where $\widehat{\epsilon}_n$ is an estimated residual. For the SAR regression equation, they suggest the use of generalized two stage least squares.

7 In this paper, we do not consider the large group interactions case so as to simplify the presentation.

8 Note that $W_nY_n = G_nX_n\beta_0 + G_n\epsilon_n$.

This assumption assures the identification of θ_0 from the moment equations $E(g_n(\theta_0)) = 0$ for a sufficiently large n. If $G_n X_n \beta_0$ and X_n are linearly dependent, which includes the case when all exogenous variables X_n are irrelevant, the additional moments in (b) will help to identify θ_0.

And the parameter space Θ of θ is assumed to have the following property:

Assumption 6. The θ_0 is in the interior of the parameter space Θ, which is a bounded subset of R^{k+1}.[9]

The parameter space of λ is usually taken to be $(-1, 1)$ when W_n is a row-normalized matrix. For the cases in which W_n is not normalized but its eigenvalues are real with its largest eigenvalue $\mu_{n,max} > 0$ and its smallest eigenvalue $\mu_{n,min} < 0$, the parameter space can be the interval $(-\frac{1}{|\mu_{n,min}|}, \frac{1}{|\mu_{n,max}|})$ (Anselin, 1988). Kelejian and Prucha (2010) allow complex eigenvalues for W_n and suggest the parameter space $(-\frac{1}{\tau_n}, \frac{1}{\tau_n})$ where τ_n is the spectral radius of W_n. These parameter spaces are designed to guarantee that the determinant of $(I_n - \lambda W_n)$ is positive. Kelejian and Prucha (2010) also allow the parameters, including λ, to depend on n as they are the resulted parameters after W_n being rescaled by a normalized factor which depends on n. If W_n is rescaled by the division with τ_n, the coefficient $\lambda_n (= \tau_n \lambda)$ can then be taken as $(-1, 1)$. For our GMM estimation, one does not need to impose a specific parameter space for the minimization of the GMM objective function because it is simply a polynomial function of θ. So the regularity condition in the preceding assumption on the parameter space is solely for the theoretical purpose of proving consistency of the GMM estimator. As we do not emphasize on any scale normalization of W_n, we simply consider θ_0 being a constant parameter vector.

The following proposition concerns about the asymptotic property of a GMM estimator in the general Hansen GMM setting with a linear transformation $a_n g_n(\theta)$ of the moment functions $g_n(\theta)$, where a_n is a matrix with a full row rank greater than or equal to the number of parameters in θ. The $a_n' a_n$ in the GMM objective function $g_n'(\theta) a_n' a_n g_n(\theta)$ is a nonnegative definite matrix, which represents a weighting matrix in this distance function. This general framework motivates the issue of optimum weighting matrix. Proposition 1 is a generalization of Proposition 2.1 in Lee (2001) to the heteroskedastic case.

Proposition 1. *Suppose that* $diag(P_{jn}) = 0$ *for* $j = 1, \ldots, m$, *and* Q_n *is a* $n \times k^*$ *IV matrix so that* $\lim_{n\to\infty} a_n E(g_n(\theta)) = 0$ *has a unique root at* θ_0 *in* Θ. *Then, under the stated Assumptions 1–6 and that* $\lim_{n\to\infty} \frac{1}{n} a_n D_n$ *exists and has the full rank* $(k+1)$, *the RGMME* $\widehat{\theta}_n$ *derived from* $\min_{\theta\in\Theta} g_n'(\theta) a_n' a_n g_n(\theta)$ *is a consistent estimator of* θ_0, *and* $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{D} N(0, \Gamma)$, *where*

$$\Gamma = \lim_{n\to\infty} \frac{1}{n} (D_n' a_n' a_n D_n)^{-1} D_n' a_n' a_n \Omega_n a_n' a_n D_n (D_n' a_n' a_n D_n)^{-1}, \tag{12}$$

$$\Omega_n = \mathrm{Var}(g_n(\theta_0))$$
$$= \begin{pmatrix} \mathrm{tr}[\Sigma_n P_{1n}(\Sigma_n P_{1n})^s] & \mathrm{tr}[\Sigma_n P_{1n}(\Sigma_n P_{2n})^s] & \ldots & 0 \\ \mathrm{tr}[\Sigma_n P_{2n}(\Sigma_n P_{1n})^s] & \mathrm{tr}[\Sigma_n P_{2n}(\Sigma_n P_{2n})^s] & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & Q_n' \Sigma_n Q_n \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^n \sum_{j=1}^n P_{1n,ij}(P_{1n,ij} + P_{1n,ji})\sigma_{ni}^2 \sigma_{nj}^2 & \ldots & 0 \\ \sum_{i=1}^n \sum_{j=1}^n P_{2n,ij}(P_{1n,ij} + P_{1n,ji})\sigma_{ni}^2 \sigma_{nj}^2 & \ldots & 0 \\ \vdots & & \vdots \\ 0 & \ldots & Q_n' \Sigma_n Q_n \end{pmatrix}, \tag{13}$$

$$D_n = -\frac{\partial E(g_n(\theta_0))}{\partial \theta'} = \begin{pmatrix} \mathrm{tr}(\Sigma_n P_{1n}^s G_n) & 0 \\ \vdots & \vdots \\ \mathrm{tr}(\Sigma_n P_{mn}^s G_n) & 0 \\ Q_n' G_n X_n \beta_0 & Q_n' X_n \end{pmatrix}. \tag{14}$$

The proof is similar to the i.i.d. case once we realize that the uniform convergence of sample averages of relevant moment functions can hold in the presence of heteroskedasticity and the central limit theorem for linear-quadratic forms by Kelejian and Prucha (1999) allows for heteroskedastic disturbances. The details of the proofs of all propositions are given in the Appendix.

From Proposition 1, the RGMME obtained from an arbitrary weighting matrix (with moment functions constructed from $\mathcal{P}_{2n}$) can be consistent (robust) against an unknown heteroskedasticity. In particular, if we construct the optimal GMM as in the i.i.d. case without taking into account the presence of heteroskedasticity, i.e., if we replace the weighting matrix $a_n' a_n$ by $(\widetilde{\Omega}_n)^{-1}$, where $\widetilde{\Omega}_n$ is an estimator of Ω_n based on an initial estimate of θ as if ϵ_{ni}'s were *i.i.d.*, the resulting GMME will still be consistent and asymptotically normal. However, the correct asymptotic covariance matrix will not be the one, $(\lim_{n\to\infty} \frac{1}{n} D_n' \Omega_n^{-1} D_n)^{-1}$, in the i.i.d. case. Instead, it will take the messier form of

$$\lim_{n\to\infty} \frac{1}{n} (D_n' \overline{\Omega}_n^{-1} D_n)^{-1} D_n' \overline{\Omega}_n^{-1} \Omega_n \overline{\Omega}_n^{-1} D_n (D_n' \overline{\Omega}_n^{-1} D_n)^{-1}, \tag{15}$$

where $\frac{1}{n}\overline{\Omega}_n$ is the probability limit of $\frac{1}{n}\widetilde{\Omega}_n$, whose value depends on the specific formula of $\frac{1}{n}\widetilde{\Omega}_n$. Furthermore, as a special case of the GMM estimation, the 2SLS estimation with $a_n = (0, (Q_n' Q_n)^{-1/2})$ and $a_n g_n(\theta) = (Q_n' Q_n)^{-1/2} Q_n' \epsilon_n(\theta)$ can be consistent from Proposition 1.[10] It can also serve as the initial consistent estimator in our GMM estimation.

In order to make asymptotically valid inferences from the RGMME, we need to find a consistent estimator of the asymptotic variance as given in (12). As in White (1980), we can consistently estimate the part $\frac{1}{n} Q_n' \sum_n Q_n$ in Ω_n in (13) without being able to estimate $\sum_n$, which involves n unknowns, consistently. The tricky part is the estimation of the other elements associated with the quadratic moment functions. Those elements consist of $\frac{1}{n}$ times a sum of n^2 terms. However, the uniform boundedness property of P_n ensures the convergence of these sums. The following proposition can be used to provide a consistent estimator for the covariance matrix Ω_n.

Proposition 2. *Under the assumed regularity conditions,* $\frac{1}{n}(\widehat{D}_n - D_n) = o_P(1)$ *and* $\frac{1}{n}(\widehat{\Omega}_n - \Omega_n) = o_P(1)$, *where* $\frac{1}{n}\widehat{D}_n$ *and* $\frac{1}{n}\widehat{\Omega}_n$ *are, respectively, estimators of* $\frac{1}{n} D_n$ *and* $\frac{1}{n}\Omega_n$ *with* θ_0 *replaced by a consistent initial estimator* $\widehat{\theta}_n$ *and* Σ_n *by* $\widehat{\Sigma}_n$, *where* $\widehat{\Sigma}_n = \mathrm{Diag}\{\widehat{\epsilon}_{n1}^2, \ldots, \widehat{\epsilon}_{nn}^2\}$ *and* $\widehat{\epsilon}_{ni}$*'s are the residuals of the model with* θ_0 *estimated by* $\widehat{\theta}_n$.

[9] For nonlinear extremum estimation methods, such as the ML method, compactness on the parameter space Θ is usually needed in order to apply some uniform laws of large numbers to demonstrate consistency of extremum estimates (Amemiya, 1985). However, for our GMM approach with linear and quadratic functions, θ appears nonlinearly in moment conditions in terms of polynomials. For $S_n^{-1}(\lambda)$, only its value evaluated at consistent estimates of λ_0 will be used. So for asymptotic analysis, the boundedness of Θ will be sufficient.

[10] Assumption 5(a) is crucial for the consistency of the 2SLSE.

4. "Optimal" RGMM estimator

From the preceding section, we see that the consistency of the RGMME is, in general, not affected by the choice of the weighting matrix, but its asymptotic variance is. By using a "wrong" weighting matrix, we'll still get the consistent estimator but the estimator may not be efficient. By the generalized Schwartz inequality, the optimal weighting matrix for the GMM estimation with the moment functions $g_n(\theta)$ is Ω_n^{-1}, the inverse of the covariance matrix for the moment functions $g_n(\theta_0)$. Proposition 3 shows that, with a consistent estimator $\widehat{\Omega}_n^{-1}$, the feasible "optimal" RGMME obtained from $\min_{\theta\in\Theta} g_n^{'}(\theta)\widehat{\Omega}_n^{-1}g_n(\theta)$ will be consistent and asymptotically normal with variance $(\lim_{n\to\infty}\frac{1}{n}D_n'\Omega_n^{-1}D_n)^{-1}$.

The variance matrix Ω_n is assumed to satisfy some conventional regularity conditions.

Assumption 7. The $\lim_{n\to\infty}\frac{1}{n}\Omega_n$ exists and is nonsingular.

Proposition 3. *Suppose that $(\frac{1}{n}\widehat{\Omega}_n)^{-1}-(\frac{1}{n}\Omega_n)^{-1}=o_p(1)$, then the feasible "optimal" ORGMME $\widehat{\theta}_{o,n}$ derived from $\min_{\theta\in\Theta} g_n^{'}(\theta)\widehat{\Omega}_n^{-1}g_n(\theta)$ has the asymptotic distribution*

$$\sqrt{n}(\widehat{\theta}_{o,n}-\theta_0)\xrightarrow{D} N\left(0,\left(\lim_{n\to\infty}\frac{1}{n}D_n'\Omega_n^{-1}D_n\right)^{-1}\right). \tag{16}$$

Similarly, a consistent estimator for the asymptotic covariance matrix is $(\frac{1}{n}\widehat{D}_n'\widehat{\Omega}_n^{-1}\widehat{D}_n)^{-1}$.

The "optimal" ORGMME here refers to the RGMME based on the optimal weighting with specified moment functions.[11] In the i.i.d. disturbances case, the best choices P_n from $\mathcal{P}_{2n}$ and Q_n are available, which are respectively known as $(G_n-\mathrm{Diag}(G_n))$ and $(G_nX_n\beta_0, X_n)$. However, for the case with an unknown heteroskedasticity, the best selection of P_n and Q_n may not be available. This is so because

$$D_n=\begin{pmatrix}\mathrm{tr}(P_{1n}^sG_n\Sigma_n) & 0\\ \vdots & \vdots\\ \mathrm{tr}(P_{mn}^sG_n\Sigma_n) & 0\\ Q_n^{'}G_nX_n\beta_0 & Q_n^{'}X_n\end{pmatrix}$$

and

$$\Omega_n=\begin{pmatrix}\mathrm{tr}(\Sigma_nP_{1n}(\Sigma_nP_{1n})^s) & \mathrm{tr}(\Sigma_nP_{1n}(\Sigma_nP_{2n})^s) & \dots & 0\\ \mathrm{tr}(\Sigma_nP_{2n}(\Sigma_nP_{1n})^s) & \mathrm{tr}(\Sigma_nP_{2n}(\Sigma_nP_{2n})^s) & \dots & 0\\ \vdots & \vdots & & \vdots\\ \mathrm{tr}(\Sigma_nP_{mn}(\Sigma_nP_{1n})^s) & \mathrm{tr}(\Sigma_nP_{mn}(\Sigma_nP_{2n})^s) & \dots & 0\\ 0 & 0 & \dots & Q_n^{'}\Sigma_nQ_n\end{pmatrix} \tag{17}$$

involve the unknown Σ_n. If a best selection were available, they would involve the matrix Σ_n but the latter has an unknown form. In practice, the selection of consistently estimated $(G_n-\mathrm{Diag}(G_n))$ and $(G_nX_n\beta_0, X_n)$ might be a desirable strategy.

Remark. The results in Propositions 1 and 3 are derived for the spatial scenario where each of the spatial units interacts with only a few neighboring ones. This is the typical case in spatial models. However, some models with social interactions, in particular, involving all members in a group setting, involve large group interactions. The large group interactions case has been studied in Lee (2004) for the ML estimation, and Lee (2007c) for a conditional ML approach. For the GMM estimation, it is in Lee (2007a) for the SAR model with homoskedastic disturbances. To simplify presentations, we have not considered the large group interactions case in this paper. However, it will be of interest to have some remarks on this scenario.

In the large group interactions scenario (Lee, 2004, 2007b,c), a spatial unit may be influenced by many neighboring units, but each of its neighbors' influence will be uniformly small in the sense that elements of $W_n=(w_{n,ij})$ are of an order $O(\frac{1}{h_n})$ uniformly in all n, i and j, where $h_n\to\infty$ as $n\to\infty$. Similar results of Propositions 1 and 3 can hold with some proper modifications and additions of the assumed regularity conditions. For the large group interactions case, while $h_n\to\infty$, it shall be assumed that $\lim_{n\to\infty}\frac{h_n}{n}=0$ in order to obtain consistent estimates. Assumption 4 needs to be strengthened in that elements of P_{jn}'s are of order $O(\frac{1}{h_n})$ uniformly in i, j and n so that their magnitudes are compatible with those of elements of W_n. With Assumption 5(a) in addition to the (modified) Assumptions 1–4, the results in Proposition 1 will be valid. The results in Proposition 3 will also be valid if Assumption 6 is replaced by that $\lim_{n\to\infty}\frac{h_n}{n}\Omega_n$ exists and nonsingular. Note that under Assumption 5(a), the quadratic moments will be dominated by the linear moments in the GMM estimation and the GMM estimates will be asymptotically equivalent to the 2SLS estimates under the large group interactions (Lee, 2007b).

However, when Assumption 5(a) fails in that $G_nX_n\beta_0$ and X_n are linearly dependent, the quadratic moments will be useful. When $G_nX_n\beta_0$ and X_n are multicollinear, there would be no (extra) IV variable available for W_nY_n or linear moments. Then λ_0 can only be estimated via the quadratic moments under the modified Assumption 5(b): $\lim_{n\to\infty}\frac{h_n}{n}\mathrm{tr}(\Sigma_nG_n^sP_{jn})\neq 0$ for some j, and $\lim_{n\to\infty}[\frac{h_n}{n}\mathrm{tr}(\Sigma_nG_n^sP_{1n}),\dots,\frac{h_n}{n}\mathrm{tr}(\Sigma_nG_n^sP_{mn})]'$ and $\lim_{n\to\infty}[\frac{h_n}{n}\mathrm{tr}(\Sigma_nG_n'P_{1n}G_n),\dots,\frac{h_n}{n}\mathrm{tr}(\Sigma_nG_n'P_{mn}G_n)]'$ are linearly independent. The divergent rate of h_n to infinity shall satisfy the condition $\lim_{n\to\infty}\frac{h_n^{1+\frac{2}{\delta}}}{n}=0$ for some $\delta>0$ such that $E|\epsilon_{n,i}|^{4+2\delta}$ are uniformly bounded in all n and i. This strengthened condition is needed in order to apply the generalized CLT for linear and quadratic form in Lee (2004). For this case, while the GMM estimates can be consistent, their rates of convergence will be of the order $O(\sqrt{\frac{n}{h_n}})$, which is lower than the $\sqrt{n}$ order of the case without multicollinearity. Interested readers can consult Lee (2007b) for more details.

5. Monte Carlo study

Some Monte Carlo experiments are designed to study the finite sample properties of the various robust and non-robust estimators. We focus on the case of group interactions. The data generating process is as follows. There are two regressors in addition to the intercept term, which are generated as $x_{ir,1}\sim N(3,1)$ and $x_{ir,2}\sim U(-1,2)$. The size of each group is determined by a uniform $U(3,20)$ variable (round to the closest integer), so the mean group size is about 11. The error terms are normally distributed with a mean of zero and their variances vary across groups. We consider several variance structures with special attention on this particular design: for each group, if the group size is greater than 10, then the variance is constructed to be the same as group size, otherwise, the variance is the square of the inverse of the group size (V-D1). This design V-D1 emphasizes a nonlinear variance structure. The variance function is decreasing and then increasing. Another simpler variance design assumes that the variance is the inverse of group size (V-D2). For the purpose of comparison,

[11] If the P_n and Q_n used involve the unknown parameters λ_0 and β_0, the feasible RGMM estimation will be carried out with λ_0 and β_0 replaced by some initial consistent estimators $\widehat{\lambda}$, $\widehat{\beta}$. The resulting feasible RGMME will have the same limiting distribution. The proof is similar to the i.i.d. case thus is omitted here. Details can be found in Proposition 2.3 in Lee (2001).

the corresponding baseline homoskedastic case has disturbances being i.i.d. $N(0, \overline{\sigma}^2)$, where $\overline{\sigma}^2$ is the mean of the variances of the heteroskedastic errors.

For each of the variance designs, several sets of true parameters are considered. Parameter design 1 (P-D1) has $\theta_0 = (\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.2, 0.8, 0.2, 1.5)$, and design 2 (P-D2) has $\theta_0 = (\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.2, 0.2, 0.2, 0.1)$. The stochastic part of the model with P-D2 becomes relatively more dominant than that of P-D1, since the deterministic regression part of the model has the smaller coefficients on the X_n's. We expect that it would be difficult to deal with P-D2 by the 2SLS approach as its regressors have much smaller effects on Y_n. In addition for $\lambda_0 = 0.2$, we also consider a stronger interaction effect model with $\lambda_0 = 0.6$. The parameter design P-D3 has $\theta_0 = (\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.6, 0.8, 0.2, 1.5)$, and P-D4 has $\theta_0 = (\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.6, 0.2, 0.2, 0.1)$.[12]

The models are estimated by the method of maximum likelihood (ML); the non-robust GMM (GMM) with $P_n = (G_n - \frac{\mathrm{tr}(G_n)}{n} I_n)$ and IV matrix $(G_n X_n \beta, X_n)$; the robust GMM (RGMM) with $P_n = (G_n - Diag(G_n))$ and IV matrix $(G_n X_n \beta, X_n)$.[13] Both the GMM and RGMM approaches will require an initial estimate in the evaluation of G_n (and β in $G_n X_n \beta$). The initial estimate used can be from a simple 2SLS or a simple first step GMM. The simple first step GMM (SGMM) uses $P_n = W_n$ and the linearly independent columns of $(W_n X_n, X_n)$ as IV's without a weighting matrix. For the simple 2SLS (2SLS), the IV's used are simply the linearly independent columns of $(W_n X_n, X_n)$. Also, for the weighting matrices in the GMM and RGMM approaches, we use the variance formulas for the i.i.d. case. For the RGMM approach, the optimal weighting based on the robust variance formula under an unknown heteroskedasticity will also be considered, which is the ORGMM. When the IV matrix $W_n^2 X_n$ in addition to $(W_n X_n, X_n)$ are used in a 2SLS estimation, it is noted as an 2SLS-2 estimation. The feasible best 2SLS with the IV matrix $(G_n X_n \beta, X_n)$, evaluated at the simple 2SLSE, will be denoted by B2SLS. For the feasible GMM and RGMM, the SGMME is usually used as the initial estimate of G_n. When the simple 2SLSE is used instead, the corresponding approaches will be denoted as GMM(2sl) and RGMM(2sl).

For each case, the results reported are based on 1000 Monte Carlo replications. The numbers of groups R are 100 and 200.[14] For the estimates of each coefficient, we report the empirical mean (Mean), the corresponding bias (Bias), the empirical standard error (SD), and the root mean square error (RMSE).

Table 1 summarizes the results from V-D1 with P-D1. The case with small coefficients of β_0's in P-D2 is reported in Table 2. The estimates reported in these two tables focus on the MLE, non-robust GMME, RGMME, ORGMME, and 2SLSE. We compare the finite sample biases of these robust and non-robust estimates, and their relative efficiency in terms of SD and RMSE. Table 3 supplements the results in Tables 1 and 2 with additional estimators, such as the 2SLS-2, B2SLS, SGMM, GMM(2sl) and RGMM(2sl) estimators, for the purpose of comparison. To economize the presentation, only results for $R = 100$ are reported in Tables 3–5. Table 4 presents the results with P-D3 and P-D4, where $\lambda_0 = 0.6$. Results for the variance design V-D2 with the four parameter sets are reported in Table 5. The salient features of results for various estimators are summarized in the following list:

- For the i.i.d. disturbances case, the MLE has some biases in λ_0 and the intercept term β_{10} when $R = 100$. These biases become small when R increases to 200. With heteroskedastic disturbances, the MLE can be biased in λ_0 and β_{10} even in the large sample $R = 200$. The bias of the estimate of λ_0 is downward. However, those biases are not statistically significant even with $R = 200$. The estimate of the intercept term is biased upward. The estimates of the regression coefficients β_{20} and β_{30} are unbiased even for the heteroskedastic cases. These patterns hold in Tables 1 and 2 for both P-D1 and P-D2 with large or small coefficients β_0's for V-D1. The features of the biases of the MLE of λ_0 hold with P-D3 and P-D4 in Table 4 under the same design V-D1.

With V-D2 (and all P-D1, P-D2, P-D3, and P-D4) in Table 5, the MLE's are essentially unbiased for all the parameters, even when there are heteroskedastic disturbances.

- In terms of bias, the GMME has similar patterns as the MLE. In terms of magnitudes of the biases, some may be slightly better than those of the MLE but are mostly similar.
- For the RGMM, the RGMME's are essentially unbiased for all the cases (in Tables 1, 2, 4 and 5).
- The 2SLSE's are consistent in theory. However, its finite sample performance in terms of bias can vary, depending on the pattern of variances of the disturbances and the parameter values. With P-D2 and P-D4 under V-D1, where β_0's are small, the 2SLSE's for λ_0 and β_{10} can have large biases even for $R = 200$ (in Tables 2 and 4). These are also accompanied by relatively large SD's. This is so regardless of whether the disturbances are i.i.d. or heteroskedastic. For the other parameter designs with larger β_0's (P-D1 in Table 1, P-D3 in Table 4 or V-D2 in Table 5), the performance of the 2SLSE's in terms of bias is satisfactory. This 2SLS uses $(W_n X_n, X_n)$ as IV's. For the design P-D2 with V-D1, the 2SLS-2 uses additional IV's $W_n^2 X_n$ may reduce the bias only a little in Table 3.
- The 2SLSE's for λ_0 and β_{10} have the largest SD and RMSE compared with those of the MLE's and the various GMME's (under V-D1 in Tables 1, 2 and 4, and under V-D2 in Table 5, for all parameter designs). With the additional IV's $W_n^2 X_n$ in 2SLS-2 (in Table 3), the SD and RMSE can be slightly reduced. In these finite samples, the SD and RMSE of the B2SLSE can even be larger than those of the 2SLSE. Under V-D1, when the coefficients β_0's are small, the biases and SD's of the various 2SLSE's for λ_0 and β_{10} are too large to be acceptable.
- When the 2SLSE is poor, it has consequences for the GMM and RGMM approaches if it is used as an initial estimate for G_n and $G_n X_n \beta$. In Table 3 with P-D2 in V-D1, the GMME(2sl) and RGMME(2sl) are poor as they have large biases and SD's in λ_0 and β_{10}. When the 2SLSE's are satisfactory for P-D1, the GMME(2sl) and RGMME(2sl) in Table 3 are comparable with the corresponding GMME and RGMME in Table 1 (in both Mean and SD).
- In terms of SD and RMSE, the GMME and MLE are similar under all the designs (as reported in Tables 1, 2, 4 and 5). The SD's of the GMME and MLE of λ_0 under heteroskedasticity are slightly larger than those under i.i.d. disturbances for V-D1. With V-D1, the RMSE's of the MLE and GMME of λ_0 under heteroskedastic misspecification are larger than those of the correctly specified i.i.d. cases. The corresponding RMSE's for the intercept term are larger but to a smaller degree. For V-D2 (in Table 5), those SD's and RMSE's are mostly similar for all parameter designs.

[12] In addition to λ_0, we also pay attention to x and its coefficients. We are interested in comparing the 2SLS and the robust GMM estimates. The 2SLS estimates might be sensitive to x and its coefficients, since the 2SLS forms estimation based only on the deterministic part of the model, which is determined by the importance of x.

[13] The matrices correspond to the best P_n and Q_n in the i.i.d. case.

[14] We have also experimented with $R = 50$. Because of space limitation, those results are not reported here but they can be found in the working paper version of this paper.

Table 1
Estimates under Designs V-D1 and P-D1. V-D1: If group size $>$ 10, variance $=$ group size, else variance $=$ $1/(\text{groupsize})^2$. True parameters P-D1: $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) =$ (0.2, 0.8, 0.2, 1.5).

	R		Homoskedasticity				Heteroskedasticity			
			Mean	Bias	SD	RMSE	Mean	Bias	SD	RMSE
ML	100	λ	0.1917	(−0.0083)	0.0542	0.0549	0.1614	(−0.0386)	0.0617	0.0728
		β_1	0.8217	(0.0217)	0.3577	0.3584	0.9081	(0.1081)	0.3651	0.3808
		β_2	0.2000	(−0.0000)	0.1010	0.1010	0.1974	(−0.0026)	0.1020	0.1021
		β_3	1.4960	(−0.0040)	0.1184	0.1184	1.4939	(−0.0061)	0.1155	0.1157
	200	λ	0.1950	(−0.0050)	0.0386	0.0389	0.1659	(−0.0341)	0.0435	0.0553
		β_1	0.8123	(0.0123)	0.2541	0.2544	0.8915	(0.0915)	0.2559	0.2717
		β_2	0.2003	(0.0003)	0.0699	0.0699	0.2003	(0.0003)	0.0724	0.0724
		β_3	1.4988	(−0.0012)	0.0812	0.0812	1.4971	(−0.0029)	0.0851	0.0852
GMM	100	λ	0.1951	(−0.0049)	0.0543	0.0545	0.1679	(−0.0321)	0.0592	0.0673
		β_1	0.8137	(0.0137)	0.3575	0.3578	0.8921	(0.0921)	0.3609	0.3725
		β_2	0.1997	(−0.0003)	0.1008	0.1008	0.1972	(−0.0028)	0.1019	0.1020
		β_3	1.4947	(−0.0053)	0.1183	0.1184	1.4924	(−0.0076)	0.1155	0.1158
	200	λ	0.1967	(−0.0033)	0.0387	0.0388	0.1707	(−0.0293)	0.0419	0.0511
		β_1	0.8083	(0.0083)	0.2539	0.2541	0.8794	(0.0794)	0.2532	0.2654
		β_2	0.2002	(0.0002)	0.0698	0.0698	0.2002	(0.0002)	0.0724	0.0724
		β_3	1.4981	(−0.0019)	0.0811	0.0811	1.4962	(−0.0038)	0.0850	0.0851
2SLS	100	λ	0.1995	(−0.0005)	0.2400	0.2400	0.1886	(−0.0114)	0.2124	0.2127
		β_1	0.8098	(0.0098)	0.7184	0.7184	0.8425	(0.0425)	0.6576	0.6590
		β_2	0.1982	(−0.0018)	0.1004	0.1004	0.1962	(−0.0038)	0.1019	0.1020
		β_3	1.4868	(−0.0132)	0.1197	0.1204	1.4868	(−0.0132)	0.1176	0.1184
	200	λ	0.1987	(−0.0013)	0.1604	0.1604	0.2033	(0.0033)	0.1238	0.1239
		β_1	0.8069	(0.0069)	0.4930	0.4931	0.7943	(−0.0057)	0.3914	0.3914
		β_2	0.1996	(−0.0004)	0.0696	0.0696	0.1998	(−0.0002)	0.0721	0.0721
		β_3	1.4943	(−0.0057)	0.0815	0.0817	1.4931	(−0.0069)	0.0850	0.0853
RGMM	100	λ	0.1952	(−0.0048)	0.0544	0.0547	0.1906	(−0.0094)	0.0686	0.0692
		β_1	0.8135	(0.0135)	0.3575	0.3578	0.8321	(0.0321)	0.3716	0.3730
		β_2	0.1997	(−0.0003)	0.1008	0.1008	0.1971	(−0.0029)	0.1019	0.1019
		β_3	1.4947	(−0.0053)	0.1183	0.1184	1.4918	(−0.0082)	0.1155	0.1158
	200	λ	0.1969	(−0.0031)	0.0387	0.0389	0.1936	(−0.0064)	0.0479	0.0484
		β_1	0.8080	(0.0080)	0.2539	0.2540	0.8182	(0.0182)	0.2596	0.2602
		β_2	0.2002	(0.0002)	0.0698	0.0698	0.2002	(0.0002)	0.0723	0.0723
		β_3	1.4981	(−0.0019)	0.0811	0.0811	1.4954	(−0.0046)	0.0850	0.0851
ORGMM	100	λ	0.1935	(−0.0065)	0.0535	0.0539	0.1943	(−0.0057)	0.0702	0.0704
		β_1	0.8033	(0.0033)	0.3565	0.3565	0.8334	(0.0334)	0.3851	0.3866
		β_2	0.2050	(0.0050)	0.1012	0.1014	0.1946	(−0.0054)	0.1015	0.1017
		β_3	1.5033	(0.0033)	0.1209	0.1210	1.4943	(−0.0057)	0.1196	0.1197
	200	λ	0.1976	(−0.0024)	0.0391	0.0391	0.1976	(−0.0024)	0.0497	0.0497
		β_1	0.8161	(0.0161)	0.2408	0.2414	0.8028	(0.0028)	0.2616	0.2616
		β_2	0.1960	(−0.0040)	0.0709	0.0710	0.2008	(0.0008)	0.0718	0.0718
		β_3	1.5080	(0.0080)	0.0825	0.0829	1.5015	(0.0015)	0.0846	0.0846

Note: For GMM estimation with the matrix G_n, an initial consistent GMM estimate is used in the evaluations of G_n and $G_nX_n\beta$.

- As for a comparison of the SGMME in Table 3 with the GMME in Tables 1 and 2, the SGMME's are less efficient in λ_0 and β_{10}.[15]
- The RGMME does not seem to lose efficiency compared with the GMME as their SD's and RMSE's are similar under i.i.d. disturbances in these finite samples, even though the RGMME might be theoretically less asymptotically efficient than the GMME. This is so for all the results in Tables 1, 2, 4 and 5 with all the variance and parameter designs.
- Under heteroskedaticity, there is no obvious dominated pattern in terms of SD comparison of the RGMME with the GMME. In terms of RMSE, with $R = 200$, the RMSE's of the RGMME's of λ_0 and β_{10} are slightly smaller than those of the GMME's (in Tables 1, 2 and 4).[16] For V-D2 in Table 5, there is no difference between these two estimators.
- The ORGMM is the RGMM which uses the robust heteroskedastic variance of the moments as the optimal weighting matrix. Comparing the results of ORGMME with those of RGMME, the results are similar overall. It does not seem that optimal weighting with a robust variance under an unknown heteroskedaticity would improve efficiency in these finite samples.

6. Tests for heteroskedasticity

6.1. The LM test for heteroskedasticity

The possible presence of heteroskedasticity can be tested with the Breusch–Pagan LM test (Breusch and Pagan, 1979), using estimated residuals $\widehat{\epsilon}_{ni}$'s of the model from MLE or GMME. The Breusch–Pagan LM test assumes the alternative hypothesis $\sigma_{ni}^2 = f(\alpha_1 + z_i\alpha_2)$, where z_i is a vector of p-dimensional exogenous variables and f is a continuously differentiable function. However, due to the local nature of the LM test, one does not need to specify the functional form of f. So the functional restriction on this test is simply a linear index structure $\alpha_1 + z_i\alpha_2$ on the form of unknown heteroskedasticity. Under the null hypothesis $H_0, \alpha_2 = 0$. Let Z_n be the $n \times (p+1)$ matrix of observations on $(1, z_i)$ and let d_n be the n-dimensional vector of $d_{ni} = \frac{\widehat{\epsilon}_{ni}^2}{\widehat{\epsilon}_n'\widehat{\epsilon}_n/n} - 1$. Then the LM test statistic is $\frac{1}{2}d_n'Z_n(Z_n'Z_n)^{-1}Z_n'd_n$, which is asymptotically $\chi^2(p)$ under H_0.

15 Additional results of the SGMME in the settings of Tables 4 and 5 can be found in the working paper version.

16 For $R = 50$, there are a few cases where the MLE or GMME have smaller RMSEs than those of RGMME. These occur when RGMME happens to have a relatively larger SD.

Table 2
Estimates under Designs V-D1 and P-D2. V-D1: if group size $>$ 10, variance $=$ group size, else variance $=$ $1/(\text{groupsize})^2$ True parameters P-D2: $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.2, 0.2, 0.2, 0.1)$.

	R		Homoskedasticity				Heteroskedasticity			
			Mean	Bias	SD	RMSE	Mean	Bias	SD	RMSE
ML	100	λ	0.1913	(−0.0087)	0.0559	0.0566	0.1589	(−0.0411)	0.0650	0.0769
		β_1	0.2084	(0.0084)	0.3318	0.3319	0.2481	(0.0481)	0.3322	0.3357
		β_2	0.2000	(0.0000)	0.1010	0.1010	0.1974	(−0.0026)	0.1020	0.1020
		β_3	0.0963	(−0.0037)	0.1183	0.1184	0.0932	(−0.0068)	0.1155	0.1157
	200	λ	0.1948	(−0.0052)	0.0397	0.0400	0.1621	(−0.0379)	0.0465	0.0600
		β_1	0.2044	(0.0044)	0.2327	0.2327	0.2405	(0.0405)	0.2367	0.2402
		β_2	0.2003	(0.0003)	0.0699	0.0699	0.2004	(0.0004)	0.0725	0.0725
		β_3	0.0989	(−0.0011)	0.0812	0.0812	0.0963	(−0.0037)	0.0852	0.0852
GMM	100	λ	0.1952	(−0.0048)	0.0562	0.0564	0.1664	(−0.0336)	0.0630	0.0714
		β_1	0.2051	(0.0051)	0.3316	0.3317	0.2410	(0.0410)	0.3309	0.3334
		β_2	0.1998	(−0.0002)	0.1009	0.1009	0.1972	(−0.0028)	0.1020	0.1020
		β_3	0.0962	(−0.0038)	0.1182	0.1183	0.0932	(−0.0068)	0.1154	0.1156
	200	λ	0.1968	(−0.0032)	0.0396	0.0397	0.1665	(−0.0335)	0.0452	0.0563
		β_1	0.2025	(0.0025)	0.2325	0.2326	0.2361	(0.0361)	0.2364	0.2391
		β_2	0.2002	(0.0002)	0.0698	0.0698	0.2003	(0.0003)	0.0724	0.0724
		β_3	0.0989	(−0.0011)	0.0812	0.0812	0.0962	(−0.0038)	0.0851	0.0852
2SLS	100	λ	0.7743	(0.5743)	0.7099	0.9131	0.8026	(0.6026)	0.7260	0.9436
		β_1	−0.4052	(−0.6052)	0.8281	1.0256	−0.4337	(−0.6337)	0.8765	1.0815
		β_2	0.1990	(−0.0010)	0.1091	0.1091	0.2003	(0.0003)	0.1067	0.1067
		β_3	0.0983	(−0.0017)	0.1257	0.1257	0.0971	(−0.0029)	0.1208	0.1208
	200	λ	0.6648	(0.4648)	0.8130	0.9365	0.6138	(0.4138)	1.6272	1.6790
		β_1	−0.2941	(−0.4941)	0.9153	1.0401	−0.2450	(−0.4450)	1.8081	1.8621
		β_2	0.2005	(0.0005)	0.0732	0.0732	0.2018	(0.0018)	0.0842	0.0842
		β_3	0.0996	(−0.0004)	0.0875	0.0875	0.0958	(−0.0042)	0.0890	0.0890
RGMM	100	λ	0.1953	(−0.0047)	0.0564	0.0566	0.1917	(−0.0083)	0.0743	0.0748
		β_1	0.2050	(0.0050)	0.3316	0.3316	0.2147	(0.0147)	0.3325	0.3328
		β_2	0.1998	(−0.0002)	0.1009	0.1009	0.1971	(−0.0029)	0.1019	0.1019
		β_3	0.0962	(−0.0038)	0.1182	0.1182	0.0932	(−0.0068)	0.1154	0.1156
	200	λ	0.1970	(−0.0030)	0.0397	0.0398	0.1924	(−0.0076)	0.0526	0.0532
		β_1	0.2023	(0.0023)	0.2325	0.2325	0.2091	(0.0091)	0.2369	0.2370
		β_2	0.2002	(0.0002)	0.0698	0.0698	0.2001	(0.0001)	0.0724	0.0724
		β_3	0.0989	(−0.0011)	0.0812	0.0812	0.0961	(−0.0039)	0.0850	0.0851
ORGMM	100	λ	0.1935	(−0.0065)	0.0557	0.0560	0.1948	(−0.0052)	0.0972	0.0973
		β_1	0.1926	(−0.0074)	0.3323	0.3324	0.2239	(0.0239)	0.3434	0.3442
		β_2	0.2048	(0.0048)	0.1009	0.1010	0.1944	(−0.0056)	0.1012	0.1014
		β_3	0.1050	(0.0050)	0.1209	0.1210	0.0965	(−0.0035)	0.1193	0.1193
	200	λ	0.1979	(−0.0021)	0.0404	0.0404	0.1971	(−0.0029)	0.0540	0.0541
		β_1	0.2117	(0.0117)	0.2275	0.2278	0.1994	(−0.0006)	0.2334	0.2334
		β_2	0.1960	(−0.0040)	0.0710	0.0712	0.2006	(0.0006)	0.0717	0.0717
		β_3	0.1087	(0.0087)	0.0824	0.0828	0.1024	(0.0024)	0.0846	0.0847

Note: For GMM estimation with the matrix G_n, an initial consistent GMM estimate is used in the evaluations of G_n and $G_nX_n\beta$.

$$\Sigma_{1n} = \begin{pmatrix} \operatorname{tr}\left[\left(G_n - \frac{\operatorname{tr}(G_n)}{n} I_n\right)^s G_n\right] + \frac{1}{\sigma_0^2}(G_nX_n\beta_0)'(G_nX_n\beta_0) & \frac{1}{\sigma_0^2}(G_nX_n\beta_0)'X_n \\ \frac{1}{\sigma_0^2}X_n'(G_nX_n\beta_0) & \frac{1}{\sigma_0^2}X_n'X_n \end{pmatrix}$$

Box III.

6.2. The Hausman-type tests

Alternative statistics may be based on the comparison of robust estimates against estimates which are asymptotically efficient under H_0. These are the Hausman-type test statistics (Hausman, 1978), which seem natural as the 2SLSE and RGMME are robust and the MLE and GMME are asymptotically efficient under H_0 for our model. The Hausman-type test does not need the assumption of a linear index form for the variance function.

The main idea of the Hausman-type test is to compare two estimators $\widehat{\theta}_n$ and $\widetilde{\theta}_n$, with $\widehat{\theta}_n$ being asymptotically efficient under the null hypothesis H_0, but inconsistent under the alternative H_1, while $\widetilde{\theta}_n$ is consistent under both H_0 and H_1. The Hausman-type test statistic is

$$\begin{aligned}&(\widehat{\theta}_n - \widetilde{\theta}_n)'\operatorname{Var}(\widehat{\theta}_n - \widetilde{\theta}_n)^-(\widehat{\theta}_n - \widetilde{\theta}_n)\\ &\quad = (\widehat{\theta}_n - \widetilde{\theta}_n)'[\operatorname{Var}(\widetilde{\theta}_n) - \operatorname{Var}(\widehat{\theta}_n)]^-(\widehat{\theta}_n - \widetilde{\theta}_n) \overset{D}{\sim} \chi^2(m),\end{aligned}$$

where $[\operatorname{Var}(\widetilde{\theta}_n) - \operatorname{Var}(\widehat{\theta}_n)]^-$ is a generalized inverse of the matrix $[\operatorname{Var}(\widetilde{\theta}_n) - \operatorname{Var}(\widehat{\theta}_n)]$ with m being its rank (see, e.g., Ruud, 2000). Asymptotically, this statistic is invariant with respect to the choice of a generalized inverse.

When ϵ_{ni}'s are i.i.d. normal, the MLE is asymptotically efficient. So is the best GMME $\widehat{\theta}_n$ obtained by setting $P_n = (G_n - \frac{\operatorname{tr}(G_n)}{n} I_n)$ and $Q_n = (G_nX_n\beta_0, X_n)$, as it is asymptotically equivalent to the MLE when ϵ_{ni}'s are i.i.d. normal. Under H_0, the asymptotic variance matrix of the MLE (or GMME) is $\operatorname{Var}(\widehat{\theta}_n) = \Sigma_{1n}^{-1}$, where Σ_{1n} is as in Box III. The corresponding RGMME $\widetilde{\theta}_n$ has $Q_n = (G_nX_n\beta_0, X_n)$ but

Table 3
Miscellaneous 2SLSE and GMME. V-D1, true parameters P-D1 and P-D2, $R = 100$.

P-D1		Homoskedasticity				Heteroskedasticity			
		Mean	Bias	SD	RMSE	Mean	Bias	SD	RMSE
2SLS-2	λ	0.1787	(−0.0213)	0.2349	0.2359	0.2058	(0.0058)	0.1839	0.1840
	β_1	0.8499	(0.0499)	0.7114	0.7132	0.8054	(0.0054)	0.5890	0.5890
	β_2	0.2037	(0.0037)	0.1008	0.1008	0.1942	(−0.0058)	0.1019	0.1021
	β_3	1.4965	(−0.0035)	0.1220	0.1221	1.4907	(−0.0093)	0.1212	0.1216
B2SLS	λ	0.1384	(−0.0616)	0.3048	0.3109	0.1462	(−0.0538)	0.2155	0.2222
	β_1	0.9728	(0.1728)	0.9019	0.9183	0.9556	(0.1556)	0.6673	0.6852
	β_2	0.1986	(−0.0014)	0.1009	0.1010	0.1962	(−0.0038)	0.1017	0.1018
	β_3	1.4861	(−0.0139)	0.1215	0.1223	1.4872	(−0.0128)	0.1174	0.1180
SGMM	λ	0.1928	(−0.0072)	0.0564	0.0569	0.1546	(−0.0454)	0.0775	0.0898
	β_1	0.8260	(0.0260)	0.3800	0.3809	0.9519	(0.1519)	0.4492	0.4742
	β_2	0.1978	(−0.0022)	0.1060	0.1061	0.1899	(−0.0101)	0.1131	0.1136
	β_3	1.4960	(−0.0040)	0.1188	0.1188	1.4930	(−0.0070)	0.1161	0.1163
GMM(2sl)	λ	0.1933	(−0.0067)	0.0549	0.0553	0.1628	(−0.0372)	0.0742	0.0830
	β_1	0.8155	(0.0155)	0.3680	0.3683	0.9024	(0.1024)	0.3916	0.4048
	β_2	0.2029	(0.0029)	0.1041	0.1042	0.1998	(−0.0002)	0.1011	0.1011
	β_3	1.4954	(−0.0046)	0.1196	0.1197	1.4983	(−0.0017)	0.1177	0.1177
RGMM(2sl)	λ	0.1936	(−0.0064)	0.0542	0.0546	0.1916	(−0.0084)	0.0702	0.0707
	β_1	0.8145	(0.0145)	0.3669	0.3671	0.8263	(0.0263)	0.3792	0.3801
	β_2	0.2029	(0.0029)	0.1041	0.1041	0.1997	(−0.0003)	0.1010	0.1010
	β_3	1.4955	(−0.0045)	0.1196	0.1197	1.4969	(−0.0031)	0.1177	0.1177
P-D2									
2SLS-2	λ	0.4245	(0.2245)	0.7650	0.7973	0.7576	(0.5576)	0.6991	0.8942
	β_1	−0.0465	(−0.2465)	0.9003	0.9334	−0.3795	(−0.5795)	0.8800	1.0536
	β_2	0.2025	(0.0025)	0.1041	0.1041	0.1982	(−0.0018)	0.1101	0.1101
	β_3	0.1039	(0.0039)	0.1230	0.1231	0.0970	(−0.0030)	0.1245	0.1245
SGMM	λ	0.1926	(−0.0074)	0.0566	0.0571	0.1536	(−0.0464)	0.0782	0.0909
	β_1	0.2142	(0.0142)	0.3507	0.3509	0.2526	(0.0526)	0.3402	0.3443
	β_2	0.1980	(−0.0020)	0.1060	0.1060	0.1978	(−0.0022)	0.1031	0.1031
	β_3	0.0960	(−0.0040)	0.1187	0.1188	0.0933	(−0.0067)	0.1154	0.1156
GMM(2sl)	λ	0.6338	(0.4338)	0.7609	0.8759	0.5280	(0.3280)	0.8385	0.9004
	β_1	−0.3063	(−0.5063)	0.8683	1.0051	−0.1971	(−0.3971)	1.0062	1.0817
	β_2	0.2138	(0.0138)	0.1112	0.1120	0.2155	(0.0155)	0.1064	0.1075
	β_3	0.1026	(0.0026)	0.1330	0.1331	0.1072	(0.0072)	0.1273	0.1275
RGMM(2sl)	λ	0.6136	(0.4136)	0.7673	0.8717	0.5517	(0.3517)	0.7100	0.7924
	β_1	−0.2862	(−0.4862)	0.8900	1.0141	−0.2113	(−0.4113)	0.8635	0.9564
	β_2	0.2141	(0.0141)	0.1115	0.1124	0.2126	(0.0126)	0.1053	0.1060
	β_3	0.1021	(0.0021)	0.1331	0.1331	0.1052	(0.0052)	0.1256	0.1257

Note: 1. The 2SLS uses $Q_n = [W_nX_n, X_n]$ as IV's. 2. The 2SLS-2 uses IV's $[W_n^2X_n, W_nX_n, X_n]$. 3. RGMM(2sl): Robust GMM estimation with the matrix G_n, and 2SLSE used as initial consistent estimate in the evaluations of G_n and $G_nX_n\beta$. 4. P-D1: $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.2, 0.8, 0.2, 1.5)$. 5. P-D2: $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.2, 0.2, 0.2, 0.1)$.

$$\Sigma_{2n} = \begin{pmatrix} \text{tr}[(G_n - \text{Diag}(G_n))^s G_n] + \dfrac{1}{\sigma_0^2}(G_nX_n\beta_0)'(G_nX_n\beta_0) & \dfrac{1}{\sigma_0^2}(G_nX_n\beta_0)'X_n \\ \dfrac{1}{\sigma_0^2}X_n'(G_nX_n\beta_0) & \dfrac{1}{\sigma_0^2}X_n'X_n \end{pmatrix}$$

Box IV.

$P_n = (G_n - \text{Diag}(G_n))$, which is consistent under both H_0 and H_1, but is not asymptotically efficient under H_0. So is the B2SLSE with $Q_n = (G_nX_n\beta_0, X_n)$. The RGMME $\widetilde{\theta}_n$ has the asymptotic variance matrix $\text{Var}(\widetilde{\theta}_n) = \Sigma_{2n}^{-1}$ where Σ_{2n} is as in Box IV, and the B2SLSE $\widetilde{\theta}_{n,b}$ has its asymptotic variance $\text{Var}(\widetilde{\theta}_{n,b}) = \Sigma_{b,n}^{-1}$ where

$$\Sigma_{b,n} = \frac{1}{\sigma_0^2}\begin{pmatrix} (G_nX_n\beta_0)'(G_nX_n\beta_0) & (G_nX_n\beta_0)'X_n \\ X_n'(G_nX_n\beta_0) & X_n'X_n \end{pmatrix}. \quad (18)$$

Under the alternative H_1 of heteroskedasticity, as the MLE and GMME $\widehat{\theta}_n$ are inconsistent but the B2SLSE $\widetilde{\theta}_{n,b}$ and RGMME $\widetilde{\theta}_n$ are consistent, these estimators can be used to form the Hausman-type test statistics.

The difference in variance matrices, $[\text{Var}(\widetilde{\theta}_n) - \text{Var}(\widehat{\theta}_n)]$, may or may not have a full rank. To investigate the rank of $[\text{Var}(\widetilde{\theta}_n) - \text{Var}(\widehat{\theta}_n)]$ and/or $[\text{Var}(\widetilde{\theta}_{n,b}) - \text{Var}(\widehat{\theta}_n)]$, the expression $\text{Var}(\widetilde{\theta}_n) - \text{Var}(\widehat{\theta}_n) = \text{Var}(\widehat{\theta}_n)[\text{Var}(\widehat{\theta}_n)^{-1} - \text{Var}(\widetilde{\theta}_n)^{-1}]\text{Var}(\widetilde{\theta}_n)$ is useful as $\text{Var}(\widehat{\theta}_n)$ and $\text{Var}(\widetilde{\theta}_n)$ are invertible. The rank of this difference in variance matrices is that of $[\text{Var}(\widehat{\theta}_n)^{-1} - \text{Var}(\widetilde{\theta}_n)^{-1}]$, i.e., the rank of the matrix of the difference in the precision matrices. From equations given in Boxes III and IV, $\text{Var}(\widehat{\theta}_n)^{-1} - \text{Var}(\widetilde{\theta}_n)^{-1} = \begin{pmatrix} \text{tr}[(\text{Diag}(G_n) - \frac{\text{tr}(G_n)}{n}I_n)^s G_n] & 0 \\ 0 & 0 \end{pmatrix}$, and, with (18), $\text{Var}(\widehat{\theta}_n)^{-1} - \text{Var}(\widetilde{\theta}_{n,b})^{-1} = \begin{pmatrix} \text{tr}[(G_n - \frac{\text{tr}(G_n)}{n}I_n)^s G_n] & 0 \\ 0 & 0 \end{pmatrix}$, both of which have rank one. Therefore, a generalized inverse of the difference in variance matrices of MLE (or GMME) vs RGMME can be

$$[\text{Var}(\widetilde{\theta}_n) - \text{Var}(\widehat{\theta}_n)]^- = \text{Var}(\widetilde{\theta}_n)^{-1}$$

Table 4
Estimates under Designs V-D1 and P-D3, P-D4. V-D1: If group size >10, variance = group size, else variance = 1/(groupsize)^2. True parameters P-D3: $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.6, 0.8, 0.2, 1.5)$ P-D4: $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.6, 0.2, 0.2, 0.1)$ $R = 100$.

			Homoskedasticity				Heteroskedasticity			
			Mean	Bias	SD	RMSE	Mean	Bias	SD	RMSE
ML	P-D3	λ	0.5950	(−0.0050)	0.0292	0.0296	0.5515	(−0.0485)	0.0370	0.0610
		β_1	0.8256	(0.0256)	0.3619	0.3628	1.0571	(0.2571)	0.3833	0.4615
		β_2	0.2001	(0.0001)	0.1010	0.1010	0.1985	(−0.0015)	0.1025	0.1025
		β_3	1.4967	(−0.0033)	0.1185	0.1186	1.5020	(0.0020)	0.1160	0.1160
	P-D4	λ	0.5950	(−0.0050)	0.0302	0.0306	0.5481	(−0.0519)	0.0393	0.0651
		β_1	0.2094	(0.0094)	0.3333	0.3334	0.3104	(0.1104)	0.3398	0.3573
		β_2	0.2001	(0.0001)	0.1010	0.1010	0.1987	(−0.0013)	0.1025	0.1025
		β_3	0.0964	(−0.0036)	0.1184	0.1184	0.0936	(−0.0064)	0.1159	0.1161
GMM	P-D3	λ	0.5975	(−0.0025)	0.0282	0.0284	0.5560	(−0.0440)	0.0362	0.0570
		β_1	0.8138	(0.0138)	0.3591	0.3594	1.0356	(0.2356)	0.3790	0.4463
		β_2	0.1998	(−0.0002)	0.1008	0.1008	0.1981	(−0.0019)	0.1023	0.1024
		β_3	1.4950	(−0.0050)	0.1185	0.1186	1.4995	(−0.0005)	0.1161	0.1161
	P-D4	λ	0.5975	(−0.0025)	0.0292	0.0293	0.5521	(−0.0479)	0.0392	0.0618
		β_1	0.2050	(0.0050)	0.3321	0.3322	0.3030	(0.1030)	0.3378	0.3532
		β_2	0.1998	(−0.0002)	0.1009	0.1009	0.1983	(−0.0017)	0.1023	0.1024
		β_3	0.0962	(−0.0038)	0.1182	0.1183	0.0936	(−0.0064)	0.1158	0.1160
2SLS	P-D3	λ	0.6002	(0.0002)	0.1273	0.1273	0.5938	(−0.0062)	0.1205	0.1206
		β_1	0.8073	(0.0073)	0.7393	0.7393	0.8447	(0.0447)	0.7156	0.7169
		β_2	0.1982	(−0.0018)	0.1005	0.1005	0.1963	(−0.0037)	0.1020	0.1021
		β_3	1.4869	(−0.0131)	0.1204	0.1211	1.4874	(−0.0126)	0.1180	0.1187
	P-D4	λ	0.8941	(0.2941)	0.3437	0.4524	0.9014	(0.3014)	0.3844	0.4885
		β_1	−0.4048	(−0.6048)	0.7736	0.9820	−0.4155	(−0.6155)	0.9210	1.1078
		β_2	0.1944	(−0.0056)	0.1053	0.1054	0.1949	(−0.0051)	0.1045	0.1046
		β_3	0.0957	(−0.0043)	0.1217	0.1217	0.0944	(−0.0056)	0.1182	0.1183
RGMM	P-D3	λ	0.5975	(−0.0025)	0.0286	0.0287	0.5950	(−0.0050)	0.0355	0.0359
		β_1	0.8137	(0.0137)	0.3596	0.3598	0.8326	(0.0326)	0.3723	0.3737
		β_2	0.1998	(−0.0002)	0.1009	0.1009	0.1972	(−0.0028)	0.1018	0.1019
		β_3	1.4950	(−0.0050)	0.1185	0.1186	1.4924	(−0.0076)	0.1157	0.1160
	P-D4	λ	0.5976	(−0.0024)	0.0296	0.0297	0.5956	(−0.0044)	0.0383	0.0386
		β_1	0.2051	(0.0051)	0.3320	0.3321	0.2152	(0.0152)	0.3325	0.3328
		β_2	0.1998	(−0.0002)	0.1009	0.1009	0.1971	(−0.0029)	0.1019	0.1020
		β_3	0.0962	(−0.0038)	0.1182	0.1183	0.0933	(−0.0067)	0.1154	0.1156
ORGMM	P-D3	λ	0.5966	(−0.0034)	0.0282	0.0284	0.5969	(−0.0031)	0.0364	0.0365
		β_1	0.8040	(0.0040)	0.3583	0.3583	0.8352	(0.0352)	0.3858	0.3874
		β_2	0.2050	(0.0050)	0.1013	0.1014	0.1944	(−0.0056)	0.1015	0.1017
		β_3	1.5038	(0.0038)	0.1211	0.1212	1.4951	(−0.0049)	0.1199	0.1200
	P-D4	λ	0.5966	(−0.0034)	0.0293	0.0295	0.5960	(−0.0040)	0.0389	0.0391
		β_1	0.1928	(−0.0072)	0.3325	0.3326	0.2266	(0.0266)	0.3395	0.3406
		β_2	0.2049	(0.0049)	0.1009	0.1010	0.1943	(−0.0057)	0.1009	0.1011
		β_3	0.1051	(0.0051)	0.1210	0.1211	0.0966	(−0.0034)	0.1192	0.1192

$$\times \begin{pmatrix} \mathrm{tr}^{-1}\left[\left(\mathrm{Diag}(G_n) - \frac{\mathrm{tr}(G_n)}{n} I_n\right)^s G_n\right] & 0 \\ 0 & 0 \end{pmatrix} \mathrm{Var}(\widehat{\theta}_n)^{-1}, \quad (19)$$

and that of the MLE (or GMME) vs B2SLSE is

$$[\mathrm{Var}(\widetilde{\theta}_{n,b}) - \mathrm{Var}(\widehat{\theta}_n)]^- = \mathrm{Var}(\widetilde{\theta}_{n,b})^{-1}$$
$$\times \begin{pmatrix} \mathrm{tr}^{-1}\left[\left(G_n - \frac{\mathrm{tr}(G_n)}{n} I_n\right)^s G_n\right] & 0 \\ 0 & 0 \end{pmatrix} \mathrm{Var}(\widehat{\theta}_n)^{-1}. \quad (20)$$

Another generalized inverse can be derived with the eigenvalue and eigenvector decomposition of the matrix $[\mathrm{Var}(\widetilde{\theta}_n) - \mathrm{Var}(\widehat{\theta}_n)]$. As this matrix has a rank of one from our preceding analysis, let $\mu > 0$ be the single nonzero eigenvalue and let the corresponding orthonormal eigenvector matrix be Γ_n. The corresponding generalized inverse of $[\mathrm{Var}(\widetilde{\theta}_n) - \mathrm{Var}(\widehat{\theta}_n)]$ is $\Gamma_n' \Lambda_n^- \Gamma_n$ where Λ_n^- is a diagonal matrix consisting of $\frac{1}{\mu}$ and zeros on the diagonal elements. This generalized inverse is numerically non-negative definite and is the Moore–Penrose generalized inverse.[17]

The Hausman-type tests by comparing MLE (or GMME) vs RGMME, and MLE (or GMME) vs B2SLSE are both asymptotically $\chi^2(1)$.

6.3. Monte Carlo results for the tests

Table 6 presents the results of the Hausman-type and LM tests for heteroskedasticity in the SAR model. The Monte Carlo experimental designs are V-D1 with P-D1 and P-D2. The corresponding ML, GMM and RGMM estimates are those in Tables 1 and 2, and the B2SLSE is in Table 3. The left panel of the table shows the results for the homoskedasticity cases, and the right panel shows those for the heteroskedasticity cases. In each panel, the first two columns present, respectively, the results for the Hausman-type tests, using MLE vs B2SLSE and MLE vs RGMME. The results for the two LM tests, one based on MLE, the other on GMME, are shown in the last two columns of each panel. The alternative hypothesis for the LM tests is $\sigma_{ni}^2 = f(\alpha_0 + z_i\alpha)$, with z_i being the group

17 On the other hand, the generalized inverses in (19) and (20) are not symmetric. With a finite sample, the generalized inverse based on the eigenvalue and eigenvector has the numerical advantage in that the derived asymptotic χ^2 test statistics will always be non-negative.

Table 5
Estimates under Design V-D2 and various parameters. V-D2: variance = 1/(group size). True parameters: P-D1, P-D2, P-D3 and P-D4; $R = 100$.

			Homoskedasticity				Heteroskedasticity			
			Mean	Bias	SD	RMSE	Mean	Bias	SD	RMSE
ML	P-D1	λ	0.1994	(−0.0006)	0.0173	0.0173	0.1984	(−0.0016)	0.0167	0.0168
		β_1	0.8016	(0.0016)	0.0533	0.0534	0.8046	(0.0046)	0.0513	0.0515
		β_2	0.2000	(0.0000)	0.0085	0.0085	0.1998	(−0.0002)	0.0084	0.0084
		β_3	1.4996	(−0.0004)	0.0100	0.0100	1.4996	(−0.0004)	0.0097	0.0097
	P-D2	λ	0.1946	(−0.0054)	0.0495	0.0498	0.1920	(−0.0080)	0.0490	0.0497
		β_1	0.2058	(0.0058)	0.0585	0.0587	0.2088	(0.0088)	0.0580	0.0587
		β_2	0.2000	(−0.0000)	0.0085	0.0085	0.1998	(−0.0002)	0.0084	0.0084
		β_3	0.0997	(−0.0003)	0.0100	0.0100	0.0996	(−0.0004)	0.0097	0.0097
	P-D3	λ	0.5996	(−0.0004)	0.0088	0.0088	0.5992	(−0.0008)	0.0086	0.0086
		β_1	0.8018	(0.0018)	0.0535	0.0536	0.8046	(0.0046)	0.0517	0.0520
		β_2	0.2000	(0.0000)	0.0085	0.0085	0.1999	(−0.0001)	0.0084	0.0084
		β_3	1.4997	(−0.0003)	0.0101	0.0101	1.4997	(−0.0003)	0.0099	0.0099
	P-D4	λ	0.5967	(−0.0033)	0.0265	0.0267	0.5951	(−0.0049)	0.0267	0.0272
		β_1	0.2068	(0.0068)	0.0605	0.0608	0.2104	(0.0104)	0.0609	0.0618
		β_2	0.2000	(0.0000)	0.0086	0.0086	0.1999	(−0.0001)	0.0085	0.0085
		β_3	0.0997	(−0.0003)	0.0100	0.0100	0.0996	(−0.0004)	0.0097	0.0097
GMM	P-D1	λ	0.1993	(−0.0007)	0.0172	0.0172	0.1984	(−0.0016)	0.0166	0.0167
		β_1	0.8019	(0.0019)	0.0531	0.0532	0.8047	(0.0047)	0.0510	0.0512
		β_2	0.2000	(−0.0000)	0.0085	0.0085	0.1998	(−0.0002)	0.0085	0.0085
		β_3	1.4995	(−0.0005)	0.0100	0.0100	1.4995	(−0.0005)	0.0097	0.0097
	P-D2	λ	0.1969	(−0.0031)	0.0494	0.0495	0.1960	(−0.0040)	0.0602	0.0604
		β_1	0.2038	(0.0038)	0.0585	0.0586	0.2051	(0.0051)	0.0685	0.0687
		β_2	0.1998	(−0.0002)	0.0085	0.0085	0.1996	(−0.0004)	0.0084	0.0085
		β_3	0.0996	(−0.0004)	0.0100	0.0100	0.0995	(−0.0005)	0.0097	0.0097
	P-D3	λ	0.5996	(−0.0004)	0.0090	0.0090	0.5991	(−0.0009)	0.0087	0.0087
		β_1	0.8021	(0.0021)	0.0541	0.0542	0.8048	(0.0048)	0.0519	0.0521
		β_2	0.2000	(−0.0000)	0.0085	0.0085	0.1998	(−0.0002)	0.0085	0.0085
		β_3	1.4996	(−0.0004)	0.0101	0.0101	1.4996	(−0.0004)	0.0099	0.0099
	P-D4	λ	0.5984	(−0.0016)	0.0257	0.0258	0.5973	(−0.0027)	0.0259	0.0261
		β_1	0.2038	(0.0038)	0.0591	0.0592	0.2064	(0.0064)	0.0592	0.0596
		β_2	0.1998	(−0.0002)	0.0085	0.0085	0.1997	(−0.0003)	0.0085	0.0085
		β_3	0.0996	(−0.0004)	0.0100	0.0100	0.0995	(−0.0005)	0.0097	0.0097
2SLS	P-D1	λ	0.2002	(0.0002)	0.0180	0.0180	0.1993	(−0.0007)	0.0176	0.0176
		β_1	0.7993	(−0.0007)	0.0550	0.0550	0.8020	(0.0020)	0.0531	0.0532
		β_2	0.2000	(−0.0000)	0.0085	0.0085	0.1998	(−0.0002)	0.0084	0.0084
		β_3	1.4996	(−0.0004)	0.0100	0.0100	1.4996	(−0.0004)	0.0097	0.0097
	P-D2	λ	0.2069	(0.0069)	0.1091	0.1093	0.1993	(−0.0007)	0.1159	0.1159
		β_1	0.1937	(−0.0063)	0.1168	0.1170	0.2020	(0.0020)	0.1243	0.1243
		β_2	0.1997	(−0.0003)	0.0086	0.0086	0.1995	(−0.0005)	0.0085	0.0085
		β_3	0.0995	(−0.0005)	0.0100	0.0100	0.0994	(−0.0006)	0.0097	0.0097
	P-D3	λ	0.6001	(0.0001)	0.0095	0.0095	0.5996	(−0.0004)	0.0093	0.0093
		β_1	0.7993	(−0.0007)	0.0562	0.0562	0.8020	(0.0020)	0.0543	0.0544
		β_2	0.2000	(−0.0000)	0.0085	0.0085	0.1998	(−0.0002)	0.0084	0.0084
		β_3	1.4996	(−0.0004)	0.0101	0.0101	1.4996	(−0.0004)	0.0099	0.0099
	P-D4	λ	0.6036	(0.0036)	0.0576	0.0577	0.5996	(−0.0004)	0.0614	0.0614
		β_1	0.1936	(−0.0064)	0.1198	0.1200	0.2022	(0.0022)	0.1280	0.1280
		β_2	0.1996	(−0.0004)	0.0087	0.0087	0.1995	(−0.0005)	0.0086	0.0086
		β_3	0.0995	(−0.0005)	0.0100	0.0100	0.0995	(−0.0005)	0.0097	0.0097
RGMM	P-D1	λ	0.1993	(−0.0007)	0.0172	0.0172	0.1984	(−0.0016)	0.0166	0.0167
		β_1	0.8019	(0.0019)	0.0532	0.0532	0.8047	(0.0047)	0.0510	0.0512
		β_2	0.2000	(−0.0000)	0.0085	0.0085	0.1998	(−0.0002)	0.0085	0.0085
		β_3	1.4995	(−0.0005)	0.0100	0.0100	1.4995	(−0.0005)	0.0097	0.0097
	P-D2	λ	0.1969	(−0.0031)	0.0495	0.0496	0.1960	(−0.0040)	0.0602	0.0603
		β_1	0.2038	(0.0038)	0.0585	0.0586	0.2050	(0.0050)	0.0685	0.0686
		β_2	0.1998	(−0.0002)	0.0085	0.0085	0.1996	(−0.0004)	0.0084	0.0085
		β_3	0.0996	(−0.0004)	0.0100	0.0100	0.0995	(−0.0005)	0.0097	0.0097
	P-D3	λ	0.5996	(−0.0004)	0.0090	0.0090	0.5991	(−0.0009)	0.0087	0.0087
		β_1	0.8021	(0.0021)	0.0542	0.0543	0.8048	(0.0048)	0.0518	0.0520
		β_2	0.2000	(−0.0000)	0.0085	0.0085	0.1998	(−0.0002)	0.0085	0.0085
		β_3	1.4996	(−0.0004)	0.0101	0.0101	1.4996	(−0.0004)	0.0099	0.0099
	P-D4	λ	0.5984	(−0.0016)	0.0260	0.0261	0.5973	(−0.0027)	0.0260	0.0261
		β_1	0.2038	(0.0038)	0.0596	0.0597	0.2063	(0.0063)	0.0592	0.0595
		β_2	0.1998	(−0.0002)	0.0085	0.0085	0.1997	(−0.0003)	0.0085	0.0085
		β_3	0.0996	(−0.0004)	0.0100	0.0100	0.0995	(−0.0005)	0.0097	0.0097
ORGMM	P-D1	λ	0.1988	(−0.0012)	0.0162	0.0162	0.2000	(−0.0000)	0.0164	0.0164
		β_1	0.8023	(0.0023)	0.0506	0.0507	0.8008	(0.0008)	0.0521	0.0521
		β_2	0.2004	(0.0004)	0.0085	0.0085	0.1997	(−0.0003)	0.0085	0.0085
		β_3	1.5003	(0.0003)	0.0102	0.0102	1.4997	(−0.0003)	0.0100	0.0100
	P-D2	λ	0.1956	(−0.0044)	0.0580	0.0581	0.1965	(−0.0035)	0.0486	0.0487

(*continued on next page*)

Table 5 (*continued*)

		Homoskedasticity				Heteroskedasticity			
		Mean	Bias	SD	RMSE	Mean	Bias	SD	RMSE
	β_1	0.2040	(0.0040)	0.0672	0.0673	−0.2049	(0.0049)	0.0590	0.0592
	β_2	0.2003	(0.0003)	0.0086	0.0086	0.1996	(−0.0004)	0.0085	0.0085
	β_3	0.1003	(0.0003)	0.0102	0.0102	0.0997	(−0.0003)	0.0100	0.0100
P-D3	λ	0.5994	(−0.0006)	0.0085	0.0085	0.6000	(−0.0000)	0.0085	0.0085
	β_1	0.8023	(0.0023)	0.0515	0.0515	0.8009	(0.0009)	0.0528	0.0528
	β_2	0.2004	(0.0004)	0.0085	0.0085	0.1997	(−0.0003)	0.0085	0.0085
	β_3	1.5004	(0.0004)	0.0103	0.0103	1.4997	(−0.0003)	0.0101	0.0101
P-D4	λ	0.5972	(−0.0028)	0.0254	0.0256	0.5982	(−0.0018)	0.0254	0.0254
	β_1	0.2050	(0.0050)	0.0582	0.0584	0.2049	(0.0049)	0.0597	0.0599
	β_2	0.2003	(0.0003)	0.0086	0.0086	0.1996	(−0.0004)	0.0085	0.0085
	β_3	0.1004	(0.0004)	0.0102	0.0102	0.0998	(−0.0002)	0.0100	0.0100

Note: P-D1: $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.2, 0.8, 0.2, 1.5)$; P-D2: $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.2, 0.2, 0.2, 0.1)$; P-D3: $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.6, 0.8, 0.2, 1.5)$; P-D4: $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (0.6, 0.2, 0.2, 0.1)$.

Table 6

Tests for Heteroskedasticity. V-D1; Two sets of true parameters: P-D1 and P-D2.

R		Empirical level				Power			
		MLE vs B2SLSE	MLE vs RGMME	LM via MLE	LM via GMME	MLE vs B2SLSE	MLE vs RGMME	LM via MLE	LM via GMME
P-D1									
50	1%	6.9	33.7	0.7	0.7	2.4	99.4 (19.5)	100	100
	5%	10.1	43.7	5.0	5.0	4.2	99.6 (91.9)	100	100
	10%	12.5	50.2	9.3	9.3	6.0	99.7 (97.0)	100	100
100	1%	6.1	32.9	1.1	1.1	2.4	100 (68.8)	100	100
	5%	10.3	43.5	3.8	3.8	4.6	100 (99.5)	100	100
	10%	12.5	51.1	8.5	8.5	7.5	100 (99.8)	100	100
200	1%	3.4	33.4	1.1	1.1	0.9	100 (100)	100	100
	5%	6.9	44.6	6.0	6.0	2.6	100 (100)	100	100
	10%	11.0	52.2	11.9	11.9	5.4	100 (100)	100	100
P-D2									
50	1%	11.7	33.5	0.7	0.7	11.5	99.8 (6.8)	100	100
	5%	15.1	44.0	5.2	5.1	15.3	99.8 (85.5)	100	100
	10%	18.7	49.6	9.4	9.5	17.0	99.9 (98.4)	100	100
100	1%	12.3	32.7	1.1	1.1	14.8	99.7 (46.8)	100	100
	5%	16.7	45.3	3.9	3.8	16.9	99.8 (99.3)	100	100
	10%	19.3	52.5	8.2	8.2	19.8	99.9 (99.5)	100	100
200	1%	17.2	35.0	1.3	1.3	15.6	100 (99.7)	100	100
	5%	21.9	46.4	5.9	5.9	18.4	100 (99.8)	100	100
	10%	24.9	53.1	11.9	12.0	21.5	100 (99.9)	100	100

Note: 1. The Hausman-type tests are $\chi^2(1)$ under the null hypothesis of homoskedasticity. 2. The LM tests are $\chi^2(1)$ under the null hypothesis of homoskedasticity. 3. The table shows the percentages of rejecting the null hypothesis in all the 1000 Monte Carlo replications, for nominal sizes 1%, 5%, 10%. 4. The numbers in parentheses for the powers of the Hausman-type test with MLE vs RGMME are the bias-adjusted empirical powers.

size.[18] As discussed in the previous subsection, it is not necessary to specify the functional form of f. The Hausman-type tests use both the Moore–Penrose generalized inverse and the generalized inverses in (19) and (20). The corresponding results are similar.[19]

The Hausman-type test using MLE vs B2SLSE has no power for the sample sizes $R = 50$ to 200. Even though its empirical levels are higher than the theoretical ones, its powers are not even larger than the empirical levels. For the Hausman-type test of MLE vs RGMME, its empirical levels are very large, showing over-rejection of the null hypothesis. It does have power even after adjusting the proper level of significance, but its large empirical levels will render this test useless. These phenomena can be understood by investigating the generalized inverse formulas in (19) and (20) and the small biases of the corresponding estimates. For the Hausman-type test using MLE vs RGMME, the test statistic is inflated by the variance difference term $\mathrm{tr}[(\mathrm{Diag}(G_n) - \frac{\mathrm{tr}(G_n)}{n} I_n)^s G_n]$. In the samples for the Monte Carlo study, this term happens to be very small, with a mean ranging from 0.26 to 1.06 for all cases. These are small even though the trace operation is a summation over n terms. Thus, it might produce a big number when its inverse is involved, which is explicit in (19). On the contrary, for the Hausman-type test using MLE vs B2SLSE, the corresponding variance difference term has mean value ranging from 150 to 670, which would give a small number after inversion. Overall, the Hausman-type tests are not reliable.

In contrast, the LM tests perform very well. The empirical levels are close to the theoretical ones and they have excellent powers.[20]

7. Application to county teenage pregnancy rates

Teenage pregnancy is one of the contexts where social interaction effects are believed to be most important. Jencks and Mayer (1990), for example, conclude that, ‘‘neighborhoods and classmates probably have a stronger effect on sexual behavior than on cognitive skills, school enrollment decisions, or even criminal

18 In the variance design V-D1, the group size variable in the variance function is nonlinear and complicated. So the linear index specification of the variance for the LM test provides only an approximation of the true variance function. Our intention is to see whether a linear index approximation can capture the alternative in its power function, since in practice we may not know the exact variance function.

19 The results of the Hausman-type tests reported in Table 6 are those with the Moore–Penrose generalized inverse.

20 This may indicate that the linear index approximation of the nonlinear variance function is valuable. The linear approximation does capture the group size variable in the variance function.

activity''. Many studies, including Hogan and Kitagawa (1985), Crane (1991), Case and Katz (1991) and Evans et al. (1992), analyze neighborhood effects in teenage pregnancy by using micro-data. It would be of interest to study the spatial effects at more aggregated levels and see how county teenage pregnancy rates are affected by each other. We suspect the possible presence of unknown heteoskedasticity in this aggregated data. Therefore, we apply the RGMM estimation procedures and compare them to other estimation methods.

The model considered is the SAR model in (1), by which we related a county's teenage pregnancy rate to those of its neighbors and its own characteristics. Following Kelejian and Robinson (1993), we focus on counties in the 10 Upper Great Plains States, including Colorado, Iowa, Kansas, Minnesota, Missouri, Montana, Nebraska, North Dakota, South Dakota, and Wyoming, which consist of 761 counties. A county's neighbors are referred to its geographically neighboring counties.

The data used are from ''Health and Healthcare in the United States— County and Metro Area Data'' (Thomas, 1999), and the 1990 US Census (US Census Bureau, 1992). The specific model is given by

$$Teen_i = \lambda \sum_{j=1}^{760} w_{ij} Teen_j + \beta_1 + Edu_i\beta_2 + Inco_i\beta_3 + FHH_i\beta_4 + Black_i\beta_5 + Phy_i\beta_6 + \epsilon_i,$$

where $Teen_i$ is the teenage pregnancy rate in county i, which is the percentage of pregnancies occurring to females of 12–17 years old. w_{ij} is the entry in the spatial weights matrix W_n, which will be zero if two counties are not neighboring counties. The neighbors of the same county are assigned an equal weight in the row-normalized spatial weights matrix. The term, $\sum_{j=1}^{760} w_{ij}Teen_j$, is simply the average of the teenage pregnancy rates of county i's neighbors. Edu_i is the education service expenditure (divided by 100), $Inco_i$ is median household income (divided by 1000), FHH_i is the percentage of female-headed households, $Black_i$ is the proportion of black population and Phy_i is the number of physicians per 1000 population, all in county i.[21] We assume that the ϵ_{ni}'s have zero mean and variances σ_{ni}^2's, and are independent across counties.

The model is estimated by 2SLS, B2SLS, ML, non-robust GMM, robust RGMM and optimal weighting RGMM procedures. The results are reported in Table 7. Consistent with the Monte Carlo results, most of the differences among the estimators are for λ_0 and the intercept, with the 2SLSE $\widehat{\lambda}_{2SLS} = 0.409$ being larger than those of the others: $\widehat{\lambda}_{B2SLS} = 0.358$, $\widehat{\lambda}_{ML} = 0.339$ and all three GMME's are 0.343 or 0.344. Thus, relative to the RGMME, the 2SLSE overestimates λ_0, and the B2SLSE improves upon the 2SLSE by decreasing the relative bias. For the intercept term, the 2SLSE is relatively smaller than the others. The estimates obtained from all the other methods are similar. For the t-statistics, we can see that those for the MLE and all the three GMME's procedures are similar, while those for 2SLSE and the B2SLSE are smaller for the estimates of λ_0 and the intercept, which reflects the inefficiency of the 2SLSE's. Furthermore, the differences between the robust and non-robust standard errors for the 2SLS's and the robust GMM estimators are notable. In particular, for all the three procedures, the non-robust standard errors for the coefficient on female-headed households are only about 60% as large as the robust ones, which is striking. Also, the larger non-robust standard errors of the coefficient on education service expenditure make it become marginally insignificant, although it should be statistically significant at the 5% level based on the robust standard errors. These distinctions could have an impact on the inferences, especially when the estimates are on the margin of being significant.

Table 7
Estimation of spatial effects for county teenage pregnancy rates.

	2SLS	B2SLS	ML	GMM	RGMM	ORGMM
λ	0.409	0.358	0.339	0.343	0.343	0.344
	(4.92)	(3.98)	(7.53)	(7.64)	(7.64)	–
	[4.83]	[4.09]	–	–	[6.86]	[6.92]
Cons	7.179	7.879	8.140	8.096	8.091	8.076
	(4.77)	(4.73)	(6.81)	(6.78)	(6.77)	–
	[4.24]	[4.89]	–	–	[6.54]	[6.57]
Edu	−0.011	−0.011	−0.011	−0.011	−0.011	−0.011
	(−1.63)	(−1.72)	(−1.75)	(−1.74)	(−1.74)	–
	[−2.37]	[−2.50]	–	–	[−2.52]	[−2.53]
Inco	−0.197	−0.204	−0.206	−0.206	−0.206	−0.206
	(−4.90)	(−4.80)	(−5.20)	(−5.20)	(−5.20)	–
	[−4.59]	[−4.94]	–	–	[−5.39]	[−5.39]
FHH	0.751	0.763	0.763	0.766	0.766	0.768
	(11.83)	(11.92)	(11.92)	(12.43)	(12.43)	–
	[7.71]	[7.86]	–	–	[8.18]	[8.25]
Black	0.138	0.145	0.145	0.147	0.147	0.147
	(2.42)	(2.58)	(2.58)	(2.64)	(2.64)	–
	[2.79]	[2.88]	–	–	[2.89]	[2.89]
Phy	−0.512	−0.523	−0.523	−0.526	−0.526	−0.527
	(−2.74)	(−2.80)	(−2.80)	(−2.81)	(−2.81)	–
	[−2.30]	[−2.69]	–	–	[−2.72]	[−2.72]

Note: 1. The explanatory variables are: Cons = intercept term, Edu = education service expenditure (divided by 100), Inco = median household income (divided by 1000), FHH = percentage of female-headed households, Black = proportion of black population, and Phy = number of physicians per 1000 population. 2. 2SLS uses (W_nX_n, X_n) as IV's; B2SLS uses $(G_nX_n\beta, X_n)$ as IV's and 2SLSE as initial estimate. 3. All GMM's use an initial SGMME in the evaluations of G_n and $G_nX_n\beta$. 4. The t-statistics in parentheses are those under i.i.d. disturbances assumption. The t-statistics for the 2SLS, B2SLS and RGMM and ORGMM estimators calculated from the robust variance formula are in square brackets. 5. The LM test statistic (via MLE) is 18.506 and the LM test statistic (via GMME) is 18.557. 6. The Hausman-type test statistic with MLE vs B2SLSE is 0.054 and the Hausman-type test statistic with MLE vs RGMME is 18.315.

Based on the various GMM and MLE results, we see that the county teenage pregnancy rates in these 10 states exhibit a strong spatial convergence, with an estimated spatial coefficient of around 0.34. Thus, about 34% of the changes in the teenage pregnancy rates of neighboring counties will be absorbed by a county's own teenage pregnancy rate.[22] All the other parameters have the expected signs. From Table 7 we can see that other significant and important determinants of county teenage pregnancy rate include median household income, proportion of female-headed households, fraction of black population and the number of physicians per 1000 population. Generally speaking, other things being equal, the larger the percentage of female-headed households or the higher the proportion of black population, the higher the county teenage pregnancy rate. As well as the number of physicians per 1000 population, household income and education service expenditure all help to reduce county teenage pregnancy rate.

We perform two Hausman-type tests using MLE vs B2SLSE and also MLE vs RGMME, and two LM tests based on MLE and non-robust GMME, using county population size as z_i in the variance

21 Some variables, such as the percentage of high school graduates, are insignificant in the preliminary study thus are dropped.

22 Our result is consistent with previous studies which also find significant neighborhood effects in teenage pregnancy. In particular, Hogan and Kitagawa (1985) find that the probabilities of becoming pregnant were about 1/3 higher for teenagers from low-quality neighborhoods and living in the West Side ghetto increased the chances by about 2/5. Crane (1991) also finds significant neighborhood influences in teenage pregnancy, especially in the very worst neighborhoods. However, in our case, county teenage pregnancy rates are aggregated from individual outcomes and are treated as continuous. Other studies, including Case and Katz (1991) and Evans et al. (1992), find insignificant neighborhood effects in teenage pregnancy.

function. The LM test statistics based on the MLE is 18.506, the one based on the GMME is 18.557, both reject the null hypothesis of homoskedasticity. However, the Hausman-type test statistics using the MLE vs B2SLSE is as small as 0.054, and the other one with the MLE vs RGMME is 18.315. From the Monte Carlo study, we observe that the Hausman-type test by comparing the MLE and B2SLSE does not have power, and the one using the MLE vs RGMME tends to over-reject the null. Thus, the Hausman-type tests might have the same weakness as in the Monte Carlo cases. Even though the LM tests may reject the null of homoskedastic errors, our overall conclusion is that even if there was any heteroskedasticity in this sample, it does not have noticeable effects on the ML and GMM coefficient estimates in this application. However, the presence of heteroskedasticity does affect the estimates of the standard errors, and consequentially, the statistical inferences.

8. Conclusion

This paper considers the GMM estimation in the presence of unknown heteroskedasticity in a SAR model where the disturbances are independent but may have heteroskedastic variances.

In the presence of heteroskedastic disturbances, the ML approach for the SAR model would in general provide an inconsistent MLE if the disturbances were treated as i.i.d. Method of Moments or GMM approaches would theoretically suffer from the inconsistency if the moment functions are designed for i.i.d. disturbances, and thus, ignore the unknown heteroskedaticity in the disturbances. In this paper, we analyze a general systematic framework in GMM estimation where the moment functions take into account the possible presence of unknown heteroskedastic disturbances. The resulted estimator RGMME is shown to be consistent and asymptotically normal. Asymptotically valid inferences can be drawn with consistently estimated covariance matrices. We also consider the optimal RGMM estimation which can improve asymptotic efficiency by the construction of a feasible optimal weighting matrix under an unknown heteroskedasticity. Statistical procedures for testing the presence of unknown heteroskedaticity are investigated.

Monte Carlo experiments are designed to study the finite sample properties of the ML, GMM, 2SLS, robust GMM and some related estimators, and the test statistics. The Monte Carlo results show that even though 2SLSE's shall be consistent in the presence of unknown heteroskedaticity, they may have large variances and biases in finite samples for cases where regressors do not have strong effects. The robust GMME has desirable properties while the biases associated with the MLE and non-robust GMME may remain in large samples, especially, for the spatial effect coefficient and the intercept term. However, the magnitudes of biases are only moderate. With moderately large sample sizes, those biases may be statistically insignificant. The Hausman-type test statistics are shown to be unreliable, but the LM test statistics have good finite sample properties.

The various approaches are applied to the study of county teenage pregnancy rates. The empirical results show a strong spatial convergence among county teenage pregnancy rates with a significant spatial effect. The LM test statistics confirm the presence of heteroskedasticity, but it has no impact on the coefficient estimates of this empirical model. However, the presence of heteroskedasticity does affect the estimates of the standard errors, and consequentially, the statistical inferences.

Acknowledgements

We appreciate having financial support for our research from the NSF under grant no. 0519204, and thank Patricia Reagan for helpful comments and the data source for our empirical study. We are grateful to the guest editors and three anonymous referees for valuable comments and suggestions on an earlier version of this paper.

Appendix. Some useful lemmas and proofs of main results

Lemma A.1. *For any two square matrices $A_n = [a_{n,ij}]$ and $B_n = [b_{n,ij}]$ of dimension n with zero diagonals, assume that ϵ_{ni}'s have zero mean and are mutually independent. Then,*

(1) $E(A_n\epsilon_n \cdot \epsilon_n' B_n \epsilon_n) = 0$,
(2) $E(A_n\epsilon_n(B_n\epsilon_n)') = A_n\Sigma_n B_n'$, *and*
(3) $E(\epsilon_n' A_n\epsilon_n \cdot \epsilon_n' B_n \epsilon_n) = \sum_{i=1}^n\sum_{j=1}^n a_{n,ij}(b_{n,ij} + b_{n,ji})\sigma_{ni}^2\sigma_{nj}^2 = \mathrm{tr}[\Sigma_n A_n(B_n'\Sigma_n + \Sigma_n B_n)]$;

where $\Sigma_n = Diag\{\sigma_{n1}^2, \ldots, \sigma_{nn}^2\}$ with $\sigma_{ni}^2 = E(\epsilon_{ni}^2)$ and $\epsilon_n = (\epsilon_{n1}, \ldots, \epsilon_{nn})'$.

Proof. (1) Because ϵ_{ni}'s are mutually independent and $b_{n,ii} = 0$,

$$E(A_n\epsilon_n \cdot \epsilon_n' B_n\epsilon_n) = A_n\sum_{i=1}^n\sum_{j=1}^n b_{n,ij}E(\epsilon_{ni}\epsilon_{nj}\epsilon_n) = A_n\sum_{i=1}^n b_{n,ii}E(\epsilon_{ni}^3) = 0.$$

(2) $E(A_n\epsilon_n(B_n\epsilon_n)') = A_nE(\epsilon_n\epsilon_n')B_n' = A_n\Sigma_n B_n'$.

(3) As $\epsilon_n' A_n\epsilon_n\epsilon_n' B_n\epsilon_n = \sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n a_{n,ij}b_{n,kl}\epsilon_{ni}\epsilon_{nj}\epsilon_{nk}\epsilon_{nl}$, the mutual independence of ϵ_{ni}s implies that $E(\epsilon_{ni}\epsilon_{nj}\epsilon_{nk}\epsilon_{nl}) \neq 0$ only if $(i = j = k = l)$, $(i = j, k = l)$, $(i = k, j = l)$, or $(i = l, j = k)$. It follows that

$$\begin{aligned} E(\epsilon_n' A_n\epsilon_n \cdot \epsilon_n' B_n\epsilon_n) &= \sum_{i=1}^n a_{n,ii}b_{n,ii}E(\epsilon_{ni}^4) + \sum_{i=1}^n\sum_{j\neq i}^n(a_{n,ii}b_{n,jj} \\ &\quad + a_{n,ij}b_{n,ij} + a_{n,ij}b_{n,ji})E(\epsilon_{ni}^2)E(\epsilon_{nj}^2) \\ &= \sum_{i=1}^n\sum_{j=1}^n(a_{n,ii}b_{n,jj} + a_{n,ij}b_{n,ij} + a_{n,ij}b_{n,ji})\sigma_{ni}^2\sigma_{nj}^2 \\ &= \mathrm{tr}[\Sigma_n A_n(\Sigma_n B_n + B_n'\Sigma_n)], \end{aligned}$$

because $a_{n,ii} = b_{n,ii} = 0$ for all i. □

The expressions in Lemma A.1 provide the formula for Ω_n in (13).

Lemma A.2. *For any square matrices $A_n = [a_{n,ij}]$ of dimension n, assume that ϵ_{ni}'s have a zero mean and are mutually independent. Then,*

(1) $E(\epsilon_n' A_n\epsilon_n) = \sum_{i=1}^n a_{n,ii}\sigma_{ni}^2 = \mathrm{tr}(\Sigma_n A_n)$,
(2)

$$\begin{aligned} E(\epsilon_n' A_n\epsilon_n)^2 &= \sum_{i=1}^n a_{n,ii}^2[E(\epsilon_{ni}^4) - 3\sigma_{ni}^4] + \left(\sum_{i=1}^n a_{n,ii}\sigma_{ni}^2\right)^2 \\ &\quad + \sum_{i=1}^n\sum_{j=1}^n a_{n,ij}(a_{n,ij} + a_{n,ji})\sigma_{ni}^2\sigma_{nj}^2 \\ &= \sum_{i=1}^n a_{n,ii}^2[E(\epsilon_{ni}^4) - 3\sigma_{ni}^4] + \mathrm{tr}^2(\Sigma_n A_n) \\ &\quad + \mathrm{tr}[\Sigma_n A_n(A_n'\Sigma_n + \Sigma_n A_n)], \end{aligned}$$

and

(3)

$$\begin{aligned}\mathrm{Var}(\epsilon_n' A_n \epsilon_n) &= \sum_{i=1}^n a_{n,ii}^2 [E(\epsilon_{ni}^4) - 3\sigma_{ni}^4] \\ &\quad + \sum_{i=1}^n \sum_{j=1}^n a_{n,ij}(a_{n,ij} + a_{n,ji})\sigma_{ni}^2 \sigma_{nj}^2 \\ &= \sum_{i=1}^n a_{n,ii}^2 [E(\epsilon_{ni}^4) - 3\sigma_{ni}^4] \\ &\quad + \mathrm{tr}[\Sigma_n A_n (A_n' \Sigma_n + \Sigma_n A_n)];\end{aligned}$$

where $\Sigma_n = Diag\{\sigma_{n1}^2, \ldots, \sigma_{nn}^2\}$ *with* $\epsilon_n = (\epsilon_{n1}, \ldots, \epsilon_{nn})'$ *and* $\sigma_{ni}^2 = E(\epsilon_{ni}^2)$.

Proof. (1) $E(\epsilon_n' A_n \epsilon_n) = \sum_{i=1}^n \sum_{j=1}^n a_{n,ij} E(\epsilon_{ni}\epsilon_{nj}) = \sum_{i=1}^n a_{n,ii}\sigma_{ni}^2 = \mathrm{tr}(\Sigma_n A_n)$.

(2) From the proof of part (3) of Lemma A.1, one has

$$\begin{aligned}E(\epsilon_n' A_n \epsilon_n)^2 &= \sum_{i=1}^n a_{n,ii}^2 [E(\epsilon_{ni}^4) - 3\sigma_{ni}^4] + \left(\sum_{i=1}^n a_{n,ii}\sigma_{ni}^2\right)^2 \\ &\quad + \sum_{i=1}^n \sum_{j=1}^n a_{n,ij}(a_{n,ij} + a_{n,ji})\sigma_{ni}^2 \sigma_{nj}^2 \\ &= \sum_{i=1}^n a_{n,ii}^2 [E(\epsilon_{ni}^4) - 3\sigma_{ni}^4] \\ &\quad + \mathrm{tr}^2(\Sigma_n A_n) + \mathrm{tr}[\Sigma_n A_n (A_n' \Sigma_n + \Sigma_n A_n)].\end{aligned}$$

(3) The result follows from (1) and (2) because $\mathrm{Var}(\epsilon_n' A_n \epsilon_n) = E(\epsilon_n' A_n \epsilon_n)^2 - E^2(\epsilon_n' A_n \epsilon_n)$. □

Lemma A.3. *Suppose that* $\{A_n\}$ *are uniformly bounded in both row and column sums and* ϵ_{ni}*'s have a zero mean and are mutually independent where its sequence of variances* $\{\sigma_{ni}^2\}$ *is bounded, and, in addition, if* $a_{n,ii} \neq 0$ *for some i, the sequence four moments* $\{\mu_{ni,4}\}$ *is bounded. Then,* $E(\epsilon_n' A_n \epsilon_n) = O(n)$, $\mathrm{var}(\epsilon_n' A_n \epsilon_n) = O(n)$, $\epsilon_n' A_n \epsilon_n = O_P(n)$, *and* $\frac{1}{n}\epsilon_n' A_n \epsilon_n - \frac{1}{n}E(\epsilon_n' A_n \epsilon_n) = o_P(1)$.

Proof. As σ_{ni}^2's are bounded, the variance matrix $\Sigma_n = Diag\{\sigma_{n1}^2, \ldots, \sigma_{nn}^2\}$ is bounded in both row and column sum norms. The product of two matrices which are uniformly bounded in the row (column) sum norm is uniformly bounded in the row (column) sum norm. Furthermore, elements of uniformly bounded in the row (or column) sum matrices are uniformly bounded.

As $\Sigma_n A_n$ are uniformly bounded in row (or column) sum norm, $E(\epsilon_n' A_n \epsilon_n) = \mathrm{tr}(\Sigma_n A_n) = O(n)$.

From Lemma A.2, the variance of $\epsilon_n' A_n \epsilon_n$ is $\sum_{i=1}^n a_{n,ii}^2(\mu_{ni,4} - 3\sigma_{ni}^4) + \mathrm{tr}[\Sigma_n A_n(A_n'\Sigma_n + \Sigma_n A_n)]$. As $\Sigma_n A_n$ is uniformly bounded in row or column sums, it implies $\mathrm{tr}(\Sigma_n A_n A_n' \Sigma_n)$ and $\mathrm{tr}(\Sigma_n A_n \Sigma_n A_n)$ are $O(n)$. In addition, if $a_{n,ii}$'s are not zero, the uniform boundedness of σ_{ni}^2 and $\mu_{ni,4}$ will guarantee that $\sum_{i=1}^n a_{n,ii}^2(\mu_{ni,4} - 3\sigma_{ni}^4)$ is $O(n)$. Hence, $\mathrm{var}(\epsilon_n' A_n \epsilon_n) = O(n)$ follows.

As $E(\epsilon_n' A_n \epsilon_n)^2 = \mathrm{var}(\epsilon_n' A_n \epsilon_n) + E^2(\epsilon_n' A_n \epsilon_n) = O(n^2)$, the generalized Chebyshev inequality implies that $P(\frac{1}{n}|\epsilon_n' A_n \epsilon_n| \geq M) \leq \frac{1}{M^2}(\frac{1}{n})^2 E(\epsilon_n' A_n \epsilon_n)^2 = \frac{1}{M^2}O(1)$ and, hence, $\frac{1}{n}\epsilon_n' A_n \epsilon_n = O_P(1)$. Finally, because $\mathrm{var}(\frac{1}{n}\epsilon_n' A_n \epsilon_n) = O(\frac{1}{n}) = o(1)$, the Chebyshev inequality implies that $\frac{1}{n}\epsilon_n' A_n \epsilon_n - \frac{1}{n}E(\epsilon_n' A_n \epsilon_n) = o_P(1)$. □

Lemma A.4. *Suppose that* A_n *is an* $n \times n$ *matrix with its column sums being uniformly bounded, elements of the* $n \times k$ *matrix* C_n *are uniformly bounded, and elements* ϵ_{ni} *of* $\epsilon_n = (\epsilon_{n1}, \ldots, \epsilon_{nn})'$ *are mutually independent with zero mean and finite third absolute moments, which are uniformly bounded for all n and i.*

Then, $\frac{1}{\sqrt{n}} C_n' A_n \epsilon_n = O_P(1)$ *and* $\frac{1}{n} C_n' A_n \epsilon_n = o_P(1)$. *Furthermore, if the limit of* $\frac{1}{n} C_n' A_n \Sigma_n A_n' C_n$ *exists and is positive definite, then* $\frac{1}{\sqrt{n}} C_n' A_n \epsilon_n \xrightarrow{D} N(0, \lim_{n\to\infty} \frac{1}{n} C_n' A_n \Sigma_n A_n' C_n)$.

Proof. Let $a_{n,j}$ denote the jth column of A_n. It follows that $\frac{1}{\sqrt{n}} C_n' A_n \epsilon_n = \frac{1}{\sqrt{n}} \sum_{j=1}^n q_{nj}\epsilon_j$ where $q_{nj} = C_n' a_{n,j}$. The first result follows from Chebyshev's inequality because $\{q_{nj}\}$ and $\{\sigma_{nj}^2\}$ are uniformly bounded and $\mathrm{var}(\frac{1}{\sqrt{n}} C_n' A_n \epsilon_n) = \frac{1}{n}\sum_{j=1}^n \sigma_{nj}^2 q_{nj} q_{nj}'$. The second result follows from the Liapounov double array CLT and the Cramer-Wold device (Billingsley, 1995, Theorem 27.3 and Theorem 29.4). To check the Liapounov condition, let α be a non-zero row vector of constants and $B_n^2 = var(\alpha C_n' A_n \epsilon_n) = \sigma^2 \alpha C_n' A_n \Sigma_n A_n' C_n \alpha'$. The assumptions imply that $\lim_{n\to\infty} \frac{1}{n} B_n^2 > 0$ and there exist constants c_1 and c_2 such that $|\alpha q_{nj}| < c_1$ and $E|\epsilon_{ni}|^3 < c_2$, for all n and j. Hence, the Liapounov condition $\sum_{j=1}^n \frac{1}{B_n^3} E(|\alpha q_{nj}\epsilon_j|^3) \leq \frac{c_1^3 c_2}{(\frac{1}{n}B_n^2)^{\frac{3}{2}} n^{\frac{1}{2}}} \to 0$ holds. □

Lemma A.5. *Suppose that* $\{A_n\}$ *is a sequence of symmetric* $n \times n$ *matrices with row and column sums uniformly bounded and* $b_n = [b_{ni}]$ *is a n-dimensional column vector such that* $\sup_n \frac{1}{n}\sum_{i=1}^n |b_{ni}|^{2+\eta_1} < \infty$ *for some* $\eta_1 > 0$. *The* $\epsilon_{n1}, \ldots, \epsilon_{nn}$ *are mutually independent, with a zero mean and moments higher than four exist such that* $E(|\epsilon_{ni}|^{4+\eta_2})$ *for some* $\eta_2 > 0$, *for all n and i, are uniformly bounded.*

Let $\sigma_{Q_n}^2$ *be the variance of* Q_n *where* $Q_n = \epsilon_n' A_n \epsilon_n + b_n'\epsilon_n - \mathrm{tr}(A_n\Sigma_n)$. *Assume that* $\frac{1}{n}\sigma_{Q_n}^2$ *is bounded away from zero. Then,* $\frac{Q_n}{\sigma_{Q_n}} \xrightarrow{D} N(0, 1)$.

Proof. See Kelejian and Prucha (2001). □

Proof of Proposition 1. For consistency of an extremum estimate, a standard approach can follow, for example, the setting in Theorem 4.1.1 of Amemiya (1985). Let $s_n(\theta) = \frac{1}{n}a_n g_n(\theta)$. The essential ingredients in that theorem are (i) a compact parameter space Θ of θ, (ii) $s_n(\theta)$ is continuous in θ, (iii) $s_n(\theta)$ converges in probability to $s(\theta)$, where $s(\theta) = \lim_{n\to\infty} \frac{1}{n}a_n g_n(\theta)$, uniformly in $\theta \in \Theta$, and (iv) $s(\theta)$ has the unique global extremum at θ_0 in Θ. The (iv) is an identification condition, which will be satisfied under our identification assumptions. For our case, the compactness of Θ can be replaced by boundedness because $s_n(\theta)$ is simply a polynomial function of θ. The continuity of $s_n(\theta)$ in (ii) is obvious. So it remains to demonstrate the uniform convergence of $s_n(\theta)$ to $s(\theta)$ in (iii). Let $a_n = (a_{n1}, \ldots, a_{nm}, a_{nx})$, where a_{nj} is jth column of the matrix, a_{nx} is a submatrix. Then let $a_{i,n}$ be the ith row of the matrix a_n. Furthermore, explicitly, denote $a_{i,n} = (a_{i,n1}, \ldots, a_{i,nm}, a_{i,nx})$ where $a_{i,nj}, j = 1, \ldots, m$, are scalars, and $a_{i,nx}$ is a row subvector with its dimension k^* as the number of rows of Q_n. It is sufficient to consider the uniform convergence of $a_{i,n}g(\theta)$ for each i. Then $a_{i,n}g_n(\theta) = \epsilon_n'(\theta)(\sum_{j=1}^m a_{i,nj}P_{jn})\epsilon_n(\theta) + a_{i,nx}Q_n'\epsilon_n(\theta)$. Because $S_n(\lambda) = S_n + (\lambda_0 - \lambda)W_n$, by expansion, $\epsilon_n(\theta) = d_n(\theta) + \epsilon_n + (\lambda_0 - \lambda)G_n\epsilon_n$ where $d_n(\theta) = (\lambda_0 - \lambda)G_n X_n \beta_0 + X_n(\beta_0 - \beta)$. It follows that $\epsilon_n'(\theta)(\sum_{j=1}^m a_{i,nj}P_{jn})\epsilon_n(\theta) = d_n'(\theta)(\sum_{j=1}^m a_{i,nj}P_{jn})d_n(\theta) + l_n(\theta) + q_n(\theta)$, where $l_n(\theta) = d_n'(\theta)(\sum_{j=1}^m a_{i,nj}P_{jn}^s)(\epsilon_n + (\lambda_0 - \lambda)G_n\epsilon_n)$ and $q_n(\theta) = (\epsilon_n' + (\lambda_0 - \lambda)\epsilon_n' G_n')(\sum_{j=1}^m a_{i,nj}P_{jn})(\epsilon_n + (\lambda_0 - \lambda)G_n\epsilon_n)$. The term $l_n(\theta)$ is linear in ϵ_n. By expansion,

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

$$+ (\lambda_0 - \lambda)^2 \frac{1}{n}(X_n\beta_0)' G_n' \left(\sum_{j=1}^m a_{i,nj} P_{jn}^s\right) G_n\epsilon_n$$
$$+ (\lambda_0 - \lambda)(\beta_0 - \beta)' \frac{1}{n} X_n' \left(\sum_{j=1}^m a_{i,nj} P_{jn}^s\right) G_n\epsilon_n$$
$$= o_P(1),$$

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,

$$\frac{1}{n}q_n(\theta) = \frac{1}{n}\epsilon_n' \left(\sum_{j=1}^m a_{i,nj} P_{jn}\right)\epsilon_n + (\lambda_0 - \lambda)\frac{1}{n}\epsilon_n' G_n' \left(\sum_{j=1}^m a_{i,nj} P_{jn}^s\right)\epsilon_n$$
$$+ (\lambda_0 - \lambda)^2 \frac{1}{n}\epsilon_n' G_n' \left(\sum_{j=1}^m a_{i,nj} P_{jn}\right) G_n\epsilon_n$$
$$= (\lambda_0 - \lambda)\frac{1}{n}\sum_{j=1}^m a_{i,nj}\mathrm{tr}(\Sigma_n G_n' P_{jn}^s)$$
$$+ (\lambda_0 - \lambda)^2 \frac{1}{n}\sum_{j=1}^m a_{i,nj}\mathrm{tr}(\Sigma_n G_n' P_{jn} G_n) + o_P(1),$$

uniformly in $\theta \in \Theta$, by Lemmas A.2 and A.3, and $E(\epsilon_n' P_{jn}\epsilon_n) = \mathrm{tr}(\Sigma_n P_{jn}) = \mathrm{tr}(\Sigma_n \cdot Diag\{P_{jn}\}) = 0$ for all $j = 1, \ldots, m$ because $Diag\{P_{jn}\} = 0$ by design. Consequently,

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

uniformly in $\theta \in \Theta$. The consistency of the GMME $\widehat{\theta}_n$ follows from this uniform convergence and the identification condition.

For the asymptotic distribution of $\widehat{\theta}_n$, by Taylor's expansion of $\frac{\partial g_n'(\widehat{\theta}_n)}{\partial\theta} a_n' a_n g_n(\widehat{\theta}_n) = 0$ at θ_0,[23]

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

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

$$\frac{1}{n}\epsilon_n'(\theta)P_{jn}^s G_nX_n\beta_0 = \frac{1}{n}d_n'(\theta)P_{jn}^s G_nX_n\beta_0 + \frac{1}{n}\epsilon_n' P_{jn}^s G_nX_n\beta_0$$
$$+ (\lambda_0 - \lambda)\frac{1}{n}\epsilon_n' G_n' P_{jn}^s G_nX_n\beta_0$$
$$= \frac{1}{n}d_n'(\theta)P_{jn}^s G_nX_n\beta_0 + o_P(1),$$

and

$$\frac{1}{n}\epsilon_n'(\theta)P_{jn}^s G_n\epsilon_n$$
$$= \frac{1}{n}d_n'(\theta)P_{jn}^s G_n\epsilon_n + \frac{1}{n}\epsilon_n' P_{jn}^s G_n\epsilon_n + \frac{1}{n}(\lambda_0 - \lambda)\epsilon_n' G_n' P_{jn}^s G_n\epsilon_n$$
$$= \frac{1}{n}\mathrm{tr}(\Sigma_n P_{jn}^s G_n) + (\lambda_0 - \lambda)\frac{1}{n}\mathrm{tr}(\Sigma_n G_n' P_{jn}^s G_n) + o_P(1),$$

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

$$\frac{1}{n}\epsilon_n'(\theta)P_{jn}^s W_nY_n = \frac{1}{n}d_n'(\theta)P_{jn}^s G_nX_n\beta_0 + \frac{1}{n}\mathrm{tr}(\Sigma_n P_{jn}^s G_n)$$
$$+ (\lambda_0 - \lambda)\frac{1}{n}\mathrm{tr}(\Sigma_n G_n' P_{jn}^s G_n) + o_P(1),$$

uniformly in $\theta \in \Theta$. At θ_0, $d_n(\theta_0) = 0$ and, hence, $\frac{1}{n}\epsilon_n'(\theta_0) P_{jn}^s W_nY_n = \frac{1}{n}\mathrm{tr}(\Sigma_n P_{jn}^s G_n) + o_P(1)$. At θ_0, $\frac{1}{n}\epsilon_n'(\theta_0)P_{jn}^s X_n = o_P(1)$. Finally, $\frac{1}{n}Q_n' W_nY_n = \frac{1}{n}Q_n' G_nX_n\beta_0 + \frac{1}{n}Q_n' G_n\epsilon_n = \frac{1}{n}Q_n' G_nX_n\beta_0 + o_P(1)$. In conclusion, $\frac{1}{n}\frac{\partial g_n(\widetilde{\theta}_n)}{\partial\theta} = -\frac{1}{n}D_n + o_P(1)$ with D_n in (14). On the other hand, Lemma A.5 implies that $\frac{1}{\sqrt{n}}a_n g_n(\theta_0) = \frac{1}{\sqrt{n}}[\epsilon_n'(\sum_{j=1}^m a_{nj}P_{jn})\epsilon_n + a_{nx}Q_n'\epsilon_n] \xrightarrow{D} N(0, \lim_{n\to\infty}\frac{1}{n}a_n\Omega_n a_n')$. The asymptotic distribution of $\sqrt{n}(\widehat{\lambda}_n - \lambda_0)$ follows. □

Proof of Proposition 2. A. The consistency of $\frac{1}{n}\widehat{\Omega}_n$: We shall show that each element in $\frac{1}{n}\widehat{\Omega}_n - \frac{1}{n}\Omega_n$ is of the order of $o_p(1)$.

(a) The consistency of some elements: One generic form of the elements in the matrix $\frac{1}{n}\Omega_n$ is $\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\sigma_{ni}^2\sigma_{nj}^2$, with $P_{\Delta n,ij} = P_{an,ij}(P_{bn,ij} + P_{bn,ji})$, note that $P_{\Delta n,ii} = 0$. We shall first show that $\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\epsilon_{ni}^2\epsilon_{nj}^2 - \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\sigma_{ni}^2\sigma_{nj}^2 = o_p(1)$, then we establish that this convergence holds when ϵ_{ni}'s are replaced by the residuals $\widehat{\epsilon}_{ni}$'s.

(i) Show that $\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\epsilon_{ni}^2\epsilon_{nj}^2 - \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\sigma_{ni}^2\sigma_{nj}^2 = o_p(1)$.

Define the $n \times n$ matrix $P_{\Delta n} = [P_{\Delta n,ij}]$. Because P_{bn} is uniformly bounded in either the row or column sum norms, its elements are uniformly bounded, i.e., there exists a constant c such that $|P_{bn,ij} + P_{bn,ji}| \le c$ for all i, j and n. Therefore $|P_{\Delta n,ij}| \le c|P_{an,ij}|$. Because P_{an} is uniformly bounded in both the row and column norms, it follows that $P_{\Delta n}$ is uniformly bounded in both the row and colum sum norms.

As $\epsilon_{ni}^2\epsilon_{nj}^2 - \sigma_{ni}^2\sigma_{nj}^2 = (\epsilon_{ni}^2 - \sigma_{ni}^2)(\epsilon_{nj}^2 - \sigma_{nj}^2) + \sigma_{ni}^2(\epsilon_{nj}^2 - \sigma_{nj}^2) + \sigma_{nj}^2(\epsilon_{ni}^2 - \sigma_{ni}^2)$, one has

$$\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}(\epsilon_{ni}^2\epsilon_{nj}^2 - \sigma_{ni}^2\sigma_{nj}^2) = Q_n + L_{n1} + L_{n2},$$

where $Q_n = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}(\epsilon_{ni}^2 - \sigma_{ni}^2)(\epsilon_{nj}^2 - \sigma_{nj}^2)$, $L_{n1} = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \sigma_{nj}^2 P_{\Delta n,ij}(\epsilon_{ni}^2 - \sigma_{ni}^2)$, and $L_{n2} = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \sigma_{ni}^2 P_{\Delta n,ij}(\epsilon_{nj}^2 - \sigma_{nj}^2)$. Define vectors $u_n = (u_{n1}, \ldots, u_{nn})$ where $u_{ni} = \epsilon_{ni}^2 - \sigma_{ni}^2$, and $C_{\sigma n} = (\sigma_{n1}^2, \ldots, \sigma_{nn}^2)$. It follows that $Q_n = \frac{1}{n}u_n' P_{\Delta n}u_n$, $L_{n1} = \frac{1}{n}u_n' P_{\Delta n}C_{\sigma n}'$, and $L_{n2} = \frac{1}{n}C_{\sigma n}P_{\Delta n}u_n$. As $E(u_n' P_{\Delta n}u_n) = \mathrm{tr}(P_{\Delta n}\Lambda_n)$ where $\Lambda_n = E(u_nu_n') = \mathrm{Diag}\{\mu_{n1,4} - \sigma_{n1}^4, \ldots, \mu_{nn,4} - \sigma_{nn}^4\}$ is a diagonal matrix, $E(u_n' P_{\Delta n}u_n) = \mathrm{tr}(\mathrm{Diag}(P_{\Delta n})\Lambda_n) = 0$ because $P_{\Delta n,ii} = 0$ for all i. It follows by Lemma A.3 that $Q_n = o_P(1)$. On the other hand, Lemma A.4 gives $L_{n1} = o_p(1)$ and $L_{n2} = o_p(1)$. Hence, we conclude the convergence in (i). Next, we'll show that the ϵ_{ni}'s can be replaced by the residuals $\widehat{\epsilon}_{ni}$'s.

(ii) Show that $\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\widehat{\epsilon}_{ni}^2\widehat{\epsilon}_{nj}^2 - \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\epsilon_{ni}^2\epsilon_{nj}^2 = o_p(1)$. Now

$$\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\widehat{\epsilon}_{ni}^2\widehat{\epsilon}_{nj}^2 - \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\epsilon_{ni}^2\epsilon_{nj}^2 = B_{n1} + B_{n2} + B_{n3},$$

where $B_{n1} = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\epsilon_{nj}^2(\widehat{\epsilon}_{ni}^2 - \epsilon_{ni}^2)$, $B_{n2} = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\epsilon_{ni}^2(\widehat{\epsilon}_{nj}^2 - \epsilon_{nj}^2)$, and $B_{n3} = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}(\widehat{\epsilon}_{ni}^2 - \epsilon_{ni}^2)(\widehat{\epsilon}_{nj}^2 - \epsilon_{nj}^2)$.

[23] Note that the Taylor's expansion of $\frac{\partial g_n'(\hat{\theta}_n)}{\partial\theta}a_n' a_n g_n(\hat{\theta}_n)$ is only to expand the component $g(\hat{\theta}_n)$ at θ_0 but not the component $\frac{\partial g_n'(\hat{\theta}_n)}{\partial\theta}$. So the second order derivative of $g_n(\theta)$ would not be needed. This simplifies our analysis.

From the model, we get

$$\widehat{\epsilon}_n = S_n(\widehat{\lambda})Y_n - X_n\widehat{\beta} = \epsilon_n + (\lambda_0 - \widehat{\lambda})G_n\epsilon_n + X_n(\beta_0 - \widehat{\beta}) + (\lambda_0 - \widehat{\lambda})G_nX_n\beta_0$$

In a scalar form, $\widehat{\epsilon}_{ni} = \epsilon_{ni} + b_{ni} + c_{ni}$, where $b_{ni} = (\lambda_0 - \widehat{\lambda})(e_{i,n}G_n\epsilon_n)$ and $c_{ni} = e_{i,n}X_n(\beta_0 - \widehat{\beta}) + (\lambda_0 - \widehat{\lambda})e_{i,n}G_nX_n\beta_0$, where $e_{i,n}$ is the ith row in the $n \times n$ identity matrix. Thus $\widehat{\epsilon}^2_{ni} = \epsilon^2_{ni} + b^2_{ni} + c^2_{ni} + 2\epsilon_{ni}b_{ni} + 2\epsilon_{ni}c_{ni} + 2b_{ni}c_{ni}$. We shall consider that all the three terms B_{nl}, $l = 1, 2, 3$, converge to zero in probability. Let's consider B_{n1}

$$B_{n1} = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\epsilon^2_{nj}(\widehat{\epsilon}^2_{ni} - \epsilon^2_{ni}) = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\epsilon^2_{nj}[b^2_{ni} + c^2_{ni} + 2\epsilon_{ni}b_{ni} + 2\epsilon_{ni}c_{ni} + 2b_{ni}c_{ni}].$$

We want to show this is $o_p(1)$. We shall pay special attention to those terms with the higher orders in ϵ's. The other remaining terms are simpler. An example of such a term is

$$\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\epsilon^2_{nj}\epsilon_{ni}b_{ni} = (\lambda_0 - \widehat{\lambda})\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\sum_{l=1}^n P_{\Delta n,ij}G_{n,il}\epsilon_{ni}\epsilon^2_{nj}\epsilon_{nl}.$$

As $\widehat{\lambda} - \lambda_0 = o_p(1)$, this will be $o_p(1)$ if $\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\sum_{l=1}^n P_{\Delta n,ij} G_{n,il}\epsilon_{ni}\epsilon^2_{nj}\epsilon_{nl}$ is stochastically bounded. By Cauchy's inequality, $E|\epsilon_{ni}\epsilon_{nl}\epsilon^2_{nj}| \le [E(\epsilon_{ni}\epsilon_{nl})^2]^{\frac{1}{2}}E^{\frac{1}{2}}(\epsilon^4_{nj}) \le E^{\frac{1}{4}}(\epsilon^4_{ni})E^{\frac{1}{4}}(\epsilon^4_{nl})E^{\frac{1}{4}}(\epsilon^4_{nj}) \le c$ for some constant c, for all i, j, l, and n because $\{\mu_{ni,4}\}$ is a bounded sequence. It follows that

$$E\left|\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\sum_{l=1}^n P_{\Delta n,ij}G_{n,il}\epsilon_{ni}\epsilon^2_{nj}\epsilon_{nl}\right| \le c\frac{1}{n}\sum_{i=1}^n\left(\sum_{j=1}^n|P_{\Delta n,ij}|\right)\left(\sum_{l=1}^n|G_{n,il}|\right) = O(1),$$

because $P_{\Delta,n}$ and G_n are uniformly bounded in row and column sums. By the Markov inequality, it implies that $\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\sum_{l=1}^n P_{\Delta n,ij}G_{n,il}\epsilon_{ni}\epsilon^2_{nj}\epsilon_{nl} = O_p(1)$.

Another term with high order ϵ's is

$$\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\epsilon^2_{nj}b^2_{ni} = (\lambda_0 - \widehat{\lambda})^2\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n P_{\Delta n,ij}G_{n,ik}G_{n,il}\epsilon^2_{nj}\epsilon_{nk}\epsilon_{nl} = o_p(1),$$

because

$$E\left|\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n P_{\Delta n,ij}G_{n,ik}G_{n,il}\epsilon^2_{nj}\epsilon_{nk}\epsilon_{nl}\right| \le c\frac{1}{n}\sum_{i=1}^n\left(\sum_{j=1}^n|P_{\Delta n,ij}|\right)\left(\sum_{k=1}^n|G_{n,ik}|\right)\left(\sum_{l=1}^n|G_{n,il}|\right) = O(1).$$

The remaining terms in B_{n1} are simpler and the same arguments with the Markov inequality shall be applicable. Thus $B_{n1} = o_p(1)$. B_{n2} has a similar structure as B_{n1}, because i is replaced by j and vice versa. So $B_{n2} = o_p(1)$.

It remains to consider B_{n3}, which is

$$B_{n3} = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}[b^2_{ni} + c^2_{ni} + 2\epsilon_{ni}b_{ni} + 2\epsilon_{ni}c_{ni} + 2b_{ni}c_{ni}] \times [b^2_{nj} + c^2_{nj} + 2\epsilon_{nj}b_{nj} + 2\epsilon_{nj}c_{nj} + 2b_{nj}c_{nj}].$$

The highest order term with ϵ's is

$$\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}b^2_{ni}b^2_{nj} = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}(e_{i,n}G_n\epsilon_n)(e_{j,n}G_n\epsilon_n)(\lambda_0 - \widehat{\lambda})^2 = (\lambda_0 - \widehat{\lambda})^2K_n,$$

where $K_n = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\sum_{k_1=1}^n\sum_{k_2=1}^n\sum_{l_1=1}^n\sum_{l_2=1}^n P_{\Delta n,ij}G_{n,ik_1} G_{n,ik_2}G_{n,jl_1}G_{n,jl_2}\epsilon_{nk_1}\epsilon_{nk_2}\epsilon_{nl_1}\epsilon_{nl_2}$. The Cauchy inequality implies that $E|\epsilon_{nk_1}\epsilon_{nk_2}\epsilon_{nl_1}\epsilon_{nl_2}| \le \mu_{nk_1,4}\mu_{nk_2,4}\mu_{nl_1,4}\mu_{nl_2,4} \le c$, for some constant c for all n. By the uniform boundedness in row and column sums for $P_{\Delta,n}$ and G_n,

$$E|K_n| \le \frac{c}{n}\sum_{i=1}^n\left(\sum_{j=1}^n|P_{\Delta n,ij}|\right)\left(\sum_{k_1=1}^n|G_{n,ik_1}|\right)\left(\sum_{k_2=1}^n|G_{n,ik_2}|\right) \times\left(\sum_{l_1=1}^n|G_{n,jl_1}|\right)\left(\sum_{l_2=1}^n|G_{n,jl_2}|\right) = O(1),$$

which implies that $K_n = O_p(1)$ by the Markov inequality. Other terms in B_{n3} can similarly be analyzed. Thus, we conclude that $B_{n3} = o_P(1)$.

Therefore, $\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\widehat{\epsilon}^2_{ni}\widehat{\epsilon}^2_{nj} - \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\epsilon^2_{ni}\epsilon^2_{nj} = o_p(1)$. Combining (i) and (ii), we have $\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\widehat{\epsilon}^2_{ni}\widehat{\epsilon}^2_{nj} - \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{\Delta n,ij}\sigma^2_{ni}\sigma^2_{nj} \xrightarrow{p} 0$.

(b) The consistency of the other elements: The other elements in the matrix $\frac{1}{n}\Omega_n$ are of the form $\frac{1}{n}Q_n^{'}\sum_n Q_n = \frac{1}{n}\sum_{i=1}^n\sigma^2_{ni}q_i'q_i$. With similar arguments in (a) or arguments as in White (1980), $\frac{1}{n}\sum_{i=1}^n\widehat{\epsilon}^2_{ni}q_i'q_i \xrightarrow{P} \frac{1}{n}\sum_{i=1}^n\sigma^2_{ni}q_i'q_i$.

In conclusion, we've shown that $\frac{1}{n}\widehat{\Omega}_n \xrightarrow{P} \frac{1}{n}\Omega_n$.

B. The consistency of $\frac{1}{n}\widehat{D}_n$: One generic form for the elements of $\frac{1}{n}D_n$ is $\frac{1}{n}\sum_{i=1}^n(P^s_{jn}G_n)_{ii}\sigma^2_{ni}$. Since $P_n's$, $G_n's$ are all uniformly bounded in both the row and column sums, so are the matrices $(P^s_{jn}G_n)'s$. Thus $\frac{1}{n}\sum_{i=1}^n(P^s_{jn}G_n)_{ii}\widehat{\epsilon}^2_i - \frac{1}{n}\sum_{i=1}^n(P^s_{jn}G_n)_{ii}\sigma^2_{ni} \xrightarrow{p} 0$ can be shown with the same arguments in part (a) above.

Together, these prove the validity of Proposition 2. □

Proof of Proposition 3. The generalized Schwartz inequality implies that the optimal weighting matrix for $a_n'a_n$ in Proposition 1 is $(\frac{1}{n}\Omega_n)^{-1}$. For consistency, consider $\frac{1}{n}g_n'(\theta)\widehat{\Omega}_n^{-1}g_n(\theta) = \frac{1}{n}g_n'(\theta)\Omega_n^{-1}g_n(\theta) + \frac{1}{n}g_n'(\theta)(\widehat{\Omega}_n^{-1} - \Omega_n^{-1})g_n(\theta)$. With $a_n = (\frac{1}{n}\Omega_n)^{-1/2}$ in Proposition 1, Assumption 6 implies that $a_0 = (\lim_{n\to\infty}\frac{1}{n}\Omega_n)^{-1/2}$ exists. Because a_0 is nonsingular, the identification condition of θ_0 corresponds to the unique root of $\lim_{n\to\infty}E(\frac{1}{n}g_n(\theta)) = 0$ at θ_0, which is satisfied by Assumption 5. Hence, the uniform convergence in probability of $\frac{1}{n}g_n'(\theta)\Omega_n^{-1}g_n(\theta)$ 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 $\frac{1}{n}g_n'(\theta)(\widehat{\Omega}_n^{-1} - \Omega_n^{-1})g_n(\theta) = o_P(1)$ uniformly in $\theta \in \Theta$. Let $\|\cdot\|$ be the Euclidean norm or the maximum row sum norm for vectors and matrices. Then, $\|\frac{1}{n}g_n'(\theta)(\widehat{\Omega}_n^{-1} - \Omega_n^{-1})g_n(\theta)\| \le (\frac{1}{n}\|g_n(\theta)\|)^2\|(\frac{\widehat{\Omega}_n}{n})^{-1} - (\frac{\Omega_n}{n})^{-1}\|$. From the proof of Proposition 1, $\frac{1}{n}[g_n(\theta) - E(g_n(\theta))] = o_P(1)$ uniformly in $\theta \in \Theta$. On the other hand, as

$$\frac{1}{n}d_n'(\theta)P_{jn}d_n(\theta) = (\lambda_0 - \lambda)^2\frac{1}{n}(X_n\beta_0)'G_n'P_{jn}G_n(X_n\beta_0) + (\lambda_0 - \lambda)\frac{1}{n}(X_n\beta_0)'G_n'P^s_{jn}X_n(\beta_0 - \beta) + (\beta_0 - \beta)'\frac{1}{n}X_n'P_{jn}X_n(\beta_0 - \beta) = O_P(1),$$

uniformly in $\theta \in \Theta$, $\frac{1}{n}E(\epsilon_n'(\theta)P_{jn}\epsilon_n(\theta)) = \frac{1}{n}d_n'(\theta)P_{jn}d_n(\theta) + (\lambda_0 - \lambda)\frac{1}{n}\text{tr}(\Sigma_n P_{jn}^s G_n) + (\lambda_0 - \lambda)^2\frac{1}{n}\text{tr}(\Sigma_n G_n' P_{jn} G_n) = O(1)$, uniformly in $\theta \in \Theta$. Similarly, $\frac{1}{n}E(Q_n'\epsilon_n(\theta)) = \frac{1}{n}Q_n'd_n(\theta) = (\lambda_0 - \lambda)\frac{1}{n}Q_n'G_nX_n\beta_0 + \frac{1}{n}Q_n'X_n(\beta_0 - \beta) = O(1)$ uniformly in $\theta \in \Theta$. These imply that $\| \frac{1}{n}E(g_n(\theta)) \| = O(1)$ uniformly in $\theta \in \Theta$. Consequently, by the Markov inequality, $\frac{1}{n} \| g_n(\theta) \| = O_P(1)$ uniformly in $\theta \in \Theta$. Therefore, $\| \frac{1}{n}g_n'(\theta)(\widehat{\Omega}_n^{-1} - \Omega_n^{-1})g_n(\theta) \|$ converges in probability to zero, uniformly in $\theta \in \Theta$. The consistency of the feasible optimum GMME $\widehat{\theta}_{o,n}$ follows.

For the limiting distribution, as $\frac{1}{n}\frac{\partial g_n(\widehat{\theta}_n)}{\partial\theta} = -\frac{D_n}{n} + o_P(1)$ from the proof of Proposition 1,

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

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

References

Amemiya, T., 1985. Advanced Econometrics. Basil Blackwell, Oxford.
Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, The Netherlands.
Billingsley, P., 1995. Probability and Measure, 3rd ed.. John Wiley and Sons, New York, NY.
Breusch, T., Pagan, A., 1979. A simple test for heteroskedaticity and random coefficient variation. Econometrica 47, 1287–1294.
Case, A.C., 1991. Spatial patterns in household demand. Econometrica 59, 953–965.
Case, A.C., Katz, L.F., 1991. The company you keep: The effects of family and neighborhood on disadvantaged youths, NBER working paper no. w3705 (NBER, Cambridge, MA).
Crane, J., 1991. The epidemic theory of ghettos and neighborhood effects on dropping out and teenage childbearing. American Journal of Sociology 96, 1226–1259.
Evans, W.N., Oates, W.E., Schwab, R.M., 1992. Measuring peer group effects: A study of teenage behavior. Journal of Political Economy 100, 966–991.
Glaeser, E.L., Sacerdote, B., Scheinkman, J.A., 1996. Crime and social interactions. Quarterly Journal of Economics 111, 507–548.
Hausman, J.A., 1978. Specification tests in econometrics. Econometrica 46, 1251–1271.
Hogan, D.P., Kitagawa, E.M., 1985. The impact of social status, family structure, and neighborhood on the fertility of black adolescents. American Journal of Sociology 90, 825–855.
Jencks, C, Mayer Jr., S., 1990. The social consequences of growing up in a poor neighborhood. In: Lynn, L.E., McGeary, M.G.H. (Eds.), Inner-city Poverty in the United States. National Academy, Washington, DC.
Kelejian, H.H., Robinson, D., 1993. A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. Papers in Regional Science 72, 297–312.
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., Prucha, I.R., 2010. Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. Journal of Econometrics 157, 53–67.
Lee, L-f., 2001. Generalized method of moments estimation of spatial autoregressive processes, Unpublished manuscript (The Ohio State University, 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.
Lee, L-f., 2007a. GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. Journal of Econometrics 137, 489–514.
Lee, L-f., 2007b. The method of elimination and substitution in the GMM estimation of mixed regressive, spatial autoregressive models. Journal of Econometrics 140, 155–189.
Lee, L-f., 2007c. Identification and estimation of spatial econometric models with group interactions, contextual factors and fixed effects. Journal of Econometrics 140, 333–374.
LeSage, J.P., 1999. The theory and practice of spatial econometrics. www.spatial-econometrics.com.
Ord, J., 1975. Estimation methods for models of spatial interaction. Journal of the American Statistical Association 70, 120–126.
Ruud, P.A., 2000. Classical Econometric Theory. Oxford University Press, New York, NY.
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.
Thomas, R.K., 1999. Health and Healthcare in the United States-County and Metro Area Data. Bernan Press, Lanham MD.
Census Bureau, U.S., 1992. Census of Population and Housing 1990, Summary Tape File 3 on CD-ROM. The Bureau Producer and Distributor, Washington, DC.
White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–838.