Improving the J Test in the SARAR Model by Likelihood-based Estimation

Peter Burridge

Available online: 14 Feb 2012

To cite this article: Peter Burridge (2012): Improving the J Test in the SARAR Model by Likelihood-based Estimation, Spatial Economic Analysis, 7:1, 75-107

To link to this article: http://dx.doi.org/10.1080/17421772.2011.647055

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Improving the J Test in the SARAR Model by Likelihood-based Estimation

PETER BURRIDGE

(Received 27 July 2010; accepted 18 May 2011)

ABSTRACT  It has been demonstrated recently that in small-to-medium samples the empirical significance levels of the asymptotic J-type tests for the SARAR model introduced by Kelejian (2008) can be controlled in many cases by the use of a bootstrap to construct a reference distribution. A feature of the popular GMM estimator in this context that deserves to receive more attention is that in small samples it will often deliver spatial parameter estimates that lie outside the invertibility region of the model. Using such illegitimate estimates to construct bootstrap samples is then problematic; the present paper finds that this practical obstacle may be removed by the use of quasi-maximum likelihood estimates that guarantee invertibility. The effects of different spatial weight patterns and sample size on the empirical significance levels and power of the tests are illustrated, and the paper demonstrates that estimation using QMLE, allied to a simple bootstrap, yields tests with reliable significance levels and reasonable power, in a majority of cases.

Optimisation du test ‘J’ dans le modèle SARAR, à l’aide d’une estimation à base de vraisemblance

RÉSUMÉ  dans des échantillons petits à moyens, il est possible, dans de nombreux cas, de contrôler les niveaux à signification empirique des tests asymptotiques introduits par Kelejian (2008) à l’aide d’un ‘bootstrap’. Dans ce contexte, une caractéristique de l’estimateur GMM, très répandu, est qu’il fournit, dans de petits échantillons, des estimations de paramètres spatiaux situés hors de la région d’inversibilité du modèle. L’emploi de telles estimations illégitimes pour la réalisation d’échantillons ‘bootstrap’ devient alors problématique; la présente communication indique que l’on peut supprimer cet obstacle pratique en utilisant le QMLE garantissant l’inversibilité. Les effets des tendances du poids spatial et de la taille des échantillons sur les niveaux d’importance et la puissance sont illustrés, et la communication démontre que le QMLE, allié à un simple ‘bootstrap’, permet de réaliser des tests offrant, dans la plupart des cas, des niveaux d’importance fiables et une puissance raisonnable.

Department of Economics and Related Studies, University of York, Heslington, York YO10 5DD, UK. Email: peter.burridge@york.ac.uk. Thanks to colleagues and participants at the June 2010 Spatial Statistics and Econometrics workshop in Orléans, especially Jesús Mur, James LeSage, and Christine Thomas-Agnan, and also to two anonymous referees, all of whose comments have improved the paper. Remaining deficiencies are of course solely the author’s responsibility.

ISSN 1742-1772 print; 1742-1780 online/12/010075-33  http://dx.doi.org/10.1080/17421772.2011.647055
© 2012 Regional Studies Association
Mejora de la prueba J en el modelo SARAR mediante la estimación basada en probabilidad

**Extracto**  En muestras entre pequeñas y medianas, los niveles de significancia empírica de las pruebas asintóticas de tipo J para el modelo SARAR introducidas por Kelejian (2008) pueden controlarse en muchos casos mediante el uso de un bootstrap. Una característica del popular estimador GMM dentro de este contexto es que en las muestras pequeñas, a menudo producirá estimaciones de parámetros espaciales que están fuera de la región de reversibilidad del modelo. No obstante, el empleo de este tipo de estimaciones ilegítimas para construir muestras bootstrap es problemático; el estudio actual muestra que este obstáculo práctico puede eliminarse mediante el uso del QMLE que garantiza la reversibilidad. Se ilustran los efectos de las pautas de peso espacial y del tamaño de la muestra sobre el poder y los niveles de significancia, y el estudio demuestra que el QMLE, aliado a un bootstrap simple, dota a las pruebas de niveles de significancia fiables y de un poder razonable, en la mayoría de los casos.

**Key Words:** J test; maximum likelihood; SARAR model; bootstrap

**JEL Classification:** C21; C52

1. Introduction

Kelejian (2008) introduces spatial extensions of the J-test of Davidson and MacKinnon (1981) for testing a null model, Model₀, against g alternatives in which it is not nested. In Kelejian’s 2008 paper, the tests are constructed for SARAR models using the GMM-type estimators of Kelejian & Prucha (1998, 1999, 2010) (KP), and in small samples as found by Burridge & Fingleton (2010) (BF) they can be too liberal to be useful in various parts of the parameter space. To correct this size-inflation problem, BF proposed use of a parametric bootstrap; this was quite effective, but was subject to difficulties when, as happened quite often in their experiments, the GMM estimators applied to the original data delivered spatial parameter estimates lying outside the invertibility region. Such estimates are, of course, illegitimate, in the sense that the underlying rationale of the model breaks...
down (see LeSage & Pace (2009, p. 26), for more discussion on this point). However, if the device of bootstrapping the J-tests is to be widely applied, this problem must be resolved, and the present paper explores an approach based on numerical maximization of a Gaussian likelihood that ensures that the estimated model is invertible. The BF paper left open a number of questions, among them how sensitive to the form of the weights the J-tests’ properties are, and to what extent the test size inflation is a small sample problem. Both questions are addressed in the present paper, and it is found, encouragingly, that J-test significance levels are not greatly influenced by the forms of weight matrices employed, for samples of size 400, while for samples of this size the significance level-inflation that occurs when there is low (or exactly zero) correlation between the regressor of a true null model and that in a false alternative model is largely absent. A disappointment, as a quick glance at the figures will show, is that there remain cases in which the likelihood-based bootstrapped tests are too liberal. However, the problem arises only for certain parameter combinations, and provided test users are aware of this phenomenon, empirical results can be interpreted with due caution.1

The next section describes the spatial models between which the J-type tests are designed to discriminate, defines the moment-based estimators, and the test statistics. In Section 3 the bootstrap is introduced, Section 4 summarizes the experimental evidence, and Section 5 concludes. To enhance readability, as many technical details as possible are relegated to the Appendix.

2. The Models, Estimators, and the J-type Tests

2.1. Null and Alternative Models

Following Kelejian (2008), the SARAR(1,1) model set-up is adopted. That is, under the null hypothesis, Model 0 is true:

$$Y = X_0 \beta_0 + \lambda_0 W_0 Y + U_0,$$

$$U_0 = \rho_0 M_0 U_0 + \varepsilon_0.$$  \hspace{1cm} (1)

Here, the matrix of exogenous variables, $X_0$, $n \times k_0$ and the dependent variable, $Y$, $n \times 1$ are each measured without error, the $n \times n$ weight matrices, $W_0$ and $M_0$ are fixed a priori, and the unobserved shock vector, $\varepsilon_0 \sim IID(0, \sigma_0^2 I_n)$ independent of the exogenous regressors, $X_0$. The parameters to be estimated are the slope coefficients, $\beta_0$, the spatial lag and error coefficients, $\lambda_0$ and $\rho_0$ and the variance, $\sigma_0^2$.

Under the alternative, the data are generated by a similar structure, Model 1:

$$Y = X_1 \beta_1 + \lambda_1 W_1 Y + U_1,$$

$$U_1 = \rho_1 M_1 U_1 + \varepsilon_1.$$  \hspace{1cm} (2)

Suppose, in addition, that in (1) both $W_0$ and $M_0$ arise from the row-standardization of symmetric matrices of non-negative elements and that $w_{ii} = m_{ii} = 0 \ (i = 1, \ldots, n)$; both matrices therefore have real eigenvalues. To economize on notation, let $C_0 = I - \lambda_0 W_0$, and $B_0 = I - \rho_0 M_0$. The matrices $C_0$ and $B_0$ are required to be non-singular, with inverses that may be expressed as power series expansions in $\lambda_0 W_0$ or $\rho_0 M_0$ as the case may be. Denoting the eigenvalues of these matrices as $\omega_i, \mu_i \ (i = 1, \ldots, n)$ necessary and sufficient
conditions are \( \hat{\lambda} \in (\omega_{\text{min}}^{-1}, \omega_{\text{max}}^{-1}) \) and \( \rho \in (\mu_{\text{min}}^{-1}, \mu_{\text{max}}^{-1}) \) with the row standardization assumed here implying that \( \omega_{\text{max}}^{-1} = \mu_{\text{max}}^{-1} = 1 \). Of course, though these assumptions are convenient numerically, they are not entirely innocuous, as discussed by Lee (2004, p. 1903) and Kelejian & Prucha (2010, pp. 55–56), and others. We assume that the alternative model parameters are such that the matrices \( B_1 = I - \rho_1 M_1 \) and \( C_1 = I - \lambda_1 W_1 \) also have inverses expressible as power series. These restrictions are important for the interpretation of the model; any parameter estimates that violate them should therefore be rejected. However, the only really satisfactory way to guarantee that parameter estimates satisfy the invertibility restrictions is to adopt an estimator that cannot deliver illegitimate parameter values. This drives the adoption of the QMLE in the experimental work described below in preference to the popular GMM estimators. As will be shown, the J-tests are much better behaved as a result in almost all the cases considered.

In introducing the J-type tests to this setting, Kelejian allows for some finite number, \( g \geq 1 \), of alternative models of the same type as (2) in which Model_0 is not nested. However, until some experience with the leading case of a single alternative, \( g = 1 \), is available, consideration of multiple alternatives might be thought ambitious. For this reason, the results described below relate only to tests for single alternatives.

### 2.2. Likelihood-based Estimation

If, in addition, in (1) the disturbance, \( \varepsilon_0 \) is Normally distributed, the log-likelihood may be written

\[
\ln L_0 = -\frac{n}{2} \ln(2\pi \sigma^2) - \frac{1}{2\sigma^2} [\mathbf{C}_0 \mathbf{Y} - \mathbf{X}_0 \beta]' \mathbf{B}_0 \mathbf{B}_0' [\mathbf{C}_0 \mathbf{Y} - \mathbf{X}_0 \beta] + \ln |\mathbf{C}_0| + \ln |\mathbf{B}_0|
\]

(3)

and we see that maximization of (3) will involve not only the sum of squared residuals but also the Jacobian term, \( \ln|\mathbf{C}_0| + \ln|\mathbf{B}_0| \). As a result, in numerical maximization of (3) the solutions for both \( \lambda \) and \( \rho \) will be driven away from the invertibility boundary by the behaviour of these determinants as the boundary is approached. It is worth noting that even if the shocks in the process that generated the data are not Normally distributed, it remains possible to estimate the model's parameters by maximizing (3), the procedure then being designated as quasi maximum likelihood estimation (QMLE). Another alternative to the Kelejian and Prucha GMM estimator is the 'optimal' moment-based estimator of Lee & Liu (2010). Since, at the sample maximum of (3) the score vector will be zero to within the numerical tolerance adopted in the optimization algorithm, it is of some interest to compare the QMLE moment conditions that will then be satisfied to those proposed by Lee & Liu (2010); in the Appendix it is shown that the latter are the same, except for one QMLE condition that is split into two separate conditions in the Lee and Liu approach. However, this slight change has an apparently significant side-effect, at least in samples of the size considered here. In Lee and Liu’s framework the extra moment condition introduces over-identification which is then handled by a two-step procedure that leads to an approximate solution to their moment conditions; however, neither the first round estimates of \( \rho \) or \( \lambda \) nor the
final GMM estimates need lie in the invertibility region. In their experiments, Lee and Liu reported a few cases in which the final estimates were outside this region even for a sample size of 490 and true values of 0.4 for both λ and ρ (Lee & Liu, 2010, Endnote 29). For smaller sample sizes or more extreme ρ or λ values one would expect to see many more such cases, which would be a problem in the present context. However, it is not necessary to introduce over-identification, and thus the potential problems introduced by an approximate multi-stage approach. The present paper avoids such difficulties simply by fully solving for the QMLE obtained by maximizing (3) by a numerical search algorithm.

2.3. Kelejian’s J-tests for the Case g = 1

Kelejian (2008) presents J-type tests based on estimates constructed using the generalized two-stage least squares estimator of Kelejian & Prucha (1998), combined with the generalized moments estimator of Kelejian & Prucha (1999). Suppose g = 1, so there is a single alternative to the null model. Write \( X_0 = [X_{01}, X_{02}] \) and \( X_1 = [X_{11}, X_{12}] \) in which \( X_{01} = X_{11} = [1, 1, \ldots, 1] \)' and \( X_{02} (n \times [k_0 - 1]) \), and \( X_{12} (n \times [k_1 - 1]) \) are the remaining non-constant exogenous regressors in the two models. Further, write \( Z_0 = [X_{01}, X_{02}; W_0; Y] \), \( Z_1 = [X_{11}, X_{12}; W_1; Y] \), and \( \gamma_1 = [\beta'_1, \lambda'_1] \) so the null model is

\[
Y = X_0 \beta_0 + \lambda_0 W_0 Y + U_0
\]

\[= Z_0 \gamma_0 + U_0.\tag{4}\]

The detailed implementation of the J-type tests using Kelejian and Prucha’s estimators is described in the Appendix, where the notation used in this section is defined more fully. The procedure involves estimating both models using the generalized spatial 2SLS method of Kelejian & Prucha (1998) incorporating estimates of \( \rho_0 \) and \( \rho_1 \) given by the non-linear GMM method of Kelejian & Prucha (1999). For Model0 the final step thus involves estimation of the transformed equation

\[
(I_n - \hat{\rho}_0 M_0)Y = (I_n - \hat{\rho}_0 M_0)(Z_0 \gamma_0 + U_0) \tag{5}
\]

\[
Y' (\hat{\rho}_0) = Z_0' (\hat{\rho}_0) \gamma_0 + \varepsilon' (\hat{\rho}_0) \quad \text{say}
\]

by the method of instrumental variables, which yields, say,

\[
Y' (\hat{\rho}_0) = Z_0' (\hat{\rho}_0) \hat{\gamma}_0 (\hat{\rho}_0) + \hat{\varepsilon}' (\hat{\rho}_0). \tag{6}
\]

Similarly, from Model1 is obtained

\[
Y' (\hat{\rho}_1) = Z_1' (\hat{\rho}_1) \hat{\gamma}_1 (\hat{\rho}_1) + \hat{\varepsilon}' (\hat{\rho}_1). \tag{7}
\]

Letting \( \hat{Y}' (\hat{\rho}_1) \) denote the fitted value from (7), the RHS of (5) can now be augmented to generate a test of the hypothesis that Model0 is true. Kelejian defines two tests:

**Test 1** A conjectured \( \chi^2 \) version, given \( g = 1 \).

Using the fitted value from (7), set up the augmented equation

\[
Y' (\hat{\rho}_0) = Z_0' (\hat{\rho}_0) \hat{\gamma}_0 + \hat{Y}' (\hat{\rho}_1) \delta + \varepsilon' (\hat{\rho}_0) \tag{8}
\]

\[= Z' \gamma^* + \varepsilon^* \]
and estimate it, using a suitably augmented matrix of instruments, to obtain a Wald test statistic for $d = 0$ in (8). That the associated test statistic has an asymptotic $\chi^2_1$ distribution under the null hypothesis remains a conjecture at the time of writing—see the first Remark under Equation 9 of Kelejian (2008). However, in the experiments reported by BF and those below, the $\chi^2_1$ distribution is found to be a reasonable approximation in most of the parameter space even in samples of modest size.

Test 2 A $\chi^2_2$ version, given $g = 1$.

Use the first step estimates (see the Appendix for details), $\hat{\gamma}_1$, to augment the RHS of (5) with both $Z_1\hat{\gamma}_1$ and $M_1Z_1\hat{\gamma}_1$, in place of the single forecast value, $\hat{Y}^*(\hat{\rho}_1)$, and with the augmented instrument set as before construct a Wald test of $d_1 = d_2 = d_0$ in the equation,

$$Y^*(\hat{\rho}_0) = Z_0^*(\hat{\rho}_0)\hat{\gamma}_0 + Z_1\hat{\gamma}_1\hat{\delta}_1 + M_1Z_1\hat{\gamma}_1\hat{\delta}_2 + \varepsilon^*(\hat{\rho}_0).$$

Kelejian proves that this second Wald statistic has an asymptotic $\chi^2_2$ distribution under appropriate conditions while the alternative one degree-of-freedom form, Test 1 above, is introduced in a remark that also raises the question of the relative efficiency of the two tests, commented on below.

2.4. Likelihood-based Versions of the J-tests

Notice, first of all, that the parameter estimates obtained above for each of the two competing models could be replaced by the QMLE. These alternative estimates could then be used to obtain the predictions from Model 1 with which the right-hand sides of (8) or (9) are augmented to generate the tests. Establishing the asymptotic distribution of the test statistics under such a modification is beyond the scope of this paper; the numerical evidence to be presented in Section 4 relates to the upper 5% tail of the sampling distributions and suggests strongly that this quantile of the finite sample distribution of the tests implemented using QMLE is in fact generally closer to the 5% quantile of the asymptotic distribution than is that of the original test. Adopting the QMLE, in place of $\hat{Y}^*(\hat{\rho}_1)$ obtained from an IV estimation of (7), we can construct the fitted value,

$$\hat{Y}^*(\hat{\rho}_1) = Z_1^*(\hat{\rho}_1)\hat{\gamma}_1$$

in which

$$Z_1^*(\hat{\rho}_1) = (I_n - \hat{\rho}_1M_1)Z_1 = \tilde{B}_1Z_1,$$

$$\hat{\gamma}_1 = [\tilde{\beta}'_1, \tilde{\lambda}_1]'$$

where $\tilde{\rho}_1, \tilde{\gamma}_1$ and $\tilde{\beta}'_1$ are the QML estimators obtained from Model 1. Similarly, define

$$Y^*(\hat{\rho}_0) = (I_n - \hat{\rho}_0M_0)Y,$$

$$Z_0^*(\hat{\rho}_0) = (I_n - \hat{\rho}_0M_0)Z_0.$$
2.4.1. The QMLE-based test statistics. The test corresponding to Test 1, above, is implemented with, in place of (8), the augmented equation

\[ Y^* (\hat{\rho}_0) = Z_0^* (\hat{\rho}_0) \gamma_0 + \tilde{Y}^* (\hat{\rho}_1) \delta + \varepsilon^*. \]  

(15)

The test statistic for the restriction, \( \delta = 0 \), is then constructed from instrumental variable estimation of \( \gamma_0 \) and \( \delta \) in (15) similarly to Test 1. To construct the test that corresponds to Test 2, both \( Z_1 \tilde{\gamma}_1 \) and \( M_1 Z_1 \tilde{\gamma}_1 \), are included in place of the single forecast value, \( \tilde{Y}^* (\hat{\rho}_1) \), in the augmented equation and

\[ Y^* (\hat{\rho}_0) = Z_0^* (\hat{\rho}_0) \gamma_0 + Z_1 \tilde{\gamma}_1 \delta_1 + M_1 Z_1 \tilde{\gamma}_1 \delta_2 + \varepsilon^* \]  

(16)

is estimated by instrumental variables similarly to Test 2, above, and the test statistic is constructed in the same way.

3. The Bootstrap Resampling Scheme

It was shown in BF that by resampling it was possible to gain control over empirical significance levels of the tests as originally formulated, in many cases. Here, it is anticipated that the same will be true of the reformulated tests, but the expectation is also that the severe inflation of significance levels for some parameter combinations that could not previously be corrected will now be either eliminated or much reduced as the procedures will now be based on invertible parameter estimates, as of course they should be. The scheme adopted is as follows:

Compute the \( J \) test statistics as above, then

(i) Using the whitened residuals from Model_0 as the building block, draw a random sample using sampling with replacement; call this random sample, \( e^* \).

(ii) Using \( \hat{\rho}_0 \) the QMLE of \( \rho \) from Model_0, generate

\[ u^* = [I - \hat{\rho}_0 M_0]^{-1} e^*. \]

(iii) Using the remaining QMLE parameter estimates,

\[ \tilde{\gamma}_0 = [\tilde{\beta}_0', \tilde{\delta}_0]' \]

generate

\[ Y^* = [I - \tilde{\lambda}_0 W_0]^{-1} (X_0 \tilde{\beta}_0 + u^*). \]

(iv) Calculate the \( J \) statistic using the \( Y^* \) sample.

(v) Repeat (ii)–(iv) the designated number of times, \( m \), to create a sample from the bootstrap distribution of the relevant \( J \) statistic.

(vi) If the proportion of the \( m \) bootstrap replicates that exceed the observed \( J \) statistic is less than the chosen significance level, reject the null hypothesis at that level.

4. Experimental Results

An interpretation of the J-type tests in which the researcher wishes to engage in model selection would have two implications: firstly, using these tests for such a purpose would entail the consideration of four possible test outcomes according to
whether Model\textsubscript{0} is rejected or not and Model\textsubscript{1} is rejected or not when the null hypothesis is reversed (which would greatly complicate the experimental reporting), and secondly a comparison with possibly more appropriate model-selection methods based on information criteria would surely be required. For these reasons the interpretation assumed throughout is that the researcher has chosen to privilege Model\textsubscript{0} so that a true hypothesis test is being conducted in which the alternative is ‘Model\textsubscript{0} is false and Model\textsubscript{1} is true’, while a failure to reject the null hypothesis will result in the retention of Model\textsubscript{0}.

The J-type statistics described in Sections 2 and 3 are not pivotal; that is to say their null distributions are not invariant to changes in the regressors, \(\mathbf{X}\), or the weight matrices, \(\mathbf{W}\), or the parameters of Model\textsubscript{0}. As a result, test performance will vary from case to case, and in particular, the true (empirical) significance level achieved will not in general be exactly equal to the nominal level (5\% in our experiments). Such variations are not a problem when empirical and nominal levels are close, but when the tests are either excessively liberal or very conservative, the interpretation of the results becomes difficult. It is therefore important to have some understanding of the reliability of nominal significance levels. Of course, tests that lack power can also mislead, and so it is also important to consider how powerful the tests are, once their significance levels are properly controlled. The experiments were set up to be informative about four related questions:

(1) Will estimating the spatial parameters by QMLE produce test statistics less prone to the severe inflation of significance levels identified for certain parts of the parameter space in BF? To answer this the experiments of BF were repeated and extended but with the tests based on the QMLE as set out in Section 2.4, and the results compared.

(2) Following on from (1), given that empirical significance levels obtained by estimating the spatial parameters, \(\lambda\) and \(\rho\), by QMLE are still variable, is the bootstrap more effective at controlling test performance on the extreme cases previously identified?

(3) Is the form of the weights matrices important? To provide some evidence on this point, in addition to the weights used in BF, weights for each of two sets of 25 regions are employed, the regions being arranged in either a 5 \times 5 square or a continuous ring.

(4) How quickly is the size inflation in certain extreme cases eliminated with increasing sample size? The cases for which the bootstrapped tests are most liberal are repeated for larger samples, with the weights for \(n = 400\) being either a block diagonal matrix formed of two copies of the weight matrices used for \(n = 200\) or corresponding to a square 20 \times 20 grid with weights inversely proportional to Euclidean distance. Although by no means general, these larger weight matrices give quite encouraging results, in that the size-inflation problem seems to be much reduced.

The experimental set-up is described in detail in the Appendix. It is the same as in BF except that use of the QMLE ensures that the estimates of both \(\lambda\) and \(\rho\) will lie in the invertible region. The results reported here are based on tests with nominal significance level 5\% and with \(m = 99\), in the bootstrap and \(s = 5,000\) replications, except where indicated. The standard error of the estimated significance levels, determined by \(s\), is therefore generally approximately 0.003 unless indicated otherwise. The comparatively small number of bootstrap samples per case (repeated
5,000 times, of course) turned out to give results virtually identical to those obtained in preliminary explorations using $m = 499$ or $999$. Thus $99$ was chosen to help limit computer time; in a one-off inferential problem one could use many thousands of bootstrap samples, of course, to maximize power, though the gains available are certainly quite trivial. The sampling variation in the results does not affect the main conclusions, and the number of replications is increased to 40,000 when greater accuracy is required. As in BF, the case in which the two models have different regressors but the same spatial weights is designated Case 1, and the case in which they have the same regressors but different spatial weights is Case 2. More complicated cases are not investigated here.

The two forms of test statistic, which here have either 1 or 2 degrees of freedom, when referred to critical values from the corresponding $X^2$ distribution, will be designated as the ‘asymptotic test’ and when referred to a bootstrap distribution, the ‘bootstrapped test’. In Case 1 there are $3(r \in 0, 1, 2) \times 5(\rho_0 \in -0.5, 0.0, 0.5, 0.90, 0.95) \times 5(\rho \in 0.0, 0.3, 0.9, 0.95) \times 5(\lambda \in 0.0, 0.3, 0.6, 0.9, 0.95) = 375$ sets of empirical size and power estimates of the asymptotic tests for each sample size considered in full ($n = 25, 26, 200$) while in Case 2 there are $2(r \in 1, 2) \times 5(\rho \in 0.0, 0.3, 0.6, 0.9, 0.95) \times 5(\lambda \in 0.0, 0.3, 0.6, 0.9, 0.95) = 50$ such sets for each sample size. Notice that the combination, $(\rho, \lambda) = (0, 0)$ appears in the experiments; in this case, neither model is ‘spatial’ and so in Case 2 the corresponding ‘power’ should equal the significance level. The empirical results reproduce this feature to within sampling error, which is a useful consistency check. In Case 2, the minimum feasible $r$ value is $r=0$ for the 1 degree-of-freedom tests, but $r=1$ for the 2 degree-of-freedom tests as explained below.

4.1. Case 1

4.1.1. Significance levels. It turns out that the richness of the instrument set used in the final step of test construction (for the QMLE-based tests) and throughout the several estimation steps (in the case of the GMM/2SLS-based tests) that is indexed by $r = 0, 1, 2$ as described in the Appendix, has a big effect on empirical significance levels. An example of such effects appears in Figures 1–3 which show the empirical sizes and powers of the asymptotics 1 $d.f.$ tests for the three sample sizes (designated CS1S25, CS1S26 and CS1S200 for example) in histograms each showing the 125 cases for a given combination of sample size and $r$. Indeed, it is clear from these summary statistics that the empirical significance levels of the tests are closest to the nominal 5% level when the minimal instrument set, $r=0$, is used. This pattern is reproduced throughout the results to varying degrees, and so to economize on space whilst delivering a clear message, most attention is given to the results for the smallest feasible instrument set. In Case 1 this is $r=0$. Turning again to Figure 1 notice that while the empirical significance levels are more tightly clustered around 0.05 for the larger sample size, the most liberal cases are further from the nominal level. This feature is present in both tests and all $r$ values. The rogue cases arise when the regressors in the true null model are (almost) orthogonal to those in the alternative, $\rho_0$ is high and $\lambda_0$ is low. As in the discussion of BF, the explanation appears to be that the asymptotic distribution of the tests is a poor approximation whenever the calculated standard error of the estimated coefficient ($\delta$) in the augmented equation is very high, and that this happens with these parameter combinations. If this phenomenon were widespread (across possible
parameter values) and persisted with increasing sample size, it would be disastrous for the usefulness of these tests. However, such evidence as is available suggests that it is a small sample problem that is largely absent for the samples of size 400 and parameter values that have been investigated here. Figure 4 shows the effect of basing inferences on the bootstrapped distribution rather than the asymptotic distribution of the 1 degree-of-freedom test calculated with $r/C^30_0$. Although not eliminated, the inflation of significance levels is much reduced, while any associated power losses are minuscule. Evidently, the simple bootstrap is useful therefore. To see the scale of the improvement that can be achieved by the use of the QMLE rather than the GMM/TLS estimator, when $r = 0$ compare Figures 5 and 6. For sample size, $n = 26$, the significance levels of the asymptotic QMLE-based tests are marginally worse than for the GMM/2SLS-based tests, but the position is reversed for the simple bootstrap versions. For $n = 200$, however, the improvement achieved is dramatic: the asymptotic QMLE-based test is comparable to the

**Figure 1.** Case 1 Asymptotic degree-of-freedom tests with minimal instruments, $r = 0$. Empirical significance levels and powers, sample sizes 25, 26 and 200.
bootstrapped GMM/2SLS-based version, and then is itself further improved when bootstrapped.

Comparing test performance when a rich instrument set is used is somewhat flattering to the QMLE-based approach, in part because it is clear that the biggest improvement in test performance in samples of the size considered here comes about by using a minimal instrument set. Nevertheless, the fact that some parameter combinations lead to excessively liberal tests, even after use of the bootstrap and QML estimation, requires some further investigation. It is important to note that in order to implement the GMM/2SLS-based tests, as in BF, whenever the estimated $r$ or $l$ values were outside the invertible region (in practice, exceeding 1.0), they were replaced by 0.97; this was quite a common occurrence. The QMLE, of course, did not need to be constrained in this way, and so could in fact be closer to the boundary in some samples. The worst parameter combination trialled from the point of view of significance level inflation is: $(r_x, r, \lambda) = (0.0, 0.95, 0.0)$; Table 1 summarizes the results obtained, revealing that the QMLE approach produces a dramatic improvement for $n=200$ when $r=1$, 2 but not otherwise. In three

Figure 2. Case 1 asymptotic 1 degree-of-freedom tests with intermediate instruments, $r=1$. Empirical significance levels and powers, with sample sizes 25, 26 and 200.
experiments with \( n = 400 \), (a) as in the table but with weights corresponding to two copies of the \( n = 200 \) matrix, (b) with the regressor in Model 1 replaced by the residual from its projection on the regressor of model zero to give exact orthogonality in the sample, and (c) as in the table but with weights for a \( 20 \times 20 \) grid proportional to the reciprocal of inter-centroid distances, each of the QMLE-based tests has essentially the correct significance level.

The main conclusion to this point seems to be that in Case 1, to achieve control over the empirical significance level, the 1 d.f. tests should be implemented with \( r = 0 \) and bootstrap critical values used if the sample is small. However, over most of the parameter space, provided \( r = 0 \), and especially if \( n \) is large, bootstrapping is not really necessary in this case.

4.1.2. Power. So far, the 2 d.f. test has not been discussed. Comparing Figures 4 and 7 it is clear that, after bootstrapping to obtain the best available control of significance levels, the 1 d.f. test has an advantage, being no more liberal than the 2 d.f. test, but having marginally greater power in each of the models tested. Further, for \( n = 200 \), the powers are virtually 100\% for all the parameter settings tested. For

---

**Figure 3.** Case 1 asymptotic 1 degree-of-freedom tests with rich instruments, \( r = 2 \). Empirical significance levels and powers, with sample sizes 25, 26 and 200.
the smaller sample sizes, the powers fall into three clusters, clearly visible in the upper right panels of Figure 4. These correspond to $\rho_x = 0.95$ (lowest power), 0.90 (moderate power) and ± 0.5 or 0.0 (high power). For relatively low levels of correlation between the regressors in the competing models (measured by $\rho_x$), size-corrected power is increasing in $\lambda$ and decreasing (though less significantly) in $\rho$ while at high levels of $\rho_x$ power is increasing in both $\rho$ and $\lambda$. No explanation is yet available for this phenomenon. For the $n = 400$ samples, all tests have empirical powers of 100%.

4.2. Case 2

In these experiments, the instruments, $H_{1,1-r}$, are used in (A17) in place of $H_{01,1-r}$ for the estimation of the augmented equation, for the reason given in BF (better power). However, the effect of instrument choice is not as simple as in Case 1. First of all, the minimal set, obtained by setting $r = 0$, is not rich enough to permit calculation of the two-degree-of-freedom test in Case 2, as shown in the Appendix.
Figure 5. Empirical significance levels of the 1 d.f. tests using GMM/2SLS estimators, with rich instrument sets ($r = 2$).
Figure 6. Case 1. Empirical significance levels of the 1 d.f. tests using QML estimators with rich instrument sets (r=2). Samples sizes 26 and 200.
For this reason, we concentrate on \( r = 1, 2 \) for this case. The main experiments relate to the weight matrices described above for \( n = 26, 200, 400 \) with a subsidiary check made using very different weight matrices with \( n = 25 \), as described below.

### 4.2.1. Significance levels

Figures 8–13 present the distributions of empirical significance level and power estimates that result from 5,000 replications of each test (either asymptotic or bootstrapped) on each parameter combination for each pair of weight matrices with either rich (\( r = 2 \)) or medium (\( r = 1 \)) sets of instruments. Looking first at the left-hand panels of Figure 8 which relates to the 1 d.f. test applied to two versions of the Cliff & Ord (1973) 26-county weights, observe that the richer instruments give slightly less dispersed empirical significance levels, and that bootstrapping improves this aspect further. If the tests were in fact pivotal, then the significance levels estimated using 5,000 replications would be approximately independent random samples from \( N(\mu, \sigma^2) \) with \( \mu = 0.05 \) and \( \sigma = (0.05 \times 0.95/5000)^{1/2} = 0.0031 \). Testing this for the lower right panel of the table, we find that the mean significance level observed, 0.0478, is significantly below 0.05; thus the bootstrap is unable to fully correct the empirical significance level. However, the results also show that the empirical significance levels obtained for this pair of weight matrices are very close to the nominal 5%. Comparing Figures 8 and 9 reveals that the empirical significance levels of the 1 degree-of-freedom and 2 degree of freedom tests are very similar for these weight matrices. The story is quite different for the two sets of EU interregional weights (\( n = 200 \)) Figures 10 and 11 again show that the bootstrap reduces the dispersion of the empirical significance levels, but those of the 2 degree-of-freedom test particularly are now heavily skewed to the right, even after application of the bootstrap.

### Table 1. Empirical significance levels, for \( (\rho_x, \rho, \lambda) = (0.0, 0.95, 0.0) \), 1 degree-of-freedom test, Case 1 using \( \chi^2 \) critical value, \( s = 40,000 \) for col 1, \( s = 5,000 \) for col 3; using bootstrap critical values, \( s = 5,000, m = 99 \) in cols 2 and 4

<table>
<thead>
<tr>
<th></th>
<th>GTLS</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \chi^2 )</td>
<td>BS</td>
<td>( \chi^2 )</td>
</tr>
<tr>
<td>0.64</td>
<td>0.45</td>
<td>0.24</td>
</tr>
<tr>
<td>0.51</td>
<td>0.37</td>
<td>0.22</td>
</tr>
<tr>
<td>0.10</td>
<td>0.06</td>
<td>0.14</td>
</tr>
<tr>
<td>n.a.</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>n.a.</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>n.a.</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>n.a.</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>n.a.</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>n.a.</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

For \( n = 400 \)

BS: simple bootstrap size
these weights it appears that to achieve the best control over the empirical significance level the investigator should use the 1 degree-of-freedom test in combination with the bootstrap; the richness of the instrument set is less important. The most compelling evidence that the bootstrap is working correctly comes from Figures 12 and 13; here, with sample size 400, the estimated empirical significance levels of the bootstrapped tests are indeed approximately $N(0.05, 0.0032)$ as they should be. The inflated empirical significance levels seen in Figures 10 and 11 are now absent, and the main contribution of the bootstrap appears to be to remove a slight negative skew from the significance levels observed for the asymptotic tests. Neither the richness of the instrument set nor the choice of degrees of freedom appears important in these cases. A final observation is that the empirical significance levels estimated for Case 2 are in general much less subject to distortions than those found in Case 1, while the weight matrices that give rise to the greatest distortions are the same, namely those for $n = 200$.

4.2.2. Power. The right-hand panels of Figures 8 and 9 show that for the $n = 26$ weights, unadjusted power is greatest using the 1 degree-of-freedom test with rich

Figure 7. Case 1 bootstrapped 2 degree-of-freedom tests with minimal instruments, $r = 0$. Empirical significance levels and powers, sample sizes 25, 26 and 200.
levels, with those for the bootstrapped versions being more concentrated around

0

medium set. For $n/C_1$

then using the richer instrument set gives a 3 degree-of-freedom test must be used, followed by the bootstrap. If this is done, weights Figures 10 and 11 show that to achieve control of significance level, the 1 degree-of-freedom tests, sample size $n=26$.

Since there are no extreme empirical significance levels present, comparisons of empirical powers are informative. The bootstrap-corrected tests do show a slight power reduction, of the order of 2%, which is not unexpected. For the rich instruments (r=2)

sample size $n=26$, all the test forms have approximately correct significance levels.

empirical significance levels

empirical significance levels

Empirical power

4% power advantage over the

empirical powers

empirical powers

empirical powers

Empirical power

Empirical power

Figure 8. Case 2 Empirical significance levels and powers 1 degree-of-freedom tests, sample size $n=26$.

instrument set, though all powers are low, the maximum observed being only 35%. Since there are no extreme empirical significance levels present, comparisons of empirical powers are informative. The bootstrap-corrected tests do show a slight power reduction, of the order of 2%, which is not unexpected. For the $n=200$ weights Figures 10 and 11 show that to achieve control of significance level, the 1 degree-of-freedom test must be used, followed by the bootstrap. If this is done, then using the richer instrument set gives a 3– 4% power advantage over the medium set. For $n = 400$, all the test forms have approximately correct significance levels, with those for the bootstrapped versions being more concentrated around
0.05, and as seen in Figure 13 the best power is obtained using the 2 degree-of-freedom test with a rich instrument set.

4.2.3. Experiments with other weight matrices. Role reversal A natural question in Case 2 is whether reversing the roles of null and alternative models affects test characteristics. An auxiliary experiment, using 40,000 replications to reduce the standard error of estimated significance levels to approximately 0.001, was conducted with two weight matrices with \( n = 25 \). One (queen’s case) was a 5 × 5 grid with row-normalised queen’s case contiguity weights, the other (ring) a continuous ring with equal weights on each of two neighbours. These two
matrices were deliberately chosen to be very different. The results clearly show that for the asymptotic tests both empirical significance levels and power differ according to which matrix is taken as the null hypothesis. Comparing the right and left panels of Figure 14 shows that the mean significance levels are closer to 0.05 when the continuous ring is the null, the difference being about 2% for the 2 degree-of-freedom test with \( r = 2 \). Fortunately, the bootstrap works well here, producing tests that are on average very slightly conservative as seen in each of

![Figure 10](https://example.com/figure10.png)

**Figure 10.** Case 2. Empirical significance levels and powers of 1 degree-of-freedom, sample size \( n = 200 \).
When the queen is the null, given that the number of trials is broadly, power is greater when the queen's case is the alternative and the ring is the null, than vice versa. Figure 16 shows the power differences plotted against $\rho$ and reveals in each case a rising trend that is consistent across $r$ values and degrees-of-freedom. Many of these differences are highly significant; for example, when $(\rho, \lambda) = (0.3, 0.9)$ and $r = 2$ the empirical significance levels and powers of the 2 degree-of-freedom tests are (0.042, 0.569) when the ring is the null, and (0.042, 0.437) when the queen is the null. Given that the number of trials is

Figure 11. Case 2. Empirical significance levels and powers of 2 degree-of-freedom tests, with sample size n = 200.
Figure 12. Case 2 Empirical significance levels and powers of 1 degree-of-freedom tests with sample size, \( n = 400 \).

40,000, a z-score for the difference in power may be computed as 
\[ z = (0.569 - 0.437) / (2 \times 0.503 \times 0.497 / 40000)^{1/2} = 37.3. \] Clearly, test power is generally greater when the null is the ring than when the null is the queen’s case rectangular grid. A more detailed inspection of the results reveals that the powers are similar when \( \rho = 0 \) or 0.3, becoming increasingly different as \( \rho \) increases, as shown in Figure 14. The results also reveal that it is the spatial lag parameter, \( \lambda \), that is most critical in determining the power of the tests while their relative powers are more influenced by the error parameter, \( \rho \), to such an extent that with \( r = 0 \) the 1 d.f. test with ring as null is three times as powerful as that with

\begin{align*}
\text{Series: CS1P400} & \quad \text{Sample: 51 75} \\
\text{Observations: 25} & \\
\text{Mean: 0.543840} & \\
\text{Median: 0.490000} & \\
\text{Maximum: 0.999000} & \\
\text{Minimum: 0.049000} & \\
\text{Std. Dev.: 0.323087} & \\
\text{Jarque-Bera: 2.152362} & \\
\text{Kurtosis: 1.738368} & \\
\text{Skewness: 0.328789} & \\
\text{Probability: 0.551844} & \\
\text{Series: CS1S400} & \quad \text{Sample: 51 75} \\
\text{Observations: 25} & \\
\text{Mean: 0.543840} & \\
\text{Median: 0.490000} & \\
\text{Maximum: 0.999000} & \\
\text{Minimum: 0.049000} & \\
\text{Std. Dev.: 0.323087} & \\
\text{Jarque-Bera: 2.152362} & \\
\text{Kurtosis: 1.738368} & \\
\text{Skewness: 0.328789} & \\
\text{Probability: 0.551844} & \\
\text{Series: BS1P400} & \quad \text{Sample: 51 75} \\
\text{Observations: 25} & \\
\text{Mean: 0.543840} & \\
\text{Median: 0.490000} & \\
\text{Maximum: 0.999000} & \\
\text{Minimum: 0.049000} & \\
\text{Std. Dev.: 0.323087} & \\
\text{Jarque-Bera: 2.152362} & \\
\text{Kurtosis: 1.738368} & \\
\text{Skewness: 0.328789} & \\
\text{Probability: 0.551844} & \\
\text{Series: BS1S400} & \quad \text{Sample: 51 75} \\
\text{Observations: 25} & \\
\text{Mean: 0.543840} & \\
\text{Median: 0.490000} & \\
\text{Maximum: 0.999000} & \\
\text{Minimum: 0.049000} & \\
\text{Std. Dev.: 0.323087} & \\
\text{Jarque-Bera: 2.152362} & \\
\text{Kurtosis: 1.738368} & \\
\text{Skewness: 0.328789} & \\
\text{Probability: 0.551844} & \\
\end{align*}
Case 2. Empirical significance levels and powers of 2 degree-of-freedom tests with sample size, $n=400$.

Figure 13. Case 2. Empirical significance levels and powers of 2 degree-of-freedom tests with sample size, $n=400$.

queen’s case as null when $\rho = 0.95$ and $\lambda > 0$. Although intriguing, these effects are of minor practical significance since it is most unlikely that the weight matrices under test would ever be as different as these, and it seems clear also that for Case 2, the richer instrument set is to be preferred.

For completeness, Table 2 records descriptive statistics for significance levels and powers for the bootstrapped 1 degree-of-freedom tests for all three $r$ values. The benefits of taking $r = 2$ here are apparent whichever weight matrix is the null. Notice that the median powers and maximum powers tell different stories. A final comment: the significant power differences may be a curiosum, attributable to the
small sample size and extreme difference in weight patterns; when the experiment was repeated using the two versions of the EU-NUTS weights with $n/C30$, no significant differences emerged.

5. Conclusions

This kind of numerical exercise cannot replace an exact algebraic account of the dependence of test performance on parameter values, regressor characteristics, weight matrices estimation method and sample size, but faced with the difficulty of providing such an analysis, a numerical approach is the next best thing. It must not be forgotten, however, that numerical evidence cannot prove the absence of
problems, but it can reveal them. The experiments relate to the case of a single alternative model, so $g = 1$, and a single non-constant explanatory variable, and either different weight matrices or different regressors, but not both. The weights for the main experiments were taken from real examples that have been used in empirical research.

In broad terms, it seems that the cases differ in ways that users of the tests should take into account when deciding how to implement them:

(i) In Case 1 the asymptotic tests based on the QMLE have virtually correct significance levels provided the dimension of the instrument set is as small as possible, and are superior to the tests using the generalized methods of moment estimators; the bootstrap appears to deliver correct significance...
levels when the minimal instrument set is used in virtually all cases, and in all cases tested with $N = 400$; there is little to choose between the one and two-degree-of-freedom statistics though the 1 d.f. test has slightly more power.

(ii) In Case 2 the fact that the 1 d.f. test can be calculated with a smaller instrument set gives it a slight practical advantage, though use of the 2 d.f. test and a rich instrument set gives the best results when $n = 200$. All results are again improved by use of the bootstrap to provide a reference distribution.
1. After this work was completed, the author became aware that a refinement of the J-test has been published by Kelejian and Piras (2011), whose results suggest that their modified tests have correct significance levels in all the cases they consider, and often better power than the tests studied here.

References


Appendix A.

A.1. QMLE versus Lee & Liu (2010)

The score vector corresponding to (3) is:

\[
\frac{\partial \ln L_0}{\partial \beta} = \frac{1}{\sigma^2} X_0' B_0' B_0 U_0,
\]

\[
\frac{\partial \ln L_0}{\partial \rho} = -\frac{1}{2\sigma^2} U_0'[2\rho M_0' M_0 - M_0 - M_0'] U_0 - Tr[M_0 B_0^{-1}],
\]

\[
\frac{\partial \ln L_0}{\partial \lambda} = \frac{1}{\sigma^2} Y' W_0' B_0' B_0 U_0 - Tr[W_0 C_0^{-1}],
\]

\[
\frac{\partial \ln L_0}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} U_0' B_0' B_0 U_0,
\]

where \( U_0 = [C_0 Y - X_0\beta] \). The score, (A1), can of course be written in various forms; for present purposes it is useful to write \( B_0 U_0 = \epsilon_0 \) or \( U_0 = B_0^{-1} \epsilon_0 \) as appropriate and to note that \( 2\rho M_0' M_0 - M_0 - M_0' = -[M_0 B_0 + B_0' M_0] \) so that

\[
\frac{\partial \ln L_0}{\partial \beta} = \frac{1}{\sigma^2} X_0' B_0' \epsilon_0, \tag{A2}
\]

\[
\frac{\partial \ln L_0}{\partial \rho} = \frac{1}{2\sigma^2} U_0'[M_0' B_0 + B_0' M_0] U_0 - Tr[M_0 B_0^{-1}]
\]

\[
= \frac{1}{2\sigma^2} \epsilon_0' B_0^{-1} [M_0' B_0 + B_0' M_0] B_0^{-1} \epsilon_0 - Tr[M_0 B_0^{-1}]
\]

\[
= \frac{1}{\sigma^2} \epsilon_0' B_0^{-1} \epsilon_0 - Tr[M_0 B_0^{-1}], \tag{A3}
\]

\[
\frac{\partial \ln L_0}{\partial \lambda} = \frac{1}{\sigma^2} Y' W_0' B_0' \epsilon_0 - Tr[W_0 C_0^{-1}]
\]

\[
= \frac{1}{\sigma^2} \epsilon_0' B_0 W_0 C_0^{-1} [X_0 \beta + B_0^{-1} \epsilon_0] - Tr[W_0 C_0^{-1}], \tag{A4}
\]

\[
\frac{\partial \ln L_0}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \epsilon_0' \epsilon_0. \tag{A5}
\]

Setting (A2), to (A5) equal to zero and solving the resulting system of equations of course delivers the maximum likelihood estimator of \( \theta = (\beta', \rho, \lambda, \sigma^2)' \). On the other hand, taking expectations in (A2), to (A5) yields the zero vector, so that the \( \theta \) that sets the score to zero can also be thought of as a method of moments estimator. In Lee & Liu (2010), on putting the optimal instruments described in their Proposition 5 into their Equation 4, a set of moment conditions for this model is obtained, written in their notation as \( E\{g(\bar{\theta})\} = E\{Q_1, Q_2, P_1 \epsilon, P_2 \bar{\epsilon} | \epsilon\} = 0 \) where, in the present notation,

\[
Q_1 = B_0 X_0, \tag{A6}
\]

\[
Q_2 = B_0 W_0 C_0^{-1} X_0 \beta, \tag{A7}
\]

\[
P_1 = B_0 W_0 C_0^{-1} B_0^{-1} - n^{-1} Tr[W_0 C_0^{-1}] \cdot I_n, \tag{A8}
\]
\[ \mathbf{P}_2 = \mathbf{M}_0 \mathbf{B}_0^{-1} - n^{-1} \mathbf{Tr}[\mathbf{M}_0 \mathbf{B}_0^{-1}] \mathbf{I}_n. \]  

(A9)

It is easily seen that with the addition of (A5) these conditions correspond to those obtained from setting the score vector equal to zero, but with (A4) split into two conditions: (A7) involving \( \mathbf{Q}_2 \) being linear in \( \varepsilon \), (A8) involving \( \mathbf{P}_1 \) being quadratic. Thus in Lee and Liu’s framework over-identification is introduced: \( \mathbf{Q}_2 \varepsilon \) has dimension \((k_0 \times 1)\) \( \mathbf{Q}_1 \varepsilon \) has dimension \((1 \times 1)\) \( \varepsilon \mathbf{P}_2 \varepsilon \) has dimension \((1 \times 1)\) and \( \varepsilon \mathbf{P}_1 \varepsilon \) has dimension \((1 \times 1)\), thus there are \( k_0 + 3 \) moment conditions for the \( k_0 + 2 \) parameters, \( \beta, \lambda, \rho \) to which must be added (A5) to determine \( \sigma^2 \). Lee & Liu (2010, Proposition 5) propose to implement such moment conditions by replacing \( \mathbf{P}_1 \) and \( \mathbf{Q}_1 \) by ‘feasible’ versions using first round \( n^{1/2} \)-consistent estimators, and then minimizing \( \hat{g}'(\theta) \hat{\Omega}^{-1} \hat{g}(\theta) \), say, in which \( \hat{\Omega} \) is a consistent stimator of the covariance matrix of \( g(\theta) \).

A.2. Kelejian’s Test Details

This material is included for completeness, in particular to explain the notation used in Section 2.3 of the paper. The non-linear GMM versions of the tests, as originally introduced by Kelejian (2008), may be defined via a series of steps.

**Step 1.** Estimate \( \gamma_0 \) in (4) and similarly \( \gamma_1 \) by instrumental variables (IV). Define the matrix,

\[ \mathbf{L}_{0,r} = [\mathbf{X}_{01} \ldots \mathbf{X}_{02} \mathbf{W}_{00} \mathbf{X}_{02} \ldots \mathbf{W}_{01} \mathbf{X}_{02}], \]

for some small integer \( r \), and construct a matrix of instruments,

\[ \mathbf{H}_{0,r} = [\mathbf{L}_{0,r} \mathbf{M}_0 \mathbf{L}_{0,r}]_{LI}, \]

(A10)

in which the subscript \( LI \) denotes a spanning set of linearly independent columns. Define the associated projection matrix, \( \mathbf{P}_{0,r} = \mathbf{H}_{0,r} (\mathbf{H}_{0,r} \mathbf{H}_{0,r})^{-1} \mathbf{H}_{0,r}' \), leading to the IV estimator,

\[ \hat{\gamma}_{0r} = [\mathbf{Z}' \mathbf{P}_{0,r} \mathbf{Z}_0]^{-1} \mathbf{Z}' \mathbf{P}_{0,r} \mathbf{Y}. \]

Similarly, for Model1 defining \( \mathbf{L}_{1,r} = [\mathbf{X}_{11} \ldots \mathbf{X}_{12} \mathbf{W}_1 \mathbf{X}_{12} \ldots \mathbf{W}_1 \mathbf{X}_{12}], \) and

\[ \mathbf{H}_{1,r} = [\mathbf{L}_{1,r} \mathbf{M}_1 \mathbf{L}_{1,r}]_{LI}, \]

(A11)

gives the projector, \( \mathbf{P}_{1,r} \), and IV estimator,

\[ \hat{\gamma}_{1r} = [\mathbf{Z}_1 \mathbf{P}_{1,r} \mathbf{Z}_1]^{-1} \mathbf{Z}_1 \mathbf{P}_{1,r} \mathbf{Y}. \]

The instruments defined above differ from those given by Kelejian, so, for future reference, define the hybrid matrix, \( \mathbf{L}_{01,r} = [[\mathbf{X}_{01} \mathbf{X}_0] \mathbf{W}_0 [\mathbf{X}_0 \mathbf{X}_1] \ldots \mathbf{W}_0 [\mathbf{X}_0 \mathbf{X}_1]]_{LI} \) and introduce \( \mathbf{H}_{01,r} = [\mathbf{L}_{01,r} \mathbf{M}_0 \mathbf{L}_{01,r}]_{LI} \), which corresponds to Kelejian’s \( \mathbf{A}_n \).

**Step 2.** Estimate \( \rho_0 \) and \( \rho_1 \) by the non-linear GMM method of Kelejian & Prucha (1999). Define the vector of residuals from IV estimation of the null model as, \( \hat{\mathbf{U}}_0 = \mathbf{Y} - \mathbf{Z}_0 \hat{\gamma}_0 \). The non-linear GMM estimator of \( \rho_0 \) is obtained as follows. Let \( \mathbf{D} \) be a fixed \( n \times n \) matrix, then \( \hat{\omega}_0 \) in (1) satisfies the second moment condition:
Obtained when sample averages and minimizing the sum of squares of the three sample moments which are easily converted into statements about the moments of lag-transformed regression estimate (1998) that yields, say, result is the generalized spatial 2SLS procedure suggested in Kelejian & Prucha.

Defining the vector of residuals from the alternative, \( \hat{U}_1 = Y - Z_1 \hat{\gamma}_1 \) we estimate \( \rho \) in a similar fashion to get \( \hat{\rho}_1 \), say.

**Step 3.** Estimate spatially lag-transformed (in feasible GLS form) regressions by IV. Using \( \hat{\rho}_0 \) from Step 2, construct the transformed regression

\[
(I_n - \hat{\rho}_0 M_0)Y = (I_n - \hat{\rho}_0 M_0)(Z_0 \hat{\gamma}_0 + U_0),
\]

and estimate this equation by IV using the same instruments as before, \( H_{0,i} \); the result is the generalized spatial 2SLS procedure suggested in Kelejian & Prucha (1998) that yields, say,

\[
Y^*(\hat{\rho}_0) = Z_0^*(\hat{\rho}_0)\hat{\gamma}_0 + \varepsilon^*(\hat{\rho}_0).
\]

Use the residual vector, \( \varepsilon^*(\hat{\rho}_0) \), to estimate the variance of the shocks, \( \hat{\sigma}_0^2 = \varepsilon^*(\hat{\rho}_0)^T\varepsilon^*(\hat{\rho}_0)/n \).

Similarly, using \( \hat{\rho}_1 \) from Step 2, construct the alternative spatially lag-transformed regression

\[
(I_n - \hat{\rho}_1 M_1)Y = (I_n - \hat{\rho}_1 M_1)(Z_1 \hat{\gamma}_1 + U_1),
\]

and estimate it by IV using the instruments, \( H_{1,i} \) to obtain

\[
Y^*(\hat{\rho}_1) = Z_1^*(\hat{\rho}_1)\hat{\gamma}_1 + \varepsilon^*(\hat{\rho}_1).
\]

Let \( \hat{Y}^*(\hat{\rho}_1) \) denote the fitted value from (A15). The RHS of (A12) is now augmented to generate a test of the hypothesis that Model_0 is true. There are two forms of the final step.

**Step 4a** (conjectured \( \chi^2 \) version, in case \( g = 1 \) as assumed here).

Using the fitted value from (A15), set up the augmented equation

\[
Y^*(\hat{\rho}_0) = Z_0^*(\hat{\rho}_0)\hat{\gamma}_0 + \hat{Y}^*(\hat{\rho}_1)\delta + \varepsilon^*(\hat{\rho}_0)
\]

\[
= Z_{0}^{**}\gamma^{**} + \varepsilon^{**}
\]
and the augmented matrix of instruments

\[ H_{r}^{**} = [H_{0,r} \hat{H}_{01,r}]_{LI} \]  

(A17)

with projection matrix \( P_{r}^{**} \), say, obtaining the IV estimator

\[ \hat{\gamma}^{**} = (Z^{**}P_{r}^{**}Z^{**})^{-1}Z^{**}P_{r}^{**}Y^{*}(\hat{\rho}_{0}) \]

with estimated asymptotic covariance matrix, \( \hat{V} = \hat{\sigma}_{0}^{2}(Z^{**}P_{r}^{**}Z^{**})^{-1} \), which is used to extract a Wald test statistic for \( H \) if \( H_{0} \) is true, letting \( l \) denote the number of elements in \( \gamma^{**} \), so that \( \hat{\gamma}^{**}(l) \) is the least estimated coefficient, and \( \hat{V}(l, l) \) its estimated variance, it is conjectured that,

\[ \frac{\hat{\gamma}^{**}(l)^{2}}{\hat{V}(l, l)} \rightarrow^{d} \chi_{(l)}^{2}. \]  

(A18)

The specification of \( H_{01,r} \) in the instrument set, (A17), is as given by Kelejian (2008); however, in the experiments of BF it was found that test power improved dramatically in some cases if \( H_{1,r} \) was used in place of \( H_{01,r} \) at this point.

**Step 4b** (\( \chi_{2}^{2} \) version, in case \( g = 1 \) as assumed here).

Augment the RHS of (A12) with both \( Z_{1}\hat{\gamma}_{1} \) and \( M_{1}Z_{1}\hat{\gamma}_{1} \), in place of the single forecast value, \( \hat{Y}^{*}(\hat{\rho}_{1}) \), and augment the instrument vector as before. Following the same line of development leads to a statistic that is asymptotically \( \chi_{(2)}^{2} \). That is, now estimate the equation

\[ Y^{*}(\hat{\rho}_{0}) = Z_{0}^{*}(\hat{\rho}_{0})\gamma_{0} + Z_{1}\hat{\gamma}_{1}\delta_{1} + M_{1}Z_{1}\hat{\gamma}_{1}\delta_{2} + \varepsilon^{l}(\hat{\rho}_{0}) \]

(A19)

using the instruments, \( H_{r}^{**} \), as above, obtaining the IV estimator,

\[ \hat{\gamma}^{l} = (Z^{l}P_{r}^{**}Z^{l})^{-1}Z^{l}P_{r}^{**}Y^{*}(\hat{\rho}_{0}), \]

with estimated asymptotic covariance matrix, \( \hat{V}^{l} = \hat{\sigma}_{0}^{2}(Z^{l}P_{r}^{**}Z^{l})^{-1} \). Writing the matrix that selects the final two elements of \( \gamma^{l} \) in the usual way as \( R = [0 \ldots 0 \ 1_{2}] \), the hypothesis to be tested is \( H_{0} : R\gamma^{l} = 0 \) and a Wald test statistic is

\[ \hat{\gamma}^{l}R^{l}[\hat{R}V^{l}R^{l}]^{-1}R^{l}\gamma^{l} \rightarrow^{d} \chi_{(2)}^{2}. \]  

(A20)

Kelejian proves (A20) under appropriate conditions while the alternative one degree-of-freedom form, (A18), is introduced in a remark that also raises the question of the relative efficiency of the two tests.

**A.3. The Instrument Set and Test Degrees of Freedom in Case 2**

When \( X_{02} \) and \( X_{12} \) coincide, then \( L_{0,r} = [X_{01};X_{02};W_{0}X_{02};\ldots;W_{0}X_{02}] \), \( H_{0,r} = [L_{0,r};M_{0}L_{0,r}]_{LI} \), while \( L_{1,r} = [X_{01};X_{02};W_{1}X_{02};\ldots;W_{1}X_{02}] \), and \( H_{1,r} = [L_{1,r};M_{1}L_{1,r}]_{LI} \); further, setting \( X_{01} = 1 \), \( M_{0} = W_{0} \) and \( M_{1} = W_{1} \), noting that the row normalization gives \( W_{1}1 = 1, j = 0,1 \) the above instrument matrices reduce to \( H_{0,0} = [1;X_{02};W_{0}X_{02}] \) and \( H_{1,0} = [1;X_{02};W_{1}X_{02}] \) so that
\[ H_0^{**} = [H_{01}^0, H_{01}^1]_{LL} = [1, \mathbf{X}_{02}, \mathbf{W}_0 \mathbf{X}_{02}, \mathbf{W}_1 \mathbf{X}_{02}]. \] Now the dimension of \( \gamma^{**} \) is 4, but that of \( \gamma^1 \) is 5, so that the 2 degree-of-freedom test cannot be calculated with \( r = 0 \). However, the 1 degree-of-freedom test can be calculated with this minimal instrument set.

**A.4. The Experimental Set Up**

**Case 1.** This is implemented with a single explanatory variable other than the constant, that is, the regressors are \( \mathbf{X}_0 = [\mathbf{X}_{01}^0, \mathbf{X}_{02}] \) where \( \mathbf{X}_{01} = \mathbf{1} \), the vector, \([1,1, \ldots, 1]^\prime\) and \( \mathbf{X}_{02} \) is a draw from \( N(0,I_n) \) and the two spatial weight matrices are equal, \( \mathbf{M}_0 = \mathbf{W}_0 \), while the alternative has the same spatial structure, \( \mathbf{M}_1 = \mathbf{W}_1 = \mathbf{W}_0 \), but the explanatory variable, \( \mathbf{X}_{02} \) is replaced by another that is in general correlated with it, constructed in the experiments as, \( \mathbf{X}_{12} = \rho_x \mathbf{X}_{02} + (1 - \rho_x^2)^{1/2} \times N(0, \mathbf{I}) \) for various values, including zero. A special case is also considered, in which \( \mathbf{X}_{12} \) is drawn as above with \( \rho_x = 0 \), replaced by the residual from its projection on \( \mathbf{X}_{02} \) giving a regressor that is exactly orthogonal to \( \mathbf{X}_{02} \) in the sample.

**Case 2.** This is implemented by having the explanatory variables, \( \mathbf{X}_0 \) and \( \mathbf{X}_1 \), the same in the two models (\( \rho_x = 1 \)), but the spatial structures differ, so that \( \mathbf{W}_1 \neq \mathbf{W}_0 \) and \( \mathbf{M}_1 \neq \mathbf{M}_0 \); for simplicity, \( \mathbf{M}_0 = \mathbf{W}_0 \) and \( \mathbf{M}_1 = \mathbf{W}_1 \neq \mathbf{W}_0 \). Setting \( \mathbf{M}_1 = \mathbf{W}_1 \), here causes no loss of identification because of the presence of the non-constant regressors, \( \mathbf{X}_{12} \).

In preliminary trials undertaken by BF, but not reported in detail, regressors were drawn from heavy tailed or very skewed distributions, or with various degrees of spatial correlation, but no major differences in the behaviour of the tests emerged.

For comparability with BF, two spatial frameworks are used for the main results, the 26 counties of Ireland, with weight matrix as employed in Cliff & Ord (1973, p. 164), and a set of 200 EU NUTS-2 regions with weight matrix \( \mathbf{W}_0 \) based on a matrix of 1s and 0s denoting contiguous and non-contiguous regions respectively, subsequently normalized so that rows sum to 1, as used by Fingleton (2007). Under Case 2, the 26 county alternative weight matrix, \( \mathbf{W}_1 \), is defined by replacing the non-zero elements of row \( i \) of the corresponding \( \mathbf{W}_0 \) by \( n_i^{-1} \) the reciprocal of the number of non-zero entries in the \( i \)th row. For the 200 EU regions, with \( w_{ij} = d_{ij}^{-2} \) for \( d_{ij} \leq 300 \) km and \( i \neq j \), \( w_{ij} = 0 \) otherwise, where \( d_{ij} \) is the straight line (Euclidean) distance between centroids of regions \( i \) and \( j \), the weights are defined as \( \mathbf{W}_{ij} = w_{ij} / \sum w_{ij}. \) Thus in these two instances the tests are having to discriminate between quite similar weight matrices. In smaller auxiliary experiments, therefore, two very different matrices with \( n = 25 \) are used: a 5 \( \times \) 5 square, with queen’s case contiguity weights, again row normalized, and a closed ribbon of 25 regions with equal weights of 0.5 on each of two neighbours. Finally, to provide a little further evidence on the effect of increasing \( n \), matrices are built from two diagonal blocks each equal to one of the matrices used for \( n = 200 \), and from a queen’s case 20 \( \times \) 20 square.

For Case 1 the instrument set uses \( r \in (0, 1, 2) \) that is, a minimal, intermediate, and a rich set while in Case 2 only \( r \in (1, 2) \) is relevant for the two-degree-of-freedom test, so, while the one-degree-of-freedom test is computable with \( r = 0 \), these results are not extensively reported. The number of replications is 5,000.
except where the context makes the extra accuracy obtained by using 40,000 especially desirable. The experiments do in fact take significant computer processor time. For example, on the equipment used in this study, 40,000 replications of each of the 50 parameter and r combinations adopted in Case 2 takes about 2 h of CPU time for the \( n = 25 \) lattices. The following parameter values are used: \((\rho_0, \lambda_0) \in (0.0, 0.3, 0.6, 0.9, 0.95)^2\) and \((\rho_0, \lambda_0) = (\rho_1, \lambda_1)\) so that empirical significance levels and powers reflect solely differences between the explanatory variables (Case 1) or weight matrices (Case 2); in Case 1 the explanatory variables observed for region \( i, X_{02i} \) and \( X_{12i} \) \((i = 1, \ldots, n)\) are drawn from a bivariate Normal distribution with variances unity and correlation \( \rho_x \in (-0.5, 0.0, 0.5, 0.9, 0.95)\); the shocks are independent standard Normal, \( v_0 \sim IIDN(0, I_n)\) and similarly \( v_1\) in each case. The matrices, \((I-\lambda_iW_i)\) and \((I-\rho_iM_i)\) \(i = 0, 1\) are non-singular at the chosen parameter values.