ESTIMATION BIAS IN SPATIAL MODELS WITH STRONGLY CONNECTED WEIGHT MATRICES

Tony E. Smith
Department of Systems and Electrical Engineering
University of Pennsylvania

January 22, 2008
(Revised June 8, 2008)

Abstract

In this paper it is shown that for both spatial lag and spatial autoregressive models with strongly connected weight matrices, maximum likelihood estimates of the spatial dependence parameter are necessarily *biased downward*. In addition, it is shown that same bias is present in general Moran tests of spatial dependency, so that positive dependencies will often *fail to be detected* when weight matrices are strongly connected. A simulated numerical example is presented to illustrate some of the practical consequences of these biases.

Key Words: bias, spatial lag models, spatial autoregressive models, Moran test

1. Introduction

In a recent simulation study, Mizruchi and Neuman (2008) have shown that for spatial lag models with strongly connected (high density) weight matrices, there is often a severe downward bias in maximum-likelihood estimates of the spatial dependency parameter.[1] This same bias is also reported by Farber, Páez and Volz (2008) in their recent simulation study of the influence of network topology on tests of spatial dependencies. Hence the central purpose of this paper is to clarify the nature of this bias from an analytical perspective. In addition, it is shown that same bias is present in both spatial autoregressive models and in the more general Moran test of spatial dependency. In all cases this bias implies that significantly positive spatial dependencies will often *fail to be detected* when weight matrices are strongly connected.

To establish these results, the analytical strategy will be to consider the extreme case of *maximally connected* weight matrices, and to obtain exact results for this case. The rest will then follow from simple continuity considerations. To avoid repetition, the analytical development of spatial regression models will focus on spatial lag models. Parallel results for spatial autoregressive models will simply be sketched. Hence to fix the ideas, we begin with the following a standard *spatial lag model* (SL) for n spatial units:

(1) $\quad y = \rho W y + X\beta + \varepsilon\ ,\ \varepsilon \sim N(0, \sigma^2 I_n)$

where $y \in R^n$ is some variable of interest and $X = [1_n, x_1, .., x_k] \in R^{n\times(k+1)}$ represents a relevant set of k explanatory variables, with $1_n = (1, .., 1)'$ denoting the unit n-vector (corresponding to the intercept term in this linear model). [Throughout the following analysis it will always be assumed that X has *full* column rank, $k+1$, so that $(X'X)^{-1}$ exists.] The unknown parameters of the model include the vector, $\beta = (\beta_0, \beta_1, .., \beta_k)'$ of beta coefficients, the variance, σ^2, of each residual in ε, and the *spatial dependence parameter*, ρ, which is of primary interest in the present analysis.

Also of major interest is the structure of the *spatial weight matrix*, W. For purposes of the present analysis, it is convenient to begin by characterizing these matrices in the following way. First we choose a fixed positive scalar, b, to serve as an *upper bound* on weight values. With respect to this bound, an n-square matrix, $W = (w_{ij} : i, j = 1, .. n)$, is designated as a *weight matrix* iff (i) $w_{ii} = 0$ and (ii) $0 \le w_{ij} \le b$ for all $i, j = 1, .., n$. As usual, condition (i) specifies that dependencies are defined only between distinct spatial units. Condition (ii) can be thought of as a normalization condition the allows each weight, w_{ij}, to be interpreted as the "degree of connectivity" between i and j, where $w_{ij} = b$ implies a maximal degree of connectivity. This is particularly appropriate for

[1] I am indebted to a referee for pointing out that similar observations were made by Bao and Ullah (2007) with respect to the second order bias of these estimates in the context of a pure spatial lag model with circular weight matrices of varying degrees of connectivity.

applications of model (1) to *social networks* among n agents. For the present, the bound b only serves as a convenient conceptual device, and can be set equal to one without loss of generality. However, the question of appropriate matrix normalizations for the estimation of ρ is of some importance, and will be addressed below.

If the class of all n-square weight matrices is denoted by $\mathbf{W}_n \subset R^{n \times n}$ (where the fixed scale parameter b is taken to be implicit), then the relevant geometry of this set can be depicted for the $n = 2$ case as follows. Observe that each matrix $W \in \mathbf{W}_n$ is of the form

$$(2) \qquad W = \begin{pmatrix} 0 & w_{12} \\ w_{21} & 0 \end{pmatrix}$$

and thus is fully characterized by the 2-vector (w_{12}, w_{21}). Hence the entire class $\mathbf{W}_2$ is seen to be equivalent to the points in the square, $[0,b]^2$, shown below:

Figure 1 here

Here the lower left hand corner corresponds to the *minimally connected weight matrix*, W_* with all zero components, and the upper right hand corner corresponds to the *maximally connected weight matrix*, W^*,[2] with all off-diagonal elements equal to b. This depiction for the 2x2 case makes it clear that W_* and W^* are the two natural extreme weight matrices in $\mathbf{W}_n$ for all n.[3] Since W_* corresponds to complete *statistical independence* in model (1), attention has naturally focused on those weight matrices, $W \in \mathbf{W}_n$, that are "sufficiently close" to W_* to inherit all of its desirable large-sample properties (such as consistency and asymptotic normality of parameter estimates). Thus most of the literature has focused on those matrices in the lower left neighborhood shown in Figure 1.

In this context, the distinguishing feature of the present analysis is that it focuses rather on the upper right neighborhood in Figure 1, which for the moment we loosely designate as "strongly connected" weight matrices.[4] Our central objective is to show not only that such weight matrices fail to share the desirable properties of the independence case, but also to determine the exact nature of this failure. Of particular interest will be the severe *downward bias* in maximum likelihood estimates of the spatial dependency parameter, ρ.

[2] This terminology is not to be confused with the graph-theoretical notion of "totally connected" which refers only to the presence of a nonzero links between all distinct node pairs.

[3] This can also be expressed in terms of the (cell-wise) matrix inequalities $W_* \leq W \leq W^*$ for all $W \in \mathbf{W}_n$

[4] Here it should be noted that maximally connected spatial weight matrices have been previously studied in a somewhat different context by Kelejian and Prucha (2002) who described them simply as models with "equal spatial weights" [see also Kelejian, Prucha and Yuzefovich (2006) and Baltagi (2006)].

To establish this result in a self-contained manner, it is convenient to begin with a detailed development of the maximum-likelihood estimation problem for the spatial lag model in Section 2 below. This is followed in Section 3 by an analysis of the maximally connected case, W^*, in the upper right corner. The results for this case are extended by continuity in Section 4 to all matrices "sufficiently close" to W^* in an appropriate sense, and some numerical illustrations are given. In Section 5 it is shown that these results are essentially the same for spatial autoregressive models. Finally it is shown in Section 6 that strong connectivity also has consequences for Moran diagnostic tests of spatial independence.

2. Maximum Likelihood Estimation for SL Models

Model (1) implies that y is multinormally distributed, and in particular that for any given data, (y, X) the *log likelihood function* for parameters (β, σ^2, ρ) takes the form:[5]

$$(3) \qquad L(\beta, \sigma^2, \rho \mid y, X) = const - \tfrac{n}{2}\ln(\sigma^2) + \ln|\det(I_n - \rho W)|$$

$$-\tfrac{1}{2\sigma^2}\left((I_n - \rho W)y - X\beta\right)'\left((I_n - \rho W)y - X\beta\right)$$

where I_n is the *n*-square identity matrix, and where all terms not involving the parameters are subsumed in *const*. As with all generalized linear models, one proceeds by first fixing the covariance parameters (in this case, ρ) and maximizing the likelihood function in β and σ^2 to produce the well-known closed form *conditional estimates*:

$$(4) \qquad \hat{\beta}_{sl}(\rho) = (X'X)^{-1}X'(I_n - \rho W)y$$

$$(5) \qquad \hat{\sigma}^2_{sl}(\rho) = (1/n)\left((I_n - \rho W)y - X\hat{\beta}_{sl}(\rho)\right)'\left((I_n - \rho W)y - X\hat{\beta}_{sl}(\rho)\right)$$

where the subscript "*sl*" denotes the SL model. These are then substituted into (3) to yield a reduced function designated as the *concentrated likelihood function,* L_{sl}, for ρ. After some simple cancelling of terms, this function takes the form:

$$(6) \qquad L_{sl}(\rho \mid y, X) = const + \ln|\det(I_n - \rho W)| - (n/2)\ln[\hat{\sigma}^2(\rho)]$$

One then maximizes this function to obtain the *maximum likelihood estimate,* $\hat{\rho}_n$, of ρ and then substitutes this value into (4) and (5) to obtain corresponding *maximum likelihood estimates,* $\hat{\beta}_n = \hat{\beta}_{sl}(\hat{\rho}_n)$ and $\hat{\sigma}^2_n = \hat{\sigma}^2_{sl}(\hat{\rho}_n)$, of β and σ^2 respectively. However, our primary interest here is in $\hat{\rho}_n$ itself.

[5] Most of the following development is quite standard, and can be found in many references including Anselin (1988) and Anselin and Bera (1998, section III.B).

To analyze the function, L_{sl}, one can make further reductions as follows [see also Anselin (1988, Section 12.1.1)]. First let

(7) $\quad M = I_n - X(X'X)^{-1}X'$

denote the *orthogonal projection* onto the complement of the span of X, so that by construction, $M = M'$,

(8) $\quad MX = X - X(X'X)^{-1}X'X = X - X = 0$, and

(9) $\quad MM = (I_n - X(X'X)^{-1}X')(I_n - X(X'X)^{-1}X') = I_n - X(X'X)^{-1}X' = M$

Then substitution of (4) and (7) into (5) yields the more compact form of the conditional variance estimate,

(10) $\quad \hat{\sigma}^2_{sl}(\rho) = (1/n)\left([I_n - X(X'X)^{-1}X'](I_n - \rho W)y\right)' \left([I_n - X(X'X)^{-1}X'](I_n - \rho W)y\right)$

$$= (1/n)\left(M(I_n - \rho W)y\right)' \left(M(I_n - \rho W)y\right)$$

$$= (1/n)\left(y'(I_n - \rho W)'M(I_n - \rho W)y\right)$$

This in turn allows the concentrated likelihood in (6) to be written as

(11) $\quad L_{sl}(\rho \mid y, X) = const + \ln|\det(I_n - \rho W)| - (n/2)\ln[y'(I_n - \rho W)'M(I_n - \rho W)y]$

where the term $-(n/2)\ln(1/n)$ has now been absorbed into the constant.

Further reduction is possible by observing that if the eigenvalues of W are denoted by $\lambda(W) = \{\lambda_i : i = 1,..,n\}$, then the corresponding eigenvalues of $(I_n - \rho W)$ are well known to be given by $\lambda(I_n - \rho W) = \{1 - \rho\lambda_i : i = 1,..,n\}$. To avoid complications in the analysis to follow, it is convenient to restrict our attention to weight matrices, W, with *real* eigenvalues (which, most importantly, includes all W which are either *symmetric* or are row normalizations of symmetric matrices). In addition, it will also be assumed that the maximum eigenvalue, $\lambda_{\max}(W)$, of W is positive.[6] (In particular this includes all *nonzero symmetric* weight matrices.) Hence we now restrict our attention to the subset:

(12) $\quad \mathbf{W}_n^+ = \{W \in \mathbf{W}_n : \lambda(W)$ *is real, and* $\lambda_{\max}(W) > 0\}$

[6] This maximum eigenvalue is always nonnegative [Horn and Johnson (1985, Th.8.1.3)], but need not be positive even when W has positive elements. Even for $n = 2$ the matrix, $W = [0\ 1; 0\ 0]$, has $\lambda(W) = \{0, 0\}$.

Given this subset, together with the fact that the determinant of any matrix is the product of its eigenvalues [Horn and Johnson (1985,Th.1.2.12)], it then follows that

$$(13) \qquad \det(I_n - \rho W) = \prod_i (1 - \rho\lambda_i) \;\Rightarrow\; \ln|\det(I_n - \rho W)| = \sum_i \ln|1 - \rho\lambda_i|$$

as long as each term, $1 - \rho\lambda_i$, on the right hand side is *nonzero*. This of course requires further restrictions on ρ. To specify these conditions, we first note that since the trace of every matrix is the sum of its eigenvalues [Horn and Johnson (1985,Th.1.2.12)], it follows that

$$(14) \qquad \sum_i \lambda_i = tr(W) = \sum_i w_{ii} = 0$$

for all $W \in \mathbf{W}_n$. But since $\lambda_{\max}(W) > 0$ for all $W \in \mathbf{W}_n^+$, this in turn implies that $\lambda_{\min}(W)$, must be *negative*. These observations together imply that for any $W \in \mathbf{W}_n^+$, all terms, $1 - \rho\lambda_i$, in (13) will be *positive* if the admissible values of ρ are restricted to the *open interval*

$$(15) \qquad [W] = \left(\frac{1}{\lambda_{\min}(W)}, \frac{1}{\lambda_{\max}(W)}\right)$$

Hence we now restrict ρ to the interval, $[W]$. Under this restriction, (13) allows (11) to be reduced to the explicit form,

$$(16) \qquad L_{sl}(\rho \mid y, X) = const + \sum_i \ln|1 - \rho\lambda_i| - (n/2)\ln[y'(I_n - \rho W)'M(I_n - \rho W)y]$$

which is more readily analyzed (and computed).

At this point one typically proceeds by observing that since $\ln|\det(I_n - \rho W)| = -\infty$ on the boundaries of $[W]$, it is reasonable to assume that L_{sl} has a well-defined differentiable maximum in the open interval $[W]$. This will be true as long as the second term in (16) is *bounded above*. To ensure this, it must of course be assumed that

$$(17) \qquad M(I_n - \rho W)y \neq 0 \;\; \textit{for all} \;\; \rho \in [W]$$

To interpret this condition, observe that model (1) can be equivalently written as $(I_n - \rho W)y = X\beta + \varepsilon$, where $(I_n - \rho W)y$ represents the value of y after spatial lag effects have been accounted for. If this variable is designated as the *effective value* of y,

$$(18) \qquad y_W(\rho) = (I_n - \rho W)y$$

in model (1), then as a parallel to classical regression, it is here assumed that for the given data vector, y, none of its effective values, $\{y_W(\rho) : \rho \in [W]\}$ is perfectly fitted by X (i.e., lies in the span of X). We designate data sets (y, X) satisfying (17) as *W-regular*.

Notice that for $\rho = 0$ this implies the usual regularity condition that $My \neq 0$. Data (y, X) satisfying only this (classical regression) condition is simply said to be *regular*.

3. Biased Estimation for the Maximally Connected Case in SL Models

Given the simple form of the concentrated likelihood function, L_{sl}, in (16), one can proceed to search for a maximum, $\hat{\rho}_n$, in the interval $[W]$ (typically by standard line search procedures). However, it turns out that for the maximally connected case, $W^* \in \mathbf{W}_n^+$, this maximization procedure is doomed to fail. Indeed, the main result of this section will be to show that even for regular data sets, L_{sl} *is always unbounded on* $[W^*]$. To establish this result, we begin by analyzing the properties of W^*. First observe that since the n-square unit matrix is constructible as the outer product, $1_n 1_n'$, the maximally connected weight matrix, $W^* \in \mathbf{W}_n^+$, can be written as:

$$(19) \qquad W^* = b \cdot (1_n 1_n' - I_n) = \begin{pmatrix} 0 & b & \cdots & b \\ b & 0 & & \vdots \\ \vdots & & \ddots & b \\ b & \cdots & b & 0 \end{pmatrix}$$

With this explicit form, the following result shows that the eigenvalues of W^* are computable in closed form:

Lemma 1. *For all* $b > 0$ *the eigenvalues of* W^* *in* (19) *are given by*

$$(20) \qquad \lambda(W^*) = \{-b, .., -b, b(n-1)\}$$

where the eigenvalue, $-b$*, has multiplicity* $n-1$.

Proof: It follows from Searle (1982, Section 12.3.d) that the eigenvalues of any matrix of the form $A = aI + c11'$ are given by

$$(21) \qquad \lambda(A) = \{a, ..., a, (a + nc)\}$$

where a has multiplicity $n-1$. Hence the eigenvalues of

$$(22) \qquad W^* = b \cdot (1_n 1_n' - I_n) = (-b) I_n + (b) 1_n 1_n'$$

are immediately seen to be those in (20).

The second (and most important) property of maximally connected weight matrices is the following identity:

Lemma 2. *If* M *is the orthogonal projection matrix in* (7) *associated with any data matrix,* $X = [1_n, x_1, .., x_k]$, *for model* (1) *then,*

(23) $M \cdot W^* = -b \cdot M = W^* \cdot M$

Proof: Simply observe from (19) and (7) that

(24) $M \cdot W^* = (I_n - X(X'X)^{-1}X') \cdot b \cdot (1_n 1_n' - I_n)$

$$= b \cdot [1_n 1_n' - I_n - X(X'X)^{-1}X' 1_n 1_n' + X(X'X)^{-1}X']$$

But since $[X(X'X)^{-1}X']X = X$ and since 1_n is the first column of X, it follows in particular that $[X(X'X)^{-1}X']1_n = 1_n$. Hence we see that

(25) $M \cdot W^* = b \cdot [1_n 1_n' - I_n - 1_n 1_n' + X(X'X)^{-1}X'] = b \cdot [X(X'X)^{-1}X' - I_n] = -b \cdot M$

Next observe that since $[X(X'X)^{-1}X']1_n = 1_n \Rightarrow 1_n' = 1_n'[X(X'X)^{-1}X']$, it also follows that

(26) $W^* \cdot M = b \cdot (1_n 1_n' - I_n) \cdot (I_n - X(X'X)^{-1}X')$

$$= b \cdot [1_n 1_n' - 1_n 1_n' X(X'X)^{-1}X' - I_n + X(X'X)^{-1}X']$$

$$= b \cdot [1_n 1_n' - 1_n 1_n' - I_n + X(X'X)^{-1}X']$$

$$= b \cdot [-I_n + X(X'X)^{-1}X'] = -b \cdot M \ .$$

One useful consequence of this result is the following:

Lemma 3. *Every regular data set,* (y, X), *is* W^**-regular.*

Proof: First observe from Lemma 2, together with the symmetry of W^* and M, that for any $\rho \in [W^*]$ and data set (y, X),

(27) $y'(I_n - \rho W^*)'M(I_n - \rho W^*)y = y'(I_n - \rho W^*)'(M - \rho MW^*)y$

$$= y'(M - \rho MW^* - \rho W^* M + \rho^2 W^* MW^*)y$$

$$= y'(M + \rho bM + \rho bM + \rho^2 b^2 M)y$$

$$= (1 + 2\rho b + \rho^2 b^2) \cdot y'My$$

$$= (1+\rho b)^2 \cdot y'My$$

But $\rho \in [W^*]$ then implies that $\rho > -1/b$ and hence that $1+\rho b > 0$. Thus W^*-regularity of (y, X) will follow if it can be shown that $y'My > 0$. But since M is an orthogonal project matrix and hence is *positive semidefinite*, it follows that $y'My \geq 0$ for all y, and moreover that $y'My = 0 \Leftrightarrow My = 0$ [Horn and Johnson (1985, p.400)]. Finally since the regularity of (y, X) implies that $My \neq 0$, it must then be true that $y'My > 0$, and thus that W^*-regularity holds.

With these properties, we are now ready to establish our main result, namely that L_{sl} is unbounded on $[W^*]$. In particular, we show that L_{sl} increases without bound as ρ approaches the lower boundary of $[W^*]$. To so, observe from (15) and (20) that this *lower boundary point*, ρ_*, is given by

(28) $$\rho_* = 1/\lambda_{\min}(W^*) = -1/b$$

With this definition we now have:

Proposition 1. *If* $W = W^*$ *in model* (1), *then for all regular data sets* (y, X) *and all decreasing sequences* (ρ_m) *in* $[W^*]$, *with* $\lim_{m\to\infty} \rho_m = \rho_*$,

(29) $$\lim_{m\to\infty} L_{sl}(\rho_m \mid y, X) = \infty$$

Proof: The strategy will be to use Lemmas 1 and 2 to show that the concentrated likelihood function in (16) is reducible to a simple analytical form for which the result is obvious. To do so, we first observe from Lemma 1 and the positivity of $\min_i\{1-\rho\lambda_i(W)\}$ on $[W]$ that for any $\rho \in [W^*]$ we must have[7]

(30) $$\sum_i \ln|1-\rho\lambda_i| = \sum_i \ln(1-\rho\lambda_i) = (n-1)\ln[1-\rho(-b)] + \ln[1-\rho b(n-1)]$$

$$= (n-1)\ln(1+\rho b) + \ln[1-\rho b(n-1)]$$

Moreover, we see from (27) above that

(31) $$-(n/2)\ln[y'(I_n - \rho W^*)'M(I_n - \rho W^*)y] = -(n/2)\ln[(1+\rho b)^2 y'My]$$

$$= -\{n\ln(1+\rho b) + (n/2)\ln(y'My)\}$$

[7] For the case of $b = 1/(n-1)$ this result appears in section 2.5 of Kelejian and Prucha (2002).

Notice also from Lemma 3 that this log expression is *well defined* for all $\rho \in [W^*]$. Hence by substituting (30) and (31) into (16) we obtain the following simple expression for the concentrated likelihood function,

$$(32) \quad L_{sl}(\rho \mid y, X) = const + \{(n-1)\ln(1+\rho b) + \ln[1-\rho b(n-1)]\}$$

$$-\{n\ln(1+\rho b) + (n/2)\ln(y'My)\}$$

$$= const - \ln(1+\rho b) + \ln[1-\rho b(n-1)]$$

where the term, $(n/2)\ln(y'My)$, not containing ρ has again been absorbed in *const*. From here we need only observe that since $\rho_* = -1/b$, it follows that for any decreasing sequence (ρ_m) in $[W^*]$, with $\lim_{m\to\infty} \rho_m = \rho_*$, we must have

$$(33) \quad \lim_{m\to\infty} L_{sl}(\rho_m \mid y, X) = const - \lim_{m\to\infty} \ln(1+\rho_m b) + \lim_{m\to\infty} \ln[1-\rho_m b(n-1)]$$

$$= const - \ln(1+\rho_* b) + \ln[1-\rho_* b(n-1)]$$

$$= const - \ln(0) + \ln(n) = \infty$$

and the result is established.[8]

Hence from a formal viewpoint, it may be concluded that *no maximum likelihood estimator of ρ exists* for model (1) when $W = W^*$.[9] This somewhat surprising in view of the fact that existence of maximum likelihood estimators for model (1) is generally assumed to hold as long as $W \in \mathbf{W}_n$ and $\rho \in [W]$. Moreover, it is interesting to note that from a practical viewpoint, such a failure would most likely not even be detected by standard software. Indeed, one would typically observe that the line-search algorithm has "converged" to some value of ρ very close to ρ_*.

To gain further insights here, it is useful to illustrate this finding with a numerical example, as shown in Figure 2 below. This is taken from the numerical simulation example presented in Section 4 below (for a sample size of $n = 50$). The "First Term" and "Second Term" shown in Figure 2 correspond, respectively, to the log-determinant expression (30) and the log-quadratic expression (31) in Proposition 1 above. Notice that the log-determinant term is always well behaved, since it is a sum of simple concave functions, $\ln(1-\rho\lambda_i)$, on $[W^*]$. Hence the "culprit" here is the log-quadratic term, which

---

[8] Note that L_{sl} is also unbounded at the upper boundary of $[W^*]$, namely $\rho^* = 1/\lambda_{\max}(W^*) = 1/[b(n-1)]$. But since $L_{sl}(\rho^* \mid y, X) = -\infty$, this is of little interest for maximum likelihood estimation.

[9] This failure of existence is an instance of the more general result of Arnold (1979, Th.3) regarding the non-existence of maximum-likelihood estimators for covariance parameters in linear models with exchangeably distributed errors. I am indebted to Federico Martellosio for pointing this out to me.

in the present case not only diverges to $+\infty$ at ρ_*, but does so at a faster rate than the corresponding divergence of the log-determinant to $-\infty$.

Figure 2 Here

Before examining the practical consequences of this result for strongly connected weight matrices, we give an alternative statement of Proposition 1 that will also prove useful for applications. Recall that our basic regularity assumption on data (y, X) was designed to avoid cases where some effective y-value, $y_W(\rho)$ was perfectly fitted by the data, X, in model (1). We now show that Proposition 1 results from the fact that for *every* data set (y, X) in model (1), if $W = W^*$, the X must yield a *perfect fit* to the "effective" y-value, $y_{W^*}(\rho_*)$, on the lower boundary of $[W^*]$. This fact depends critically on the presence of an intercept term in model (1) [as should already be apparent from the proof of Lemma 2]. Hence it is now convenient to make this intercept term explicit by rewriting model (1) as

$$(34) \qquad y = \rho W y + \beta_0 1_n + \tilde{X}\tilde{\beta} + \varepsilon \ , \ \varepsilon \sim N(0, \sigma^2)$$

where $\tilde{X} = [x_1, .., x_k]$ and $\tilde{\beta} = (\beta_1, .., \beta_k)'$. For the particular case of W^* it then follows that for any choice of $\tilde{X}$,

$$(35) \qquad y = \rho W^* y + \beta_0 1_n + \tilde{X}\tilde{\beta} + \varepsilon$$

$$\Rightarrow (I_n - \rho W^*) y = \beta_0 1_n + \tilde{X}\tilde{\beta} + \varepsilon$$

$$\Rightarrow y_{W^*}(\rho) = \beta_0 1_n + \tilde{X}\tilde{\beta} + \varepsilon$$

With this notation, recall from Lemma 3 and (17) that for any regular data set $(y, X) = (y, [1_n, \tilde{X}])$ there exists no $\rho \in [W^*]$ such that the effective value $y_{W^*}(\rho)$ is a perfect fit to X, i.e., such that

$$(36) \qquad y_{W^*}(\rho) = \beta_0(\rho) 1_n + \tilde{X}\tilde{\beta}(\rho)$$

for some choice of $[\beta_0(\rho), \tilde{\beta}(\rho)]$. But even when this is true, it turns out that condition (36) *always* fails at the lower boundary value, ρ_*, of $[W^*]$ as we now show:[10]

[10] The following result is essentially contained in Theorem 1 of Kelejian, Prucha and Yuzefovich (2006), where it is employed to study the consistency properties of 2SLS estimation in the case of equal spatial weights.

Proposition 2. *If* $W = W^*$ *in model* (1)*, then for any data set* (y, X),

(37) $$y_{W^*}(\rho_*) = \beta_0(\rho_*) \cdot 1_n + \tilde{X}\tilde{\beta}(\rho_*)$$

where $\beta_0(\rho_*) = 1_n' y$ *and* $\tilde{\beta}(\rho_*) = 0$.

Proof: First recall from (28) that for any positive bound b,

(38) $$\rho_* = 1/\lambda_{\min}(W^*) = -1/b < 0$$

Hence it follows that,

(39) $$y_{W^*}(\rho_*) = (I_n - \rho_* W^*)y = \{I_n - (-1/b)[b \cdot (1_n 1_n' - I_n)]\}y$$

$$= [I_n + (1_n 1_n' - I_n)]y = 1_n 1_n' y = (1_n' y) \cdot 1_n + X \cdot 0$$

$$= \beta_0(\rho_*) \cdot 1_n + X\beta(\rho_*)$$

and the result is established.

The advantage of viewing this result in terms of perfect fits is that it provides information about the bias of other parameter estimates. For when $\hat{\rho}_n \approx \rho_*$, expression (37) suggests that one should have $\hat{\beta}_0 \approx 1_n' y$ and $\hat{\beta}_j \approx 0$ for all $j = 1,..,k$. Moreover, since a "perfect fit" necessarily implies zero variance of residuals, this in turn suggests that $\hat{\sigma}^2 \approx 0$.[11] We shall explore the practical consequences of these findings in the next section.

4. Consequences for Strongly Connected Weight Matrices in SL Models

The above results show that for the extreme case of maximally connected matrices, we can obtain an exact analytical formulation of the bias inherent in maximum likelihood estimation for spatial lag models. This in turn suggests that such bias should be inherited by matrices, $W \in \mathbf{W}_n^+$ that are "close" to W^* in some appropriate sense. To do so, it is convenient to endow $\mathbf{W}_n^+$ with a matrix norm that will allow some explicit measure of "closeness". Here there are many choices. For example, the ℓ_1*-norm* of any matrix $A = (a_{ij}) \in R^{n \times n}$ is $\|A\|_1 = \sum_{ij} |a_{ij}|$, and the ℓ_2*-norm* (*Euclidean norm*) of A is $\|A\|_2 = \left(\sum_{ij} a_{ij}^2\right)^{1/2}$.[12] However, for our present purposes, the following scaled version of the ℓ_1-norm is useful for weight matrices, $W \in \mathbf{W}_n^+$,

[11] These results illustrate the more general finding of Arnold (1979, p.196) regarding the inconsistency of standard parameter estimates for linear models with exchangeably distributed errors.
[12] Many other choices are illustrated in Horn and Johnson (1985, Section 5.6).

$$(40)\qquad \|W\|_{rc} = \frac{\|W\|_1}{\|W^*\|_1} = \tfrac{1}{b\cdot n(n-1)}\sum_{ij} w_{ij} = \tfrac{1}{n(n-1)}\sum_{ij}(w_{ij}/b)\,.$$

which we designate as the *relative connectivity norm*.[13] If (w_{ij}/b) denotes the relative connectivity between units (agents) i and j, then this is simply the average of these relative connectivities over all distinct (i, j) pairs. In the case of binary matrices, $W \in \mathbf{W}_n^+$, this is easily seen to reduce to the graph-theoretic notion of *average link density*. Given this norm (or any other matrix norm), the induced *distance* between W and W^* is then given by

$$(41)\qquad \|W - W^*\|_{rc} = \tfrac{1}{b\cdot n(n-1)}\sum_{ij}|\, w_{ij} - b\,| = \tfrac{1}{n(n-1)}\sum_{ij}[1-(w_{ij}/b)]$$

Next, we observe that up to this point the actual *magnitude* of ρ has not been considered. All that has been asserted is that for any given weight matrix, W, these values must lie in the open interval $[W]$ of expression (15), and that this interval contains zero (so that both positive and negative values of ρ are always possible.) But to gain further insight, it is useful to evaluate this interval in specific cases. In the numerical illustration below we shall use a sample size, $n = 50$. Hence, by setting the bound at $b = 1$, it follows from Lemma 1 that for the maximally connected matrix, $W^* \in \mathbf{W}_{50}^+$, we have $\lambda_{\min}(W^*) = -1$ and $\lambda_{\max}(W^*) = 49$. The corresponding bounds on ρ for this case are thus seen to be

$$(42)\qquad \rho \in \left(-1/b\ ,1/b(n-1)\right) = \left(-1\,,\,\tfrac{1}{49}\right) \approx (-1\,,\,.02)$$

which, from a practical viewpoint, is seen to offer little room for positive spatial dependencies at all. Since positive dependencies are by far the most interesting for practical applications, it is clear that a better choice of b should be considered. Here the most natural choice is to set $b = 1/(n-1)$, so that under this normalization we obtain

$$(43)\qquad \lambda_{\max}(W^*) = b(n-1) = (n-1)/(n-1) = 1$$

This will ensure that the interval, $[0,1)$, of nonnegative ρ-values used for most applications actually lies in $[W^*]$. For $n = 50$ we then have

$$(44)\qquad [W^*] = \left(-1/b\,,\,1\right) = \left(-(n-1)\,,\,1\right) = (-49\,,\,1)$$

[13] Since every positive scaling of a norm yields another norm, the first equality shows that this is indeed a matrix norm.

One additional feature of this normalization that is particularly useful for the present analysis is that[14]

(45) $[0,1) \subset [W]$ *for all* $W \in \mathbf{W}_n^+$

So this same interval of ρ-values is available for every choice of $W \in \mathbf{W}_n^+$.[15]

Given this normalization, the objective of this section is to extend the bias results for maximally connected weight matrices W^* in Proposition 1 to all weight matrices, $W \in \mathbf{W}_n^+$, that are *strongly connected* in the sense that they are "sufficiently close" to W^* in the relative connectivity norm. To do so, it is convenient to introduce the following additional conventions. First, for any given $W \in \mathbf{W}_n^+$ and data set (y, X) for model (1), we shall write the maximum likelihood estimator for ρ as $\hat{\rho}_W(y, X)$. As pointed out above, this estimator can fail to exist even when (y, X) is W-regular. But for weight matrices close to W^* (in relative connectivity), it should be clear that if (y, X) is W-regular, then a differentiable maximum, $\hat{\rho}_W(y, X)$, fails to exist only when L_{sl} is unbounded at the lower boundary of $[W]$. In such cases, we simply set $\hat{\rho}_W(y, X)$ equal to this lower boundary, so that $\hat{\rho}_W(y, X)$ can be treated as a well-defined value for each W. Next, to quantify the possible bias of these estimates, it is convenient to focus only on the most important case of *positive* dependencies in model (1), i.e., $\rho > 0$, and to quantify various degrees of underestimation by inequalities of the form,

(46) $\hat{\rho}_W(y, X) < \rho/(1+\alpha)$

where parameter $\alpha > 0$ can be interpreted as a *bias factor*. For example, a bias factor of $\alpha = 1$ would imply that $\hat{\rho}_W(y, X)$ is less than half the true value of ρ. More generally, higher bias factors correspond to more severe underestimation of ρ. With these conventions, we now have the following consequence of Proposition 1:

> **Proposition 3.** *For any regular data set* (y, X) *with* $n \geq 3$ *and any given value,* $\rho_0 \in (0,1)$*, of the spatial dependence parameter for model* (1)*, there exists for each choice of bias factor,* $\alpha \in (0,1)$*, a sufficiently small* $\varepsilon = \varepsilon(\alpha, \rho_0, y, X) > 0$ *such that for all* $W \in \mathbf{W}_n^+$*,*

(47) $\left\| W - W^* \right\|_{rc} < \varepsilon \Rightarrow \hat{\rho}_W(y, X) < \rho_0/(1+\alpha)$

---

[14] Expression (45) follows from the fact that $0 \leq W \leq W^* \Rightarrow \lambda_{\max}(W) \leq \lambda_{\max}(W^*) = 1 \Rightarrow 1/\lambda_{\max}(W) \geq 1$ [see Horn and Johnson (1985, Corollary 8.1.19)].

[15] An alternative normalization that also shares this property is to set b equal to the reciprocal of the smallest row or column sum, as proposed by Kelejian and Prucha (2008, Lemma 2). Though less standard than the present convention, this normalization has the advantage of being much easier to compute for large weight matrices.

Proof Sketch: The proof of this result is rather technical, and is deferred to the Appendix. But the basic idea is simple. Observe from Figure 2 that not only does the concentrated likelihood function, L_{sl}, diverge to $+\infty$ at ρ_*, but in fact its derivative is *everywhere negative* in $[W^*]$. Hence, if we now write the concentrated likelihood function as, $L_{sl}(\rho \mid y, X, W)$, to emphasize its dependence on W [as well as data (y, X)], then the strategy of the proof is to show that the corresponding derivative, $L'_{sl}(\rho \mid y, X, W)$, with respect to ρ is *continuous* in W at the point W^*. Using this continuity property, it is then possible to show that for any choice of bias factor, α, when W is sufficiently close to W^* [i.e., when ε in (47) is sufficiently small], one can guarantee that $L'_{sl}(\rho \mid y, X, W)$ will be negative for all $\rho \in [W]$ with $\rho \geq \rho_0/(1+\alpha)$, and thus that $L_{sl}(\rho \mid y, X, W)$ can only achieve a maximum on $[\rho_*, \rho/(1+\alpha))$.

In other words, for any degree of bias, $\alpha > 0$, there is some threshold level of "strong connectivity", $\|W - W^*\| < \varepsilon$, which is sufficient to ensure this degree of bias. The proof sketched above also shows (from the persistence of negative slopes) that under conditions of *no* spatial dependence (i.e, $\rho = 0$) this null hypothesis will tend to be falsely rejected in favor of *negative* dependencies $(\rho < 0)$ for strongly connected weight matrices. Moreover, in cases where such dependencies are indeed negative, the strength of these dependencies will tend to be overestimated. But as with all such continuity results, Proposition 3 still leaves open the question of how "strong" this connectivity must be in order to see a substantial effect. While such questions can only be answered definitively by extensive simulations, it is nonetheless possible to illustrate the potential significance of these results by means of a "typical" example.[16]

Here we set $n = 50$, $k = 2$, and construct x-data (x_1, x_2) by simulating two uniformly distributed random vectors, so that $X = [1_{50}, x_1, x_2]$. Model (1) was then parameterized with $\beta = (\beta_0, \beta_1, \beta_2)' = (1, 2, 3)$ and standard deviation $\sigma = 1$. Again for sake of illustration the single value, $\rho = .5$, was chosen to represent (substantial) positive spatial dependency in model (1). To analyze the effects of strong connectivity, only symmetric binary weight matrices were used in order to allow an *average link density* interpretation of the matrix norm in (40). A number of matrices, W, with different *average link densities*, $d = \|W\|_{rc} \in (0,1)$ were randomly sampled. In particular, the values $d \in \{.30, .50, .80, .90, .95, .99\}$ were chosen for study. For each d a matrix, $W_{(d)} \in \mathbf{W}_n^+$,

[16] This example is only meant to illustrate the practical consequences of the analytical results above. As mentioned in the introduction, more extensive and systematic simulations can be found in both Mizruchi and Neuman (2008) and Farber, S., A. Páez and E. Volz (2008).

was randomly sampled from the distribution independently assigning $w_{ij} = 1$ with probability d and $w_{ij} = 0$ otherwise.[17]

In order to make the results at different density levels more readily comparable, each matrix W_d was normalized in the same manner as W^*, by dividing W_d by its maximum eigenvalue. This rescaling ensures that the positive values of ρ in each simulated model are *exactly the same*, namely $\rho \in (0,1)$.[18] For each of these matrices, 1000 y-vectors were then simulated for model (1), and corresponding maximum-likelihood estimates $\{\hat{\rho}_d(s) : s = 1,..,1000\}$ were computed.[19] Perhaps the simplest way to summarize these results is to compare the sample mean values of $\hat{\rho}_d$ for each of these densities with the true value, $\rho = .50$, as in column 2 of Table 1 below.

Table 1 Here

As expected, one sees underestimation in all cases, with steadily increasing severity for higher densities. For comparison, the maximally connected case, $d = 1$, has been added to show this extreme case is vastly worse than all others. But nonetheless, one can see the continuity properties in Proposition 3 at work. Underestimation becomes quite severe as connectivity density increases. Note also that in Table 1 the corresponding ρ-intervals, $[W_d]$, in (15) above are given in column 4 (column 3 will be discussed below).

To provide a fuller comparison, selected histograms of $\{\hat{\rho}_d(s) : s = 1,..,1000\}$ are shown for the cases $d = .50,.80,.90,.99$ in Figure 3 below.[20] Here the true value, $\rho = .50$, is

Figure 3 Here

indicated by a bold arrow in each case to facilitate the visual comparison of these estimates. So at average link-density levels of at least 80% ($d \geq .80$) there is a substantial downward bias in ρ estimates. Another way to see this is to ask what fraction of these estimates are reported as significantly different from zero in the standard two-sided tests using asymptotic z-values.[21] Here, for a true value of $\rho = .50$, only the upper 42% of sample estimates at $d = .80$ are significantly different from zero. When the density is

[17] Note that density values, d, can only be approximated by this sampling procedure. However, repeated samples at each density level yielded variations that were too small to warrant reporting. In all cases the matrix, $W_{(d)}$ chosen had an average link density well within .01 of d.

[18] The normalization, $b = 1/(n-1) = 1/49$, used above has the theoretical advantage of preserving all relative connectivity relationships. But the present scaling to unit maximum eigenvalues is a more typical normalization used in practice. For comparison, calculations were also done for the $1/49$ scaling, and produced even more dramatic underestimation results than those presented here.

[19] The estimation was done in Matlab using a modified version of the LeSage (1999) suite of programs.

[20] Cases $d = .30$ and. $d = .95$ are, respectively, very similar to $d = .50$ and $d = .90$, and are omitted.

increased to $d = .90$ this drops to less than 15%. Further investigations of such significance questions will be taken up in Section 5 below.

Finally, it is of interest to recall from the discussion following Proposition 2 above that this underestimation of ρ has consequences for the bias of other parameter estimates that are at least qualitatively predictable. While it is difficult to place magnitudes on the degree of these biases, they can at least be illustrated for the simulations of model (1) above. The mean estimates for all parameters are shown in Table 2 below (where the means for $\hat{\rho}$ have been repeated from Table 1):

Table 2 Here

Recall from Proposition 2 that for the "perfect fit" case in the last row of Table 2, one would predict an intercept coefficient, $\hat{\beta}_0 \approx 1_n' y$. In the present case, the mean value of $1_n' y$ was about 351, which is in clear agreement with Table 2. Hence for strongly connected weight matrices this is seen to result in extreme overestimation of β_0 in the present case. It is also interesting to note that while the limiting estimates of $\beta = (\beta_1, \beta_2)'$ and σ^2 in Table 2 also agree with the zero values predicted by Proposition 2, these biases seem to disappear much more rapidly as link density decreases. However, it is worth noting that even a slight downward bias in $\hat{\sigma}^2$ (and hence $\hat{\sigma}$) can have potentially serious consequences for testing, where it can lead to erroneous significance of beta parameters.

5. Extension to Spatial Autoregressive Models

The results above demonstrate that strong connectivity of weight matrices can lead to severe bias in the estimation of spatial dependencies in spatial lag models. Hence it is natural to ask whether similar behavior is exhibited by spatial autoregressive models. Our main result here is to show that with respect to spatial dependence parameters, the results for these two models are essentially *identical*. To establish this, we begin by formulating this model and sketching the parallel maximum likelihood estimation problem for this case. As a parallel to model (1), the standard *spatial autoregressive model* (SAR)[22] for n spatial units:

(48) $\quad y = X\beta + u\ ,\ u = \rho W u + \varepsilon,\ \ \varepsilon \sim N(0, \sigma^2 I_n)$

[21] Note that for tests of positive ρ it is theoretically more appropriate to consider a one-sided test ($\rho > 0$). But such results are not reported in standard spatial regression software.

[22] This is also known as the *spatial errors model*, to emphasize the spatial dependence among errors.

where now the *spatial dependence parameter*, ρ, and *spatial weight matrix*, W, characterize possible spatial dependencies among the residuals rather than the dependent variable, y.[23] If one solves for u and writes this model in reduced form as

$$(49) \qquad y = X\beta + (I_n - \rho W)^{-1}\varepsilon, \;\; \varepsilon \sim N(0, \sigma^2 I_n)$$

then it becomes clear that ρ and W directly influence the covariance structure of the residuals, ε. Again y is multinormally distributed, where the *log likelihood function* for parameters (β, σ^2, ρ) in (3) now takes the form:

$$(50) \qquad L(\beta, \sigma^2, \rho \mid y, X) = const - \tfrac{n}{2}\ln(\sigma^2) + \ln|\det(I_n - \rho W)|$$

$$- \tfrac{1}{2\sigma^2}(y - X\beta)'(I_n - \rho W)'(I_n - \rho W)(y - X\beta)$$

The parallel between (3) and (50) is even more clear when one solves for the conditional estimates of β and σ^2 given ρ,

$$(51) \qquad \hat{\beta}_{sar}(\rho) = [X'(I_n - \rho W)'(I_n - \rho W)X]^{-1}X'(I_n - \rho W)'(I_n - \rho W)y$$

$$(52) \qquad \hat{\sigma}^2_{sar}(\rho) = (1/n)(y - X\hat{\beta}_{sar}(\rho))'(I_n - \rho W)'(I_n - \rho W)(y - X\hat{\beta}_{sar}(\rho))$$

and substitutes into (50) to obtain the *concentrated likelihood function*, L_{sar}, for ρ. Again, after cancelling of terms, this function reduces to

$$(53) \qquad L_{sar}(\rho \mid y, X) = const + \ln|\det(I_n - \rho W)| - (n/2)\ln[\hat{\sigma}^2(\rho)]$$

which is seen to be identical in form to (6).[24] Hence these concentrated likelihood functions differ only with respect to their corresponding conditional variance estimates in (5) and (52). However, for the special case of maximally connected weight matrices, W^*, it turns out that these conditional variance estimates are *identical*, as we now show.

To do so, we begin with the following preliminary result on a certain class of orthogonal projections, which are exemplified by the key projection, $X(X'X)^{-1}X'$, embodied in expression (7) for M. If for any matrix, $A \in R^{n \times k}$, of full column rank $k \leq n$ we now let $P_A = A(A'A)^{-1}A'$ denote the orthogonal projection of $R^{n \times k}$ into the span of A (so that by definition, $P_A A = A$) then we have the following useful condition for equality between such projections:

[23] Although the spatial dependence parameter in this model acts on residuals rather than y, we choose to keep the same notation, ρ, in order to emphasize the parallels between these two models.

[24] In particular the constant terms, *const*, are also easily shown to be identical.

Lemma 4. *For any matrices,* $A, B \in R^{n \times k}$ *, of full column rank,*

(54) $P_A = P_B \Leftrightarrow P_A B = B$

Proof: Since $P_B B = B$, it follows at once that $P_A = P_B \Rightarrow P_A B = P_B B = B$. So we need only show the converse. To do so, observe first that

(55) $P_A B = B \Rightarrow P_A B[(B'B)^{-1}B'] = B[(B'B)^{-1}B'] \Rightarrow P_A P_B = P_B$

Moreover, it also follows that

(56) $B = P_A B = A(A'A)^{-1}A'B = A(A'A)^{-1}(A'B)$

$$\Rightarrow B'B = B'A(A'A)^{-1}(A'B) \Rightarrow |B'B| = |B'A| \cdot |A'A|^{-1} \cdot |A'B|$$

$$\Rightarrow |A'B|^2 = |B'B| \cdot |A'A| > 0$$

and hence that $A'B$ is nonsingular. Thus by the first line of (56) we have

(57) $P_B B = B \Rightarrow P_B A(A'A)^{-1}(A'B) = A(A'A)^{-1}(A'B) \Rightarrow P_B A = A$

where the last implication follows by post-multiplication of both sides by the inverse of the nonsingular matrix $(A'A)^{-1}(A'B)$. Hence by the argument in (55)

(58) $P_B A = A \Rightarrow P_B P_A = P_A \Rightarrow P_A P_B = P_A$

where here the last implication follows by taking transposes of both sides and using the symmetry of P_A and P_B. Hence it be concluded from (55) and (58) that

(59) $P_A = P_A P_B = P_B$

and the result is established.

With this result, we now have the following key identity between SL models (1) and SAR models (49) for the case of maximally connected weight matrices.

Proposition 4. *If* $W = W^*$ *in models* (1) *and* (49)*, then the concentrated likelihood functions* L_{sl} *and* L_{sar} *are identical for all* $\rho \in [W^*]$.

Proof: To establish this result, it is clear from (6) and (53) that it suffices to show that the conditional variance estimates in (10) and (52) are identical on $[W^*]$. But if for notational convenience we now let

(60) $B_\rho = I_n - \rho W^* = I_n - \rho b(1_n 1_n' - I_n)$

[where $b = 1/(n-1)$ is one possibility] then by the first line of (10), it follows that for the SL model (1),

$$(61)\quad \hat{\sigma}_{sl}^{\,2}(\rho) = (1/n)\left([I_n - X(X'X)^{-1}X']B_\rho y\right)'\left([I_n - X(X'X)^{-1}X']B_\rho y\right)$$

$$= (1/n)\left\|[I_n - X(X'X)^{-1}X']B_\rho y\right\|^2$$

$$= (1/n)\left\|(I_n - P_X)B_\rho y\right\|^2$$

To compare this with the SAR model (49), observe from (51) that

$$(62)\quad \hat{\beta}_{sar}(\rho) = (X'B_\rho'B_\rho X)^{-1}X'B_\rho'B_\rho y = [(B_\rho X_\rho)'(B_\rho X)]^{-1}(B_\rho X)'B_\rho y$$

and hence from (52) that

$$(63)\quad \hat{\sigma}_{sar}^2(\rho) = (1/n)(y - X\hat{\beta}_{sar}(\rho))'B_\rho'B_\rho(y - X\hat{\beta}_{sar}(\rho)) = (1/n)\left\|B_\rho(y - X\hat{\beta}_{sar}(\rho))\right\|^2$$

$$= (1/n)\left\|B_\rho(y - X[(B_\rho X_\rho)'(B_\rho X)]^{-1}(B_\rho X)'B_\rho y)\right\|^2$$

$$= (1/n)\left\|\{I_n - (B_\rho X)[(B_\rho X)'(B_\rho X)]^{-1}(B_\rho X)'\}B_\rho y)\right\|^2$$

$$= (1/n)\left\|(I_n - P_{B_\rho X})B_\rho y\right\|^2$$

In this form it is clear that the result will follow if it can be shown that

$$(64)\quad P_X = P_{B_\rho X} \quad \textit{for all } \rho \in [W^*]$$

But since $X = [1_n, \tilde{X}]$ and $P_X X = X$ together imply that $P_X 1_n = 1_n$, we must have

$$(65)\quad P_X(B_\rho X) = bP_X(1_n 1_n' - I_n)X = b(P_X 1_n)1_n' X - bP_X X$$

$$= b(1_n 1_n')X - bX = b(1_n 1_n' - I_n)X = B_\rho X$$

and may conclude from Lemma 4 that (64) holds for all $\rho \in [W^*]$.

Hence it follows at once from Proposition 4 that for maximally connected weight matrices, W^*, it will always be true that the maximum likelihood estimates, $\hat{\rho}_n$, of ρ in corresponding SL and SAR models are *identical*. This in turn implies that Proposition 1 must hold in tact if the SL model in (1) is replaced by SAR model in (49). Hence the

same type of continuity argument in Proposition 2 can be used to show that the spatial dependence parameter, ρ, in SAR models will be underestimated for strongly connected weight matrices.

Rather than repeat such arguments here, we simply report the corresponding estimation results for the SAR model based on the same data X, parameters (β, σ^2, ρ), and weight matrices, W_d, $d \in \{.30,.50,.80,.90,.95,.99,1.00\}$ used in Section 3 above. The results for ρ are displayed in column 3 of Table 1 in that section and show that, as predicted by Proposition 4, these estimates converge to the same extreme value as d approaches unity. However it is also clear that (at least in this particular example) the underestimation of ρ is even more severe than for the SL model above.

Table 3 Here

The results for other parameter estimates are shown in Table 3 above. Notice first that all mean beta estimates appear to be remarkably accurate – even in the maximally connected case. This is explained by the well known fact that for the SAR model, $\hat{\beta}$ is *always* an unbiased estimator of β for a correctly specified model, since

$$(66) \quad E(\hat{\beta} \mid X) = [X'(I_n - \rho W)'(I_n - \rho W)X]^{-1} X'(I_n - \rho W)'(I_n - \rho W)E(y \mid X)$$

$$= [X'(I_n - \rho W)'(I_n - \rho W)X]^{-1} X'(I_n - \rho W)'(I_n - \rho W)X\beta = \beta$$

Notice also that there is some slight underestimation of residual variance, as in the case of SL models. This does not appear to be too severe (in the present example). But again, even slight underestimation of variance can lead to erroneous significant of beta parameters. Moreover, in the extreme case of maximal connectivity these estimates are in fact completely unstable, as can be seen by the dependency of $\hat{\beta}$ on $\hat{\rho}$ in the conditional beta estimator of (62) above. If we set

$$(67) \quad \hat{\rho}_n = 1/\lambda_{\min}(W^*) = -1/b$$

in this extreme case, then

$$(68) \quad B_{\hat{\rho}_n} = I_n - (-1/b)[b(1_n 1_n' - I_n)] = 1_n 1_n'$$

together with $1_n' 1_n = n$ implies that

$$(69) \quad \hat{\beta}_{sar}(\hat{\rho}_n) = (X'B_{\hat{\rho}_n}' B_{\hat{\rho}_n} X)^{-1} X'B_{\hat{\rho}_n}' B_{\hat{\rho}_n} y = (X'1_n 1_n' X)^{-1} X'1_n 1_n' y$$

Hence if there is at least one explanatory variable other than the intercept (i.e., if $k \geq 1$) then the matrix, $X'1_n 1_n' X$, is *singular* and the inverse in (69) will not exist. In practice

however, what typically happens is that estimation algorithms converge to values close to $-1/b$ which will yield well-defined answers. In the case illustrated above, where $-1/(1/49) = -49$, even values of -48.999 continue to produce reasonable looking estimates on average.

6. Consequences for Moran Tests of Spatial Autocorrelation

Aside from the above consequences for spatial regression models such as SL and SAR models, strong connectivity of weight matrices has broader implications for diagnostic analyses of spatial autocorrelation. This is most evident with respect to the single most widely used test for spatial autocorrelation, namely the *Moran Test*. In particular, suppose that one considers the null hypothesis of independence ($\rho = 0$), under which both SL and SAR models reduce to the standard linear model:

(70) $y = X\beta + \varepsilon \ , \ \varepsilon \sim N(0, \sigma^2 I_n)$

If one constructs the standard maximum-likelihood (OLS) estimates of β under this hypothesis,

(71) $\hat{\beta} = (X'X)^{-1} X' y$

and forms the corresponding vector of *residual estimates*:

(72) $\hat{\varepsilon} = y - \hat{y} = y - X\hat{\beta}$

then for any given candidate choice of a spatial weight matrix, W, the associated *Moran statistic*, I_W, is defined by [see for example Anselin (1988, Section 8.1.1)]:

(73) $I_W = \alpha_W \dfrac{\hat{\varepsilon}' W \hat{\varepsilon}}{\hat{\varepsilon}'\hat{\varepsilon}}$

where the positive constant, $\alpha_W = n / \|W\|_1 = n / \sum_{ij} w_{ij}$, plays no substantive role in the analysis to follow. This can be expressed in a more convenient form (again following Anselin) by noting from (71) and (72) that

(74) $\hat{\varepsilon} = y - X(X'X)^{-1} X'y = [I_n - X(X'X)^{-1} X']y = My$

and hence from (9) that I_W can be equivalently written as

(75) $I_W = \alpha_W \dfrac{y'MWMy}{y'My}$

Under the hypothesis of independence in (70), the mean and variance of I_W are well known to be [Anselin (1988, Section 8.1.1)]:

(76) $$E(I_W) = \frac{\alpha_W tr(MW)}{n-(k+1)}$$

(77) $$\text{var}(I_W) = \frac{(\alpha_W)^2\{tr(MWMW') + tr(MWMW) + [tr(MW)]^2\}}{[n-(k+1)]\cdot[n-(k-1)]} - [E(I_W)]^2$$

In this setting, our main result is to show that for *maximally connected* weight matrices, W^*, this Moran statistic is *degenerate*.[25] In particular it is *completely concentrated at the mean*, $E(I_{W^*})$, and hence can never detect spatial autocorrelation. To establish this result, we first note from (75) that this statistic is only meaningful for data sets (y, X) with $y'My \neq 0$. But since, $y'My = 0 \Leftrightarrow My = 0$ (as shown in the proof of Lemma 3 above), this is in turn equivalent to the condition that $My \neq 0$. Hence, for purposes of this section we again assume *regularity* of (y, X). In addition, we employ the normalization convention, $b = 1/(n-1)$, for W^* so that $\lambda_{\max}(W^*) = 1$. Finally, for each regular data set, (y, X), we let $I_{W^*}(y, X)$ denote the corresponding sample value of I_{W^*} in (75). With these conventions, we have the following result:

Proposition 5. *For all regular data sets* (y, X),

(78) $$I_{W^*}(y, X) = E(I_{W^*})$$

Proof: First observe that since $b = 1/(n-1) \Rightarrow \left\|W^*\right\|_1 = n(n-1)[1/(n-1)] = n$, it follows that

(79) $$\alpha_{W^*} = n \Big/ \left\|W^*\right\|_1 = n/n = 1$$

and hence that the Moran statistic for this case reduces to

(80) $$I_{W^*}(y, X) = \frac{y'MW^*y}{y'My}$$

Thus we see from Lemma 3 that

(81) $$I_{W^*}(y, X) = \frac{y'(-bM)y}{y'My} = -\frac{1}{n-1}\cdot\frac{y'My}{y'My} = -\frac{1}{n-1}$$

[25] This degeneracy is also an instance of the more general result in Arnold (1979, Th.5) for the class of invariant test statistics for linear models with exchangeably distributed errors. A more explicit version relating to the present case is given in Martellosio (2008, Props. 3.4 and 3.6).

and may conclude that I_{W^*} is indeed concentrated at a single value. To show that this value is precisely the mean, $E(I_{W^*})$, under independence, we first note that since the trace of orthogonal projection (symmetric idempotent) matrix, M, is equal to the dimension of its image space [Searle (1982, Section 12.2)], and since the dimension of the complement of the span of X is $n-(k+1)$, it follows that

(82) $$tr(M) = n-(k+1)$$

This in turn implies from Lemma 3 that

(83) $$tr(MW^*) = tr(-bM) = \left(-\frac{1}{n-1}\right)tr(M) = -\frac{n-(k+1)}{n-1}$$

and hence from (76) and (79) that

(84) $$E(I_{W^*}) = \frac{tr(MW^*)}{n-(k+1)} = -\frac{1}{n-1}$$

Thus the result follows from (81) and (84).

As a consequence of Proposition 5, it follows that (with probability one)[26] the realized value of I_{W^*} is precisely its *expected value* under independence. Hence no evidence for spatial dependence can ever be detected in this extreme case. More generally, the same type of continuity argument used in Proposition 3 above shows that for weight matrices, W, that are sufficiently close to W^* [say in terms of the relative connectivity norm] it must be true that the possible values of I_W are concentrated close to the mean $E(I_W)$. So again this statistic should have little ability to detect spatial dependence.

To make these ideas more concrete, we choose to focus on the standard z-test for Moran statistics found in most software. If the standard deviation of I_W under independence is denoted by $\sigma(I_W) = \text{var}(I_W)^{1/2}$, then it is well known [Cliff and Ord(1981, Section 8.5.1)] that the standardized z-value

(85) $$Z_W = \frac{I_W - E(I_W)}{\sigma(I_W)}$$

is approximately distributed $N(0,1)$ for large n. Hence one can use this distribution theory to test the hypothesis of spatial independence with respect to weight matrix W.[27]

[26] It is a simple matter to show for any X, the set of y with $My = 0$ has probability measure zero.

[27] It is worth noting here that the *exact* distribution of I_W under independence has been obtained by Tiefelsdorf and Boots (1995). However, most statistical packages rely on the asymptotic approximation above.

To study the behavior of this test for strongly connected weight matrices, we shall focus only on the simulation results in Section 3 above based on the SL model. Here it was assumed that $\rho = .5$ and hence that a substantial degree of positive spatial dependence is present. To determine whether this dependence can be detected by the Moran statistic for a given weight matrix, W, it suffices to compute $I_W(y, X)$ for simulated data sets from model (1), and then examine the frequency distribution of z-values, $Z_W(y, X)$, generated by this data. For a one-sided test of $\rho > 0$ at the $\alpha = .05$, one need only count the fraction of z-values above $z_\alpha = 1.65$ to determine the power of this test to detect positive spatial dependence, given the true value $\rho = .5$. For the 1000 simulated values at each link density level in Section 3 above, the resulting estimated power levels are shown in Table 4 below.

Table 4 Here

Here it is clear that at link densities above .80 the distribution is so concentrated around the null mean, $E(I_W)$, that even a dependency level of $\rho = .5$ is detectable less than 10% of the time.[28] It is also of interest to note that even though the distribution of I_W concentrates at the null mean as link density approaches 1, the power levels do not appear to fall to zero in Table 4. The reason for this is that concentration of I_W values drives the *variance* in (76) to zero (as can easily be verified by the same calculations as for the mean in the proof of Proposition 5[29]). Hence when I_W is highly concentrated, the standardized value, Z_W, becomes unstable (as it approaches the limiting indeterminate values 0/0 for W^*).

7. Concluding Remarks

In this paper it has been shown that presence of strongly connected spatial weight matrices can introduce serious biases into both the estimation and testing of spatial autocorrelation. Hence one is led to ask whether there is any simple intuitive explanation for this. One possibility relates to the notion of "effective sample size". It has long been observed that the presence of statistical dependencies essentially reduce the amount of information gained from each individual observation. For example, the observation of a sequence of perfectly correlated coin tosses will offer no more information than the observation of only the first toss, no matter how long the sequence is. Hence it can be argued that in so far as strong spatial connectivity reflects strong dependencies among

[28] It is also of interest to note that the .054 value for density .99 is consistent with a limiting value of $\alpha = .05$ for the maximally connected case, as implied by results of Martellosio (2008, Prop.3.5).

[29] Note in particular from Lemma 2 that for W^*, $tr(MW^*MW^*) = tr[(-bM)(-bM)] = b^2 tr(MM)$ $= b^2 tr(M) = [n-(k+1)]/(n-1)^2$.

units (or agents), there should be less statistical information available for estimation or tests of hypotheses.

But while this argument has intuitive appeal, and is no doubt true to some extent, it fails to explain, for example, why maximum-likelihood methods should systematically *underestimate* the ρ parameter in SL and SAR models. In this paper it has been shown that much can be learned by studying the extreme case of maximally connected weight matrices, W^*. In particular, both concentrated likelihood functions and Moran statistics reduce to particularly simple forms in this case, and can be studied in detail. But even in this extreme case, the subtlety of the underestimation question above is underscored by the fact that quite different arguments were used to bound the values for each term in the concentrated log likelihood function. In particular, both the eigenvalue structure of W^* and the relation of W^* to the regression projection operator, $I_n - X(X'X)^{-1}X'$, were involved. So in some respects, these results serve to raise as many theoretical questions as they answer.

Even more important are questions relating to the practical consequences of these results. While the single simulation example presented here is very suggestive, it can provide no definitive guidelines for applications. Hence the actual severity of these biases can only be determined by more extensive and systematic simulation studies, as already begun by Mizruchi and Neuman (2008) and Farber, Páez and Volz (2008).

References.

Anselin, L. (1988) *Spatial Econometrics*: *Methods and Models*, Kluwer: Netherlands.

Anselin, L. and A. Bera (1998), "Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics." In A. Ullah and D. Giles (eds.), *Handbook of Applied Economic Statistics*, New York: Marcel Dekker, pp. 237–289.

Arnold, S.F. (1979) "Linear models with exchangeably distributed errors", *Journal of the American Statistical Association*, 74: 194-199.

Baltagi, B.H., (2006) "Random effects and spatial autocorrelation with equal weights", *Econometric Theory*, 22: 973-984.

Bao, Y. and A. Ullah (2007) "Finite sample properties of maximum likelihood estimator in spatial models", *Journal of Econometrics*, 137: 396-413.

Farber, S., A. Páez and E. Volz (2008) "Topology and dependency tests in spatial and network autoregressive models", forthcoming in *Geographical Analysis* (currently available at *http://www.science.mcmaster.ca/geo/faculty/paez/publications.html*).

Horn, R.A. and C.R. Johnson, (1985) *Matrix Analysis*, Cambridge University Press: Cambridge.

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

Kelejian, H.H. and I.R. Prucha (2002) "2SLS and OLS in a spatial autoregressive model with equal spatial weights", *Regional Science and Urban Economics*, 32: 601-707.

Kelejian, H.H., I.R. Prucha and Y. Yuzefovich, (2006) "Estimation problems in models with spatial weighting matrices which have blocks of equal elements", *Journal of Regional Science*, 46: 507-515.

Kelejian, H.H. and I.R. Prucha (2008) "Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances", forthcoming in the *Journal of Econometrics*.

LeSage, J. (1999) *Spatial Econometrics Toolbox, http://www.spatial-econometrics.com*.

Martellosio, F. (2008) "Testing for spatial autocorrelation: the regressors that make the power disappear", *Working Paper*, Department of Economics, University of Reading, Reading RG6 6AW, UK.

Mizruchi, M.S. and E.J Neuman (2008) "The effect of density on the levels of bias in the network autocorrelation model", forthcoming in *Social Networks* (currently available at *http://www-personal.umich.edu/~mizruchi/dens6.pdf*.)

Ord, K. (1975) "Estimation methods for models of spatial interaction", *Journal of the American Statistical Association*, 70: 120-126.

Searle, S.R., (1982) *Matrix Algebra Useful for Statistics*, Wiley: New York.

Tiefelsdorf, M. and B.N. Boots (1995) "The exact distribution of Moran's I", *Environment and Planning A*, 27: 985-999.

Acknowledgements:

The author is grateful to Mark Mizruchi, Eric Neuman, Oleg Smirnov, Harry Kelejian and Federico Martellosio for many helpful comments on an earlier draft of this paper.

APPENDIX: Proof of Proposition 3

As stated in the proof sketch for Proposition 3, our strategy will be to establish *negativity* of the concentrated likelihood derivative with respect to ρ on a given interval in $[W]$. For purposes of analysis, we now drop the SL model subscript and write the concentrated likelihood function as $L_W(\rho \mid y, X)$, so that by (16),[30] and the positivity of $\min_i\{1-\rho\lambda_i(W)\}$ on $[W]$, it follows that for all $\rho \in [W]$,

$$\text{(A.1)} \quad L_W(\rho \mid y, X) = const + \sum_{i=1}^{n} \ln(1-\rho\lambda_i) - (n/2)\ln[y'(I_n - \rho W)'M(I_n - \rho W)y]$$

Since we are only interested in nonnegative feasible ρ-values for each weight matrix, W, it is convenient to define this interval as

$$\text{(A.2)} \quad [W]_+ = \{\rho \in [W] : \rho \geq 0\} = [0, 1/\lambda_{\max}(W))$$

In these terms, the relevant domain, D, for the function, $L_W(\rho \mid y, X)$, with arguments (ρ, W) is given by

$$\text{(A.3)} \quad D = \{(\rho, W) \in R \times R^{n\times n} : \rho \in [W]_+, W \in \mathbf{W}_n^+\}$$

[where the data (y, X) in (A.1) is taken to be fixed]. Our interest then focuses on the partial derivative of this function with respect to ρ, which we write as

$$\text{(A.4)} \quad L'_W(\rho \mid y, X) = -\sum_{i=1}^{n} \frac{\lambda_i}{(1-\rho\lambda_i)} - (n/2)\frac{(\partial/\partial\rho)[y'(I_n - \rho W)'M(I_n - \rho W)y]}{y'(I_n - \rho W)'M(I_n - \rho W)y}$$

To analyze this function, observe first that since,

$$\text{(A.5)} \quad \tfrac{\partial}{\partial\rho}[y'(I_n - \rho W)'M(I_n - \rho W)y] = \tfrac{\partial}{\partial\rho}[y'My - 2\rho y'MWy + \rho^2 y'W'MWy]$$

$$= -2y'MWy + 2\rho y'W'MWy$$

$$= 2[\rho y'W'MWy - y'MWy]$$

and since (9) together with the symmetry of M implies both[31]

$$\text{(A.6)} \quad y'W'MWy = y'W'MMWy = (MWy)'MWy = \|MWy\|^2 \geq 0 \quad \text{, and}$$

[30] Expressions such as (16) refer to the text, in contrast to Appendix expressions such as (A.16).

[31] Here the vector norm $\|x\|$ is simply Euclidean norm, which is of course identical to the matrix norm $\|A\|_2$ applied to $n \times 1$ matrices.

(A.7) $y'(I_n - \rho W)'M(I_n - \rho W)y = y'(I_n - \rho W)'MM(I_n - \rho W)y = \left\|M(I_n - \rho W)y\right\|^2 \geq 0$

it follows that (A.4) reduces to

(A.8) $L'_w(\rho \mid y, X) = (n)\dfrac{y'MWy - \rho \cdot \left\|MWy\right\|^2}{\left\|M(I_n - \rho W)y\right\|^2} - \sum_{i=1}^{n} \dfrac{\lambda_i(W)}{[1 - \rho\lambda_i(W)]}$

where the notation, $\lambda_i(W)$, now reflects the dependence of these eigenvalues on W. We are interested in the continuity properties of this function with respect to weight matrices, W, in the neighborhood of W^*. But, as with the concentrated likelihood function itself, it is clear from (A.8) that for any W, the function, $L'_W(\cdot \mid y, X)$, is only well defined if $\left\|M(I_n - \rho W)y\right\| > 0$ for all $\rho \in [W]_+$. In Lemma 3 it was shown that every regular data set (y, X) is W^*-regular, and hence that positivity is ensured for the case $W = W^*$. Thus part of the continuity argument to follow will be to show that this positivity property is necessarily inherited by all W sufficiently close to W^*, and moreover that for regular data sets, (y, X), the values $\left\|M(I_n - \rho W)y\right\|$ are uniformly bounded away from zero on some neighborhood of W^*. Thus it may be assumed for the moment that $L'_W(\cdot \mid y, X)$ is well defined. To establish W-continuity of this function at W^*,[32] it is convenient to consider each term separately by letting

(A.9) $U_W(\rho \mid y, X) = (n)\dfrac{y'MWy - \rho \cdot \left\|MWy\right\|^2}{\left\|M(I_n - \rho W)y\right\|^2}$

and

(A.10) $V_W(\rho \mid y, X) = -\sum_{i=1}^{n} \dfrac{\lambda_i(W)}{[1 - \rho\lambda_i(W)]}$

be the associated component functions defined on D. Hence W-continuity of L_W at W^* will be implied by W-continuity of both U_W and V_W. at W^*. But for V_W in particular, W-continuity follows from well known results. Since $\rho \in [W]_+ \subset [W]$ for all $W \in \mathbf{W}_n^+$, it follows that $V_W(\rho \mid y, W)$ is always well defined on D, and hence that W-continuity of this function is guaranteed by continuity of each eigenvalue function, $\lambda_i(W)$. But this follows at once from the results of Horn and Johnson (1985, Appendix D), where it is shown that the roots of polynomials are continuous functions of their coefficients (in this

[32] By W-continuity of a function $f_w(\rho)$ *at* W we mean that for any sequence (W_m) in $\mathbf{W}_n^+$, $\lim_{m\to\infty} W_m = W \Rightarrow \lim_{m\to\infty} f(\rho \mid W_m) = f(\rho \mid W)$ for each $\rho \in [W]$. The function $f_w(\rho)$ is then said to be W-continuous if W-continuity holds at every point, $W \in \mathbf{W}_n^+$.

case the components, w_{ij}, of W).[33] However, as mentioned above, W-continuity of $U_W(\rho \mid y, X)$ is complicated by the need to ensure that $\|M(I_n - \rho W)y\|$ remains positive. Hence it is convenient to focus on the component function

(A.11) $g(W \mid y, X) = y'MWy$

in the numerator of $U_W(\rho \mid y, X)$, with domain $\mathbf{W}_n^+$. The reason for this choice can be seen by expanding the denominator of $U_W(\rho \mid y, X)$ as:

$$\text{(A.12)} \quad \|M(I_n - \rho W)y\|^2 = y'(M - \rho MW)'(M - \rho MW)y$$

$$= y'My + 2\rho y'MWy + \rho^2 (MWy)'(MWy)$$

$$= \|My\|^2 - 2\rho g(W \mid y, X) + \rho^2 \|MWy\|^2$$

Since the first term is positive for regular data (y, X) and since the last term is always nonnegative, it follows that if $g(W \mid y, X) \le 0$, then the second term will be nonnegative for all $\rho \in [W]_+$ and hence that

$$\text{(A.13)} \quad \|M(I_n - \rho W)y\|^2 \ge \|My\|^2 > 0$$

Thus nonpositivity of $g(W \mid y, X)$ will guarantee that $\|M(I_n - \rho W)y\|$ is *uniformly bounded away from zero* for $\rho \in [W]_+$, and hence that $U_W(\rho \mid y, X)$ is well defined. Moreover, if it can be established that $g(W \mid y, X)$ is actually *negative*, then by (A.9) it will also follow that $U_W(\rho \mid y, X)$ is negative for all $\rho \in [W]_+$ – which will help to establish not only W-continuity W^*, but also the desired negativity of $L'_W(\cdot \mid y, X)$ on appropriate intervals. To specify these intervals, we now fix some "true" spatial dependence value, $\rho_0 \in (0,1)$, in Model (1), and for each bias factor, $\alpha > 0$, let

$$\text{(A.14)} \quad \rho_\alpha = \rho_0 / (1 + \alpha)$$

denote the bound on underestimation in expression (47) of Proposition 3. Then our first objective is to establish conditions on W that will ensure negativity of $g_W(y, X)$ for all $\rho \in [W]_+$ with $\rho \ge \rho_\alpha$.

To do so, recall first from Lemma 2 [and the scaling $b = 1/(n-1)$] that

[33] At this point it is also worth noting that in both establishing continuity and drawing inferences based on continuity, it will make no difference which norm in being used. Indeed, since every n-square matrix norm is a vector norm on $R^{n \times n}$, and since *all* vector norms on a finite-dimensional space are topologically equivalent [Horn and Johnson (1985, Corollary 5.4.6)], questions of convergence are independent of the choice of norms. Only the "sizes" of neighborhoods will differ.

(A.15) $g_{W^*}(y,X) = y'MW^*y = -by'My = -\dfrac{y'My}{n-1}$

which is always negative for regular data (y,X). Hence noting that for any $W \in \mathbf{W}_n^+$,

(A.16) $g_W(y,X) \le g_{W^*}(y,X) + |\, g_W(y,X) - g_{W^*}(y,X)\,|$

$$= -\frac{y'My}{n-1} + |\, y'MWy - y'MW^*y\,|$$

it follows that $g_W(y,X)$ will also be negative if the second term is sufficiently small. But by the Cauchy-Schwarz inequality

(A.17) $|\, y'MWy - y'MW^*y\,| = |\, y'M(W-W^*)y\,| \le \|y\| \cdot \|M(W-W^*)y\|$

Moreover, by the submultiplicative property of the matrix norm $\|A\|_2$,[34]

(A.18) $\|M(W-W^*)y\| \le \|M(W-W^*)\|_2 \cdot \|y\| \le \|M\|_2 \cdot \|W-W^*\|_2 \cdot \|y\|$

and by the norm inequality, $\|A\|_2 \le \|A\|_1$ [Horn and Johnson (1985, p.315)] and the definition of the relative connectivity norm, $\|W\|_{rc}$, in expression (40),

(A.19) $\|W-W^*\|_2 \le \|W-W^*\|_1 = \|W^*\|_1 \cdot \|W-W^*\|_{rc}$

Hence combining (A.17) through (A.19) together with the fact that

(A.20) $\|W^*\|_1 = \sum_{ij} w_{ij}^* = n(n-1)/(n-1) = n$

we see that

(A.21) $|\, y'MWy - y'MW^*y\,| \le n\|y\|^2 \|M\|_2 \|W-W^*\|_{rc}$

Hence for any $\varepsilon > 0$

(A.22) $\|W-W^*\|_{rc} < \varepsilon \;\Rightarrow\; |\, y'MWy - y'MW^*y\,| < \left(n\|y\|^2 \|M\|_2\right)\varepsilon$

[34] See Horn and Johnson (1985, section 5.6). Recall also from footnote 21 that by convention, $\|y\| = \|y\|_2$.

which implies at once that $g_W(y,X)$ is W-continuous at W^*. More important is the fact that for sufficiently small $\varepsilon > 0$, $g_W(y,X)$ will be negative. In particular, from (A.16) and (A.22) it suffices to require that

(A.23) $\left(n\|y\|^2 \|M\|_2\right)\varepsilon - \dfrac{y'My}{n-1} < 0$

Hence if we now let

(A.24) $\varepsilon_1 = \varepsilon_1(y,X) = \dfrac{y'My}{\|y\|^2 \cdot \|M\|_2\, n(n-1)}$

then it must follow that

(A.25) $\|W - W^*\|_{rc} < \varepsilon_1 \;\Rightarrow\; g_W(y,X) < 0$

Thus $g_W(y,X)$ is *negative* for W closer to W^* than ε_1. This in turn implies that (A.13) holds, and thus ensures that $U_W(\rho \mid y,X)$ is well defined for all $\rho \in [W]_+$. More importantly, it then follows from (A.9) that

(A.26) $\|W - W^*\|_{rc} < \varepsilon_1 \;\Rightarrow\; U_W(\rho,y,X) < 0$ *for all* $\rho \in [W]_+$.

Hence it only remains to be shown that if W sufficiently close to W^* then $V_W(\rho \mid y,X)$ is also nonpositive for all $\rho \geq \rho_\alpha$. To do so we again start in a manner similar to (A.16) by looking at perturbations of $V_W(\rho \mid y,X)$ around $V_{W^*}(\rho \mid y,X)$. To do so, observe first from the continuity of each eigenvalue function, $\lambda_i(W)$, that for any $e > 0$ there is some $\varepsilon_2(e) > 0$ such that

(A.27) $\|W - W^*\|_{rc} < \varepsilon_2(e) \;\Rightarrow\; |\lambda_i(W) - \lambda_i(W^*)| < e \;,\; i = 1,..,n$

$$\Rightarrow\; \lambda_i(W) = \lambda_i(W^*) + e_i \;,\; |e_i| < e \;,\; i = 1,..,n$$

Notice moreover that without loss of generality we may assume that $\varepsilon_2(\cdot)$ is an increasing function [so that smaller values of e are associated with smaller values of $\varepsilon_2(e)$]. Condition (A.27) in turn implies from Lemma 1 together with $b = 1/(n-1)$ that for $\|W - W^*\|_{rc} < \varepsilon_2(e)$ we must have,

(A.28) $V_W(\rho \mid y,X) = -\sum_{i=1}^{n} \dfrac{\lambda_i(W^*) + e_i}{1 - \rho\left(\lambda_i(W^*) + e_i\right)}$

$$= -\sum_{i=1}^{n-1} \frac{[-1/(n-1)]+e_i}{1-\rho\left([-1/(n-1)]+e_i\right)} - \frac{1+e_n}{1-\rho(1+e_n)}$$

$$= \sum_{i=1}^{n-1} \frac{1-(n-1)e_i}{(n-1+\rho)-(n-1)\rho e_i} - \frac{1+e_n}{(1-\rho)-\rho e_n}$$

Note that when $W = W^*$ so that $e_i \equiv e = 0$ this reduces to the expression, $(n-1)/(n-1+\rho) - 1/(1-\rho)$, which is clearly negative for $\rho \in [W^*]_+$ and $n \geq 3$. So negativity of $V_W(\rho \mid y, X)$ should continue to hold for e sufficiently small. To determine an explicit bound on e, observe first that each term of the summation in (A.28) can be written as a function of the form $f(x) = (1-ax)/(b-cx)$. But one can readily verify by differentiating $f(x)$ that

(A.29) $f'(x) = (c-ab)/(b-cx)^2$

By evaluating the numerator of (A.29) in terms of (A.28) we have

(A.30) $c - ab = (n-1)\rho - (n-1)(n-1+\rho) = -(n-1)^2 < 0$

so that each term $i = 1,..,n-1$ is seen to be decreasing in e_i. Hence we can bound these expressions above in terms of e by observing that

(A.31) $-e \leq e_i \Rightarrow \dfrac{1-(n-1)e_i}{(n-1+\rho)-(n-1)\rho e_i} \leq \dfrac{1-(n-1)(-e)}{(n-1+\rho)-(n-1)\rho(-e)}$

$$= \frac{1+(n-1)e}{(n-1+\rho)+(n-1)\rho e}$$

Similarly, the last term in (A.28) [ignoring the minus sign in front] can also be written in the same form, where in this case, $a = -1, b = 1-\rho, c = \rho$ imply that

(A.32) $c - ab = \rho - (-1)(1-\rho) = 1 > 0$

so that the last term is increasing in e_n, and can be bounded below in terms of e by observing that

(A.33) $-e \leq e_n \Rightarrow \dfrac{1+e_n}{(1-\rho)-\rho e_n} \geq \dfrac{1+(-e)}{(1-\rho)-\rho(-e)} = \dfrac{1-e}{(1-\rho)+\rho e}$

By applying these inequalities to (A.28), we now have

$$\text{(A.34)}\quad V_W(\rho \mid y, X) \le \sum_{i=1}^{n-1} \frac{1+(n-1)e}{(n-1+\rho)+(n-1)\rho e} - \frac{1-e}{(1-\rho)+\rho e}$$

$$= (n-1)\frac{1+(n-1)e}{(n-1+\rho)+(n-1)\rho e} - \frac{1-e}{(1-\rho)+\rho e}$$

Hence it remains to find a value of e sufficient small to ensure the right hand side is negative. To do so, observe that [35]

$$\text{(A.35)}\quad (n-1)\frac{1+(n-1)e}{(n-1+\rho)+(n-1)\rho e} - \frac{1-e}{(1-\rho)+\rho e} < 0$$

$$\Leftrightarrow\ (n-1)\frac{1+(n-1)e}{(n-1+\rho)+(n-1)\rho e} < \frac{1-e}{(1-\rho)+\rho e}$$

$$\Leftrightarrow\ \{(n-1)(1+(n-1)e)\}\cdot((1-\rho)+\rho e) < \{(n-1+\rho)+(n-1)\rho e\}(1-e)$$

Notice that the bracketed factors on each side are by far the largest. Hence we focus on values of e that will yield the desired inequality for these terms:

$$\text{(A.36)}\quad (n-1)(1+(n-1)e) < (n-1+\rho)+(n-1)\rho e$$
$$\Leftrightarrow\ [(n-1)^2-(n-1)\rho]e < (n-1+\rho)-(n-1) = \rho$$

$$\Leftrightarrow\ e < \frac{\rho}{(n-1)(n-1-\rho)}$$

Similarly for the remaining two factors, it follows that

$$\text{(A.37)}\quad (1-\rho)+\rho e < 1-e \ \Leftrightarrow\ e < \frac{\rho}{1+\rho}$$

But since

$$\text{(A.38)}\quad \frac{\rho}{(n-1)(n-1-\rho)} < \frac{\rho}{1+\rho}$$

for $n \ge 3$, it follows that (A.35) will hold for e satisfying (A.36). Hence if we now set

---

[35] The last line of the argument in (A.35) assumes that $(1-\rho)+\rho e > 0$. But for $\rho \in [W]_+$ we must have $0 < 1-\rho\lambda_{\max}(W) = 1-\rho[\lambda_{\max}(W^*)+e_n] = 1-\rho(1+e_n)$. Also, by footnote 7 in the text we must have $W \le W^* \Rightarrow \lambda_{\max}(W) \le \lambda_{\max}(W^*) \Rightarrow e_n \le 0$. Hence, $0 < 1-\rho(1+e_n) = (1-\rho)-\rho e_n = (1-\rho)+\rho \mid e_n \mid \le (1-\rho)+\rho e$.

(A.39) $e(\rho) = \dfrac{\rho}{(n-1)(n-1-\rho)}$

then it follows from (A.27) that

(A.40) $\left\|W - W^*\right\|_{rc} < \varepsilon_2[e(\rho)] \;\Rightarrow\; V_W(\rho \mid y, X) < 0$

To complete the argument, observe that since $\varepsilon_2(\cdot)$ was chosen to be increasing and since $e(\cdot)$ is increasing for $\rho > 0$, it follows that for all $\rho \in [W]_+$,

(A.41) $\rho \geq \rho_\alpha \;\Rightarrow\; \varepsilon_2[e(\rho_\alpha)] \leq \varepsilon_2[e(\rho)]$

Hence if we now let $\varepsilon = \varepsilon(\alpha, \rho_0, y, X)$ in Proposition 3 be defined by

(A.42) $\varepsilon = \min\{\varepsilon_1, \varepsilon_2[e(\rho_\alpha)]\}$

then by (A.40) it follows that

(A.43) $\left\|W - W^*\right\|_{rc} < \varepsilon \;\Rightarrow\; V_W(\rho \mid y, X) < 0 \;\; \textit{for all } \rho \in [W]_+ \textit{ with } \rho \geq \rho_\alpha$

Finally, (A.43) together with (A.26) and (A.8) through (A.10) are seen to imply that

(A.44) $\left\|W - W^*\right\|_{rc} < \varepsilon \;\Rightarrow\; L'_W(\rho \mid y, X) < 0 \;\; \textit{for all } \rho \in [W]_+ \textit{ with } \rho \geq \rho_\alpha$

and the result is established.

Figures for the Text:

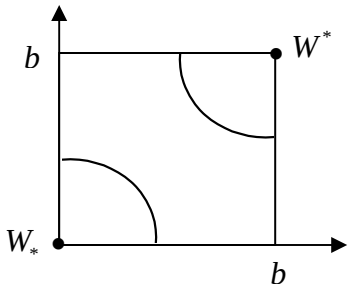

Figure 1. 2x2 Weight Matrices

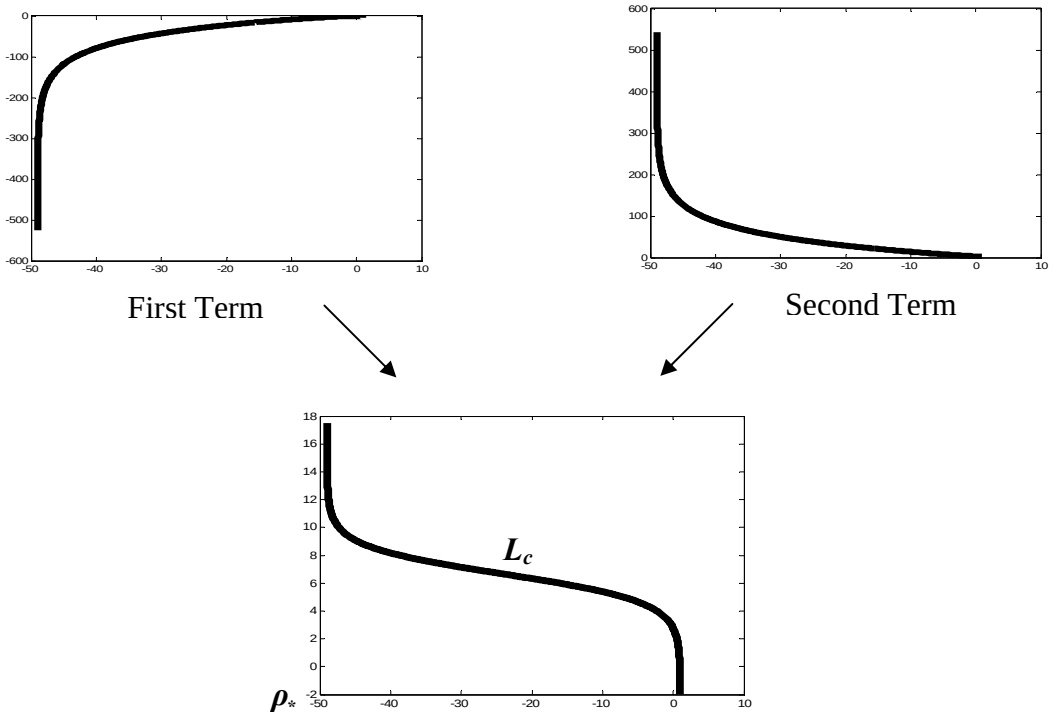

Figure 2. Concentrated Likelihood Function for W^*

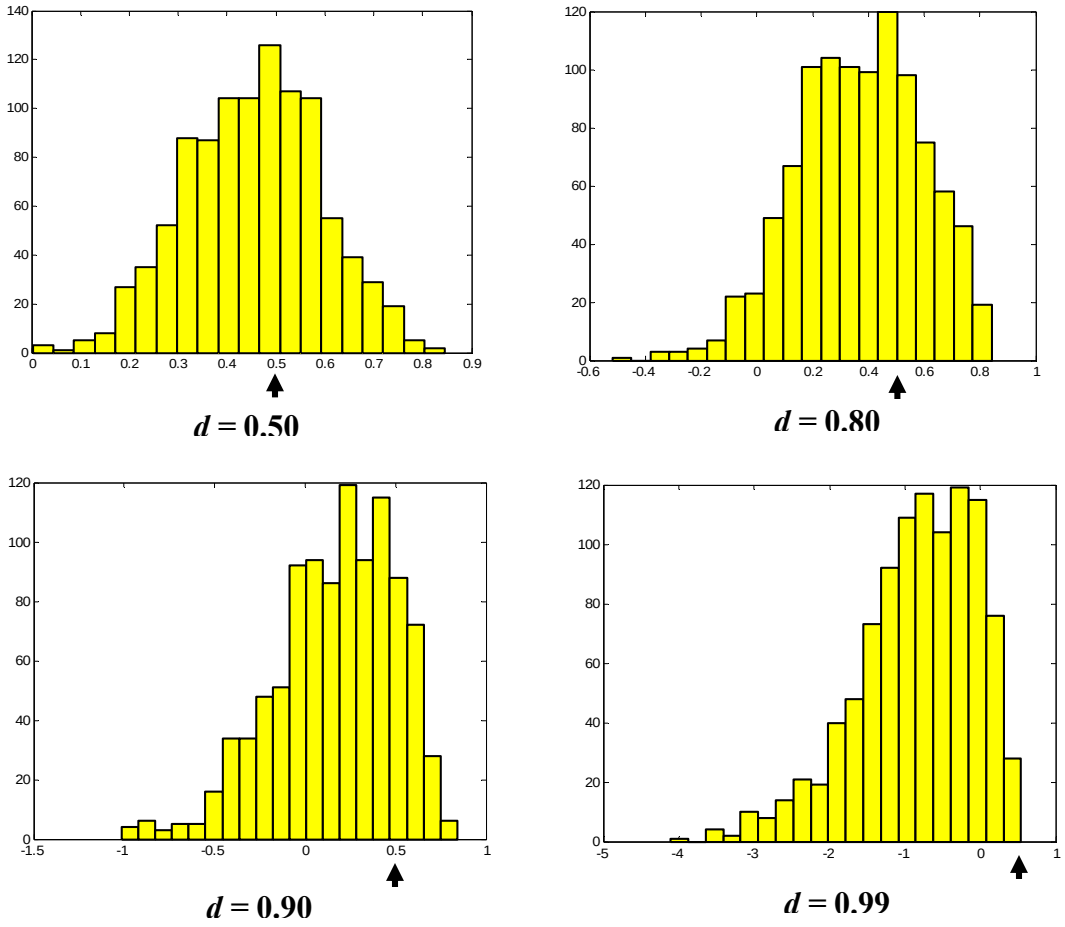

Figure 3. Histograms of Rho Estimates

Tables for the Text:

Av. Link Density	Mean $\hat{\rho}$ for SL models	Mean $\hat{\rho}$ for SAR models	ρ -Interval
.30	0.481	0.195	(-2.49 , 1)
.50	0.454	-0.038	(-3.51 , 1)
.80	0.369	-0.801	(-6.31 , 1)
.90	0.168	-1.880	(-9.02 , 1)
.95	0.033	-2.281	(- 10.7, 1)
.99	-0.830	-6.363	(- 18.9, 1)
1.00	-48.999	-48.999	(-49.0 , 1)

Table 1. Mean Estimates of Rho

Av. Link Density	Mean $\hat{\rho}$ (ρ = .5)	Mean $\hat{\beta}_0$ (β_0 = 1)	Mean $\hat{\beta}_1$ (β_1 = 2)	Mean $\hat{\beta}_2$ (β_2 = 3)	Mean $\hat{\sigma}^2$ (σ^2 = 1)
.30	0.481	1.138	1.978	3.003	0.92887
.50	0.454	1.336	1.946	2.9973	0.92644
.80	0.369	1.942	1.922	3.0143	0.91802
.90	0.168	3.384	1.944	2.9201	0.91908
.95	0.033	4.302	1.985	2.9209	0.90547
.99	-0.830	10.330	1.994	3.012	0.89564
1.00	-48.999	351.020	.00004	.00006	3.8e-010

Table 2. Mean Values of Parameter Estimates for the SL Model

Av. Link Density	Mean $\hat{\rho}$ (ρ = .5)	Mean $\hat{\beta}_0$ (β_0 = 1)	Mean $\hat{\beta}_1$ (β_1 = 2)	Mean $\hat{\beta}_2$ (β_2 = 3)	Mean $\hat{\sigma}^2$ (σ^2 = 1)
.30	0.195	1.064	2.010	2.939	0.937
.50	-0.038	1.040	1.960	2.958	0.933
.80	-0.801	0.956	2.039	2.039	0.904
.90	-1.880	0.997	2.011	3.006	0.864
.95	-2.281	0.998	1.994	3.037	0.823
.99	-6.363	1.047	1.945	3.032	0.706
1.00	-48.999	1.025	1.985	2.999	0.159

Table 3. Mean Values of Parameter Estimates for the SAR Model

Av. Link Density	Sample Mean	Null Mean	Power (ρ = .5)
.30	0.0416	-0.0182	0.383
.50	-0.0046	-0.0183	0.137
.80	-0.0152	-0.0184	0.091
.90	-0.0190	-0.0187	0.059
.95	-0.0187	-0.0189	0.055
.99	-0.0201	-0.0190	0.054

Table 4. Power of Moran for a Test at ρ = .5