Spatial Econometrics

Cem Ertur*

November 12, 2014
— Version 2.0 —

1 Spatial Autoregressive models of order 1

1.1 SAR specification : the mixed regressive spatial autoregressive model

The SAR specification adds the spatially lagged endogenous variable to the standard explanatory variables on the right hand side of the classical cross section regression model. For reasons of generality that will be discussed below, we allow the elements of W_N, X_N and ε_N to depend on the sample size N, that is, to form triangular arrays following (**?**, p. 53-67).[1] The model is then:

$$y_{i,N} = \beta_0 + \rho \sum_{j=1}^{N} w_{ij,N} y_{j,N} + \sum_{k=1}^{K} x_{ik,N} \beta_k + \varepsilon_{i,N} \quad \text{for } i = 1, ..., N \tag{1}$$

or in more compact matrix form:

$$y_N = \beta_0 \iota_N + \rho W_N y_N + X_N \beta + \varepsilon_N \tag{2}$$

where y_N is the $N \times 1$ vector of the dependent variable; ι_N is the $N \times 1$ unit vector; $W_N y_N$ is the $N \times 1$ vector called the spatially lagged endogenous variable; W_N is the $N \times N$ interaction matrix with non negative, non stochastic and finite weights such that $w_{ij,N} \geq 0$ for $i \neq j$ and $w_{ij,N} = 0$ for

*Laboratoire d'Economie d'Orléans (UMR 7322 CNRS), UFR Droit Economie Gestion, Université d'Orléans. E-mail : cem.ertur@univ-orleans.fr, web site : http://certur.free.fr/.

[1] Moreover we allow for the parameters to depend on the sample size as well, although we will not index them by the sample size N to keep notations as simple as possible.

$i = j$; ρ is the spatial autoregressive parameter measuring the magnitude of the interaction or spatial autocorrelation between neighboring spatial units; X_N is the $N \times K$ matrix of the K non-stochastic explanatory variables; β is the $K \times 1$ vector of the associated parameters and finally the disturbance $\varepsilon_{i,N}$ of the $N \times 1$ error vector ε_N are independently and identically distributed $(0,\ \sigma^2_{\varepsilon,N})$.

If the interaction matrix is normalized, as is often the case in the literature, to have row sums of unity, with weights $w^*_{ij,N} = w_{ij,N} / \sum_j w_{ij,N}$, then the i^{th} row of the vector $W^*_N y_N$, can be interpreted as the spatially weighted average of the neighboring values of i, i.e. $[W^*_N y_N]_i = \sum_j w^*_{ij,N} y_{j,N}$. Note that, in matrix form, $W^*_N = D_N W_N$, where $D_N = \text{diag}(1/\sum_j w_{ij,N})$ is a diagonal matrix containing the inverse of the row sums of W_N.

1.2 SDM specification: the Spatial Durbin model

A straightforward generalization of the SAR model consists to add the spatially lagged exogenous variables to the specification to obtain the so-called Spatial Durbin Model or SDM:

$$y_{i,N} = \beta_0 + \rho \sum_{j=1}^{N} w_{ij,N} y_{j,N} + \sum_{k=1}^{K} \beta_k x_{ik,N} + \sum_{k=1}^{K} \gamma_k \sum_{j=1}^{N} w_{ij,N} z_{ik,N} + \varepsilon_{i,N} \quad (3)$$

for $i = 1, ..., N$, or in more compact matrix form:

$$y_N = \beta_0 \iota_N + \rho W_N y_N + X_N \beta + W_N Z_N \gamma + \varepsilon_N \quad (4)$$

where y_N is the $N \times 1$ vector of the dependent variable, $W_N y_N$ is the spatially lagged dependent variable, ι_N is the $N \times 1$ unit vector and $X = [X_{1,N}, \ldots, X_{k,N}, \ldots, X_{K,N}]$ is the $N \times K$ matrix of the explanatory variables, with the $N \times 1$ vector $X_{k,N}$ representing the k^{th} explanatory variable; $\beta = (\beta_1, \ldots, \beta_k, \ldots, \beta_K)'$ is the $K \times 1$ vector of the associated coefficients, with $\beta_{k,N}$ the coefficient associated to the k^{th} explanatory variable. Note that Z_N may include explanatory variables not included in X_N, supposed affecting y_N only through their spatial lag: Z_N may then be a $N \times (K + M)$ matrix $Z_N = [X_N\ \widetilde{Z}_N]$ where X_N is the previous $N \times K$ matrix of the explanatory variables and $\widetilde{Z}_N$ is the $N \times M$ matrix of extra explanatory variables not included in X_N which are supposed to affect y_N only through their spatial lag. $W_N Z_N$ is the $(N \times (K + M))$ matrix of the spatially lagged explanatory variables and γ is the $(K + M) \times 1$ vector of the associated coefficients. For ease of exposition, let us write the SDM as a SAR model as

follows:

$$y_N = \rho W_N y_N + \widetilde{X}_N b + \varepsilon_N \tag{5}$$

where $\widetilde{X}_N = [\iota \ X_N \ W_N Z_N]$ and $b_N = [\beta_0, \beta', \gamma']'$ with Z_N being either X_N or $Z_N = [X_N \ \widetilde{Z}_N]$.

All the following results will be derived for the SAR model without any loss of generality, having in mind that any SDM model may be written as a SAR model with a generalized set of explanatory variables including their spatial lag. However some specific issues with regard to the estimation of the SDM model will be stressed when needed, also the spatial multipliers will be presented in details using the SDM model for more clarity.

1.3 Basic results

For simplicity and without loss of generality, consider the SAR model as defined in equation (**??**), with X_N including now the constant term:

$$y_N = \rho W_N y_N + X_N \beta + \varepsilon_N \tag{6}$$

1.3.1 Spatial filter

The spatial filter is defined as $(I_N - \rho W_N)$. Note that if ρ was known, we could filter out spatial autocorrelation from y_N and then use the classical regression model as follows:

$$(I_N - \rho W_N) y_N = X_N \beta + \varepsilon_N \tag{7}$$

Of course ρ is usually unknown and has to be estimated as well as β and $\sigma^2_{\varepsilon,N}$.

1.3.2 Reduced form

If $(I_N - \rho W_N)$ is invertible, the reduced form of the SAR specification is obtained as:

$$y_N = (I_N - \rho W_N)^{-1} X_N \beta + (I_N - \rho W_N)^{-1} \varepsilon_N \tag{8}$$

Therefore, we need to precisely define the invertibility condition for $(I_N - \rho W)$ which is needed to write the reduced form: $(I_N - \rho W_N)$ is invertible if $\det(I_N - \rho W_N) \neq 0$. Note first that if $\rho = 0$, $(I_N - \rho W_N)$ is non singular.

Now consider $\rho \neq 0$, we have:[2]

$$\det(I_N - \rho W_N) = \det[(-\rho)(W_N - \frac{1}{\rho} I_N)] \tag{9}$$
$$= (-\rho)^N \det(W_N - \frac{1}{\rho} I_N)$$

Therefore $\det(I_N - \rho W_N) \neq 0$ and $(I_N - \rho W_N)$ is non singular if $\rho^{-1} \notin \{\nu_{1,N}, ..., \nu_{N,N}\}$, where $\nu_{1,N}, ..., \nu_{N,N}$ denote the eigenvalues of W_N, i.e. if ρ^{-1} is not an eigenvalue of W_N. The parameter space for ρ will be specified and discussed in more details below.

Let us also note that the reduced form equation of y_N can also be represented as:

$$y_N = X_N\beta + \rho G_N X_N \beta + (I_N - \rho W_N)^{-1}\varepsilon_N \tag{10}$$

because $I_N + \rho G_N = (I_N - \rho W_N)^{-1}$, where $G_N = W_N(I_N - \rho W_N)^{-1}$. This alternative expression of the reduced form plays an important role with respect to the identification conditions in Lee (2004) as we will see below.

The reduced form has two important implications. First, in (conditional) mean, the value of the dependent variable in the cross-sectional unit i will be affected not only by the values taken by the explanatory variable in the cross-sectional unit i, but also by those in all other cross-sectional units through the inverse spatial transformation $(I_N - \rho W_N)^{-1}$, which is a full matrix. This is the so-called *global spatial multiplier effect* or *global interaction effect*. Second, a random shock in a specific cross-sectional unit i does not only affect the value of the dependent variable in the same unit, but also has an impact on the values of dependent variable in all other cross-sectional units through the same inverse spatial transformation. This is the so-called *"spatial diffusion" process of random shocks.*[3]

1.3.3 Variance-covariance matrix

Assuming that X_N is nonstochastic, the mathematical expectation of y_N is $\mathrm{E}(y_N) = (I_N - \rho W_N)^{-1} X_N \beta$. The variance-covariance matrix of y_N is in

[2] Recall that λ is an eigenvalue of A if and only if $\det(\lambda I_N - A) = 0$. Note that $\det(\lambda I_N - A) = (-1)^n \det(A - \lambda I_N)$. The roots of $\det(A - \lambda I_N)$ are the same as those of $\det(\lambda I_N - A)$ (**?**, p. 38-39). The convention chosen here insures that the leading coefficient of λ^n is always $+1$.

[3] This is actually not a properly defined global interaction or diffusion process as we do not take into account the dynamics in the time dimension, which is of course necessary. This kind of process can only be analysed in the framework of space-time panel data models. In the cross-section models that we have here, interaction or diffusion is considered as instantaneous.

turn obtained as:

$$\mathrm{V}(y_N) = \sigma^2_{\varepsilon,N}(I_N - \rho W_N)^{-1}(I_N - \rho W_N)'^{-1} \tag{11}$$

This variance-covariance matrix is in general a full matrix, its structure is such that every location is correlated with every other location in the system, but closer locations more so. It is also interesting to note that the diagonal elements in equation (**??**), the variances at each location, are related to the neighborhood structure and are therefore not constant, inducing heterogeneity even though the initial error process is homoscedastic.

Moreover, let $\underline{\mathrm{X}}_N = [W_N y_N \; X_N]$ be the regressor matrix and $\delta = [\rho \; \beta']'$ the vector of coefficients in equation (**??**). The OLS estimator $\hat{\delta}_{OLS}$ of δ is $\hat{\delta}_{OLS} = (\underline{\mathrm{X}}'_N\underline{\mathrm{X}}_N)^{-1}\underline{\mathrm{X}}'_N y_N$. It follows that $\hat{\delta}_{OLS} = \delta + (\underline{\mathrm{X}}'_N\underline{\mathrm{X}}_N)^{-1}\underline{\mathrm{X}}'_N\varepsilon_N$. The consistency of the OLS estimator may depend on the limiting behaviour of $(1/N)(W_N y_N)'\varepsilon_N$, which is a component of $(1/N)\underline{\mathrm{X}}'\varepsilon_N$.[4]

Let us first consider the spatially lagged endogenous variable:

$$W_N y_N = G_N X_N \beta + G_N \varepsilon_N \tag{12}$$

where $G_N = W_N(I_N - \rho W_N)^{-1}$. Let us note that $W_N y_N$ is correlated with the error term ε_N in finite samples, because, in general:

$$\begin{aligned}
\mathrm{E}[(W_N y_N)'\varepsilon_N] &= \mathrm{E}[G_N X_N \beta + G_N \varepsilon_N)'\varepsilon_N] \\
&= \mathrm{E}[\beta' X'_N G'_N \varepsilon_N + \varepsilon'_N G'_N \varepsilon_N] \\
&= \beta' X'_N G'_N \,\mathrm{E}[\varepsilon_N] + \mathrm{E}[\varepsilon'_N G'_N \varepsilon_N] \\
&= \mathrm{E}[\varepsilon'_N G'_N \varepsilon_N] \neq 0
\end{aligned} \tag{13}$$

Indeed,

$$\begin{aligned}
E[\varepsilon'_N G'_N \varepsilon_N] &= \mathrm{E}[\mathrm{tr}(\varepsilon'_N G'_N \varepsilon_N)] \\
&= E[\mathrm{tr}(G'_N \varepsilon_N \varepsilon'_N)] \\
&= \mathrm{tr}\, G'_N \,\mathrm{E}(\varepsilon_N \varepsilon'_N) \\
&= \sigma^2_{\varepsilon,N} \,\mathrm{tr}\, G'_N = \sigma^2_{\varepsilon,N} \,\mathrm{tr}\, G_N \neq 0
\end{aligned} \tag{14}$$

generally $\mathrm{tr}\, G_N = \mathrm{tr}\, W_N(I_N - \rho W_N)^{-1}$ will not be equal to zero if $\rho \neq 0$, implying that OLS estimators will be biased in finite samples. Because of this non zero expectation and in the event that:

$$\lim_{N\to\infty}(1/N)(W_N y_N)'\varepsilon_N = \lim_{N\to\infty}(\sigma^2_{\varepsilon,N}/N)\,\mathrm{tr}\, G_N \neq 0$$

the OLS estimator is inconsistent, unless $(1/N)\,\mathrm{tr}\, G_N$ can become smaller and converge to zero as N tends to infinity (see **?**, for more details).

[4]The other component is $(1/N)X'\varepsilon_N$, which converges in probability to zero, and does not create any problem for OLS consistency.

1.4 Hypotheses

More specifically, the interaction matrix and the autoregressive parameter are generally assumed to satisfy the following assumptions mainly based on Kelejina and Prucha (1998, 1999, 2010, p. 55-56) and Lee (2002, 2004, p.1899-1925). In order to distinguish the true parameters from other possible values in the parameter space when needed, we denote ρ_0, β_0, and σ_0^2 as the true parameters which generate an observed sample as suggested by Lee (2004), especially which it comes to maximum likelihood estimation.

Assumption 1 (a) The disturbances $\{\varepsilon_{i,N} : 1 \leq i \leq N, N \geq 1\}$ are identically distributed. Moreover, for each sample size N, they are jointly independently distributed with $\mathrm{E}(\varepsilon_{i,N}) = 0$ and $\mathrm{E}(\varepsilon_{i,N}^2) = \sigma_{\varepsilon,N}^2$, where $0 < \sigma_{\varepsilon,N}^2 < b$ with $b < \infty$. (b) Finally, $\mathrm{E}(|\varepsilon_{i,N}|^{4+\eta})$ for some $\eta > 0$ exists, that is, a moment higher than the fourth exists.

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

Assumption 3 (a) All diagonal elements of W_N are zero. (b) $\rho \in (-\underline{a}_N^\rho, \bar{a}_N^\rho)$ with $0 < \underline{a}_N^\rho, \bar{a}_N^\rho \leq a^\rho < \infty$. (c) The matrix $I_N - \rho W$ is non singular for all $\rho \in (-\underline{a}_N^\rho, \bar{a}_N^\rho)$.

Assumption 4 The row and column sums of the sequences of matrices W_N and $(I_N - \rho W_N)^{-1}$ at $\rho = \rho_0$ are bounded uniformly (in absolute value).

Let A_N be a square matrix, we say that the row and column sums of the sequences of matrices A_N is bounded uniformly in absolute value if there exists a constant $c < \infty$ that does not depend on N such that:

$$\|A_N\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{N} |a_{ij,N}| < c, \quad \|A_N\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{N} |a_{ij,N}| < c, \text{ for all } N$$

Note that this condition is identical to the condition that the sequences of the maximum row sum matrix norms $\|A_N\|_\infty$ and the maximum column sum matrix norms $\|A_N\|_1$ are bounded (Horn and Johnson, 1985, p.294-295).

Assumption 5 $(I_N - \rho W_N)^{-1}$ are uniformly bounded in either row or column sums, uniformly in ρ in a compact parameter space Λ. The true ρ is in the interior of Λ.

Assumption 6 Sufficient condition for global identification:
The $\lim_{n\to\infty} \frac{1}{N}[X_N \;\; G_N X_N \beta]'[X_N \;\; G_N X_N \beta]$ exists and is non singular, where $G_N = W_N(I_N - \rho W_N)^{-1}$.

Under this set of assumptions, one can prove $\sqrt{N}$-rate of convergence and asymptotic normality of the Maximum Likelihood and Quasi Maximum Likelihood estimators. The latter is appropriate when the estimator is derived from a Normal Likelihood but the disturbances in the model are not truly normally distributed. Actually **?** used a set of less restrictive assumptions to proof the results (see below).

Discussion
Assumption 1 allows the error term to depend on the sample size N, i.e. to form *triangular arrays*. Note that even if the error term does not depend on N, the elements of y_N would still depend on N since the elements of the inverse of $I_N - \rho W_N$ would generally depend on N. A *triangular array* (tableau triangulaire) of random variables is a doubly indexed sequence in which each row (column) is only as long as the row's (column's) index. For example, the first element of the vector y will be different if $N = 10$ and $N = 15$. This implies that these elements and the vector y should be indexed by N:

$$y_N = (y_{1N}, y_{2N}, \ldots, y_{NN})$$

Our samples for y for $N = 1, 2, 3, \ldots$ are then (in rows):

$$\begin{array}{r|lll} N=1 & y_{11} & & \\ N=2 & y_{12} & y_{22} & \\ N=3 & y_{13} & y_{23} & y_{33} \\ \ldots & & \ldots & \end{array}$$

where $y_{11} \neq y_{12} \neq y_{13}$, $y_{22} \neq y_{23}$ etc.

The triangular nature of the variables, which leads to certain statistical problems, especially with respect to the relevant Central Limit Theorem to apply to get the asymptotic properties of the maximum likelihood estimators, has, as far as we know, only been recognized by (**?**) or Lee (2002, 2004). Because statistics involving quadratic forms of ε_N will be present in the estimation, the existence of the fourth order moment of $\varepsilon_{i,N}$ will guarantee finite variances for the quadratic forms. The higher than the fourth moment condition in Assumption 1 is needed in order to apply a Central Limit Theorem due to **?**.

Assumption 2 rules out multicollinearity among the regressors of X_N. The nonstochastic nature of X_N and its uniform boundedness conditions in

Assumption 2 are for simplicity. They can be relaxed without any problem: if the elements of X_N are stochastic and have unbounded ranges, conditions in Assumption 2 can be replaced by some finite moment conditions.

Assumption 3, 4 and 5 deserve some particular attention as they are properly specific to spatial regression models. Assumptions 3(a) is clearly a normalization rule: it implies that no unit is viewed as its own neighbor. Assumption 3(b) underlines that the autoregressive parameter ρ depends on the sample size N as underlined by Kelejian and Prucha (2010, p.54-55). It also defines the parameter space for ρ as an interval around zero such that $(I_N - \rho W_N)$ is non-singular for values of ρ in that interval. This Assumption will be discussed in greater details in the next section. Assumption 3(c) ensures that y_N is uniquely defined in reduced form in equation (**??**). In other words, Assumption 4 guarantees that the system (**??**) has an equilibrium and Y_N has mean $(I_N - \rho W)^{-1} X_N \beta$ and variance $\sigma_\varepsilon^2 (I_N - \rho W)^{-1} (I_N - \rho W)'^{-1}$.

Assumptions 3 and 4 imply that the row and column sums of the variance-covariance matrix of y_N in equation (**??**) are uniformly bounded in absolute value as N goes to infinity, thus limiting the degree of correlation between the elements of y_N. If A_N and B_N are two matrices conformable for multiplication and whose row and column sums are uniformly bounded in absolute value. Then the row and column sums of $A_N B_N$ are also uniformly bounded in absolute value **?**. Indeed the extent of correlation is limited in virtually all large sample analysis. Making an analogy to the time series literature, these assumptions ensure that the process for the dependent variable exhibit a fading memory.

Consider the following simple illustration: let $\{X_i\}$, $i = 1, ..., N$ be a random sample, where $\mathrm{E}(X_i) = \mu$, $\mathrm{V}(X_i) = \sigma^2$ for all i and $\mathrm{cov}(X_i, X_j) = a\sigma^2$ with $0 < a < 1$ for $i \neq j$. Consider now the sample mean: $\bar{X}_N$, then clearly $\mathrm{E}(\bar{X}_N) = \mu$, but $\mathrm{V}(\bar{X}_N) = \frac{\sigma^2}{N^2}[N^2 a + N(1-a)]$ and $\lim_{N\to\infty} \mathrm{V}(\bar{X}_N) = a\sigma^2 \neq 0$. Therefore, $\bar{X}_N$ is not consistent in mean square. For $\bar{X}_N$ to be consistent in mean square, the extent of correlation must be limited so as its variance goes to zero when the sample size goes to infinity, which implies that the covariances go to zero for large $|i - j|$.

In practice, interactions matrices, specially spatial weight matrices, are often specified to be row normalized. In many of these cases, no spatial unit is assumed to be a neighbor to more than a given number q of other spatial units. That is, for every j the number of $w_{ij} \neq 0$ is less than or equal to q. In other words, each spatial unit has a *limited* number of neighbors regardless of the sample size N. Clearly, in such cases, the spatial weight matrix W_N is *sparse* for large N and Assumption 3 is satisfied. Also, in other cases, the spatial weight matrix does not contain zeros, but the weights are

formulated such that they decline rapidly as a function of some measure of distance between neighbors. Again, in such cases Assumption 3 will typically be satisfied for W_N.

The uniform boundedness condition of $(I_N - \rho W)^{-1}$ at ρ_0 in Assumption 4 implies that the sequence of matrices $(I_N - \rho W)^{-1}$ are uniformly bounded in both row and column sums uniformly in a neighborhood of ρ_0.

Assumption 5 is needed to deal with the nonlinearity of $\ln|(I_N - \rho W)^{-1}|$ as a function of ρ in the log-likelihood function (Lee, 2004, Appendix A). It is stronger than Assumption 4.

If $||W|| \leq 1$ for all N, where $||.||$ is a matrix norm, then the sequence of matrices $||(I_N - \rho W)^{-1}||$ are uniformly bounded in any subset of $(-1, 1)$ bounded away from the boundary. In particular if W is row-normalized $(I_N - \rho W)^{-1}$ is uniformly bounded in row sums norm uniformly in any closed subset of $(-1, 1)$. For this case, Λ in Assumption 5 can be taken as a single closed set contained in $(-1, 1)$ for all N. For the case where W is not row-normalized but its eigenvalue are real, since the Jacobian $|(I_N - \rho W)^{-1}|$ will be positive if $-1/\nu_{min} < \rho < 1/\nu_{max}$, where ν_{min} and ν_{max} are the minimum and maximum eigenvalues of W, Λ can be a closed interval contained in $(-1/\nu_{min}, 1/\nu_{max})$ for all N (see below section 1.5 parameter space for a complete discussion on this issue). Note finally that Assumption 5 rules out consideration of models where the true ρ_0 is close to -1 or 1. Assumption 5 is not required by Kelejian & Prucha (1998, 1999) in the context of Instrumental Variables or Generalized Method of Moments estimation approaches where only Assumption 4 is used.

Assumption 6 requires that the generated regressors $G_N X_N \beta$ and X_N are not asymptotically multicollinear. It is a sufficient condition for global identification of the vector of unknown parameters $(\beta', \rho, \sigma^2)'$ (see Lee, Theorems 3.1 and 3.2, p.1905-1906, 2004). Indeed, note that the reduced form equation of the SAR model can be represented as follows, replacing Y_N from the reduced form equation (**??**) in equation (**??**):

$$Y_N = X_N \beta + \rho G_N X_N \beta + (I_N - \rho W_N)^{-1} \varepsilon_N \tag{15}$$

using the fact that $I_N + \rho G_N = (I_N - \rho W_N)^{-1}$. It should be noted that the presence of X_N and the linear independence of $G_N X_N \beta$ and X_N are the crucial conditions for the asymptotic results, especially the $\sqrt{N}$-rate of convergence of the vector of unknown parameters.

However $G_N X_N \beta$ and X_N can be linearly dependent if $\beta = 0$, which corresponds to the pure SAR model. In this case $\beta = 0$, then $G_N X_N \beta = 0$ and, hence, the set of $G_N X_N \beta$ and X_N is linearly dependent. $G_N X_N \beta$ and

X_N can also be linearly dependent if W_N is row-normalized and the relevant regressor is only a constant term. To see this let $X_N = [\iota_N \quad X_{1N}]$ and conformably $\beta = [\beta_0, \quad \beta_1']$, where $\beta_1 = 0$. Consequently as $X_N\beta = \iota_N\beta_0$, $G_NX_N\beta = (\beta_0/(1-\rho))\iota_N$, because $W_N\iota_N = \iota_N$ implies that $(I_N - \rho W_N)\iota_N = (1-\rho)\iota_N$ and $G_NX_N\beta\iota_N = (1/(1-\rho))\iota_N$. The multicollinearity of $G_NX_N\beta$ and X_N is equivalent to the columns of $G_NX_N\beta$ lying in the space spanned by the columns of X_N, i.e., $M_NG_NX_N\beta = 0$ where $M_N = I_N - X_N(X_N'X_N)^{-1}X_N'$. It is also possible that even though $G_NX_N\beta$ and X_N are linear independent for finite N, they become asymptotically multicollinear as N goes to infinity. This case corresponds to:[5]

$$\lim_{N\to\infty}(1/N)[G_NX_N\beta]'M_NG_NX_N\beta = 0$$

To take into account those multicollinearity cases, Lee (2004, p.1907) replaces Assumption 6 by the following assumption:

Assumption 6' (Lee, 2004)
$\lim_{N\to\infty}(1/N)[G_NX_N\beta]'M_N[G_NX_N\beta] = 0$

and again proofs consistency and asymptotic normality of the ML and QML estimators (Theorems 4.1 and 4.2, p.1908) under a more general set of assumptions on the interaction matrix (see Lee, 2004, for more details on some particular cases). Assumptions 2 and 3 in Lee (2004, p.1902) are as follows:

Assumption 7 (Lee, 2004)
The elements $w_{ij,N}$ of W_N are at most of order h_N^{-1}, denoted by $O(1/h_N)$, uniformely in all i, j, where the rate sequence can be bounded or divergent. As a normalization, the elements of the main diagonal of W_N, denoted $w_{ii,N} = 0$ for all i.

Assumption 8 (Lee, 2004) The ratio $h_N/N \to 0$ as N goes to infinity.

Assumptions 7 and 8 are general in that they cover interaction matrices where elements are not restricted to be nonnegative and those that might not be row-normalized.

Standard interaction matrices where neighboring units are defined by only a few adjacent ones (binary interaction matrices based, for example, on first order contiguity, on k-nearest neighbors with k small compared

[5] From the partitioned matrix formula, the $\lim_{N\to\infty}\frac{1}{N}[X_N \quad G_NX_N\beta]'[X_N \quad G_NX_N\beta]$ is nonsingular if and only if $\lim_{N\to\infty}(1/N)X_N'X_N$ (see Assumption 2) and $\lim_{N\to\infty}(1/N)[G_NX_N\beta]'M_N[G_NX_N\beta]$ are nonsingular.

to the sample size N, on band-distance with a relatively small threshold), commonly used in the applied literature, satisfy assumptions 7 and 8.

Assumption 7 means that, for some real constant c, there exists a finite integer N_c such that for all $N > N_c$, $|h_N w_{ij,N}| < c$ for all i, j.

- This assumption is always satisfied if $\{h_N\}$ is a bounded sequence.
- For models with a few neighboring units, $\{h_N\}$ would be bounded.

It is for example the case if W_N is row-normalized such that its ith row $w_{i,N} = (d_{i1}, d_{i1}, ..., d_{i1})/\sum_{j=1}^{N} d_{ij}$, where $d_{ij} \geq 0$, represents a function of the distance between the ith and jth units in some relevant space. For a row-normalized interaction matrix, as d_{ij} are non-negative constants and uniformly bounded, if the row-sums $\sum_{j=1}^{n} d_{ij}$, for $i = 1, ..., N$ are uniformly bounded away from zero at the rate h_N, in the sense that $\sum_{j=1}^{N} d_{ij} = O(h_N)$ uniformly in i and $\liminf_{N\to\infty} h_N^{-1} \sum_{j=1}^{N} d_{ij} > c$, where c is a positive constant independent of i and N, the implied row-normalized interaction matrix will have the property ascribed by assumption 7.

Assumption 8 excludes the cases where $\sum_{j=1}^{N} d_{ij}$, for $i = 1, ..., N$, diverges to infinity at a rate equal to or faster than the rate of the sample size N, because the ML estimator would likely be inconsistent for those cases.

- An important case that $\{h_N\}$ might diverge to infinity and still satisfies Assumptions 7 and 8 is that of Case (1991).

In this model, neighbors refer to farmers who live in the same district. Suppose that there are R districts and there are m farmers in each district (for simplicity). The sample size is $N = mR$. It is assumed that in a district, each neighbor of a farmer is given equal weights. In that case $W_N = I_N \otimes B_m$, where $B_m = (\iota_m \iota_m' - I_m)/(m-1)$, $\otimes$ is the Kronecker product, and ι_m is a m-dimensional column vector of ones. For this exemple $h_n = m - 1$ and $h_N/N = m - 1/mR = O(1/R)$.

If the increase of the sample size N is partly generated by the increase of the number of farmers m, then h_N goes to infinity. If the number of district R is finite, h_N is divergent at the N rate; but if R goes also to infinity, h_N is divergent at a rate slower than the N rate. In other words if the sample size N increases by increasing both R and m, then h_N goes to infinity and h_N/N goes to zero as N goes to infinity. Therefore the ML estimator would still be consistent in this case. However if the number of districts R is finite, then h_N is divergent at the N rate and the ML estimator is no more consistent.

Consider for simplicity the case where data is only collected in one district, the interaction matrix is then $W_N = (\iota_N \iota_N' - I_N)/(N-1)$. In this

case, $h_N = N - 1$ is $O(N)$, i.e. divergent at the N rate, and Lee (2004) shows then that the ML estimator is no more consistent. Therefore if h_N diverges to infinity at a rate equal to or faster than the sample size N, the ML estimators would be inconsistent.

Whether $\{h_N\}$ is a bounded or divergent sequence has interesting implications on the OLS approach. The OLS estimators of β and ρ are inconsistent when $\{h_N\}$ is bounded, but they can be consistent when $\{h_N\}$ is divergent (see Lee, 2002, for details).

1.5 Parameter space (Kelejian and Prucha, 2010)

Assumption 3(b) defines the parameter space for the autoregressive parameter. In the existing literature the parameter space for the autoregressive parameter is typically taken to be the interval $(-1, 1)$, or a subset thereof, and the autoregressive parameter is assumed not to depend on the sample size. However, in applications it is typically found that for un-normalized interaction matrices, $I_N - \rho W_N$ is singular for some values of $\rho \in (-1, 1)$.

To avoid this situation, many applied researchers normalize each row of their interaction matrices in such a way that $I_N - \rho W_N$ is non-singular for all $\rho \in (-1, 1)$. We now discuss the implications of various normalizations of the interaction matrix.

Suppose c_N denotes a scalar normalization factor. Clearly, this normalization factor may depend on the sample size. For example, some of our results below relate to the case in which c_N corresponds to the maximal row or column sum of the absolute values of the elements of W_N. Given such a normalizing factor, an equivalent specification of model (1) for y_N is obtained if ρW_N is replaced by $\rho^* W_N^*$ where $\rho^* = c_N \rho$ and $W_N^* = W_N / c_N$. It is important to observe that even if ρ and its corresponding parameter space do not depend on N, ρ^* and its implied parameter space will depend on the sample size as a result of the normalization of the interaction matrix.

It is for this reason that we allow in Assumption 3 for the elements of the interaction matrices, and the autoregressive parameters and the corresponding parameter spaces to depend on N.[6] Of course, Assumption 3 also covers the case where the true data generating process corresponds to a model where autoregressive parameters do not depend on N.

[6] To keep the notation as simple as possible we will no more index all the variables by the sample size N.

1.5.1 General results

Having defined in Assumption 3, the parameter space for ρ as an interval around zero such that $(I_N - \rho W)$ is non-singular for values of ρ in that interval, the following Theorem gives the bounds for that interval (**?**, **?**).

Theorem 1 (Kelejian and Prucha, 2010, p. 56)
Let τ denote the spectral radius of W, i.e.,

$$\tau = \max\{|\nu_1|, ..., |\nu_N|\}$$

where $|\nu_1|, ..., |\nu_N|$ denote the modulus of the eigenvalues of W. Then $(I_N - \rho W)$ is nonsingular for all values of ρ in the interval $(-1/\tau, 1/\tau)$.

Proof. Consider that for $\rho \neq 0$, $\det(I_N - \rho W) = \det[(-\rho)(W - \frac{1}{\rho} I_N)] = (-\rho)^N \det(W - \frac{1}{\rho} I_N)$. Consequently $(I_N - \rho W)$ is non singular for values of $\rho^{-1} \notin \{\nu_1, ..., \nu_N\}$, i.e. if ρ^{-1} is not an eigenvalue of W. In particular $(I_N - \rho W)$ is nonsingular for $|\rho^{-1}| > \tau$. Rewriting the last inequality as $|\rho| < \tau^{-1}$ completes the proof. ■

Remark 1 Note that if $|\rho| < \tau^{-1}$, $(I_N - \rho W)^{-1}$ can be expanded into an infinite series as:

$$(I_N - \rho W)^{-1} = I_N + \rho W + \rho^2 W^2 + ... + \rho^r W^r + ... = \sum_{r=0}^{\infty} \rho^r W^r \qquad (16)$$

Theorem 2 (Kelejian and Prucha, 2010, p.56)
Let

$$r = \max_{1 \leq i \leq N} \sum_{j=1}^{N} |w_{ij}|, \qquad c = \max_{1 \leq j \leq N} \sum_{i=1}^{N} |w_{ij}|$$

and let

$$\tau^* = \min(r, c)$$

Then $\tau \leq \tau^*$ and consequently $I_N - \rho W$ is non-singular for all values of ρ in the interval $(-1/\tau^*, 1/\tau^*)$.

Proof. Note first that r is the maximum row sum matrix norm and c is the maximum column sum matrix norm of W. As an immediate consequence of Gershgorin's Theorem (Horn and Johnson, 1985, p.344-346).: the spectral radius is the greatest lower bound for the values of all matrix norms of W. We then have $\tau = \max\{|\nu_1|, ..., |\nu_N|\} \leq \tau^*$. The result now follows from Theorem 1. ■

Note however that the specification of the parameter space is here somewhat more restrictive than the previous one, i.e. $(-1/\tau^*, 1/\tau^*) \subset (-1/\tau, 1/\tau)$.

It is possible to consider the specification of the parameter space for ρ in some special cases such as when W is a row normalized interaction matrix or when W is a symmetric, not row normalized interaction matrix.

1.5.2 Row normalized W matrix

Theorem 3
If W is a row normalized interaction matrix, then $(I_N - \rho W)^{-1}$ exists for all $|\rho| < 1$.

Proof. (1) Consider the case where W is normalized to have row sums of unity. All the eigenvalues of the row normalized W matrix are then less than 1: indeed, in this case $r = 1$, then $|\nu_i| \leq 1$ for all i. Using Theorem 2, $(I_N - \rho W)$ is nonsingular for all values of ρ in the interval $(-1, 1)$. ■

Proof. (2) Alternatively, let us prove this result using Gershgorin's Theorem.

Theorem 4 (Gershgorin's Theorem)
Let A be a $(N \times N)$ square matrix with elements a_{ij}. Let

$$R_i = \sum_{j=1, j\neq i}^{N} |a_{ij}|, \qquad C_j = \sum_{i=1, i\neq j}^{N} |a_{ij}|$$

Then each eigenvalue of A lies in at least one of the N circles defined by:

$$|\nu - a_{ii}| \leq R_i, \ i = 1, ..., N$$

and hence in the union of these circles. Since the spectrum of A equals the spectrum of A', the deleted absolute row sums can be replaced by the deleted absolute column sums, so that also each eigenvalue of A lies in at least one of the N circles defined by:[7]

$$|\nu - a_{jj}| \leq C_j, \ j = 1, ..., N$$

and hence in their union.

[7] The set of distinct eigenvalues, denoted by $\sigma(A)$, is called the spectrum of A.

Consider now W which has $w_{ij} \geq 0$ and $w_{ii} = 0$ and let

$$r = \max_i \sum_{j=1}^{N} w_{ij} = \max_i R_i, \qquad c = \max_j \sum_{i=1}^{N} w_{ij} = \max_j C_j,$$

Then the eigenvalues of W satisfy for $i = 1, ..., N$

$$|\nu_i| \leq r, |\nu_i| \leq c, \quad i = 1, \ldots, N$$

If W is row normalized, $r = 1$ and so $|\nu_i| \leq 1$.

Next, let Q be a nonsingular matrix that triangularize W as follows:

$$QWQ^{-1} = G_\nu, \qquad G_\nu = \begin{pmatrix} \nu_1 & \ldots & * \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \nu_N \end{pmatrix}$$

This is always possible, e.g. G_ν may represent the Jordan normal form (Horn and Johnson, 1995, p.119-128). Then

$$\begin{aligned} \det(I_N - \rho W) &= \det(Q^{-1}Q(I_N - \rho W)) = \det(Q(I_N - \rho W)Q^{-1}) \\ &= \det(QQ^{-1} - \rho QWQ^{-1}) = \det(Q^{-1}Q - \rho G_\nu) \\ &= \det(I_N - \rho G_\nu) \\ &= (1 - \rho\nu_1)(1 - \rho\nu_2)\ldots(1 - \rho\nu_N) = \Pi_{i=1}^{N}(1 - \rho\nu_i) \end{aligned}$$

and $\det(I_N - \rho W) \neq 0$ for any $|\rho| < 1$, since $|\rho\nu_i| \leq |\rho| < 1$. ■

Remark 2 Moreover it is clear that $\det(I_N - \rho W) > 0$ for any $|\rho| < 1$, meaning that $I_N - \rho W$ is definite positive with all eigenvalues strictly positive.

Remark 3 If W is not row normalized $I_N - \rho W$ will generally be singular for certain values of $|\rho| < 1$. Then using Theorems 1 or 2 it is always possible to normalize the interaction matrix in such a way that the inverse of $I_N - \rho W$ will exist in an easily established region.

Consider for exemple the following model where W is not row normalized:

$$y = \rho W y + X\beta + \epsilon = \rho^* W^* y + X\beta + \epsilon \tag{17}$$

where $\rho^* = \tau^*\rho$, $W^* = \frac{W}{\tau^*}$ and $\tau^* = \min(r, c)$ defined in Theorem 2. Note that $|I_N - \rho^* W^*| \neq 0$ for:

$$|\rho^*| < \frac{1}{\min\left(\frac{r}{\tau^*}, \frac{c}{\tau^*}\right)} = \frac{1}{\frac{1}{\tau^*}\min\left(r, c\right)} = 1 \tag{18}$$

So if the model is normalized using τ^* and ρ^* is taken to be the parameter, $(I_N - \rho^* W^*)^{-1}$ exists for all $|\rho^*| < 1$. One would then estimate ρ^* as a parameter, and since $\rho^* = \tau^* \rho$, one would estimate ρ as $\hat{\rho} = \hat{\rho}^*/\tau^*$.

We could also use τ, the spectral radius of W, defined in Theorem 1, for an alternative normalization. Note however that τ^* is much easier to compute than τ, especially for large sample sizes.

1.5.3 Symmetric W matrix

Theorem 5
Consider the case where W is symmetric, all the eigenvalues of W are then real. Assume that W is not row normalized. Let ν_{max} and ν_{min} be the largest and the smallest eigenvalues of W. Assume as will typically be case if all the eigenvalues of W are real, that $\nu_{max} > 0$ and $\nu_{min} < 0$. Then $(I_N - \rho W)$ is nonsingular for all values of ρ in the interval $(\nu_{min}^{-1}, \nu_{max}^{-1})$.

Proof. If $\rho = 0$, $(I_N - \rho W)$ is nonsingular. If $\rho \neq 0$ we have:

$$|I_N - \rho W| = (1 - \rho\nu_1)(1 - \rho\nu_2) \dots (1 - \rho\nu_N) = \Pi_{i=1}^{N}(1 - \rho\nu_i)$$

so $(I_N - \rho W)$ is nonsingular unless ρ is equal to the inverse of an eigenvalue $\nu_1^{-1}, \nu_2^{-1} \dots \nu_N^{-1}$, i.e ρ^{-1} is equal to an eigenvalue. Thus $(I_N - \rho W)$ is nonsingular if $\rho^{-1} < \nu_{min}$ or $\rho > \nu_{min}^{-1}$ and $\rho^{-1} > \nu_{max}$ or $\rho < \nu_{max}^{-1}$ and therefore if $\rho \in (\nu_{min}^{-1}, \nu_{max}^{-1})$ ■

Remark 4 Consider now the normalization by τ of a symmetric not row-normalized matrix W. As all the eigenvalues of this matrix are real so that $\nu_{max} > 0$ and $\nu_{min} < 0$. We know that $(I_N - \rho W)$ is nonsingular for all values of ρ in the interval $(1/\nu_{min}, 1/\nu_{max})$. Consider $\rho^* = \tau\rho$ and $W^* = W/\tau$, eigenvalues of the normalized matrix W^* are given by $\nu^* = \nu/\tau$ and $(I_N - \rho^* W^*)$ is nonsingular for all values of ρ^* in the interval $(\tau/\nu_{min}, \tau/\nu_{max})$. Suppose that $|\nu_{max}| > |\nu_{min}|$, then the previous interval becomes $(\nu_{max}/\nu_{min}, 1)$. Suppose now that $|\nu_{max}| < |\nu_{min}|$, then $\tau = |\nu_{min}|$ and the previous interval becomes $(-1, |\nu_{min}|/\nu_{max})$. Note that those two intervals are less restrictive than the interval $(-1, 1)$: $(-1, 1) \subset (\nu_{max}/\nu_{min}, 1)$ and $(-1, 1) \subset (-1, |\nu_{min}|/\nu_{max})$.

Remark 5 (Similarity) Consider the case where the row normalized matrix is computed from a symmetric matrix, the row normalized matrix is no more symmetric and may have complex eigenvalues. However in this case, it

will have the same real eigenvalues and determinant as a similar symmetric matrix.[8]

Theorem 6
For symmetric matrices similar to row normalized matrices where $\nu_{max} = 1$, $(I_N - \rho W)$ is nonsingular for all values of ρ in the interval $(\nu_{min}^{-1}, 1)$.

Proof. Consider $R = I_N - \rho DW$, where DW is row normalized and non-symmetric. However, in this case DW has the same real eigenvalues as the symmetric but not row normalized matrix $D^{1/2}WD^{1/2}$. Consider the similarity transformation:

$$\begin{aligned} D^{-1/2}RD^{1/2} &= D^{-1/2}(I_N - \rho DW)D^{1/2} = I_N - \rho D^{-1/2}DWD^{1/2} \\ &= I_N - \rho D^{1/2}WD^{1/2} \end{aligned}$$

Then $I_N - \rho DW$ and $I_N - \rho D^{1/2}WD^{1/2}$ have then the same real eigenvalues and determinant. As DW is row normalized $\nu_{max} = 1$ and $I_N - \rho D^{1/2}WD^{1/2}$ is nonsingular for all values of ρ in the interval $(\nu_{min}^{-1}, 1)$. ■

Remark 6 (Strictly diagonally dominant matrix) Consider again $R = I_N - \rho DW$ where DW is row normalized and non-symmetric and $|\rho| < 1$, then R is strictly diagonally dominant. This means that the diagonal element of R (which equals 1) strictly exceeds the sum of the other elements in the row (which equals ρ since W is row normalized). Strictly diagonally dominant matrices are invertible, therefore R is invertible.

Remark 7 (LeSage and Pace, 2009, p.88-89) Consider the case where W has complex eigenvalues: W is row normalized and is not similar to a symmetric matrix. If a real matrix has complex eigenvalues, these come in complex conjugate pairs. Consider the determinant of $I_N - \rho W$:

$$\det(I_N - \rho W) = \Pi_{i=1}^{N}(1 - \rho\nu_i) = [\Pi_{i=3}^{N}(1 - \rho\nu_i)](1 - \rho\nu_1)(1 - \rho\nu_2)$$

where, without loss of generality, one of the complex conjugate pairs of eigenvalues appears in ν_1 and the other in ν_2. If the product $(1-\rho\nu_1)(1-\rho\nu_2)$ equals 0, this would lead to a zero determinant which would imply that $(I_N - \rho W)$ is singular. The question is then what values of ρ could lead to a singular $(I_N - \rho W)$? Let $\nu_1 = r + ic$ and $\nu_2 = r - ic$ where r is the real

[8] Recall that a Matrix B is said to be similar to a matrix A if there exist a nonsingular matrix S such that $B = S^{-1}AS$. Note also that if A and B are similar, then they have the same eigenvalues, counting multiplicity (Horn and Johnson, p.44-45).

part of ν_1 and ν_2, c is the imaginary part of ν_1 and ν_2 and i is the square root of -1, so that $i^2 = -1$. Assume that $c \neq 0$. Consider now:

$$\begin{aligned} &(1-\rho\nu_1)(1-\rho\nu_2) = 0 \\ &(1-\rho r-\rho ic)(1-\rho r+\rho ic) = 0 \\ &1-2\rho r+\rho^2 r^2-\rho^2 i^2 c^2 = 0 \\ &1-2\rho r+\rho^2(r^2+c^2) = 0 \end{aligned} \tag{19}$$

The discriminant of this quadratic equation in ρ is $d = -4c^2 < 0$. Therefore the quadratic equation will have two complex roots. This means that a real ρ cannot result as a root of this quadratic equation. In other words, complex conjugate eigenvalues do not affect whether $(I_N - \rho W)$ is singular. Only pure real eigenvalues can affect the singularity of $(I_N - \rho W)$.

Consequently, for W with complex eigenvalues, the interval for ρ which guarantees the non singularity of $(I_N - \rho W)$ is $(\nu_{min}^{-1}, 1)$ where ν_{min} is here the most negative purely real eigenvalue of W.

1.5.4 Definite positiveness of $(I - \rho W)$

We know that:

$$\det(I_N - \rho W) = \Pi_{i=1}^{N}(1-\rho\nu_i)$$

It is straightforward to see that $\det(I_N - \rho W) > 0$ if $|\rho| < 1/\tau$, where $\tau = \max\{|\nu_1|, ..., |\nu_N|\}$ as defined in Theorem 1, as well as if $|\rho| < 1/\tau^*$ where $\tau^* = \min(r, c)$ with $r = \max_{1\leq i\leq N}\sum_{j=1}^{N}|w_{ij}|$ and $c = \max_{1\leq j\leq N}\sum_{i=1}^{N}|w_{ij}|$ as defined in Theorem 2. It will also be the case if W is a row normalized matrix with $|\rho| < 1$ as in Theorem 3.

1.6 Estimation of the SAR model

Let us consider again the first order Spatial Autoregressive Model:

$$y = \rho W y + X\beta + \varepsilon \tag{20}$$

The variance-covariance matrix for y is easily seen to be equal to:

$$\mathrm{V}(y) = \sigma^2(I_N - \rho W)^{-1}(I_N - \rho W')^{-1} \tag{21}$$

Note that, in general, this matrix will be full and it's main diagonal will not be constant, inducing complete heteroscedasticity.

It also follows from the reduced form (**??**) that the spatially lagged variable Wy is, in general, correlated with the error term since:

$$E[(Wy)'\varepsilon] = \sigma^2 \operatorname{tr} W(I_N - \rho W)^{-1} \neq 0 \tag{22}$$

Therefore the OLS estimator will be biased in finite samples and inconsistent under Assumptions 1-5. However, as highlighted by Lee (2002), in some cases, it may still be consistent and even be asymptotically efficient relative to some other estimators.

1.6.1 Ordinary Least Squares estimation

Let us first consider the special cases where the OLS estimator is consistent (Lee, 2002). Those cases are often neglected in the literature but are nevertheless important, especially when it comes to economic applications.

As highlighted by Ertur and Koch (2011), traditionally, connectivity has been understood as geographical proximity, and various weights matrices based on geographical space have thus been used in the spatial econometric literature, such as contiguity, nearest neighbors and geographical distance-based matrices. However the definition is in fact much broader and can be generalized to any network structure to reflect any kind of interactions between observations. This is why we prefer to use the terms *interaction* matrix for W. We also propose to generalize spatial econometric methods to give birth to a new field of econometrics that we call "Econometrics of interactions".

As also underlined by Durlauf et al. (2005, p.643-645), what really matters when adapting these methods to, for instance, growth econometrics is the identification of the appropriate notion of space and of the appropriate similarity or interaction measure. By analogy to Akerlof (1997), countries may be considered as located in some general socio-economic and institutional or political space, defined by a range of factors. Implementation of spatial methods thus requires accurate identification of their localisation in such a general space.[9]

In this framework, it could be the case that the interaction structure is much denser between units than when considering for instance binary interaction matrices based on geographical space (first order contiguity, nearest neighbors, band-distance). In such a framework, Lee (2002) shows that the

[9]Ideally, such a matrix should be theory-based, but this is beyond the scope of the present section. Note that, the field of 'Econometrics of interactions' has been proposed in a forthcoming paper by Behrens, Ertur and Koch (2012), who furthermore derive the interaction matrix from their structural theoretical model.

OLS estimator can be consistent regardless of whether or not the disturbances are spatially correlated an even that it is asymptotically relatively efficient compared to other estimators.

Lee (2004) focuses on OLS estimation of SAR models where the interaction matrix is row-normalized and has the property that the row sum diverges to infinity for each unit $i = 1, ..., N$.

More specifically, consider the row-normalized interaction matrix based on some function d_{ij} of the distance between units i and j in some relevant space: $w_{ij} = d_{ij}/\sum_j d_{ij}$ and replace Assumptions 6 and 7 by the following one:

Assumption 9 The elements d_{ij}, in W_N are nonnegative constants and are uniformly bounded. The sums, $\sum_j d_{ij}$, are uniformly bounded away from zero at the rate h_N, where $\lim_{N\to\infty} h_N = \infty$.

This assumption requires that each unit in the limit has infinitely many neighbors. For economic applications where either the neighbors of any unit are dense in a relevant compact characteristic space or each unit is influenced by many of its neighboring units (which represents a significant portion of the total population of units), it is likely that $\sum_j d_{ij}$ will diverge and $(1/N)\sum_j d_{ij}$ will converge as N becomes large.

For example, in Case et al. (1993), interactions are defined in an ethnic space. In the most relevant interaction matrix used in this study, $d_{ij} = 1/|r_i - r_j|$, where r_i is the proportion of state i's population that is of African descent. As no state in the United States has a zero proportion of African-Americans in its population and that no couple of states share the same proportion of African-Americans in their population, d_{ij} would be positive. As there is a distribution of the population of African-Americans across regions of US states, $(1/N)\sum_j d_{ij}$ will be bounded away from zero and $\sum_j d_{ij}$ will be likely to possess the N rate of divergence in this example. Therefore it is likely that the OLS estimator is consistent is this setting.

The consistency of the OLS estimator depends crucially on $\lim_{N\to\infty} h_N = \infty$. Moreover, Lee (2002) shows that as long as h_N tends to infinity at a rate faster than $\sqrt{N}$, i.e. if $\lim_{N\to\infty}(h_N/\sqrt{N}) = \infty$, the OLS estimator will converge in probability to its true value at the usual $\sqrt{N}$ rate and its limiting distribution will be normal. Under this assumption, Lee (2004) also shows that the OLS estimator has the same limiting distribution as that of the Maximum Likelihood estimator and can therefore be asymptotically efficient.

1.6.2 Instrumental variables or two-stage least squares estimation

As the spatially lagged variable Wy is, in general, correlated with the error term ε, a classical instrumental variables method (IV) or, equivalently, two-stage least squares method (2SLS), have been proposed by Anselin (1988) and Kelejian and Prucha (1998) among others.

Let H denote the $N \times p$ matrix of instruments used to estimate the SAR model and let $Z = [Wy\ X]$ denote the matrix of regressors in equation (**??**), where we exclude spatially lagged explanatory variables:

$$y = Z\delta + \varepsilon \tag{23}$$

where $\delta = [\rho\ \beta']'$. Consider the following Assumptions in addition to Assumptions 1-4:

Assumption 10 The instrument matrices H have full column rank $p \geq k+1$ (for all N large enough). They are composed of a subset of the linearly independent columns of $(X, WX, W^2X, ...)$.

Assumption 11 Let H be a matrix of instruments:

(a) $\lim_{N\to\infty} n^{-1}H'H = Q_{HH}$ where Q_{HH} is finite and nonsingular.

(b) $\lim_{N\to\infty} n^{-1}H'Z = Q_{HZ}$ where Q_{HZ} is finite and has full column rank.

In the 2SLS approach, the predicted value of Z in a regression on the instruments H is obtained using OLS in the first stage, as $Z = H\theta + \eta$ with $\hat{\theta} = (H'H)^{-1}H'Z$:

$$\widehat{Z} = Z\hat{\theta} = H(H'H)^{-1}H'Z = P_H Z \tag{24}$$

with H an $N \times p$ matrix of instruments, including the exogenous variables X, with $p \geq k+1$. Note that it is straightforward to see that $P_H = H(H'H)^{-1}H'$ is the orthogonal projection matrix associated with H and is symmetric and idempotent, hence singular. This projection does not affect X but yields:

$$\widehat{Wy} = H(H'H)^{-1}H'Wy \tag{25}$$

We replace, in the second stage, Z by $\widehat{Z}$ in equation (**??**) to get the 2SLS estimator:

$$\hat{\delta}_{2SLS} = (\widehat{Z}'\widehat{Z})^{-1}\widehat{Z}'y \tag{26}$$

or equivalently, in full expression, the IV estimator:

$$\hat{\delta}_{IV} = [Z'H(H'H)^{-1}H'Z]^{-1}Z'H(H'H)^{-1}H'y \tag{27}$$

Inference on δ is then based on the asymptotic variance-covariance matrix:

$$AsyVar(\hat{\delta}_{IV}) = \hat{\sigma}^2_{\varepsilon}[Z'H(H'H)^{-1}H'Z]^{-1} \tag{28}$$

Discussion

In practice, **?** suggest using $[X\ WX\ W^2X]$ as instruments for Z, but they were aware that the ideal instruments would be $\mathrm{E}(Z) = [X\ W\,\mathrm{E}(y)]$, where $\mathrm{E}(y) = (I - \rho W)^{-1}X\beta$ using the reduced form, as also stressed in the discussions in **?** and **?**. In principle, the problem is to approximate $\mathrm{E}(Z)$ as closely as possible. Assumption 10 assumes that H contains, at least, the linearly independent columns of X and WX, which ensures that $\widehat{Z} = (X, \widehat{Wy})$ with $\widehat{Wy} = P_H Wy$.

Following **?**, suppose furthermore that W is row-normalized, all its eigenvalues are therefore less than or equal to one in absolute value. Then observing that $|\rho| < 1$, it is readily seen that:

$$\begin{aligned}\mathrm{E}(y) &= (I_N - \rho W)^{-1}X\beta \\ &= \left[\sum_{r=0}^{\infty}\rho^r W^r\right]X\beta, \quad W^0 = I_N\end{aligned} \tag{29}$$

Consequently, in this case, $W\,\mathrm{E}(y)$ is seen to be formed as a linear combination of the columns of the matrices $X, WX, W^2X, ...$ It is for this reason that they postulate in Assumption 10 that H is composed of a subset of the linearly independent columns of those matrices. In practice that subset might be the linearly independent columns of $[X\ WX\ W^2X]$, or if the number of regressors is large, just those of $[X\ WX]$.[10]

Assumption 11 will ensure that the estimators remain well defined asymptotically. More specifically, under this set of assumptions, it is possible to show that the estimators of ρ and β are consistent and asymptotically normally distributed (**?**).

Two points should nevertheless be noted.

First, Assumption 11b rules out models in which all of the parameters associated to the exogenous regressors, including the intercept parameter, if an intercept is present, are zero, i.e. $\beta = 0$. This case is of limited interest in practice since the mean of y would then be zero.

[10]Note that this is still valid if $|\rho| < \tau^{-1}$, where τ is the spectral radius of W, see equation (**??**) and does not require row-normalization.

Second, if W is row-normalized, Assumption 11b will not be satisfied if the only non-zero element of β corresponds to the constant term. In this case Assumption 11b requires that the generation of y involves at least one non-constant regressor. This implies that if W is row-normalized, the hypothesis that all the slopes are zero cannot be tested using 2SLS estimators.

Let ι_N be the $N \times 1$ vectors of unit elements. Also suppose that the first column of X is ι_N and that the remaining columns of X are denoted by the $N \times (k-1)$ matrix X_1 so that $X = [\iota \; X_1]$. Partition β accordingly as $\beta = (\beta_0 \; \beta_1')'$. Then the SAR model can be expressed as:

$$y = \beta_0 \iota + X_1 \beta_1 + \rho W y + \varepsilon \tag{30}$$

If W is row-normalized, we know that $W\iota_N = \iota_N$. Now if $\beta_1 = 0$, then it follows from (**??**) that:

$$\mathrm{E}(Wy) = W \left[\sum_{r=0}^{\infty} \rho^r W^r \right] \iota_N \beta_0 = \frac{\beta_0}{1-\rho} \iota_N \tag{31}$$

Thus the mean of Wy is not linearly independent of ι_N and the matrix Q_{HZ} does not have full column rank and thus Assumption 11b is violated. This means that at least one nonconstant regressor in X must have a significant coefficient in order that valid instruments can be generated. As the 2SLS estimators are based on the existence of relevant nonconstant regressors, it is impossible to test the joint significance of X with those estimators.

If W is not row-normalized, then is general $W\iota_N$ will be linearly independent of ι_N and Assumption 11b would not require the existence of a relevant nonconstant regressor in the generation of y.

Lee (2003) suggests a best 2SLS (B2SLS) estimator, which requires using $W(I_N - \rho W)^{-1} X$ and X as the matrix of instruments. But the proposed method has exactly the same limitation as the previous one and requires at least a relevant nonconstant regressor in X.

In addition, Lee (2007) notes that even if valid instruments do exist, the 2SLS estimator may be inefficient relative to the Maximum Likelihood (ML) estimator. Comparing the limiting variance matrices of ML and 2SLS estimators, he observes that neither the 2SLS nor the B2SLS estimators have the same limiting distribution as the ML estimator. He then suggests a best Generalized Method of Moments (GMM) estimator, incorporating some other moment conditions in addition to those based on X in order to improve upon the efficiency of the 2SLS estimator. Moreover the proposed method does not have the previous limitation, i.e. it is applicable even if all regressors are irrelevant.

1.6.3 Maximum Likelihood estimation

The simultaneity embedded in the Wy term must be explicitly accounted for in the maximum likelihood estimation framework, as first outlined by **?**. More recently, **?** presents a comprehensive investigation of the asymptotic properties of the maximum likelihood estimators of SAR models.

Under the hypothesis of normality of the error term, $\varepsilon \sim N(0, \sigma_\epsilon^2 I_N)$, the log-likelihood function for the SAR model is given by:

$$\begin{aligned} \ln L(\beta', \rho, \sigma_\epsilon^2) &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma_\epsilon^2) + \ln|I - \rho W| \\ &\quad -\frac{1}{2\sigma_\epsilon^2}\left[(I-\rho W)y - X\beta\right]'\left[(I-\rho W)y - X\beta\right] \end{aligned} \quad (32)$$

An important aspect of this log-likelihood function is the Jacobian of the transformation, which is the determinant of the $(N \times N)$ *full* matrix $(I_N - \rho W)$ for our model.

Recall that moving from the likelihood for the Normal error vector to the likelihood for the observed dependent variable, the Jacobian of the transformation needs to be inserted. In the SAR model, this corresponds to:

$$\left|\frac{\partial \varepsilon}{\partial y}\right| = \left|\frac{\partial(y - \rho W y - X\beta)}{\partial y}\right| = |I_N - \rho W| \quad (33)$$

Note that this Jacobian reduces to a scalar 1 in the standard regression model, since $|\partial(y - X\beta)/\partial y| = |I_N| = 1$. Maximizing the log-likelihood function (**??**) is therefore not equivalent to minimizing weighted least squares, i.e. the last term in the log-likelihood function, as in the standard linear regression model since it ignores the Jacobian term. This illustrates informally how weighted least squares will not yield a consistent estimator in the SAR model, due to the endogeneity in the Wy term. The log-Jacobian also implies constraints on the parameter space for ρ, which must be such that $|I_N - \rho W| > 0$.

Let us write the usual first order conditions for the maximization of the log-likelihood function (**??**):

$$\begin{aligned} \frac{\partial \ln L}{\partial \beta'} &= X'(I_N - \rho W)y - X'X\beta = 0 \quad (34) \\ \frac{\partial \ln L}{\partial \rho} &= -\operatorname{tr}[W(I_N - \rho W)^{-1}] + \frac{1}{\sigma_\varepsilon^2}[(I_N - \rho W)y - X\beta]'Wy = 0 \\ \frac{\partial \ln L}{\partial \sigma_\varepsilon^2} &= -\frac{N}{\sigma_\varepsilon^2} + \frac{1}{\sigma_\varepsilon^4}[(I_N - \rho W)y - X\beta]'[(I_N - \rho W)y - X\beta] = 0 \end{aligned}$$

Note that here we have used the following result to compute the partial derivative of the determinant of $I - \rho W$ with respect to ρ:

$$\begin{aligned} \frac{\partial ln|I_N - \rho W|}{\partial \rho} &= \mathrm{tr}(I_N - \rho W)^{-1} \frac{\partial (I_N - \rho W)}{\partial \rho} \\ &= \mathrm{tr}(I_N - \rho W)^{-1}(-W) \\ &= -\mathrm{tr}[W(I_N - \rho W)^{-1}] \end{aligned} \tag{35}$$

The presence of the Jacobian term could complicate the computation of the maximum likelihood estimators which involves the repeated evaluation of this determinant. However **?** suggests that it can be expressed as a function of the eigenvalues ω_i of the spatial weights matrix W:

$$|I_N - \rho W| = \prod_{i=1}^{N}(1 - \rho\omega_i) \quad \Longrightarrow \quad \ln|I_N - \rho W| = \sum_{i=1}^{N} \ln(1 - \rho\omega_i) \tag{36}$$

This expression simplifies considerably the computations since the eigenvalues of W have to be computed only once at the outset of the numerical optimization procedure.

From the usual first-order conditions, the maximum likelihood estimators of β and σ^2, given ρ, are obtained as:

$$\hat{\beta}_{ML}(\rho) = (X'X)^{-1}X'(I - \rho W)y \tag{37}$$

$$\hat{\sigma}^2_{ML}(\rho) = \frac{1}{N}\left[(I - \rho W)y - X\hat{\beta}_{ML}(\rho)\right]'\left[(I - \rho W)y - X\hat{\beta}_{ML}(\rho)\right] \tag{38}$$

Note that, for convenience:

$$\hat{\beta}_{ML}(\rho) = \hat{\beta}_O - \rho\hat{\beta}_L \tag{39}$$

where $\hat{\beta}_O = (X'X)^{-1}X'y$ and $\hat{\beta}_L = (X'X)^{-1}X'Wy$. Define $\hat{e}_O = y - X\hat{\beta}_O$ and $\hat{e}_L = Wy - X\hat{\beta}_L$, it can be then easily seen that:

$$\hat{\sigma}^2_{ML}(\rho) = \left[\frac{(\hat{e}_O - \rho\hat{e}_L)'(\hat{e}_O - \rho\hat{e}_L)}{N}\right] \tag{40}$$

Substitution of (**??**) and (**??**) in the log-likelihood function (**??**) yields a concentrated log-likelihood as a non-linear function of a single parameter ρ:

$$\begin{aligned} \ln L(\rho) &= -\frac{N}{2}\left[\ln(2\pi) + 1\right] + \sum_{i=1}^{N} \ln(1 - \rho\omega_i) \\ &\quad -\frac{N}{2}\ln\left[\frac{(\hat{e}_O - \rho\hat{e}_L)'(\hat{e}_O - \rho\hat{e}_L)}{N}\right] \end{aligned} \tag{41}$$

where $\hat{e}_O$ and $\hat{e}_L$ are the estimated residuals in a regression of y on X and Wy on X, respectively. A maximum likelihood estimate for ρ is obtained from a numerical optimization of the concentrated log-likelihood function (**??**).[11]

The estimation procedure can then be described as follows. We begin by first regressing by OLS y on X which yields $\hat{\beta}_O$ and Wy on X which yields $\hat{\beta}_L$. We then compute the estimated residuals $\hat{e}_O$ and $\hat{e}_L$. Given those, we maximize, using a numerical optimization routine, the concentrated log-likelihood function to find ρ. Given $\hat{\rho}_{ML}$ we can then compute $\hat{\beta}_{ML} = \hat{\beta}_O - \hat{\rho}\hat{\beta}_L$ and $\hat{\sigma}^2_{ML} = [(1/N)(\hat{e}_O - \hat{\rho}\hat{e}_L)'(\hat{e}_O - \hat{\rho}\hat{e}_L)]$.

Under the regularity conditions described for instance in Lee (2004, p.1902-1904), it can be shown that the maximum likelihood estimators have the usual asymptotic properties, including consistency, normality, and asymptotic efficiency.

The asymptotic variance-covariance matrix follows as the inverse of the information matrix. Defining $G_\rho = W(I_N - \rho W)^{-1}$ to simplify notation, we have then:

$AsyVar[\beta', \rho, \sigma^2] =$

(42)

$$\begin{bmatrix} \frac{1}{\sigma^2}X'X & \frac{1}{\sigma^2}(X'G_\rho X\beta)' & 0 \\ \frac{1}{\sigma^2}X'G_\rho X\beta & tr\left[(G_\rho + G'_\rho)G_\rho\right] + \frac{1}{\sigma^2}(G_\rho X\beta)'(G_\rho X\beta) & \frac{1}{\sigma^2}trG_\rho \\ 0 & \frac{1}{\sigma^2}trG_\rho & \frac{N}{2\sigma^4} \end{bmatrix}^{-1}$$

The estimated asymptotic variance-covariance matrix will then be used for asymptotic inference on the parameters under the regularity conditions.

1.6.4 Quasi Maximum Likelihood estimation

The Quasi Maximum Likelihood (QML) estimator is derived from a normal likelihood but the disturbances in the model are not truly normally distributed. The QML estimators are identical to ML estimators, but their asymptotic variance-covariance matrix will take into account non-normality of the disturbance term. Let $\theta = (\beta', \rho, \sigma_\varepsilon)'$ and let θ_0 be the true value of θ. The variance-covariance matrix of $1/\sqrt{(N)}\partial L(\theta_0)/\partial\theta$ is:

$$\mathrm{E}\left(\frac{1}{\sqrt{N}}\frac{\partial L(\theta_0)}{\partial\theta}.\frac{1}{\sqrt{N}}\frac{\partial L(\theta_0)}{\partial\theta'}\right) = -\mathrm{E}\left(\frac{1}{N}\frac{\partial^2 L(\theta_0)}{\partial\theta\partial\theta'}\right) + \Omega_\theta \qquad (43)$$

[11] The reader unfamiliar with spatial econometrics methods can refer to LeSage and Pace (2009). LeSage also provides Matlab routines for estimating such models in his Econometrics Toolbox (http://www.spatial-econometrics.com).

where

$$\Sigma_{\theta,N} = -\,\mathrm{E}\left(\frac{1}{N}\frac{\partial^2 L(\theta_0)}{\partial\theta\partial\theta'}\right)$$
$$= \begin{bmatrix} \frac{1}{\sigma^2}X'X & \frac{1}{\sigma^2}(X'G_\rho X\beta)' & 0 \\ \frac{1}{\sigma^2}X'G_\rho X\beta & tr\left[(G_\rho + G'_\rho)G_\rho\right] + \frac{1}{\sigma^2}(G_\rho X\beta)'(G_\rho X\beta) & \frac{1}{\sigma^2}trG_\rho \\ 0 & \frac{1}{\sigma^2}trG_\rho & \frac{N}{2\sigma^4} \end{bmatrix} \tag{44}$$

and

$$\Omega_{\theta,N} = \begin{bmatrix} 0 & * & * \\ \frac{\mu_3}{\sigma^4}\sum_{i=1}^N G_{ii}x_i & \frac{2\mu_3}{\sigma^4}\sum_{i=1}^N G_{ii}G_i X\beta + \frac{(\mu_4-3\sigma^4)}{\sigma^4}\sum_{i=1}^N G_{ii}^2 & * \\ \frac{\mu_3}{2\sigma^6}\iota'_N X & \frac{1}{2\sigma^6}\left[\mu_3\iota'_N GX\beta + (\mu_4 - 3\sigma^4)\,\mathrm{tr}\,G\right] & \frac{(\mu^4-3\sigma_4)}{4\sigma^8} \end{bmatrix} \tag{45}$$

is a symmetric matrix with $\mu_j = \mathrm{E}(\varepsilon_i^j), j = 2,3,4$, being respectively, the second, third and fourth moments of ε, where G_i is the ith row of G, G_{ij} is the (i,j)th entry of $G = W(I_N - \rho W)^{-1}$, and x_i is the ith row of X. Assumption 6 is sufficient to guarantee that the average Hessian matrix is nonsingular for large enough N. If ε is normally distributed, $\Omega_\theta = 0$. Lee proofs then the following theorem (theorem 3.2, 2004, p.1906):

Theorem 7
Under previous Assumptions $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma_\theta^{-1} + \Sigma_\theta^{-1}\Omega_\theta\Sigma_\theta^{-1})$, where
$\Omega_\theta = \lim_{N\to\infty}\Omega_{\theta,N}$
and
$\Sigma_\theta = -\lim_{N\to\infty}\mathrm{E}\left(\frac{1}{N}\frac{\partial^2 L(\theta_0)}{\partial\theta\partial\theta'}\right)$ which are assumed to exist.

If the disturbance term is normally distributed, then $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma_\theta^{-1})$.

The estimation of the asymptotic variance of $\hat{\theta}$ is trivial. The Σ_θ can be estimated by (**??**) evaluated at $\hat{\theta}$. The Ω_θ can be estimated by (**??**), μ_3 and μ_4 in Ω_θ can be estimated by the third and fourth order empirical moments based on estimated residuals $\hat{\varepsilon}$ (Lee, 2004).

1.7 SEM specification: the spatial error model

Let us consider the following cross section regression model with spatially autocorrelated errors:

$$y = X\beta + \varepsilon \qquad \varepsilon = \lambda W\varepsilon + u \tag{46}$$

where y is the $N \times 1$ vector of the dependent variable, X is the $N \times K$ matrix of the K non-stochastic explanatory variables; β is the $K \times 1$ vector of the associated parameters. In this specification, the $N \times 1$ error vector ε follows a first order SAR process: λ is the spatial autoregressive parameter and W is the $N \times N$ interaction matrix. Finally the disturbance terms u_i in the $N \times 1$ error vector u are independently and identically distributed $(0,\ \sigma_u^2)$.

1.7.1 Reduced form

If $I_N - \lambda W$ is invertible,[12] noting that $\varepsilon = (I_N - \lambda W)^{-1}u$, the reduced form of the SEM specification is obtained as:

$$y = X\beta + (I_N - \lambda W)^{-1}u \tag{47}$$

This reduced form implies that a random shock in a specific cross-sectional unit i does not only affect the value of the dependent variable in the same unit, but also has an impact on the values of dependent variable in all other cross-sectional units through the inverse spatial transformation $(I_N - \lambda W)^{-1}$. This is the so-called *spatial "diffusion" process of random shocks*. The SEM specification shares this property with the SAR model, but does not share the *global spatial multiplier effect* that we highlighted for the SAR process.

Note that this model may also be interpreted as a constrained Spatial Durbin Model:

$$(I_N - \lambda W)y = (I_N - \lambda W)(\beta_0 \iota_N + X\beta) + u \tag{48}$$

$$y = (I_N - \lambda W)\beta_0 \iota_N + \lambda W y + X\beta - \lambda W X\beta + u \tag{49}$$

Using a row-normalized interaction matrix W, one gets:

$$y = \frac{\beta_0}{1-\lambda}\iota_N + \lambda W y + X\beta - \lambda W X\beta + u \tag{50}$$

which is a Spatial Durbin Model:

$$y = \frac{\beta_0}{1-\lambda}\iota_N + \lambda W y + X\beta + WX\gamma + \varepsilon \tag{51}$$

with non linear constraints $\gamma = -\lambda\beta$.

[12] The invertibility conditions are the same as for the SAR model.

1.7.2 Spatial filter

If λ was known, we could filter out spatial autocorrelation from both of y and X and then use the classical regression model as follows:

$$(I_N - \lambda W)y = (I_N - \lambda W)X\beta + u \tag{52}$$

Of course λ is usually unknown and has to be estimated as well as β and $\sigma^2_{u,N}$.

1.7.3 Variance-covariance matrix

Assuming that X is non stochastic, the mathematical expectation of y is trivially $\mathrm{E}(y) = X\beta$. Its variance-covariance matrix is in turn obtained as:

$$\mathrm{V}(y) = \mathrm{V}(\varepsilon) = \sigma^2_u(I_N - \lambda W)^{-1}(I_N - \lambda W)'^{-1} = \sigma^2_u\Omega_\varepsilon(\lambda) \tag{53}$$

This is exactly the result we obtained for the SAR model. Note again that in general, this matrix will be full and it's main diagonal will not be constant. Thus spatial autocorrelation in the error term induces complete heteroscedasticity. This a case of non spherical error term, where OLS estimators are consistent but inefficient.

1.8 Estimation of the SEM model

1.8.1 GLS and FGLS estimation

The GLS estimator is given by:

$$\hat{\beta}_{GLS} = [X'\Omega^{-1}_\varepsilon(\lambda)X]^{-1}X'\Omega^{-1}_\varepsilon(\lambda)y \tag{54}$$

Clearly, this estimator would be BLUE if λ were known, we would have:

$$\mathrm{E}(\hat{\beta}_{GLS}) = \beta \text{ and } \mathrm{V}(\hat{\beta}_{GLS}) = \sigma^2_u[X'\Omega^{-1}_\varepsilon(\lambda)X]^{-1} \tag{55}$$

Of course as λ is unknown, the GLS estimator is not feasible. Nevertheless let's have a closer look at this estimator:

$$\begin{aligned}
\hat{\beta}_{GLS} &= [X'(I_N - \lambda W)'(I_N - \lambda W)X]^{-1}X'(I_N - \lambda W)'(I_N - \lambda W)y \\
\hat{\beta}_{GLS} &= [(X - \lambda WX)'(X - \lambda WX)]^{-1}(X - \lambda WX)'(y - \lambda Wy) \\
\hat{\beta}_{GLS} &= (X'_L X_L)^{-1}X'_L y_L
\end{aligned}$$

where $X_L = (I_N - \lambda W)X = X - \lambda WX$ and $y_L = (I_N - \lambda W)y = y - \lambda Wy$ may be interpreted as spatial counterparts to the Cochrane-Orcutt transformation in Times Series. The GLS estimator is thus the OLS estimator on spatially filtered variables, if λ were known.

Using the feasible GLS estimator requires to have a consistent estimator for λ that could be used to transform the model by the spatial Cochrane-Orcutt procedure and then estimate β by OLS:

$$\begin{aligned} \hat{\beta}_{FGLS} &= [X'\Omega_{\varepsilon}^{-1}(\hat{\lambda})X]^{-1}X'\Omega_{\varepsilon}^{-1}(\hat{\lambda})y \quad (56) \\ \hat{\beta}_{FGLS} &= (\tilde{X}_L'\tilde{X}_L)^{-1}\tilde{X}_L'\tilde{y}_L \quad (57) \end{aligned}$$

where $\tilde{X}_L = X - \hat{\lambda}WX$ and $\tilde{y}_L = y - \hat{\lambda}Wy$.

The OLS estimator for β is consistent in the spatially autocorrelated error model, the errors can then be estimated consistently. However the OLS estimator for λ in the SAR model for the errors is not consistent as we previously showed, but **?** propose a consistent Generalized Method of Moments (GMM) estimator for λ.

1.8.2 Maximum likelihood estimation

Consider again the spatial error model (SEM) under Assumptions 1-4 and under the additional hypothesis of normality of the error term $u \sim N(0, \sigma_u^2 I_N)$:

$$y = X\beta + \varepsilon \qquad \varepsilon = \lambda W\varepsilon + u \quad (58)$$

We then have:

$$\varepsilon \sim N(0, \sigma_u^2\Omega_{\varepsilon}(\lambda)) \text{ and } y \sim N(X\beta, \sigma_u^2\Omega_{\varepsilon}(\lambda)) \quad (59)$$

Observing that $|\sigma_u^2\Omega_{\varepsilon}(\lambda)| = |\sigma_u^2|^N|I - \lambda W|^{-2}$, the log-likelihood function for the SEM model is then obtained as:

$$\begin{aligned} \ln L(\beta', \lambda, \sigma_{\epsilon}^2) &= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\sigma_u^2\Omega_{\varepsilon}(\lambda)| - \frac{1}{2\sigma_u^2}(y - X\beta)'\Omega_{\varepsilon}(\lambda)^{-1}(y - X\beta) \\ \ln L(\beta', \lambda, \sigma_{\epsilon}^2) &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma_u^2) + \ln|I - \lambda W| \quad (60) \\ &\quad -\frac{1}{2\sigma_u^2}(y - X\beta)'\Omega_{\varepsilon}(\lambda)^{-1}(y - X\beta) \end{aligned}$$

Noting that $\Omega_\varepsilon(\lambda)^{-1} = (I - \lambda W)'(I - \lambda W)$, the log-likelihood may then be written as follows:

$$\begin{aligned}
\ln L(\beta', \lambda, \sigma_\epsilon^2) &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma_u^2) + \ln|I - \lambda W| \\
&\quad -\frac{1}{2\sigma_u^2}\left[(I - \lambda W)(y - X\beta)\right]'\left[(I - \lambda W)(y - X\beta)\right] \\
\ln L(\beta', \lambda, \sigma_\epsilon^2) &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma_u^2) + \ln|I - \lambda W| \\
&\quad -\frac{1}{2\sigma_u^2}\left[(I - \lambda W)y - (I - \lambda W)X\beta\right]'\left[(I - \lambda W)y - (I - \lambda W)X\beta\right] \\
\ln L(\beta', \lambda, \sigma_\epsilon^2) &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma_u^2) + \ln|I - \lambda W| \\
&\quad -\frac{1}{2\sigma_u^2}\left[y_L - X_L\beta\right]'\left[y_L - X_L\beta\right]
\end{aligned} \tag{61}$$

where $X_L = (I_N - \lambda W)X$ and $y_L = (I_N - \lambda W)y$. Suppose now that λ is known, the first order conditions for β and σ_u^2 are:

$$\frac{\partial \ln L}{\partial \beta'} = -\frac{1}{2\sigma_u^2}(X_L'y_L + 2\beta X_L'X_L) = 0 \tag{62}$$

$$\frac{\partial \ln L}{\partial \sigma_u^2} = -\frac{N}{\sigma_u^2} + \frac{1}{\sigma_u^4}(y_L - X_L\beta)'(y_L - X_L\beta) = 0 \tag{63}$$

It follows that the maximum likelihood estimators for β and σ_u^2, given λ, are obtained as:

$$\hat{\beta}_{ML}(\lambda) = (X_L'X_L)^{-1}X_L'y_L \tag{64}$$

$$\hat{\sigma}^2_{ML}(\lambda) = \frac{1}{N}(y_L - X_L\beta)'(y_L - X_L\beta) \tag{65}$$

Substitution of (**??**) and (**??**) in the log-likelihood function (**??**) yields a concentrated log-likelihood function as a non-linear function of a single parameter λ:

$$\ln L = -\frac{N}{2}[1 + \ln(2\pi)] - \frac{N}{2}\ln\left(\frac{\hat{u}'\hat{u}}{N}\right) + \sum_{i=1}^{N}\ln(1 - \lambda\omega_i) \tag{66}$$

where $\hat{u} = y_L - X_L\hat{\beta}_{ML}(\lambda)$, then $\hat{u}'\hat{u} = y_L'Y_L - y_L'X_L(X_L'X_L)^{-1}X_L'y_L$, where $y_L = y - \lambda Wy$ and $X_L = X - \lambda WX$ are the spatially filtered variables. A maximum likelihood estimate for λ is obtained from a numerical optimization of the concentrated log-likelihood function (**??**).

The estimation procedure is hence more complicated than for the SAR model since $\hat{u}$ in the concentrated log-likelihood function is indirectly also a function of λ as $\hat{\beta}$ is obtained for a value for λ. Therefore a one-time optimization of the concentrated log-likelihood function with respect to λ does not suffice to obtain ML estimates of all the unknown parameters. An iterative approach is thus required. This would essentially alternate back and forth between the estimation of λ conditional upon a vector of residuals u generated for a value of β, and an estimation of β and σ^2 conditional upon a value for λ until numerical convergence is obtained. This procedure can be described as follows. We first regress y on X by OLS and compute the initial set of residuals $\hat{u} = y - X\beta_{OLS}$. Given $\hat{u}$, we then find λ that maximizes the concentrated log-lokelihood function. Given λ, we carry out FGLS that yields $\hat{\beta}_{FGLS}$. We then compute a new set of residuals $\hat{u} = y - X\hat{\beta}_{FGLS}$. If the numerical convergence criterion is met, that is, if values for both the residuals and $\hat{\beta}_{FGLS}$ fail to change from one iteration to the next, given $\hat{u}$ and $\hat{\lambda}_{ML}$ we compute $\hat{\sigma}^2_{ML}$, else we go back to the maximization of the concentrated log-likelihood function to get a new λ.

As for the SAR model, under the regularity conditions described for instance in Lee (2004, p.1902-1904), it can be shown that the maximum likelihood estimators have the usual asymptotic properties, including consistency, normality, and asymptotic efficiency.

The asymptotic variance-covariance matrix follows as the inverse of the information matrix. Defining $G_\lambda = W(I - \lambda W)^{-1}$ to simplify notation, we have then:

$\text{AsyVar}[\beta', \lambda, \sigma^2] =$

(67)

$$\begin{bmatrix} \frac{1}{\sigma^2} X_L' X_L & 0 & 0 \\ 0 & G_\lambda^2 + \text{tr}\, G_\lambda' G_\lambda & \frac{1}{\sigma^2} tr G_\lambda \\ 0 & \frac{1}{\sigma^2} tr G_\lambda & \frac{N}{2\sigma^4} \end{bmatrix}^{-1}$$

Due to the block-diagonal form of the asymptotic variance matrix, knowledge of the precision of λ does not affect the precision of the β estimates. Consequently, if the latter is the primary interest, the complex inverse and trace expressions in (**??**) need not be computed.

2 Spatial or interaction multipliers

2.1 The spatial Durbin Model

Let us consider the general case of the Spatial Durbin Model SDM:

$$y_i = \beta_0 + \rho \sum_{j=1}^{N} w_{ij} y_j + \sum_{k=1}^{K} x_{ik}\beta_k + \sum_{k=1}^{K} w_{ij} z_{ik}\gamma_k + \varepsilon_i \text{ for } i = 1, ..., N \quad (68)$$

or in matrix form:

$$y = \beta_0 \iota_N + \rho W y + X\beta + WZ\gamma + \varepsilon \quad (69)$$

where y is the $N \times 1$ vector of the dependent variable, Wy is the spatially lagged dependent variable, ι_N is the $N \times 1$ unit vector and X is the $N \times K$ matrix of the explanatory variables; β is the $K \times 1$ vector of the associated coefficients. Note that Z may include explanatory variables not included in X, supposed affecting y only through their spatial lag: Z may then be a $N \times (K + M)$ matrix $Z = [X\ \widetilde{Z}]$ where X is the previous $N \times K$ matrix of the explanatory variables and $\widetilde{Z}$ is the $N \times M$ matrix of extra explanatory variables not included in X which are supposed to affect y only through their spatial lag. WZ is the $(N \times (K + M))$ matrix of the spatially lagged explanatory variables and γ is the $(K + M) \times 1$ vector of the associated coefficients.

Suppose for simplicity that Z only includes X, the model may then be written as follows:

$$y = \beta_0 \iota_N + \rho W y + X\beta + WX\gamma + \varepsilon \quad (70)$$

or

$$y = \beta_0 \iota_N + \rho W y + \sum_{k=1}^{K} (I_N \beta_k + W\gamma_k) X_k + \varepsilon \quad (71)$$

and the reduced form is then:

$$y = (I - \rho W)^{-1} \iota_N \beta_0 + \sum_{k=1}^{K} (I - \rho W)^{-1} (I_N \beta_k + W\gamma_k) X_k + (I - \rho W)^{-1} \varepsilon \quad (72)$$

Note that if W is row normalized then $W^q \iota_N = \iota_N$ for $q \geq 0$ and we have:

$$\begin{aligned} (I_N - \rho W)^{-1} \iota_N &= (I_N + \rho W + \rho^2 W^2 + \ldots) \iota_N \\ &= (1 + \rho + \rho^2 + \ldots) \iota_N = \frac{1}{1 - \rho} \iota_N \end{aligned} \quad (73)$$

$$y = \frac{\beta_0}{1-\rho}\iota_N + \sum_{k=1}^{K}(I_N - \rho W)^{-1}(I_N\beta_k + W\gamma_k)X_k + (I_N - \rho W)^{-1}\epsilon \quad (74)$$

Let us take the partial derivatives of y relative to X_k for $k = 1, \ldots, K$:

$$\begin{aligned}\frac{\partial y}{\partial X_k'} = S_k(W) &= (I_N - \rho W)^{-1}(I_N\beta_k + W\gamma_k) \\ &= (I_N + \rho W + \rho^2 W^2 + \ldots)(I_N\beta_k + W\gamma_k)\end{aligned}$$

where $(I_N - \rho W)^{-1}$ is the so-called *global spatial multiplier* or *global interaction multiplier*.

Recall that $(I_N - \rho W)^{-1} = I_N + \rho G_\rho$, where $G_\rho = W(I_N - \rho W)^{-1}$. Therefore those partial derivatives could also be represented by:

$$\frac{\partial y}{\partial X_k'} = S_k(W) = (I_N + \rho G_\rho)(I_N\beta_k + W\gamma_k) \quad (75)$$

which yields an alternative decomposition of partial derivatives of y relative to X_k for $k = 1, \ldots, K$. Let us define, as the *impact matrix* associated to the k^{th} explanatory variable, the $N \times N$ matrix $S_k(W) = (I_N - \rho W)^{-1}(I_N\beta_k + W\gamma_k)$. $S_k(W)$ is a $N \times N$ full matrix whose elements are:

$$S_k(W) = \begin{pmatrix} S_k(W)_{11} & S_k(W)_{12} & \ldots & S_k(W)_{1N} \\ S_k(W)_{21} & S_k(W)_{22} & & S_k(W)_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ S_k(W)_{N1} & S_k(W)_{N2} & \ldots & S_k(W)_{NN} \end{pmatrix} \quad (76)$$

The partial derivatives of y_i relative to x_{ik} or x_{jk} for $i, j = 1, \ldots, N$, $j \neq i$ and for $k = 1, \ldots, K$ are then:

$$\frac{\partial y_i}{\partial x_{ik}} = S_k(W)_{ii}, \qquad \frac{\partial y_i}{\partial x_{jk}} = S_k(W)_{ij} \quad (77)$$

In general $S_k(W)_{ii} \neq 0$ and $S_k(W)_{ij} \neq 0$ for $i, j = 1, \ldots, N$, $j \neq i$ and for $k = 1, \ldots, K$. The diagonal elements of this matrix, $\text{diag}(S_k(W))$, represent the *direct impacts* including "own spillover" effects, which are inherently heterogenous in presence of spatial autocorrelation due to differentiated interaction terms in the W matrix. This type of heterogeneity is called *interactive heterogeneity*, in opposition to standard individual heterogeneity in panel data models (Debarsy and Ertur, 2010). The off-diagonal elements of the impact matrix represent *indirect impacts*: they are collected in the matrix $Q_k(W) = S_k(W) - \text{diag}(S_k(W))$.

Note that the own derivative for country i includes the feedback effects where country i affects country j and country j also affects country i as well as longer paths which might go from country i to j to k and back to i. The magnitude of those direct effects will depend on: (1) the degree of interaction between countries, which is governed by the W matrix, (2) the parameter ρ, measuring the strength of spatial correlation between countries and (3) the parameters β_k and γ_k.

Moreover, considering column j, we note that a variation Δx_{jk} of the k^{th} explanatory variable in spatial unit j differently affects each of the spatial units of the sample:

$$\begin{pmatrix} S_k(W)_{1j} \\ S_k(W)_{2j} \\ \vdots \\ S_k(W)_{jj} \\ \vdots \\ S_k(W)_{Nj} \end{pmatrix}$$

The sum down the j^{th} column yields the total impact on y_i $(i = 1, ..., N)$ for all the N spatial units of the sample of a change of x_{jk} in spatial unit j. The total impacts, direct and indirect, from each of the units $j = 1, ..., N$ are then collected in the row vector $\iota'_N S_k(W)$. However it may be of interest to distinguish direct and indirect effects in applied papers where the direct impacts are actually higher than each of the indirect impacts and potentially higher than their sum (if the impat matrix is strictly diagonally dominant). The total indirect impacts from each of the units $j = 1, ..., N$ may then be usefully collected in the row vector $\iota'_N Q_k(W)$.

Considering row i, we note that an identical variation ΔX_k of the k^{th} explanatory variable across all the units of the sample differently affects spatial unit i:

$$\begin{pmatrix} S_k(W)_{i1} & S_k(W)_{i2} & \dots & S_k(W)_{ii} & \dots & S_k(W)_{iN} \end{pmatrix}$$

The sum across the i^{th} row represents the total impact on y_i of an identical change of x_{jk} $(j = 1, ..., N)$ across all the N spatial units in the sample. The total impacts, direct and indirect, on each of the units $i = 1, ..., N$ are then collected in the column vector $S_k(W)\iota_N$. Again note that the total indirect impacts on each of the units $i = 1, ..., N$ may be collected in the row vector $Q_k(W)\iota_N$.

Of course, in both cases the corresponding elements of main diagonal of $S_k(W)$ may be taken into account or not in those sums, depending on the inclusion or the exclusion of the direct effects in those computations.

Given the complexity and the amount of the information available in such impact matrices, LeSage and Pace (2009) suggest some useful summary scalar measures. The average direct impact, including feedback effects, is defined as:

$$N^{-1}\,\mathrm{tr}(S_k(W))$$

whereas the average global impact is defined as:

$$N^{-1}\iota_N' S_k(W)\iota_N$$

where ι_N is the $N \times 1$ sum vector. Finally the average indirect impact is, by definition, the difference between the average global impact and the average direct impact:

$$N^{-1}\iota_N' S_k(W)\iota_N - N^{-1}\,\mathrm{tr}(S_k(W)) = N^{-1}\iota_N' Q_k(W)\iota_N$$

2.2 Special cases

- **The spatial autoregressive model (SAR)**
  Let us now consider the specification corresponding to the SAR model, excluding the spatial lags of the explanatory variables from the specification ($\rho \neq 0$, $\beta \neq 0$ and $\gamma = 0$), the partial derivatives of y relative to X_k for $k = 1, \ldots, K$ are then simply:

  $$\frac{\partial y}{\partial X_k'} = S_k(W) = (I_N - \rho W)^{-1} I_N \beta_k = (I_N + \rho W + \rho^2 W^2 + \ldots) I_N \beta_k \tag{78}$$

  The diagonal elements of this *impact matrix* represent the direct effects including "own spillover" effects whereas the off-diagonal terms represent indirect effects. Note that the magnitude of pure feedback effects are then given by $S_k(W)_{ii} - \beta_k$, where β_k could be interpreted as representing the direct impact of the explanatory variable if there was no spatial autocorrelation, i.e. if ρ was equal to zero. Note also that in this special case, using a row-normalized interaction matrix W, since then $(I_N - \rho W)^{-1}\iota_N = \frac{1}{1-\rho}\iota_N$, the total impacts on each of the units $i = 1, ..., N$ collected in the column vector $S_k(W)\iota_N$ may be written as follows:

  $$S_k(W)\iota_N = (I_N - \rho W)^{-1}\beta_k \iota_N = \frac{\beta_k}{1-\rho}\iota_N$$

  The average global impact of a variation of the k^{th} explanatory variable simplifies then to:

  $$N^{-1}\iota_N' S_k(W)\iota_N = N^{-1}\frac{\beta_k}{1-\rho}\iota_N'\iota_N = \frac{\beta_k}{1-\rho} \tag{79}$$

- **The cross regressive model or SLX model**
  Consider now the following simpler model, which does not include the spatially lagged endogenous variable ($\rho = 0$) but includes exogenous variables together with spatially lagged exogenous variables ($\beta \neq 0$ and $\gamma \neq 0$) which may be estimated by OLS under the usual set of assumptions:

$$y = \beta_0 \iota_N + \sum_{k=1}^{K} (I_N \beta_k + W\gamma_k) X_k + \varepsilon \tag{80}$$

  Let us take the partial derivatives of y relative to X_k for $k = 1, \ldots, K$:

$$\frac{\partial y}{\partial X_k'} = S_k(W) = I_N \beta_k + W\gamma_k \tag{81}$$

  Again, the diagonal elements of this *impact matrix* represent the direct effects (as in the standard OLS model), whereas the off-diagonal terms represent indirect effects. If β_k can be interpreted as the direct effect as in the standard a-spatial setting, one must be cautious when interpreting γ_k as the spillover effect. The individual impacts actually depend on the interaction structure of the "spatial weight matrix" W.

  For a simple row-standardized first order contiguity matrix, a very special case indeed, the individual impact of a variation of X_k in spatial unit j on y in spatial unit i is $\frac{w_{ij}}{\sum_j w_{ij}} \gamma_k$. Since W is a binary row-standardised matrix, the impact is constant in each row and the sum of each of the rows of $W\gamma_k$ is identical and equal to γ_k. This means that if one increases X_k by the same amount ΔX_k in each of the "neighbors" of unit i, then the total impact on y in unit i is $\gamma_k \Delta X_k$ and is identical for all units. But as a row-standardised matrix is no more symetric, the sum as each of the columns will not be identical for all units. The sum of a column can be interpreted in turn as the total impact of an increase of X_k in unit j on y for all of the units of the sample. These sums will in general be different for each of the columns, but they are still representing spillover effects, different from the identical spillover effect represented by the sums of the rows.

  Furthermore all of those results bear on the row-standardization of the W matrix which is just a standardization method among others and perhaps not the best one (see Kelejian and Prucha, 2010). Contrary to the Spatial Durbin Model, here the impacts are only *local*, coming from neighboring units, W playing the role of a *local spatial multiplier* or *local interaction multiplier*.

- **The spatially autocorrelated error model (SEM)**
  Let us finally consider the following regression model with spatially autocorrelated errors:

$$y = \beta_0 \iota_N + X\beta + \varepsilon \qquad \varepsilon = \lambda W\varepsilon + u \tag{82}$$

  which may be written as follows:

$$y = \beta_0 \iota_N + X\beta + (I_N - \lambda W)^{-1}u \tag{83}$$

  Let us take the partial derivatives of y relative to X_k for $k = 1, \ldots, K$:

$$\frac{\partial y}{\partial X_k'} = S_k(W) = I_N\beta_k \tag{84}$$

  which is exactly the same result as in the standard a-spatial regression model, where there are no spatial spillovers. Note that this model may also be interpreted as a constrained Spatial Durbin Model:

$$(I_N - \rho W)y = (I_N - \lambda W)(\beta_0 \iota_N + X\beta) + u \tag{85}$$

$$y = (I_N - \rho W)\beta_0 \iota_N + \lambda Wy + X\beta - \lambda WX\beta + u \tag{86}$$

  As before, using a row-normalized interaction matrix W, one gets:

$$y = \frac{\beta_0}{1-\lambda}\iota_N + \lambda Wy + X\beta - \lambda WX\beta + u \tag{87}$$

  which is the Spatial Durbin Model:

$$y = \frac{\beta_0}{1-\lambda}\iota_N + \lambda Wy + X\beta + WX\gamma + \varepsilon \tag{88}$$

  with non linear constraints : $\gamma = -\lambda\beta$.

2.3 Impact of a random shock

2.3.1 The spatial Durbin model

Reconsider the SDM model, in matrix form:

$$y = \beta_0 \iota_N + \rho Wy + X\beta + WX\gamma + \varepsilon \tag{89}$$

the reduced form is then:

$$y = (I - \rho W)^{-1}\iota_N\beta_0 + (I - \rho W)^{-1}(X\beta + WX\gamma) + (I - \rho W)^{-1}\varepsilon \tag{90}$$

This specification also implies that a shock affecting one unit propagates to all the other units of the sample through the spatial transformation $(I - \rho W)^{-1}$. Note that the derivative of y with respect to ε is:

$$\frac{\partial y}{\partial \varepsilon'} = (I - \rho W)^{-1} = (I_N + \rho W + \rho^2 W^2 + \ldots) \tag{91}$$

which is the so-called *global spatial multiplier* or the *interaction multiplier*. Let us define, as the *impact matrix* associated to the random shock, the $(N \times N)$ matrix $U(W) = (I_N - \rho W)^{-1}$. $U(W)$ is a $(N \times N)$ full matrix whose elements are:

$$U(W) = \begin{pmatrix} U(W)_{11} & U(W)_{12} & \ldots & U(W)_{1N} \\ U(W)_{21} & U(W)_{22} & & U(W)_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ U(W)_{N1} & U(W)_{N2} & \ldots & U(W)_{NN} \end{pmatrix} \tag{92}$$

The partial derivatives of y_i relative to ε_i or ε_j for $i, j = 1, \ldots, N$, $j \neq i$ are then:

$$\frac{\partial y_i}{\partial \varepsilon_i} = U(W)_{ii}, \qquad \frac{\partial y_i}{\varepsilon_j} = U(W)_{ij} \tag{93}$$

In general $U(W)_{ii} \neq 0$ and $U(W)_{ij} \neq 0$ for $i, j = 1, \ldots, N$, $j \neq i$.

The diagonal elements of this matrix, $\text{diag}(U(W))$, represent the *direct impacts* of a unitary random shock including "own spillover" effects. Again those impacts are heterogenous due to differentiated interaction terms in the W matrix. The off-diagonal elements of the impact matrix represent *indirect impacts* of the unitary random shock.

Note also that the own derivative for unit i includes as previously the feedback effects where the unitary random shock on unit i affects unit j and unit j also affects unit i as well as longer paths which might go from unit i to j to k and back to i. The magnitude of those direct effects will now depend on: (1) the degree of interaction between countries, which is governed by the W matrix and (2) the parameter ρ, measuring the strength of spatial correlation between units. The magnitude of pure feedback effects are given by $U(W)_{ii} - 1$.

Considering column j, we note that an unitary random shock in spatial unit j differently affects each of the spatial units of the sample. It represents the emission side of the spatial diffusion process. Considering row i, we note that unitary random shocks across all the units of the sample differently affects spatial unit i. It represents the reception side of the spatial diffusion process.

2.3.2 Simulating the impacts of heterogenous random shocks

Let a_i be the magnitude of the shock affecting unit i and $\hat{\varepsilon}^i$ be the $(N \times 1)$ vector containing the estimated error of the model with a shock on the error term in unit i:

$$\hat{\varepsilon}^i = (\hat{\varepsilon}_1, ..., \hat{\varepsilon}_i + a_i, ..., \hat{\varepsilon}_N)' \tag{94}$$

Therefore, the $(N \times 1)$ vector y^{i*} of the simulated dependent variable with a shock in unit i is:

$$y^{*i} = (I - \hat{\rho}W)^{-1}(\iota_N \hat{\beta}_0 + X\hat{\beta} + WX\hat{\gamma}) + (I - \hat{\rho}W)^{-1}\hat{\varepsilon}^i \tag{95}$$

$$y^{*i} = (I - \hat{\rho}W)^{-1}\widetilde{X}\hat{\delta} + (I - \hat{\rho}W)^{-1}\hat{\varepsilon}^i \tag{96}$$

where $\widetilde{X} = [\iota_N \;\; X \;\; WX]$, $\hat{\delta} = [\hat{\beta}_0 \;\; \hat{\beta} \;\; \hat{\gamma}]'$ and $\hat{\beta}_0$, $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\rho}$ are the Maximum Likelihood estimates of the unknown parameters in equation (**??**).

Furthermore, let Y^* be the $(N \times N)$ matrix where each column i represents the simulated dependent variable for all units in the sample with a shock in unit i:

$$Y^* = [y^{*1} \; ... \; y^{*n}] = (I - \hat{\rho}W)^{-1}[\widetilde{X}\hat{\delta} \; ... \; \widetilde{X}\hat{\delta}] + (I - \hat{\rho}W)^{-1}\hat{\varepsilon}^* \tag{97}$$

where $\hat{\varepsilon}^* = [\hat{\varepsilon}^1 \; ... \; \hat{\varepsilon}^N]$ is a $(N \times N)$ matrix. Given the definition of $\hat{\varepsilon}^i$, the matrix $\hat{\varepsilon}^*$ can be rewritten as follows:

$$\hat{\varepsilon}^* = \begin{bmatrix} \hat{\varepsilon}_1 + a_1 & \hat{\varepsilon}_1 & \dots & \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 & \hat{\varepsilon}_2 + a_2 & \dots & \hat{\varepsilon}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\varepsilon}_N & \hat{\varepsilon}_N & \dots & \hat{\varepsilon}_N + a_N \end{bmatrix}$$

$$\hat{\varepsilon}^* = \iota_N' \otimes \hat{\varepsilon} + A \tag{98}$$

where ι_N is the unit vector of dimension N and A is a diagonal matrix of order N, whose i^{th} diagonal element corresponds to a_i. Therefore, we may write:

$$Y^* = (I - \hat{\rho}W)^{-1}(\iota_N' \otimes \widetilde{X}\hat{\delta}) + (I - \hat{\rho}W)^{-1}(\iota_N' \otimes \hat{\varepsilon} + A) \tag{99}$$

which may be also written:

$$Y^* = \iota_N' \otimes (I - \hat{\rho}W)^{-1}\widetilde{X}\hat{\delta} + \iota_N' \otimes (I - \hat{\rho}W)^{-1}(\hat{\varepsilon} + A) \tag{100}$$

Finally, we compute the impact of a shock on unit i on all values of the dependent variable by calculating the difference between the simulated dependent

variable Y^* and the matrix of observed dependent variable $Y = \iota'_N \otimes y$ with $y = (I - \hat{\rho}W)^{-1}(\widetilde{X}\hat{\delta} + \hat{\varepsilon})$:

$$Y^* - Y = (I - \hat{\rho}W)^{-1}A \tag{101}$$

Note that when $A = I$ we retrieve the previous result on the impact of a unitary random shock. The implementation of shocks different from unity and different from each others just requires the post-multiplication of the impact matrix associated to the random shock by the matrix A.

2.3.3 Special cases

The SAR model, with $\gamma = 0$ is a straightforward special case of the previous development. The pure SAR model with $\beta_0 = \beta = \gamma = 0$ is also a trivial special case.

It must be stressed that the spatial diffusion of a random shock is the only property of the spatially autocorrelated error model (SEM), in which there are otherwise no spatial spillovers (see **?**, **?**, for some examples).