CITY AND REGIONAL PLANNING 870.03

Notes on Spatial Econometric Models

Philip A. Viton

March 2, 2010

Contents

1 Introduction

Following Bivand, Pebesma and Gómez-Rubio (2008) we can distinguish three types of spatial analysis:

1. Spatial point processes: we observe a (typically complete) set of data points in space. We ask whether the data exhibits any spatial clustering patterns, or whether it exhibits what is referred to as Complete Spatial Randomness (CSR). For example, consider a pattern of the locations of cancers in an urban area: are they random, or are they clustered (and if, so, is there any factor which is associated with the clustering)?

2. Geostatistical data: we observe data at a typically limited set of spatial points. Based on these observations, we are interested in interpolating the data to unobserved points. For example, we observe air quality at a set of monitoring stations. What does this tell us about air quality in the region as a whole?

3. Areal data: data is observed distributed into pre-defined spatial regions. A typical question is whether what happens in one region is influenced by what happens in other regions. For example: we observe crime data by state. Does the crime rate in one state depend on the rate in other (neighboring) states?

This note discusses only the third problem, in the context of several standard spatial econometric models. Our setting is a dataset of spatial regions: these could be countries, states, census tracts, zip codes etc. For the most part we shall assume that our dataset contains a single observation on each region (ie, an observation at a single point in time), comprising a spatial cross-section. We suppose that there are R regions in the dataset.

2 Spatial Autocorrelation

The essence of spatial analysis is that "space matters", ie what happens in one region is related to what happens in neighboring regions. This has been made more precise in what Tobler (1979) refers to as the First Law of Geography: "Everything is related to everything else, but closer things more so". One way to approach this is via the notion of spatial autocorrelation. According to Anselin and Bera (1998),

> Spatial autocorrelation can be loosely defined as the coincidence of value similarity with locational similarity. In other words, high or low values for a random variable tend to cluster in space (positive spatial autocorrelation) or locations tend to be surrounded by neighbors with very dissimilar values (negative spatial autocorrelation). Of the two types of spatial autocorrelation, positive autocorrelation is by far the more intuitive. Negative spatial autocorrelation implies a checkerboard pattern of values and does not always have a meaningful substantive interpretation.

Formally, if y_i and y_j are realizations of a random variable y indexed by spatial locations, then we have spatial autocorrelation if

$$\mathsf{Corr}(y_i, y_j) = \mathsf{E}(y_i y_j) - \mathsf{E}(y_i)\mathsf{E}(y_j) \neq 0$$

The problem that this raises can be seen from two perspectives. First, if we draw a sample of R locations from a spatially autocorrelated random process, then effectively we have a sample size of *one*. (Compare a single draw from a bivariate normal density: we get two values, but there is still only one draw). Unless we have a panel-data setting, there is effectively no way to increase that sample size. Secondly, the covariance (correlation) matrix for our sample is $R \times R$, and clearly it will be impossible to estimate that many terms. The only possibility is to impose some *a priori* structure on the problem to make it manageable. One obvious way to do this is to assume some systematic pattern of spatial covariances (autocorrelation), parsimoniously parametrized.

Spatial autocorrelation is something like temporal autocorrelation, but more complicated. The reason is that temporal autocorrelation can only go one way: what happens at one time can be influenced only what has happened in the past. But spatial autocorrelation can potentially go in any direction: it is like saying that what happens at any one point in time can be influenced by both the past

and the future. For this reason, we cannot simply transfer models of temporal autocorrelation to the spatial context.

3 Spatial Neighbors and Weights

3.1 Spatial neighbors

The basis for most models is an indicator of whether one region is a spatial neighbor of another; or equivalently, which regions are neighbors of a given region. This is a square symmetric $R \times R$ matrix with (i, j) element equal to 1 if regions i and j are neighbors of one another (or more generally, are spatially related), and zero otherwise. By convention, the diagonal elements of this "spatial neighbors" matrix are set to zero. As LeSage (1998) points out, there is an embarrassingly large number of ways to construct such a matrix. These include:[1]

- Linear contiguity: i and j are neighbors if they share (part of) a common eastern or western border.[2]
- Rook contiguity (so called after the movement of the chess piece): two regions are neighbors if they share (part of) a common border (on any side). In implementation, this is usually based on a small "snap distance" and says that two regions share a common border if that border is longer than the snap distance.
- Bishop continuity: two regions are spatial neighbors if they meet at a "point". This is the spatial analog of two elements of a graph meeting at a vertex. In implementation, one says that two regions are neighbors in this sense if their common border is shorter than the "snap distance" referred to above.
- Queen contiguity: this is the union of Rook and Bishop contiguity. Two regions are neighbors in this sense if they share any part of a common border, no matter how short.

[1] There is a good review of the possibilities in Kelejian and Robinson (1995, appendix A). See also LeSage and Pace (2009): the LeSage reference in the text is an early draft of this book.

[2] Can anyone explain why it is ever interesting to privilege the east/west borders over the north/south ones in a land/spatial context? I don't see it.

One can go further and define "second order" measures of contiguity: these would count as neighbors regions sharing a border with a first-order neighbor according to each of the criteria listed (so, we might have second-order rook neighbors, etc). Another approach is distance-based, and considers two regions as neighbors if, for example, their population-weighted centroids are within some given distance of one another. This can obviously be expanded in several ways: one could use different distances, or different weights for computing the centroids.

It is important not to proceed too mechanically here, but rather to consider neighbors in light of the practical problem being studied. For example (using a case mentioned by LeSage), suppose that two zip code regions meet only at a very short border (effectively, at a point); but that one is considered as a bedroom community for the other. Then in most cases one would not want to use a rook measure, because the two regions would not turn out to be neighbors. An exception might be where there is no direct transportation route between the two: for example, they are separated be a river, and the nearest bridge involves gong through a third region. In this case rook contiguity might be exactly what is wanted. It all depends on the context.

One can construct spatial neighbors matrices "by hand", but of course this is tedious and error prone. The spatial analysis software described in section 8 can construct the matrices for you, given, for example GIS data on the boundaries of the regions, as represented in an ESRI shapefile.

.

3.2 Spatial weights

While in principle one could use spatial neighbors matrices directly, in practice one generally uses slightly transformed matrices, usually referred to as spatial weights matrices. The most common transformation is called "row-standardization" in which the rows of the neighbors matrix are made to sum to unity. Let $\tilde{W}$ with elements $\tilde{w}_{ij}$ be a spatial neighbors matrix. To row-standardize this, we divide each element in a row by the sum of the elements in the row. Thus a spatial weights matrix W, with element w_{ij} is defined by

$$w_{ij} = \tilde{w}_{ij} / \sum_j \tilde{w}_{ij}$$

(or, slightly more precisely, in light of the fact that region i may have no neighbors, ie may be an *island*, $W_{ij} = \tilde{w}_{ij}/\max(1, \sum \tilde{w}_{ij})$.[3] The matrix W is also referred to as a *row-stochastic* matrix, since, if there are no islands, each element is between zero and 1, and the rows sum to one, like probabilities. Note that, unlike the neighbors matrix, the spatial weights matrix is no longer symmetric.

To see why row-standardization is intuitively attractive, consider any observation x, an $R \times 1$ vector, and look at Wx. Since W (and $\tilde{W}$) are $R \times R$ and x is $R \times 1$, Wx is the $R \times 1$ vector $\tilde{W}x$, with entry (row) k as the sum of those elements which are neighbors of region k, divided divided by $\sum \tilde{w}_{kj}$, the count (total number) of its the neighbors. Thus the k-th entry of Wx is an average of x, where the average is taken only over the neighbors of region k. (Recall that the diagonal elements of $\tilde{W}$ are zero, which means that the data for region k itself does not enter this computation).

Example: Consider an $R = 4$ region system, with the following neighbor relations:

Region	**Neighbors**
1	2, 3, 4
2	1, 3
3	1, 2, 4
4	1, 3

Then

$$\tilde{W} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

In this system, region 1 has three neighbors, and the rest have two, so $\sum_k \tilde{w}_{jk} = (3, 2, 3, 2)$ and the row-standardized weights matrix is:

$$W = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

[3] In fact, some computational routines can fail if the data contains islands, and it is best (where possible) to use a dataset without them. That is why one often sees analyses conducted at the level of the 48 contiguous US states plus the District of Columbia, excluding Alaska and Hawaii.

Now if $x = (x_1, x_2, x_3, x_4)'$, then

$$Wx = \begin{pmatrix} \frac{1}{3}(x_2 + x_3 + x_4) \\ \frac{1}{2}(x_1 + x_3) \\ \frac{1}{3}(x_1 + x_2 + x_4) \\ \frac{1}{2}(x_1 + x_3) \end{pmatrix}$$

which is indeed the vector of the average x over the neighbors of each region. Note that in the case of a spatially weighted disturbance term, say u, Wu will incorporate the desired "backward and forward" autocorrelation that distinguishes spatial from temporal autocorrelation.

Another possibility is what has been referred to as distance-based contiguity: let d_{ij} be the distance between (centroids of) regions i and j. We define

$$\tilde{w}_{ij} = 1 \quad \text{if } d_{ij} \ < d$$

and zero otherwise, for a pre-specified d. In other words, regions i and j are spatial neighbors when they are within d distance units of one another. Yet another possibility starts by defining b_{ij} as the proportion of the length of the border of region i (ie, its perimeter) that is shared with region j, and then takes

$$\tilde{w}_{ij} = \frac{b_{ij}^{\alpha}}{d_{ij}^{\beta}}$$

The two parameters α and β can be either pre-specified (one possibility is to take $\alpha = 1$ and $\beta = 2$, leading to a gravity-model formulation) or estimated: in the latter case we clearly have a non-linear problem. In a non-regional application to a problem studied in sociology or political science, we could take $\tilde{w}_{ij} = 1$ if individuals i and j belong to the same social network, and study their behavior on that assumption. In practice, most regional-type research begins with the simple $0/1$ row-standardized contiguity matrix.

4 Is there a spatial problem?

The first question an analyst must ask is: does the data exhibit spatial autocorrelation or not? If not, then clearly we can be content with the standard non-spatial models we have been studying so far. But if spatial autocorrelation is present, we will need to formulate more complicated models.

4.1 Moran's Test

The most common test for the existence of spatial autocorrelation is due to Patrick Moran, and is usually referred to as Moran's-I.[4] The statistic is defined for a particular data (or residual) vector x by

$$I = \frac{R}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

where $\bar{x}$ is the mean of x.[5] Moran's statistic has expectation

$$\mathsf{E}(I) = \frac{-1}{R-1}$$

and there is a complicated expression for the variance:

$$\mathsf{V}(I) = \frac{RS_4 - S_3 S_1 (1-2R)}{(R-1)(R-2)(R-3)(\sum_i \sum_j w_{ij})^2}$$

where

$$\begin{aligned}
S_1 &= \frac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2 \\
S_2 &= \sum_i \left(\sum_j w_{ij} + \sum_j w_{ji} \right)^2 \\
S_3 &= \frac{R^{-1} \sum_i (x_i - \bar{x})^4}{\left(R^{-1} \sum_i (x_i - \bar{x})^2 \right)^2} \\
S_4 &= (R^2 - 3R + 3) S_1 - R S_2 + 3 (\sum_i \sum_j w_{ij})^2
\end{aligned}$$

It can be shown that under the null hypothesis of *no spatial autocorrelation*, Moran's statistic is asymptotically normal. So the statistic

$$I^* = \frac{I - \mathsf{E}(I)}{\sqrt{\mathsf{V}(I)}}$$

[4] Note that here I obviously does not refer to the identity matrix. There is usually no problem in keeping the two separate, though if you work with matrix-based definitions of Moran's test, there's room for ambiguity.

[5] For the special case where $\bar{x} = 0$ (this includes the case of the OLS residuals) and where the spatial weights matrix is row-stochastic (so $\sum_j w_{ij} = 1$ for all i) the Moran statistic takes the simple form $I = x'Wx/x'x$.

is asymptotically standard normal (under the null), and we can use it just like a p-value.

A few points to bear in mind about this:

- The normality is an asymptotic result. One should be wary about this in a setting where the number of regions is small. On the other hand, Anselin and Florax (1995) has a simulation-based discussion of the properties of the Moran statistic in small samples, and conclude that it performs quite well.
- The computation of the statistic is relative to a given choice of the spatial weights W. If in fact the pattern of spatial autocorrelation is generated by a different set of weights, then the test can give spurious results.
- The statistic assumes that the any other trends in the data have been eliminated. If this is not so, the test can again give spurious results. Bivand et al. (2008, p. 260) give an example where there is no spatial autocorrelation in the data, but there is a simple linear trend, and where Moran's test incorrectly indicates the presence of spatial autocorrelation. (De-trending the data resolves the problem).

There are other tests for spatial autocorrelation in the literature, for example Geary's C statistic, defined as

$$C = \frac{R-1}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}$$

Note that this is not quite the inverse of Moran's statistic. Moran's statistic is usually said to be a measure of global spatial autocorrelation, while Geary's statistic is more sensitive to local autocorrelation.

5 Spatial Cross-Sectional Models

Our starting point is the linear-in-parameters cross-sectional model

$$y = X\beta + u$$

with the error term u classical.

5.1 Spatial Durbin Model

The spatial Durbin model is

$$y = X\beta + WX\theta + u$$

which just adds average-neighbor values of the independent variables to the specification. Example: the level of crime in region j depends on the intensity of policing in j as well as on the intensity in neighboring jurisdictions. Apart from potential problems of multicollinearity (recall that row-wise, X and WX are for different regions because the diagonal elements of W are zero), this model poses no problems for us.

5.2 SAR — The Spatial Autoregressive Model

This model says that levels of the *dependent* variable y depend on the levels of y in neighboring regions. It is thus a formulation of the idea of a spatial spillover: for example, the number of crimes in a zip-code region may depend on the number of crimes in adjacent (neighboring) zip-code regions. The formal model is

$$y = \lambda Wy + X\beta + u$$

with u assumed to be classical. Note that λWy makes sense since the diagonal elements of W are zero, which implies that we do not have the circular specification that y_j on the left is influenced by the same y_j on the right. Clearly we would not want to run OLS on this model, since the presence of y on both the left and right sides means that we have a correlation-between-errors-and-regressors problem, and the resulting estimates will be biased and inconsistent. But we can easily obtain the reduced form as[6]

$$\begin{aligned} y &= \lambda Wy + X\beta + u \\ (I - \lambda W)y &= X\beta + u \\ y &= (I - \lambda W)^{-1}X\beta + (I - \lambda W)^{-1}u \end{aligned}$$

(assuming that the inverse exists). We immediately see a couple of potential problems here. First, the new error term $u^* = (I - \lambda W)^{-1}u$ is no longer homoskedas-

[6]Note that from now on we revert to the convention that I is the identity matrix, and not Moran's statistic.

tic. Second, and probably more fundamentally, the model is no longer linear-in-parameters because of the new unknown parameter λ.[7]

5.3 SEM — The Spatial Error Model

In this model, the spatial influence comes only through the error terms: we have

$$\begin{aligned} y &= X\beta + u \\ u &= \rho W u + \upsilon \end{aligned}$$

with υ assumed to be normal with $\mathsf{E}(\upsilon) = 0$, $\mathsf{E}(\upsilon\upsilon') = \sigma^2 I$ (ie, completely classical). Solving the error specification for u we find

$$\begin{aligned} (I - \rho W)u &= \upsilon \\ u &= (I - \rho W)^{-1}\upsilon \end{aligned}$$

so that the model is

$$y = X\beta + (I - \rho W)^{-1}\upsilon$$

This is conceptually simpler than the SAR case because the only problems are heteroskedasticity and non-linearity in ρ.[8]

5.4 The General Spatial Model

It is possible to combine the SAR and SEM models:

$$\begin{aligned} y &= \lambda W_1 y + u \\ u &= \rho W_2 u + \upsilon \end{aligned}$$

where υ is classical and both W_1 and W_2 are spatial weights matrices. (The description of this as The General Spatial Model is due to LeSage). One motivation for this is as follows: suppose we have estimated a SAR model. We then test the

[7] In fact, *very* nonlinear in λ, as you can see even for the relatively small and straightforward W used in the example on page 6: here $(I - \lambda W)^{-1}$ involves cubic powers of λ. On the other hand, the determinant is surprisingly simple, though nonlinear in λ: it is $-\frac{9}{2\lambda^3+7\lambda^2-9}$.

[8] Note that the literature contains a certain amount of variation in naming these models: what I have been calling the SAR model may be referred to as the Spatial Lag Model; and what I call the SEM model is sometimes referred to as the spatial autocorrelated (SAR) model.

residuals for spatial autocorrelation using (say) Moran's test. If we cannot reject the hypothesis that the residuals are (still) spatially autocorrelated, then this model, which allows for both sources, may be appropriate.

A problem is the choice of forms (structure) for W_1 and W_2. Theory provides no guides; and if one naively takes $W_1 = W_2 = W$ then one can run into identification problems. Thus, this model does not appear to be used much in practice.

5.5 Interpretation

In the context of the SAR model, considering $(I - \lambda W)$, one can show that

$$(1/\zeta_{\max}) \leq \lambda \leq (1/\zeta_{\min})$$

where $\zeta_{\max}$ and $\zeta_{\min}$ are the largest and smallest eigenvalues of the spatial weights matrix W. It can also be shown that if W is row-stochastic:

$$-1 \leq \lambda \leq 1$$

but note that there is no theoretical requirement that W actually be row-stochastic. Now consider

$$(I - \lambda W)^{-1}$$

If $|\lambda| < 1$, this can be expanded in a power series as

$$(I - \lambda W)^{-1} = I + \lambda W + \lambda^2 W^2 + \lambda^3 W^3 + \ldots$$

(To see this formally, post-multiply both sides by $(I - \lambda W)$; then do the multiplication on the right and note that all terms except I cancel). So suppose that W is row-stochastic with no islands. Now, many analysts believe that in a spatial context, it is natural to assume that $\lambda \geq 0$: for example, it would be odd to suppose that an increase in crime in neighboring census tracts leads to a decrease in crime in this tract. Finally, think of the series W, W^2, etc. W is the spatial weights of the neighbors of a given region. W^2 is the weights of the neighbors of the neighbors, W^3 is the weights of the neighbors of the neighbors of the neighbors, etc. So when we look at

$$(I - \lambda W)^{-1} X\beta$$

we are examining

$$X\beta + \lambda W X\beta + \lambda^2 W^2 X\beta + \lambda^3 W^3 X\beta + \ldots$$

which is the sum of a series of decreasing influences of the entire spatial system. In the case of the SEM we are looking at

$$(I - \rho W)^{-1} u$$

which by the same reasoning can be interpreted as an autoregressive-in-space error structure, with decreasing intensity as we move further away.

6 Spatial Panel Data Models

In this context, observations are indexed by i (location) and t (time). We assume balanced panels. Most formulations of the model adopt an unobserved heterogeneity perspective: we have an unobserved time-invariant covariate α_i. If α_i is correlated with the observed covariates x_{it} then we cannot absorb α_i into the disturbance term (because then it would induce correlation between errors and regressors). Thus, in the case of the SAR model, we write

$$y_{it} = \lambda \sum_j w_{itj} y_{ij} + \alpha_i + x_{it}\beta + u_{it}$$

where w_{itj} are the elements of the (i,t) row of W and x_{it} is the (i,t) row of X, *excluding a constant*. This is the (spatial) fixed-effects model. As in non-spatial fixed effects, the problem comes from the need to estimate the α's, which function as region-specific intercepts. (Typically in these models T is much smaller than R: this is another distinction from non-spatial panels). If the number of regions is relatively small (a few dozen, perhaps as many at 50) then the model can be set up with a bunch of dummy variables and estimated without problems. (In practice it is a bit more difficult to set up, because, unlike the case of non-spatial panels, we are assuming that the data is grouped by time period: ie all R observations on the entire cross-section in the first period, followed by R observations on the entire cross-section in the second period, etc: this is in order to make use of the spatial weights matrix W meaningful).[9] But what if we have upwards of 3000 regions (eg, the set of all US counties)? We would be attempting to estimate more than 3000 intercepts, which is extremely problematic.

Fortunately, the problem can be solved in the spatial context in much the same way as in the non-spatial context: by de-meaning the data, region-wise. However,

[9] That is, suppose we organized the data as for non-spatial panels. Then the first row of data would be region 1 in period 1, and the second would be region 1 in period 2, etc. In this case, Wy would not give the desired average (cross-sectional) spatial effect.

as Anselin, Gallo and Jayet (2008) observe, the computation of the means is complicated by the spatial dependencies (the W matrix), and must be done carefully.[10] But given a correct de-meaning, then just as in the non-spatial context, we obtain a regression equation without the fixed effects (the α's). We can then estimate this equation, and reverse the transformation to recover estimates of the intercepts if they are of interest to us. (Remember that one standard interpretation of the fixed effects model is that the unobserved covariate is of no interest to us, ie is a nuisance parameter). We would also need to compute estimates of the standard errors of the intercepts; but this can be done, again relatively straightforwardly. We can also formulate and estimate spatial two-way fixed effects models where there are both temporally invariant and spatially invariant unobserved covariates.

Returning to the case of a temporally invariant covariate (α_i), if it can be assumed *not* to be correlated with the observed *x*'s then we can in principle absorb it into the disturbance term, resulting in the spatial random effects model. The problem is then to formulate a set of reasonable but parsimonious assumptions about the joint distribution of the two components of the error term. This is more difficult than in the non-spatial context, partly because we cannot just import the non-spatial assumptions wholesale. For example, Elhorst (2009) describes a model in which we have

$$\begin{aligned} y_{it} &= x_{it}\beta + u_{it} \\ u_{it} &= \alpha_i + \varepsilon_{it} \\ \varepsilon_{it} &= \lambda W \varepsilon_{it} + v_{it} \\ v_{it} &= \rho v_{i,t-1} + z_{it} \end{aligned}$$

where $\alpha_i \sim N(0, \sigma_\alpha^2)$ is a set of region-specific effects, $z_{it} \sim N(0, \sigma^2 I)$, $v_{i0} \sim N(0, \sigma_z^2/1-\rho^2)$ (v_{i0} is the initial value of the temporally autocorrelated disturbance vector) and α and z are assumed independent. Note that this is very general: it has both spatial error autocorrelation via the parameter λ and temporal AR(1) autocorrelation via the parameter ρ. See also Kapoor, Kelejian and Prucha (2007).

7 Estimation

In this section we focus on estimation of the cross-sectional models only. For spatial panels see, eg, Elhorst (2009).

[10] As it happens, publication of Anselin et al. (2008) resulted in a substantial revision of the MATLAB code for spatial panels described in the final section of these notes.

7.1 ML Estimation of the SAR model

The SAR model

$$y = \lambda W y + X\beta + u$$

can be written as

$$\begin{aligned} (I - \lambda W)y &= X\beta + u \\ Ay &= X\beta + u \end{aligned}$$

with $u \sim N(0, \sigma^2 I)$, $A = (I - \lambda W)$, and W a row-standardized spatial weights matrix. The model is usually estimated by maximum-likelihood. The log-likelihood function is

$$\ln L(\beta, \lambda, \sigma) = -(R/2)\ln\pi - (R/2)\ln\sigma^2 + \ln\|A\| - (1/2\sigma^2)(Ay - X\beta)'(Ay - X\beta)$$

where $\|A\|$ is the determinant of A. Anselin (1988, ch. 12) suggests a way to do the estimation. Focussing first on β it is straightforward to show that the ML estimator is given by

$$\begin{aligned} b &= (X'X)^{-1}X'Ay \\ &= (X'X)^{-1}X'y - \lambda(X'X)^{-1}X'Wy \\ &= b_0 - \lambda b_L \end{aligned}$$

where $b_0 = (X'X)^{-1}X'y$ and $b_L = (X'X)^{-1}X'Wy$. Inspection shows that b_0 is the coefficient vector from the OLS regression of y on X, while b_L is from the OLS regression of Wy on X. So if λ is known, we could compute the ML estimate of β. Next, write the residuals of these two OLS regressions as

$$\begin{aligned} e_0 &= y - Xb_0 \\ e_L &= Wy - Xb_L \end{aligned}$$

It can be shown that the ML estimate of σ^2 is

$$s^2 = (1/R)(e_0 - \lambda e_L)'(e_0 - \lambda e_L)$$

so once again we could estimate σ^2 if λ were known.

We can now use all this to write down a version of the log-likelihood function in terms of λ only: the result is the *concentrated* log-likelihood, $\ln L^*$ which can be seen to be

$$\ln L^* = C - (R/2)\ln[(1/R)(e_o - \lambda e_L)'(e_o - \lambda e_L)] + \ln\|A\|$$

where C does not involve any unknown parameters. We can now maximize $\ln L^*$ with respect to λ and obtain the ML estimate of this parameter, and work backwards. In detail, the estimation steps are:

1. Regress y on X : this gives b_0. Compute the residual $e_0 = y - Xb_0$
2. Regress Wy on X : this gives b_L. Compute the residual $e_L = Wy - Xb_L$
3. Find the λ that maximizes the concentrated log-likelihood function. Call it $\hat{\lambda}$.
4. Given $\hat{\lambda}$, compute $b = b_o - \hat{\lambda} b_L$ and $s^2 = (1/R)(e_0 - \hat{\lambda} e_L)'(e_0 - \hat{\lambda} e_L)$

Note that steps (1) and (2) are simply OLS linear estimation problems; and that step (3) is a *one-parameter* nonlinear optimization problem.

One problem with this is that, due to the stepwise nature of the estimation process, you do not get estimates of the (joint) covariance matrix of all the estimated parameters. However, because they are maximum-likelihood estimates, we know that they are asymptotically efficient, meaning that for large samples the covariance matrix attains the Cramer-Rao lower bound, given by

$$-\mathsf{E}(\partial^2 \ln L / \partial\theta \partial\theta')^{-1}$$

where $\theta = (\beta, \lambda, \sigma^2)$. This in turn can be estimated by the numerical Hessian of the log likelihood.

7.2 ML Estimation of the SEM model

Estimation of the cross-sectional SEM model

$$\begin{aligned} y &= X\beta + u \\ u &= \rho W u + \upsilon \\ & \quad \upsilon \text{ classical} \end{aligned}$$

(so that $\mathsf{E}(\upsilon\upsilon') = \sigma^2 I$) which we can write as

$$\begin{aligned} y &= X\beta + u \\ Bu &= \upsilon \\ B &= (I - \rho W) \\ & \quad \upsilon \text{ classical} \end{aligned}$$

is somewhat more complicated. The log-likelihood is

$$\ln L = -(R/2)\ln\pi - (R/2)\ln\sigma^2 + \ln\|B\| - (1/2\sigma^2)(y - X\beta)'B'B(y - X\beta)$$

and, as previously described, $B\upsilon$ is heteroskedastic. We could estimate β using GLS, and an estimate of σ^2 is similar to the SAR case. The concentrated log-likelihood is

$$\ln L^* = C - (R/2)\ln[(1/R)e'B'Be] + \ln\|I - \rho W\|$$

where $e = y - Xb_{\text{GLS}}$. The problem is that b_{GLS} itself depends on ρ (unlike the SAR case). Anselin suggests an iterative procedure, essentially as follows:

1. Regress y on X. Call the coefficient estimate b_{OLS} and compute the residual vector $e = y - Xb_{\text{OLS}}$
2. Use this e in the concentrated log-likelihood, and optimize to find $\hat{\rho}$.
3. Use $\hat{\rho}$ to compute the GLS estimator b_{GLS} and then a new residual vector $e = y - Xb_{\text{GLS}}$
4. First time or if the residuals have not converged: go back to step 2 and re-estimate ρ. Otherwise: go to step 5.
5. At this point we have a converged estimate of ρ (say, $\hat{\rho}$) and the associated residual vector e, and a GLS estimator of β. We can now estimate σ^2 by $(1/R)e'B'Be$.

7.3 Computational Considerations

It is important to understand that there are real computational difficulties in the case where your data represents many regions. For example, suppose you have data by US counties, of which there are about 3000 ($= 3 \times 10^3$). Then W is a matrix of about 9×10^6 (9 million) elements. Clearly it is going to be very difficult to work with arrays this size. However, it turns out that for this particular case (the US counties), and with a rook definition of contiguity, only about $12,500$ elements are non-zero (LeSage and Pace (2009)). (A county in Ohio will not be a neighbor of a county in California, or indeed of counties in most other states). So if we could just keep track of which elements are not zero, we could save considerable space. That is what sparse-matrix representations do. Instead of keeping track of the entire

matrix, they record, for each non-zero element only, its row, its column, and its value. The result, for US counties, is that we need to keep track of only about $37,500$ (= 12,500 $\times$ 3) numbers, rather than 9 million. Of course, for all this to be practically useful, we need routines to work with sparse matrix representations, eg to be able to multiply or invert them, without having to expand them into their full ("dense") forms.

A related problem is that even if W can be represented as a sparse matrix, $(I - \lambda W)$ is not sparse and is also $R \times R$. However, note from the likelihood functions, that all we really need is the log determinant of $(I - \lambda W)$ and not $(I - \lambda W)$ itself. For small R one can compute the log determinant directly. For larger R one approach is to note that

$$\ln \| I - \lambda W \| = \ln \left(\prod_{i=1}^{R} (1 - \lambda \zeta_i) \right)$$

where the ζ_i are the eigenvalues of W, which are specific to W and do not change as estimates of λ change. The advantage of this is that the eigenvalues need to be computed only once. The disadvantage is that the computation can be time-consuming, even once: for example, Kelejian and Prucha (1999, footnote 12) report for a case of $R = 1500$ and with the average number of spatial neighbors (non-zero entries in a row) equal to 10, computing the eigenvalues took 22 minutes; however, this was on what is now considered very old and slow hardware. This slowness motivated a search for a method to compute the log determinant that did not need to compute the eigenvalues. Nowadays, with improved hardware and more RAM, this is less of a problem. Still, there is room for improvement: a popular approach is a Monte Carlo method due to Barry and Pace (1999).

8 Software

Most standard statistics packages do not contain estimation routines for spatial econometric models.[11] However, most packages do contain a built-in programming language, so that in principle you could implement the estimation routines yourself, following the sketches in the previous section. But obviously this is a second-best solution: if you anticipate doing lots of work with spatial data, it is worth while learning how to use a package that has spatial estimation routines built in.

[11] I know that Limdep does not. Stata may contain an add-on for spatial econometrics, but I don't know for sure.

8.1 The R system

R has probably the most extensive set of spatial statistics support, not limited to the econometric models discussed in this note. It is a free and open-source system, available at `http://www.r-project.org/` for most computer platforms (including Windows, the Mac, and most flavors of Unix). The downside it that it has a fairly steep learning curve (at least I find that it has) and that it is command-driven system: no point-and-click menus etc.

The core R system is supplemented by a very extensive package system, providing routines to do all sorts of statistical and data-analysis tasks. R has become the preferred platform for professional statisticians, so it is extraordinarily complete: about the only thing I have found it lacking is estimation of the mixed-logit (random-parameters logit) model. It has some of the best graphics available for any system, even commercial.

There are three packages of interest in our spatial context: `sp`, `spdep` and `splm`.[12] The `sp` package contains routines for constructing spatial neighbors and weights matrices, and can manipulate ESRI shapefiles. `spdep` has cross-sectional econometric estimation routines, using the weights matrices from `sp`,while `splm` has routines for some panel-data models, also using weights matrices from `sp`.[13] Bivand et al. (2008) describes the main facilities provided by `sp` and `spdep`: this book is available for free download (for OSU faculty and students) from the Springer website.[14] Chapter 9 describes the construction of spatial weights matrices, and Chapter 10 illustrates some of the estimation routines. Briefly, the cross-sectional estimation routines are:

[12] There is a general view of the facilities available for spatial analysis in R at `http://cran.r-project.org/web/views/Spatial.html`. See also `http://geodacenter.asu.edu/r-spatial-projects`.

[13] The panel-data routines are not available as a standard R package. You can install them from the R command line via `install.packages("splm", repos="http://R-Forge.R-project.org")`, but it may be easier to install it by hand after getting the materials from `http://r-forge.r-project.org/projects/splm/`. Just put the zip file somewhere on your hard disk, then, from within R do Packages -> Insall from local zip files.

[14] In fact, most of the books in the "Use R!" series are available for download as PDF files. This includes a very helpful book (Spector (2007)) on managing data in R, including facilities for import and export. I have produced a version of Bivand et al. (2008) re-formatted to put 2 book pages on one physical (landscape) page. The re-formatted version is available on the (restricted, OSU people only) site: `http://facweb.knowlton.ohio-state.edu/pviton/courses/misc/bivand-crop.zip`.

Function	Estimates
`lm`	Basic non-spatial linear regression model
`lagsarlm`	Cross-sectional spatial autoregressive (SAR) model
`lagsarlm` with `type="mixed"`	Cross-sectional spatial Durbin model
`errorsarlm`	Cross-sectional spatial error (SEM) model

The panel-data models are estimated with package `splm`. The package allows estimation of both fixed-effects and random effects models: for random effects, the specification is as described in section 6. The routines are:

Function	Estimates
`spfelm` with `effects="spfe"`	Panel-data spatial fixed-effects model
`spfelm` with `effects="tpfe"`	Panel-data spatial time-effects model
`spfelm` with `effects="sptpfe"`	Panel-data spatial 2-way fixed-effects model
`spreml` with `errors="semssre"`	Panel-data spatial random effects model

and there are variants of the random effects model for special cases of no spatial autocorrelation ($\lambda = 0$) and/or no time autocorrelation ($\rho = 0$).

It is not clear how effective these routines are with large cross-sections, where the computational problems can be significant: in this case you may want to consider Matlab, see below. For the panel data models in package `splm`, the documentation may not be available if you use an R version earlier than 2.10 (because the R system changed the way it processes help files). The package still works, though, and I have a PDF version of a pre-print of a paper by Giovanni Millo and Gianfranco Piras which is to be submitted to the *Journal of Statistical Software*; this gives the essential documentation. I have placed a copy on the CRP 780.03 restricted website.

8.2 Matlab

Matlab is a general-purpose matrix-oriented programming language, available on an essentially free year-to-year license to OSU students and faculty. Licenses expire at the beginning of each academic year, after which the system will not start until you obtain a new license. For students, this means that once you leave OSU you will need to obtain a commercial license in order to continue (or even revise) your work (unless you go to an institution which also has a Matlab license).

Unlike R, Matlab is not primarily a data-analysis system. In particular, it has no facilities for referring to variables by name: what you do in Matlab is read your data into a matrix, and then refer to columns of that matrix by column number. So a typical regression command might be "regress column 12 of matrix `DATA` on columns 1,3 6 7 and 9 of `DATA`". This is obviously less convenient than a system which allows you to say "regress `crime` against `pop`, `income`, and `age`".

James LeSage has written a general-purpose econometrics toolbox for Matlab, available free from `www.spatial-econometrics.com`. Besides containing estimation routines for standard single-equation and simultaneous equations linear models, it also contains routines for estimating spatial models.[15] Because of Matlab's efficient sparse matrix storage facilities, it is relatively fast to estimate even very large models (for example, panel-data models involving the set of 3000 US counties). The main functions are:

Function	**Estimates**
`ols(y,X)`	Basic non-spatial linear model
`sar(y,X,W)`	Spatial cross-sectional autoregressive (SAR) model
`sem(y,X,W),`	Spatial cross-sectional error (SEM) model
`sac(y,X,W1,W2)`	General cross-sectional spatial model
`sem_panel_FE(y,X,W1,W2,T)`	spatial fixed-effects panel data model
`sem_panel_RE(y,X,W1,W2,T)`	spatial random-effects panel data model

with the `W`'s being row-standardized spatial weights matrices. The panel-data models (for balanced panels only) are due to J. Paul Elhorst, and are as specified in section 6. It is very important to obtain the latest (2009) versions of the panel-data code: earlier versions did the de-meaning incorrectly. See Anselin et al. (2008) and Elhorst (2009). The spatial toolbox comes with a few examples, and I have found it relatively easy to adapt the examples to my own estimation needs without having to learn too much about Matlab.

8.3 GeoDa

This is a free Windows program written by Luc Anselin for the construction for spatial weights matrices and estimation of the cross-sectional SAR and SEM mod-

[15] Ken Train has also produced Matlab code for estimating the mixed-logit model, enhancing the attractiveness of this platform for computationally intensive estimation tasks.

els. It is available at `http://geodacenter.asu.edu/`. If you are running Windows Vista it is important to obtain the version referred to as OpenGeoDa, which is the only version to run under this version of the Windows operating system. Anselin and his colleagues at Arizona State University support a general spatial analysis mailing list called The Openspace List, which entertains questions on all forms of spatial analysis and econometrics (not limited to GeoDa; Roger Bivand is a frequent contributor). There is apparently also under development an enhanced set of spatial analysis tools based on the (free and open-source) Python programming language. Note that the R package `sp` can read spatial-weights matrices produced by GeoDa.

GeoDa is very fast and easy to use — it is based on a standard Windows point-and-click-type interface — but personally I find it a little too much of a "black box" for my taste. Users cannot extend the estimation routines, so that if you need to do something not provided by the system, you're out of luck. My feeling is that an investment in learning a bit of either R or Matlab programming will in the long run be more productive.

References

Anselin, Luc, *Spatial Econometrics: Methods and Models*, Dordrecht, The Netherlands: Kluwer Academic Publishers, 1988.

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

____ and Raymond J. G. M. Florax, "Small Sample Properties of Tests for Spatial Dependence in Regression Models: Some Further Results," in Luc Anselin and Raymond J. G. M. Florax, eds., *New Directions in Spatial Econometrics*, Berlin: Springer, 1995, pp. 21–74.

____ , Julie Le Gallo, and Hubert Jayet, "Spatial Panel Econometrics," in László Mátyás and Patrick Sevestre, eds., *The Econometrics of Panel Data*, New York, N.Y.: Springer, 2008, pp. 625–659.

Barry, R. P. and R. K. Pace, "Monte Carlo Estimates of the Log Determinant of Large Sparse Matrices," *Linear Algebra and Its Applications*, 1999, *289*, 41–54.

Bivand, Roger S., Edzer J. Pebesma, and Virgilio Gómez-Rubio, *Applied Spatial Data Analysis with R* Use R!, Secaucus, N.J.: Springer, 2008.

Elhorst, J. Paul, "Spatial Panel Data Models," in Mandfred M. Fischer and Arthur Getis, eds., *Handbook of Applied Spatial Analysis*, Berlin: Springer, 2009.

Kapoor, M., Harry H. Kelejian, and Ingmar R. Prucha, "Panel Data Models with Spatially Autocorrelated Error Components.," *Journal of Econometrics*, 2007, *140* (1), 97–130.

Kelejian, Harry H. and Dennis P. Robinson, "Spatial Correlation: A Suggested Alternative to the Autoregressive Model," in Luc Anselin and Raymond J. G. M. Florax, eds., *New Directions in Spatial Econometrics*, Berlin: Springer, 1995, pp. 75–95.

____ and Ingmar R. Prucha, "A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model," *International Economic Review*, 1999, *40* (2), 509–533.

LeSage, James P., *Spatial Econometrics*, www.spatial-econometrics.org, 1998.

LeSage, James P. and R. Kelly Pace, *Introduction to Spatial Econometrics*, Boca Raton: CRC Press, 2009.

Spector, Phil, *Data Manipulation With R* Use R!, Secaucus, N.J.: Springer, 2007.

Tobler, W., "Cellular Geography," in S. Gale and G. Olssohn, eds., *Philosophy in Geography*, Dordrecht: Reidel, 1979, pp. 379–386.