Selecting the W Matrix: Parametric vs Non Parametric Approaches

Jesus Mur, Marcos Herrera, Manuel Ruiz

University of Zaragoza. email: jmur@unizar.es

University of Zaragoza. email: mherreragomez@gmail.com

Technical University of Cartagena, email: manuel.ruiz@upct.es

Abstract

In spatial econometrics, it is customary to specify a weighting matrix, the so-called W matrix, just choosing one matrix from the different types of matrices a user is considering (Anselin, 2002). In general, this selection is made a priori, depending on the user's judgment. This decision is extremely important because if matrix W is miss-specified in some way, parameter estimates are likely to be biased and they will be inconsistent in models that contain some spatial lag. Also, for models without spatial lags but where the random terms are spatially autocorrelated, the obtaining of robust standard estimates of the errors will be incorrect if W is miss-specified. Goodness-of-fit tests may be used to chose between alternative specifications of W. Although, in practice, most users impose a certain W matrix without testing for the restrictions that the selected spatial operator implies. In this paper, we aim to establish a nonparametric procedure where the chosen by objective criteria. Our proposal is related with the Theory of Information. Specifically, the selection criterion that we propose is based on objective information existing in the data, which does not depend on the investigator's subjectivity: it is a measure of conditional entropy. We compare the performance of our criteria against some other alternative like the J test of Davidson and McKinnon or a likelihood ratio obtained in a maximum likelihood framework.

Introduction

The weighting matrix is a very characteristic element of spatial models and, frequently, is cause of dispute in relation to what is it and how should it be specified. Half a century has gone after the pioneering works of Moran (1948) or Whittle (1954), where the terms "join" or "link" were used preferently. The work of Ord (1975) is very important in the conversion of this matrix in a key element for modelling spatial data. The matrix has received considerable attention afterward (Anselin, 2002) but, from our point of view, we do not have still a totally convincing answer to both questions.

The weighting matrix is a spatial operator usually taken for granted in applied work. According to our view, this is a very optimistic position because exist a great uncertainty that characterizes its specification. Take for example the principle of *allotopy* as stated by Ancot et al. (1982): "often what happens in a region is related with other phenomena located in distinct and remote parts of the space". The problem is identify which ones. A time series analyst faces similar problems although he has clear indications: due to the nature of economic dynamics, you must look to the past and take into account also the frequency of the data. None of the two questions is free from controversy when the data proceed form a spatial cross-section. The Space is irregular and heterogeneous (nothing to do with the monotonous succession of time) and the influences may be of any type across Space. Nearness, as claimed by Tobler (1970) is just one possibility.

We agree with Haining (2003, p.74) in the sequence of actions: "*The first step of quantifying the structure of spatial dependence in a data set is to define for any set of point or area objects the spatial relationships that exist between them*". This is what Anselin (1988, p. 16) designates "*the need to determine which other units in the spatial system have an influence on the particular unit under consideration (...) expressed in the topological notions of neighborhood and nearest neighbor*". This first step is crucial, absolutely, but it might be not so simple as just writing a binary or row-standardized weighting matrix. In some cases, we might have enough information to fully specify a weighting matrix. In other cases, this matrix will be a mere hypothesis. We suspect that the second situation is, by far, the most common among practitioners.

According to our view, the weighting matrix results from a problem of underidentification that affects, in general, to most of the spatial models. Paelinck (1979, p.20) acknowledges that there is an identification problem in

the interdependent specifications used to model spatial behaviors. In terms of Lesage and Pace (2009, p.8), a unrestricted spatial autoregressive process:

$$\left.\begin{array}{c} y_i = \alpha_{ij} y_j + \alpha_{ik} y_k + x_i \beta + \varepsilon_i \\ y_j = \alpha_{ji} y_i + \alpha_{jk} y_k + x_j \beta + \varepsilon_j \\ y_k = \alpha_{ki} y_i + \alpha_{kj} y_j + x_k \beta + \varepsilon_k \\ \varepsilon_i; \varepsilon_j; \varepsilon_k \sim N\left(0; \sigma^2\right) \end{array}\right\} \quad (1)$$

"would be of little practical usefulness since it would result in a system with many more parameters than observations. The solution to the over-parametrization problem that arises when we allow each dependence relation to have relation-specific parameters is to impose structure on the spatial dependence parameters". This is the reason why we need a spatial weighting matrix. In spite of the efforts of Folmer and Oud (2008), trying to advance towards what they call a structural approach to the the weighting matrix, or the arguments of Paci and Usai (2009) in favor of the use of proxies for spillover effects, this is a point of consensus in the literature.

The question of specifying a matrix seems complex, although usual practice has favoured simple solutions. By large, the dominant approach involves an exogenous treatment of the problem. Nearby or neighboring spatial units are identified as contiguous in a square binary connectivity matrix as, for example, in the traditional physical adjacency criteria, the m-nearest neighbors or the great circle distance. Afterwards, the binary matrix can be normalized in some way (by rows, columns, according to the total sum). Other matrices are constructed using some given function of the geographical distance between the centroids of the spatial units; the inverse of the distance between the two points is the most common measure of distance and the matrix may also be normalized. Geography may be substituted by another domain in order to obtain others measures of distance. Recently they have appear various endogenous procedures like the AMOEBA algorithm (Getis and Aldsdat, 2004, Aldsdat and Getis, 2006), the CCC method of Mur and Paelinck (2010) or the entropy-based approach of Esteban et al (2009). Although very different between them, the basic idea of the endogenous approaches appeared in the works of Kooijman (1976) and Openshaw (1977): exploit the information contained in the raw data, or in the residuals of the model, in order estimate the weighting matrix. This is feasible if we have a panel of spatial data like in Conley and Molinari (2007), Bhattacharjee and Jensen-Butler (2006) and Beenstock et al (2010) but

risky in the case of a single cross-section (the problem is data mining). Finally, there are different well-known approaches which combine strong a prioris about the channels of interaction with endogenous inferential algorithms (Bodson and Peters, 1975, Dacey, 1965).

Bavaud (1998, p.153), given this state of affairs, is clearly skeptic, *"there is no such thing as "true", "universal" spatial weights, optimal in all situations*" and continues by stating that the weighting matrix "*must reflect the properties of the particular phenomena, properties which are bound to differ from field to field*". We share his skepticism. What does this means? That, at the end, the problem of selecting a weighting matrix among the different possibilities is a problem of model selection. In fact, different weighting matrices result in different spatial lags of the endogenous or the exogenous variables included in the model. Different equations with different regressors amounts to a model selection problem, even when the weighting matrix appears in the equation of the errors. This is the direction that we want to explore in the present paper as an alternative way to deal with the uncertainty of specifying the spatial weighting matrix.

Section 2 continues with a revision of the techniques of model selection that seem to fit better into our problem. We present our own non-parametric procedure in Section 3. Section 4 discusses a large Monte Carlo experiment in which we compare the small sample behaviour of the most promising techniques. Section 5 concludes summarizing the most interesting results of our work.

Choosing a Weighting Matrix

The model of (1) can be written in matrix form:

$$y = \Gamma y + x\beta + \varepsilon \tag{2}$$

where y and ε are $(n \times 1)$ vectors, x is a $(n \times k)$ matrix, β is a $(k \times 1)$ vector of parameters and Γ is a $(n \times n)$ matrix of interaction coefficients. The model is underidentified. A solution, perhaps the most popular, consists of introducing some structure in the matrix Γ, parametrizing the spatial interaction coefficients as, for example: $\Gamma = \rho W$, ρ is a parameter and W a matrix of weights. The term $y_W = Wy$ that, consequently, appears on the right hand side (rhs, form

now on) of the equation is called the spatial lag of the endogenous variable. At this point it is worth to highlight a couple of questions:

(i) The weighting matrix can be constructed in different ways following, for example, some interaction hypothesis. Each hypothesis will result in a different weighting matrix leading to a different spatial lag. In sum, different weighting matrices amounts to different models.

(ii) There are some general guidelines about how to specify a weighting matrix using concepts like nearness, accessibility, influence, etc. Different models might require different interaction channels that are not necessarily known. This implies uncertainty and diffuse priors.

Corrado and Fingleton (2011) discuss the construction of the weighting matrix on theoretical grounds (that is, they wonder, among other things, about the information that the weights of a weighting matrix should contain). We prefer to focus on the statistical treatment of such uncertainty.

Let us assume that we have a set of N linearly independent weighting matrices, $\Upsilon = \{W_1; W_2; \ldots; W_N\}$. Usually N corresponds to a small number of different competing matrices but in some cases this number may be quite large, reflecting a situation of great uncertainty. As said, each matrix generates a different spatial lag and a different spatial model. These matrices may be related by different restrictions, resulting in a series of nested models; if the matrices are not related, the sequence of spatial models will be non-nested.

Two weighting matrices may be nested, for example, in the cases of binary rook-type and queen-type movements: all the links of the first matrix are contained in the second matrix which include also some other non-zero links. Discriminating between these two matrices is not difficult using the techniques for selecting between nested models. For example, in a maximum-likelihood approach (we would need the assumption of normality) it may be enough with a likelihood ratio or a Lagrange Multiplier. The last one is very simple as appear in the Appendix 1.

For the case of non-nested matrices, we may find several proposals in the literature. Anselin (1984) provides the appropriate Cox-statistic for the case of:

$$\left.\begin{array}{l} H_0 : y = \rho_1 W_1 y + x_1 \beta_1 + \varepsilon_1 \\ H_A : y = \rho_2 W_2 y + x_2 \beta_2 + \varepsilon_2 \end{array}\right\} \tag{3}$$

that Leenders (2002) converts into the J-test using an augmented regression like the following:

$$y = (1-\alpha)\left[\rho_1 W_1 y + x_1\beta_1\right] + \alpha\left[\hat{\rho}_2 W_2 y + x_2\hat{\beta}_2\right] + \nu \tag{4}$$

being $\hat{\rho}_2$ and $\hat{\beta}_2$ the corresponding maximum-likelihood estimates (ML from now on) of the respective parameters on a separate estimation of H_A and generalizes also to the comparison of a null model against N different models. Kelejian (2008) maintains the approach of Leenders although in a $SARAR$ framework, which requires GMM estimators:

$$\begin{aligned} y &= \rho_i W_i y + x_i\beta_i + u_i = Z_i\gamma_i + u_i \\ u_i &= \lambda_i M_i u_i + v_i \end{aligned} \tag{5}$$

with $i = 1, 2, ..., N$, $Z_i = (W_i y, x_i)$ and $\gamma_i = (\rho_i, \beta)$. The J-test for selecting a weighting matrix corresponds to the case where $x_i = x$; $W_i = M_i$ but $W_i \neq W_j$. In order to obtain the test we need the estimation of an augmented regression, similar to that of (4):

$$y(\hat{\lambda}) = S(\hat{\lambda})\eta + \varepsilon \tag{6}$$

where $S(\hat{\lambda}) = \left[Z(\hat{\lambda}), F\right]$, $Z(\lambda) = (I - \lambda W)$ (the same for y), being $\hat{\lambda}$ the estimate of λ for the model of the null. Moreover $F = [Z_1\hat{\gamma}_1, Z_2\hat{\gamma}_2, \ldots, Z_N\hat{\gamma}_N, W_1 Z_1\hat{\gamma}_1, W_2 Z_2\hat{\gamma}_2, \ldots, W_N Z_N\hat{\gamma}_N]$. The equation of 6 can be estimated by 2SLS using a matrix of instruments: $\hat{S} = \left[\hat{Z}(\hat{\lambda}), \hat{F}\right]$, where $\hat{F} = PF$ (similar for$Z(\hat{\lambda})$) with $P = H\left(H'H\right)^{-1}H$ and $H = \left[x, Wx, W^2x\right]$. Under the null that, let say, model 0 is correct the 2SLS estimate of η is asymptotically normal:

$$\hat{\eta} \sim N\left[\eta_0; \sigma_\epsilon^2\left(\hat{S}'\hat{S}\right)^{-1}\right] \tag{7}$$

where $\eta_0 = [\gamma'; 0]$. The J test means that the last $2N$ parameters of vector η are zero. Define $\hat{\delta} = A\hat{\eta}$ where A is a $2N \times (k+1+2N)$ matrix corresponding to the null hypothesis: $H_0 :\ A\eta = 0$, then the J tests can be formulated as a Wald statistic:

$$\hat{\delta}'\hat{V}^{-1}\hat{\delta} \sim \chi^2(2N) \tag{8}$$

being $\hat{V}$ the estimated sample covariance of $\hat{\delta}$.

Burridge and Fingleton (2010) show that the asymptotic chi-square distribution for the J-test, under the null, may be a poor approximation. They advocate for a bootstrap resampling procedure that appears to improve slightly both size and power. There remain implementation problems related to the use of consistent estimates for the parameters of (5) in the corresponding augmented regression. Kelejian (2008) proposes construct the test using GMM-type estimators and Burridge (2011) suggests a mixture between GMM and likelihood-based moment conditions which controls more effectively the size of the test. Piras and Lozano (2010) present new evidence on the use of the J-test that relates the power of the test to a judicious selection of the instruments.

The problem of model selection has been treated very often, and very successfully, from a Bayesian perspective (Leamer, 1978); this includes the case of selecting a weight matrix in a spatial model by Hepple (1985 a, b). The Bayesian approach, although highly demanding in terms of information, is appealing and powerful. The method appears well documented in Lesage and Pace (2009). The same as with the J-test, the starting point is a finite set of alternative models, $M = \{M_1; M_2; \ldots; M_N\}$. The specification of each model coincides (regressors, structure of dependence, etc.) but not for the spatial weighting matrix. Denote by θ the vector of k parameters. Then, the joint probability of the set of N models, k parameters and n observations corresponds to:

$$p\left(M, \theta, y\right) = \pi\left(M\right) \pi\left(\theta \middle| M\right) L\left(Y \middle| \theta, M\right) \tag{9}$$

where $\pi\left(M\right)$ refers to the priors of the models, usually $\pi\left(M\right) = 1/N$; $\pi\left(\theta \middle| M\right)$ reflects the priors of the vector of parameters conditional to the model and $L\left(y \middle| \theta, M\right)$ is the likelihood of the data conditioned on the parameters and models. Using the Bayes' rule:

$$p\left(M, \theta | y\right) = \frac{p\left(M, \theta, y\right)}{p\left(y\right)} = \frac{\pi\left(M\right) \pi\left(\theta \middle| M\right) L\left(y \middle| \theta, M\right)}{p\left(y\right)} \tag{10}$$

The posterior probability of the models, conditioned to the data, results from the integration of (7) over the parameter vector θ:

$$p\left(M | y\right) = \int p\left(M, \theta | y\right) d\theta \tag{11}$$

This is the measure of probability needed in order to compare different weighting matrices. Lesage and Pace (2009) discuss the case of a Gaussian SAR model:

$$\left.\begin{array}{c} y = \rho_i W_i y + X_i \beta_i + \varepsilon_i \\ \varepsilon_i \sim i.i.d. \mathcal{N}(0; \sigma_\epsilon^2) \end{array}\right\} \tag{12}$$

The log-marginal likelihood of (9) is:

$$p\left(M|y\right) = \int \pi_\beta \left(\beta|\, \sigma^2\right) \pi_\sigma \left(\sigma^2\right) \pi_\rho \left(\rho\right) L\left(y|\, \theta, M\right) d\beta d\sigma^2 d\rho \tag{13}$$

They assume independence between the priors assigned to β and σ^2, Normal-Inverse-Gamma (NIG in what follows) conjugate priors, and that for ρ, a $Beta(d, d)$ distribution. The calculations are not simple and, finally, '*we must rely on univariate numerical integration over the parameter ρ to convert this* (expression 13) *to the scalar expression necessary to calculate $p\left(M|Y\right)$ needed for model comparison purposes*' (Lesage and Pace, 2009, p 172). The SEM case is solved in Lesage and Parent (2007); to our knowledge, the $SARAR$ model of (5) remains still unsolved.

The techniques of model selection may also be useful here, specially if we have no preferences for a given weighting matrix; in other words, if we do not have a W matrix in the null hypothesis. There is a huge literature on model selection for nested and non-nested models with different purposes and criteria. In our case, we are looking for the most appropriate weighting matrix in order to better fit the data, so the Kullback-Leibler information criterion may be a good standard. The Akaike information criterion is simple to obtain and assures a certain trade-off between fit and parsimony (Akaike, 1974). The expression of the statistic is very well-known:

$$AIC_i = -2L\left(\hat{\theta}; y\right) + q(k) \tag{14}$$

being $L\left(\hat{\theta}; y\right)$ the log-likelihood of the model at the maximum-likelihood estimates, $\hat{\theta}$,and $q(k)$ a penalty function that depends on the number of unknown parameters. The most common specification for the penalty is simply $q(k) = 2k$. The decision rule is to select the model, weighting matrix in our case, that produces the lowest AIC.

Recently Hansen (2007) introduced another perspective to the problem of model selection that is related to the confidence of the practitioner in the

alternatives. In general, the criteria that minimize the mean-square estimation error achieve a certain balance between bias, due to misspecification errors, and variance due to parameter estimation. The optimal criterion would select the estimator with the lowest risk. This is what happens with most of the selection criteria as, for example, the AIC or the $SBIC$ statistic; also with the Bayesian concept of posterior probability, which combines prior with sampling information. The procedure of the J test is a classical decision problem solved, using only sampling information, with the purpose of minimizing the type II error assuring a given type I error.

Expressed in another way, given our collection of weighting matrices $W = \{W_1; W_2; \ldots; W_N\}$, all of which are referred to a spatial model, the purpose is to select the matrix, W_n, which, combined with the other terms of the model, produces a vector of estimates,$\hat{\theta}_n(W_n)$, that minimizes the risk. Hansen (2007) show that further reductions in the mean-squared error can be attained by averaging across estimators. The averaging estimator for θ is:

$$\hat{\theta}(W) = \sum_{n=1}^{N} \varpi^n \hat{\theta}_n(W_n) \tag{15}$$

As stated in Hansen and Racine (2010), the collection of weights, $\{\varpi^n; n = 1, 2, ..., N\}$ should be non-negative and lie on the unit simplex of $\mathbb{R}^{\mathbb{N}}; \sum_{n=1}^{N} \varpi^n = 1$.

A Non-Parametric Proposal for Choosing a Weighting Matrix

The purpose of this section is to present a new non-parametric procedure for selecting a weighting matrix. The selection criterion is based on the information content existing in the Space for the relation we are working with; this relation may be, or not, of a causal type. The measure of information that we use is based on a reformulation of the traditional entropy indices in terms of what is called *symbolic entropy*, and it does not depend on the priors of the practitioner.

As explained in Matilla and Ruiz (2008), the idea is, first, to transform the series into a sequence of symbols which should capture the relevant information. Then we translate the inference to the space of symbols using appropriate techniques.

Beginning with the symbolization process, assume that $\{x_s\}_{s\in S}$ and $\{y_s\}_{s\in S}$ are two spatial processes, where S is a set of locations in Space. Denote by $\Gamma_n = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ the set of symbols defined by the practitioner; σ_i, for $i = 1, 2, \ldots, l$, is a symbol. Symbolizing a process is defining a map

$$f : \{x_s\}_{s\in S} \to \Gamma_l \tag{16}$$

such that each element x_s is associated to a single symbol $f(x_s) = \sigma_{i_s}$ with $i_s \in \{1, 2, \ldots, l\}$. We say that location $s \in S$ is of the $\sigma_i - type$, relative to the series $\{x_s\}_{s\in S}$, if and only if $f(x_s) = \sigma_{i_s}$. We call f the *symbolization map*. The same process can be followed for the series y_s.

Denote by$\{Z_s\}_{s\in S}$ a bivariate process as:

$$Z_s = \{x_s, y_s\} \tag{17}$$

For this case, we define the set of symbols Ω_l as the direct product of the two sets Γ_l, that is, $\Omega_l^2 = \Gamma_l \times \Gamma_l$ whose elements are of the form $\eta_{ij} = \left(\sigma_i^x, \sigma_j^y\right)$. The symbolization function of the bivariate process would be

$$g : \{Z_s\}_{s\in S} \to \Omega_l^2 = \Gamma_l \times \Gamma_l \tag{18}$$

defined by

$$g\left(Z_s = (x_s, y_s)\right) = \left(f(x_s), f(y_s)\right) = \eta_{ij} = \left(\sigma_i^x, \sigma_j^y\right) \tag{19}$$

We say that s is $\eta_{ij} - type$ for $Z = (x, y)$ if and only if s is $\sigma_i^x - type$ for x and $\sigma_j^y - type$ for y.

In the following, we are going to use the following symbolization function f . Let M_e^x be the median of the univariate spatial process $\{x_s\}_{s\in S}$ and define the indicator function

$$\tau_s = \begin{cases} 1 & if \quad x_s \geq M_e^x \\ 0 & otherwise \end{cases} \tag{20}$$

Let $m \geq 2$ be the embedding dimension, defined by the practitioner. For each $s \in S$, let N_s be the set formed by the $(m-1)$ neighbours s. We use the term $m - surrounding$ to denote the set formed by each s and N_s, such that $m - surrounding\ x_m(s) = \left(x_s, x_{s_1}, \ldots, x_{s_{m-1}}\right)$. We define the indicator function for each s_i with $i = 1, 2, \ldots, m-1$:

$$\iota_{ss_i} = \begin{cases} 0 & if \quad \tau_s \neq \tau_{s_i} \\ 1 & otherwise \end{cases} \tag{21}$$

Finally, we have a symbolization map for the spatial process $\{x_s\}_{s\in S}$ as $f : \{x_s\}_{s\in S} \to \Gamma_m$, where:

$$f(x_s) = \sum_{i=1}^{m-1} \iota_{ss_i} \tag{22}$$

$\Gamma_m = \{0, 1, \ldots, m-1\}$. The cardinality of Γ_m is equal to m.

Moreover, we need to introduce some fundamental definitions:

Definition 1: The Shannon entropy, $h(x)$, of a discrete random variable x is: $h(x) = -\sum_{i=1}^{n} p(x_i)\, ln(p(x_i))$.

Definition 2: The entropy $h(x, y)$ of a pair of discrete random variables (x, y) with joint distribution $p(x, y)$ is: $h(x, y) = -\sum_x \sum_y p(x, y)\, ln(p(x, y))$.

Definition 3: Conditional entropy $h(x|y)$ with distribution $p(x, y)$ is defined as: $h(x|y) = -\sum_x \sum_y p(x, y)\, ln(p(x|y))$.

The last index, $h(x|y)$, is the entropy of x that remains when y has been observed.

These entropy measures can be adapted to the empirical distribution of the symbols. Once the series has been symbolized, for a embedding dimension $m \geq 2$, we can calculate the absolute and relative frequency of the collections of symbols $\sigma^x_{i_s} \in \Gamma_l$ and $\sigma^y_{j_s} \in \Gamma_l$.

The absolute frequency of symbol σ^x_i is:

$$n_{\sigma^x_i} = \#\{s \in S | s \quad is \quad \sigma^x_i - type \quad for \quad x\} \tag{23}$$

Similarly, for series $\{y_s\}_{s\in S}$, the absolute frequency of symbol σ^y_j is:

$$n_{\sigma^y_j} = \#\left\{s \in S | s \quad is \quad \sigma^y_j - type \quad for \quad y\right\} \tag{24}$$

Next, the relative frequencies can also be estimated:

$$p(\sigma^x_i) \equiv p_{\sigma^x_i} = \frac{\#\{s \in S | s \quad is \quad \sigma^x_i - type \quad for \quad x\}}{|S|} = \frac{n_{\sigma^x_i}}{|S|} \tag{25}$$

$$p\left(\sigma_j^y\right) \equiv p_{\sigma_j^y} = \frac{\#\left\{s \in S | s \quad is \quad \sigma_j^y - type \quad for \quad y\right\}}{|S|} = \frac{n_{\sigma_j^y}}{|S|} \tag{26}$$

where $|S|$ denotes the cardinal of set S; in general $|S| = R$.

Similarly, we calculate the relative frequency for $\eta_{ij} \in \Omega_l^2$:

$$p\left(\eta_{ij}\right) \equiv p_{\eta_{ij}} = \frac{\#\left\{s \in S | s \quad is \quad \eta_{ij} - type\right\}}{|S|} = \frac{n_{\eta_{ij}}}{|S|} \tag{27}$$

Finally, the *symbolic entropy* for the $two - dimensional$ spatial series $\{Z_s\}_{s \in S}$ is:

$$h_Z\left(m\right) = -\sum_{\eta \in \Omega_m^2} p\left(\eta\right) ln\left(p\left(\eta\right)\right) \tag{28}$$

We can obtain the marginal symbolic entropies as

$$h_x\left(m\right) = -\sum_{\sigma^x \in \Gamma_m} p\left(\sigma^x\right) ln\left(p\left(\sigma^x\right)\right) \tag{29}$$

$$h_y\left(m\right) = -\sum_{\sigma^y \in \Gamma_m} p\left(\sigma^y\right) ln\left(p\left(\sigma^y\right)\right) \tag{30}$$

In turn, we can obtain the symbolic entropy of y, conditioned by the occurrence of symbol σ^x in x as:

$$h_{y|\sigma^x}\left(m\right) = -\sum_{\sigma^y \in \Gamma_m} p\left(\sigma^y | \sigma^x\right) ln\left(p\left(\sigma^y | \sigma^x\right)\right) \tag{31}$$

We can also estimate the conditional symbolic entropy of y_s given x_s:

$$h_{y|x}\left(m\right) = \sum_{\sigma^x \in \Gamma_m} p\left(\sigma^x\right) h_{y|\sigma^x}\left(m\right) \tag{32}$$

Now we can move to the problem of choosing a weighting matrix for the relationship between variables x and y. This selection will be made among a finite set of weighting matrices, relevant for the relationship between the two processes. Let us denote by $\mathcal{W}\left(x, y\right) = \{W_\jmath | \jmath \in \mathcal{J}\}$ this set of matrices, where $\mathcal{J}$ is a set of indices. We refer to $\mathcal{W}\left(x, y\right)$ as the spatial-dependence structure set between x and y.

Denote by $\mathcal{K}$ a subset of Γ_m and let $W \in \mathcal{W}\left(x, y\right)$ be a member of the set of matrices. We can define

$$\mathcal{K}_W^x = \left\{\sigma^x \in \mathcal{K} | \sigma^x \text{ is admissible for } Wx\right\}. \tag{33}$$

where *admissible* indicates that the probability of occurrence of the symbol is positive.

By Γ_m^x we denote the set of symbols that are admissible for $\{x_s\}_{s\in S}$. Let $W_0 \in \mathcal{W}(x,y)$ be the most informative weighting matrix for the relationship between x and y. Given the spatial process $\{y_s\}_{s\in S}$, there is a subset $\mathcal{K} \subseteq \Gamma_m$ such that $p\left(\mathcal{K}_{W_0}^x|\sigma^y\right) > p\left(\mathcal{K}_W^{*x}|\sigma^y\right)$ for all $\mathcal{K}^* \subseteq \Gamma_m$, $W \in \mathcal{W}(x,y) \setminus \{W_0\}$ and $\sigma^y \in \Gamma_m^y$. Then

$$\begin{aligned} h_{W_0x|y}(m) &= -\sum_{\sigma^y\in\Gamma^y} p(\sigma^y)\left[\sum_{\sigma^x\in\mathcal{K}_{W_o}^x} p(\sigma^x|\sigma^y)\ln\left(p(\sigma^x|\sigma^y)\right)\right] \leq \qquad (34)\\ &\leq -\sum_{\sigma^y\in\Gamma^y} p_{\sigma^y}\left[\sum_{\sigma^x\in K_W^{*x}} p(\sigma^x|\sigma^y)\ln\left(p(\sigma^x|\sigma^y)\right)\right] = h_{Wx|y}(m) \end{aligned}$$

We have thus proved the following theorem.

Theorem 1: *Let $\{x_s\}_{s\in S}$ and $\{y_s\}_{s\in S}$ two spatial processes. For a fixed embedding dimension $m \geq 2$, with $m \in \mathbb{N}$, if the most important weighting matrix that reveals the spatial-dependence structure between x and y is $W_0 \in \mathcal{W}(x,y)$ then*

$$h_{W_0x|y}(m) = \min_{W\in\mathcal{W}(x,y)} \left\{h_{Wx|y}(m)\right\}. \qquad (35)$$

Monte Carlo Experiment

In this section, we generate a large number of samples from differents data generation process (D.G.P.) to study the performance of different proposals: J test, Bayesian approach, averaging estimator and conditional symbolic entropy.

Our principal interest is to detect the weighting matrix more informative between different alternatives. For this, we having the explanatory variable, x, the same in the all models, but the spatial structures differ, so that $W_0 = W_i$, where i is the matrix for the $i-th$ alternative model.

A great variety of alternative of weighting matrices are possible for our study, however we restrict our attention to k-nearest neighbors and weights distance-

based. Also, we can work with different models: Spatial autoregressive process (SAR) or spatial error model (SEM) or SARAR(p,q).

Each experiment starts by obtaining a random map in a hypothetical two-dimensional space. This irregular map is reflected on the corresponding normalized W matrix. In the first case, W is based on a matrix of 1s and 0s denoting contiguous and non-contiguous regions, respectively, subsequently normalized so that rows sum to 1. For the second case, distance-based weigth, W is constructed using $w_{ij} = d_{ij}^{-2}$ for $d_{ij} < D$, where D is a cut-distance, and $d_{ij} = 0$ otherwise, denoting d_{ij}as the straigh-line (Euclidean) distance between regions i ans j.

The following global parameters are involved in the $D.G.P.$:

$$N \in \{100, 300, 600, 1000\}\,,\ \rho \in \{0.1; 0.3; 0.5; 0.7; 0.9\}\,,\ m \in \{4, 5, 6, 7, 8\} \quad (36)$$

where N is the sample size, ρ is the spatial autocorrelation parameter and m is usually known as the *embedding dimension*. Briefly, the latter corresponds to the set made by each observation and its $m-1$ neighbours.

In the experiment, we want to simulate both linear and non-linear relations between the variables x and y.

In the first case, linearity, we control the relation by, for instance, the coefficient of determination expected from the equation. Based on a specification like this:

$$y = \beta x + \theta W x + \varepsilon, \quad (37)$$

the strength of the relation can be deduced by the expected $R^2_{y/x}$ coefficient.

Under equation (37), the expected coefficient of determination between the variables is equal to (assuming an unit variance of x and in ε as well as incorrelation between the two variables):

$$R^2_{y/x} = \frac{\beta^2 + \left(\theta^2/_{m-1}\right)}{\beta^2 + \left(\theta^2/_{m-1}\right) + 1}$$

We have considered different values for this coefficient:

$$R^2_{y/x} \in \{0.3; 0.5; 0.7; 0.9\} \quad (38)$$

For simplicity, in all cases we maintain $\beta = 0.5$. The spatial lag parameter of x, θ, is obtained by deduction: $\theta = \sqrt{\frac{(1-m)(\beta^2(1-R^2)-R^2)}{1-R^2}}$.

Having defined the values of the parameters involved in the simulation, we can present the different processes used in the analysis.

To analyze the empirical size, we have considered that the variables are distributed as follows:

$$\begin{aligned} y &\sim \mathcal{N}(0,1) \\ x &\sim \mathcal{N}(0,1) \end{aligned} \tag{39}$$

Two linear and two non-linear models have been contemplated for statistical power. The latter are obtained by applying different non-linear transformations to the variable y with respect to the corresponding linear case.

Linear Models

DGP1

$$y = \beta x + \theta W x + \varepsilon \tag{40}$$

DGP2

$$y = (I - \rho W)^{-1}(\theta W x + \varepsilon) \tag{41}$$

Non-Linear Models

DGP3

$$y = {}^{1}\!/_{(\beta x + \theta W x + \varepsilon)} \tag{42}$$

DGP4

$$y = {}^{1}\!/_{\left[(I-\rho W)^{-1}(\theta W x + \varepsilon)\right]} \tag{43}$$

In all cases: $x \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{N}(0,1)$ and $Cov(x, \varepsilon) = 0$.

[TO BE COMPLETED]

Conclusions

[TO BE COMPLETED]

Appendix 1. A Lagrange Multiplier for discriminating between two weighting matrices

Let us assume a given spatial model of an autoregressive type with a normally distributed error term:

$$y = \rho W y + x\beta + \varepsilon;\ \varepsilon \sim iidN(0; \sigma^2) \tag{44}$$

We deal with the problem of choosing between two weighting matrices, one of which is nested in the other. For example, we need to decide if the ring formed by the 3 nearest neighbors is enough or do we need to use the 4 nearest neighbors ring. The question is to decide if some weights might be zero. In that case, the nesting weighting matrix may be splitted into two matrices: $W = W_1 + W_0$. The null hypothesis is that the weights in W_0 are not relevant in the model of 44, which becomes:

$$y = \rho W_1 y + x\beta + \varepsilon;\ \varepsilon \sim iidN(0; \sigma^2) \tag{45}$$

The model of the alternative may be writen as:

$$y = \rho_1 W_1 y + \rho_0 W_0 y + x\beta + \varepsilon;\ \varepsilon \sim iidN(0; \sigma^2) \tag{46}$$

Strictly speaking, the parameters ρ_0 and ρ_1 must also coincide although the important point is that if ρ_0 is zero, the W_0 weighting matrix dissapears from the specification. Accordingly, we propose the following null and anternative hypothesis:

$$\left.\begin{array}{l} H_0 : \rho_0 = 0 \\ H_A : \rho_0 \neq 0 \end{array}\right\} \tag{47}$$

Assuming normality in the error terms, the Lagrange Multiplier is the following:

$$LM_{W_0} = \left(\frac{y' W_0^{'} \hat{\varepsilon}_{W_1}}{\hat{\sigma}^2} - tr\left(B_1 W_0\right) \right)^2 \hat{\sigma}^2_{g(\rho_0)} \sim \chi^2(1) \qquad (48)$$

$\hat{\sigma}^2$ is the maximum-likelihood estimation of σ^2 obtained from the model of 46 under the null of 47; $\hat{\varepsilon}_{W_1}$ is the vector of residuals from the model of the null where only intervenes the matrix W_1 . B_1 is the matrix $B_1 = (I - \hat{\rho}_1 W_1)^{-1}$ where the maximum likelihood estimation of parameter $\hat{\rho}_1$is used. The second term of the expression, $\hat{\sigma}^2_{g(\rho_0)}$, refers to the inverse of the estimated variance of the element of the score corresponding to the null hypothesis of 47. Its composition is the following:

$\hat{\sigma}^2_{g(\rho_0)} = I^{-1}_{\rho_0\rho_0} + I^{-1}_{\rho_0\rho_0} I^{'}_{\theta\rho_0} I^{-1}_{\theta\theta_0} I_{\theta\rho_0} I^{-1}_{\rho_0\rho_0}$

$\bullet I_{\rho_0\rho_0} = \frac{\hat{y}' W_0^{'} W_0 \hat{y}}{\hat{\sigma}^2} + tr\left(B_1^{'} W_0 + B_1^{'} W_0^{'}\right) B_1 W_0$

$\bullet I^{-1}_{\theta\theta_0} = \left[I^{-1}_{\theta\theta} - I^{'}_{\theta\rho_0} I^{-1}_{\rho_0\rho_0} I_{\theta\rho_0} \right]$

$\bullet I^{'}_{\theta\rho_0} = \frac{1}{\hat{\sigma}^2} \left[\begin{array}{ccc} x' W_0 \hat{y} & \hat{y}' W_0^{'} W_1 \hat{y} + \hat{\sigma}^2 tr\left(W_0 B_1 B_1^{'} W_1^{'} + B_1 W_0 B_1 W_1\right) & tr B_1 W_0 \end{array} \right]$

$$\bullet I^{-1}_{\theta\theta} = \frac{1}{\hat{\sigma}^2} \left[\begin{array}{ccc} x'x & x' W_1 \hat{y} & 0 \\ & \frac{\hat{y}' W_1^{'} W_1 \hat{y}}{\hat{\sigma}^2} + tr\left(B_1^{'} W_1 + B_1^{'} W_1^{'}\right) B_1 W_1 & tr B_1 W_1 \\ & & \frac{R}{2\hat{\sigma}^2} \end{array} \right]$$

Obviously, $I^{-1}_{\theta\theta_0}$ is the covariance matrix of the restricted maximum-likelihood estimates, under the null of 47, of the vector $\theta^{'} = \left[\begin{array}{ccc} \beta & \rho_1 & \sigma^2 \end{array} \right]^{'}$; $I^{-1}_{\theta\theta}$ is the covariance matrix of the unrestricted maximum likelihood estimates of vector θ in the model of 47. $I_{\theta\rho_0}$ is the covariance vector between the maximum-likelihood estimates of the coefficients of the null model, $\theta^{'} = \left[\begin{array}{ccc} \beta & \rho_1 & \sigma^2 \end{array} \right]^{'}$, and the parameter of the null hypothesis, ρ_0. Given a significance level for the test, α, the decision rule for testing the hypothesis of 47 is:

$If \quad 0 \leq LM_{W_0} \leq \chi^2_\alpha(1) \quad Do \quad not \quad reject \quad H_0,$

$If \quad LM_{W_0} > \chi^2_\alpha(1) \quad Reject \quad H_0.$

References

[1] Aldstadt, J. and A. Getis (2006): Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters. *Geographical Analysis* 38 327-343.

[2] Ancot, L, J. Paelinck, L. Klaassen and W Molle (1982): Topics in Regional Development Modelling. In M. Albegov, Å. Andersson and F. Snickars

(eds, pp.341-359), *Regional Development Modelling in Theory and Practice*. Amsterdam: North Holland.

[3] Anselin L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer.

[34] Anselin, L. (2002): Under the Hood: Issues in the Specification and Interpretation of Spatial Regression Models. *Agricultural Economics* 17 247–267.

[4] Bavaud, F. (1998): Models for Spatial Weights: a Systematic Look. *Geographical Analysis* 30 153-171.

[5] Beenstock M., Ben Zeev N. and Felsenstein D (2010). Nonparametric Estimation of the Spatial Connectivity Matrix using Spatial Panel Data. *Working Paper*, Department of Geography, Hebrew University of Jerusalem.

[6] Bhattacharjee A, Jensen-Butler C (2006): Estimation of spatial weights matrix, with an application to diffusion in housing demand. *Working Paper*, School of Economics and Finance, University of St.Andrews, UK.

[7] Bodson, P. and D. Peters (1975): Estimation of the Coeffcients of a Linear Regression in the Presence of Spatial Autocorrelation: An Application to a Belgium Labor Demand Function. *Environment and Planning A* 7 455-472.

[29] Burridge, P. (2011): Improving the J test in the SARAR model by likelihood-based estimation. *Working Paper*; Department of Economics and Related Studies, University of York .

[28] Burridge, P. and Fingleton, B. (2010): Bootstrap inference in spatial econometrics: the J-test. *Spatial Economic Analysis* 5 93-119.

[8] Conley, T. and F. Molinari (2007): Spatial Correlation Robust Inference with Errors in Location or Distance. *Journal of Econometrics*, 140 76-96.

[31] Corrado, L. and B. Fingleton (2011): Where is Economics in Spatial Econometrics? Working Paper; Department of Economics, University of Strathclyde.

[9] Dacey M. (1965): A Review on Measures of Contiguity for Two and k-Color Maps. In J. Berry and D. Marble (eds.): *A Reader in Statistical Geography*. Englewood Cliffs: Prentice-Hall.

[10] Fernández E., Mayor M. and J. Rodríguez (2009): Estimating spatial autoregressive models by GME-GCE techniques. *International Regional Science Review*, 32 148-172.

[11] Folmer, H. and J. Oud (2008): How to get rid of W? A latent variable approach to modeling spatially lagged variables. *Environment and Planning A* 40 2526-2538

[12] Getis A, and J. Aldstadt (2004): Constructing the Spatial Weights Matrix Using a Local Statistic Spatial. *Geographical Analysis*, 36 90-104.

[13] Haining, R. (2003): *Spatial Data Analysis*. Cambridge: Cambridge University Press.

[25] Hepple, L. (1995a): Bayesian Techniques in Spatial and Network Econometrics: 1 Model Comparison and Posterior Odds. *Environment and Planning A,* 27, 447–469.

[26] Hepple, L. (1995b): Bayesian Techniques in Spatial and Network Econometrics: 2 Computational Methods and Algorithms. *Environment and Planning A,* 27, 615–644.

[27] Kelejian, H (2008): A spatial J-test for Model Specification Against a Single or a Set of Non-Nested Alternatives. *Letters in Spatial and Resource Sciences*, 1 3-11.

[14] Kooijman, S. (1976): Some Remarks on the Statistical Analysis of Grids Especially with Respect to Ecology. *Annals of Systems Research* 5.

[32] Hansen, B. (2007): Least Squares Model Averaging. *Econometrica* 75 1175-1189.

[33] Hansen, B. and J. Racine (2010): Jackknife Model Averaging. *Working Paper*, Department of Economics, McMaster University.

[23] Leamer, E (1978): *Specification Searches: Ad Hoc Inference with Non Experimental Data*. New York: John Wiley and Sons, Inc.

[30] Leenders, R (2002): Modeling Social Influence through Network Autocorrelation: Constructing the Weight Matrix. *Social Networks*, 24 21-47.

[15] Lesage, J. and K. Pace (2009): *Introduction to Spatial Econometrics*. Boca Raton: CRC Press.

[24] Lesage, J. and O. Parent (2007): Bayesian Model Averaging for Spatial Econometric Models. *Geographical Analysis* 39 241-267.

[16] Matilla, M. and M. Ruiz (2008): A non-parametric independence test using permutation entropy. *Journal of Econometrics*, 144 139-155.

[17] Moran, P. (1948): The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society B* 10 243-251.

[18] Mur, J. and J Paelinck (2010): Deriving the W-matrix via p-median complete correlation analysis of residuals. *The Annals of Regional Science*, DOI: 10.1007/s00168-010-0379-3.

[19] Openshaw, S. (1977): Optimal Zoning Systems for Spatial Interaction Models. *Environment and Planning A* 9, 169-84.

[20] Ord K. (1975): Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association.* 70 120-126.

[35] Paci, R. and S. Usai (2009): Knowledge flows across European regions. *The Annals of Regional Science*, 43 669-690.

[21] Paelinck, J and L. Klaassen (1979): *Spatial Econometrics*. Farnborough: Saxon House

[22] Piras, G and N Lozano (2010): Spatial J-test: some Monte Carlo evidence. *Statistics and Computing,* DOI: 10.1007/s11222-010-9215-y.

[23] Tobler W. (1970): A computer movie simulating urban growth in the Detroit region. *Economic Geography,* 46 234-240.

[24] Whittle, P. (1954): On Stationary Processes in the Plane. *Biometrika*, 41 434-449.