Efficient estimation of spatial autoregressive models

Théophile Azomahou* and Yves Kuhry

February 2001

BETA-Theme, Université Louis Pasteur, Strasbourg
Department of Economics
61, avenue de la Forêt Noire
F-67085 Strasbourg Cedex; France

Abstract

This paper considers estimating general spatial autoregressive models using the generalized method of moments (GMM). I propose nonparametric estimates of the optimal instruments based on conditional second moment restrictions. We show that these instruments are optimal over all possible instruments, especially over those usually suggested in a spatial context. We provide a consistent nonparametric estimator of sample autocovariances function for irregularly spaced spatial stochastic processes. We then derive a consistent estimator for the asymptotic covariance matrix. Finally we investigate the asymptotic distribution of the GMM estimator.

JEL Classification: C13; C14; C50

Keywords: Instrumental variables; Spatial dependence; Nonparametric estimation

*Tel.: (+33) 390 242 092; fax: (+33) 390 242 071; e-mail: `azomahou@cournot.u-strasbg.fr`

1 Setup: overview and motivation

General forms of dependence are rarely allowed for in cross-sectional data, although routinely permitted in time-series. A major reason is that time-series typically come with a time label giving the data a natural ordering and a structure which is absent in cross-sectional data. In practice, this makes estimation and inference given minimal assumptions about dependence more difficult with cross-sections than with time-series.

Spatial dependence is defined as a special case of non-zero covariance structure for cross-sectional observations at pairs of locations. Contrary to time-series models, until quite recently, little attention has been paid to models for spatial data. The problem follows from the fact that spatial processes are characterized by a system of dependence which, by nature, involves multi-directional motion whereas the dependence in time is uni-directional. This particular characteristic of spatial stochastic processes precludes the simple transposition of time-series methodologies.

Examples of recent empirical studies that explicitly incorporate spatial dependence concern, among others, the study of spatial patterns in household demand (Case, 1991), the analysis of innovation decisions (Hautsch and Klotz, 1999), the forecasting of cigarette demand using panel data (Baltagi and Li, 1999), real wages reaction to local and aggregate unemployment rates over time (Ziliak et al., 1999) and the estimation of a hedonic model for residential sales transactions (Bell and Bockstael, 2000).

Two structures of spatial dependence mostly encountered in the literature (spatial dependence across observations for the response variable and spatial autocorrelation in error terms) are sources of econometric and statistical inference problems. As outlined by Anselin (1988, p. 59), the asymptotic properties of the ordinary least squares estimator for the model with spatial residual autocorrelation are more in line with the time-series analogue than for the autoregressive model. Indeed, parameter estimates will still be unbiased, but inefficient due to the nondiagonal structure of the covariance matrix of the error terms. In light of this, Kelejian and Prucha (1997) have shown that, for a spatial error model, the attempt to use a two-stage least squares procedure based on the Cochrane-Orcutt (1949) transformation leads to inconsistent estimates. Kelejian and Prucha demonstrated that the response function associated with the two-stage least squares procedure problem is not a separating function, that is, Amemiya's (1985, p. 246) rank condition is violated. This raises identification issues for the spatial parameter.

In a regression context and from a methodological viewpoint, when a spatial lag of the response variable is used as additional regressor, the spatial dynamics involved induce the problem of endogeneity. Indeed, contrary to time-series models where the lagged dependent variable is uncorrelated with the error if there is no serial residual correlation, this correlation occurs in the spatial context regardless of the properties of the disturbance. As a result, the ordinary least squares (OLS) estimator will be biased as well as inconsistent; see e.g., Azomahou and Lahatte (2000) for a detailed proof. This inconsistency is mentioned in papers presenting alternative estimation procedures and has motivated the application of maximum likelihood procedures.

In order to clarify estimation issues due to the incorporation of spatial dynamics in a cross-sectional regression, let us consider for example the so-called general linear spatial model, that is, a first order spatial autoregressive model with first order spatial autoregressive disturbance

$$\begin{aligned} y_n &= \lambda W_n y_n + X_n\beta + \epsilon_n, \\ \epsilon_n &= \rho M_n \epsilon_n + u_n, \end{aligned} \tag{1}$$

where y_n is the $n \times 1$ vector of observations on the dependent variable, W_n and M_n are non-stochastic $n \times n$ spatial weighting matrices, that is to say, W_n and M_n express for each (row) observation those (columns) locations that belong to its neighborhood set as non-zero elements. Clearly, W_n and M_n are matrices of a graph (Berge, 1983).[1] For the normalization of this model it is usually assumed that all diagonal elements of W_n and M_n are zero. For stability purpose, it is further assumed that $|\lambda| < 1$ and $|\rho| < 1$, ρ and λ being the scalar autoregressive coefficients. To achieve the stability, one may also assume in general that the column and row sums of W_n and M_n are uniformly bounded in absolute value, which ensures the stationarity of the underlying spatial process. β is a $k \times 1$ vector of regression parameters, ϵ_n is a $n \times 1$ vector of regression disturbances and u_n is an $n \times 1$ vector of innovations. The variables $W_n y_n$ and $M_n \epsilon_n$ are referred to as spatial lags. For generality, we allow the elements of the matrices and vectors to depend on the number of observations n, that is to form a triangular array. In the notation above, an index n denotes the size of the sample. This notation allows us to fix ideas on elements of the model which depend on n.

[1] To illustrate these considerations, let us consider a spatial process $\{Z(t) :\in T\}$ where $T \subseteq \mathbb{R}^d$ is an index set of a countable collection of regularly or irregularly scattered spatial sites and these sites are supplemented with a neighborhood structure. Such a neighborhood structure is generally modeled either by a connectivity matrix (say W_n, where W_n is a $n \times n$ matrix, with elements $w_{ij} = 1$ if sites i and j are juxtaposed, $w_{ij} = 0$ if not; n is the number of sites) or by a graph-theoretic formalism (the sites become vertices, which are connected with edges for contiguous objects). Such spatial processes are called lattices. The most important application of lattice models is statistical modeling of spatial images, which is widespread in astronomical image processing. See, e.g., Cressie (1991) for further discussions.

The structure of (1) is such that it contains both a spatial lag of the response variable as additional regressor and a disturbance term that is spatially autoregressive. Let us use the notations $A_n = I - \lambda W_n$ and $B_n = I - \rho M_n$. If we assume u_n to be normally distributed with mean zero and finite variance, the log-likelihood function of (1) for the joint vector of observations y_n is computed as

$$L(.,\theta) = -\frac{n}{2} \cdot \ln 2\pi - \frac{1}{2} \cdot \ln|\Omega_n| + \ln|B_n| + \ln|A_n| - \frac{1}{2}\nu_n'\nu_n, \quad (2)$$

with

$$\nu_n'\nu_n = (A_n y_n - X_n\beta)' B_n' \Omega_n^{-1} B_n (A_n y_n - X_n\beta) \quad (3)$$

defined as a sum of squares of appropriately transformed disturbances, where $\theta = (\rho, \beta', \lambda, \sigma_u^2)'$ and $\Omega := E(u_n u_n')$ is a diagonal error covariance matrix with $u_n = B_n \epsilon_n$. The core of the log-likelihood expression (2)-(3) is the quadratic form in the disturbances, which may or may not lead to a well-behaved optimization problem. Indeed, the determinants $|\Omega_n|$, $|A_n|$ and $|B_n|$ can cause problems in this respect. Two issues are linked to the maximum likelihood procedure (2)-(3): the compactness of the parameter space and computational issues in large samples. We consider them in turn.

To simplify the discussion, let us consider only the case $\ln|B_n| = \ln|I - \rho M_n|$, which turns out to be the log of the determinant for the likelihood in a spatial error specification. For this likelihood function to be defined, ρ must be such that $(I - \rho M_n)$ is positive definite. Singular points occur when the determinant of the Jacobian $|I - \rho M_n|$ equals zero, or equivalently, when $|\upsilon I - M_n| = 0$, with $\upsilon = 1/\rho$. Since this expression defines υ as an eigenvalue for M_n, the problem in relation (2)-(3) is undefined at every value of ρ such that $1/\rho$ equals an eigenvalue of M_n. Hence we observe that the values of ρ that make (2)-(3) undefined are directly related to the eigenvalues of M_n and these eigenvalues will change with the addition of new observations. As a result, with no further restrictions, the problem is characterized by a noncontinuous parameter space. Moreover, in the case of the frequently used row-standardized weighting matrix, the singular points in the parameter space could be computed and avoided. However, their existence complicates proofs of consistency theorems as the parameter space is not generically compact.

Despite this issue, assume that the log-likelihood function can be solved numerically. Brute force computation of the estimator involves the repeated evaluation of the determinant of the $n \times n$ matrix $(I - \rho M_n)$, which is of course cumbersome for large samples. A way to estimate ρ consists for example in using a grid-search algorithm after concentrating the log-likelihood function with respect to ρ. Then, feasible generalized least squares estimates of β and σ_u^2 can be obtained conditioned on the estimated value of ρ. Again, large samples matter since the Jacobian is an $n \times n$ matrix, the determinant of which must be repeatedly evaluated in searching over the ML estimate of ρ.

To minimize this computational burden, Ord (1975) suggested to approximate the troublesome term $\ln|I - \rho M_n|$ as $\sum_{i=1}^{n} \ln(|1 - \rho \upsilon_i|)$, where υ_i denotes the i-th eigenvalue of M_n. Since M_n is a known matrix, this approach has the advantage that the eigenvalues of M_n have to be computed only once (at the outset of the numerical optimization procedure used in finding the ML estimates) and not repeatedly at each numerical iteration. However, this still leaves the task of finding the eigenvalues of M_n. In many cases, it will be practically impossible to compute those eigenvalues accurately. Since spatial weighting matrices are not typically symmetric, an accurate calculation of the ML estimator may not be feasible in many cases even for moderate sample sizes. This shortcoming is of consequence because ML procedures are often challenging when the sample size is large.[2] Furthermore, the ML procedure requires distributional assumptions that the investigator may not wish to specify.

To overcome these difficulties, the spatial weighting matrix is generally row standardized. That is, M_n is transformed so that the elements in each row sum to unity. As a result and assuming that all diagonal elements of M_n are zero, all eigenvalues of M_n will be less than or equal to one, implying that no positive roots of $|I - \rho M_n| = 0$ will be found between 0 and 1, and no negative roots will be found between 0 and $-1/|\upsilon_{\min}|$, see, e.g., Kelejian and Robinson (1995) for proofs. Given the maximum and minimum eigenvalues for the row standardized weighting matrices, the statistical problem will always be well defined over the range $[-1/|\upsilon_{\min}|, 1]$. Unfortunately, there is not necessarily a good reason to row-standardize relying on the underlying economic story, although there is a great temptation to interpret ρ as an autocorrelation coefficient; see e.g., Kelejian and Robinson (1995) for a discussion.

All these issues motivate the recent development of alternative estimators initiated by Kelejian and Prucha (1998, 1999). These estimators are based on generalized method of moments and generalized spatial two-stage least squares procedures with instrumental variables, and have the advantage of being computationally simple even in the case of extremely large samples. However, we can outline some limits.[3]

The first is related to the mechanism of selection of the instruments. Indeed, as we

[2] The precision of numerical procedures required to calculate eigenvalues decreases rapidly as the size of the spatial weighting increases. Kelejian and Prucha (1999a) found that eigenvalues of nonsymmetric (e.g. row-standardized) spatial weighting matrices of dimensions over 400 could not be reliably computed.

[3] From a theoretical side, Conley (1999) developed a GMM-based approach to deal consistently with lattice models by correcting for spatial correlations in the error terms due to locations.

have seen, in a first step of their estimation procedure Kelejian and Prucha consider as instruments a full column rank ($p \geq k+1$) matrix, say H_n, that is composed of a subset of the linearly independent columns of $(X_n, W_nX_n, W_n^2X_n, \cdots, M_nX_n, M_nW_nX_n, M_nW_n^2X_n, \cdots,)$ where the subset contains at least the linearly independent columns of (X_n, M_nX_n). They note that in practice, that subset might be the linearly independent columns of $(X_n, W_nX_n,\ W_n^2X_n, M_nX_n, M_nW_nX_n, M_nW_n^2X_n)$, or if the number of regressors is large, just those of $(X_n, W_nX_n,\ M_nX_n, M_nW_nX_n)$.

Ideally, the instruments should be strongly related to the original variable, and asymptotically uncorrelated with the error terms, although there is no general rule to achieve this. While asymptotically the number of instruments does not matter, in finite samples, practical limits will result from problems with multicolinearity and concerns about degrees of freedom. The requirement of asymptotic uncorrelatedness between instruments and the error term is harder to assess. In the simple case considered by Kelejian and Prucha (1999a) where the instruments may consist of fitted values, the lack of correlation can be shown analytically. However, it should be noted that the subset of instruments considered involves powers of spatial weighting matrices. In the case of row-standardized weighting matrix, a well known result is that elements of powers will tend to zero when the power increases. Anselin (1980) already suggested instruments of the same nature. Even if using a spatially lagged predicted value of a regression of the dependent variable on the non-spatial regressors as instruments is acceptable, spatial lags of the explanatory variables in the model may lead to problems with multicolinearity. In all cases, these instruments are not optimal, as we will show later.

Another problem is that of the assumptions (namely assumption 7b Kelejian and Prucha (1998)) that rule out models in which all of the parameters corresponding to the exogeneous regressors – including the intercept parameter if an intercept is present – are zero.[4] Note that in such situation, the mean of the response variable is zero and hence, this case may be of limited interest in practice. Moreover, if the spatial weighting matrix is row standardized, the concerned assumption will fail if the only nonzero element of the vector of parameters to be estimated corresponds to the constant term. The reason is that the resulting matrix will no longer have full column rank as hypothesized in the assumption. Thus, the assumption requires that the generation of the vector of observations on the response variable involves at least

[4]This assumption says that the instruments matrix H_n satisfy $\text{plim}_{n\to\infty} H_n' Z_n = Q_{HZ}$ and $\text{plim}_{n\to\infty} H_n' M_n Z_n = Q_{HMZ}$, where Q_{HZ} and Q_{HMZ} are finite and have full column rank, $Z_n = (X_n, W_n y_n)$, X_n being a n by k matrix of regressors, W_n and M_n are spatial weighting matrices and y_n denotes the regressand; furthermore, $Q_{HZ} - \rho Q_{HMZ} = \text{plim}_{n\to\infty} H_n'(I - \rho M_n) Z_n$ has full column rank where $|\rho| < 1$.

one nonconstant regressor. One implication of this is that if the spatial weighting spatial matrix in the regression model is row standardized, the null hypothesis that all slopes are zero cannot be tested.

The second limit concerns the moment conditions of the estimation procedure. Let $\epsilon_{n,i}$, $\bar{\epsilon}_{n,i}$ and $\bar{\bar{\epsilon}}_{n,i}$ denote respectively the i-th elements of ϵ_n, $\bar{\epsilon}_n = M_n\epsilon_n$ and $\bar{\bar{\epsilon}}_n = M_n^2\epsilon_n$. Similarly, let $u_{n,i}$ and $\bar{u}_{n,i}$ be the i-th elements of u_n and $\bar{u}_n = M_n u_n$. Then, the spatial error specification implies the three moments of interest $E(u_n' u_n/n) = \sigma_u^2$, $E(\bar{u}_n'\bar{u}_n/n) = \sigma_u^2 n^{-1}\mathrm{tr}(W_n' W_n)$ and $E(\bar{u}_n' u_n/n) = 0$. These moments lead to a three-equation system

$$\Im_n[\rho, \rho^2, \sigma_u^2]' = \wp_n, \tag{4}$$

with

$$\Im_n = \frac{1}{n}\begin{bmatrix} 2E(\epsilon_n'\bar{\epsilon}_n) & -E(\bar{\epsilon}_n'\bar{\epsilon}_n) & n \\ 2E(\bar{\bar{\epsilon}}_n'\bar{\epsilon}_n) & -E(\bar{\bar{\epsilon}}_n'\bar{\bar{\epsilon}}_n) & \mathrm{tr}(W_n' W_n) \\ E(\epsilon_n'\bar{\bar{\epsilon}}_n + \bar{\epsilon}_n'\bar{\epsilon}_n) & -E(\bar{\epsilon}_n'\bar{\bar{\epsilon}}_n) & 0 \end{bmatrix},$$

and

$$\wp_n = n^{-1}[E(\epsilon_n'\epsilon_n), E(\bar{\epsilon}_n'\bar{\epsilon}_n), E(\epsilon_n'\bar{\epsilon}_n)]',$$

Kelejian and Prucha (1999a) suggested two estimators of ρ and σ_u^2. Essentially, these estimators are based on estimated values of $\Im_n$ and $\wp_n$. Let us denote the sample analogue of (4) by

$$\bar{\wp}_n = \bar{\Im}_n\pi + e_n, \quad \pi = (\rho, \rho^2, \sigma_u^2)', \tag{5}$$

where e_n can be viewed as a vector of i.i.d. regression residuals. The simplest of the two estimators of ρ and σ_u^2 is given by the first and third elements of the OLS estimator $\tilde{\pi}_n$ for π obtained by regressing $\bar{\wp}_n$ against $\bar{\Im}_n$. Since $\bar{\Im}_n$ is a square matrix,

$$\tilde{\pi}_n = \bar{\Im}_n^{-1}\bar{\wp}_n. \tag{6}$$

It should be noticed that $\tilde{\pi}_n$ is based on an overparametrization, in that it does not use the information that the second element of π is the square of the first. The second set of estimators of ρ and σ_u^2 considered by Kelejian and Prucha, say, $\tilde{\tilde{\rho}}_n$ and $\tilde{\tilde{\sigma}}_u^2$, is more efficient, and is defined as the nonlinear least squares estimator:

$$\hat{\pi} = \arg\min_{\pi\in\times}(\bar{\wp}_n - \bar{\Im}_n\pi)'(\bar{\wp}_n - \bar{\Im}_n\pi). \tag{7}$$

Since the construction of $\hat{\pi}$ follows from the moment conditions (4), the estimator (7) is termed by the authors as a generalized moment estimator. The standard error of the spatial parameter ρ following from (7) or (6) is difficult to derive and is not provided in Kelejian and Prucha (1999a) for the following reasons.

First of all, note that the observations of the nonlinear or the ordinary least squares regression that determine the GMM estimates for ρ and σ_u^2 are sample averages, and thus the degrees of freedom considerations appropriate for a linear or nonlinear regression model, where the observations correspond to single data points do not apply here. The reasons for not deriving standard error for ρ are several: (i) this derivation is "challenging"; (ii) for the model, ρ can be viewed as a nuisance parameter with regard to the distribution of the estimators for β and λ (the parameters corresponding to the exogeneous regressors and the spatial lag respectively), and thus all one needed was to establish the consistency of estimators: the parameters β and λ are typically the ones of primary interest; (iii) while having the distribution of the estimator for ρ would allow for a direct test for spatial autocorrelation, the presence of spatial autocorrelation can be tested with, say, the Moran I test.[5] However, it is a serious shortcoming to be unable to make statistical inference directly from parameter estimates. Now let us review in detail some arguments.

Before going into more details, it should be noted that the primary interest of Kelejian and Prucha (1998, 1999) was to obtain a computationally simple consistent estimator for ρ, and thus for the disturbance variance covariance matrix, which could then be utilized in a feasible generalized least squares procedure for the regression coefficients. Since all they needed for the feasible and the true GLS estimator to have the same asymptotic distribution was consistency of $\hat{\rho}$, the asymptotic distribution of the estimator for ρ was not important, as it does not affect the limiting distribution of the feasible GLS estimator.

The difficulties in deriving the limiting distribution can already be seen by looking at the OLS form of the GMM estimator. To derive the asymptotic distribution we would have to establish the limiting distribution of $n^{1/2}\bar{\Im}'_n\bar{\wp}_n$. Recall that $\bar{\Im}_n$ and $\bar{\wp}_n$ are respectively the sample analogue of $\Im_n$ and $\wp_n$. Assuming for a moment that the true disturbances u_n are known, the elements of this vector are products of quadratic forms in the original error terms. Thus this is not the usual situation one encounters in running OLS on observations that correspond to a single time period. The distribution is derivable, but requires an appropriate Central Limit Theorem. In that respect the CLT for quadratic forms where the elements can depend on the sample size given in Kelejian and Prucha (1999b) may be useful.

Furthermore, in comparing the procedure of Kelejian and Prucha with ML it seems fair to note that at this point there is no formal result available concerning the asymptotic distribution of the ML estimator for the regression parameters and ρ. Of course, one would expect the usual expressions for the variance covariance

[5]See, e.g., Moran (1948) for details on the Moran I statistics.

matrix of the ML estimator also to apply here under reasonable assumptions, but so far that has not been demonstrated formally.

In view of these issues, this study addresses several concerns. Firstly, we propose moment conditions based on the Cochrane-Orcutt (1949) type transformation of the general spatial model. These moments allow us to define a more general heteroskedastic form. The conditional moments restrictions use second moment information which may improve the efficiency of the estimators. Secondly, following Newey (1986), we construct optimal nonparametric instruments based on nearest neighbor procedures. We show that these instruments are optimal over all possible instruments, especially over those usually suggested in a spatial context. Since we are working with dependent observations, it is suitable to include this dependence in the estimation of the asymptotic covariance matrix of parameters. We provide a consistent nonparametric estimator of sample autocovariance functions for an irregularly spaced spatial stochastic process, and we show that this estimator converges in probability. We then derive an estimator for the asymptotic covariance matrix, and we prove that this estimator is consistent in norm L_2. Finally we investigate the asymptotic distribution of the GMM estimator.

The study is organized as follows. Section 2 introduces the model and describes the conditional second moment restrictions. The model is based on a Cochrane-Orcutt type transformation of the mixed regressive spatial autoregressive specification with heteroskedasticity of a general form. Section 3 presents the estimators and gives conditions for its consistency. Section 4 provides concluding remarks.

2 Conditional spatial autoregressive model

We consider the same structure as equation (1). Let $z_n = (x_n, W_n y_n)$ and $\delta = (\beta', \lambda)'$ be the vector of parameters in the first equation of (1). Then, relation (1) can be rewritten more compactly as

$$\begin{aligned} y_n &= z_n \delta + \epsilon_n, \\ \epsilon_n &= \rho M_n \epsilon_n + u_n. \end{aligned} \tag{8}$$

If we assume that $E(u_n) = 0$, $E(u_n u_n') = \sigma_u^2$, then we have $E(\epsilon_n \epsilon_n') = \sigma_u^2 (I - \rho M_n)^{-1}$ $(I - \rho M_n')^{-1}$, implying a form of spatial heteroskedasticity related to spatial error models.[6] Applying a Cochrane-Orcutt type transformation to this model, that is

[6]The invertibility of forms $I - \lambda W_n$ can be ensured by several conditions. For example, if all eigenvalues of W_n are less than or equal to one in absolute value, $|\lambda| < 1$ implies that all eigenvalues of λW_n are strictly below one in absolute value, which ensures that $(I - \lambda W_n)^{-1} = \sum_{i=0}^{\infty} \lambda^i W_n^i$.

first replacing ϵ_n by $(I-\rho M_n)^{-1}u_n$ in the first part of (8), then premultiplying the resulting relation by $(I-\rho M_n)$ yields,

$$y_n = f(\bar{x}_n, \gamma_0) + u_n, \tag{9}$$

with

$$f(\bar{x}_n, \gamma_0) = \rho M_n y_n + (z_n - \rho M_n z_n)\delta,$$

where $\gamma = (\delta', \rho)'$, $\bar{x}_n = (M_n y_n, z_n, M_n z_n)$. Now set $\bar{z}_n = (y_n, \bar{x}_n)$ and suppose that there is a known skedastic function $h(\bar{x}_n, \gamma, \eta)$ of γ with an additional parameter η such that

$$E(u_n|\bar{x}_n) = 0, \quad E(u_n^2|\bar{x}_n) = h(\bar{x}_n, \gamma, \eta). \tag{10}$$

Let us denote by $\varphi(\bar{z}_n, \theta_0)$ an $s \times 1$ residual vector of functions with $\theta = (\gamma', \eta)' = (\beta', \rho, \lambda, \eta)'$ being a $q \times 1$ vector of parameters. The skedastic function $h(\bar{x}_n, \gamma, \eta)$ could be used to construct a weighted least squares estimator for the parameters. In order to incorporate the heteroskedasticity information in estimating the model where only the conditional first and second moments are restricted, we add the conditional variance residual as an additional conditional moment restriction specifying

$$\varphi(\bar{z}_n, \theta_0) = \begin{bmatrix} y_n - f(\bar{x}_n, \gamma) \\ \{y_n - f(\bar{x}_n, \gamma)\}^2 - h(\bar{x}_n, \gamma, \eta) \end{bmatrix}. \tag{11}$$

Remark 1 *We verify easily that (11) is a moment condition. Indeed, denote the first block of (11) as $\varphi_{11} = u_n$ and the second block as $\varphi_{12} = u_n^2 - h(\cdot)$. Given the assumptions of the model we have $E(\varphi_{11}|\bar{x}_n) = E(u_n|\bar{x}_n) = 0$ and $E(\varphi_{12}|\bar{x}_n) = E(u_n^2|\bar{x}_n) - h(\bar{x}_n, \gamma, \eta) = 0$. The conditional second moment restrictions model considered here is one where the true distribution of data satisfies*

$$E\left[\varphi(\bar{z}_n, \theta_0)|\bar{x}_n\right] = (0, 0)', \tag{12}$$

where θ_0 denotes the true value of the parameters.

The conditional moment restriction (12) implies that $\varphi(\bar{z}_n, \theta_0)$ is uncorrelated with functions of $\bar{x}_n$, which yields an infinity of unconditional moment restrictions $E[A(\bar{x}_n)\varphi(\bar{z}_n, \theta_0)] = (0, 0)'$. Such restrictions can be used to estimate θ_0 by setting the sample cross-product of $\varphi(\bar{z}_n, \theta_0)$ with functions of $\bar{x}_n$ close to zero. As pointed

The claim that all eigenvalues of W_n are less than or equal to one in absolute value, given W_n is row standardized, follows from Geršgorin's theorem. See, e.g., Horn and Johnson (1985, pp. 296–344) for details on this purpose.

out by Newey (1986), the additional conditional moment restriction that may be exploited by instrumental variables estimators using residuals (11) will result in estimators that are at least as asymptotically efficient as the heteroskedasticity corrected least squares estimator, and more efficient in some cases. To state this estimator resulting from the use of (11), let $A(\bar{x}_n)$ denote an $r \times s$ matrix of functions of $\bar{x}_n$, with $r > q$. Then $E[A(\bar{x}_n)\varphi(\bar{z}_n, \theta_0)] = (0,0)'$, which suggests a GMM estimator or a nonlinear IV estimator. This estimator is obtained by minimizing a quadratic form in the sample moments. Let $S_n \overset{\text{a.s.}}{\to} S_0$ be an $r \times r$ positive definite symmetric matrix and $\psi(\theta) = \sum_{i=1}^{n} A(\bar{x}_{n,i})\varphi(\bar{z}_{n,i}, \theta)/n$. The GMM estimator is then:

$$\hat{\theta} = \arg\min_{\theta \in \times} \psi(\theta)' S_n \psi(\theta). \tag{13}$$

To discuss the asymptotic efficiency of $\hat{\theta}$, one needs to compute its asymptotic covariance matrix. In the case of independent and identically distributed (i.i.d.) observations, this is straightforward. With dependent data, as is the case here, the task is more cumbersome. The expression of the asymptotic covariance matrix will look the same as that for i.i.d. observations, but must include sample spatial autocovariances.

First, apply the usual mean-value expansion argument to the first-order conditions

$$\frac{\partial}{\partial \theta'} \psi(\hat{\theta})' S_n \psi(\hat{\theta}) = 0,$$

for $\hat{\theta}$. Then suppose that a uniform law of large numbers leads to $\partial\psi(\hat{\theta})/\partial\theta' \overset{\text{p}}{\to} E[A(\bar{x}_n)\partial\varphi(\bar{z}, \theta)/\partial\theta'] =: \Gamma_n$ for any $\hat{\theta} \overset{\text{p}}{\to} \theta_0$. Furthermore, assume $\{\bar{z}_{n,i}\}_{i=1}^{n}$ to be spatially dependent observations. A corollary of the Lindeberg-Feller central limit theorem for triangular arrays (see, e.g., Billingsley (1979, p. 319)) then gives $\sqrt{n}\psi(\theta_0) \overset{\text{d}}{\to} \mathcal{N}(0, V_n)$, where

$$V_n = E\left[A(\bar{x}_n)\frac{\partial}{\partial\theta'}\varphi(\bar{z}, \theta)\frac{\partial}{\partial\theta'}\varphi(\bar{z}, \theta)' A(\bar{x}_n)'\right].$$

Expanding $\psi(\hat{\theta})$ around θ_0, solving for $\sqrt{n}(\hat{\theta} - \theta_0)$, replacing the estimated average by probability limits and applying the continuous mapping theorem (see, e.g., Serfling (1980, p. 24)) yields the asymptotic normal distribution and asymptotic covariance matrix:

$$n(\hat{\theta} - \theta_0) = -\left(\Gamma_n' S_0 \Gamma_n\right)^{-1} \Gamma_n' S_0 \sqrt{n}\psi(\theta_0) + o_p(1),$$

and

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{\text{d}}{\to} \mathcal{N}\left(0, \left(\Gamma_n' S_0 \Gamma_n\right)^{-1} \Gamma_n' S_0 V_n S_0 \Gamma_n \left(\Gamma_n' S_0 \Gamma_n\right)^{-1}\right). \tag{14}$$

The above asymptotic covariance matrix depends on both S_0 and $A(\bar{x}_n)$. As shown by Hansen (1982), the optimal choice of S_0 that minimizes the asymptotic variance is $S_0 = V_n^{-1}$. This can be computed using $\hat{S}_0 = \hat{V}_n^{-1}$ as a consistent estimator $\hat{V}_n$ of V_n.

As we have pointed out earlier, V_n might account for the dependence between observations and include autocovariances between $A(\bar{x}_n)\varphi(\bar{z}_n, \theta_0)$. One result of this chapter proposes in the next section a nonparametric estimation to fit the sample autocovariances. We show that this estimator is consistent.

Our second focus is the optimal asymptotic variance minimizer selection of instruments $A(\bar{x}_n)$. The material used here is in direct line of Newey (1986). To describe this choice, let

$$D(\bar{x}_n, \theta_0) = E\left[\frac{\partial}{\partial \theta_0}\varphi(\bar{z}_n, \theta_0)|\bar{x}_n\right], \tag{15}$$

and

$$\Omega(\bar{x}_n \theta_0) = E\left[\varphi(\bar{z}_n, \theta_0)\varphi(\bar{z}_n, \theta_0)'|\bar{x}_n\right]. \tag{16}$$

From (9), (11) and (15), the components of $D(\bar{x}_n, \theta_0)$ are computed as

$$\begin{aligned}
E\left[\frac{\partial}{\partial \beta'}f(\bar{x}_n, \gamma)\right] &= E\left[\frac{\partial}{\partial \lambda}f(\bar{x}_n, \gamma)\right] = z_n(I - \rho M_n), \\
E\left[\frac{\partial}{\partial \rho}f(\bar{x}_n, \gamma)\right] &= E\left[M_n(y_n - z_n\delta)\right] = 0, \\
E\left[\frac{\partial}{\partial \eta}f(\bar{x}_n, \gamma)\right] &= 0,
\end{aligned}$$

where the expectations are taken with respect to $\bar{x}_n$. We obtain

$$D(\bar{x}_n, \theta_0) = \begin{bmatrix} z_n(I - \rho M_n) & 0 & z_n(I - \rho M_n) & 0 \\ \frac{\partial}{\partial \beta'}h(\bar{x}_n, \gamma, \eta) & \frac{\partial}{\partial \rho}h(\bar{x}_n, \gamma, \eta) & \frac{\partial}{\partial \lambda}h(\bar{x}_n, \gamma, \eta) & \frac{\partial}{\partial \eta}h(\bar{x}_n, \gamma, \eta) \end{bmatrix}. \tag{17}$$

The asymptotic covariance matrix of $\hat{\theta}$ is then

$$\Phi = \left[A(\bar{x}_n)D(\bar{x}_n)\right]^{-1}E\left[A(\bar{x}_n)\Omega(\bar{x}_n)A(\bar{x}_n)'\right]\left[D(\bar{x}_n)'A(\bar{x}_n)'\right]^{-1}, \tag{18}$$

where θ_0 is omitted in notations in order to ease presentation (we also omit it in the sequel when there is no confusion). Amemiya (1977) showed that the optimal instruments are

$$A^*(\bar{x}_n) = FD(\bar{x}_n)'\Omega(\bar{x}_n)^{-1}, \tag{19}$$

where F is a nonsingular matrix of constants which can be chosen to be equal to the identity matrix; see Chamberlain (1987) for a proof of the IV formulation used here. For this choice of instruments, (18) reduces to

$$\Phi^*_{A^*} = \left(E\left[D(\bar{x}_n)'\Omega(\bar{x}_n)^{-1}D(\bar{x}_n)\right]\right)^{-1}. \tag{20}$$

A fact that is stated below is that the instruments defined by (19) are optimal over all possible instruments and as a result, over those frequently used in the studies on spatial econometrics. The reason is that (19) produces optimal instruments with respect to the criterion of the smallest asymptotic variance within the class of IV estimators. Before stating this result, let us make the following remarks.

Remark 2 *Generally the optimal instruments $A^*(\bar{x}_n)$ are not feasible as they depend on unknown parameters. An estimator $\hat{A}^*(\bar{x}_n)$ of $A^*(\bar{x}_n)$ can then be used instead with $\hat{S}_0 = \left(\sum_{i=1}^n \hat{A}^*(\bar{x}_{n,i})\hat{A}^*(\bar{x}_{n,i})'/n\right)^{-1}$. The resulting IV estimator is*

$$\hat{\theta} = \underset{\theta \in \times}{arg\ min} \bar{\psi}(\theta)'\hat{S}_0^{-1}\bar{\psi}(\theta), \tag{21}$$

with $\bar{\psi}(\theta) = \sum_{i=1}^n \varphi(\bar{z}_{n,i}, \theta)\hat{A}^*(\bar{x}_{n,i})/n$.

Remark 3 *If y_n is continuous, it seems reasonable to make the following assumptions on the third and fourth moment: $E(u_n^3|\bar{x}_n) = 0$ and $E(u_n^4|\bar{x}_n) = \kappa_0 h(\bar{x}_n, \gamma_0, \eta_0)^2$. Two points should then be stressed. First, the question whether additional moment restrictions can improve the estimators efficiency is answered by comparing the asymptotic variance of the heteroskedasticity corrected least squares estimator*

$$\left(E\left[\{E(u_n^2|\bar{x}_n)\}^{-1}\frac{\partial}{\partial\gamma'}f(\bar{x}_n, \gamma)\frac{\partial}{\partial\gamma'}f(\bar{x}_n, \gamma)'\right]\right)^{-1},$$

with the block of the bound of (20) corresponding to γ. As noticed by Newey (1986), the two are equal if $E(u_n^3|\bar{x}_n) = 0$ and γ does not enter the variance function. In other words, the conditional moment bound will in general be lower than the asymptotic variance of the optimally weighted least squares if either $E(u_n^3|\bar{x}_n) \neq 0$ or γ does enter the variance function.

Remark 4 *Secondly, the fourth moment specification is more general than assuming the error to be normally distributed as it allows for a free parameter κ_0 that quantifies the degree of kurtosis.*[7] *If we have consistent estimates $\hat{\gamma}$ and $\hat{\eta}$ for γ and η, κ_0 can be estimated by the sample variance of $\{y_n - f(\bar{x}_n, \hat{\gamma})\}^2/h(\bar{x}_n, \hat{\gamma}, \hat{\eta})$. The estimated optimal instruments can then be formed as $\hat{A}^*(\bar{x}_n) = \hat{D}(\bar{x}_n)'\hat{\Omega}(\bar{x}_n)^{-1}$ with $\hat{D}(\bar{x}_n) = \hat{D}(\bar{x}_n\hat{\theta})$ and*

$$\hat{\Omega}(\bar{x}_n) = \begin{bmatrix} h(\bar{x}_n, \hat{\gamma}, \hat{\eta}) & 0 \\ 0 & \hat{\kappa}h(\bar{x}_n, \hat{\gamma}, \hat{\eta})^2 \end{bmatrix} \tag{22}$$

[7] Note however that this general specification comes with the strong assumption that the free parameter κ_0 is constant.

The zero off-diagonal block in (22) follows from the fact that $\mathrm{cov}[(u_n, u_n^2)|\bar{x}_n] = E[(u_n u_n^2)|\bar{x}_n] = 0$. To establish our first proposition, we make use of the following assumptions.

Assumption 1 *There exists a set of instruments $H_n^*(\bar{x}_n)$ formed as the orthogonal projection onto the space spanned by the columns of H_n subset of $(X_n, W_n X_n, W_n^2 X_n, \cdots, M_n X_n, M_n W_n X_n, M_n W_n^2 X_n, \cdots)$.*

Assumption 2 *There exists an asymptotic covariance matrix $\Phi^*_{H_n^*}$ associated with the instruments H_n^* verifying usual regularity conditions such that the estimator has the following asymptotic distribution*

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \Phi^*_{H_n^*}\right), \tag{23}$$

with

$$\Phi^*_{H_n^*} = \left(G_n' P_n G_n\right)^{-1} G_n' P_n V_n P_n G_n \left(G_n' P_n G_n\right)^{-1},$$

where

$$G_n = E\left[H_n^*(\bar{x}_n) \frac{\partial}{\partial \theta} \varphi(\bar{z}_n, \theta_0) | \theta\right],$$

and

$$V_n = E\left[H_n^*(\bar{x}_n) \varphi(\bar{z}_n, \theta_0) \varphi(\bar{z}_n, \theta_0)' H_n^*(\bar{x}_n)'\right].$$

Assumption 1 is a general formulation of instruments typically suggested in spatial literature; see e.g. Kelejian and Prucha (1999a) and Anselin and Bera (1998). As discussed in the introduction, these instruments consist mainly in spatial lags of explanatory variables and a powers of spatial weighting matrices. This assumption is not strong in that we do not need the matrix of instruments to have full column rank (at the expense of working with generalized inverses). Indeed, to the matrix of instruments, we will associate an asymptotic covariance matrix of parameter estimators. So, we only need the asymptotic covariance matrix to be invertible. Assumption 2 is a standard one which states the regularity conditions under which there exists an asymptotic covariance matrix $\Phi^*_{H_n^*}$ for a consistent $\tilde{\theta}$ of θ_0, based on instruments defined by $H_n^*(\bar{x}_n)$ such that the distribution in (23) holds. These two assumptions allow us to compare the asymptotic covariance matrices $\Phi^*_{H_n^*}$ and $\Phi^*_{A_n^*}$.

Proposition 1 *Given Assumptions 1 and 2 and the maintained assumptions of the model, $\Phi^*_{A_n^*} << \Phi^*_{H_n^*}$.*

Proof. See the appendix.

Proposition 1 states that the mechanism of selection of the instruments leads to the best choice of possible instruments. This result is of two practical interests. First, in the spatial context, we do not need to compute more than one spatial lag of the explanatory or/and dependent variables to be used as instruments. In the case where $W_n = M_n$ one might compute at most the second power of the spatial weighting matrix. The problem of higher order spatial weighting matrices, the elements of which may tend towards zero is then avoided. The second practical interest of this result is that, it provides optimal patterns for instruments. We now state the convergence of various estimators used in the analysis.

3 Estimators

In the previous section we saw that to make efficient inferences about the population parameter value θ_0 (such as to compute confidence intervals or test statistics), it is useful to have a consistent asymptotic covariance estimator. Since we are working with spatial autoregressive processes, spatial dependence may be accounted for, and the middle matrix V_n in relation (14) might include autocovariances between $A(\bar{x}_n)\varphi(\bar{z}_n, \theta_0)$ for different spatial units. This slightly complicates the analysis.

3.1 Nonparametric estimation of spatial autocovariances

Contrary to time-series, there is typically no natural order for arranging a spatial sample, due to its multi-directional motion. The spatial autocovariance function cannot then be approximated from the sample analogue as is the case for time-series. We will outline the reason later. At this stage we have two possibilities. First, we can suppose that the underlying data generating process follows from regularly scattered spatial sites. However, in practice spatial data rarely exhibit such a characteristic. Instead, we consider the case of irregularly scattered spatial data which seems more natural.

Let $s \in \mathbb{R}^d$ be a generic data location in a d-dimensional Euclidean space and $\{U(s), s \in \mathbb{D} \subset \mathbb{R}^d\}$ be a spatial stationary process with zero mean and finite variance. One wishes to have a consistent estimation of the associated autocovariance function $C_n(\tau)$. In the sequel, index n will be omitted to ease presentation when it is not necessary.

Definition 1 *For a distance lag $\tau \in \mathbb{R}^d$, we define the autocovariance function as*

$$C_n(\tau) = cov\left[U(s+\tau), U(s)\right] \equiv E\left[U(s+\tau)U(s)\right]. \tag{24}$$

Definition 2 *Let us define the random variable $Z(\tau)$ as*

$$Z(\tau) = U(s+\tau)U(s). \tag{25}$$

The cloud of points denoted N_U constructed from the n realizations $u(s)$ of $U(s)$ allows us to define a cloud of points denoted N_Z associated with the $n(n-1)/2$ realizations $z(\tau)$ of $Z(\tau)$ by setting

$$z(\tau) = u(s+\tau)u(s). \tag{26}$$

The difficulty here is that one has only one observation for a distance τ between two points of the sample. It is then not possible to construct an estimator of $E[Z(\tau)|s]$ as an empirical mean of a large sample as is the case for time-series data. To overcome this issue, we propose to fit the cloud of points N_Z by a smooth function which will be a consistent approximation of relation (24). The intuition is the following.

Assuming $C_n(\tau)$ to be regular and the sample points to be uniformly distributed, $C_n(\tau)$ can be estimated by smoothing the scatter plot N_Z by using the kernel method. Intuitively, the smoothing function should provide a consistent nonparametric estimation for $C_n(\tau)$ which is then robust to spatial correlation of unknown form. Several assumptions are needed to construct this estimator and to prove its consistency.

Assumption 3 *$\{U(s), s \in [a,b]\}$ is a spatial stationary process with zero mean. Furthermore, assume that there exists a spatial autocovariance function $C_n(\tau)$ for $U(s)$ such that $C_n(\tau)$ is $\mathcal{C}^\infty$.*

Assumption 4 *We define the following sets:*
(i) $I = \{i_1, \cdots, i_n\}$ is a set of n coordinates with i_j having a uniform distribution $\mathcal{U}_{[a,b]}$,
(ii) $\Lambda = \{\tau = |i-i'|/(i,i') \in I\}$ with $\max\{0, \tau - \Delta(n)\} < i - i' < \Delta(n) + \tau$, where $\Delta(n)$ denotes the maximum distance or the maximum allowable error,
(iii) $\Psi_n^\tau = \{(i,i') \in I \times I / \, |i-i'| \in B(\tau, \Delta(n)), \tau \in \Lambda\}$, where $B(\tau,\Delta(n))$ is an open ball, and n here indicates the number of coordinates,
(iv) $\lim_{n\to\infty} \Delta(n) = 0$.

Assumption 5 *There exists a function K defined onto $[0, \Delta(n)]$, positive, strictly decreasing, with $K(0) = 1$ and $K(\Delta(n)) = 0$.*

Proposition 2 *Given Assumptions 3–5, a consistent nonparametric estimator $\hat{C}_n(\tau)$ of $C_n(\tau)$ is such that*

$$\hat{C}_n(\tau) = \sum_{(i,i')\in\Psi_n^\tau} K\left(\big||i-i'| - \tau\big|\right) U(i)U(i') \xrightarrow[n\to\infty]{p} C_n(\tau). \tag{27}$$

Proof. See the appendix.

The estimator (27) is nonparametric in that the form of the spatial correlation between the $U(s)$ at different τ does not need to be specified. The interest of this result is that it provides a consistent estimation of spatial covariance of unknown form directly from the n data points. Note how this contrasts with a parametric framework. Indeed, in such a situation, one might estimate $n(n-1)/2$ parameters from n data points. Since it is impossible to estimate parametrically such covariance terms or correlations directly from n data points due to identification issues, it is necessary to impose sufficient constraints on the n by n spatial interaction matrix such that a finite number of parameters characterizing the correlation can be efficiently estimated.

3.2 Variance-covariance matrix estimation

Once we have $\hat{C}_n(\tau)$ at hand, a consistent estimator of V_n can be formed using for example nearest neighbor procedures $(k-nn)$. See e.g. Härdle (1990) for details on the $k-nn$ method. See also e.g. Newey (1990) for the choice of the weights and the choice of the number of nearest neighbors k used here.

First of all note from (27) that $\hat{C}_n(\tau)$ is symmetric by construction. Now, for positive integers m and k with $m > k$, let ϖ_{mk} be constants satisfying $\varpi_{mk} \geq 0$, $\varpi_{mk} = 0$ and $\sum_{m=1}^{k} \varpi_{mk} = 1$. The weights are defined as ω_{ij} for $i \neq j$ with the restriction of uniform boundedness. Then a consistent estimator $\hat{V}_n$ of V_n is

$$\hat{V}_n = \hat{C}_n(0) + \frac{1}{2} \sum_{\substack{(i,j)\in\Psi_n^\tau \\ j\neq i}} \omega\left(||i-j|-\tau|\right) \hat{C}_n(\tau), \tag{28}$$

with $\hat{C}_n(0)$ denoting the variance at zero lags. We may expect that the estimator $\hat{V}_n$ formed by smoothing sample autocovariances with weights $\omega(.)$ that approach one as $n \to \infty$, should be consistent. Note that even if the estimator (28) is obtained in a similar fashion to that of Newey and West (1987), it is different in two respects. First, the dependence concerned here is spatial and not temporal, which leads to the construction of the particular sample autocovariances estimator in (27). Second, the problem of the bound on the number of sample autocovariances does not matter since $\hat{C}_n(\tau)$ is nonparametrically estimated. The consistency of (28) is stated below.

Proposition 3 *Given Assumptions 3–5*

$$\left\{ \hat{C}_n(0) + \frac{1}{2} \sum_{\substack{(i,j)\in\Psi_n^\tau \\ j\neq i}} \omega\left(||i-j|-\tau|\right) \hat{C}_n(\tau) \right\} - V_n \xrightarrow[n\to\infty]{L_2} 0. \tag{29}$$

Proof. See the appendix.

We now provide the nonparametric estimation of the optimal instruments and the asymptotic distribution of $\hat{\theta}$. This is achieved using nearest neighbor procedures to estimate the conditional expectations that make up the optimal instruments and then replace these estimates by the definition of the optimal instruments.

A nearest neighbor estimator of the conditional covariance $\Omega(\bar{x}_{ni})$ at $\bar{x}_{ni}$ that excludes the own observation is formed as

$$\hat{\Omega}(\bar{x}_{n,i}) = \sum_{j=1}^{n} \omega_{ij} \varphi(\bar{z}_{n,j}, \hat{\theta}) \varphi(\bar{z}_{n,j}, \hat{\theta})'. \tag{30}$$

For the specification we are concerned with, i.e. relation (8), the functional form is assumed to be known. As a result, some components of $D(\bar{x}_n, \theta)$ (especially the first row-block) in (17) have known functional form. Assuming the heteroskedastic function $h(.)$ to be known, we can take advantage of this in computing a consistent estimator $\hat{D}(\bar{x}_n, \theta)$ of $D(\bar{x}_n, \theta)$. With these estimators at hand, one can combine them to form an estimator of optimal instruments such as

$$\hat{A}^*(\bar{x}_n) = \hat{D}(\bar{x}_n)' \hat{\Omega}(\bar{x}_n)^{-1}. \tag{31}$$

An efficient estimator of θ_0 can then be derived as described in relation (13). To prove the asymptotic distribution of the GMM estimator, we impose some usual regularity conditions.

Assumption 6 *For a consistent estimator $\hat{\theta}$ of θ in (17), the row and column sums of the element of $D(\bar{x}_n, \theta_0)$ are uniformly bounded in absolute value.*

Assumption 7 *(i) The vector of parameters θ is such that the true value θ_0 is an element of the interior of Θ which is compact; (ii) there exists a neighborhood $\mathcal{N}$ of θ_0 and $\xi(\bar{z}_n)$ such that almost surely, $\varphi(\bar{z}_n, \theta)$ is continuous on Θ, $\mathcal{C}^\infty$ on $\mathcal{N}$, $\sup_{\theta\in\Theta} \|\varphi(\bar{z}_n, \theta)\| \leq \xi(\bar{z}_n)$, $\sup_{\theta\in\Theta} \|\partial\varphi(\bar{z}_n, \theta)/\partial\theta\| \leq \xi(\bar{z}_n)$; (iii) for $l = 1, \cdots, d$, there exists ν_l such that $\xi_l(\bar{z}_n) \equiv \max \|\partial^l \varphi(\bar{z}_n, \theta)/\partial\theta^l\|$ satisfies $E[\xi_l(\bar{z}_n)^{\nu_l}] < \infty$.*

Assumption 8 *(i) $E[A(\bar{x}_n)A(\bar{x}_n)']$ exists and is nonsingular, and there exists a unique solution to $E[A(\bar{x}_n)\varphi(\bar{z}_n, \theta)] = 0$ at $\theta = \theta_0$; (ii) there is a function $K : \mathbb{N} \to \mathbb{R}$ such that $K(n)/n \to 0$ and $K(n)^2/n \to \infty$.*

Proposition 4 *Given Assumptions 6–8 and Propositions 2–4, for $\hat{\theta}$ in equation (13)*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \Phi^*_{A^*_n}\right), \tag{32}$$

where

$$\left(\frac{1}{n}\sum_{i=1}^{n} \hat{D}(\bar{x}_i)'\hat{\Omega}(\bar{x}_i)^{-1}\hat{D}(\bar{x}_i)\right)^{-1} = \Phi^*_{A^*_n} + o_p(1).$$

Proof. See the appendix.

Remark 5 *Observe that by Assumption 8, the estimator in equation (21) is an IV estimator with $\hat{S}_0 = \left(\sum_{i=1}^{n} \hat{A}^*_{n,i}\hat{A}^*_{n,i}/n\right)^{-1} \xrightarrow{p} S_0 = \left(E[A^*A^{*\prime}]\right)^{-1}$. So $\hat{m}'(\theta)\hat{S}_0\hat{m}(\theta)$ converges uniformly in probability toward $m'(\theta)S_0m(\theta)$ which is continuous and has a unique solution at $\theta = \theta_0$.*

Propositions 3 and 4 state the consistency of the nonparametric estimation sample autocovariances and the asymptotic distribution of parameter estimates, respectively. These estimators are conceptually simple in that they rely on well known econometric literature. In this sense, their rationale is obvious. Also, they are relatively simple to implement and can be used for empirical purpose, even in large samples.

4 Conclusion

In this paper, we suggest efficient estimation procedures of the general spatial autoregressive model based on optimal nonparametric instruments. The structure of the model is such that the conditional restrictions use second moment information that allows a general form of heteroskedasticity. To incorporate the dependence between observations at different locations in space, we first propose a consistent nonparametric estimation of sample autocovariance terms for irregularly scattered spatial processes. This is achieved using nearest neighbor procedures to estimate the conditional expectations that make up the optimal instruments, and then replace these estimates into the definition of the optimal instruments.

Two main directions of future work appear promising. First, it should be noted that our interest was in establishing a simple way to compute a consistent estimator for sample autocovariances as defined in (27). It should be of interest to extend the study to the asymptotic properties of this estimator. Secondly, the finite sample properties of estimators in both Propositions 3 and 4 remain to be derived via for example Monte Carlo experiments.

Acknowledgements

This paper is taken from Théophile Azomahou Ph.D. thesis (European Ph.D.) at Université Louis Pasteur of Strasbourg. Part of the study was written when his was visiting the Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain-La-Neuve. We are grateful to François Laisney for his detailed constructive criticisms and continued encouragement in developing this work. We also wish to thank the Econometrics seminar participants at Université Louis Pasteur of Strasbourg. We would especially like to thank Ingmar Prucha for clarification of some aspects of his study co-authored with Harry Kelejian. Discussions with Gaël Giraud are gratefully acknowledged. We are responsible for any remaining mistakes.

Appendix proofs

Proof of Proposition 1

The proof is based on the comparison of the asymptotic covariance matrices $\Phi^*_{A^*_n}$ and $\Phi^*_{H^*_n}$. It is shown that $\Phi^*_{H^*_n} - \Phi^*_{A^*_n}$ is minimized by A^*_n in the sense of the comparison of positive semi-definite matrices over all possible H^*_n. Furthermore, note that since $\Phi^*_{A^*_n}$ does not depend on F, it suffices to show the result for $F = I$ in relation (19).

Let $\vartheta_{H^*_n} = G'PH^*_n\varphi(\bar{z}_n, \theta_0)$ and $\vartheta_{A^*_n} = A^*_n\varphi(\bar{z}_n, \theta_0)$. In all expectations formulae we will use iterated expectations. Then we have $E[\vartheta_{H^*_n}\vartheta_{A^*_n}{}'] = E[G'PH^*_n \varphi(\bar{z}_n, \theta_0)\varphi(\bar{z}_n, \theta_0)'A^*_n{}'] = G'PE[H^*_n\Omega A^*_n{}']$. From relation (19), $D = \Omega A^{*\prime}_n$ and $E[\vartheta_{H^*_n} \vartheta_{A^*_n}{}'] = G'PE[H^*_nD]$. From Assumption 2 it follows that $E[\vartheta_{H^*_n}\vartheta_{A^*_n}{}'] = G'PE[H^*_n \partial\varphi(\bar{z}_n, \theta_0)/\partial\theta] = G'PG$. One can also easily verify that $E[\vartheta_{H^*_n}\vartheta_{A^*_n}{}'] = E[\vartheta_{A^*_n}\vartheta_{H^*_n}{}']$. Now, observe that

$$E[\vartheta_{H^*_n}\vartheta_{H^*_n}{}'] = E[G'PH^*_n\varphi(\bar{z}_n, \theta_0)\varphi(\bar{z}_n, \theta_0)'H^{*\prime}_nP'G] = G'PVPG$$

and

$$E[\vartheta_{A^*_n}\vartheta_{A^*_n}{}']^{-1} = \left[E(E[D'\Omega^{-1}\varphi(\bar{z}_n, \theta_0)\varphi(\bar{z}_n, \theta_0)'\Omega^{-1}D])\right]^{-1}, \tag{33}$$

and from relation (20), is $\Phi^*_{A^*_n}$. Then we can rewrite

$$\Phi^*_{H^*_n} - \Phi^*_{A^*_n} = (G'PG)^{-1}G'PVPG(G'PG)^{-1} - \Phi^*_{A^*_n}, \tag{34}$$

as

$$(E[\vartheta_{H^*_n}\vartheta_{A^*_n}{}'])^{-1}E[\vartheta_{H^*_n}\vartheta_{H^*_n}{}'](E[\vartheta_{A^*_n}\vartheta_{H^*_n}{}'])^{-1} - (E[\vartheta_{A^*_n}\vartheta_{A^*_n}{}'])^{-1}.$$

Finally, setting

$$\zeta = \left(E[\vartheta_{H^*_n}\vartheta_{A^*_n}{}']\right)^{-1}\left\{\vartheta_{H^*_n} - E[\vartheta_{H^*_n}\vartheta_{A^*_n}{}'](E[\vartheta_{A^*_n}\vartheta_{A^*_n}{}'])^{-1}\vartheta_{A^*_n}\right\},$$

we obtain

$$\Phi^*_{H_n^*} - \Phi^*_{A_n^*} = E(\zeta\zeta'), \tag{35}$$

which is positive semi-definite. This proves that $\Phi^*_{A_n^*}$ is a lower bound for the asymptotic variance of all IV estimators considered. $\square$

Proof of Proposition 2

By Assumption 4 we have $\Delta(n) \to 0$. Also, define

$$\lim_{n\to\infty} \Psi_n^\tau = \{(i,i') \in I \times I / \left|i - i'\right| = \tau, \tau \in \Lambda\} \equiv \Psi_\infty^\tau, \tag{36}$$

where we recall that $\Lambda = \{\tau = |i - i'|/(i,i') \in I\}$. Since $i - i' = \tau + \delta_i \in B(\tau, \Delta(n))$, where δ_i is a small variation (an increment), it follows that $\lim_{n\to\infty} |\delta_i| \leq \lim_{n\to\infty}$ $\Delta(n) = 0$. That is, as given in Assumption 5, by definition of function $K(.)$, $\lim_{n\to\infty} K(\delta_i) = 1$. It then follows that

$$\lim_{n\to\infty} \hat{C}(\tau) = \sum_{(i,i')\in\Psi_\infty^\tau} U(i)U(i'). \tag{37}$$

To complete the proof, it remains to be shown that one can select $\Delta(n)$ such that the expectation of the cardinality of Ψ_∞^τ converges toward infinity for $n \to \infty$. The conclusion will follow from the law of large numbers. Let $\Lambda_n^\tau = B(\tau, \Delta(n)) \cap \Lambda$, that is, the set of sample distances between τ and $\tau \pm \Delta(n)$. Since points coordinates are uniformly distributed, so are the associated distances. As a result, the expectation of $\#\Lambda_n^\tau$ is given by

$$E\left(\#\Lambda_n^\tau\right) = \#\Lambda \frac{\mu\left(B(\tau, \Delta(n))\right)}{\mu\left([a,b]\right)}, \tag{38}$$

where $\mu(.)$ denotes Lebesgue's measure and $\#$ is the cardinality symbol. The cardinality of $\#\Lambda$ is computed as

$$\#\Lambda = \binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2}. \tag{39}$$

Now observe that $\mu(B(\tau, \Delta(n)))$ is the size of $B(\tau, \Delta(n))$, that is $2\Delta(n)$, and $\mu([a,b])$ denotes the length of $[a,b]$, that is $b - a$. Then, we obtain

$$E\left(\#\Lambda_n^\tau\right) = \frac{n(n-1)}{2}\frac{2\Delta(n)}{b-a}. \tag{40}$$

Finally, one can choose for example $\Delta(n) = 1/\sqrt{n}$ which implies that

$$\#\Psi_\infty^\tau = \lim_{n\to\infty} E\left(\#\Lambda_n^\tau\right) = +\infty. \square \tag{41}$$

Proof of Proposition 3

For notational convenience, we will suppress the set under the sum symbol. For $i = i'$ we have $\tau = 0$. Then $C_n(0) = \sum K(0)U(i)U(i')$. By Assumption 5, $K(0) = 1$. Then $C_n(0) = \sum_i U^2(i)$. One deduces that $V_n = \sum_i U^2 + \sum_{(i,j)} \omega_{ij}(.)(\sum_{(i,i')} K(.)U(i)U(i'))$. To show that $\hat{V}_n$ converges in probability toward V_n, let us apply Chebychev's inequality, that is for every $\eta > 0$,

$$P\left[\|\hat{V}_n - V_n\| > \eta\right] \leq \frac{E\|\hat{V}_n - V_n\|^2}{\eta^2}. \tag{42}$$

The job turns out to prove that $E\|\hat{V}_n - V_n\|^2 = o_p(1)$. To show this, decompose $\hat{V}_n$ and V_n as $\hat{V}_{n,1} = \hat{C}_n(0)$, $\hat{V}_{n,2} = \hat{C}_n(\tau)$, $V_{n,1} = C_n(0)$ and $V_{n,2} = C_n(\tau)$. Then we have

$$\begin{aligned} \left\|\hat{V}_n - V_n\right\| &= \left\|\hat{V}_{n,1} + \hat{V}_{n,2} - (V_{n,1} + V_{n,2})\right\|, \\ &= \left\|(\hat{V}_{n,1} - V_{n,1}) + (\hat{V}_{n,2} - V_{n,2})\right\|. \end{aligned} \tag{43}$$

Using the triangular inequality we obtain

$$\left\|\hat{V}_n - V_n\right\| \leq \left\|\hat{V}_{n,1} - V_{n,1}\right\| + \left\|\hat{V}_{n,2} - V_{n,2}\right\|. \tag{44}$$

Note that $\hat{V}_{n,1} \xrightarrow{\text{p}} V_{n,1}$ since $C_n(0)$ is a constant limit for $\hat{C}_n(0)$. From Proposition 2 we known that $\hat{V}_{n,2} \xrightarrow{\text{p}} V_{n,2}$. From the last part of (43) and using Minkoswski inequality we have

$$\begin{aligned} \left(\sum_{i=1}^{n} \|\hat{V}_{n,i} - V_{n,i}\|^2/n\right)^{1/2} &= \left(\sum_{i=1}^{n} \|(\hat{V}_{n,1i} - V_{n,i1}) + (\hat{V}_{n,2i} - V_{n,i2})\|^2/n\right)^{1/2} \\ &\leq \left(\sum_{i=1}^{n} \|\hat{V}_{n,1i} - V_{n,1i}\|^2/n\right)^{1/2} \\ &+ \left(\sum_{i=1}^{n} \|\hat{V}_{n,2i} - V_{n,2i}\|^2/n\right)^{1/2} \\ &= o_p(1) + o_p(1) = o_p(1). \end{aligned}$$

The conclusion follows as we have shown that each term of the inequality converges in norm L_2 as $n \to \infty$. $\square$

Proof of Proposition 4

The proof of Proposition 4 makes use of the two following intermediate lemmata.

Lemma 1 $\sum_{i=1}^{n} \|[\hat{A}^*(\bar{x}_{n,i}) - A^*(\bar{x}_{n,i})]\xi(\bar{z}_{n,i})\|^2/n = o_p(1)$.

Proof:
Let $\hat{A}^*_{n,i} = \hat{A}^*(\bar{x}_{n,i})$, $A^*_{n,i} = A^*(\bar{x}_{n,i})$ and $\xi_{n,i} = \xi(\bar{z}_{n,i})$. By Assumption 7 (iii), $E[\xi_1(\bar{z}_n)^2] < \infty$. Then, using respectively, the triangular inequality, the fact for any regular matrices A and B, $\|AB\| \le \|A\|\|B\|$, the Cauchy-Schwarz and the Markov inequalities, we obtain

$$
\begin{aligned}
\left\|\sum_{i=1}^{n}(\hat{A}^*_{n,i} - A^*_{n,i})\xi(\bar{z}_n)/n\right\| &\le \sum_{i=1}^{n} \left\|\hat{A}^*_{n,i} - A^*_i\right\| \|\xi(\bar{z}_n)\|/n, \\
&\le \sum_{i=1}^{n} \left\|\hat{A}^*_{n,i} - A^*_i\right\| \xi_{n,1i}/n, \\
&\le \left(\sum_{i=1}^{n} \left\|\hat{A}^*_{n,i} - A^*_{n,i}\right\|^2/n\right)^{1/2} \left(\sum_{i=1}^{n} \xi^2_{n,1i}/n\right)^{1/2}, \\
&= o_p(1)O_p(1) = o_p(1).
\end{aligned}
$$

The first two inequalities follow from the triangular inequality and the third follows from the Cauchy-Schwarz inequality. The fact that $(\sum_{i=1}^{n} \xi^2_{n,1i}/n)^{1/2}$ is bounded in probability, i.e. is $O_p(1)$, is ensured by the Markov inequality.[8] The conclusion of the Lemma follows as the product of terms $O_p(1)$ and $o_p(1)$ in the same probability space is $o_p(1)$. $\square$

Lemma 2 $\sum_{i=1}^{n}\{\hat{A}^*(\bar{x}_{n,i}) - A^*(\bar{x}_{n,i})\}\varphi(\bar{z}_{n,i},\theta)\sqrt{n} = o_p(1)$.

Proof:
We keep the notations of Lemma 1. Let $\hat{m}(\theta) = \sum_{i=1}^{n} \hat{A}^*_{n,i}\varphi(\bar{z}_{n,i},\theta)/n, m(\theta) = \sum_{i=1}^{n} A^*_{n,i}\varphi(\bar{z}_{n,i},\theta)/n, \hat{m}_\theta(\theta) = \partial\hat{m}(\theta)/\partial\theta$, and $m_\theta(\theta) = \partial m(\theta)/\partial\theta$. Applying an analogue step to that of Lemma 1 we obtain

$$
\begin{aligned}
\sup_{\theta\in\Theta}\|\hat{m}(\theta) - m(\theta)\| &\le \sup_{\theta\in\Theta}\left(\sum_{i=1}^{n} \left\|\hat{A}^*_{n,i} - A^*_{n,i}\right\|/n\right)\left(\sum_{i=1}^{n} \xi^2_{n,i}/n\right), \\
&\le \sup_{\theta\in\Theta}\left(\sum_{i=1}^{n} \left\|\hat{A}^*_{n,i} - A^*_{n,i}\right\|^2/n\right)^{1/2}\left(\sum_{i=1}^{n} \xi^2_{n,i}/n\right)^{1/2}, \\
&= o_p(1)O_p(1) = o_p(1).
\end{aligned}
$$

Furthermore, by the law of large numbers $\sup_{\theta\in\Theta}\|m(\theta) - E[m(\theta)]\| \xrightarrow{\text{P}} 0$. So by the triangular inequality, $\sup_{\theta\in\Theta}\|\hat{m}(\theta) - E[m(\theta)]\| \xrightarrow{\text{P}} 0$. Applying a similar reasoning to

[8] Let us recall the Markov inequality. If X is a random variable and if $E(|X|^r) < \infty$ for $r > 0$ not necessarily an integer, then $P(|X| \ge \eta) < E(|X|^r)/\eta^2$.

the derivatives $m_\theta(\theta)$ of $m(\theta)$ and by Assumption 7, $\sup_{\theta\in\mathcal{N}}\|\hat{m}_\theta(\theta)-E[m_\theta(\theta)]\|\xrightarrow{\text{p}}0$. Using the continuous mapping theorem and Lindeberg-Levy theorems for $\varphi(\bar{z}_n,\theta)$ we have

$$\begin{aligned}\sum_{i=1}^n \hat{A}^*(\bar{x}_{n,i})\varphi(\bar{z}_{n,i},\theta)\sqrt{n} &= \sum_{i=1}^n(\hat{A}^*_{n,i}-A^*_{n,i})\varphi(\bar{z}_{n,i},\theta)\sqrt{n}\\ &\quad+\sum_{i=1}^n A^*(\bar{x}_{n,i})\varphi(\bar{z}_{n,i},\theta)\sqrt{n},\\ &= o_p(1)+\sum_{i=1}^n A^*(\bar{x}_{n,i})\varphi(\bar{z}_{n,i},\theta)\sqrt{n}\xrightarrow{\text{d}}\mathcal{N}(0,\Phi^*_{A^*_n}).\ \square\end{aligned}$$

We are now able to prove Proposition 4. Recall that $\hat{A}^*_{n,i}=\hat{D}'_{n,i}\hat{\Omega}^{-1}_{n,i}$. Then, $\sum_{i=1}^n\|\hat{A}^*_{n,i}-A^*_{n,i}\|^2/n=\sum_{i=1}^n\|\hat{D}'_{n,i}\Omega^{-1}_{n,i}-D_{n,i}\Omega^{-1}_{n,i}\|^2/n$. Suppose that Assumptions 6–8 are satisfied. Let C denotes a generic constant that can takes different values in different appearances. Following Newey (1990), note that the estimates of the optimal instruments take the form $\hat{A}^*_{n,i}=\hat{D}'_{n,i}\hat{\Omega}^{-1}_{n,i}=(\hat{T}_{n,i}+\hat{G}_{n,i})'\hat{\Omega}^{-1}_{n,i}$, where $\hat{T}_{n,i}$ is a matrix of trend terms, and $\hat{G}_{n,i}$ is a matrix consisting of the non-parametric estimates of the trend residuals. By Lemma A.1 in Newey (1990), $\sum_{i=1}^n\|\hat{G}_{n,i}-G_{n,i}\|^2/n=o_p(1)$ and $\sum_{i=1}^n\|\hat{T}_{n,i}-T_{n,i}\|^2/n=o_p(1)$. Then, by Lemma 1 and using the Markov inequality, we obtain

$$\begin{aligned}\sum_{i=1}^n\left\|\hat{A}^*_{n,i}-A^*_{n,i}\right\|^2/n &\le \sum_{i=1}^n C\left(\left\|(\hat{T}_{n,i}-T_{n,i})\hat{\Omega}^{-1}\right\|^2+\left\|(\hat{G}_{n,i}-G_{n,i})\hat{\Omega}^{-1}\right\|^2\right)/n\\ &\le C\|\hat{\Omega}^{-1}\|^2\sum_{i=1}^n\left(\left\|(\hat{T}_{n,i}-T_{n,i})\right\|^2+\left\|(\hat{G}_{n,i}-G_{n,i})\right\|^2\right)/n\\ &= O_p(1)o_p(1)=o_p(1).\end{aligned}$$

We may conclude that $\sum_{i=1}^n\|\hat{A}^*_{n,i}\hat{D}_{n,i}-A^*_{n,i}D_{n,i}\|/n=o_p(1)$. Again, by the law of large numbers $\sum_{i=1}^n A^*_{n,i}D_{n,i}/n=\sum_{i=1}^n D'_{n,i}\Omega^{-1}_{n,i}D_{n,i}/n\xrightarrow{\text{p}}{\Phi^*_{A^*_n}}^{-1}$. The asymptotic distribution in Proposition 4 follows by Lemma 2. $\square$

References

Amemiya, T., 1977. The maximum likelihood and nonlinear three-stage least squares estimator in the general nonlinear simultaneous equations model. Econometrica pp. 955–968.

Amemiya, T., 1985. Advanced Econometrics. Cambridge: Harvard University Press.

Anselin, L., 1980. Estimation Methods for Spatial Autoregressive Structures. Ithaca, NY: Cornell University, Regional Science Dissertation and Monograph Series 8.

Anselin, L., 1988. Spatial Econometrics Methods and Models. Dordrecht: Kluwer.

Anselin, L., Bera, A., 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. in Handbook of Applied Economic Statistics, A. Truchmuche (eds.) pp. 237–289.

Azomahou, T., Lahatte, A., 2000. On the inconsistency of the ordinary least squares estimator for spatial autoregressive processes. Working Paper BETA, Université Louis Pasteur, Strasbourg .

Baltagi, B. H., Li, D., 1999. Prediction in the panel data model with spatial correlation. Texas A&M University, Department of Economics .

Bell, K. P., Bockstael, N. E., 2000. Applying the generalized-moments estimation approach to spatial problems involving microlevel data. The Review of Economics and Statistics pp. 72–82.

Berge, C., 1983. Graphes. Paris: Gauthier-Villars.

Billingsley, P., 1979. Probability and Measure. New York: Wiley.

Case, A., 1991. Spatial patterns in household demand. Econometrica pp. 953 –965.

Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. Journal of Econometrics pp. 305–334.

Cochrane, D., Orcutt, G., 1949. Application of least squares regressions to relationships containing autocorrelated error terms. Journal of the American Statistical Association pp. 32–61.

Conley, T., 1999. Generalized method of moments estimation with cross sectional dependence. Journal of Econometrics pp. 1–45.

Cressie, N., 1991. Statistics for Spatial Data. New York: Wiley-Interscience.

Hansen, L., 1982. Large sample properties of generalized method of moments estimators. Econometrica pp. 1029–1054.

Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press.

Hautsch, N., Klotz, S., 1999. Estimating the neighborhood influence on decision makers: Theory and an application on the analysis of innovation decisions. University of Konstanz .

Horn, R., Johnson, C., 1985. Matrix Analysis. Cambridge: Cambridge University Press.

Kelejian, H., Prucha, I., 1997. Estimation of spatial regression models with autoregressive errors by two-stage least squares procedures: A serious problem. International Regional Science Review pp. 103–111.

Kelejian, H., Prucha, I., 1998. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbance. Journal of Real Estate Finance and Economics pp. 99–121.

Kelejian, H., Prucha, I., 1999a. A generalized moments estimator for the autoregressive parameter in a spatial model. International Economic Review pp. 509–533.

Kelejian, H., Prucha, I., 1999b. On the asymptotic distribution of the moran i test with applications. Working Paper p. 41.

Kelejian, H. H., Robinson, D. P., 1995. Spatial Correlation: A suggested Alternative to the Autoregressive Model. Berlin: Springer-Verlag.

Moran, P., 1948. The interpretation of statistical maps. Journal of the Royal Statistical Society pp. 243–251.

Newey, W. K., 1986. Efficient estimation of models with conditional moments restrictions. Handbook of statistics pp. 419–454.

Newey, W. K., 1990. Efficient instrumental variables estimation of nonlinear models. Econometrica pp. 809–837.

Newey, W. K., West, K. D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica pp. 703–708.

Ord, K., 1975. Estimation methods for models of spatial interaction. Journal of American Statistical Association pp. 120–126.

Serfling, R., 1980. Approximation Theorems of Mathematical Statistics. New York: Wiley.

Ziliak, J. P., Wilson, B. A., Stones, J. A., 1999. Spatial dynamics and heterogeneity in the cyclicality of real wages. The Review of Economics and Statistics pp. 227–236.