Estimating Nonlinearities in Spatial Autoregressive Models

Nicolas Debarsy, Vincenzo Verardi

Estimating Nonlinearities in Spatial Autoregressive Models*

Nicolas Debarsy† and Vincenzo Verardi‡

January 13, 2010

Abstract

In spatial autoregressive models, the functional form of autocorrelation is assumed to be linear. In this paper, we propose a simple semiparametric procedure, based on Yatchew's (1998) partial linear least squares, that relaxes this restriction. Simple simulations show that this model outperforms traditional SAR estimation when nonlinearities are present. We then apply the methodology on real data to test for the spatial pattern of voting for independent candidates in US presidential elections. We find that in some counties, votes for "third candidates" are non-linearly related to votes for "third candidates" in neighboring counties, which pleads for strategic behavior.

Keywords: Spatial econometrics, semiparametric estimations

JEL Classification: C14, C21

*We would like to thank all our colleagues at CRED, ECARES, and CERPE and in particular Marjorie Gassner and Darwin Ugarte for useful comments.

†Corresponding author, CERPE, Facultés Universitaires Notre Dame de la Paix de Namur. E-mail: ndebarsy@fundp.ac.be. Nicolas Debarsy is Doctoral Researcher of the FNRS and gratefully acknowledges their financial support.

‡CRED, Facultés Universitaires Notre Dame de la Paix de Namur; ECARES, CKE, Université Libre de Bruxelles. E-mail: vverardi@fundp.ac.be. Vincenzo Verardi is Associated Researcher of the FNRS and gratefully acknowledges their financial support.

1 Introduction

In spatial econometrics, autoregressive models (or SAR) have been developed to estimate how changes in a given variable spread to the neighborhood (see Anselin, 1988 or LeSage and Pace, 2009, for further details). This effect is generally assumed to be linear, which is restrictive. The objective of this note is to propose a simple estimation method, based on Yatchew's (1998) difference estimator, that allows for non-linear spatial interdependence.

The structure of the paper is the following: after this short introduction we propose in section 2 a procedure to estimate a semiparametric spatial autoregressive model. Furthermore we recommend a test that allows to understand if assuming linearity in the autoregressive component is legitimate. In section 3 we present some simple simulations to assess the performance of the procedure and we present an empirical example based on US presidential elections in section 4. Section 5 concludes.

2 Estimation method

2.1 Nonlinear spatial autoregressive model

The general form of a linear first order spatial autoregressive model is

$$y_i = x_i\theta + \rho W y_i + \varepsilon_i, \quad i = 1, ..., N \tag{1}$$

where y_i is the value taken by the dependent variable and x_i is the row vector of characteristics of individual i. Wy_i measures the value of y in the neighborhood $N_{(i)}$ of individual i and is defined as $Wy_i = \sum_{j \in N_{(i)}} w_{ij} y_j$ where w_{ij} models spatial interactions between i and j. Column vector θ and the autoregressive spatial parameter ρ are the coefficients to be estimated. Given the linearity of (1), a unit change in Wy is associated to a ρ units change in the conditional expectation of y, whatever the value of Wy. This assumption could be relaxed by considering a more general model of the type

$$y_i = x_i\theta + f(Wy_i) + \varepsilon_i, \quad i = 1, ..., N \tag{2}$$

where f is any bounded and differentiable function. This model can be easily estimated extending Yatchew's (1998) semiparametric procedure to the case of a spatial autoregressive model by assuming (i) that Wy is drawn from a continuous distribution with a convex finite support, (ii) that the first derivative of f is bounded by a constant L and (iii) that the ε_i are independent and identically distributed ($i.i.d$) with mean 0 and variance σ^2_ε. It is well-known in the spatial econometric literature that Wy is endogenous. To deal with this, we suggest using Newey et al. (1999) procedure which consists in introducing the estimated residuals ($\hat{\eta}_i$) of a first step equation where the troublesome variable Wy is regressed on the exogenous variables in (2) and a set of instruments. The instruments we use are standard in the spatial econometric literature and are the aggregate value of the exogenous variables measured in the neighborhood (i.e. Wx and W^2x). To simplify notation, we consider $\hat{\eta}_i$ as an additional exogenous variable in the matrix x in (2). If the coefficient associated to $\hat{\eta}_i$ is not significant, the endogeneity bias is negligible.

The rationale underlying the semi-parametric method is simple: suppose that we rearrange the observations by sorting them in increasing order according to variable Wy. By taking first differences, we have:

$$y_i - y_{i-1} = (z_i - z_{i-1})\delta_{diff} + [f(Wy_i) - f(Wy_{i-1})] + (\varepsilon_i - \varepsilon_{i-1}), \quad i = 2, ..., N \tag{3}$$

with $z_i = [x_i\ \eta_i]$, $\delta'_{diff} = [\theta'_{diff}\ \gamma_{diff}]$ and γ_{diff} the coefficient associated to $(\eta_i - \eta_{i-1})$. Increasing the number of observations (which, broadly speaking, means filling the finite support interval of Wy with values) will cause the difference $Wy_i - Wy_{i-1}$ to shrink at a rate of $1/N$. Since the first derivative of f is assumed bounded by L, $|f(Wy_i) - f(Wy_{i-1})| \leq L\,|Wy_i - Wy_{i-1}|$. The shrinkage of $(Wy_i - Wy_{i-1})$ will induce $f(Wy_{i-1})$ to cancel out with $f(Wy_i)$. Reordering and taking differences thus allows to estimate the θ vector of parameters consistently with LS, whatever the functional form of f, as soon as its first derivative is bounded.

Note that this simple estimator is inefficient (it has a Gaussian efficiency of 66.7%). To increase efficiency, Yatchew (1998) suggests to use higher order differences. In our setup, this generalization is written as:

$$\sum_{j=0}^{D_m} d_j y_{t-j} = \left(\sum_{j=0}^{D_m} d_j z_{i-j} \right) \delta_{diff} + \sum_{j=0}^{D_m} d_j f\left(W y_{i-j}\right) + \sum_{j=0}^{D_m} d_j \varepsilon_{i-j}, \quad i = D_m + 1, ..., N \quad (4)$$

where D_m ($\in \mathbb{N}_0^+$) is the order of differencing and $d_0, ..., d_m$ are differencing weights. Two conditions are imposed on $d_0, ..., d_m$. The first is $\sum_{j=0}^{m} d_j = 0$, which ensures that the nonparametric spatial component part is partialled out. The second, $\sum_{j=0}^{m} d_j^2 = 1$, guarantees that the variance of the transformed residual in (4) is σ_ε^2. Yatchew (1998) shows that, with D_m sufficiently large, the estimator approaches asymptotic efficiency.[1]

As far as inference is concerned, Yatchew (1998) shows that $\hat{\theta}_{diff}$ has the approximate sampling distribution

$$\hat{\theta}_{diff} \sim N\left(\theta, \left(1 + \frac{1}{2D_m}\right) \frac{\sigma_\varepsilon^2}{N \sigma_u^2}\right) \quad (5)$$

where σ_u^2 is the conditional variance of z given Wy. The standard errors of the estimated parameters can thus be easily computed. Concerning the relevance of the nonlinear effect of the variable Wy , Yatchew (1998) developed a simple test based on the comparison of the scale of the residuals of the difference equation (s_{diff}^2) with that of the LS regression where the function f is assumed to be linear (s_{lin}^2). The underlying idea of the test is that if nonlinearities exist, a linear approximation of the tested relation will lead to an overestimation of the variance of the residuals.

The proposed test statistic is:

$$V = \frac{\sqrt{D_m N}(s_{lin}^2 - s_{diff}^2)}{s_{diff}^2} \quad (6)$$

[1] When $D_m = 1$, $d_0 = \frac{1}{\sqrt{2}}$ and $d_1 = -\frac{1}{\sqrt{2}}$, equation (4) boils down to equation (3).

which is asymptotically distributed as a $N(0,1)$. A rejection of the null would suggest a non-linear relation between Wy and y.

In our setup, the estimation of s^2_{lin} cannot be based on LS since it is biased and inconsistent when an autoregressive term is present. We therefore estimate it using the residuals of (1), the SAR model, estimated by maximum likelihood.

Finally, since $\hat{\theta}_{diff}$ is a consistent estimator of θ, the relation between y and Wy can be assessed by running a non-parametric estimation of the partialled out $\tilde{y}_i$ (which is $\tilde{y}_i = y_i - x_i\hat{\theta}_{diff}$) on Wy. This variable still contains the information on the spatial dependence of interest, but is filtered of the influence of the control variables. The non-parametric estimator we consider is cubic spline. Alternative such as kernel regression methods can be used. They lead to very similar results.

To assess the performance of the proposed methodology, we present some simple simulations in the following section.

3 Simulations

The four following data generating processes (DGP) are considered:

a) $y_i = x_i\theta + \varepsilon_i$

b) $y_i = 0.75Wy_i + x_i\theta + \varepsilon_i$

c) $y_i = 0.75Wy_i - 0.4\left(Wy_i\right)^2 + x_i\theta + \varepsilon_i$

d) $y_i = \left(\frac{1}{1+\exp(-2Wy_i)} - 0.5\right) + x_i\theta + \varepsilon_i$

where x_i is a 1×3 vector whose elements are drawn from three independent $N(0,4)$, θ is a 3×1 vector of ones and $\varepsilon_i \sim N(0, 0.1)$. The simulated sample size is 300. The x-coordinates are generated from a $U[0,20]$ and the y-coordinates from a $U[0,50]$. Spatial weights are

$$w_{ij} = \begin{cases} \frac{1/b_{ij}}{\sum_j 1/b_{ij}} \text{ if } b_{ij} < \bar{b} \\ 0 \text{ otherwise} \end{cases}$$

where b_{ij} are all pairwise distances. Parameter $\bar{b}$ (the threshold value above which the interaction between i and j is assumed to be negligible) is set to 5. By convention, $w_{ii} = 0$. All the models are fitted assuming $D_m = 1$.

To illustrate the fitting performance of the proposed estimation procedure, we generate four samples according to the DGPs discussed above and present the scatter plots, the non-parametric fit (thick plain line) and the true DGP (thin dashed line) in Figure 1. As expected, the results are unambiguous.

[INSERT FIGURE 1 HERE]

In the case of no spatial autocorrelation (panel a), no clear pattern emerges and the non-parametric curve lies close to the horizontal line (the true DGP). In the three other cases (panels b, c and d), the nonparametric estimation of the autocorrelation matches the true functional form quite well. The last two panels (c and d) shed doubt on the appropriateness of a linear approximation for the spatial component.

To check the performance of the V-statistic in detecting nonlinearities, we replicate the four DGPs described above 1000 times. Each time a new error term is randomly drawn and a new dependent variable is generated. The design space is kept unchanged. We then compute the percentage of rejection of the null (at 5%). Results are presented in Table 1.

Table 1: Test for linearity

Spatial autocorrelation	absent	linear	quadratic	sigmoid
% Rejection	5.2%	4.8%	100%	100%

The null is always rejected in the quadratic and the sigmoid configurations. The size of the test is about 5% for the absent and linear cases. Though we are aware that

these simulations are too simple to provide a clear assessment of the quality and power of the test, it seems that it could be an interesting complementary tool to assess the linearity of potential spatial autocorrelation.

4 Application

In this section, we present an illustrative example of the nonlinear SAR model. The objective of the analysis is to study the voting behavior for independent candidates (focusing on US presidential elections of 2000) in a given county as a function of the votes cast by this candidate in neighboring counties (which are assumed to be well anticipated by electors). The hypothesis tested is that voters will not vote for the third candidate if they believe that it might help the candidate they dislike the most win the elections (as occurred, for instance, in the first round of the French presidential elections of 2002 when Le Pen, the extreme right wing candidate, obtained 16.86% of the votes and qualified for the second round at the expense of the socialist candidate).

Hence, if interested voters anticipate that the third candidate will collect a limited number of votes, they will vote for him to declare their dissatisfaction with the political establishment. Furthermore, the resulting share of votes is expected to increase jointly with the share of votes in the neighborhood as the message sent will be stronger. However, if voters perceive that the third candidate will obtain so many votes that it results in jeopardizing the political scenario they will stop voting for him. We therefore expect to observe a concave-shaped spatial autoregressive component in the vote for the third candidate. To test for this, we estimate the relation

$$y_i = x_i\theta + f(Wy_i) + \varepsilon_i \text{ for } i = 1, ..., N \tag{7}$$

where variable y is the log of the vote share cast by outsider candidates and i indexes counties. The control variables (x) are those generally considered in this type of regressions i.e. (i) the log of the votes share of independents in the previous elections,

(ii) the total population, (iii) the log of the vote share for Republicans, (iv) the average per capita income, (v) the log of the ratio between democrats and republican votes for the previous elections and (vi) the ethnic composition (i.e. the proportion of blacks, whites, asians, native indians others and pacific island which is the reference).[2] The order of difference considered is 10 and the weighting matrix is defined as follows: counties located in different states do not interact.[3] Within states, we assign a spatial weight proportional to the inverse of the distance between counties' centroids which implicitly assumes that individuals are better informed on closer counties. Data come from Polidata, a national demographic and political data consulting firm in the US.

We concentrate our analysis only on swing states (since the described theory only holds in these)[4] and focus on Oregon, Pennsylvania, Tennessee and Wisconsin where the concaved-shaped relation is clear.[5]

[INSERT FIGURE 2 HERE]

For the state of Pennsylvania, we observe that the sincere voting behavior occurs as long as Wy is smaller than 2.7%. Indeed, in this situation, we observe that an increase in the vote share for independents in the neighborhood induces an increase in the vote share for independents in the considered county. This is probably because voters think that their vote will strengthen the message conveyed by the neighbors on the dissatisfaction with the political establishment. However, when the votes for outsiders become more numerous in the neighborhood (around 2.7%), the "votes for change" start decreasing. This could be explained by the fact that voters realize that they should vote strategically to prevent the candidate they dislike the most to win the elections. This theory seems to hold well in Oregon and Tennessee, where the concaved-shaped relation is evident. For Wisconsin, we observe a similar behavior except that

[2]Including the log of the vote share for the Democrats instead does not affect results.

[3]This assumption relies on the majoritarian system in place.

[4]In 2000 swing states were Tennessee, Nevada, Ohio, Missouri, New Hampshire, Florida, New Mexico, Wisconsin, Iowa, Oregon, Minnesota and Pennsylvania.

[5]The graphs for all the other States are available from the authors upon request.

a first threshold is observed when Wy is around 2.3% and a decrease takes place at 2.45%.

As far as inference is concerned, the V-statistic is larger than 2 (in absolute value) in all States. Its value is 4.19 for Oregon, 10.10 for Pennsylvania, 25.53 for Tennessee and 40.39 for Wisconsin. This clearly rejects the linearity assumption in all the cases considered, which confirms the impression given by the graphs.

5 Conclusion

In spatial econometrics, the SAR is one of the most commonly used models. In this paper, we propose a simple generalization of it for the case of a non-linear spatial autoregressive component. We present some simulations and a simple empirical application to show the usefulness of the procedure.

References

[1] Anselin L., 1988, Spatial Econometrics, Methods and Models. (Kluwer Academic Publishers, Dordrecht).

[2] LeSage J. and K. Pace, 2009, Introduction to Spatial Econometrics. (CRC Press/Taylor and Francis Group, London).

[3] Newey, W.K., Powell, J.L. and F. Vella, 1999, Nonparametric estimation of triangular simultaneous equation models, Econometrica, 67, 565-603.

[4] Yatchew, A., 1998, Nonparametric Regression Techniques in Economics, Journal of Economic Literature, 36(2), 669-721.

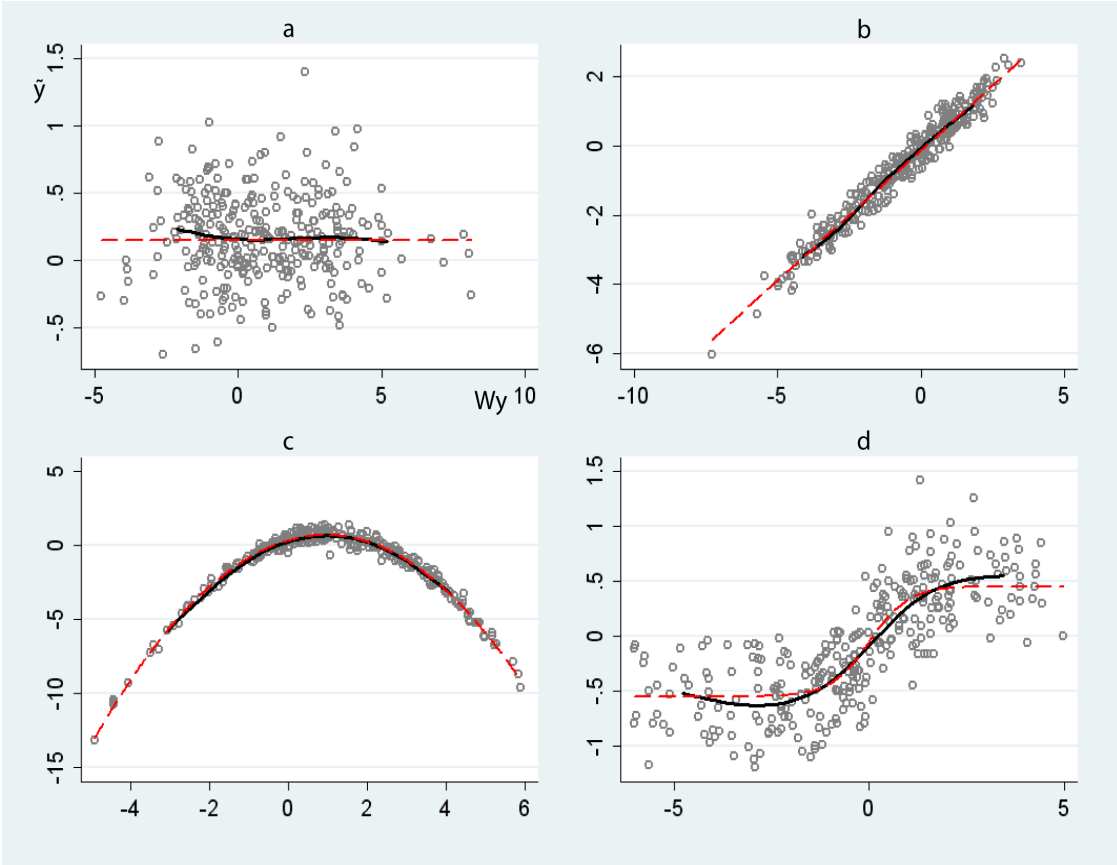

Figure 1: Non-parametric fit of spatial autocorrelation

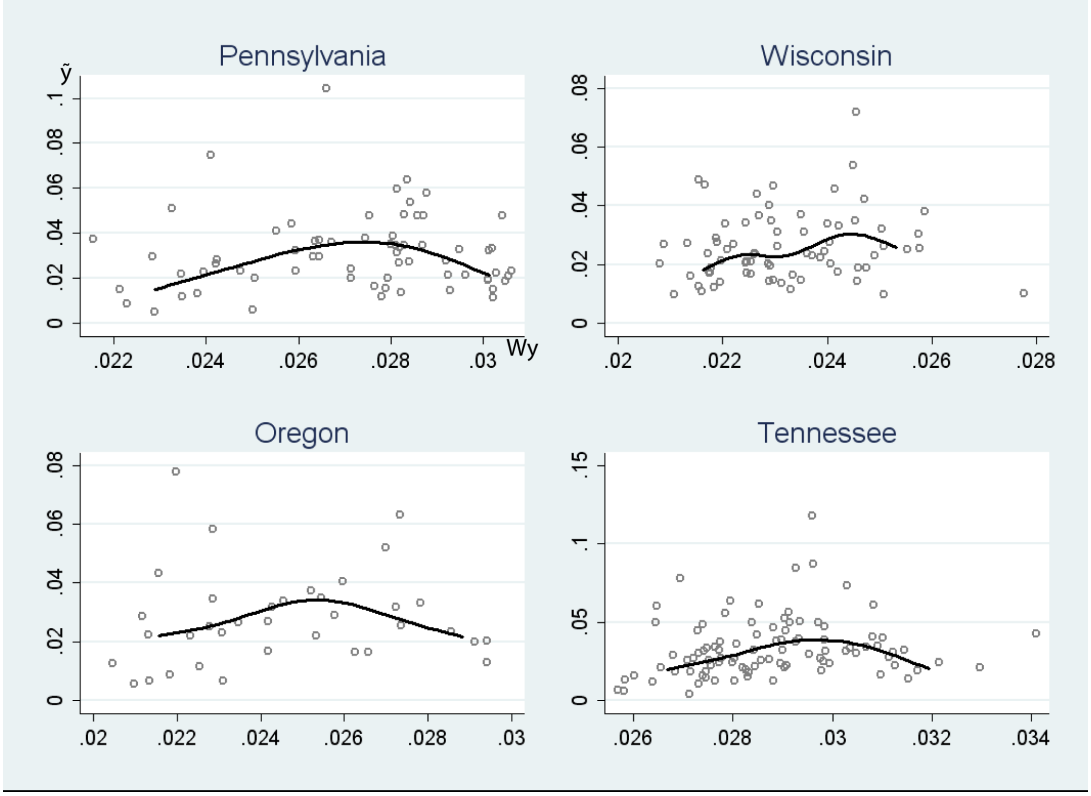

Figure 2: Nonlinear SAR by state