ARTICLE IN PRESS

Economic Modelling xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.elsevier.com/locate/econmod

# Statistical inference of partially linear varying coefficient spatial autoregressive models

Chuanhua Wei*, Shuang Guo, Shufen Zhai

*Department of Statistics, School of Science, Minzu University of China, Beijing 100081, PR China*

ARTICLE INFO

*JEL classification:*
C12
C13
C14
C21

*Keywords:*
Partially linear varying coefficient model
Spatial autoregressive models
Local linear method
Profile quasi-maximum likelihood approach
Generalized likelihood ratio test

ABSTRACT

This paper proposes a semiparametric partially linear varying coefficient spatial autoregressive model, which is a generalization of standard spatial autoregressive model and partially linear spatial autoregressive model. To estimate the unknown spatial lag parameter, constant coefficients and coefficient functions, a profile quasi-maximum likelihood approach based on the local-linear method is introduced. To test the existence of the spatial effects, a generalized likelihood ratio test statistic is proposed, and a residual-based bootstrap procedure is used to derive the p-value of the test. Some simulations are conducted to examine the performance of our proposed procedures and the results are satisfactory. Furthermore, a real-world example is given to demonstrate the application of the proposed procedures.

## 1. Introduction

Spatial econometrics is a subfield of econometrics that deals with the incorporation of spatial effects in econometric methods. Two broad classes of spatial effects may be distinguished, referred to as spatial dependence and spatial heterogeneity, see Anselin (1988) and Lesage and Pace (2009) for details. One useful approach to deal with spatial dependence is the spatial autoregressive model, which adds a weighted average of nearby values of the dependent variable to the base set of explanatory variables. The traditional spatial autoregressive model is very attractive due to its simplicity in estimation and interpretation, inference and application of this model can be found in Cliff and Ord (1973); Anselin (1988); Kelejian and Prucha (1999), and Lee (2004, 2007)). However, the parametric structure of spatial autoregressive model is highly sensitive to model misspecification, so it may not be adequate in many complex situations. To capture the underlying relationships between the response variables and their associated covariates, some nonparametric and semiparametric spatial autoregressive models have been proposed in recent years. Nonparametric and partially linear spatial autoregressive model have been introduced by Gress (2004). Su and Jin (2010) developed a profile quasi-maximum likelihood estimation approach for partially linear spatial autoregressive model, and studied the asymptotic properties of the proposed estimators. Li and Mei (2013, 2016)) applied the generalized likelihood ratio test method and the bootstrap procedure to test the nonparametric component and the parametric component of the partially linear spatial autoregressive models. Su (2012) studied a nonparametric spatial autoregressive model that the spatially lagged response variable enters the model linearly while the covariates enter the model nonparametrically. Malikov and Sun (2015) proposed a flexible semiparametric varying coefficient spatial autoregressive model in which both spatial lag parameter and regression coefficients are permitted to be nonparametric functions of some contextual variables to allow for potential nonlinearities and parameter heterogeneity in the spatial relationship. Sun (2016) studied a spatial varying coefficient models with nonparametric spatial weights, which allows the data to determine unknown spatial weights.

In this paper, we will propose a new semiparametric spatial autoregressive model. The main motivation to propose the model is to analyse the Boston housing data set, which was given in Harrison and Rubinfeld (1978), corrected for a few minor errors by Gilley and Pace (1996) and augmented with longitude and latitude by Pace and Gilley (1997). The data set consists of the median value of owner-occupied homes in 506 census tracts in the Boston Standard Metropolitan Statistical Area in 1970, together with 13 related variables which might explain the variation of housing value. For this data set, Fan and Huang (2005) proposed the following partially linear varying coefficient model

$$Y_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\alpha}(U_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1.1}$$

* Corresponding author.
*E-mail address:* chweisd@163.com (C. Wei).

http://dx.doi.org/10.1016/j.econmod.2017.04.015
Received 4 July 2016; Received in revised form 28 February 2017; Accepted 20 April 2017

where $Y_i$, $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^{\mathrm{T}}$, $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iq})^{\mathrm{T}}$ and $U_i$ are the observations of the response and associated explanatory variables. $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ is a vector of $p$-dimensional unknown parameters and $\boldsymbol{\alpha}(\cdot) = (\alpha_1(\cdot), \alpha_2(\cdot), \ldots, \alpha_q(\cdot))^{\mathrm{T}}$ is a $q$-dimensional vector of unknown functions, $\varepsilon'_i s$ are independent and identically distributed random errors with zero mean and finite variance $\sigma^2$. On the other hand, the presence of both spatial dependence (also known as spatial autocorrelation) or spatial heterogeneity (also referred to as spatial nonstationary) in the housing market has been emphasized by a large numbers of literatures. Some type spatial autoregressive models were applied to analysis the above dataset by Pace and Gilley (1997), Su and Yang (2001) and Li and Mei (2016). Now, we combine the above two types of models and build the following partially linear varying coefficient spatial autoregressive model

$$Y_i = \rho \sum_{j=1}^{n} w_{ij} Y_j + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\alpha}(U_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1.2}$$

where $\mathbf{W} = (w_{ij})$, $1 \le i, j \le n$ is a specified $n \times n$ spatial weight matrix. The definition of spatial weight matrix $\mathbf{W}$ is a fundamental issue in using spatial econometrics method. They are based on the geographic arrangement of the observations, or contiguity. Weights are non-zero when two locations share a common boundary, or are within a given distance of each other. More generally, $\mathbf{W}$ matrices can be specified based on geographical distance decay, economic distance, the structure of a social network, more examples can be found in Anselin (1988) and Lesage and Pace (2009).

Model (1.2) is flexible enough to include a variety of existing models. When $\boldsymbol{\alpha}(\cdot) = \boldsymbol{\alpha}$, model (1.2) becomes the standard spatial autoregressive model. When $q$=1 and $\mathbf{Z}_i = 1$, the model is the partially linear spatial autoregressive model studied by Su and Jin (2010). When $\mathbf{X}_i = 0$, the model is varying coefficient spatial autoregressive model. When $\rho = 0$, the model reduces to the partially linear varying coefficient models, which was studied by Fan and Huang (2005) and have been applied in many different applications, see Li and Racine (2007); Cai et al. (2009); Cai and Xiong (2012); Sun et al. (2013); Wang et al. (2009) and references therein.

In this paper, we consider the estimating and testing problem of the model (1.2). Following Su and Jin (2010), combining the profile least-squares approach of Fan and Huang (2005) for the standard partially linear varying coefficient and quasi-maximum likelihood approach for the traditional spatial autoregressive models, the profile quasi-maximum likelihood approach is introduced to estimate the unknown spatial lag parameter, constant coefficients and coefficient functions of model (1.2). The second question that we addressed is to test the existence of the spatial effects. This leads to the following testing problem

$$H_0\colon \rho = 0 \quad \text{VS} \quad H_1\colon \rho \neq 0. \tag{1.3}$$

For the standard linear spatial autoregressive models, likelihood ratio test and Rao's Score (Lagrange Multiplier ) test can be applied for the problem (1.3); details can be found in Anselin (1988). However, the test problem (1.3) is a semiparametric hypothesis versus another semiparametric hypothesis testing problem. Many traditional tests cannot be directly applied to the above hypothesis. For this kind of testing problem, Li and Mei (2016) applied the generalized likelihood ratio (GLR) technique of Fan and Huang (2005) and Fan et al. (2001) to the testing problems in the parametric component of the partially linear spatial autoregressive models. Following Li and Mei (2016), we develop the GLR test procedure for the testing problem (1.3) of model (1.2).

The rest of this paper is organized as follows. In Section 2, the profile quasi-maximum likelihood method is proposed to estimate the model (1.2). In Section 3, the test statistic is proposed and the residual-based bootstrap procedure is suggested to derive the $p$-value of the test. Simulations are conducted in Section 4 to examine the finite sample performance of the proposed procedures. As an application example, the Boston house price data are analyzed by the proposed methods in Section 5. Conclusion is presented in Section 6.

## 2. Profile quasi-maximum likelihood method

Let us work with the matrix notation. Denote

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{X}_1^{\mathrm{T}} \\ \mathbf{X}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{X}_n^{\mathrm{T}} \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1^{\mathrm{T}} \\ \mathbf{Z}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{Z}_n^{\mathrm{T}} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{Z}_1^{\mathrm{T}} \boldsymbol{\alpha}(U_1) \\ \mathbf{Z}_2^{\mathrm{T}} \boldsymbol{\alpha}(U_2) \\ \vdots \\ \mathbf{Z}_n^{\mathrm{T}} \boldsymbol{\alpha}(U_n) \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Then model (1.2) can be written as

$$\mathbf{Y} = \rho \mathbf{W} \mathbf{Y} + \mathbf{X} \boldsymbol{\beta} + \mathbf{M} + \boldsymbol{\varepsilon}. \tag{2.1}$$

Pretending that the error distribution is normal, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, denote $\boldsymbol{\varepsilon}(\boldsymbol{\delta}) = \mathbf{Y} - \rho \mathbf{W} \mathbf{Y} - \mathbf{X} \boldsymbol{\beta} - \mathbf{M}$ and $\mathbf{A}(\rho) = \mathbf{I}_n - \rho \mathbf{W}$, where $\boldsymbol{\delta} = (\boldsymbol{\beta}^{\mathrm{T}}, \rho)^{\mathrm{T}}$, and $\mathbf{I}_n$ is the identity matrix of order $n$. For model (2.1), the log-likelihood function is

$$\log L_n(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\alpha}(U_1), \ldots, \boldsymbol{\alpha}(U_n)) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) + \log|\mathbf{A}(\rho)| - \frac{\boldsymbol{\varepsilon}(\boldsymbol{\delta})^{\mathrm{T}} \boldsymbol{\varepsilon}(\boldsymbol{\delta})}{2\sigma^2}, \tag{2.2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \rho, \sigma^2)$.

In the following, we will apply the profile least-squares approach of Fan and Huang (2005) and Su and Ullah (2006) to estimate the parameter $\boldsymbol{\theta}$ and the unknown coefficient functions $\boldsymbol{\alpha}(\cdot)$. If the parameters $\boldsymbol{\beta}$ and $\rho$ are known, then model (1.1) can be written as

$$Y_i^* - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta} = \alpha_1(U_i) Z_{i1} + \cdots + \alpha_q(U_i) Z_{iq} + \varepsilon_i, \tag{2.3}$$

where $(Y_1^*, Y_2^*, \ldots, Y_n^*)^{\mathrm{T}} = \mathbf{A}(\rho) \mathbf{Y} = \mathbf{Y} - \rho \mathbf{W} \mathbf{Y}$. Clearly, model (2.3) is a standard varying coefficient model. Many procedures have been proposed to estimate the unknown varying coefficient functions. We will apply the local linear approach to model (2.3). Assume $\{\alpha_j(\cdot), j = 1, 2, \ldots, q\}$ have continuous second order derivatives. Then, for any given $u$ in a small neighborhood of $u_0$, one can approximate $\alpha_j(\cdot)$ locally by a linear function

$$\alpha_j(u) \approx \alpha_j(u_0) + \alpha'_j(u_0)(u - u_0), \quad j = 1, 2, \ldots, q,$$

where $\alpha'_j(u) = \partial \alpha_j(u)/\partial u$. This leads to the following weighted local least-squares problems: find $\{(\alpha_j(u_0), \alpha'_j(u_0)), j = 1, 2, \ldots, q\}$ to minimize

$$\sum_{i=1}^{n} \left[ (Y_i^* - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}) - \sum_{j=1}^{q} \{\alpha_j(u_0) + \alpha'_j(u_0)(U_i - u_0)\} Z_{ij} \right]^2 K_h(U_i - u_0) \tag{2.4}$$

where $K$ is a kernel function, $h$ is a bandwidth and $K_h(\cdot) = K(\cdot/h)/h$.

Let

$$\mathbf{D}_{u_0} = \begin{bmatrix} \mathbf{Z}_1^{\mathrm{T}} & (U_1 - u_0) \mathbf{Z}_1^{\mathrm{T}} \\ \mathbf{Z}_2^{\mathrm{T}} & (U_2 - u_0) \mathbf{Z}_2^{\mathrm{T}} \\ \vdots & \vdots \\ \mathbf{Z}_n^{\mathrm{T}} & (U_n - u_0) \mathbf{Z}_n^{\mathrm{T}} \end{bmatrix}, \mathbf{S} = \begin{bmatrix} (\mathbf{Z}_1^{\mathrm{T}}\ \mathbf{0}_{1\times q}) \{\mathbf{D}_{u_1}^{\mathrm{T}} \mathbf{K}_{u_1} \mathbf{D}_{u_1}\}^{-1} \mathbf{D}_{u_1}^{\mathrm{T}} \mathbf{K}_{u_1} \\ (\mathbf{Z}_2^{\mathrm{T}}\ \mathbf{0}_{1\times q}) \{\mathbf{D}_{u_2}^{\mathrm{T}} \mathbf{K}_{u_2} \mathbf{D}_{u_2}\}^{-1} \mathbf{D}_{u_2}^{\mathrm{T}} \mathbf{K}_{u_2} \\ \vdots \\ (\mathbf{Z}_n^{\mathrm{T}}\ \mathbf{0}_{1\times q}) \{\mathbf{D}_{u_n}^{\mathrm{T}} \mathbf{K}_{u_n} \mathbf{D}_{u_n}\}^{-1} \mathbf{D}_{u_n}^{\mathrm{T}} \mathbf{K}_{u_n} \end{bmatrix}.$$

and

$$\mathbf{K}_{u_0} = \mathrm{diag}\{K_h(U_1 - u_0), K_h(U_2 - u_0), \ldots, K_h(U_n - u_0)\}.$$

The solution to the problem (2.4) is given by

$$[\overline{\alpha}_1(u_0), \ldots, \overline{\alpha}_q(u_0), \overline{\alpha}'_1(u_0), \ldots, \overline{\alpha}'_q(u_0)]^{\mathrm{T}} = \{\mathbf{D}_{u_0}^{\mathrm{T}} \mathbf{K}_{u_0} \mathbf{D}_{u_0}\}^{-1} \mathbf{D}_{u_0}^{\mathrm{T}} \mathbf{K}_{u_0} [\mathbf{A}(\rho) \mathbf{Y} - \mathbf{X} \boldsymbol{\beta}]. \tag{2.5}$$

We take $u_0$ to be each of $U_1, U_2, \ldots, U_n$, then we can obtain the estimators of $\boldsymbol{\alpha}(U_j)$, $j = 1, 2, \ldots, n$. Therefore, we can define the estima-