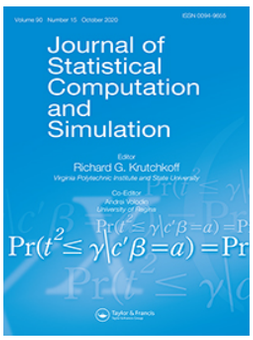

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Variable selection of partially linear varying coefficient spatial autoregressive model

Tizheng Li , Qingyan Yin & Jialong Peng

To cite this article: Tizheng Li , Qingyan Yin & Jialong Peng (2020) Variable selection of partially linear varying coefficient spatial autoregressive model, Journal of Statistical Computation and Simulation, 90:15, 2681-2704, DOI: 10.1080/00949655.2020.1788560

To link to this article: https://doi.org/10.1080/00949655.2020.1788560

Published online: 03 Jul 2020.

Submit your article to this journal

Article views: 136

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=gscs20

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2020, VOL. 90, NO. 15, 2681–2704
https://doi.org/10.1080/00949655.2020.1788560

Taylor & Francis
Taylor & Francis Group

Check for updates

Variable selection of partially linear varying coefficient spatial autoregressive model

Tizheng Li, Qingyan Yin and Jialong Peng

Department of Mathematics, School of Science, Xi'an University of Architecture and Technology, Xi'an, People's Republic of China

ABSTRACT

The partially linear varying coefficient spatial autoregressive model is a recently proposed semi-parametric spatial autoregressive model, in which some of the explanatory variables have varying coefficients while the remained explanatory variables possess constant ones. Although some estimation methods have been proposed for the partially linear varying coefficient spatial autoregressive model, the problem of selecting important explanatory variables in the parametric component of such model has not been addressed to date. In this paper, we propose a penalized profile least squares method to address this problem. Different from the existing estimation methods, the proposed method can simultaneously select the significant explanatory variables in the parametric component and estimate the corresponding nonzero regression coefficients. Furthermore, we provide a computationally feasible algorithm to obtain the penalized profile least squares estimator. The finite sample performance of the proposed variable selection method is evaluated through some simulation studies and illustrated by a real data example.

ARTICLE HISTORY
Received 6 August 2019
Accepted 24 June 2020

KEYWORDS
Spatial dependence; partially linear varying coefficient spatial autoregressive model; least squares method; profile quasi-maximum likelihood method; SCAD

MATHEMATICS SUBJECT CLASSIFICATION (2010)
91B72; 62G05

1. Introduction

Many classical statistical analyses using cross-sectional data are based on the assumption that the observations are independent in space. However, many phenomena usually exhibit spatial dependence such as real estate pricing, economic growth, price competition among firms, strategic interaction among governments, technology adoption, spread of epidemic disease, industrial organization, social interaction and so on. For example, the price of a house in one place may be influenced by those of its neighbours, and this dependence sometimes cannot be explained by its explanatory variables. In the context of strategic interaction among governments, a local government's spending on some welfare programmes might be affected by those of its neighbours. In the study of social interaction, a student's behaviour can be directly influenced by his or her friends' behaviours. The possibility of spatial dependence among observations has fuelled increased spatial modelling and data analysis activities among statistical and econometric literature.

CONTACT Tizheng Li tizhengli@xauat.edu.cn Department of Mathematics, School of Science, Xi'an University of Architecture and Technology, Xi'an 710055, People's Republic of China

© 2020 Informa UK Limited, trading as Taylor & Francis Group

Spatial autoregressive models, which extend the traditional regression models by taking into account a spatially lagged term of the response variable, have been one of the most important statistical tools for modelling spatial dependence of spatial data. Among the spatial autoregressive models, the linear spatial autoregressive model has received much attention from theoretical and applied researchers because it is powerful in explanation and easy to be fitted. The linear spatial autoregressive model has been extensively studied in the estimation [1–4], test [5–7] and variable selection [8–10], and has been successfully applied to a variety of fields, especially the regional science and economics, for spatial data analysis.

Although the linear spatial autoregressive model prevails in practice, there is an increasing trend for many researchers to recognize and thus start addressing the importance of nonlinearity in modelling spatial dependence. For example, in modelling hedonic agricultural land prices, Kostov [11] noticed that there is a great deal of uncertainty concerning the functional form of the hedonic price function and the hedonic price function is generally nonlinear. Basile and Gress [12] proposed a semi-parametric spatial auto-covariance specification of the growth model for the European economy, where the regression part is non-parametrically specified except for a fixed spatial autoregressive parameter. They found that the estimate of the spatial autoregressive parameter in their model is considerably smaller than that in the parametric specifications, suggesting that incorrect functional form might 'increase' spatial dependence. In other words, even if there is no spatial correlation in the 'true' model, the model specification with incorrect functional form can 'create' it. Other researchers [13–15] also considered some flexible functional forms to account for certain forms of nonlinearities in their studies. Most of these studies introduce a parametric transformation (for example, the Box–Cox transformation) on the response variable or/and the explanatory variables. Nevertheless, parametric functional form transformation can at most provide certain protection against some specific nonlinear forms. In the absence of a priori information and theoretical foundation, it is generally advisable to consider more flexible functional forms.

Motivated by the idea of non-parametric modelling, Su [16] proposed non-parametric spatial autoregressive model where the spatially lagged term of the response variable enters the model linearly while all the explanatory variables enter the model in a non-parametric way, and developed a semi-parametric generalized method of the moment to estimate the spatial autoregressive parameter and the regression function. The non-parametric spatial autoregressive model makes no assumption on the form of the regression function and lets the data determine a functional form tailored to the data, hence it carries no risk of model misspecification which often occurs in traditional parametric spatial autoregressive models. Furthermore, it can provide useful insights for further parametric modelling. However, the non-parametric spatial autoregressive model may fail to incorporate some important prior information and the convergence rate of the resulting estimator of the regression function deteriorates as the number of the explanatory variables increases, which is referred to as the 'curse of dimensionality' in the literature of non-parametric regression. Moreover, even if a precise non-parametric estimator of the regression function can be obtained, it becomes very difficult to present and interpret the non-parametric estimator in empirical applications when the number of the explanatory variables is larger than two.

One strategy to attenuate the 'curse of dimensionality' is the partially linear spatial autoregressive model proposed by Su and Jin [17], in which the spatially lagged term of the

response variable and some of the explanatory variables enter the model linearly, whereas the remained explanatory variables are incorporated in the model non-parametrically. In contrast with parametric spatial autoregressive models, the partially linear spatial autoregressive model has a non-parametric component, thus effectively lessening the possibility of either inconsistent estimate or misleading inference due to misspecification. Meanwhile, the parametric component of the partially linear spatial autoregressive model can in general be estimated with a convergence rate comparable to what we would obtain by using correctly specified parametric spatial autoregressive models. However, the partially linear spatial autoregressive model is still subject to the 'curse of dimensionality' when the number of the explanatory variables incorporated in its non-parametric component is large.

Another strategy is the varying coefficient spatial autoregressive model proposed by Li and Chen [18], in which the regression coefficients in classical linear spatial autoregressive model are replaced by unknown functions of certain variable which is usually called as smoothing variable in the literature of non-parametric regression. However, in reality, some of the regression coefficients may be constant rather than varying. In such situation, mistakenly treating constant coefficients as varying ones will degrade the efficiency of estimation and inference. In order to improve estimation and inference efficiency, one should consider the partially linear varying coefficient spatial autoregressive model in which some of the explanatory variables have varying coefficients while the remained explanatory variables possess constant ones, which can be regarded as a special case of the varying coefficient spatial autoregressive model. The sample form of the partially linear varying coefficient spatial autoregressive model is as follows:

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

where Y_i $(i = 1, \ldots, n)$ are observations of response variable Y, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$, $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iq})^{\mathrm{T}}$ and U_i $(i = 1, \ldots, n)$ are, respectively, observations of explanatory variables $X_1, \ldots, X_p$, $Z_1, \ldots, Z_q$ and smoothing variable U, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of regression coefficients, $\mathbf{a}(\cdot) = (a_1(\cdot), \ldots, a_q(\cdot))^{\mathrm{T}}$ is the vector of coefficient functions, w_{ij} $(i, j = 1, \ldots, n; i \neq j)$ are pre-specified exogenous spatial weights that determine the structure of neighbourhood among spatial units, ρ is the spatial autoregressive parameter that measures the intensity of spatial dependence among the observations of the response variable, and ε_i $(i = 1, \ldots, n)$ are the independent and identically distributed error terms with mean 0 and finite variance σ^2.

The specification of the spatial weight matrix $\mathbf{W} = (w_{ij})_{i,j=1}^{n}$ is a fundamental issue in using model (1). The spatial weight matrix is usually specified based on contiguity relationship or geographic distance among spatial units. For example, the element w_{ij} of $\mathbf{W}$ is non-zero when spatial units i and j share a common boundary or are within a given distance of each other. However, in some practical applications especially in the field of economics, it is better to construct the spatial weight matrix by using economic distance, social distance or socioeconomic distance. The diagonal elements of the spatial weight matrix are usually set to be zero in order to prevent it from predicting itself. When specifying the spatial weight matrix $\mathbf{W}$, it is a common practice to have $\mathbf{W}$ being row-normalized such that

its ith row is

$$\mathbf{w}_i = (w_{i1}, \ldots, w_{in}) = (d_{i1}, \ldots, d_{in}) / \sum_{k=1}^{n} d_{ik},$$

where d_{ij} is a positive number which reflects the contiguity relationship or certain distance between spatial units i and j. In this case, the weighting operation $\sum_{j \neq i} w_{ij} Y_j$ can be treated as a weighted average of the neighbouring values of Y_i and the parameter space of the spatial autoregressive parameter ρ is the interval $(-1, 1)$.

Wei et al. [19] and Sun et al. [20], respectively, proposed the profile quasi-maximum likelihood estimation method and the semi-parametric series-based least squares estimation method for model (1). Furthermore, Wei et al. [19] developed a generalized likelihood ratio test statistic to test the existence of the spatial effects in model (1).

In recent years, variable selection or model selection has become an important and fundamental issue in data analysis because high-dimensional data are commonly encountered in various applied fields such as epidemiology, genetics, economics and finance. It is well known that the traditional variable selection procedures such as Akaika information criterion (AIC), Bayesian information criterion (BIC), risk inflation criterion (RIC) and stepwise regression can be extremely computationally intensive in the analysis of the high-dimensional data. To address this computational challenge, various penalized likelihood or least-squares methods have been well developed and become a promising alternative. With appropriately selected penalty function, the penalized method would automatically shrink the small regression coefficients to zero and remove the associated explanatory variables from the current model, hence serve as the purpose of variable selection. Recently, the penalized method has been extended to many important non-parametric and semi-parametric regression models such as the additive model [21], the varying coefficient model [22], the partially linear varying coefficient model [23], the partially linear model [24], the single-index model [25], the partially linear single-index model [26] and the partially linear additive model [27]. Furthermore, the penalized method has also been extended to spatial regression models. For example, Wang and Zhu [28] developed a penalized least squares method with various penalty functions for a spatial linear model in which the error process is assumed to be a second-order stationary random field. Huang et al. [29] proposed a spatial LASSO (least absolute shrinkage and selection operator) method to simultaneously select explanatory variables and neighbourhood structures in a spatial linear regression model with GIS (geographic information systems) layers. Zhu et al. [30] proposed an adaptive LASSO method to simultaneously select explanatory variables and spatial neighbourhood structures in spatial linear model with Gaussian process errors. Chu et al. [31] developed a penalized maximum likelihood method to simultaneously select explanatory variables and estimate the corresponding non-zero parameters in spatial linear model with Gaussian process errors. More recently, Nandy et al. [32] proposed a spatially weighted l_2 error norm with a group LASSO type penalty function to select the non-zero components in spatial additive regression model. Wang [33] developed a partially adaptive group L_r ($r \geq 1$) penalized M-type method to simultaneously choose the explanatory variables with non-zero varying and constant coefficients in spatial semi-parametric varying coefficient regression model.

In contrast, the research on the variable selection for spatial autoregressive models are relatively rare and are mainly limited to parametric spatial autoregressive models. For example, Wu and Sun [8] and Liu et al. [9], respectively, developed penalized least squares method and penalized quasi-maximum likelihood method for variable selection of the linear spatial autoregressive model. The studies mentioned above only focus on the case that the number of the explanatory variables is fixed. Xie et al. [10] studied the problem of variable selection in linear spatial autoregressive model with a diverging number of parameters, constructed a penalized estimator based on the instrumental variable and SCAD (smoothly clipped absolute deviation) penalty function, and derived the asymptotic properties of the resulting penalized estimator. However, to the best of our knowledge, there has been little work on variable selection of the partially linear varying coefficient spatial autoregressive model, which greatly limits the scope of application of partially linear varying coefficient spatial autoregressive model.

In this paper, we propose a penalized profile least squares method, which is the combination of the penalized least squares method and the profile quasi-maximum likelihood method proposed by Wei et al. [19], to select the important explanatory variables in the parametric component of model (1). Unlike the existing estimation methods, the proposed method can simultaneously select the significant explanatory variables in parametric component and estimate the corresponding non-zero regression coefficients. Furthermore, we provide a computationally feasible algorithm to obtain the penalized profile least squares estimator. Extensive simulation studies are conducted to assess the finite sample performance of the proposed variable selection method and the simulation results show that the proposed variable selection method performs well in finite samples. As an illustration, we apply the proposed variable selection method to analyse the well-known Boston housing price dataset.

The remainder of this paper is organized as follows. In Section 2, we propose a class of variable selection method for the partially linear varying coefficient spatial autoregressive model, in which the issue of selecting the penalty function is also discussed. In Section 3, we discuss in detail some issues related to the practical implementation of the proposed variable selection method. Extensive simulation studies are conducted in Section 4 to assess the finite sample performance of the proposed variable selection method. In Section 5, we apply the proposed variable selection method to analyse the Boston housing price dataset. The paper is then concluded with some remarks.

2. The methodology

Throughout this paper, let $\mathbf{I}_n$ be an identity matrix of size n, $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^{\mathrm{T}}$, $\mathbf{M} = (\mathbf{Z}_1^{\mathrm{T}}\mathbf{a}(U_1), \ldots, \mathbf{Z}_n^{\mathrm{T}}\mathbf{a}(U_n))^{\mathrm{T}}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$, and $\mathbf{T}(\rho) = \mathbf{I}_n - \rho\mathbf{W}$ for any value of ρ. Then the model (1) can be expressed in form of matrix as

$$\mathbf{Y} = \rho\mathbf{W}\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{M} + \boldsymbol{\varepsilon}. \tag{2}$$

If ρ and $\mathbf{a}(\cdot)$ were both known, we would obtain the least squares estimator of $\boldsymbol{\beta}$ by minimizing

$$\frac{1}{2}\|\mathbf{T}(\rho)\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{M}\|^2 \tag{3}$$

with respect to $\boldsymbol{\beta}$, where $\|\cdot\|$ denotes the Euclidean norm. In order to select the significant explanatory variables in the parametric component of model (1), we adopt the penalized least squares method. Let $p_{\lambda_j}(\cdot)$ be a pre-specified penalty function with a tuning parameter λ_j, then the penalized least squares function can be defined as

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{T}(\rho)\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{M}\|^2 + n\sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|). \tag{4}$$

Minimizing $Q(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ yields the penalized least squares estimator of $\boldsymbol{\beta}$. Throughout this paper, we allow different components of the regression coefficient vector $\boldsymbol{\beta}$ to have different tuning parameters. This is reasonable because the tuning parameters for the regression coefficients whose values are close to zero should be much larger than those for the regression coefficients whose values are far away from zero. Thus, we can simultaneously unbiasedly estimate the regression coefficients whose values are far away from zero, and quickly shrink the regression coefficients whose values are close to zero toward zero by setting different tuning parameters for different regression coefficients.

Before we pursue this issue further, let us briefly discuss the issue of selecting the penalty function. Various penalty functions have been proposed and widely used in the literature on variable selection for linear regression model. Let us take the penalty function to be the L_0 penalty function defined by $p_{\lambda_j}(|\beta_j|) = 0.5\lambda_j{}^2 I(|\beta_j|\neq 0)$, where $I(\cdot)$ denotes the indicator function. Since $\sum_{j=1}^{p} I(|\beta_j|\neq 0)$ equals the number of non-zero regression coefficients in the linear regression model, many popular variable selection criteria such as AIC, BIC and RIC can be derived from a penalized least squares problem with the L_0 penalty function by specifying the value of λ_j to be $\sigma\sqrt{2/n}$, $\sigma\sqrt{\log(n)/n}$ and $\sigma\sqrt{\log(p)/n}$, respectively, even though these criteria were motivated by different principles. However, Fan and Li [34] pointed out several drawbacks of using the L_0 penalty function such as the expensive computational cost and the high instability. To avoid these drawbacks, Tibshirani [35] proposed the LASSO technique, which can be viewed as the solution of the penalized least squares problem with the L_1 penalty function defined as $p_{\lambda_j}(|\beta_j|) = \lambda_j|\beta_j|$. Frank and Friedman [36] considered the L_q $(0 < q < 1)$ penalty function defined by $p_{\lambda_j}(|\beta_j|) = \lambda_j|\beta_j|^q$, which yields the Bridge regression technique. Fan and Li [34] systematically studied the issue of choosing the penalty function and suggested using the SCAD penalty function because it simultaneously satisfies the mathematical conditions for unbiasedness, sparsity and continuity, the details can be found in Fan and Li [34]. Therefore, we only focus on the SCAD penalty function throughout this paper. The SCAD penalty function is of the following form:

$$p_{\lambda_j}(|\beta_j|) = \begin{cases} \lambda_j|\beta_j|, & |\beta_j|\leq\lambda_j, \\ -\dfrac{\beta_j^2 - 2a\lambda_j|\beta_j| + \lambda_j^2}{2(a-1)}, & \lambda_j < |\beta_j| < a\lambda_j, \\ \dfrac{(a+1)\lambda_j^2}{2}, & |\beta_j|\geq a\lambda_j, \end{cases}$$

where a and λ_j are tuning parameters. For the first tuning parameter a, Fan and Li [34] suggested that $a = 3.7$ is a reasonable choice. The SCAD penalty function is continuously differentiable on $(-\infty, 0)\cup(0, +\infty)$ but singular at 0. Its derivative vanishes outside $[-a\lambda_j, a\lambda_j]$. As a consequence, the penalized likelihood or least squares method with SCAD penalty function can produce sparse solution and unbiased estimates of the large parameters. More details on SCAD penalty function can be found in Fan and Li [34].

However, (4) is not yet ready for optimization because ρ and $\mathbf{a}(\cdot)$ are both unknown. To address this issue, we first use the profile quasi-maximum likelihood method proposed by Wei et al. [19] to estimate ρ and $\mathbf{a}(\cdot)$, then substitute the resulting estimators into (4) and finally minimize (4) with respect to $\boldsymbol{\beta}$. The profile quasi-maximum likelihood method proposed by Wei et al. [19] can be described as follows.

For model (2), it follows from Wei et al. [19] that the Gaussian quasi log-likelihood function is

$$\log L(\mathbf{a}(\cdot), \boldsymbol{\beta}, \rho, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) + \log(|\mathbf{T}(\rho)|) - \frac{1}{2\sigma^2} \times [\overline{\mathbf{Y}}(\rho) - \mathbf{X}\boldsymbol{\beta} - \mathbf{M}]^{\mathrm{T}}[\overline{\mathbf{Y}}(\rho) - \mathbf{X}\boldsymbol{\beta} - \mathbf{M}], \tag{5}$$

where $\overline{\mathbf{Y}}(\rho) = \mathbf{T}(\rho)\mathbf{Y}$. Since the unknown coefficient function vector $\mathbf{a}(\cdot)$ is present in Equation (5), Wei et al. [19] proposed the following two-step procedure to obtain the estimator of the parameter vector $(\boldsymbol{\beta}^{\mathrm{T}}, \rho, \sigma^2)^{\mathrm{T}}$.

In the first step, fix $\boldsymbol{\beta}$ and ρ and re-write model (2) as

$$\overline{\mathbf{Y}}(\rho) - \mathbf{X}\boldsymbol{\beta} = \mathbf{M} + \boldsymbol{\varepsilon}. \tag{6}$$

The local linear smoothing method can be used to estimate the coefficient function vector $\mathbf{a}(\cdot)$ in model (6).

Assume that all the coefficient functions $a_1(\cdot), \ldots, a_q(\cdot)$ have continuous second-order derivatives. Then for any given u_0 in the domain of the smoothing variable U, it follows from Taylor's expansion that

$$a_j(u) \approx a_j(u_0) + a_j'(u_0)(u - u_0), \quad j = 1, \ldots, q$$

in a neighbourhood of u_0. The local linear smoothing method finds $\mathbf{a}(u_0)$ and $\mathbf{a}'(u_0) = (a_1'(u_0), \ldots, a_q'(u_0))^{\mathrm{T}}$ to minimize the following locally weighted least squares function

$$\sum_{i=1}^{n}\left[\overline{Y}_i(\rho) - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{Z}_i^{\mathrm{T}}\mathbf{a}(u_0) - (U_i - u_0)\mathbf{Z}_i^{\mathrm{T}}\mathbf{a}'(u_0)\right]^2 K_h(U_i - u_0), \tag{7}$$

where $\overline{Y}_i(\rho)$ is the ith element of $\overline{\mathbf{Y}}(\rho)$ and $K_h(\cdot) = K(\cdot/h)/h$ with $K(\cdot)$ being a kernel function and h being the bandwidth.

Let $\boldsymbol{\Psi}(u_0) = (\mathbf{a}(u_0)^{\mathrm{T}}, \mathbf{a}'(u_0)^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{W}(u_0) = \mathrm{diag}(K_h(U_1 - u_0), \ldots, K_h(U_n - u_0))$, and

$$\mathbf{Z}(u_0) = \begin{pmatrix} \mathbf{Z}_1^{\mathrm{T}} & (U_1 - u_0)\mathbf{Z}_1^{\mathrm{T}} \\ \vdots & \vdots \\ \mathbf{Z}_n^{\mathrm{T}} & (U_n - u_0)\mathbf{Z}_n^{\mathrm{T}} \end{pmatrix}.$$

With the above notations, the solution of the weighted least squares problem (7), that is, the estimator of $\boldsymbol{\Psi}(u_0)$, can be expressed as

$$\widehat{\boldsymbol{\Psi}}(u_0;\boldsymbol{\beta},\rho) = [\mathbf{Z}(u_0)^{\mathrm{T}}\mathbf{W}(u_0)\mathbf{Z}(u_0)]^{-1}\mathbf{Z}(u_0)^{\mathrm{T}}\mathbf{W}(u_0)[\,\overline{\mathbf{Y}}(\rho) - \mathbf{X}\boldsymbol{\beta}].$$

Consequently, the estimator of the coefficient function vector $\mathbf{a}(u)$ at u_0 is

$$\widehat{\mathbf{a}}(u_0;\boldsymbol{\beta},\rho) = (\mathbf{I}_q, \mathbf{0}_{q\times q})[\mathbf{Z}(u_0)^{\mathrm{T}}\mathbf{W}(u_0)\mathbf{Z}(u_0)]^{-1}\mathbf{Z}(u_0)^{\mathrm{T}}\mathbf{W}(u_0)[\,\overline{\mathbf{Y}}(\rho) - \mathbf{X}\boldsymbol{\beta}], \tag{8}$$

where $\mathbf{0}_{q\times q}$ is a $q\times q$ zero matrix.

Taking u_0 in (8) to be $U_1,\ldots,U_n$, respectively, we can obtain $\widehat{\mathbf{M}}(\boldsymbol{\beta},\rho)$, the fitted vector of $\mathbf{M}$ given $\boldsymbol{\beta}$ and ρ, as

$$\widehat{\mathbf{M}}(\boldsymbol{\beta},\rho) = \begin{pmatrix} \mathbf{Z}_1^{\mathrm{T}}\widehat{\mathbf{a}}(U_1;\boldsymbol{\beta},\rho) \\ \vdots \\ \mathbf{Z}_n^{\mathrm{T}}\widehat{\mathbf{a}}(U_n;\boldsymbol{\beta},\rho) \end{pmatrix} = \mathbf{S}[\,\overline{\mathbf{Y}}(\rho) - \mathbf{X}\boldsymbol{\beta}], \tag{9}$$

where

$$\mathbf{S} = \begin{pmatrix} (\mathbf{Z}_1^{\mathrm{T}}, \mathbf{0}_{1\times q})\left[\mathbf{Z}(U_1)^{\mathrm{T}}\mathbf{W}(U_1)\mathbf{Z}(U_1)\right]^{-1}\mathbf{Z}(U_1)^{\mathrm{T}}\mathbf{W}(U_1) \\ \vdots \\ (\mathbf{Z}_n^{\mathrm{T}}, \mathbf{0}_{1\times q})\left[\mathbf{Z}(U_n)^{\mathrm{T}}\mathbf{W}(U_n)\mathbf{Z}(U_n)\right]^{-1}\mathbf{Z}(U_n)^{\mathrm{T}}\mathbf{W}(U_n) \end{pmatrix}.$$

In the second step, replacing $\mathbf{M}$ by $\widehat{\mathbf{M}}(\boldsymbol{\beta},\rho)$ in (5) yields profile quasi log-likelihood function

$$\begin{aligned}\log L(\boldsymbol{\beta},\rho,\sigma^2) = &-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) + \log(|\mathbf{T}(\rho)|) - \frac{1}{2\sigma^2} \\ &\times [\,\overline{\mathbf{Y}}(\rho) - \mathbf{X}\boldsymbol{\beta} - \widehat{\mathbf{M}}(\boldsymbol{\beta},\rho)]^{\mathrm{T}}[\,\overline{\mathbf{Y}}(\rho) - \mathbf{X}\boldsymbol{\beta} - \widehat{\mathbf{M}}(\boldsymbol{\beta},\rho)].\end{aligned}$$

Given ρ, maximizing $\log L(\boldsymbol{\beta},\rho,\sigma^2)$ with respect to $\boldsymbol{\beta}$ and σ^2 gives the profile quasi-maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2, respectively, as

$$\widehat{\boldsymbol{\beta}}(\rho) = \left[\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})\mathbf{X}\right]^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})\mathbf{T}(\rho)\mathbf{Y} \tag{10}$$

and

$$\begin{aligned}\widehat{\sigma}^2(\rho) &= \frac{1}{n}\left[(\mathbf{I}_n - \mathbf{S})\left(\mathbf{T}(\rho)\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\rho)\right)\right]^{\mathrm{T}}\left[(\mathbf{I}_n - \mathbf{S})\left(\mathbf{T}(\rho)\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\rho)\right)\right] \\ &= \frac{1}{n}\mathbf{Y}^{\mathrm{T}}(\mathbf{T}(\rho))^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})^{\mathrm{T}}\mathbf{H}(\mathbf{I}_n - \mathbf{S})\mathbf{T}(\rho)\mathbf{Y},\end{aligned} \tag{11}$$

where $\mathbf{H} = \mathbf{I}_n - (\mathbf{I}_n - \mathbf{S})\mathbf{X}[\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})^{\mathrm{T}}$. Substituting $\widehat{\boldsymbol{\beta}}(\rho)$ and $\widehat{\sigma}^2(\rho)$ into $\log L(\boldsymbol{\beta},\rho,\sigma^2)$ yields the profile quasi log-likelihood function of ρ as

$$\log L(\rho) = -\frac{n}{2}[\log(2\pi) + 1] - \frac{n}{2}\log\left(\widehat{\sigma}^2(\rho)\right) + \log(|\mathbf{T}(\rho)|). \tag{12}$$

Maximizing $\log L(\rho)$ yields the profile quasi-maximum likelihood estimator $\widehat{\rho}$ of ρ. Then, substituting $\widehat{\rho}$ into $\widehat{\boldsymbol{\beta}}(\rho)$ and $\widehat{\sigma}^2(\rho)$ gives the final profile quasi-maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2 as $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}(\widehat{\rho})$ and $\widehat{\sigma}^2 \equiv \widehat{\sigma}^2(\widehat{\rho})$, respectively, and finally the local linear estimator of $\mathbf{a}(u_0)$ as $\widehat{\mathbf{a}}(u_0) \equiv \widehat{\mathbf{a}}(u_0;\widehat{\boldsymbol{\beta}},\widehat{\rho})$.

Replacing ρ and $\mathbf{M}$ in (4) by $\widehat{\rho}$ and $\widehat{\mathbf{M}}(\boldsymbol{\beta}, \widehat{\rho})$, we obtain the penalized profile least squares function

$$Q_{\mathrm{P}}(\boldsymbol{\beta}) = \frac{1}{2}\|(\mathbf{I}_n - \mathbf{S})\mathbf{T}(\widehat{\rho})\mathbf{Y} - (\mathbf{I}_n - \mathbf{S})\mathbf{X}\boldsymbol{\beta}\|^2 + n\sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|). \tag{13}$$

Thus, we can obtain the penalized profile least squares estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{P}}$ of $\boldsymbol{\beta}$ by minimizing $Q_{\mathrm{P}}(\boldsymbol{\beta})$.

3. Practical implementation of the proposed variable selection method

In this section, we discuss some issues related to the practical implementation of the proposed variable selection method such as how to obtain the proposed penalized profile least squares estimator and the selection of the bandwidth h and the tuning parameter vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$.

3.1. Local quadratic approximation of penalty function

Since many penalty functions, including the commonly used L_1 and SCAD penalty functions, are irregular at the origin and may not have a second derivative at some points, direct implementation of the Newton–Raphson algorithm may be very difficult. Following Fan and Li [34], we locally approximate the penalty function $p_{\lambda_j}(|\beta_j|)$ by a quadratic function at every step of the iteration. Specifically, given an initial value $\boldsymbol{\beta}^{(0)}$ that is close to the minimizer of $Q_{\mathrm{P}}(\boldsymbol{\beta})$, if $|\beta_j^{(0)}|$ is very close to 0, then set $\widehat{\beta}_{j\mathrm{P}} = 0$. Otherwise, the first derivative of the penalty function is locally approximated by

$$[p_{\lambda_j}(|\beta_j|)]' = p'_{\lambda_j}(|\beta_j|)\frac{\beta_j}{|\beta_j|} \approx p'_{\lambda_j}(|\beta_j^{(0)}|)\frac{\beta_j}{|\beta_j^{(0)}|}.$$

In other words, for $\beta_j \approx \beta_j^{(0)}$, the penalty function can be locally approximated by a quadratic function as

$$p_{\lambda_j}(|\beta_j|) \approx p_{\lambda_j}(|\beta_j^{(0)}|) + \frac{1}{2}\left[p'_{\lambda_j}(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\right]\left(\beta_j^2 - \beta_j^{(0)2}\right).$$

With the aid of the local quadratic approximation, the solution of the penalized profile least squares problem (13) can be found by iteratively computing

$$\boldsymbol{\beta}^{(1)} = [\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})\mathbf{X} + n\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(0)})]^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})\mathbf{T}(\widehat{\rho})\mathbf{Y}, \tag{14}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(0)}) = \mathrm{diag}(p'_{\lambda_1}(|\beta_1^{(0)}|)/|\beta_1^{(0)}|, \ldots, p'_{\lambda_p}(|\beta_p^{(0)}|)/|\beta_p^{(0)}|)$.

Hence, we have the following iterative algorithm:

(Step 1) Given an initial value $\boldsymbol{\beta}^{(0)}$ that is close to the minimizer of $Q_{\mathrm{P}}(\boldsymbol{\beta})$, if $|\beta_j^{(0)}| < \tau$, then set $\widehat{\beta}_{j\mathrm{P}} = 0$.

(Step 2) Set $\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(0)}$ and obtain $\boldsymbol{\beta}^{(k+1)}$ by (14).
(Step 3) Iterate Step 2 until convergence and denote the final estimator of $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}_{\mathrm{P}}$.

In the practical implementation of the above iterative algorithm, we take the profile quasi-maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ as the initial value $\boldsymbol{\beta}^{(0)}$ of $\boldsymbol{\beta}$ and the value of τ to be 10^{-3}.

3.2. Selection of bandwidth and tuning parameters

Bandwidth selection. Since it is difficult to choose the optimal value of the bandwidth h, we select, like that in Su and Jin [17] and Su [16], the optimal value of h via the following two-step procedure. (i) Choose an initial value of h by using the simple rule of thumb (ROT) method, that is, $h = s_u n^{-1/5}$, where s_u is the standard deviation of the observations $U_1, \ldots, U_n$ of the smoothing variable U, to obtain a preliminary estimator $(\overline{\boldsymbol{\beta}}^{\mathrm{T}}, \overline{\rho}\,)^{\mathrm{T}}$ of $(\boldsymbol{\beta}^{\mathrm{T}}, \rho)^{\mathrm{T}}$ in model (2). (ii) Use the leave-one-out least squares cross-validation (CV) method to choose the optimal value of h by regressing $\mathbf{Y} - \overline{\rho}\,\mathbf{W}\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}$ on $\{U_i, \mathbf{Z}_i\}_{i=1}^n$ with the local linear smoothing method. Specifically, let $\overline{Y}_i$ denote the ith element of $\mathbf{Y} - \overline{\rho}\,\mathbf{W}\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}$ and $\overline{\mathbf{a}}_{-i,h}(U_i)$ be the leave-one-out local linear estimator of $\mathbf{a}(U_i)$ by leaving the observation $(U_i, \mathbf{Z}_i, \overline{Y}_i)$ out in the profile quasi-maximum likelihood estimation procedure and by using the bandwidth h, then the optimal value of the bandwidth h is chosen to minimize

$$\mathrm{CV}(h) = \sum_{i=1}^{n} \Big[\overline{Y}_i - \mathbf{Z}_i^{\mathrm{T}} \overline{\mathbf{a}}_{-i,h}(U_i) \Big]^2. \tag{15}$$

Selection of tuning parameters. It is well known that the tuning parameter plays a very important role in penalized least squares or likelihood method. In the past decades, it has been well understood that the generalized cross-validation (GCV) selection method has the asymptotic behaviour similar to that of AIC, and has been widely used. However, Wang et al. [37] pointed out that the tuning parameter selected by GCV might not be able to consistently identify the true model, and they also verified that the SCAD penalized method with the tuning parameter selected by a BIC-type criterion can identify the true model consistently. Following the advocation of Wang et al. [37], we use the BIC selector to choose the optimal value of the tuning parameter λ. Specifically, the BIC statistic is defined as

$$\mathrm{BIC}(\boldsymbol{\lambda}) = \log\{\mathrm{RSS}(\boldsymbol{\lambda})/n\} + \{\log(n)/n\} e(\boldsymbol{\lambda}), \tag{16}$$

where $e(\boldsymbol{\lambda}) = \mathrm{tr}\{(\mathbf{I}_n - \mathbf{S})\mathbf{X}[\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})\mathbf{X} + \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})]^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{S})^{\mathrm{T}}\}$ is the effective number of parameters of the penalized profile least squares estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ of $\boldsymbol{\beta}$ with tuning parameter vector $\boldsymbol{\lambda}$ and $\mathrm{RSS}(\boldsymbol{\lambda}) = \|(\mathbf{I}_n - \mathbf{S})\mathbf{T}(\widehat{\rho})\mathbf{Y} - (\mathbf{I}_n - \mathbf{S})\mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|^2$ is the residual sum of squares. Obviously, the optimization problem over a p-dimensional space is quite difficult especially when the number of the explanatory variables $X_1, \ldots, X_p$ is large. However, it is expected that the choice of λ_j's should satisfy that the tuning parameters for the regression coefficients whose values are close to zero should be much larger than those for the regression coefficients whose values are far away from zero. Hence, in practice, we suggest taking $\lambda_j = \lambda/\widehat{\beta}_j$ $(j = 1, \ldots, p)$, where $\widehat{\beta}_j$ is the jth element of the profile

quasi-maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Then, (16) reduces to an easily solved one-dimensional minimization problem. In fact, such a choice of the tuning parameters, in some sense, is the same rationale behind the adaptive LASSO proposed by Zou [38], and works quite well from our simulation experience.

4. Simulation studies

In this section, we conduct some simulation studies to evaluate the finite sample performance of the proposed variable selection method. For ease of presentation, we call the proposed variable selection method with the SCAD penalty function as the SCAD method throughout this section.

4.1. Spatial layout and design of experiment

The spatial layout used in our simulation studies is taken as a square region with the length of each side being l units. This type of spatial layout is of wide application backgrounds in the field of remote sensing. The $l\times l$ lattice squares in the region, which leads to a sample size of $n = l^2$, are designed as the spatial units at which the observations of the response variable and the explanatory variables are made. These n spatial units are labelled by 1 to n with the order from left to right and from bottom to top.

Given the above spatial layout, the spatial weight matrix $\mathbf{W}$ are constructed based on the Rook contiguity and the exponential decay function of the distance between spatial units, respectively. For the Rook contiguity, the standardized spatial weight matrix $\mathbf{W}$ is generated as follows: (i) Let $w_{ij} = 1$ if spatial unit j shares a common edge with spatial unit i and let $w_{ij} = 0$ otherwise; (ii) divide each element w_{ij} by the corresponding row sum to form the standardized spatial weight matrix $\mathbf{W}$. For the latter way, the element w_{ij} of the spatial weight matrix $\mathbf{W}$ is taken as $w_{ij} = \exp(-d_{ij})/\sum_{k=1}^{n}\exp(-d_{ik})$, where d_{ij} is the Euclidean distance between spatial units i and j.

We simulate $N = 500$ data sets from the following partially linear varying coefficient spatial autoregressive model

$$Y_i = \rho\sum_{j\neq i} w_{ij}Y_j + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + Z_{i1}a_1(U_i) + Z_{i2}a_2(U_i) + \varepsilon_i, \quad i = 1,\ldots,n, \tag{17}$$

where w_{ij} is the (i,j)th element of $\mathbf{W}$, $U \sim U(0,1)$, $\mathbf{Z} = (Z_1, Z_2)^{\mathrm{T}}$ with $Z_1{\equiv}1$ and $Z_2{\sim}N(0,1)$, $\mathbf{X} = (X_1,\ldots,X_p)^{\mathrm{T}}$ follows a normal distribution $N(\mathbf{0},\boldsymbol{\Sigma})$ in which the diagonal elements of $\boldsymbol{\Sigma}$ are 1 and the other elements of $\boldsymbol{\Sigma}$ are γ, and $\varepsilon{\sim}N(0,1)$. As for the coefficient functions $a_1(u)$ and $a_2(u)$, we consider the following two groups of functions:

G1: $a_1(u) = 2u - 1, a_2(u) = (1-2u)^3$,
G2: $a_1(u) = \sin(2\pi u), a_2(u) = 3.5[\exp(-(4u-1)^2) + \exp(-(4u-3)^2)] - 1.5$.

Obviously, the variation pattern of the second group of coefficient functions is more complex than that of the first group of coefficient functions. Therefore, we select the above two groups of the coefficient functions to empirically assess the effect of the complexity of the coefficient functions on the performance of the proposed variable selection method.

To assess the selection performance of the proposed variable selection method, we use the average number of the true zero regression coefficients that are correctly identified as zero, the average number of the true non-zero regression coefficients that are incorrectly identified as zero, the proportion of excluding any relevant explanatory variables, the proportion of correctly selecting all relevant explanatory variables and the proportion of including all relevant explanatory variables and some irrelevant explanatory variables, which are abbreviated as 'C', 'IC', 'U-f', 'C-f' and 'O-f', respectively. To measure the accuracy of an estimator of $\boldsymbol{\beta}$, we employ the commonly used mean square error (MSE) criterion. Specifically, suppose that $\widetilde{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$, then the MSE of $\widetilde{\boldsymbol{\beta}}$ is defined as

$$\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}) = \frac{1}{N}\sum_{l=1}^{N} \|\widetilde{\boldsymbol{\beta}}_{(l)} - \boldsymbol{\beta}\|^2,$$

where $\widetilde{\boldsymbol{\beta}}_{(l)}$ is the lth estimator of $\boldsymbol{\beta}$ obtained based on the lth simulated data set. Furthermore, according to the comments of the associate editor and the referee, it is helpful to compare the estimation performance of the proposed variable selection method with that of the existing profile quasi-maximum likelihood method. To this goal, we employ the following relative mean square error (RMSE) index which is defined by

$$\mathrm{RMSE} = \frac{\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_{\mathrm{P}})}{\mathrm{MSE}(\widehat{\boldsymbol{\beta}})},$$

where $\widehat{\boldsymbol{\beta}}$ is either the unpenalized estimator of $\boldsymbol{\beta}$ obtained by using the profile quasi-maximum likelihood method based on the full model or the oracle estimator of $\boldsymbol{\beta}$ obtained by using the profile quasi-maximum likelihood method based on the true model. Obviously, when the value of the index RMSE is smaller than 1, it means that the estimator in numerator is more accurate than that in denominator. Thus, the value of the index RMSE measures the estimation accuracy of the proposed penalized estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{P}}$ relative to that of $\widehat{\boldsymbol{\beta}}$. In what follows, we use the symbols $\mathrm{RMSE}_{\mathrm{o}}$ and $\mathrm{RMSE}_{\mathrm{u}}$ to denote, respectively, the RMSE of the proposed penalized estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{P}}$ of $\boldsymbol{\beta}$ relative to the oracle estimator and the unpenalized estimator of $\boldsymbol{\beta}$.

In order to evaluate the influence of the tuning parameters λ_j's on the finite sample performance of the proposed variable selection method, we also select the values of the tuning parameters λ_j's by using the AIC criterion in addition to the BIC criterion proposed in Section 3.2. The AIC statistic is of the following form:

$$\mathrm{AIC}(\boldsymbol{\lambda}) = \log\{\mathrm{RSS}(\boldsymbol{\lambda})/n\} + (2/n)e(\boldsymbol{\lambda}),$$

where $\mathrm{RSS}(\boldsymbol{\lambda})$ and $e(\boldsymbol{\lambda})$ are defined in Section 3.2.

Throughout the simulations, the kernel function is taken to be the popularly used Gaussian kernel function $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$ and the value of the bandwidth h is chosen by using the bandwidth selection procedure proposed in Section 3.2.

4.2. Simulation results with analysis

In the first simulation study, we set $\boldsymbol{\beta} = (2, 1, 1.5, 0.5, 0, 0, 0, 0, 0, 0)^{\mathrm{T}}$ and $\gamma = 0$, which implies that $p = 10$ and the explanatory variables $X_1, \ldots, X_p$ are mutually independent.

Table 1. Simulation results of the first simulation study under the first group of coefficient functions.

W	ρ	l	Method	$RMSE_o$	$RMSE_u$	C	IC	U-f(%)	R-f(%)	O-f(%)
Rook	0	10	SCAD-BIC	1.7611	0.4927	5.720	0.010	1.0	78.0	21.0
			SCAD-AIC	2.9593	0.6674	4.768	0	0	38.2	61.8
		15	SCAD-BIC	1.2898	0.4260	5.886	0	0	90.4	9.6
			SCAD-AIC	2.4943	0.5986	5.020	0	0	48.6	51.4
	0.3	10	SCAD-BIC	1.6577	0.5021	5.710	0.012	1.2	79.0	19.8
			SCAD-AIC	2.7061	0.6695	4.760	0.002	0.2	40.0	59.8
		15	SCAD-BIC	1.3365	0.4517	5.904	0	0	91.2	8.8
			SCAD-AIC	2.1261	0.6138	5.060	0	0	50.8	49.2
	0.6	10	SCAD-BIC	1.8009	0.4920	5.662	0.002	0.2	75.0	24.8
			SCAD-AIC	3.0022	0.6798	4.694	0	0	36.8	63.2
		15	SCAD-BIC	1.3628	0.4439	5.862	0	0	88.8	11.2
			SCAD-AIC	2.2288	0.6021	5.074	0	0	50.2	49.8
	0.9	10	SCAD-BIC	1.6168	0.5078	5.744	0.016	1.6	80.0	18.4
			SCAD-AIC	2.8722	0.6820	4.788	0.004	0.4	41.2	58.4
		15	SCAD-BIC	1.3588	0.4700	5.850	0	0	87.8	12.2
			SCAD-AIC	2.2397	0.6346	5.042	0	0	48.8	51.2
Exponential	0	10	SCAD-BIC	1.9190	0.5035	5.612	0.008	0.8	71.6	27.6
			SCAD-AIC	3.0418	0.6850	4.704	0	0	35.6	64.4
		15	SCAD-BIC	1.3056	0.4460	5.876	0	0	89.6	10.4
			SCAD-AIC	2.0942	0.6064	5.046	0	0	48.2	51.8
	0.3	10	SCAD-BIC	1.5195	0.4795	5.752	0.012	1.2	79.8	19.0
			SCAD-AIC	2.5496	0.6497	4.788	0	0	40.8	59.2
		15	SCAD-BIC	1.3218	0.4353	5.874	0	0	89.4	10.6
			SCAD-AIC	2.3796	0.6134	5.008	0	0	48.0	52.0
	0.6	10	SCAD-BIC	1.6704	0.4934	5.678	0.014	1.4	75.8	22.8
			SCAD-AIC	2.7455	0.6683	4.680	0	0	35.4	64.6
		15	SCAD-BIC	1.3013	0.4235	5.910	0	0	91.8	8.2
			SCAD-AIC	2.9246	0.6182	4.934	0	0	48.8	51.2
	0.9	10	SCAD-BIC	2.2088	0.4820	5.646	0.006	0.6	72.6	26.8
			SCAD-AIC	3.5089	0.6861	4.624	0.004	0.4	35.6	64.0
		15	SCAD-BIC	1.3819	0.4473	5.862	0	0	88.2	11.8
			SCAD-AIC	2.3450	0.6317	4.968	0	0	45.0	55.0

Furthermore, we take the value of the spatial autoregressive parameter ρ to be 0, 0.3, 0.6 and 0.9, respectively, to see the effect of the intensity of the spatial dependence on the performance of the proposed method. When $\rho = 0$, that is, there are no spatial effects in model (1), it reduces to the popularly used partially linear varying coefficient model. Thus, we take $\rho = 0$ to evaluate whether the proposed variable selection method still works in this special case. The simulation results under the given two groups of coefficient functions are reported in Tables 1 and 2, respectively, in which 'SCAD-AIC' and 'SCAD-BIC' stand for the proposed SCAD method with the tuning parameter being selected by AIC and BIC criteria, respectively.

We summarize some empirical findings from Tables 1 and 2 as follows. First, the SCAD-BIC significantly outperforms the SCAD-AIC because it not only gives smaller RMSE, but also has a higher probability to identify the four relevant explanatory variables and smaller probability of including irrelevant explanatory variables, which is consistent with the theoretical finding obtained by Wang et al. [37]. Second, all $RMSE_u$ ratios of the proposed penalized estimator to the unpenalized estimator are much less than 1 and can be as small as 0.45, which clearly shows that the proposed variable selection method indeed gives more accurate estimate of the regression coefficient vector $\boldsymbol{\beta}$ than the existing profile quasi-maximum likelihood method. This justifies the importance of the proposed penalized profile least squares method. Third, when the sample size n increases from 100 to 225,

Table 2. Simulation results of the first simulation study under the second group of coefficient functions.

W	ρ	l	Method	$RMSE_o$	$RMSE_u$	C	IC	U-f(%)	R-f(%)	O-f(%)
Rook	0	10	SCAD-BIC	1.9239	0.5341	5.626	0.018	1.8	71.8	26.4
			SCAD-AIC	2.9811	0.7130	4.636	0.002	0.2	32.8	67.0
		15	SCAD-BIC	1.4865	0.4561	5.802	0	0	84.8	15.2
			SCAD-AIC	2.4633	0.6207	4.944	0	0	43.6	56.4
	0.3	10	SCAD-BIC	1.8601	0.4963	5.638	0.006	0.6	75.8	23.6
			SCAD-AIC	3.0906	0.7011	4.616	0.002	0.2	35.0	64.8
		15	SCAD-BIC	1.4816	0.4569	5.866	0	0	88.8	11.2
			SCAD-AIC	2.4789	0.6284	5.018	0	0	47.0	53.0
	0.6	10	SCAD-BIC	1.8532	0.5204	5.632	0.020	2.0	70.8	27.2
			SCAD-AIC	2.9186	0.7114	4.596	0.002	0.2	32.6	67.2
		15	SCAD-BIC	1.4109	0.4565	5.848	0	0	87.2	12.8
			SCAD-AIC	2.6284	0.6266	5.040	0	0	45.4	54.6
	0.9	10	SCAD-BIC	2.0478	0.5418	5.646	0.024	2.4	70.8	26.8
			SCAD-AIC	3.2197	0.7136	4.638	0.004	0.4	32.6	67.0
		15	SCAD-BIC	1.3483	0.4462	5.866	0	0	89.8	10.2
			SCAD-AIC	2.2977	0.6184	4.956	0	0	47.8	52.2
Exponential	0	10	SCAD-BIC	1.7998	0.5256	5.596	0.020	2.0	69.8	28.2
			SCAD-AIC	2.9651	0.7100	4.540	0.002	0.2	33.8	66.0
		15	SCAD-BIC	1.5383	0.4561	5.834	0	0	86.6	13.4
			SCAD-AIC	2.5401	0.6170	5.010	0	0	47.6	52.4
	0.3	10	SCAD-BIC	1.9464	0.5055	5.618	0.010	1.0	72.6	26.4
			SCAD-AIC	3.0175	0.7070	4.530	0	0	30.2	69.8
		15	SCAD-BIC	1.3503	0.4330	5.886	0	0	90.6	9.4
			SCAD-AIC	2.2981	0.6193	4.970	0	0	49.2	50.8
	0.6	10	SCAD-BIC	1.9004	0.5131	5.650	0.012	1.2	73.6	25.2
			SCAD-AIC	2.9964	0.7050	4.606	0.002	0.2	34.4	65.4
		15	SCAD-BIC	1.3384	0.4445	5.834	0	0	86.6	13.4
			SCAD-AIC	2.3691	0.6269	4.938	0	0	44.6	55.4
	0.9	10	SCAD-BIC	2.0697	0.5294	5.580	0.014	1.4	67.2	31.4
			SCAD-AIC	4.4550	0.7056	4.578	0.002	0.2	31.4	68.4
		15	SCAD-BIC	1.4287	0.4455	5.850	0	0	88.4	11.6
			SCAD-AIC	2.4943	0.6256	4.946	0	0	47.4	52.6

both SCAD-BIC and SCAD-AIC perform better, while the performance of the SCAD-BIC becomes much closer to that of the oracle estimator. Fourth, we can see from Tables 1 and 2 that the spatial weight matrix **W** has little influence on the finite sample performance of the proposed variable selection method. Furthermore, by comparing the simulation studies under the considered four values of the spatial autoregressive parameter ρ, we can conclude that the spatial autoregressive parameter ρ is of little influence on the finite sample performance of the proposed variable selection method. This means that the proposed variable selection method works quite well no matter whether there are spatial effects in model (1) or not and no matter how strong the spatial autoregressive parameter ρ is. Fifth, by comparing the simulation results in Table 2 with those in Table 1, we can find that the simulation results under the two groups of coefficient functions have no evident difference, this indicates that the degree of complexity of coefficient functions has little impact on the performance of the proposed variable selection method. Thus, we only consider the first group of coefficient functions in the following simulation studies to save the space.

In the second simulation study, we take the value of γ to be 0.25, 0.50 and 0.75, respectively, to assess the effect of the intensity of the dependence among the explanatory variables $X_1, \ldots, X_p$ on the finite sample performance of the proposed variable selection method.

Table 3. Simulation results of the second simulation study under the first group of coefficient functions.

W	γ	l	Method	$RMSE_o$	$RMSE_u$	C	IC	U-f(%)	R-f(%)	O-f(%)
Rook	0.25	10	SCAD-BIC	2.0415	0.5006	5.658	0.020	2.0	73.0	25.0
			SCAD-AIC	3.3025	0.6741	4.740	0.004	0.4	37.2	62.4
		15	SCAD-BIC	1.4918	0.4163	5.878	0	0	90.0	10.0
			SCAD-AIC	3.4338	0.5986	5.048	0	0	50.6	49.4
	0.50	10	SCAD-BIC	2.2794	0.5501	5.670	0.092	9.2	69.4	21.4
			SCAD-AIC	3.6236	0.7021	4.752	0.024	2.4	34.2	63.4
		15	SCAD-BIC	1.4730	0.4242	5.876	0.002	0.2	88.4	11.4
			SCAD-AIC	2.9984	0.6447	5.016	0	0	44.2	55.8
	0.75	10	SCAD-BIC	3.7877	0.6678	5.632	0.334	33.4	47.4	19.2
			SCAD-AIC	4.6739	0.7600	4.718	0.128	12.8	22.8	64.4
		15	SCAD-BIC	1.7172	0.4386	5.862	0.036	3.6	85.0	11.4
			SCAD-AIC	3.6666	0.6666	4.960	0.004	0.4	36.6	63.0
Exponential	0.25	10	SCAD-BIC	1.8498	0.4969	5.742	0.030	3.0	77.4	19.6
			SCAD-AIC	3.1965	0.6751	4.770	0.006	0.6	39.4	60.0
		15	SCAD-BIC	1.4974	0.4448	5.830	0	0	86.4	13.6
			SCAD-AIC	2.6832	0.6300	4.998	0	0	46.4	53.6
	0.50	10	SCAD-BIC	2.8731	0.5352	5.600	0.060	6.0	66.0	28.0
			SCAD-AIC	4.3514	0.6924	4.782	0.020	2.0	31.6	66.4
		15	SCAD-BIC	1.4243	0.4108	5.868	0	0	88.6	11.4
			SCAD-AIC	2.8053	0.6262	4.996	0	0	43.6	56.4
	0.75	10	SCAD-BIC	3.0327	0.6270	5.664	0.286	28.6	51.8	19.6
			SCAD-AIC	4.0270	0.7589	4.732	0.156	15.6	23.6	60.8
		15	SCAD-BIC	1.7575	0.4534	5.832	0.048	4.8	81.6	13.6
			SCAD-AIC	3.9832	0.6887	4.862	0.010	1.0	31.6	67.4

The remainder of the experimental design is kept to be the same as that in the first simulation study except that we fix $\rho = 0.9$ in this simulation study. The simulation results are shown in Table 3.

We can see by comparing the results in Table 3 with those in Table 1 that, as the dependence among $X_1, \ldots, X_p$ increases, both SCAD-BIC and SCAD-AIC tend to select a model with less relevant explanatory variables and worse estimation accuracy. However, as the sample size n increases from 100 to 225, the performance of the SCAD-BIC becomes very close to that of the SCAD-BIC under the situation that $X_1, \ldots, X_p$ are mutually independent. This shows that the SCAD-BIC still works well in the case of dependence among $X_1, \ldots, X_p$ provided that the sample size is moderate.

In the third simulation study, we consider model (17) with a diverging number of parameters $\boldsymbol{\beta}$, which is quite different from the first two simulation studies. In this case, the dimension of the full model is diverging, while the dimension of the true model is fixed to be four. Specifically, the dimension of $\boldsymbol{\beta}$ is $p = \lfloor 2n^{1/2} \rfloor$, where $\lfloor x \rfloor$ is the largest integer that is smaller than or equal to x, this means that the value of p is 20 and 30 for $n = 100$ and $n = 225$, respectively, the four non-zero elements of $\boldsymbol{\beta}$ are still taken as $(2, 1, 1.5, 0.5)$, and the other components of $\boldsymbol{\beta}$ are taken to be 0. The simulation results are summarized in Table 4.

By comparing the results in Table 4 with those in Table 1, we obtain the following empirical findings. First, as the dimension of $\boldsymbol{\beta}$ increases, the index O-f becomes larger, while the index R-f becomes smaller, which indicates that the proposed variable selection method tends to select a model with more irrelevant explanatory variables, especially for the small sample size $n = 100$. Second, when the dimension of $\boldsymbol{\beta}$ increases, the value of the index $RMSE_o$ becomes larger, while the value of the index $RMSE_u$ becomes smaller. This means

Table 4. Simulation results of the third simulation study under the first group of coefficient functions.

W	ρ	l	Method	$RMSE_o$	$RMSE_u$	C	IC	U-f(%)	R-f(%)	O-f(%)
Rook	0	10	SCAD-BIC	2.7472	0.3037	15.228	0.026	2.6	61.4	36.0
			SCAD-AIC	6.5340	0.5672	11.984	0	0	16.2	83.8
		15	SCAD-BIC	1.5322	0.1479	25.728	0	0	84.0	16.0
			SCAD-AIC	6.7785	0.4121	20.936	0	0	19.4	80.6
	0.3	10	SCAD-BIC	2.7261	0.2942	15.230	0.022	2.2	61.4	36.4
			SCAD-AIC	6.1719	0.5354	12.264	0	0	16.0	84.0
		15	SCAD-BIC	1.4786	0.1512	25.734	0	0	83.6	16.4
			SCAD-AIC	6.9195	0.4160	21.084	0	0	19.8	80.2
	0.6	10	SCAD-BIC	2.9575	0.2821	15.238	0.012	1.2	62.6	36.2
			SCAD-AIC	7.7285	0.5617	11.906	0.002	0.2	18.0	81.8
		15	SCAD-BIC	1.5900	0.1526	25.722	0	0	83.0	17.0
			SCAD-AIC	6.4641	0.4080	21.232	0	0	19.8	80.2
	0.9	10	SCAD-BIC	2.5727	0.2997	15.230	0.016	1.6	61.2	37.2
			SCAD-AIC	6.1860	0.5697	11.938	0.002	0.2	13.8	86.0
		15	SCAD-BIC	1.6965	0.1481	25.712	0	0	83.6	16.4
			SCAD-AIC	6.7301	0.4126	21.074	0	0	16.2	83.8
Exponential	0	10	SCAD-BIC	3.7140	0.2861	15.214	0.016	1.6	62.6	35.8
			SCAD-AIC	8.5207	0.5603	11.890	0	0	14.6	85.4
		15	SCAD-BIC	1.7877	0.1544	25.698	0	0	83.2	16.8
			SCAD-AIC	6.2486	0.4029	21.312	0	0	22.4	77.6
	0.3	10	SCAD-BIC	2.4952	0.2923	15.186	0.026	2.6	63.0	34.4
			SCAD-AIC	6.7568	0.5632	11.944	0.006	0.6	16.8	82.6
		15	SCAD-BIC	1.5502	0.1454	25.752	0	0	84.0	16.0
			SCAD-AIC	5.3948	0.3917	21.498	0	0	22.4	77.6
	0.6	10	SCAD-BIC	7.3846	0.2941	15.310	0.028	2.8	63.6	33.6
			SCAD-AIC	11.0580	0.5683	11.930	0.002	0.2	17.0	82.8
		15	SCAD-BIC	1.6472	0.1523	25.724	0	0	83.8	16.2
			SCAD-AIC	6.6733	0.4272	20.866	0	0	19.4	80.6
	0.9	10	SCAD-BIC	2.4890	0.3013	15.214	0.020	2.0	60.6	37.4
			SCAD-AIC	5.6172	0.5575	12.102	0.004	0.4	18.0	81.6
		15	SCAD-BIC	1.5021	0.1508	25.748	0	0	86.8	13.2
			SCAD-AIC	6.1193	0.4102	21.104	0	0	20.8	79.2

that the proposed variable selection method is apt to give an estimate of $\boldsymbol{\beta}$ with worse accuracy as the dimension of $\boldsymbol{\beta}$ increases. This also means that the proposed variable selection method can significantly improve the estimation performance of the existing profile quasi-maximum likelihood method especially when the dimension of $\boldsymbol{\beta}$ is diverging. Third, when the sample size n increases from 100 to 225, the performance of the SCAD-BIC improves dramatically, but the improvement on the performance of the SCAD-AIC is rather limited. This simulation study shows that the proposed variable selection method can be still used in the situation where the number of regression coefficients $\boldsymbol{\beta}$ diverges at a rate much slower than the sample size n as long as the BIC **criterion is** used to choose the tuning parameters and the sample size is moderate.

In the fourth simulation study, we empirically investigate the influence of the bandwidth h on the finite sample performance of the proposed variable selection method. To this goal, we take the value of the bandwidth h to be $\frac{2}{3}h_{\text{opt}}$, h_{opt} and $\frac{3}{2}h_{\text{opt}}$, which correspond to 'under-smoothing', 'right-smoothing' and 'over-smoothing', respectively, where h_{opt} is the optimal value of the bandwidth h selected by the procedure described in Section 3.2. Because the simulation studies under the four values of the spatial autoregressive parameter ρ are very similar, we only report the simulation studies under $\rho = 0.9$ in Table 5 given the limited space.

Table 5. Simulation results of the fourth simulation study under the first group of coefficient functions.

W	h	l	Method	$RMSE_o$	$RMSE_u$	C	IC	U-f(%)	R-f(%)	O-f(%)
Rook	$\frac{2}{3}h_{opt}$	10	SCAD-BIC	1.8421	0.5168	5.688	0.012	1.2	77.4	21.4
			SCAD-AIC	3.1201	0.7091	4.690	0	0	37.0	63.0
		15	SCAD-BIC	1.3651	0.4408	5.878	0	0	89.6	10.4
			SCAD-AIC	2.5090	0.6182	4.998	0	0	47.2	52.8
	h_{opt}	10	SCAD-BIC	1.6720	0.5035	5.708	0.010	1.0	78.4	20.6
			SCAD-AIC	2.7101	0.6720	4.738	0	0	38.6	61.4
		15	SCAD-BIC	1.3285	0.4549	5.906	0	0	91.6	8.4
			SCAD-AIC	2.1958	0.6154	5.068	0	0	51.2	48.8
	$\frac{3}{2}h_{opt}$	10	SCAD-BIC	1.7334	0.4818	5.670	0.002	0.2	75.4	24.4
			SCAD-AIC	2.7478	0.6486	4.782	0	0	39.4	60.6
		15	SCAD-BIC	1.2918	0.4369	5.878	0	0	90.0	10.0
			SCAD-AIC	2.2177	0.6015	5.076	0	0	50.6	49.4
Exponential	$\frac{2}{3}h_{opt}$	10	SCAD-BIC	1.8277	0.5339	5.710	0.016	1.6	77.4	21.0
			SCAD-AIC	3.2824	0.7239	4.746	0.004	0.4	38.4	61.2
		15	SCAD-BIC	1.3934	0.4749	5.850	0	0	87.4	12.6
			SCAD-AIC	2.3479	0.6488	5.024	0	0	48.4	51.6
	h_{opt}	10	SCAD-BIC	1.8948	0.4990	5.618	0.006	0.6	71.8	27.6
			SCAD-AIC	3.0112	0.6905	4.708	0	0	34.4	65.6
		15	SCAD-BIC	1.3220	0.4454	5.878	0	0	89.8	10.2
			SCAD-AIC	2.1168	0.6064	5.062	0	0	49.6	50.4
	$\frac{3}{2}h_{opt}$	10	SCAD-BIC	1.4437	0.4596	5.778	0.012	1.2	81.2	17.6
			SCAD-AIC	2.4106	0.6205	4.842	0	0	42.6	57.4
		15	SCAD-BIC	1.2932	0.4270	5.884	0	0	90.0	10.0
			SCAD-AIC	2.3629	0.6056	5.030	0	0	47.6	52.4

We can see from Table 5 that the simulation results under the three considered values of the bandwidth h have not evident difference, which shows that the proposed variable selection method is quite robust to the bandwidth h.

In the fifth simulation study, we investigate the influence of the non-normality of the error distribution on the finite sample performance of the proposed variable selection method. To this end, we consider the following three non-normality error distributions whose scales are so adjusted that they all have mean zero and variance one:

(I) Uniform distribution $U(-\sqrt{3}, \sqrt{3})$.
(II) Transformed and centralized chi-square distribution $\frac{1}{2}\chi^2(2) - 1$, where $\chi^2(2)$ denotes the random variable of a chi-square distribution with 2 degrees of freedom .
(III) Transformed t distribution $\frac{1}{\sqrt{2}}t(4)$, where $t(4)$ represents the random variable of a t distribution with 4 degrees of freedom.

The simulation results under the three non-normality error distributions are summarized in Tables 6–8, respectively.

By comparing the simulation results in Tables 6–8 with those in Table 1, we can conclude that the non-normality of the error distribution has little influence on the finite sample performance of the proposed variable selection method.

5. A real data example

In this section, we apply the proposed variable selection method to analyse the Boston housing price data set. The data set consists of the median value (MV) of owner-occupied

Table 6. Simulation results of the fifth simulation study under the first group of coefficient functions and the error distribution $U(-\sqrt{3}, \sqrt{3})$.

W	ρ	l	Method	$RMSE_o$	$RMSE_u$	C	IC	U-f(%)	R-f(%)	O-f(%)
Rook	0	10	SCAD-BIC	1.6905	0.4886	5.732	0.006	0.6	78.6	20.8
			SCAD-AIC	2.9304	0.6808	4.688	0	0	38.0	62.0
		15	SCAD-BIC	1.3065	0.4489	5.878	0	0	89.6	10.4
			SCAD-AIC	2.1828	0.6211	5.014	0	0	48.8	51.2
	0.3	10	SCAD-BIC	1.7302	0.5000	5.670	0.002	0.2	75.6	24.2
			SCAD-AIC	2.5893	0.6529	4.852	0	0	41.8	58.2
		15	SCAD-BIC	1.4134	0.4486	5.856	0	0	88.8	11.2
			SCAD-AIC	2.4083	0.6042	5.076	0	0	50.2	49.8
	0.6	10	SCAD-BIC	1.8157	0.5251	5.646	0.008	0.8	74.8	24.4
			SCAD-AIC	2.8390	0.6800	4.782	0	0	38.2	61.8
		15	SCAD-BIC	1.3918	0.4430	5.868	0	0	89.2	10.8
			SCAD-AIC	2.4083	0.6139	5.020	0	0	50.4	49.6
	0.9	10	SCAD-BIC	1.9303	0.4915	5.740	0.012	1.2	81.0	17.8
			SCAD-AIC	3.6759	0.6715	4.756	0.004	0.4	40.4	59.2
		15	SCAD-BIC	1.2785	0.4434	5.884	0	0	90.2	9.8
			SCAD-AIC	2.3657	0.6174	5.014	0	0	49.4	50.6
Exponential	0	10	SCAD-BIC	2.0640	0.5113	5.614	0.004	0.4	73.6	26.0
			SCAD-AIC	3.3304	0.6753	4.746	0	0	39.4	60.6
		15	SCAD-BIC	1.3354	0.4528	5.874	0	0	89.0	11.0
			SCAD-AIC	2.2503	0.6212	5.044	0	0	50.2	49.8
	0.3	10	SCAD-BIC	2.3611	0.5135	5.614	0.008	0.8	70.6	28.6
			SCAD-AIC	3.4448	0.6757	4.724	0	0	37.2	62.8
		15	SCAD-BIC	1.3449	0.4249	5.876	0	0	89.2	10.8
			SCAD-AIC	2.4907	0.6101	4.970	0	0	46.4	53.6
	0.6	10	SCAD-BIC	1.9962	0.5026	5.694	0.018	1.8	76.0	22.2
			SCAD-AIC	3.1594	0.6569	4.824	0.004	0.4	41.6	58.0
		15	SCAD-BIC	1.3803	0.4445	5.880	0	0	89.8	10.2
			SCAD-AIC	2.3568	0.6226	4.992	0	0	45.4	54.6
	0.9	10	SCAD-BIC	1.6703	0.4948	5.664	0.010	1.0	75.0	24.0
			SCAD-AIC	2.7617	0.6675	4.714	0	0	34.8	65.2
		15	SCAD-BIC	1.4855	0.4553	5.872	0	0	89.8	10.2
			SCAD-AIC	2.3581	0.6110	5.128	0	0	51.2	48.8

homes in 506 census tracts in the Boston Standard Metropolitan Statistical Area in 1970, together with 13 variables which might explain the variation of housing value (Harrison and Rubinfeld [39]). Like that in Fan and Huang [40], the response variable and the explanatory variables used here are: MV (the median value of owner-occupied homes in $ 1000), CRIM (the per capita crime rate by town), RM (the average number of rooms per dwelling), TAX (the full value property tax rate per $10,000), NOX (the nitrogen oxide concentration in parts per 10 million), PTRATIO (the pupil-teacher ratio by town school district), AGE (the proportion of owner-occupied homes built prior to 1940) and LSTAT (the proportion of population that is in the lower status). For simplicity of notation, the response variable MV and the six explanatory variables CRIM, RM, TAX, NOX, PTRATIO and AGE are denoted by Y, X_2, X_3, X_4, X_5, X_6 and X_7, respectively.

Taking $X_1 = 1$ and $U = \sqrt{\text{LSTAT}}$, Fan and Huang [40] first used the following varying coefficient model:

$$Y = a_1(U) + a_2(U)X_2 + a_3(U)X_3 + a_4(U)X_4 + a_5(U)X_5 + a_6(U)X_6 + a_7(U)X_7 + \varepsilon \tag{18}$$

to fit the data set. Then they applied the GLR test proposed by Fan et al. [41] to see whether each coefficient function varies significantly and found that the coefficients of PTRATIO

Table 7. Simulation results of the fifth simulation study under the first group of coefficient functions and the error distribution $\frac{1}{2}\chi^2(2) - 1$.

W	ρ	l	Method	$RMSE_o$	$RMSE_u$	C	IC	U-f(%)	R-f(%)	O-f(%)
Rook	0	10	SCAD-BIC	1.7409	0.5103	5.746	0.010	1.0	79.0	20.0
			SCAD-AIC	2.6552	0.6774	4.838	0.004	0.4	42.0	57.6
		15	SCAD-BIC	1.2976	0.4273	5.908	0	0	92.4	7.6
			SCAD-AIC	2.5030	0.5831	5.100	0	0	53.6	46.4
	0.3	10	SCAD-BIC	1.9784	0.5248	5.656	0.024	2.4	73.0	24.6
			SCAD-AIC	3.0838	0.6855	4.700	0	0	37.4	62.6
		15	SCAD-BIC	1.4782	0.4355	5.882	0	0	89.8	10.2
			SCAD-AIC	2.5443	0.6262	4.946	0	0	47.6	52.4
	0.6	10	SCAD-BIC	1.6374	0.5064	5.700	0.008	0.8	76.4	22.8
			SCAD-AIC	2.8284	0.6881	4.694	0	0	38.0	62.0
		15	SCAD-BIC	1.3814	0.4238	5.876	0	0	90.0	10.0
			SCAD-AIC	2.7649	0.6068	4.950	0	0	49.0	51.0
	0.9	10	SCAD-BIC	1.8789	0.5174	5.642	0.014	1.4	74.6	24.0
			SCAD-AIC	2.7838	0.6682	4.820	0.002	0.2	39.8	60.0
		15	SCAD-BIC	1.2688	0.4266	5.906	0	0	92.4	7.6
			SCAD-AIC	2.3773	0.6045	4.998	0	0	49.8	50.2
Exponential	0	10	SCAD-BIC	1.8763	0.5122	5.690	0.018	1.8	75.2	23.0
			SCAD-AIC	2.9847	0.6717	4.726	0	0	40.0	60.0
		15	SCAD-BIC	1.3125	0.4400	5.884	0	0	90.4	9.6
			SCAD-AIC	2.4725	0.6108	5.022	0	0	48.2	51.8
	0.3	10	SCAD-BIC	1.8121	0.5221	5.680	0.018	1.8	73.6	24.6
			SCAD-AIC	2.6037	0.6709	4.852	0.004	0.4	40.2	59.4
		15	SCAD-BIC	1.2523	0.4326	5.908	0	0	92.6	7.4
			SCAD-AIC	2.1890	0.6077	5.014	0	0	50.4	49.6
	0.6	10	SCAD-BIC	1.6625	0.5027	5.688	0.010	1.0	74.8	24.2
			SCAD-AIC	2.4699	0.6555	4.840	0.002	0.2	41.6	58.2
		15	SCAD-BIC	1.3653	0.4465	5.868	0	0	89.8	10.2
			SCAD-AIC	2.3487	0.6195	5.026	0	0	48.8	51.2
	0.9	10	SCAD-BIC	1.8214	0.4885	5.638	0.014	1.4	73.2	25.4
			SCAD-AIC	3.0918	0.6759	4.688	0.002	0.2	34.2	65.6
		15	SCAD-BIC	1.5634	0.4367	5.854	0	0	88.6	11.4
			SCAD-AIC	2.5285	0.6059	5.040	0	0	47.6	52.4

and AGE do not vary significantly at the significance level 1%. Thus they set the coefficients of PTRATIO and AGE to be constants and employed the partially linear varying coefficient model

$$Y = a_1(U) + a_2(U)X_2 + a_3(U)X_3 + a_4(U)X_4 + a_5(U)X_5 + \beta_1 X_6 + \beta_2 X_7 + \varepsilon \quad (19)$$

to fit the data set again. On the other hand, the presence of the spatial dependence in the housing market has been extensively emphasized in the literature. By incorporating a spatially lagged term of the response variable into model (19), Wei et al. [19] built a partially linear varying coefficient spatial autoregressive model to fit the data set. However, the model built by Wei et al. [19] does not consider the interaction and the quadratic terms of the explanatory variables X_6 and X_7, which may lead to modelling bias. To this purpose, we employed a saturated partially linear varying coefficient spatial autoregressive model

$$\begin{aligned} Y_i = \rho \sum_{j \neq i} w_{ij} Y_j + a_1(U_i) + a_2(U_i)X_{i2} + a_3(U_i)X_{i3} + a_4(U_i)X_{i4} + a_5(U_i)X_{i5} \\ + \beta_1 X_{i6} + \beta_2 X_{i7} + \beta_3 X_{i6}^2 + \beta_4 X_{i6}X_{i7} + \beta_5 X_{i7}^2 + \varepsilon_i,\ i = 1, \ldots, 506 \end{aligned} \quad (20)$$

Table 8. Simulation results of the fifth simulation study under the first group of coefficient functions and the error distribution $\frac{1}{\sqrt{2}}t(4)$.

W	ρ	l	Method	RMSE_o	RMSE_u	C	IC	U-f(%)	R-f(%)	O-f(%)
Rook	0	10	SCAD-BIC	1.8680	0.5110	5.700	0.018	1.8	78.0	20.2
			SCAD-AIC	2.8463	0.6765	4.792	0.002	0.2	37.4	62.4
		15	SCAD-BIC	1.3103	0.4288	5.874	0	0	90.8	9.2
			SCAD-AIC	2.6112	0.6060	4.954	0	0	46.8	53.2
	0.3	10	SCAD-BIC	2.1600	0.4952	5.718	0.014	1.4	78.4	20.2
			SCAD-AIC	3.8055	0.6787	4.736	0.002	0.2	35.6	64.2
		15	SCAD-BIC	1.3944	0.4592	5.860	0	0	88.0	12.0
			SCAD-AIC	2.2885	0.6237	5.010	0	0	48.4	51.6
	0.6	10	SCAD-BIC	1.7301	0.5292	5.666	0.026	2.4	73.8	23.8
			SCAD-AIC	2.5272	0.6733	4.800	0.004	0.4	38.2	61.4
		15	SCAD-BIC	1.3134	0.4482	5.872	0	0	88.8	11.2
			SCAD-AIC	2.4823	0.6354	4.890	0	0	43.6	56.4
	0.9	10	SCAD-BIC	1.6901	0.5056	5.692	0.014	1.4	75.2	23.4
			SCAD-AIC	2.8863	0.6826	4.736	0.002	0.2	38.8	61.0
		15	SCAD-BIC	1.3198	0.4611	5.866	0	0	88.6	11.4
			SCAD-AIC	2.0460	0.6013	5.096	0	0	53.2	46.8
Exponential	0	10	SCAD-BIC	1.7645	0.4889	5.688	0.014	1.4	76.6	22.0
			SCAD-AIC	2.8802	0.6630	4.702	0	0	39.2	60.8
		15	SCAD-BIC	1.2045	0.4418	5.900	0	0	91.0	9.0
			SCAD-AIC	2.1843	0.6002	5.098	0	0	53.4	46.6
	0.3	10	SCAD-BIC	1.7876	0.4762	5.714	0.020	2.0	77.8	20.2
			SCAD-AIC	2.9853	0.6665	4.690	0.008	0.8	37.4	61.8
		15	SCAD-BIC	1.2525	0.4445	5.872	0	0	90.0	10.0
			SCAD-AIC	2.0334	0.6094	5.010	0	0	47.4	52.6
	0.6	10	SCAD-BIC	2.0782	0.4817	5.714	0.008	0.8	77.4	21.8
			SCAD-AIC	3.2613	0.6800	4.716	0	0	36.8	63.2
		15	SCAD-BIC	1.3117	0.4384	5.898	0	0	92.2	7.8
			SCAD-AIC	2.3376	0.6231	5.006	0	0	47.8	52.2
	0.9	10	SCAD-BIC	1.7879	0.5234	5.704	0.016	1.6	76.8	21.6
			SCAD-AIC	2.7519	0.6754	4.776	0.002	0.2	42.2	57.6
		15	SCAD-BIC	1.2939	0.4388	5.884	0	0	90.2	9.8
			SCAD-AIC	2.3064	0.6105	5.088	0	0	50.0	50.0

Table 9. A summary of symbols.

Symbol	Description
$(\boldsymbol{\beta}, \rho, \sigma^2)$	Parameters in partially linear varying coefficient spatial autoregressive model
$\mathbf{a}(\cdot)$	Coefficient function vector in partially linear varying coefficient spatial autoregressive model
$\log L(\mathbf{a}(\cdot), \boldsymbol{\beta}, \rho, \sigma^2)$	Quasi log-likelihood function
$\log L(\boldsymbol{\beta}, \rho, \sigma^2)$	Profile quasi log-likelihood function
$\log L(\rho)$	Centred quasi log-likelihood function of ρ
$Q(\boldsymbol{\beta})$	Penalized least squares function of $\boldsymbol{\beta}$
$Q_{\text{P}}(\boldsymbol{\beta})$	Penalized profile least squares function of $\boldsymbol{\beta}$
$\widehat{\mathbf{a}}(u; \boldsymbol{\beta}, \rho)$	Local linear estimator of $\mathbf{a}(u)$ given $\boldsymbol{\beta}$ and ρ
$\widehat{\boldsymbol{\beta}}(\rho)$	Profile quasi-maximum likelihood estimator of $\boldsymbol{\beta}$ given ρ
$\widehat{\sigma}^2(\rho)$	Profile quasi-maximum likelihood estimator of σ^2 given ρ
$\widehat{\rho}$	Profile quasi-maximum likelihood estimator of ρ
$\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}(\widehat{\rho})$	Final profile quasi-maximum likelihood estimator of $\boldsymbol{\beta}$
$\widehat{\sigma}^2 \equiv \widehat{\sigma}^2(\widehat{\rho})$	Final profile quasi-maximum likelihood estimator of σ^2
$\widehat{\mathbf{a}}(u) \equiv \widehat{\mathbf{a}}(u; \widehat{\boldsymbol{\beta}}, \widehat{\rho})$	Final local linear estimator of $\mathbf{a}(u)$
$\widehat{\boldsymbol{\beta}}_{\text{P}}$	Penalized profile least squares estimator of $\boldsymbol{\beta}$
$p_{\lambda_j}(\cdot)$	Penalty function with a tuning parameter
$K(\cdot)$	Kernel function used in the local linear smoothing method
h	Bandwidth used in the local linear smoothing method
$\text{CV}(\cdot)$	Cross-validation (CV) statistic
$\text{BIC}(\cdot)$	Bayesian information criterion (BIC) statistic
$\text{AIC}(\cdot)$	Akaika information criterion (AIC) statistic

to fit the data set, where $w_{ij} = \exp(-d_{ij})/\sum_{l\neq i}\exp(-d_{il})$ with d_{ij} being the Euclidean distance in terms of the Cartesian coordinates of census tracts i and j.

We applied the proposed variable selection method to identify the significant explanatory variables in the parametric component of model (20), and estimate the unknown coefficient functions and parameters in the selected model. The kernel function used here was still taken to be the Gaussian kernel function, and the bandwidth h was selected by using the simple rule of thumb (ROT) method, that is, $h = s_u n^{-1/5}$, where s_u is the standard deviation of the observations $U_1, \ldots, U_n$ of the smoothing variable U. The tuning parameter λ was selected by the BIC approach proposed in Section 3.2.

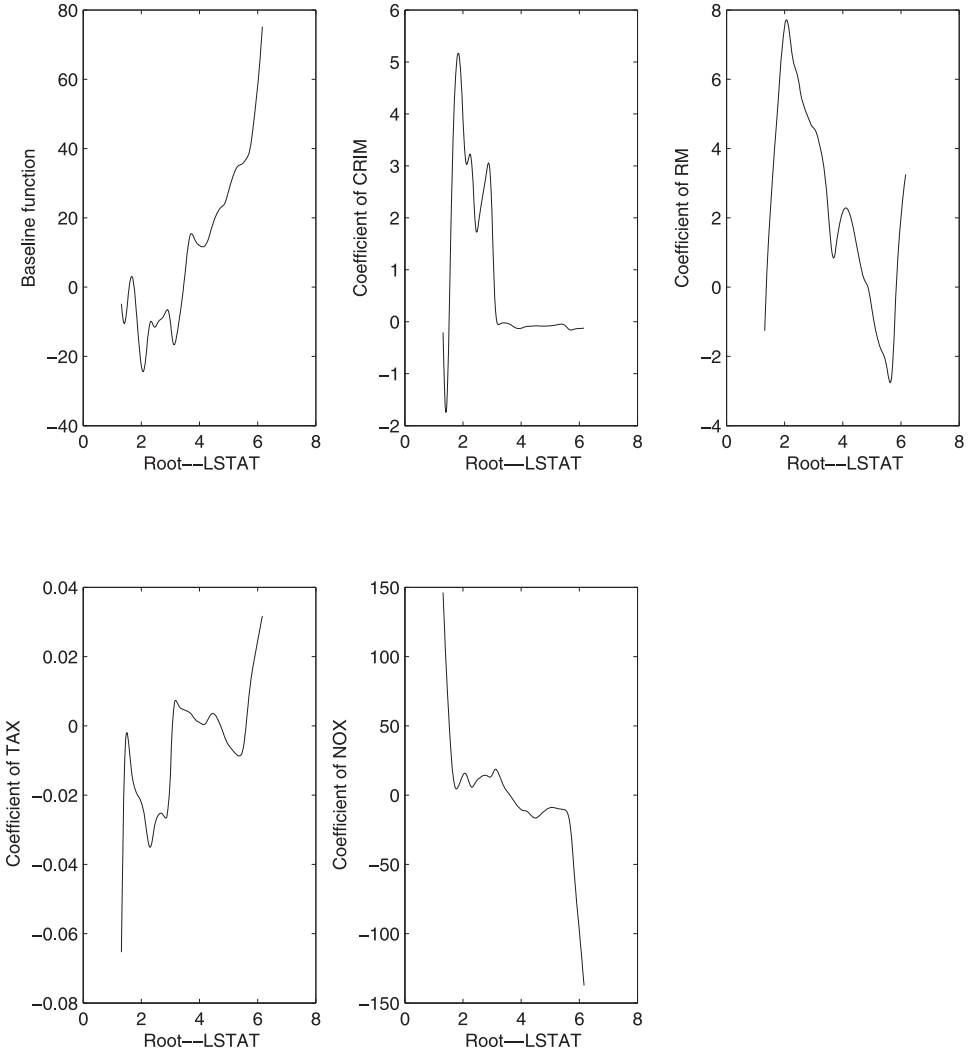

Figure 1. Estimated curves of functional coefficients in the selected model for Boston housing price dataset.

The proposed variable selection method selected the terms X_6^2 and X_7 into model (20). The final linear component of the regression function is estimated as $-0.0155X_7 - 0.0145X_6^2$, which indicates that both X_7 and X_6^2 have negative impact on the house price. The estimate of the spatial autoregressive parameter ρ is 0.298, which means that the house prices among census tracts do affect each other. This is a true phenomenon in the real world. Finally, we gave the estimated curves of the coefficient functions in Figure 1. From Figure 1, we can see clearly that the associations between the housing price and the explanatory variables X_2, X_3, X_4 and X_5 do vary with the smoothing variable U.

6. Concluding remarks

In this paper, we proposed a penalized profile least-squares method to select the important explanatory variables in the parametric component of partially linear varying coefficient spatial autoregressive model. Compared to the existing estimation methods, the proposed penalized profile least squares method can simultaneously select significant explanatory variables in the parametric component and estimate the corresponding non-zero regression coefficients. The simulation results and real data analysis both indicate that the proposed penalized profile least squares method works well in selecting relevant explanatory variables and estimating the corresponding non-zero regression coefficients in the parametric component of partially linear varying coefficient spatial autoregressive model.

Some interesting future research topics related to this paper should be mentioned. First, although our simulation studies show that the proposed variable selection method indeed works well in finite samples, the theoretical properties of the proposed variable selection method need to be further investigated. Second, in this paper, we only considered the case where the number of regression coefficients p is fixed. In some situations, the number of regression coefficients may increase with the sample size n. In other cases, the number of regression coefficients may be much greater than the sample size n. How to extend our method to these cases deserves further study.

Third, in this paper, we only focused on the problem of selecting the explanatory variables with non-zero constant coefficients in partially linear varying coefficient spatial autoregressive model. However, to further improve estimation accuracy and model interpretability, one needs to study the problem of selecting the explanatory variables with non-zero varying coefficients in partially linear varying coefficient spatial autoregressive model. More generally, how to develop a variable selection method to simultaneously select the explanatory variables with non-zero constant and varying coefficients in partially linear varying coefficient spatial autoregressive model is an interesting but challenging research topic.

Fourth, it is noted that the spatial weight matrix in this paper is assumed to be exogenous. This exogenous assumption is reasonable if the spatial weight matrix is constructed based on contiguity relationship or geographic distance among spatial units. However, in some practical applications especially in the field of economics, it is a common practice to construct the spatial weight matrix by using economic or socioeconomic distances. In this case, the spatial weight matrix is likely to be endogenous. Thus, it is appealing and necessary to study the variable selection problems of the partially linear varying coefficient spatial autoregressive model with an endogenous spatial weight matrix .

6.1. A summary of symbols

According to a suggestion of the associate editor, we add a summary of symbols to make the symbols used in this paper more clear (Table 9).

Acknowledgments

The authors are grateful to Editor Richard G. Krutchkoff, the associate editor, and the anonymous referee for their constructive comments that greatly improved the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by the National Natural Science Foundation of China [grant numbers 11671317 and 11972273], the National Statistical Science Project [grant number 2019LY36], the Scientific Research Foundation of Education Department of Shaanxi Province of China [grant number 17JK0423] and the China Scholarship Council [grant number 201808615033].

References

[1] Lee LF. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. Econometrica. 2004;72:1899–1925.
[2] Kelejian HH, Prucha IR. Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. J Econom. 2010;157:53–67.
[3] Lin X, Lee LF. GMM estimation of spatial autoregressive models with unknown heteroskedasticity. J Econom. 2010;157:34–52.
[4] Drukker DM, Egger P, Prucha IR. On two-step estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors. Econom Rev. 2013;32:686–733.
[5] Anselin L. Rao' score tests in spatial econometrics. J Stat Plan Inference. 2001;97:113–139.
[6] Baltagi BH, Yang ZL. Heteroscedasticity and non-normality robust LM tests of spatial dependence. Reg Sci Urban Econ. 2013;43:725–739.
[7] Yang ZL. LM tests of spatial dependence based on bootstrap critical values. J Econom. 2015;185:33–39.
[8] Wu YQ, Sun Y. Shrinkage estimation of the linear model with spatial interaction. Metrika. 2017;80:51–68.
[9] Liu X, Chen JB, Cheng SL. A penalized quasi-maximum likelihood method for variable selection in the spatial autoregressive model. Spat Stat. 2018;25:86–104.
[10] Xie TF, Cao RY, Du J. Variable selection for spatial autoregressive models with a diverging number of parameters. Stat Papers. 2020;61:1125–1145.
[11] Kostov P. A spatial quantile regression hedonic model of agriculture land prices. Spatial Econom Anal. 2009;4:53–72.
[12] Basile R, Gress B. Semi-parametric spatial auto-covariance models of regional growth in Europe. Région et Dévelop. 2005;21:93–118.
[13] Baltagi BH, Li D. LM tests for functional form and spatial error correlation. Int Reg Sci Rev. 2001;24:194–225.
[14] Yang ZL, Li CW, Tse YK. Functional form and spatial dependence in dynamic panels. Econ Lett. 2006;91:138–145.
[15] Xu XB, Lee LF. A spatial autoregressive model with a nonlinear transformation of the dependent variable. J Econom. 2015;186:1–18.
[16] Su LJ. Semi-parametric GMM estimation of spatial autoregressive models. J Econom. 2012;167:543–560.

[17] Su LJ, Jin SN. Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. J Econom. 2010;157:18–33.
[18] Li KM, Chen JB. Profile maximum likelihood estimation of semi-parametric varying coefficient spatial lag model. J Quant Tech Econom. 2013;30:85–98.
[19] Wei CH, Guo S, Zhai SF. Statistical inference of partially linear varying coefficient spatial autoregressive models. Econ Model. 2017;64:553–559.
[20] Sun YQ, Zhang YQ, Huang JH. Estimation of a semi-parametric varying-coefficient mixed regressive spatial autoregressive model. Econom Stat. 2019;9:140–155.
[21] Huang J, Horowitz JL, Wei FR. Variable selection in nonparametric additive models. Ann Stat. 2010;38:2282–2313.
[22] Wang HS, Xia YC. Shrinkage estimation of the varying coefficient model. J Am Stat Assoc. 2009;104:747–757.
[23] Li RZ, Liang H. Variable selection in semiparametric regression modeling. Ann Stat. 2008;36:261–286.
[24] Xie HL, Huang J. SCAD-penalized regression in high-dimensional partially linear models. Ann Stat. 2009;37:673–696.
[25] Peng H, Huang T. Penalized least squares for single index models. J Stat Plan Inference. 2011;141:362–1379.
[26] Liang H, Liu X, Li RZ, et al. Estimation and testing for partially linear single-index models. Ann Stat. 2010;38:3811–3836.
[27] Liu X, Wang L, Liang H. Estimation and variable selection for semiparametric additive partially linear models. Stat Sin. 2011;21:1225–1248.
[28] Wang HN, Zhu J. Variable selection in spatial regression via penalized least squares. Canadian J Stat. 2009;37:607–624.
[29] Huang H-C, Hsu N-J, Theobald DM, et al. Spatial LASSO with applications to GIS model selection. J Comput Graph Stat. 2010;19:963–983.
[30] Zhu J, Huang H-C, Reyes PE. On selection of spatial linear models for lattice data. J Roy Stat Soc B. 2010;72:389–402.
[31] Chu TJ, Zhu J, Wang HN. Penalized maximum likelihood estimation and variable selection in geostatistics. Ann Stat. 2011;39:2607–2625.
[32] Nandy S, Lim CY, Maiti T. Additive model building for spatial regression. J R Stat Soc B. 2017;79:779–800.
[33] Wang KN. Variable selection for spatial semivarying coefficient models. Ann Inst Stat Math. 2018;70:323–351.
[34] Fan JQ, Li RZ. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96:1348–1360.
[35] Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc B. 1996;58:267–288.
[36] Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. Technometrics. 1993;35:109–148.
[37] Wang HS, Li RZ, Tsai CL. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika. 2007;94:553–568.
[38] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101:1418–1429.
[39] Harrison D, Rubinfeld DL. Hedonic housing prices and the demand for clean air. J Environ Econ Manage. 1978;5:81–102.
[40] Fan JQ, Huang T. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. Bernoulli. 2005;11:1031–1057.
[41] Fan JQ, Zhang CM, Zhang J. Generalized likelihood ratio statistic and Wilks phenomenon. Ann Stat. 2001;29:153–193.