mathematics

Article

Variable Selection for the Spatial Autoregressive Model with Autoregressive Disturbances

Xuan Liu [1] **and Jianbao Chen** [2,*]

[1] School of Mathematics and Information Science, Nanchang Normal University, Nanchang 330032, China; liuxuanyg@163.com
[2] College of Mathematics and Statistics, Fujian Normal University, Fuzhou 350117, China
[*] Correspondence: jbchen@fjnu.edu.cn

Abstract: Along with the rapid development of the geographic information system, high-dimensional spatial heterogeneous data has emerged bringing theoretical and computational challenges to statistical modeling and analysis. As a result, effective dimensionality reduction and spatial effect recognition has become very important. This paper focuses on variable selection in the spatial autoregressive model with autoregressive disturbances (SARAR) which contains a more comprehensive spatial effect. The variable selection procedure is presented by using the so-called penalized quasi-likelihood approach. Under suitable regular conditions, we obtain the rate of convergence and the asymptotic normality of the estimators. The theoretical results ensure that the proposed method can effectively identify spatial effects of dependent variables, find spatial heterogeneity in error terms, reduce the dimension, and estimate unknown parameters simultaneously. Based on step-by-step transformation, a feasible iterative algorithm is developed to realize spatial effect identification, variable selection, and parameter estimation. In the setting of finite samples, Monte Carlo studies and real data analysis demonstrate that the proposed penalized method performs well and is consistent with the theoretical results.

Keywords: spatial; variable selection; SCAD; penalized method

Citation: Xuan, L.; Jianbao, C. Variable Selection for the Spatial Autoregressive Model with Autoregressive Disturbances. *Mathematics* **2021**, *9*, 1448. https://doi.org/10.3390/math9121448

Academic Editor: María Purificación Galindo Villardón

Received: 8 May 2021
Accepted: 17 June 2021
Published: 21 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Spatial econometric models are mainly used to deal with spatial dependent data in applications. Spatial dependence across sectional units may concern a spatial autocorrelation in a dependent variable or disturbance term. The first form of dependence is usually defined by a spatial autoregressive (SAR) model and another by a spatial error model (SEM). In fact, both spatial dependencies may be reflected in a spatial autoregressive model with autoregressive disturbances (SARAR). These models were first introduced by Cliff and Ord [1], which have aroused wide concern, see, e.g., the research by Kelejian and Prucha [2], Lee [3], Arraiz et al. [4], and the books by Anselin [5] and Cressie [6].

In practice, explanatory variables are needed to be chosen from a number of variables during the initial data analysis. How to select significant variables to keep in the final model becomes very important for further analysis. Therefore, variable selection has received increasing attention in statistical modeling and inference. However, the study of variable selection in spatial econometric models is not as sufficient as that in classical linear models due to the complexity caused by spatial dependence. The main goal of our analysis is to fill some gaps in this area to a certain degree. We mainly focus on a variable selection method for the SARAR model based on a penalized quasi-likelihood method and investigate its oracle property. Furthermore, a feasible algorithm is given for realizing these procedures.

The methods of variable selection for classical linear models have been developed rapidly since the Akaike information criterion (AIC) was proposed by Akaike [7]. Then, similar methods based on the information criterion have progressed remarkably, such as the Bayesian information criterion (BIC) [8], risk inflation criterion (RIC) [9], etc. Using these criteria, the best subset selection became the standard method to select covariants for a long time. Although they are practically useful, the common drawback is the lack of stability and incorporating stochastic errors from each stage of variable selection as noted by Liang and Li [10]. Moreover, it may require a comparison of all possible submodels. This is a combinational problem with NP-complexity [11]. In order to overcome these drawbacks, penalized methods of variable selection have been proposed in recent years, including least absolute shrinkage and selection operator (LASSO) [12], smoothly clipped absolute deviation (SCAD) penalty [13], elastic-net (ENet) [14], adaptive LASSO [15], minimax concave penalty (MCP) [16], and so on. These methods can select significant variables and estimate unknown parameters simultaneously. Fan and Li [13] established the oracle property in the sense that the penalized estimator behaves the same as the ordinary least squares estimator as we know the true linear model, which can be used to assess the efficiency of the penalized estimator. In the Bayesian framework, some developments include Mitchell and Beauchamp [17], Raftery et al. [18], Jiang [19], etc. Other related methods can be found in Chen et al. [20] and Steel [21].

Along with the rapid development of the geographic information system (GIS), variable selection for the spatial econometric models has become a new concern in the last 10 years or so. Based on the Bayesian idea, LeSage and Parent [22] developed the Bayesian model averaging (BMA) technique for the SAR model and SEM. Some extension works include LeSage and Fischer [23] and Cuaresma et al. [24,25]. In order to avoid the complex calculation of marginal likelihoods in BMA, Piribauer [26] used stochastic search variable selection (SSVS) prior to deal with the identification of the SAR model. Generally, it is challenging to extend the penalized methods to data that are dependent either over time or across space, as variable selection involves not only regression coefficients but also autocorrelation coefficients [27]. In recent years, Liu et al. [28] gave an efficient variable selection procedure for the SAR model and obtained the large sample properties by a penalized quasi-likelihood method. Using SCAD penalty and instrumental variable, Xie et al. [29] considered variable selection in the SAR model with a diverging number of parameters. They showed that the SCAD penalty in the SAR model for variable selection also has a nice oracle property as in the classical linear model.

High dimensional spatial data may lead to complex and multiple spatial dependencies. However, the existing methods are constrained by dimension and spatial heterogeneity, which brings great challenges to the application of traditional spatial econometric models. Although the technology of dimension reduction by eliminating redundant information through variable selection in classical linear models is being gradually developed and the research on variable selection in spatial lag models has been completed, it is still difficult to effectively solve the problem of variable selection with spatial heterogeneity in error terms. The SARAR model has both a dependent variable spatial effect and spatial error term, it can reflect spatial effect information and describe spatial heterogeneity relatively comprehensively. Moreover, once the spatial effect of error is ignored, it will lead to model recognition errors, reduce the estimation error and prediction accuracy, and bring concerns to the application research. In light of the above considerations and the excellent performance of penalized methods, we studied the variable selection of spatial cross-section data based on the SARAR model. The main contributions are as follows: (1) For high-dimensional spatial heterogeneous data, a penalty quasi-likelihood method is proposed to solve the problem of dimensionality reduction of explanatory variables and the identification of two kinds of spatial effects. (2) Using the idea of step-by-step transformation, a new iterative numerical algorithm is proposed to avoid the influence of spatial heterogeneity. (3) Simulation and case analysis will help practitioners in related fields to use reasonably. (4) The proposed method can provide a useful reference for the study of variable selection in semi-parametric and nonparametric spatial regression models.

The remainder of this paper is as follows. Section 2 presents a penalized quasi-likelihood method in the SARAR model. Section 3 introduces a feasible algorithm to complete a variable selection procedure. Section 4 provides a Monte Carlo study to investigate the finite sample performance. Section 5 illustrates the proposed method through an application of the Boston housing data. Summary and discussion is stated in Section 6. Appendixes A and B contain some assumptions and proofs of theorems.

2. Model and Variable Selection

2.1. The SARAR Model

The SARAR model can be specified as:

$$\begin{aligned} \boldsymbol{Y}_n &= \rho_1 \boldsymbol{W}_{1n} \boldsymbol{Y}_n + \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{U}_n, \\ \boldsymbol{U}_n &= \rho_2 \boldsymbol{W}_{2n} \boldsymbol{U}_n + \boldsymbol{E}_n, \end{aligned} \tag{1}$$

where $\boldsymbol{Y}_n$ denotes an $n \times 1$ vector of observations on the dependent variable, $\boldsymbol{X}_n$ is an $n \times k$ matrix of observations on k exogenous explanatory variables, $\boldsymbol{W}_{1n}$ and $\boldsymbol{W}_{2n}$ are known $n \times n$ spatial weight matrices, $\boldsymbol{\beta}$ is a k-dimensional parameter vector of regression coefficients, ρ_1 and ρ_2 are scalar spatial autoregressive coefficients with $|\rho_1| < 1$ and $|\rho_2| < 1$, $\boldsymbol{U}_n$ is an $n \times 1$ vector of regression disturbances, both $\boldsymbol{W}_{1n}\boldsymbol{Y}_n$ and $\boldsymbol{W}_{2n}\boldsymbol{U}_n$ are the spatial lag term and spatial error lag term respectively, and $\boldsymbol{E}_n = (e_1, \cdots, e_n)^{\mathrm{T}}$ is an n-dimensional vector of i.i.d. innovations with zero mean and finite variance σ^2. Note that this model is also known as the Cliff–Ord model or the SARAR(1,1) model. The SAR model and SEM are corresponding to $\rho_2 = 0$ and $\rho_1 = 0$, respectively.

Let $\boldsymbol{\theta}_0 = (\sigma_0^2, \rho_{10}, \rho_{20}, \boldsymbol{\beta}_0^{\mathrm{T}})^{\mathrm{T}} = (\theta_{1,0}, \theta_{2,0}, \cdots, \theta_{k+3,0})^{\mathrm{T}}$ be the true value of $\boldsymbol{\theta}$, and $\boldsymbol{\theta} = (\sigma^2, \rho_1, \rho_2, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} = (\theta_1, \theta_2, \cdots, \theta_{k+3})^{\mathrm{T}}$. Denote $\boldsymbol{S}_{1n}(\rho_1) = \boldsymbol{I}_n - \rho_1 \boldsymbol{W}_{1n}$, $\boldsymbol{S}_{2n}(\rho_2) = \boldsymbol{I}_n - \rho_2 \boldsymbol{W}_{2n}$, $\boldsymbol{E}_n(\gamma) = \boldsymbol{S}_{2n}(\rho_2)(\boldsymbol{S}_{1n}(\rho_1)\boldsymbol{Y}_n - \boldsymbol{X}_n\boldsymbol{\beta})$, where $\gamma = (\rho_1, \rho_2, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$. According to the idea of quasi-maximum likelihood estimation [3], we can write the log-quasi-likelihood function of the model (1) as

$$\ln L_n(\boldsymbol{\theta}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + \ln|\boldsymbol{S}_{1n}(\rho_1)| + \ln|\boldsymbol{S}_{2n}(\rho_2)| - \frac{1}{2\sigma^2}\boldsymbol{E}_n^{\mathrm{T}}(\gamma)\boldsymbol{E}_n(\gamma), \tag{2}$$

where $L_n(\boldsymbol{\theta})$ is the quasi-likelihood function of the model (1).

2.2. Penalized Method

The spatial econometric research has shown that it is inappropriate to use the the ordinary least squares estimation (OLS) method directly for SAR models. In the case of the SARAR model, the OLS estimators of the spatial autoregressive coefficients are biased and inconsistent. Therefore, the penalized least squared method can not be directly used for variable selection in this model. Considering a good performance of the quasi-maximum likelihood estimation in the SARAR model, the penalized quasi-likelihood method deserves priority. We start with a penalized quasi-likelihood function for the model (1) defined as:

$$J(\boldsymbol{\theta}) = -\ln L_n(\boldsymbol{\theta}) + n\sum_{j=2}^{k+3} p_{\lambda_j}(|\theta_j|), \tag{3}$$

where $p_\lambda(\cdot)$ is the SCAD penalty function defined by Fan and Li [13] as:

$$p'_\lambda(\vartheta) = \lambda\left\{\mathrm{I}(\vartheta \le \lambda) + \frac{(a\lambda - \vartheta)_+}{(a-1)\lambda}\mathrm{I}(\vartheta > \lambda)\right\} \text{ for } \vartheta > 0 \text{ and some } a > 2.$$

For comparison, we also introduce the following two popular penalty functions.

1. HARD thresholding penalty function:

$$p_\lambda(|\vartheta|) = \lambda^2 - (|\vartheta| - \lambda)^2\mathrm{I}(|\vartheta| < \lambda).$$

2. L_1 penalty function:

$$p_\lambda(|\vartheta|) = \lambda|\vartheta|.$$

In fact, the L_1 penalty function corresponds to the LASSO [12]. The AIC and BIC correspond to the penalty functions $p_\lambda(\vartheta) = n^{-1}\mathrm{I}(\vartheta \neq 0)$ and $p_\lambda(\vartheta) = n^{-1}\log(n)\mathrm{I}(\vartheta \neq 0)$ respectively because $\sum\limits_j \mathrm{I}(\vartheta_j \neq 0)$ gives the size of the selected submodel.

In the classical linear models, Fan and Li [13] proposed that a perfect variable selection method should possess the following three properties:

(1) Unbiasedness: The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias;
(2) Sparsity: The resulting estimator automatically sets small estimated coefficients to zero to reduce model complexity;
(3) Continuity: The resulting estimator is continuous in data to avoid instability in the model prediction.

Under some regular conditions, they showed that variable selection via the SCAD penalty function possesses above properties, but the other penalty functions proposed above may not satisfy the three properties simultaneously. Related references can be seen in Fan and Li [13], and Wang and Zhu [30] for more information.

2.3. Main Results

Note that it may be chaotic in the arrangement of the original non-zero elements of $\boldsymbol{\theta}_0$. Re-labeling $\boldsymbol{\theta}_0$ can put the non-zero elements in the front together and separate them from the zero elements, which is convenient for the concise expression of the theorems and proofs. Therefore, denote $\boldsymbol{\theta}_0 = \left(\boldsymbol{\theta}_{10}^{\mathrm{T}}, \boldsymbol{\theta}_{20}^{\mathrm{T}}\right)^{\mathrm{T}}$, where we assume that $\boldsymbol{\theta}_{10}$ is a vector containing s nonzero elements and $\boldsymbol{\theta}_{20} = \mathbf{0}$ is a $(k+3-s)$-dimensional zero vector. $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\theta}}_1^{\mathrm{T}}, \hat{\boldsymbol{\theta}}_2^{\mathrm{T}}\right)^{\mathrm{T}}$ is the penalized quasi-likelihood estimator of $\boldsymbol{\theta}$. The theorems stated below give some satisfactory properties of a large sample.

Theorem 1. *Suppose that $\sqrt{n}a_n = o(1), b_n = o(1)$, and the assumptions in Appendix A hold. Then there is a local minimizer $\hat{\boldsymbol{\theta}}$ of $J(\boldsymbol{\theta})$ such that:*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p\left(n^{-1/2}\right),$$

where $a_n = \max_{2\leq j\leq s}\left\{p'_{\lambda_{jn}}(|\theta_{j,0}|)\right\}, b_n = \max_{2\leq j\leq s}\left\{\left|p''_{\lambda_{jn}}(|\theta_{j,0}|)\right|\right\}$.

For the SCAD penalty function, the $p''_{\lambda_{jn}}(\cdot)$ exists at any non-zero point by choosing a proper λ_{jn}. Theorem 1 shows that there is a local minimizer of $J(\boldsymbol{\theta})$ which is a $\sqrt{n}$ consistent penalized quasi-likelihood estimator by choosing appropriate regularization parameter λ_{jn}.

Theorem 2. *Suppose that the assumptions in Appendix A hold, $\lim\limits_{n\to\infty}\lambda_{jn} = 0$, $\lim\limits_{n\to\infty}\sqrt{n}\lambda_{jn} = \infty$, $p_{\lambda_{jn}}(|\delta|)$ satisfies $\liminf\limits_{n\to\infty}\liminf\limits_{\delta\to 0^+} p'_{\lambda_{jn}}(\delta)/\lambda_{jn} > 0$. Then with probability approaching one, the $\sqrt{n}$ consistent local minimizer $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\theta}}_1^{\mathrm{T}}, \hat{\boldsymbol{\theta}}_2^{\mathrm{T}}\right)^{\mathrm{T}}$ in Theorem 1 must satisfy:*

(i) Sparsity: $\hat{\boldsymbol{\theta}}_2 = \mathbf{0}$;
(ii) Asymptotic normality:

$$\sqrt{n}\left\{(\boldsymbol{\Sigma}_{n1}(\boldsymbol{\theta}_{10}) + \boldsymbol{\Lambda})(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) + \boldsymbol{d}\right\} \xrightarrow{d} N\{\mathbf{0}, \boldsymbol{\Sigma}_1(\boldsymbol{\theta}_{10}) + \boldsymbol{\Omega}_1(\boldsymbol{\theta}_{10})\},$$

where $\boldsymbol{\Sigma}_{n1}(\boldsymbol{\theta}_{10})$, $\boldsymbol{\Sigma}_1(\boldsymbol{\theta}_{10})$, and $\boldsymbol{\Omega}_1(\boldsymbol{\theta}_{10})$ denote the first s upper-left submatrix of $\boldsymbol{\Sigma}_n(\boldsymbol{\theta}_0)$, $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = \lim\limits_{n\to\infty}\boldsymbol{\Sigma}_n(\boldsymbol{\theta}_0)$, and $\boldsymbol{\Omega}(\boldsymbol{\theta}_0) = \lim\limits_{n\to\infty}\boldsymbol{\Omega}_n(\boldsymbol{\theta}_0)$ respectively, and $\boldsymbol{\Sigma}_n(\boldsymbol{\theta}_0)$, $\boldsymbol{\Omega}_n(\boldsymbol{\theta}_0)$ are denoted in notations.

Theorem 2 shows that the proposed method can identify the SAR model ($\rho_1 \neq 0, \rho_2 = 0$), SEM ($\rho_1 = 0, \rho_2 \neq 0$), SARAR model ($\rho_1 \neq 0, \rho_2 \neq 0$), select explanatory variables and estimate unknown parameters simultaneously. Similar to the analysis of Fan and Li [13], if $\lambda_{jn} \to 0$ as $n \to \infty$, then $a_n \to 0$ for both SCAD and HARD thresholding penalty functions. Moreover, we obtain that $\boldsymbol{\Lambda} \to \mathbf{0}$ and $\boldsymbol{d} \to \mathbf{0}$ as $n \to \infty$. Thus, under regular conditions, the responding oracle property of the penalized quasi-likelihood estimators can be obtained. That is, the penalized quasi-likelihood estimators perform asymptotically as well as the ordinary quasi-likelihood estimators for nonzero parameters when knowing the correct submodel. However, for the LASSO penalty function, some conditions in Theorem 2 can not be satisfied.

3. Algorithm Design and Implementation

In this section, we consider the implementation of the proposed procedures. Since the penalized quasi-likelihood function $J(\boldsymbol{\theta})$ is nonconcave, it is challenging to get the global optimum solution. The study by Liu et al. [28] proposed: The existing algorithms, such as local quadratic approximation (LQA) algorithm [13] and local linear approximation (LLA) algorithm [31], can not be used directly to the SAR model. Similarly, those algorithms also do not give the correct minimizer of $J(\boldsymbol{\theta})$ for the SARAR model. Hence, we design the following iterative algorithm.

Initialization:

$$\boldsymbol{\theta}^{(0)} = \left(\sigma^{(0)}, \rho_1^{(0)}, \rho_2^{(0)}, \boldsymbol{\beta}^{(0)}\right). \tag{4}$$

Iteration:

$$\text{Find } \boldsymbol{\beta}^{(p+1)} \text{ by } \mathop{\arg\min}_{\boldsymbol{\beta} \in R^k} \left\{ l_1(\boldsymbol{\beta}) = J\left(\sigma^{(p)}, \rho_1^{(p)}, \rho_2^{(p)}, \boldsymbol{\beta}\right) \right\}, \tag{5}$$

$$\text{Find } \left(\rho_1^{(p+1)}, \rho_2^{(p+1)}\right) \text{ by } \mathop{\arg\min}_{\rho_1, \rho_2 \in (-1,1)} \left\{ l_2(\rho_1, \rho_2) = J\left(\sigma^{(p)}, \rho_1, \rho_2, \boldsymbol{\beta}^{(p+1)}\right) \right\}, \tag{6}$$

$$\text{Find } \sigma^{(p+1)} \text{ by } \mathop{\arg\min}_{\sigma \in (0,\infty)} \left\{ l_3(\sigma) = J\left(\sigma, \rho_1^{(p+1)}, \rho_2^{(p+1)}, \boldsymbol{\beta}^{(p+1)}\right) \right\}. \tag{7}$$

Iterate (5) to (7) until the successive value satisfies $||\hat{\boldsymbol{\theta}}^{(q+1)} - \hat{\boldsymbol{\theta}}^{(q)}|| < \varepsilon$, where $\hat{\boldsymbol{\theta}}^{(q)} = \left(\hat{\sigma}^{(q)}, \hat{\rho}_1^{(q)}, \hat{\rho}_2^{(q)}, \hat{\boldsymbol{\beta}}^{(q)^{\mathrm{T}}}\right)^{\mathrm{T}}$ and ε is a given tolerance value. In the following simulation, we let ε be 10^{-4}. Denote the final estimate of $(\sigma^2, \rho_1, \rho_2, \boldsymbol{\beta})$ as $\left(\hat{\sigma}^2, \hat{\rho}_1, \hat{\rho}_2, \hat{\boldsymbol{\beta}}\right)$, then $\hat{\boldsymbol{\theta}} = \left(\hat{\sigma}^2, \hat{\rho}_1, \hat{\rho}_2, \hat{\boldsymbol{\beta}}^{\mathrm{T}}\right)^{\mathrm{T}}$.

In (4), the initial value of $\boldsymbol{\theta}$ is the quasi-maximum likelihood estimate based on the log-quasi-likelihood function of the model (1). In (5), we note that if both autoregressive coefficients ρ_1 and ρ_1 are known in the SARAR model (1), then we can transform it as the following linear model $\boldsymbol{Y}_n^* = \boldsymbol{X}_n^* \boldsymbol{\beta} + \boldsymbol{E}_n$, where $\boldsymbol{Y}_n^* = \boldsymbol{S}_{2n}(\rho_2)\boldsymbol{S}_{1n}(\rho_1)\boldsymbol{Y}_n$, $\boldsymbol{X}_n^* = \boldsymbol{S}_{2n}(\rho_2)\boldsymbol{X}_n$. Therefore, the LQA algorithm can be used to complete this step as in the classical linear models. In (6), the optimization problem of bivariate functions can be solved by the Nelder–Mead method [32]. In (7), by using the partial derivative, the unique minimum point is:

$$\sigma^{(p+1)} = \frac{1}{n} \boldsymbol{E}_n^{\mathrm{T}}\left(\gamma^{(p+1)}\right) \boldsymbol{E}_n\left(\gamma^{(p+1)}\right),$$

where $\gamma^{(p+1)} = \left(\rho_1^{(p+1)}, \rho_2^{(p+1)}, \boldsymbol{\beta}^{(p+1)^{\mathrm{T}}}\right)^{\mathrm{T}}$. Figure 1 presents a flowchart of the proposed algorithm.

To implement the above algorithm, the tuning parameters need to be chosen. For the SCAD penalty function, we set $a = 3.7$ as recommended by Fan and Li [13]. Moreover, it is desirable to select a proper data-driven method to estimate all tuning parameters $\lambda_2, \cdots, \lambda_{k+3}$. Wang et al. [33] proved that the optimal tuning parameter in the SCAD penalty can be determined by BIC for the linear regression models. Thus, we can select $\boldsymbol{\lambda} = (\lambda_2, \cdots, \lambda_{k+3})^{\mathrm{T}}$ by the following Bayesian information criterion:

$$\mathrm{BIC}(\boldsymbol{\lambda}) = -2\ln L_n\big(\hat{\boldsymbol{\theta}}\big) + \alpha(\boldsymbol{\lambda})\log n,$$

where $\alpha(\boldsymbol{\lambda}) = \sum_{j=1}^{k+3} I\big(\hat{\theta}_j \neq 0\big)$. Then $\boldsymbol{\lambda}$ is set to be $\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}}\{\mathrm{BIC}(\boldsymbol{\lambda})\}$.

In fact, minimizing the BIC over a $k+2$-dimensional space is an unduly onerous task for a large k. To save computation time, one may use the same tuning parameter for all penalty functions. However, the experiments, though not given for saving space, show that the spatial regression coefficient ρ_2 is easy to compress to 0 even if the sample size is medium. Intuitively, we should use different tuning parameters for spatial regression coefficients ρ_j $(j = 1, 2)$ and regression coefficients β_j $(j = 1, \cdots, k)$ because the range of ρ_j $(j = 1, 2)$ are known before estimation, but the range of β_j $(j = 1, \cdots, k)$ are not. Thus, we set $\lambda_2 = \lambda_3$, and $\lambda_4 = \cdots = \lambda_{k+3}$ to optimize the results. It should be pointed out that we can prove the consistency of the BIC criterion under more stringent conditions, such as the bounded derivative of the quasi-likelihood function and $\alpha(\lambda)$. However, it is very difficult to prove the consistency under some mild conditions and will be left for further study.

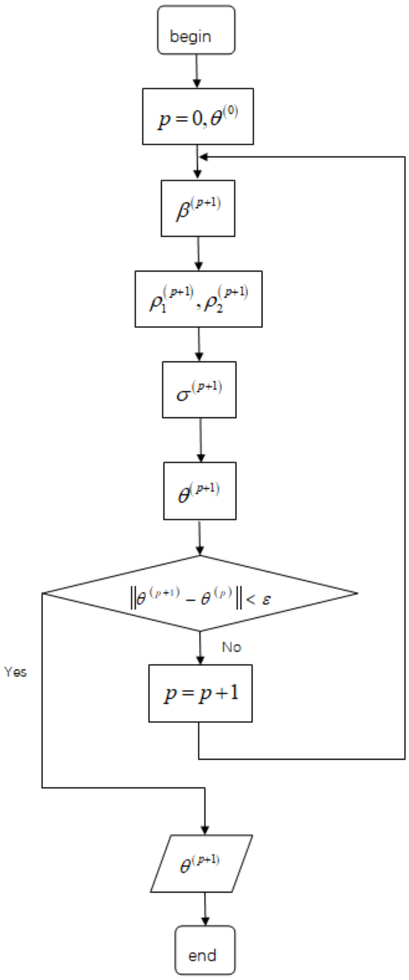

Figure 1. The algorithm flowchart.

4. Numerical Simulation

In this section, we conduct some Monte Carlo experiments to evaluate the finite sample performance of the proposed variable selection method in the SARAR model using R codes.

4.1. Simulation Sampling

The sample data is generated by model (1). We consider eight explanatory variables following an 8-dimensional normal distribution with zero mean and covariance matrix (σ_{ij}), where $\sigma_{ij} = 0.5^{|i-j|}$. The spatial autoregressive coefficients are set to be $(\rho_1, \rho_2) = (0.7, 0.7), (0.7, 0.3), (0.7, 0), (0, 0.7)$, and $(0, 0)$. For simplicity, let $\boldsymbol{W}_{1n} = \boldsymbol{W}_{2n} = \boldsymbol{I}_R \otimes \boldsymbol{B}_m$, where $\boldsymbol{B}_m = (1/(m-1))(\boldsymbol{l}_m \boldsymbol{l}_m^{\mathrm{T}} - \boldsymbol{I}_m)$, $\otimes$ is the Kronecker product, and $\boldsymbol{l}_m$ is an m-dimensional column vector of ones [3,34], which is called the Case spatial weight matrix. To observe the influence of different spatial weight matrices, the Rook spatial weight matrix is introduced, in which w_{ij} is set to be 1 when the regions share a common boundary and set to be 0 for other cases. For the Case spatial weight matrix, we take $m = 3$ and different values of R, where $R = 10, 20, 60$, then corresponding sample sizes are $n = 30, 60, 180$. For the Rook spatial weight matrix, we use the grid square area to generate it according to whether the edges are adjacent. To ensure that the region is square, the value of n is the square of the integer value and $n = 36, 64, 196$. The regression coefficients are assumed to be $\boldsymbol{\beta} = (3, 2, 0, 0, 1, 0, 0, 0)^{\mathrm{T}}$. The innovation e_i follows a normal distribution with mean 0 and variance $\sigma^2 = 1, 1.5$.

4.2. Simulation Results

For each case, we do 100 repetitions. The average number of zero coefficients which are correctly identified is denoted as "C". The label "I" indicates the average number of non-zero coefficients incorrectly shrunk to zero. To measure the estimation accuracy of $\boldsymbol{\theta}$, we compare the estimation accuracy using the medians of squared error (SE) as in Liang and Li [10], which is defined as:

$$\mathrm{SE} = \left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|^2 = \sum_{i=1}^{k+3} \left(\hat{\theta}_i - \theta_{i,0}\right)^2,$$

where $\hat{\boldsymbol{\theta}}_n = \left(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_{k+3}\right)^{\mathrm{T}}$ is the estimate of $\boldsymbol{\theta}_0$. In Tables 1–3, Oracle implies the results of variable selection knowing zero parameters. Moreover, other penalty functions, such as HARD and LASSO, are introduced in the penalized quasi-likelihood function for comparison.

Tables 1–3 clearly show that there are similar performances for variable selection under both different spatial weight matrices. In other words, the proposed method is not sensitive to the change of the spatial weight matrix. As we expected, all penalty functions can reduce their mSE (the median of SE) and give close results of Oracle with the increase of sample size. In most cases, the SCAD penalty produces the lowest mSE, the HARD penalty has a little bigger than the SCAD penalty, and the LASSO penalty produces the largest mSE. Moreover, if there are spatial effects for both the spatial lag term and the spatial error lag term ($\rho_1 \neq 0$ and $\rho_2 \neq 0$), the mSE is often relatively large; if only one of the spatial lag term and the spatial error lag term is related to the spatial effect ($\rho_1 \neq 0, \rho_2 = 0$, and $\rho_1 = 0, \rho_2 \neq 0$), the value of the mSE is usually smaller, especially when there is no spatial effect ($\rho_1 = 0$ and $\rho_2 = 0$). However, like most of the existing results of variable selection, the mSE will be less accurate in all cases if the variance σ^2 of the innovation becomes large. In terms of C and I, we can see that the average number of correctly identifying zero-valued coefficients approaches the true value and the average number of incorrectly identifying zero-valued coefficients approaches 0 as the sample size n increases. These simulation results accord with the theoretical analysis. The SCAD and HARD penalties have good performance about C, there is little difference between them in most cases. They can converge rapidly to the real number of 0 except a LASSO penalty with a low convergence rate, which may imply that both SCAD and HARD tend to give smaller models than LASSO. In the case of small samples, the LASSO penalty has the lowest value of the I in most cases. However, their differences quickly disappear in large samples for all penalties. These results are similar to those obtained by Fan and Li [13]. It is worth noting that when ρ_2 is small, it is easy to compress to 0, and then produce a larger error rate I in the setting of small samples.

Table 1. Simulation results of variable selection with Case spatial weight matrix.

$\sigma^2 = 1$	*n* = 30			*n* = 60			*n* = 180		
Method	**C**	**I**	**mSE**	**C**	**I**	**mSE**	**C**	**I**	**mSE**
$\rho_1 = 0.7$ $\rho_2 = 0.7$									
SCAD	3.8300	0.1100	0.4329	4.6000	0.0100	0.0941	5.0000	0.0000	0.0272
Hard	4.2900	0.1100	0.4961	4.6200	0.0100	0.1048	4.9700	0.0000	0.0276
LASSO	2.7300	0.1600	0.7452	3.0400	0.0600	0.2209	3.4800	0.0000	0.0635
Oracle	5.0000	0.0000	0.0909	5.0000	0.0000	0.0300	5.0000	0.0000	0.0090
$\rho_1 = 0.7$ $\rho_2 = 0.3$									
SCAD	3.8900	0.2500	0.4627	4.7300	0.0500	0.1218	5.0000	0.0000	0.0284
Hard	4.2800	0.2900	0.4861	4.6800	0.0700	0.1208	4.9100	0.0000	0.0332
LASSO	2.7800	0.3200	0.5947	3.3700	0.0600	0.1729	3.7800	0.0000	0.0546
Oracle	5.0000	0.0000	0.1111	5.0000	0.0000	0.0348	5.0000	0.0000	0.0100
$\rho_1 = 0.7$ $\rho_2 = 0.0$									
SCAD	5.1500	0.0200	0.3423	5.7000	0.0000	0.0983	5.9600	0.0000	0.0273
Hard	5.4000	0.0200	0.4259	5.5700	0.0000	0.1088	5.9400	0.0000	0.0276
LASSO	4.1900	0.0100	0.4503	4.3800	0.0000	0.1585	4.6000	0.0000	0.0598
Oracle	6.0000	0.0000	0.0461	6.0000	0.0000	0.0175	6.0000	0.0000	0.0045
$\rho_1 = 0.0$ $\rho_2 = 0.7$									
SCAD	5.1000	0.1600	0.4964	5.6800	0.0200	0.1268	5.9100	0.0000	0.0206
Hard	5.1500	0.1700	0.4975	5.7000	0.0300	0.1288	5.9300	0.0000	0.0271
LASSO	4.0000	0.1000	0.5633	4.1600	0.0100	0.1472	4.7300	0.0000	0.0487
Oracle	6.0000	0.0000	0.0761	6.0000	0.0000	0.0260	6.0000	0.0000	0.0071
$\rho_1 = 0.0$ $\rho_2 = 0.0$									
SCAD	5.9800	0.0200	0.3283	6.4600	0.0000	0.0951	6.9200	0.0000	0.0264
Hard	6.3700	0.0200	0.3164	6.6300	0.0000	0.1010	6.9400	0.0000	0.0265
LASSO	4.8200	0.0000	0.4337	4.9600	0.0000	0.1572	5.2500	0.0000	0.0561
Oracle	7.0000	0.0000	0.0340	7.0000	0.0000	0.0163	7.0000	0.0000	0.0040

Table 4 shows the results of ignoring spatial effects by the LQA algorithm [13] under the same context as in Table 1. In terms of I, when there are two spatial effects ($\rho_1 \neq 0$ and $\rho_2 \neq 0$), the number of incorrect zero in Table 1 is much lower than those in Table 4. When only one spatial effect exists ($\rho_1 \neq 0, \rho_2 = 0$, or $\rho_1 = 0, \rho_2 \neq 0$), the number of incorrect zero in Table 4 decreases slightly compared to the first case and is also larger than that in Table 1. When there is no spatial effect ($\rho_1 = 0$ and $\rho_2 = 0$), the results of our algorithm are close to that of the LQA algorithm. Meanwhile, turning attention to the C, our algorithm can identify more true zeros than the LQA algorithm as long as the spatial effect exists ($\rho_1 \neq 0$). Although we are surprised to find that the value of C and I under the LQA algorithm seem to be getting close to the correct values with slow speed as the sample size increases for the SCAD and HARD penalties, the mSE reflected the estimation errors of their parameters are large and outrageous. This is in line with our intuition: Ignoring both spatial effects, the LQA algorithm is implemented on the wrong model and easily leads to a large estimated deviation. Moreover, the LQA algorithm is affected by the initial estimation. In simulation, the initial estimation is the quasi-maximum likelihood estimation, which is equal to the least square estimation (including the observation value of dependent variable $\boldsymbol{Y}_n$). If the strong spatial effect about ρ_1 is ignored, the observation value of the dependent variable will deviate from the requirement of unbiased estimation seriously, which will lead to a great deviation of the initial estimation. With the influence of iteration, the accumulated error of final estimation will be extraordinary. However, when the spatial effects disappear, we can see that both algorithms have similar good performances, which indicates that no matter whether there are spatial effects, the proposed algorithm still has a satisfactory performance in a finite sample.

Table 2. Simulation results of variable selection with the Case spatial weight matrix.

$\sigma^2 = 1.5$	*n* = 30			*n* = 60			*n* = 180		
Method	**C**	**I**	**mSE**	**C**	**I**	**mSE**	**C**	**I**	**mSE**
$\rho_1 = 0.7$									
$\rho_2 = 0.7$									
SCAD	3.9500	0.1300	0.8425	4.7100	0.0200	0.1536	5.0000	0.0000	0.0461
HARD	4.3100	0.1300	0.8495	4.6200	0.0200	0.1666	4.9600	0.0000	0.0471
LASSO	2.9100	0.3400	1.6368	2.9900	0.1400	0.4662	3.5100	0.0000	0.1328
Oracle	5.0000	0.0000	0.1838	5.0000	0.0000	0.0548	5.0000	0.0000	0.0168
$\rho_1 = 0.7$									
$\rho_2 = 0.3$									
SCAD	4.0800	0.4000	0.8675	4.7300	0.0400	0.1829	5.0000	0.0000	0.0462
HARD	4.2200	0.3800	0.8529	4.7100	0.0800	0.1902	4.9000	0.0000	0.0536
LASSO	2.6700	0.2700	0.9862	3.2100	0.0700	0.2822	3.9700	0.0000	0.0987
Oracle	5.0000	0.0000	0.1709	5.0000	0.0000	0.0666	5.0000	0.0000	0.0188
$\rho_1 = 0.7$									
$\rho_2 = 0.0$									
SCAD	5.1700	0.0700	0.7204	5.6700	0.0200	0.1654	5.9700	0.0000	0.0462
HARD	5.2900	0.0900	0.8144	5.6000	0.0300	0.1816	5.9300	0.0000	0.0491
LASSO	4.1300	0.0200	0.7540	4.2700	0.0000	0.2520	4.7800	0.0000	0.1036
Oracle	6.0000	0.0000	0.1110	6.0000	0.0000	0.0394	6.0000	0.0000	0.0103
$\rho_1 = 0.0$									
$\rho_2 = 0.7$									
SCAD	4.9600	0.2000	0.7713	5.6900	0.0300	0.1865	5.9000	0.0000	0.0408
HARD	4.9900	0.2100	0.8786	5.6800	0.0400	0.1888	5.9200	0.0000	0.0460
LASSO	3.5500	0.0900	0.8639	4.1300	0.0100	0.2551	4.9900	0.0000	0.0862
Oracle	6.0000	0.0000	0.1497	6.0000	0.0000	0.0541	6.0000	0.0000	0.0140
$\rho_1 = 0.0$									
$\rho_2 = 0.0$									
SCAD	5.9700	0.0700	0.6812	6.5000	0.0200	0.1697	6.9100	0.0000	0.0446
HARD	6.3000	0.0700	0.6552	6.6100	0.0000	0.1714	6.9300	0.0000	0.0449
LASSO	4.8200	0.0200	0.6933	4.9200	0.0000	0.2337	5.5200	0.0000	0.1020
Oracle	7.0000	0.0000	0.0765	7.0000	0.0000	0.0367	7.0000	0.0000	0.0100

Considering the complexity of the asymptotic covariance matrix of $\boldsymbol{\theta}$, we use the traditional bootstrap method in which the sample size of the resampled observations is 100 to obtain the standard deviations of parameter estimates. The parameter vector $\boldsymbol{\theta}$ is estimated by our algorithm. SD indicates the median absolute deviation of 100 estimated coefficients in the 100 simulations, which can be regarded as an estimate of the true standard deviation of $\boldsymbol{\theta}$. Using the bootstrap, we calculate a median of estimated standard deviations, denoted as SDm, and estimate its standard deviation by median absolute deviation, denoted as SDmad.

Table 5 provides the numerical simulation results of nonzero coefficients under $\rho_1 = 0.7, \rho_2 = 0.3$, $\sigma^2 = 1$, $n = 30$, and $n = 60$ with the Case spatial weight matrix. The simulation results show that the bootstrap estimated standard deviation becomes increasingly accurate when sample size n increases. In most cases, the SD, SDm, and SDmad obtained by the SCAD and HARD penalties are smaller than that obtained by the LASSO penalty, which shows that the LASSO penalty does not appear to be as stable as the SCAD and HARD penalties. Furthermore, when the σ^2 increases and is away from 1, the estimation of the standard deviation will be less accurate although the results are not presented. In one world, the LASSO penalty generally lags behind the SCAD and HARD penalties concerning the accuracy of estimates. For saving space, the other cases, such as $\rho_1 = 0.7, \rho_2 = 0.7$, or $\sigma^2 = 1.5$, have similar results and are omitted.

Table 3. Simulation results of variable selection with the Rook spatial weight matrix.

$\sigma^2 = 1$	n = 36			n = 64			n = 196		
Method	**C**	**I**	**mSE**	**C**	**I**	**mSE**	**C**	**I**	**mSE**
$\rho_1 = 0.7$									
$\rho_2 = 0.7$									
SCAD	4.4300	0.1000	0.2802	4.6200	0.0000	0.1142	4.9900	0.0000	0.0290
HARD	4.5600	0.1200	0.2998	4.6200	0.0200	0.1316	4.9600	0.0000	0.0343
LASSO	3.2300	0.1000	0.4090	3.3500	0.0100	0.1982	3.8400	0.0000	0.0573
Oracle	5.0000	0.0000	0.2415	5.0000	0.0000	0.1121	5.0000	0.0000	0.0206
$\rho_1 = 0.7$									
$\rho_2 = 0.3$									
SCAD	4.3400	0.2400	0.2987	4.5500	0.1100	0.1290	4.9800	0.0000	0.0370
HARD	4.3900	0.2800	0.3229	4.5900	0.1500	0.1495	4.9600	0.0000	0.0375
LASSO	3.2800	0.2700	0.3923	3.2900	0.1200	0.1935	3.8600	0.0000	0.0591
Oracle	5.0000	0.0000	0.2584	5.0000	0.0000	0.1254	5.0000	0.0000	0.0368
$\rho_1 = 0.7$									
$\rho_2 = 0.0$									
SCAD	5.5100	0.0000	0.2082	5.6800	0.0000	0.0939	5.9800	0.0000	0.0264
HARD	5.5900	0.0000	0.2430	5.6700	0.0000	0.0959	5.9600	0.0000	0.0264
LASSO	4.3200	0.0000	0.3124	4.3500	0.0000	0.1748	4.8500	0.0000	0.0489
Oracle	6.0000	0.0000	0.0293	6.0000	0.0000	0.0173	6.0000	0.0000	0.0053
$\rho_1 = 0.0$									
$\rho_2 = 0.7$									
SCAD	5.1200	0.1100	0.3238	5.5200	0.0000	0.1168	5.9000	0.0000	0.0300
HARD	5.2000	0.1600	0.3712	5.5300	0.0100	0.1240	5.9200	0.0000	0.0307
LASSO	4.0100	0.1700	0.4305	4.3500	0.0200	0.1922	4.7100	0.0000	0.0717
Oracle	6.0000	0.0000	0.0616	6.0000	0.0000	0.0396	6.0000	0.0000	0.0121
$\rho_1 = 0.0$									
$\rho_2 = 0.0$									
SCAD	6.1900	0.0000	0.1688	6.5600	0.0000	0.0955	6.9300	0.0000	0.0242
HARD	6.5300	0.0100	0.1852	6.5600	0.0000	0.1106	6.8800	0.0000	0.0259
LASSO	5.0700	0.0000	0.3146	5.2900	0.0000	0.1681	5.7100	0.0000	0.0435
Oracle	7.0000	0.0000	0.0226	7.0000	0.0000	0.0140	7.0000	0.0000	0.0044

Table 4. Simulation results of variable selection when we ignore spatial effects.

$\sigma^2 = 1$	$n = 30$			$n = 60$			$n = 180$		
Method	**C**	**I**	**mSE**	**C**	**I**	**mSE**	**C**	**I**	**mSE**
$\rho_1 = 0.7$									
$\rho_2 = 0.7$									
SCAD	4.1500	1.1900	2894.0	4.5000	0.9900	4159.8	4.7800	0.4500	5025.6
HARD	1.9300	0.3700	2153.9	2.7700	0.4200	3697.3	4.1700	0.2300	4870.4
LASSO	0.2500	0.1500	2110.4	0.0000	0.0000	3449.5	0.0000	0.0000	4762.6
$\rho_1 = 0.7$									
$\rho_2 = 0.3$									
SCAD	4.2300	0.5700	73.698	4.4700	0.3400	105.62	4.7800	0.0400	122.92
HARD	3.9200	0.4300	70.815	4.3700	0.3400	101.09	4.7400	0.0500	122.92
LASSO	1.7600	0.2100	79.049	0.4900	0.1100	99.844	0.0000	0.0000	117.50
$\rho_1 = 0.7$									
$\rho_2 = 0.0$									
SCAD	4.2600	0.4400	39.324	4.5500	0.1900	51.2130	4.7700	0.0100	52.666
HARD	4.0500	0.4100	37.080	4.5400	0.2000	50.641	4.8600	0.0100	52.667
LASSO	1.9700	0.1900	40.984	1.1200	0.0600	48.713	0.0000	0.0000	49.750
$\rho_1 = 0.0$									
$\rho_2 = 0.7$									
SCAD	3.9200	0.1000	0.7908	4.6600	0.0000	0.5977	4.8600	0.0000	0.5517
HARD	4.3100	0.0700	0.8395	4.7200	0.0000	0.6243	4.8900	0.0000	0.5618
LASSO	2.4900	0.0200	0.7809	2.8300	0.0000	0.7512	3.1800	0.0000	0.6659
$\rho_1 = 0.0$									
$\rho_2 = 0.0$									
SCAD	3.5500	0.0200	0.3353	4.6600	0.0000	0.1115	4.9900	0.0000	0.0260
HARD	4.1400	0.0200	0.4382	4.6600	0.0000	0.1125	4.9000	0.0000	0.0276
LASSO	2.4900	0.0000	0.4980	2.8600	0.0000	0.1672	3.3100	0.0000	0.0533

Table 5. Standard deviations of estimates of the nonzero regression coefficients.

Method	$n = 30$			$n = 60$		
	SD	**SDm**	**SDmad**	**SD**	**SDm**	**SDmad**
SCAD						
σ^2	0.2006	0.1634	0.0325	0.1091	0.1282	0.0154
ρ_1	0.0226	0.0227	0.0032	0.0169	0.0131	0.0012
ρ_2	0.1106	0.1851	0.0236	0.0565	0.0895	0.0078
β_1	0.1225	0.1299	0.0152	0.1110	0.0886	0.0079
β_2	0.1495	0.1374	0.0172	0.1161	0.0890	0.0085
β_5	0.2074	0.1466	0.0205	0.1111	0.0871	0.0092
HARD						
σ^2	0.1788	0.1498	0.0325	0.1022	0.1251	0.0157
ρ_1	0.0222	0.0233	0.0031	0.0165	0.0130	0.0012
ρ_2	0.0850	0.1156	0.0134	0.0542	0.0684	0.0062
β_1	0.1287	0.1297	0.0172	0.1105	0.0874	0.0097
β_2	0.1816	0.1342	0.0170	0.1157	0.0877	0.0081
β_5	0.1930	0.1340	0.0207	0.1114	0.0867	0.0084
LASSO						
σ^2	0.2104	0.1609	0.0350	0.1236	0.1354	0.0173
ρ_1	0.0257	0.0252	0.0027	0.0190	0.0137	0.0013
ρ_2	0.1242	0.1782	0.0227	0.0686	0.0895	0.0063
β_1	0.1830	0.1532	0.0246	0.0974	0.0994	0.0108
β_2	0.1954	0.1570	0.0264	0.1135	0.0961	0.0094
β_5	0.1921	0.1485	0.0265	0.1161	0.0943	0.0107

5. Data Example

Now, we consider a real example for the application and performance of the proposed variable selection method in the SARAR model.

5.1. The Sample Data

We consider the Boston housing data set which was originally given by Harrison and Rubinfeld [35] and has been used by many authors, for example, Pace and Gilley [36,37], and so on. The data set contains 506 census tracts with 14 nonconstant independent variables. It can be found in the spdep library of R. Similar to the analysis of Harrison and Rubinfeld [35], the dependent variable is set to be $\log(\text{MEDV})$ and the explanatory variables are assumed as RM^2, AGE, $\log(\text{DIS})$, $\log(\text{RAD})$, TAX, PTRATIO, $(B - 0.63)^2$, $\log(\text{LSTAT})$, CRIM, ZN, INDUS, CHAS, and NOX^2. Table 6 gives the interpretation of all abbreviated variables. For subsequent analysis, the data are centralized and standardized. The spatial weight matrix is constructed with rook contiguity: The weight is 1 if two different areas share a common boundary, and 0 otherwise. Then the matrix is row-normalized as is usually carried out in practice.

Table 6. Variables used in the analysis.

Variable	**Description**
MEDV	The median value of owner-occupied homes. Source: 1970 U.S. Census.
CRIM	Crime rate by town. Source: FBI (1970).
ZN	Proportion of a town's residential land zoned for lots greater than 25,000 square feet. Source: Metropolitan Area Planning Commission (1972).
INDUS	Proportion nonretail business acres per town. Source: Harrison and Rubinfeld (1978).
CHAS	Charles River dummy: =1 if tract bounds the Charles River; =0 if otherwise. Source: 1970 U.S. Census.
NOX	Nitrogen oxide concentrations in pphm (annual average concentration in parts per hundred million). Source: TASSIM.
RM	Average number of rooms in owner units. Source: 1970 U.S. Census.
AGE	Proportion of owner units built prior to 1940. Source: 1970 U.S. Census.

Table 6. *Cont.*

Variable	Description
DIS	Weighted distances to five employment centres in the Boston region. Source: Harrison and Rubinfeld (1978).
RAD	Index of accessibility to radial highways. It was calculated on a town basis. Source: MIT Boston Project.
TAX	Full value property tax rate ($/$10,000). Source: Massachusetts Taxpayers Foundation (1970).
PTRATIO	The number of students divided by the number of teachers in town school district. Source: Massachusetts Dept. of Education (1971–1972).
B	Black proportion of population. Source: 1970 U.S. Census.
PART	Proportion of population that is lower status = $\frac{1}{2}$ (proportion of adults without some high school education and proportion of male workers classified as laborers). Source: 1970 U.S. Census.

5.2. Spatial Dependence Test

In spatial data analysis, the Moran's I statistic (Moran I) is usually used to test spatial dependence. Table 7 shows the value of the Moran's I in the Boston housing data. It is 0.7644 with a p-value 2.2×10^{-16}, which implies that the MEDV has a strong spatial correlation. It is well known that the Moran's I reflects the degree of spatial autocorrelation and can not effectively identify specific spatial autoregressive models due to the existence of different spatial correlations. Fortunately, the popular Lagrange multiplier diagnostics can help us to complete this specification for several different spatial autoregressive models. This test method avoids the optimization of the nonlinear function and is easy to implement. Using the spdep package in R, we can obtain the desired results for identification. From Table 7, it is obvious to see that the p-value in each case is very small, which implies that the Boston housing data can be modeled by spatial models. However, the values of test statistics and p-values suggest that the SARAR model is the best choice among these spatial models to fit the Boston housing data. Moreover, previous studies have used multiple hypothesis tests to judge spatial effects and select explanatory variables, and then determine the model. It is difficult to prove the relevant theoretical properties. Based on the proposed variable selection method, the SARAR model can not only be used to identify different spatial effects and select explanatory variables simultaneously, but also has a good theoretical guarantee. Therefore, we will use the SARAR model for variable selection in this data.

Table 7. Moran's I test and Lagrange multiplier diagnostics for spatial dependence.

Terms	Values of Test Statistics	p-Values
Moran I	0.7644	$<2.2e{-16}$
LMerr	186.57	$<2.2e{-16}$
LMlag	190.71	$<2.2e{-16}$
SARMA	228.32	$<2.2e{-16}$

Note: LMerr represents the test results of the SEM; LMlag represents the test results of the SAR model; and SARMA represents the test results of the SARAR model.

5.3. Model Selection and Estimation

Under a SARAR model, the results are reported in Table 8, where the quasi-maximum likelihood estimate (QMLE) and penalized quasi-likelihood estimate (PQLE) via the SCAD, HARD, and LASSO penalties are listed to assess the performance of variable selection.

The QMLE demonstrates that there are four variables that show a relatively small impact on the MEDV, including ZN, INDUS, CHAS, and AGE. These variables in other studies also show a small effect on the MEDV, such as Harrison and Rubinfeld [35], Pace and Gilley [36], and so on. Moreover, variables with positive effects include ZN, INDUS, RM^2, $\log(\text{RAD})$, $(B - 0.63)^2$, while others have negative effects. As we expected, the parameter estimates obtained by the penalized method are close to the QMLE, and both nonzero estimates keep the same sign. Moreover, the four insignificant variables (ZN, INDUS, CHAS, and AGE) are penalized to zero under different penalty functions. Therefore, these penalties produce the same selection results in this setting. However, BIC in Table 8 shows that the SCAD and HARD penalties are preferable to the LASSO penalty. Interestingly, although the spatial correlation coefficients ρ_1 and ρ_2 are also penalized by different penalty functions, they do not shrink to zero and have similar results with the QMLE. From the perspective of model specification, we can say that the penalty method recognizes the spatial autoregressive relationship.

For comparison, the Boston housing data is also fitted by a classical linear regression model and the related results are presented in Table 9. The QMLE shows that there are three unimportant variables, including ZN, INDUS, and AGE. Moreover, variables with positive effects include ZN, INDUS, CHAS, RM^2, AGE, $\log(\text{RAD})$, $(B-0.63)^2$, while others have negative effects. In addition, all penalties also produce the same selection results in this model. According to the QMLE, these penalties can also select important variables and shrink unimportant variables to zero. Based on the BIC, the SCAD and HARD penalties also outperform the LASSO penalty in this setting.

Although both models have similar selection results, the differences between them are quite obvious. For the QMLE, the estimated coefficient of AGE is negative in the SARAR model, a plausible result, but it is positive in the linear model, which seems implausible. For the PQLE, it is easy to see that the CHAS disappears in the SARAR model while it is relatively important in the linear model. Furthermore, the meaning of the parameter estimation in these two models is also distinctly different. The interpretation of parameter estimates in the SARAR model will become richer and more complicated than that in the linear model because of the spatial autocorrelation [38]. As we expected, the BIC for the SARAR model is far less than that for the classical linear model, which indicates that the SARAR model has a better fitting effect than the classical linear model in such data.

Table 8. Parameter estimates using quasi-maximum likelihood and penalized estimates via SCAD, HARD, and LASSO under a SARAR model.

Terms	QMLE	SCAD	HARD	LASSO
CRIM	−0.1405	−0.1240	−0.1410	−0.1346
ZN	0.0221	–	–	–
INDUS	0.0280	–	–	–
CHAS	−0.0058	–	–	–
NOX^2	−0.1037	−0.0223	−0.1046	−0.0560
RM^2	0.1721	0.1657	0.1643	0.1624
AGE	−0.0372	–	–	–
log(DIS)	−0.2082	−0.1127	−0.1851	−0.1415
log(RAD)	0.1750	0.1124	0.1680	0.1039
TAX	−0.1958	−0.1583	−0.1757	−0.1137
PTRATIO	−0.1030	−0.0816	−0.1071	−0.0830
(B−0.63)^2	0.0865	0.0713	0.0827	0.0652
log(LSTAT)	−0.3998	−0.4317	−0.4167	−0.3900
ρ_1	0.2805	0.2695	0.2776	0.3691
ρ_2	0.4145	0.4444	0.4107	0.2430
σ^2	0.1182	0.1197	0.1190	0.1230
BIC	489.81	474.18	467.36	477.00

Table 9. Parameter estimates using quasi-maximum likelihood and penalized estimates via SCAD, HARD, and LASSO under a classical linear model.

Terms	QMLE	SCAD	HARD	LASSO
CRIM	−0.2537	−0.2541	−0.2539	−0.2420
ZN	0.0047	–	–	–
INDUS	0.0051	–	–	–
CHAS	0.0573	0.0565	0.0578	0.0554
NOX^2	−0.2178	−0.2157	−0.2158	−0.1852
RM^2	0.1367	0.1383	0.1383	0.1418
AGE	0.0085	–	–	–
log(DIS)	−0.2529	−0.2570	−0.2567	−0.2211
log(RAD)	0.2035	0.2013	0.2011	0.1569
TAX	−0.1744	−0.1708	−0.1704	−0.1373
PTRATIO	−0.1663	−0.1667	−0.1666	−0.1575
$(B-0.63)^2$	0.0692	0.0689	0.0693	0.0665
log(LSTAT)	−0.5496	−0.5467	−0.5464	−0.5440
σ^2	0.1951	0.1952	0.1952	0.1961
BIC	696.27	677.68	677.67	680.24

6. Summary and Discussion

In theory, the proposed penalized quasi-likelihood method can identify two kinds of spatial effects, select significant explanatory variables, and estimate unknown parameters simultaneously. The penalized estimators has consistency, sparsity, and normality, which show that the penalty estimation of the coefficient of the significant variable with an unknown zero coefficient is as good as that of the significant variable with a known zero coefficient. In application, the proposed method is consistent with the theoretical results, which can effectively penalize the coefficients of insignificant variables to zero, identify the appropriate spatial regression model, and improve the interpretability of the results due to the decrease of the variable dimension.

From the analysis results of theory and application, it can be seen that the proposed method can effectively achieve a variable selection and identify spatial effects. At the same time, due to the complexity and time consumption of high-dimensional matrix inverse operation, we also find that the optimization efficiency of the penalty quasi- likelihood function still has room for further improvement. Therefore, this method is suitable for the case of a medium sample size and variable dimension not exceeding the sample size. When the sample size is large enough, the penalty GMM method can be considered to improve the operation speed. Once the dimension of the variable exceeds the sample size, our proposed method will not be applicable. Even so, the proposed method can also be used as a basis for future research, such as a new feature selection in spatial data.

In conclusion, it is significant to extend this model to other high dimensional parameter regression models, such as spatial Durbin models, dynamic panel data models, or super high dimensional nonparametric spatial regression models or semi-parametric spatial regression models, such as varying-coefficient spatial regression models, single index spatial regression models, additive spatial regression models, etc. These contents are optional for further research.

Author Contributions: Conceptualization, J.C.; methodology, X.L.; software, X.L.; validation, J.C.; formal analysis, J.C.; investigation, J.C.; resources, J.C.; data curation, X.L.; writing—original draft preparation, X.L.; writing—review and editing, J.C.; visualization, J.C.; supervision, J.C.; project administration, J.C.; funding acquisition, J.C. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the NSF of Fujian Province, China (2020J01170), Fujian Normal University Innovation Team Foundation, China (IRTL1704), Key Science and Technology Projects of Jiangxi Provincial Department of Education, China (GJJ202603), and Nanchang Normal University PhD Research Foundation, China (NSBSJJ2020006).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Assumptions

The following regular conditions are needed for the large sample properties of the penalized quasi-likelihood estimator.

Assumption A1. *The $\{e_i\}, i = 1, \cdots, n$, are independent identically distributed with $E(e_i) = 0$ and $var(e_i) = \sigma^2$. The moment $E\left(|e_1|^{4+v}\right)$ exists for a $v > 0$.*

Assumption A2. *The elements $w_{1n,ij} = O(1/h_n), w_{1n,ii} = 0$ in $\boldsymbol{W}_{1n}$, $w_{2n,ij} = O(1/h_n), w_{2n,ii} = 0$ in $\boldsymbol{W}_{2n}$, where $i, j = 1, 2, \cdots, n$, and $h_n/n \to 0$ as $n \to \infty$.*

Assumption A3. *The matrix $\boldsymbol{S}_{1n}$ and $\boldsymbol{S}_{2n}$ are nonsingular.*

Assumption A4. *The sequences of matrices $\{\boldsymbol{W}_{1n}\}$, $\{\boldsymbol{W}_{2n}\}$, $\left\{\boldsymbol{S}_{1n}^{-1}\right\}$, and $\left\{\boldsymbol{S}_{2n}^{-1}\right\}$ are uniformly bounded in both row and column sums [39].*

Assumption A5. *The $\lim\limits_{n\to\infty} n^{-1}\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{X}_n$ exists and is nonsingular. The elements of $\boldsymbol{X}_n$ are uniformly bounded constants for all n.*

Assumption A6. *The row and column sums of $\left\{\boldsymbol{S}_{in}^{-1}(\rho_i)\right\}$ are uniformly bounded, uniformly in ρ_i in a closed subset Λ of $(-1, 1)$ and the true ρ_{i0} is an interior point of Λ, $i = 1, 2$.*

Assumption A7. *As $n \to \infty$, $n^{-1}(\boldsymbol{X}_n, \boldsymbol{G}_{1n}\boldsymbol{X}_n\boldsymbol{\beta}_0)^{\mathrm{T}}(\boldsymbol{X}_n, \boldsymbol{G}_{1n}\boldsymbol{X}_n\boldsymbol{\beta}_0)$ and $n^{-1}(\boldsymbol{X}_n, \boldsymbol{G}_{2n}\boldsymbol{X}_n\boldsymbol{\beta}_0)^{\mathrm{T}}$ $(\boldsymbol{X}_n, \boldsymbol{G}_{2n}\boldsymbol{X}_n\boldsymbol{\beta}_0)$ exist and are nonsingular.*

Assumption A8. *The $\lim\limits_{n\to\infty} \boldsymbol{\Sigma}_n(\boldsymbol{\theta}_0)$ and $\lim\limits_{n\to\infty} \boldsymbol{\Omega}_n(\boldsymbol{\theta}_0)$ exist.*

Assumption A9. *The third derivatives $\left(\partial^3 L_n(\boldsymbol{\theta})\right)/\left(\partial\theta_j\partial\theta_l\partial\theta_m\right)$ exist for all $\boldsymbol{\theta}$ in an open set Θ that contains the true parameter point $\boldsymbol{\theta}_0$. Furthermore, there are functions M_{jlm} such that $\left|n^{-1}\left(\partial^3 \ln L_n(\boldsymbol{\theta})\right)/\left(\partial\theta_j\partial\theta_l\partial\theta_m\right)\right| \leq M_{jlm}$ for all $\boldsymbol{\theta} \in \Theta$, where $E\left(M_{jlm}\right) < \infty$ for j, l, m.*

Assumption A1 provides an essential condition for the use of the central limit theorem in Kelejian and Prucha [40]. Assumption A2 describes the dynamic relation between the spatial weight matrix and sample size n. If $\{h_n\}$ is a bounded sequence, Assumption A2 is easily satisfied. In the Case model [34] where h_n may diverge to infinity also satisfies Assumption A2. Assumption A3 can guarantee the existence of mean and variance of independent variable. Assumption A4 implies that the variance of $\boldsymbol{Y}_n$ is bounded as n goes to infinity. Similar conditions have been adopted in Kelejian and Prucha [40] and Lee [3]. Assumption A5 can exclude the multicollinearity of the regressors $\boldsymbol{X}_n$. Assuming that the regressors are uniformly bounded is convenient for analysis. If not, it can be replaced by stochastic regressors with certain finite moment conditions [3]. Assumption A6 is deals well with the nonlinearity of $\ln|\boldsymbol{S}_{1n}(\rho_1)|$ and $\ln|\boldsymbol{S}_{2n}(\rho_2)|$in the log-quasi-likelihood function. Assumption A7 means that $\boldsymbol{G}_{kn}\boldsymbol{X}_n\boldsymbol{\beta}_0$ and $\boldsymbol{X}_n$ are not asymptotically multicollinear with $k = 1, 2$. It is an identification condition of $\boldsymbol{\theta_0}$. Assumptions A8 and A9 are applied for Taylor expansion of the log-quasi-likelihood function and asymptotic normality of the estimator.

Appendix B. Proofs of Theorems 1 and 2

The following Lemmas are used for proofs of Theorems 1 and 2.

Lemma A1. *Under Assumptions A1–A7, we have:*

$$\frac{1}{\sqrt{n}}\frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = O_p(1).$$

Lemma A2. *Under Assumptions A1–A8, we have:*

$$\frac{1}{n}\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}} = E\left(\frac{1}{n}\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}}\right) + o_p(1).$$

Lemma A3. *Suppose that* $\liminf_{n\to\infty} \liminf_{\delta\to 0^+} p'_{\lambda_n}(\delta)/\lambda_n > 0$, $\lim_{n\to\infty} \lambda_n = 0$, $\lim_{n\to\infty} \sqrt{n}\lambda_n = \infty$, *and Assumptions 1–9 hold. Then with probability approaching one,*

$$J\left\{\begin{pmatrix} \boldsymbol{\theta}_1 \\ \mathbf{0} \end{pmatrix}\right\} = \min_{\|\boldsymbol{\theta}_2\| \le Cn^{-1/2}} J\left\{\begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix}\right\},$$

where $\boldsymbol{\theta}_1$ *satisfies* $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{10}\| = O_P\left(n^{-1/2}\right)$ *and C is a constant.*

Proof of Lemma A1. It follows from a straightforward calculation that:

$$\frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{1}{2\sigma_0^4}\left(\boldsymbol{E}_n^{\mathrm{T}}\boldsymbol{E}_n - n\sigma_0^2\right) \\ \frac{1}{\sigma_0^2}\boldsymbol{E}_n^{\mathrm{T}}\boldsymbol{S}_{2n}\boldsymbol{W}_{1n}\boldsymbol{Y}_n - \mathrm{tr}(\boldsymbol{G}_{1n}) \\ \frac{1}{\sigma_0^2}\left(\boldsymbol{E}_n^{\mathrm{T}}\boldsymbol{G}_{2n}\boldsymbol{E}_n - \sigma_0^2\mathrm{tr}(\boldsymbol{G}_{2n})\right) \\ \frac{1}{\sigma_0^2}(\boldsymbol{S}_{2n}\boldsymbol{X}_n)^{\mathrm{T}}\boldsymbol{E}_n \end{pmatrix}. \tag{A1}$$

By (A1) and some operational properties of related matrices in [3], we have:

$$\frac{1}{\sqrt{n}}\frac{\partial \ln L_n(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} = \frac{1}{\sqrt{n}\sigma_0^2}(\boldsymbol{S}_{2n}\boldsymbol{X}_n)^{\mathrm{T}}\boldsymbol{E}_n = O_p(1).$$

Note that:

$$\begin{aligned}
\mathrm{var}\left(\frac{1}{\sqrt{n}}\frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial \sigma^2}\right) &= \frac{1}{4n\sigma_0^8}\mathrm{var}\left(\boldsymbol{E}_n^{\mathrm{T}}\boldsymbol{E}_n\right) = O(1), \\
\mathrm{var}\left(\frac{1}{\sqrt{n}}\frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial \rho_1}\right) &\le \frac{2}{n\sigma_0^2}(\boldsymbol{S}_{2n}\boldsymbol{G}_{1n}\boldsymbol{X}_n\boldsymbol{\beta}_0)^{\mathrm{T}}(\boldsymbol{S}_{2n}\boldsymbol{G}_{1n}\boldsymbol{X}_n\boldsymbol{\beta}_0) \\
&\quad + \frac{2}{n\sigma_0^4}\mathrm{var}\left(\boldsymbol{E}_n^{\mathrm{T}}\boldsymbol{S}_{2n}\boldsymbol{G}_{1n}\boldsymbol{S}_{2n}^{-1}\boldsymbol{E}_n\right) = O(1), \\
\mathrm{var}\left(\frac{1}{\sqrt{n}}\frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial \rho_2}\right) &= \frac{1}{n\sigma_0^4}\mathrm{var}\left(\boldsymbol{E}_n^{\mathrm{T}}\boldsymbol{G}_{2n}\boldsymbol{E}_n\right) = O(1).
\end{aligned}$$

By the Chebyshev inequality, we obtain:

$$\frac{1}{\sqrt{n}}\frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial \sigma^2} = O_p(1), \frac{1}{\sqrt{n}}\frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial \rho_1} = O_p(1), \frac{1}{\sqrt{n}}\frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial \rho_2} = O_p(1).$$

□

Proof of Lemma A2. Note that:

$$\begin{aligned}
\frac{1}{n}\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial^2\sigma^2} &= \frac{1}{2\sigma_0^4} - \frac{1}{n\sigma_0^6}\boldsymbol{E}_n^{\mathrm{T}}\boldsymbol{E}_n,\\
\frac{1}{n}\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} &= -\frac{1}{n\sigma_0^2}(\boldsymbol{S}_{2n}\boldsymbol{X}_n)^{\mathrm{T}}\boldsymbol{S}_{2n}\boldsymbol{X}_n,\\
\frac{1}{n}\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial\rho_1^2} &= -\frac{1}{n\sigma_0^2}\Big(\sigma_0^2\mathrm{tr}\Big(\boldsymbol{G}_{1n}^2\Big) + (\boldsymbol{S}_{2n}\boldsymbol{W}_{1n}\boldsymbol{Y}_n)^{\mathrm{T}}\boldsymbol{S}_{2n}\boldsymbol{W}_{1n}\boldsymbol{Y}_n\Big),\\
\frac{1}{n}\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial\rho_2^2} &= -\frac{1}{n\sigma_0^2}\Big(\sigma_0^2\mathrm{tr}\Big(\boldsymbol{G}_{2n}^2\Big) + (\boldsymbol{G}_{2n}\boldsymbol{E}_n)^{\mathrm{T}}\boldsymbol{G}_{2n}\boldsymbol{E}_n\Big),\\
\frac{1}{n}\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\beta}\partial\rho_1} &= -\frac{1}{n\sigma_0^2}(\boldsymbol{S}_{2n}\boldsymbol{X}_n)^{\mathrm{T}}\boldsymbol{S}_{2n}\boldsymbol{W}_{1n}\boldsymbol{Y}_n,\\
\frac{1}{n}\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\beta}\partial\rho_2} &= -\frac{1}{n\sigma_0^2}\Big((\boldsymbol{S}_{2n}\boldsymbol{X}_n)^{\mathrm{T}}\boldsymbol{G}_{2n} + (\boldsymbol{W}_{2n}\boldsymbol{X}_n)^{\mathrm{T}}\Big)\boldsymbol{E}_n,\\
\frac{1}{n}\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\beta}\partial\sigma^2} &= -\frac{1}{n\sigma_0^4}(\boldsymbol{S}_{2n}\boldsymbol{X}_n)^{\mathrm{T}}\boldsymbol{E}_n,\\
\frac{1}{n}\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial\rho_1\partial\rho_2} &= -\frac{1}{n\sigma_0^2}\bigg(\Big(\boldsymbol{S}_{2n}^{-1}\boldsymbol{E}_n\Big)^{\mathrm{T}}\Big(\boldsymbol{S}_{2n}^{\mathrm{T}}\boldsymbol{W}_{2n} + \boldsymbol{S}_{2n}\boldsymbol{W}_{2n}^{\mathrm{T}}\Big)\boldsymbol{W}_{1n}\boldsymbol{Y}_n\bigg),\\
\frac{1}{n}\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial\rho_1\partial\sigma^2} &= -\frac{1}{2n\sigma_0^4}\Big((\boldsymbol{S}_{2n}\boldsymbol{W}_{1n}\boldsymbol{Y}_n)^{\mathrm{T}}\boldsymbol{E}_n + \boldsymbol{E}_n^{\mathrm{T}}\boldsymbol{S}_{2n}\boldsymbol{W}_{1n}\boldsymbol{Y}_n\Big),\\
\frac{1}{n}\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial\rho_2\partial\sigma^2} &= -\frac{1}{2n\sigma_0^4}\boldsymbol{E}_n^{\mathrm{T}}\boldsymbol{G}_{2n}^{s}\boldsymbol{E}_n.
\end{aligned}$$

Then, similar to the proof of Theorem 3.2 in [3], we can obtain Lemma 2. □

Proof of Theorem 1. Let $z_n = n^{-1/2} + a_n$. As demonstrated by Fan and Li [13], it suffices to prove that for any given $\eta > 0$, there is a positive constant C such that:

$$P\bigg\{\inf_{\|\boldsymbol{u}\|=C} J(\boldsymbol{\theta}_0 + z_n\boldsymbol{u}) > J(\boldsymbol{\theta}_0)\bigg\} \geq 1 - \eta. \tag{A2}$$

(A2) shows that there is a local minimizer in a bounded closed domain $\{\boldsymbol{\theta}_0 + z_n\boldsymbol{u} : ||u|| \leq C\}$ for continuous function $J(\boldsymbol{\theta})$ with probability at least $1-\eta$. Consequently, there exists a local minimizer $\hat{\boldsymbol{\theta}}_0$ such that $||\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0|| = O_p(z_n)$.

By $p_{\lambda_n}(0) = 0, z_n = o(1)$ and the Taylor expansion, we have:

$$\begin{aligned}
\frac{J(\boldsymbol{\theta}_0 + z_n\boldsymbol{u}) - J(\boldsymbol{\theta}_0)}{n} \geq& -n^{-1}z_n\left(\frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}_n(\boldsymbol{\theta}_0)\boldsymbol{u}z_n^2\{1 + o_p(1)\}\\
&+ \sum_{j=1}^{s}\Big[z_n d_j u_j + z_n^2 v_j u_j^2\{1 + o(1)\}\Big]\\
=& J_1 + J_2 + J_3,
\end{aligned}$$

where,

$$\begin{aligned}
J_1 &= -n^{-1}z_n\left(\frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}\boldsymbol{u},\\
J_2 &= \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}_n(\boldsymbol{\theta}_0)\boldsymbol{u}z_n^2\{1 + o_p(1)\},\\
J_3 &= \sum_{j=1}^{s}\Big[z_n d_j u_j + z_n^2 v_j u_j^2\{1 + o(1)\}\Big].
\end{aligned}$$

From Lemma 1, $b_n = o(1)$, and $O_p(n^{-1/2}z_n) = O_p(z_n^2)$, it follows that $J_1 = \|\boldsymbol{u}\| \cdot O_p\left(z_n^2\right)$, $J_2 = \|\boldsymbol{u}\|^2 \cdot O_p\left(z_n^2\right)$, and J_3 is bounded by $\|\boldsymbol{u}\| \cdot O_p\left(z_n^2\right) + \|\boldsymbol{u}\|^2 \cdot o_p\left(z_n^2\right)$. Thus, J_1 and J_3 can be dominated by J_2 uniformly with a sufficiently large $\|\boldsymbol{u}\| = C$ when $n \to \infty$. Hence, (A2) holds. Note that $z_n = o(n^{-1/2})$. This completes the proof of Theorem 1. □

Proof of Lemma A3. It suffices to prove that, for any $\boldsymbol{\theta}_1$ satisfying $||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{10}|| = O_p\left(n^{-1/2}\right)$ and $||\boldsymbol{\theta}_2|| \leq Cn^{-1/2}$, and $j = s+1, \cdots, k+3$, with probability tending to 1 as $n \to \infty$, $\partial J(\boldsymbol{\theta})/\partial\theta_j$ and θ_j have the same signs for $\theta_j \in (-Cn^{-1/2}, Cn^{-1/2})$.

For $\theta_j \neq 0$ and $j = s+1, \cdots, k+3$,

$$\frac{\partial J(\boldsymbol{\theta})}{\partial\theta_j} = -\frac{\partial \ln L_n(\boldsymbol{\theta})}{\partial\theta_j} + np'_{\lambda_{jn}}(|\theta_j|)\mathrm{sgn}(\theta_j).$$

By the Taylor expansion, we have:

$$\begin{aligned}\frac{\partial \ln L_n(\boldsymbol{\theta})}{\partial\theta_j} =& \frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial\theta_j} + \sum_{l=1}^{k+3} \frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial\theta_j\partial\theta_l}(\theta_l - \theta_{l,0}) \\ &+ \sum_{l=1}^{k+3}\sum_{m=1}^{k+3} \frac{\partial^3 \ln L_n(\boldsymbol{\theta}^*)}{\partial\theta_j\partial\theta_l\partial\theta_m}(\theta_l - \theta_{l,0})(\theta_m - \theta_{m,0}),\end{aligned}$$

where $\boldsymbol{\theta}^*$ lies between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. Under $||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{10}|| = O_p\left(n^{-1/2}\right)$, $||\boldsymbol{\theta}_2|| \leq Cn^{-1/2}$ and Assumption A9, we can obtain by Lemmas 1 and 2 that $n^{-1}\partial \ln L_n(\boldsymbol{\theta})/\partial\theta_j$ is of order $O_p(n^{-1/2})$. Thus,

$$\frac{\partial J(\boldsymbol{\theta})}{\partial\theta_j} = n\lambda_{jn}\left\{\lambda_{jn}^{-1}p'_{\lambda_{jn}}(|\theta_j|)\mathrm{sgn}(\theta_j) + O_P\left(n^{-1/2}\lambda_{jn}^{-1}\right)\right\}.$$

Note that $\liminf_{n\to\infty}\liminf_{\delta\to 0^+}\lambda_{jn}^{-1}p'_{\lambda_{jn}}(\delta) > 0$ and $\lim_{n\to\infty} n^{-1/2}\lambda_{jn}^{-1} = 0$. The sign of the derivative is the same as that of θ_j for a sufficiently large n. This shows that the minimizer attains at $\boldsymbol{\theta}_2 = \mathbf{0}$. Lemma 3 is proven. □

Proof of Theorem 2. Lemma 3 shows that part (i) holds. Next, we give the proof of part (ii). By Theorem 1, there is a $\sqrt{n}$ consistent local minimizer of $J\{(\boldsymbol{\theta}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}\}$ denoted as $\hat{\boldsymbol{\theta}}_1$, which satisfies:

$$\left.\frac{\partial J(\boldsymbol{\theta})}{\partial\theta_j}\right|_{\boldsymbol{\theta}=(\hat{\boldsymbol{\theta}}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}} = 0 \quad \text{for} \quad j = 1, \cdots, s. \tag{A3}$$

Note that $\theta_1 = \sigma^2$. By the Taylor expansion, we have:

$$\begin{aligned}\frac{\partial J(\boldsymbol{\theta})}{\partial\theta_j} =& \frac{\partial \ln L_n(\boldsymbol{\theta})}{\partial\theta_j} - np'_{\lambda_{jn}}(|\theta_j|)\mathrm{sgn}(\theta_j)\mathrm{I}(j \neq 1) \\ =& \frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial\theta_j} + \sum_{l=1}^{s}\left\{\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial\theta_j\partial\theta_l} + o_p(1)\right\}(\theta_l - \theta_{l,0}) \\ &- n\left[p'_{\lambda_{jn}}(|\theta_{j,0}|)\mathrm{sgn}(\theta_{j,0}) + \left\{p''_{\lambda_{jn}}(|\theta_{j,0}|) + o_p(1)\right\}(\theta_j - \theta_{j,0})\right]\mathrm{I}(j \neq 1),\end{aligned} \tag{A4}$$

where $\mathrm{I}(j \neq 1)$ is an indicator function.

Moreover, it follows from (A3) and (A4) that:

$$\frac{\partial \ln L_n(\boldsymbol{\theta}_0)}{\partial\theta_j} = -\sum_{l=1}^{s}\left\{\frac{\partial^2 \ln L_n(\boldsymbol{\theta}_0)}{\partial\theta_j\partial\theta_l}\right\}(\hat{\theta}_l - \theta_{l,0}) + nv_j(\hat{\theta}_j - \theta_{j,0}) + nd_j + o_p(\sqrt{n}). \tag{A5}$$

Note that (A1) can be written as:

$$\begin{pmatrix} 0 \\ \frac{1}{\sigma_0^2}\left(\boldsymbol{S}_{2n}\boldsymbol{G}_{1n}\boldsymbol{X}_n\boldsymbol{\beta}_0\right)^{\mathrm{T}}\boldsymbol{E}_n \\ 0 \\ \frac{1}{\sigma_0^2}\left(\boldsymbol{S}_{2n}\boldsymbol{X}_n\right)^{\mathrm{T}}\boldsymbol{E}_n \end{pmatrix} + \begin{pmatrix} \frac{1}{2\sigma_0^4}\left(\boldsymbol{E}_n^{\mathrm{T}}\boldsymbol{E}_n - n\sigma_0^2\right) \\ \frac{1}{\sigma_0^2}\left(\boldsymbol{E}_n^{\mathrm{T}}\boldsymbol{S}_{2n}\boldsymbol{G}_{1n}\boldsymbol{S}_{2n}^{-1}\boldsymbol{E}_n - \sigma_0^2\mathrm{tr}(\boldsymbol{G}_{1n})\right) \\ \frac{1}{\sigma_0^2}\left(\boldsymbol{E}_n^{\mathrm{T}}\boldsymbol{G}_{2n}\boldsymbol{E}_n - \sigma_0^2\mathrm{tr}(\boldsymbol{G}_{2n})\right) \\ \mathbf{0} \end{pmatrix}.$$

Then, by (A5), Slutsky's theorem and the central limit theorem of the linear-quadratic form [40], we can obtain:

$$\sqrt{n}\left[(\boldsymbol{\Sigma}_{n1}(\boldsymbol{\theta}_{10}) + \boldsymbol{\Lambda})\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right) + \boldsymbol{d}\right] \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_1(\boldsymbol{\theta}_{10}) + \boldsymbol{\Omega}_1(\boldsymbol{\theta}_{10})).$$

This completes the proof. □

References

1. Cliff, A.D.; Ord, J.K. *Spatial Autocorrelation*; Pion Ltd.: London, UK, 1973.
2. Kelejian, H.H.; Prucha, I.R. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *J. Real. Estate. Finac.* **1998**, *17*, 99–121. [CrossRef]
3. Lee, L.F. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* **2004**, *72*, 1899–1925. [CrossRef]
4. Arraiz, I.; Drukker, D.M.; Kelejian, H.H.; Prucha, I.R. A spatial Cliff-Ord-type model with heteroskedastic innovations: Small and large sample results. *J. Regional. Sci.* **2010**, *50*, 592–614. [CrossRef]
5. Anselin, L. *Spatial Econometrics: Methods and Models*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1988.
6. Cressie, N. *Statistics for Spatial Data*; John Wiley and Sons: New York, NY, USA, 1993.
7. Akaike, H. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **1973**, *60*, 255–265. [CrossRef]
8. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
9. Foster, D.P; George, E.I. The risk inflation criterion for multiple regression. *Ann. Stat.* **1994**, *22*, 1947–1975. [CrossRef]
10. Liang, H.; Li, R. Variable selection for partially linear models with measurement errors. *J. Am. Stat. Assoc.* **2009**, *104*, 234–248. [CrossRef]
11. Huo, X.; Ni, X. When do stepwise algorithms meet subset selection criteria? *Ann. Stat.* **2007**, *35*, 870–887. [CrossRef]
12. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **1996**, *58*, 267–288. [CrossRef]
13. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
14. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **2005**, *67*, 301–320. [CrossRef]
15. Zou, H. The adaptive Lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
16. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef]
17. Mitchell, T.J.; Beauchamp, J.J. Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.* **1988**, *83*, 1023–1032. [CrossRef]
18. Raftery, A.E.; Madigan, D.; Hoeting, J.A. Bayesian model averaging for linear regression models. *J. Am. Stat. Assoc.* **1997**, *92*, 179–191. [CrossRef]
19. Jiang, W.X. Bayesian variable selection for high dimensional generalized linear models: Convergence rates for the fitted densities. *Ann. Stat.* **2007**, *35*, 1487–1511. [CrossRef]
20. Chen, Y.; Du, P.; Wang, Y. Variable selection in linear models. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *6*, 1–9. [CrossRef]
21. Steel, M.F. Model averaging and its use in economics. *J. Econ. Lit.* **2020**, *58*, 644–719. [CrossRef]
22. LeSage, J.P.; Parent, O. Bayesian model averaging for spatial econometric models. *Geogr. Anal.* **2007**, *39*, 241–267. [CrossRef]
23. LeSage, J.P.; Fischer, M. Spatial growth regressions, model specification, estimation, and interpretation. *Spat. Econ. Anal.* **2008**, *3*, 275–304. [CrossRef]
24. Cuaresma, J.C.; Doppelhofer, G.; Feldkircher, M. The determinants of economic growth in European regions. *Reg. Stud.* **2014**, *48*, 44–67. [CrossRef]
25. Cuaresma, J.C.; Doppelhofer, G.; Huber, F.; Piribauer, P. Human capital accumulation and long-term income growth projections for European regions. *J. Regional. Sci.* **2018**, *58*, 81–99. [CrossRef]
26. Piribauer, P. Heterogeneity in spatial growth clusters. *Empir. Econ.* **2016**, *51*, 659–680. [CrossRef]
27. Zhu, J.; Huang, H.; Reyes, P.E. On selection of spatial linear models for lattice data. *J. R. Statist. Soc. B* **2010**, *72*, 389–402. [CrossRef]
28. Liu, X.; Chen, J.; Cheng, S. A penalized quasi-maximum likelihood method for variable selection in the spatial autoregressive model. *Spat. Stat.* **2018**, *25*, 86–104. [CrossRef]
29. Xie, T.; Cao, R.; Du, J. Variable selection for spatial autoregressive models with a diverging number of parameters. *Stat. Pap.* **2020**, *61*, 1125–1145. [CrossRef]
30. Wang, H.; Zhu, J. Variable selection in spatial regression via penalized least squares. *Can. J. Stat.* **2009**, *37*, 607–624. [CrossRef]

31. Zou, H.; Li, R. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat.* **2008**, *36*, 1509–1533.
32. Nelder, J.A.; Mead, R. A simplex method for function minimization. *Comput. J.* **1965**, *7*, 308–313. [CrossRef]
33. Wang, H.; Li, R.; Tsai, C.L. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **2007**, *94*, 553–568. [CrossRef] [PubMed]
34. Case, A.C. Spatial patterns in household demand. *Econometrica* **1991**, *59*, 953–965. [CrossRef]
35. Harrison, D.H.; Rubinfeld, D.L. Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manag.* **1978**, *5*, 81–102. [CrossRef]
36. Pace, R.K.; Gilley, O.W. Using the spatial configuration of the data to improve estimation. *J. Real. Estate. Financ.* **1997**, *14*, 333–340. [CrossRef]
37. Tang, Q. Robust estimation for functional coefficient regression models with spatial data. *Statistics* **2014**, *48*, 388–404. [CrossRef]
38. LeSage, J.; Pace, R. *Introduction to Spatial Econometrics*; CRC Press: Boca Raton, FL, USA, 2009.
39. Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge University Press: Cambridge, UK, 1985.
40. Kelejian, H.H.; Prucha, I.R. On the asymptotic distribution of the Moran I test statistic with applications. *J. Econom.* **2001**, *104*, 219–257. [CrossRef]