

Article

Variable Selection of Heterogeneous Spatial Autoregressive Models via Double-Penalized Likelihood

Ruiqin Tian ¹, Miaojie Xia ¹  and Dengke Xu ^{2,*} 
¹ School of Mathematics, Hangzhou Normal University, Hangzhou 311121, China; ruiqintian@163.com (R.T.); xiamiaojie@stu.hznu.edu.cn (M.X.)

² College of Economics, Hangzhou Dianzi University, Hangzhou 310018, China

* Correspondence: xudengke1983@163.com

Abstract: Heteroscedasticity is often encountered in spatial-data analysis, so a new class of heterogeneous spatial autoregressive models is introduced in this paper, where the variance parameters are allowed to depend on some explanatory variables. Here, we are interested in the problem of parameter estimation and the variable selection for both the mean and variance models. Then, a unified procedure via double-penalized quasi-maximum likelihood is proposed, to simultaneously select important variables. Under certain regular conditions, the consistency and oracle property of the resulting estimators are established. Finally, both simulation studies and a real data analysis of the Boston housing data are carried to illustrate the developed methodology.

Keywords: heterogeneous spatial autoregressive models; double-penalized quasi-maximum likelihood; variable selection; SCAD; tuning parameters



Citation: Tian, R.; Xia, M.; Xu, D. Variable Selection of Heterogeneous Spatial Autoregressive Models via Double-Penalized Likelihood. *Symmetry* **2022**, *14*, 1200. <https://doi.org/10.3390/sym14061200>

Academic Editor: Jinyu Li

Received: 10 May 2022

Accepted: 7 June 2022

Published: 10 June 2022

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



Copyright: © 2022 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

In the field of modern economics and statistics, the spatial autoregressive (SAR) model has always been a hot topic for economists and statisticians. Theories and methods for estimation as well as other inferences based on linear SAR models and their extensions have been deeply studied. Among them, there is a lot of wonderful literature based on linear SAR models, such as Cliff and Ord [1], Anselin [2], and Anselin and Bera [3]. Other research results for linear SAR models can also be referred to: Xu and Lee [4], Liu et al. [5], Xie et al. [6], Xie et al. [7], and so on. In order to capture the nonlinear relationship between response variables and some explanatory variables, a variety of semiparametric SAR models have recently been proposed and studied in depth. For example, based on a partially linear spatial autoregressive model, Su and Jin [8] developed a profile quasi-maximum likelihood-estimation method and discussed the asymptotic properties of the obtained estimators. Du et al. [9] developed partially linear additive spatial autoregressive models and proposed an estimation method via combining the spline approximations with an instrumental-variables-estimation method. Cheng and Chen [10] discussed partially linear single-index spatial autoregressive model as well as obtained the consistency and asymptotic normality of the proposed estimators, under some mild conditions. Other research results of semiparametric SAR models can also be seen in Wei et al. [11], Hu et al. [12], and so on. Previous research on various SAR models focused mainly on the homoskedasticity assumption, that the variance of the unobservable error, conditional on the explanatory variables, is constant. It is well-known that if the innovations are heteroskedastic, most existing statistical inference methods under homoskedasticity assumption could lead to incorrect inference, see Lin and Lee [13]. Therefore, many researchers want to relax the homoskedasticity assumption for spatial autoregressive models, by allowing different variances for each unobservable error. For example, Dai et al. [14] developed a Bayesian local-influence analysis for heterogeneous spatial autoregressive models. However, the variance terms in these above-mentioned heterogeneous spatial autoregressive models are assumed fixed and do not depend on the regression variables. Furthermore, in many application fields,

such as economics and quality management, it is a topic of interest to model the variance itself, which is helpful to identify the factors that affect the variability in the observations. Thus, we intend to propose a class of heterogeneous spatial autoregressive models where the variance parameters are modelled in terms of covariates.

In addition, many authors have done a lot of research on joint mean and variance models. For example, Wu and Li [15] proposed a variable-selection procedure via a penalized maximum likelihood for the joint mean and dispersion models of the inverse Gaussian distribution. Xu and Zhang [16] discussed a Bayesian estimation for semiparametric joint mean and variance models, based on B-spline approximations of nonparametric components. Zhao et al. [17] studied variable selection for beta-regression models with varying dispersion, where both the mean and the dispersion are modeled by explanatory variables. Li et al. [18] proposed an efficient unified variable selection of the joint location, scale, and skewness models, based on a penalized likelihood method. Zhang et al. [19] developed a Bayesian quantile-regression analysis for semiparametric mixed-effects double-regression models, on the basis of the asymmetric Laplace distribution for the errors.

As we all know, variable selection is the most important problem to be considered in regression analysis. In practice, a large number of variables should be used in the initial analysis, but many of them may not be important and should be excluded from the final model, in order to improve the accuracy of the prediction. When the number of predictive variables is large, traditional variable selection methods such as stepwise regression and best-subset selection are not computationally feasible. Therefore, various shrinkage methods have been proposed in recent years and have gained much attention, such as the LASSO (Tibshirani [20]), the adaptive LASSO (Zou [21]), and the SCAD (Fan and Li [22]). Based on these shrinkage methods, the variable selection for SAR models (see Liu et al. [5]; Xie et al. [6]; Xie et al. [7]; Luo and Wu [23]) and the variable selection for other models without spatial dependence (see, for example, Li and Liang [24]; Zhao and Xue [25]; Tian et al. [26]) have been studied extensively in recent years. To the best of our knowledge, most existing variable-selection methods in spatial-data analysis are limited, to only select the mean explanatory variables in the literature, and little work has been done to select the variance explanatory variables.

Therefore, in this paper we aim to make variable selection of heterogeneous spatial autoregressive models (heterogeneous SAR models), based on penalized quasi-maximum likelihood, using different penalty functions. The proposed variable-selection method simultaneously selects important explanatory variables in a mean model and a variance model. Furthermore, it can be proven that this variable selection procedure is consistent, and the obtained estimators of regression coefficients have oracle property under certain regular conditions. This indicates that the penalized estimators work as well as if the subset of true zero coefficients were already known. A simulation and a real data analysis of a Boston housing price are used to illustrate the proposed variable selection method.

The outline of the paper is organized as follows. Section 2 introduces new heterogeneous spatial autoregressive models. Then, we propose a unified variable-selection procedure for the joint models via the double-penalized quasi-maximum likelihood method. Section 3 gives the theoretical results of the resulting estimators. The computation of the penalized quasi-maximum likelihood estimator as well as the choice of the tuning parameters are presented in Section 4. The finite-sample performance of the method is investigated, based on some simulation studies, in Section 5. Section 6 gives a real data analysis of a Boston housing price to illustrate the proposed method. Some conclusions as well as a brief discussion are given in Section 7. Some assumptions and the technical proofs of all the asymptotic results are provided in Appendix A.

2. Variable Selection via Penalized Quasi-Maximum Likelihood

2.1. Heterogeneous SAR Models

The classical spatial autoregressive models have the following form:

$$Y = \rho WY + X\beta + \varepsilon, \quad (1)$$

where $Y = (y_1, y_2, \dots, y_n)^T$ is an n -dimensional observation vector on the dependent variable; $|\rho| < 1$ is an unknown spatial parameter; and W is a specified spatial weight matrix of known constants with zero diagonal elements. Let $X = (x_1, x_2, \dots, x_n)^T$ is an $n \times p$ matrix whose i th row $x_i^T = (x_{i1}, \dots, x_{ip})$ is the observation of the explanatory variables and $\beta = (\beta_1, \dots, \beta_p)^T$ be a $p \times 1$ vector of unknown regression parameters in the mean model; ε is an n -dimensional vector of independent identically distributed disturbances with zero mean and finite variance σ^2 .

Furthermore, similar to Xu and Zhang [16], we consider variance heterogeneity in the models and assume an explicit variance modeling related to other explanatory variables, that is:

$$\sigma_i^2 = h(z_i^T \gamma), \quad (2)$$

where $z_i^T = (z_{i1}, \dots, z_{iq})$ is the observation of explanatory variables associated with the variance of y_i and $\gamma = (\gamma_1, \dots, \gamma_q)^T$ is a $q \times 1$ vector of regression parameters in the variance model. There might be some components of z_i which coincide with some x_i 's. In addition, $h(\cdot)$ is a known function, for the identifiability of the models, so we always assume that $h(\cdot)$ is a monotone function and $h(\cdot) > 0$, which takes into account the positiveness of the variance. For example, the researcher can take $h(x) = \exp(x)$ in general. So, this paper considers the following heterogeneous SAR models:

$$\begin{cases} Y = \rho WY + X\beta + \varepsilon, \\ \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2) \\ \sigma_i^2 = \exp(z_i^T \gamma), \\ i = 1, 2, \dots, n. \end{cases} \quad (3)$$

According to the idea of the quasi-maximum likelihood estimators (Lee [27]), the log-likelihood function of the model (3) is

$$\ell_n(\rho, \beta, \gamma | Y, X, Z) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^n z_i^T \gamma + \ln|S| - \frac{1}{2} e^T \Sigma^{-1} e, \quad (4)$$

where $e = SY - X\beta$, $S = I_n - \rho W$, and I_n is an $(n \times n)$ identity matrix.

2.2. Penalized Quasi-Maximum Likelihood

In order to obtain the desired sparsity in the resulting estimators, we propose the penalized quasi-maximum likelihood

$$\mathcal{L}(\rho, \beta, \gamma) = \ell_n(\rho, \beta, \gamma) - n \sum_{j=1}^p p_{\lambda_j^{(1)}}(|\beta_j|) - n \sum_{k=1}^q p_{\lambda_k^{(2)}}(|\gamma_k|). \quad (5)$$

For notational simplicity, we rewrite (5) in the following

$$\mathcal{L}(\theta) = \ell_n(\theta) - n \sum_{j=2}^s p_{\lambda_n}(|\theta_j|), \quad (6)$$

where $\theta = (\theta_1, \dots, \theta_s)^T = (\rho, \beta_1, \dots, \beta_p; \gamma_1, \dots, \gamma_q)^T$ with $s = p + q + 1$, and $p_{\lambda_n}(\cdot)$ is a given-penalty function with the tuning parameters λ_n . The data-driven criteria, such as cross validation (CV), generalized cross-validation (GCV), or the BIC-type tuning-parameter selector (see Wang et al. [28]), can be used to choose the tuning parameters, which is described in Section 4. Here we use the same penalty function $p(\cdot)$ for all the regression coefficients but with different tuning parameters λ_n , for the mean parameters and the variance parameters, respectively. Note that the penalty functions and tuning parameters are not necessarily the same for all the parameters. For example, we wish to keep some important variables in the final model and, therefore, do not want to penalize their coefficients.

In this paper, we mainly use the smoothly clipped absolute deviation (SCAD) penalty, with a first derivative that satisfies

$$p'_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\}$$

in which $a = 3.7$ is taken in our work. The details about SCAD can be seen in Fan and Li [22].

The penalized quasi-maximum likelihood estimator of θ is denoted by $\hat{\theta}$, which maximizes the function $\mathcal{L}(\theta)$ in (6) with respect to θ . The technical details and an algorithm for calculating the penalized quasi-maximum likelihood estimator $\hat{\theta}$ are provided in Section 4.

3. Asymptotic Properties

We next study the asymptotic properties of the resulting penalized quasi-maximum likelihood estimators. We first introduce some notations. Let θ_0 denote the true value of θ . Furthermore, let $\theta_0 = (\theta_{01}, \dots, \theta_{0s})^T = ((\theta_0^{(1)})^T, (\theta_0^{(2)})^T)^T$. For ease of presentation and without loss of generality, it is assumed that $\theta_0^{(1)}$ is an $s_1 \times 1$ nonzero regression coefficient and that $\theta_0^{(2)} = 0$ is an $(s - s_1) \times 1$ vector of the zero-valued regression coefficient. Let

$$a_n = \max_{2 \leq j \leq s} \{ |p'_{\lambda_n}(|\theta_{0j}|)|, \theta_{0j} \neq 0 \}$$

and

$$b_n = \max_{2 \leq j \leq s} \{ p''_{\lambda_n}(|\theta_{0j}|) : \theta_{0j} \neq 0 \}.$$

Theorem 1. Suppose that $a_n = O_p(n^{-\frac{1}{2}})$, $b_n \rightarrow 0$, and $\lambda_n \rightarrow 0$ as $n \rightarrow \infty$. Under the conditions (C1)–(C10) in Appendix A, with probability tending to 1 there must exist a local maximizer $\hat{\theta}$ of the penalized quasi-maximum likelihood function $\mathcal{L}(\theta)$ in (6), such that $\hat{\theta}$ is a $\sqrt{n}$ -consistent estimator of θ_0 .

The following theorem gives the asymptotic normality of $\hat{\theta}$. Let

$$A_n = \text{diag}(0, p''_{\lambda_n}(|\theta_{01}^{(1)}|), \dots, p''_{\lambda_n}(|\theta_{0s_1}^{(1)}|)),$$

$$B_n = (0, p'_{\lambda_n}(|\theta_{01}^{(1)}|)\text{sgn}(\theta_{01}^{(1)}), \dots, p'_{\lambda_n}(|\theta_{0s_1}^{(1)}|)\text{sgn}(\theta_{0s_1}^{(1)}))^T,$$

where $\theta_{0j}^{(1)}$ is the j th component of $\theta_0^{(1)}$ ($1 \leq j \leq s_1$).

Furthermore, denote

$$\begin{aligned} \mathcal{I}_n(\theta_0) &= -E \left(\frac{1}{n} \frac{\partial^2 \ell_n(\theta_0)}{\partial \theta \partial \theta^T} \right) \\ &= \begin{pmatrix} \frac{1}{n} [G_n X \beta_0]^T \Sigma^{-1} (G_n X \beta_0) + \text{tr}(G_n^s G_n) & \frac{1}{n} X^T \Sigma^{-1} X & * \\ \frac{1}{n} X^T \Sigma^{-1} (G_n X \beta_0) & 0 & 0 \\ \frac{1}{n} \sum_{i=1}^n g_{n,ii} Z_i^T & 0 & \frac{1}{2n} Z^T Z \end{pmatrix}, \end{aligned}$$

$$\mathcal{J}_n(\theta_0) = \begin{pmatrix} \mathcal{J}_1 & * & * \\ \frac{1}{n} X^T \Sigma^{-2} G_n \mu_3 & 0 & \frac{1}{2n} X^T \text{diag}(\mu_3) \Sigma^{-2} Z \\ \mathcal{J}_2 & * & \frac{1}{4n} [Z^T \Sigma^{-2} \text{diag}(\mu_4) Z - Z^T Z] \end{pmatrix},$$

where $\mathcal{J}_1 = \frac{1}{n} [\mu_4^T \Sigma^{-2} 1_n \text{tr}(G_n^2) + 2(G_n X \beta_0)^T G_n \Sigma^{-2} \mu_3 - 3 \text{tr}(G_n^2)]$, $\mathcal{J}_2 = \frac{1}{2n} [Z^T \Sigma^{-2} \text{diag}(\mu_3)(G_n X \beta_0) + Z^T \Sigma^{-2} G_n \mu_4 - 3 \sum_{i=1}^n g_{n,ii} Z_i^T]$, $G_n = W A^{-1}$, and $G_n^s = G_n + G_n^T$, thus, $\mathcal{I}_n(\theta_0)$ is called as the average Hessian matrix (the information matrix when $\{e_i\}$ are normal). In addition, $\mathcal{J}_n(\theta_0)$ is a symmetric matrix, 1_n is an n -dimensional column vector of ones, $\mu_j = E(e_i^j)$, $j = 2, 3, 4$, G_{in} is the i th row of G_n , and

$g_{n,ij}$ is the (i, j) th entry of G_n . Besides, $\text{diag}(C)$ represents a diagonal matrix with C , where C is an arbitrary vector.

Theorem 2. Suppose that the penalty function $p_\lambda(t)$ satisfies

$$\liminf_{n \rightarrow \infty} \liminf_{t \rightarrow 0^+} \frac{p'_{\lambda_n}(t)}{\lambda_n} > 0,$$

and under the same mild conditions as these given in Theorem 1, if $\lambda_n \rightarrow 0$ and $\sqrt{n}\lambda_n \rightarrow \infty$ as $n \rightarrow \infty$, and $\mathcal{J}(\theta_0) = \lim_{n \rightarrow \infty} \mathcal{J}_n(\theta_0)$, then the $\sqrt{n}$ -consistent estimator $\hat{\theta} = ((\hat{\theta}^{(1)})^T, (\hat{\theta}^{(2)})^T)^T$ in Theorem 1 must satisfy

- (i) $\hat{\theta}^{(2)} = 0$ with probability tending to 1.
- (ii) $\sqrt{n}(\mathcal{I}_n^{(1)}(\theta_0^{(1)}) + A_n)\{(\hat{\theta}^{(1)} - \theta_0^{(1)}) + (\mathcal{I}_n^{(1)}(\theta_0^{(1)}) + A_n)^{-1}B_n\} \xrightarrow{\mathcal{L}} \mathcal{N}_{s_1}(0, \mathcal{I}_1(\theta_0^{(1)}) + \mathcal{J}_1(\theta_0^{(1)}))$,

where $\mathcal{I}_n^{(1)}(\theta_0^{(1)})$, $\mathcal{I}_1(\theta_0^{(1)})$, and $\mathcal{J}_1(\theta_0^{(1)})$ are the first super-left submatrix of $\mathcal{I}_n(\theta_0)$, $\mathcal{I}(\theta_0) = \lim_{n \rightarrow \infty} \mathcal{I}_n(\theta_0)$ and $\mathcal{J}(\theta_0)$. In addition, " $\xrightarrow{\mathcal{L}}$ " denotes convergence in distribution.

4. Computation

4.1. Algorithm

Since $\mathcal{L}(\theta)$ is irregular at the origin, the commonly used gradient method is not applicable. Now, an iterative algorithm is developed based on the local quadratic approximation of the penalty function $p_{\lambda_n}(\cdot)$, as in Fan and Li [22].

Firstly, note that the first two derivatives of the log-likelihood function $\ell_n(\theta)$ are continuous. Around a fixed point $\theta_0 = (\rho_0, \beta_0^T, \gamma_0^T)^T$, we approximate the log-likelihood function by

$$\ell_n(\theta) \approx \ell_n(\theta_0) + \left[\frac{\partial \ell_n(\theta_0)}{\partial \theta} \right]^T (\theta - \theta_0) + \frac{1}{2} (\theta - \theta_0)^T \left[\frac{\partial^2 \ell_n(\theta_0)}{\partial \theta \partial \theta^T} \right] (\theta - \theta_0).$$

Moreover, given an initial value t_0 , the penalty function $p'_{\lambda_n}(t)$ can be approximated by a quadratic function

$$p_{\lambda_n}(|t|) \approx p_{\lambda_n}(|t_0|) + \frac{1}{2} \frac{p'_{\lambda_n}(|t_0|)}{|t_0|} (t^2 - t_0^2), \text{ for } t \approx t_0.$$

Therefore, we can approximate the penalized quasi-maximum likelihood function (6) by the following formula

$$\mathcal{L}(\theta) \approx \ell_n(\theta_0) + \left[\frac{\partial \ell_n(\theta_0)}{\partial \theta} \right]^T (\theta - \theta_0) + \frac{1}{2} (\theta - \theta_0)^T \left[\frac{\partial^2 \ell_n(\theta_0)}{\partial \theta \partial \theta^T} \right] (\theta - \theta_0) - \frac{n}{2} \theta^T A_{n\lambda}(\theta_0) \theta,$$

where

$$A_{n\lambda}(\theta_0) = \text{diag} \left\{ 0, \frac{p'_{\lambda_n(1)}(|\beta_{01}|)}{|\beta_{01}|}, \dots, \frac{p'_{\lambda_n(p)}(|\beta_{0p}|)}{|\beta_{0p}|}, \frac{p'_{\lambda_n(2)}(|\gamma_{01}|)}{|\gamma_{01}|}, \dots, \frac{p'_{\lambda_n(q)}(|\gamma_{0q}|)}{|\gamma_{0q}|} \right\},$$

$\theta = (\theta_1, \dots, \theta_s)^T = (\rho, \beta_1, \dots, \beta_p; \gamma_1, \dots, \gamma_q)^T$ and $\theta_0 = (\theta_{01}, \dots, \theta_{0s})^T = (\rho_0, \beta_{01}, \dots, \beta_{0p}; \gamma_{01}, \dots, \gamma_{0q})^T$. Accordingly, the quadratic maximization problem for $\mathcal{L}(\theta)$ leads to a solution iterated by

$$\theta_{new} \approx \theta_0 + \left\{ \frac{\partial^2 \ell_n(\rho_0, \beta_0, \gamma_0)}{\partial \theta \partial \theta^T} - n A_{n\lambda}(\theta_0) \right\}^{-1} \left\{ n A_{n\lambda}(\theta_0) \theta_0 - \frac{\partial \ell_n(\rho_0, \beta_0, \gamma_0)}{\partial \theta} \right\}.$$

Secondly, based on the log-likelihood function (4) we can obtain the score functions

$$U(\theta) = \frac{\partial \ell_n(\rho, \beta, \gamma)}{\partial \theta} = (U_1^T(\rho), U_2^T(\beta), U_3^T(\gamma))^T,$$

where $U_1(\rho) = \frac{\partial \ell_n(\rho, \beta, \gamma)}{\partial \rho} = Y^T W^T \Sigma^{-1} (SY - X\beta) - \text{tr}(G)$, $U_2(\beta) = \frac{\partial \ell_n(\rho, \beta, \gamma)}{\partial \beta} = X\Sigma^{-1} (SY - X^T \beta)$, $U_3(\gamma) = \frac{\partial \ell_n(\rho, \beta, \gamma)}{\partial \gamma} = -\frac{1}{2} \sum_{i=1}^n z_i + \frac{1}{2} \sum_{i=1}^n \frac{(\bar{y}_i - x_i^T \beta)^2}{\sigma_i^2} z_i$, $SY = (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_n)^T$. Denote

$$H(\theta) = \frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \theta \partial \theta^T} = \begin{pmatrix} \frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \rho \partial \rho^T} & \frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \rho \partial \beta^T} & \frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \rho \partial \gamma^T} \\ \frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \beta \partial \rho^T} & \frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \beta \partial \beta^T} & \frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \beta \partial \gamma^T} \\ \frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \gamma \partial \rho^T} & \frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \gamma \partial \beta^T} & \frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \gamma \partial \gamma^T} \end{pmatrix},$$

where $\frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \rho \partial \rho^T} = -Y^T W^T \Sigma^{-1} WY - \text{tr}(G^2)$, $\frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \beta \partial \beta^T} = -X\Sigma^{-1} X^T$, $\frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \gamma \partial \gamma^T} = -\frac{1}{2} \sum_{i=1}^n \frac{(\bar{y}_i - x_i^T \beta)^2}{\sigma_i^2} z_i z_i^T$, $\frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \beta \partial \rho^T} = -X^T \Sigma^{-1} WY$, $\frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \beta \partial \gamma^T} = -\sum_{i=1}^n \frac{(\bar{y}_i - x_i^T \beta)}{\sigma_i^2} x_i z_i^T$, $\frac{\partial^2 \ell_n(\rho, \beta, \gamma)}{\partial \gamma \partial \rho^T} = -\sum_{i=1}^n y_{w_i} e_i z_i / \sigma_i^2$, here, $Y_w = WY = (y_{w_1}, y_{w_2}, \dots, y_{w_n})^T$. Finally, we give the following Algorithm 1, which can summarize the computation of penalized quasi-maximum likelihood estimators of the parameters in the heterogeneous SAR Models.

Algorithm 1

Step 1. The ordinary quasi-maximum likelihood estimators (without penalty) $\beta^{(0)}, \gamma^{(0)}, \rho^{(0)}$ of β, γ, ρ are taken as their initial values.

Step 2. $\theta^{(m)} = \{\rho^{(m)}, \beta^{(m)T}, \gamma^{(m)T}\}^T$ are given as current values, then update them by

$$\theta^{(m+1)} = \theta^{(m)} + \left\{ H(\theta^{(m)}) - nA_{n\lambda}(\theta^{(m)}) \right\}^{-1} \left\{ nA_{n\lambda}(\theta^{(m)})\theta^{(m)} - U(\theta^{(m)}) \right\}.$$

Step 3. Repeat Step 2 above until $|\theta^{(m+1)} - \theta^{(m)}| < \epsilon$, where ϵ is a given small number, such as 10^{-5} .

4.2. Choosing the Tuning Parameters

The penalty function $p_{\lambda_n^{(l)}}(\cdot)$ involves the tuning parameters $\lambda_n^{(l)}$ ($l = 1, 2$) that control the amount of the penalty. Many selection criteria, such as CV, GCV, and BIC selection can be used to select the tuning parameters. Wang et al. [28] suggested using a BIC for the SCAD estimator in linear models and partially linear models and proved that its model selection consistency property, i.e., the optimal parameter chosen by the BIC, can identify the true model with the probability tending to one. Hence, their suggestion will be adopted in this paper. Nevertheless, in real application, how to simultaneously select a total of $p + q$ shrinkage parameters $\{\lambda_{ni}, i = 1, \dots, p + q\}$ is challenging. To bypass this difficulty, we follow the idea of Li et al. [18], and simplify the tuning parameters as follows,

- (i) $\lambda_{nj}^{(1)} = \frac{\lambda_n}{|\bar{\beta}_j^{(0)}|}, j = 1, \dots, p;$
- (ii) $\lambda_{nk}^{(2)} = \frac{\lambda_n}{|\bar{\gamma}_k^{(0)}|}, k = 1, \dots, q$

where $\tilde{\beta}_j^{(0)}$ and $\tilde{\gamma}_k^{(0)}$ are, respectively, the j th element and k th element of the unpenalized estimates $\tilde{\beta}^{(0)}$ and $\tilde{\gamma}^{(0)}$. Consequently, the original $p + q$ dimensional problem about λ_{ni} becomes a one-dimensional problem about λ_n . λ_n can be selected according to the following BIC-type criterion

$$BIC_{\lambda_n} = -\frac{2}{n} \ell_n(\hat{\rho}, \hat{\beta}, \hat{\gamma}) + df_{\lambda} \times \frac{\log(n)}{n},$$

where $0 \leq df_{\lambda} \leq s$ is simply the number of nonzero coefficients of $\hat{\theta}_{\lambda_n}$, and, here, $\hat{\theta}_{\lambda_n} = (\hat{\rho}, \hat{\beta}^T, \hat{\gamma}^T)^T$ is the estimate of θ for a given λ_n .

The tuning parameter can be obtained as

$$\hat{\lambda}_n = \arg \min_{\lambda_n} BIC_{\lambda_n}.$$

From our simulation studies, we found that this method works well.

5. Simulation Study

In this section we conduct a simulation study to assess the small sample performance of the proposed procedure. In this simulation, we choose m to be 5 and R to be 30, 40, 60, thus $n = R \times m$ to be 150, 200, and 300. X and Z are, respectively, generated from multivariate normal distribution with zero mean vector and covariance matrix being $\Sigma_0 = (c_{ij})$ where $c_{ij} = 0.5^{|i-j|}$, $i = 1, \dots, 8$, $j = 1, \dots, 8$. Besides, let the spatial parameter $\rho = 0.5, 0, -0.5$, which represents different spatial dependencies; $\beta = (3, 1.8, 2.2, 0, 0, 0, 0, 0)^T$ and the structure of the variance model is $\log(\sigma_i^2) = z_i^T \gamma$ with $\gamma = (1, 1, 1, 0, 0, 0, 0, 0)^T$. In these simulation, the weight matrix is taken to be $W = I_R \otimes H_m$, $H_m = (l_m l_m^T - I_m) / (m - 1)$, where l_m is an m -dimensional vector with all elements being 1, $\otimes$ means Kronecker product.

We generated $M = 500$ random samples of size $n = 150, 200$, and 300 , respectively. For each random sample, the proposed variable-selection method, based on penalized quasi-maximum likelihood with SCAD and ALASSO penalty functions, is considered. The unknown tuning parameters $\lambda^{(l)}$, ($l = 1, 2$) for the penalty function are chosen by the BIC criterion in the simulation. The average number of the estimated zero coefficients with 500 simulation runs is reported in Tables 1–3. Note that “C” in the tables means the average number of zero regression coefficients that are correctly estimated as zero, and “IC” depicts the average number of non-zero regression coefficients that are erroneously set to zero. The performance of estimators $\hat{\beta}$, and $\hat{\gamma}$ is assessed by the mean square error (MSE), defined as

$$MSE(\hat{\beta}) = E(\hat{\beta} - \beta_0)^T (\hat{\beta} - \beta_0), MSE(\hat{\gamma}) = E(\hat{\gamma} - \gamma_0)^T (\hat{\gamma} - \gamma_0).$$

The simulation results are reported in Tables 1–3.

Table 1. Variable selections for β and γ under different sample sizes when $\rho = 0.5$.

β	n	SCAD			ALASSO		
		MSE	C	IC	MSE	C	IC
γ	150	0.0043	4.9360	0	0.0041	4.9760	0
	200	0.0029	4.9480	0	0.0027	4.9900	0
	300	0.0016	4.9760	0	0.0015	4.9980	0
	n	MSE	C	IC	MSE	C	IC
	150	0.0903	4.8960	0.0040	0.0970	4.9000	0.0040
	200	0.0633	4.9040	0	0.0631	4.9400	0
	300	0.0342	4.9580	0	0.0352	4.9900	0

From Tables 1–3, we can make the following observations: (i) as n increases, the performances of variable-selection procedures become better and better. For example, the values in the column labeled ‘MSE’ of β and γ become smaller and smaller; the values in the col-

umn labeled ‘C’ become closer and closer to the true number of zero regression coefficients in the models. (ii) Based on different spatial parameters, the results of variable selection are similar. (iii) Two different penalty functions (SCAD and ALASSO) are used in this paper, and both methods perform almost equally well. (iv) Under the same circumstances, the ‘MSE’ of the mean parameters is smaller than the variance parameters’, which are very common in parametric estimation due to the fact that lower-order moments are easier to estimate than higher-order moments’.

Table 2. Variable selections for β and γ under different sample sizes when $\rho = 0$.

β	n	SCAD			ALASSO		
		MSE	C	IC	MSE	C	IC
γ	150	0.0041	4.9180	0	0.0038	4.9700	0
	200	0.0027	4.9520	0	0.0026	4.9900	0
	300	0.0016	4.9760	0	0.0016	5.0000	0
	n	MSE	C	IC	MSE	C	IC
	150	0.0903	4.9040	0	0.0955	4.9080	0
	200	0.0632	4.9200	0	0.0645	4.9460	0
	300	0.0387	4.9260	0	0.0383	4.9760	0

Table 3. Variable selections for β and γ under different sample sizes when $\rho = -0.5$.

β	n	SCAD			ALASSO		
		MSE	C	IC	MSE	C	IC
γ	150	0.0046	4.9080	0	0.0044	4.9560	0
	200	0.0032	4.9240	0	0.0030	4.9800	0
	300	0.0017	4.9720	0	0.0017	5.0000	0
	n	MSE	C	IC	MSE	C	IC
	150	0.0919	4.8780	0	0.1033	4.8860	0
	200	0.0600	4.9220	0	0.0636	4.9500	0
	300	0.0363	4.9420	0	0.0372	4.9800	0

6. Real Data Analysis

In this section, the proposed variable selection method is used to analyze the Boston housing price data that has been analyzed by many authors, for example, Pace and Gilley [29], Su and Yang [30], and so on. The database can be found in the `spdep` library of R. The dataset contains 14 variables with 506 observations, and a detailed description of these variables is listed in Table 4.

Table 4. Description of the variables in Boston housing data.

Variables	Description
CRIM(X_1)	Per capita crime rate by town
ZN(X_2)	Proportion of residential land zoned for lots over 25,000 sq. ft.
INDUS(X_3)	Proportion of non-retail business acres per town
CHAS(X_4)	Charles River dummy variable (=1 if tract bounds river; 0 otherwise)
NOX(X_5)	Nitric-oxides concentration (parts per 10 million)
RM(X_6)	Average number of rooms per dwelling
AGE(X_7)	Proportion of owner-occupied units built prior to 1940
DIS(X_8)	Weighted distances to five Boston employment centres
RAD(X_9)	Index of accessibility to radial highways
TAX(X_{10})	Full-value property-tax rate per USD 10,000
PTRATIO(X_{11})	Pupil-teacher ratio by town
B(X_{12})	$1000(Bk - 0.63)^2$, where Bk is the proportion of blacks by town
LSTAT(X_{13})	% of the population with lower status
MEDV(Y)	Median value of owner-occupied homes in USD 1000s

This dataset has been used by Pace and Gilley [29] on the basis of spatial econometric models, and longitude–latitude coordinates for the census tracts are added to the dataset. In this paper, we take MEDV as the response variable, and the other 13 variables in Table 4 are treated as explanatory variables. Similar with Pace and Gilley [29] as well as Su and Yang [30], we use the Euclidean distances in terms of longitude and latitude to generate weight matrix $W = (w_{ij})$, where

$$w_{ij} = \max\left(1 - \frac{d_{ij}}{d_0}, 0\right),$$

d_{ij} is the Euclidean distance and d_0 is the threshold distance, which is set to be 0.05 as in Su and Yang [30]. Thus, a spatial weight matrix is used with 19.1% nonzero elements. In addition, Z-variables in the variance model are taken to be the same as X-variables in the mean model. Then, the heterogeneous SAR models are considered here as follows:

$$\begin{cases} Y_i = \rho \sum_{j=1}^n w_{ij} Y_j + \sum_{k=1}^{13} X_{ik} \beta_k + \varepsilon_i, \\ \sigma_i^2 = \exp\left(\sum_{k=1}^{13} Z_{ik} \gamma_k\right), \\ i = 1, 2, \dots, 506. \end{cases} \quad (7)$$

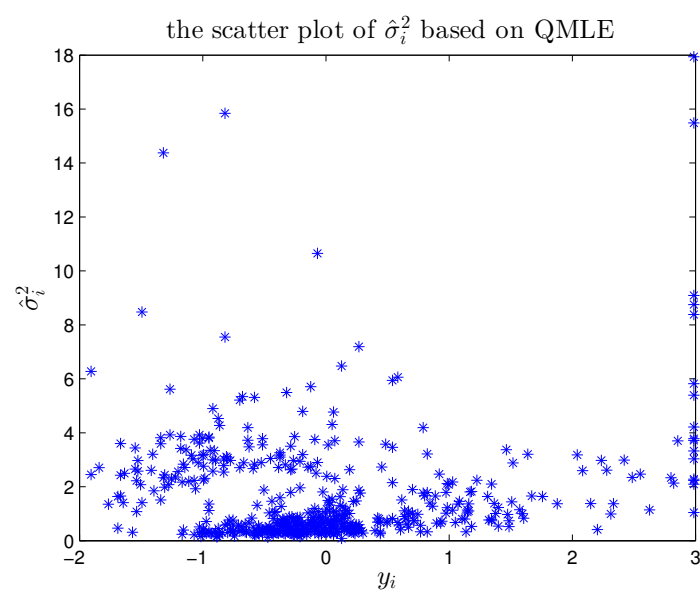
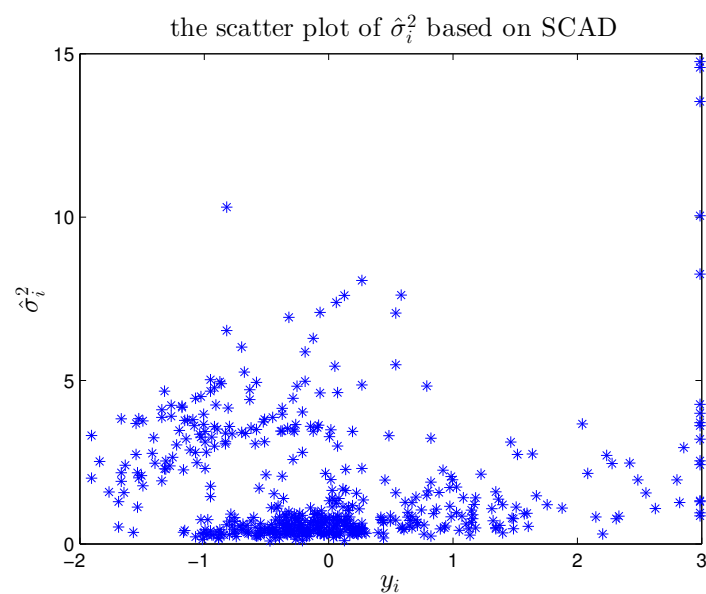
The ordinary quasi-maximum likelihood estimators (QMLE) and the penalized quasi-maximum likelihood estimators using the SCAD and ALASSO penalty functions are all considered. The tuning parameter was selected by the BIC. The estimated spatial parameter and the regression coefficients by different penalty-estimation methods are presented in Table 5. From Table 5, we can clearly see the following facts. (i) As expected, the spatial parameters are estimated very close to each other, and the other estimates are slightly different among different methods. (ii) Both the SCAD and ALASSO methods can eliminate many unimportant variables in joint mean and variance models. Concretely, the ALASSO method can select the same variables with the SCAD method in the mean model, and the SCAD method can select two more variables than the ALASSO method in the variance model. (iii) The important explanatory variables, selected by the proposed methods, are basically consistent with existing research results, for example, the regression coefficient of X_{11} is negative in the mean model, which reveals that the housing price would decrease as the pupil–teacher ratio increases. In addition, based on the estimate of γ 's, we can obtain the estimates of σ_i^2 based on different methods and list the scatter plot of $\hat{\sigma}_i^2$ in Figures 1–3, which shows that the heteroscedasticity modeling for this dataset is reasonable.

Table 5. Penalized quasi-maximum likelihood estimators for β and γ .

<i>Coefficient</i>	QMLE	SCAD	ALASSO
$\beta_1(X_1)$	−0.0882	0	0
$\beta_2(X_2)$	0.0699	0	0
$\beta_3(X_3)$	−0.0427	0	0
$\beta_4(X_4)$	0.0038	0	0
$\beta_5(X_5)$	−0.0145	0	0
$\beta_6(X_6)$	0.4641	0.5709	0.6023
$\beta_7(X_7)$	−0.1509	−0.1788	−0.1489
$\beta_8(X_8)$	−0.2076	−0.1287	−0.1175
$\beta_9(X_9)$	0.2008	0.1440	0.0502
$\beta_{10}(X_{10})$	−0.2031	−0.2049	−0.1502
$\beta_{11}(X_{11})$	−0.1073	−0.0750	−0.0829
$\beta_{12}(X_{12})$	0.1440	0.1638	0.1033
$\beta_{13}(X_{13})$	−0.0597	0	0
$\gamma_1(Z_1)$	0.1265	0	0
$\gamma_2(Z_2)$	0.1020	0	0

Table 5. Cont.

Coefficient	QMLE	SCAD	ALASSO
$\gamma_3(Z_3)$	−0.0830	0	0
$\gamma_4(Z_4)$	0.0823	0	0
$\gamma_5(Z_5)$	−0.4000	−0.3673	0
$\gamma_6(Z_6)$	0.1113	0	0
$\gamma_7(Z_7)$	0.3637	0.3461	0.2357
$\gamma_8(Z_8)$	−0.4004	−0.3877	−0.2381
$\gamma_9(Z_9)$	0.7366	0.8952	0.9143
$\gamma_{10}(Z_{10})$	0.2920	0.2105	0
$\gamma_{11}(Z_{11})$	−0.3533	−0.4085	−0.2723
$\gamma_{12}(Z_{12})$	−0.0141	0	0
$\gamma_{13}(Z_{13})$	−0.4262	−0.4586	−0.3929
ρ	0.2531	0.2881	0.2809

Figure 1. The scatter plot of $\hat{\sigma}_i^2$ based on the ordinary quasi-maximum likelihood estimators (QMLE).Figure 2. The scatter plot of $\hat{\sigma}_i^2$ based on the penalized quasi-maximum likelihood using SCAD.

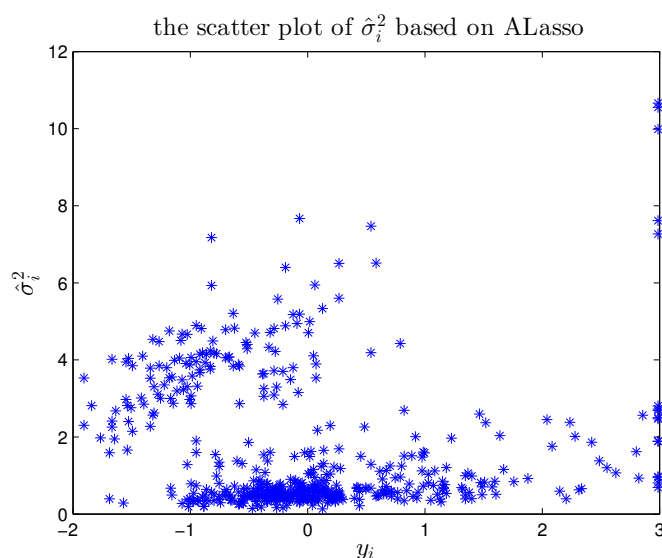


Figure 3. The scatter plot of $\hat{\sigma}_i^2$ based on the penalized quasi-maximum likelihood using ALASSO.

7. Conclusions and Discussion

Within the framework of heterogeneous spatial autoregressive models, we proposed a variable-selection method on the basis of a penalized quasi-maximum likelihood approach. Like the mean, the variance parameters may be dependent on various explanatory variables of interest, so that simultaneous variable selection to the mean and variance models becomes important to avoid the modeling biases and reduce the model complexities. We have proven that the proposed penalized quasi-maximum likelihood estimators of the parameters in the mean and variance models are asymptotically consistent and normally distributed under some mild conditions. Simulation studies and a real data analysis of the Boston housing data are conducted to illustrate the proposed methodology. The results show that the proposed variable selection method is highly efficient and computationally fast.

Furthermore, there are several interesting issues that merit further research. For example, (i) it is interesting to increase the model flexibility by introducing nonparametric functions in the context of spatial autoregressive model and to study the variable selection for both the parametric component and the nonparametric component; (ii) a possible extension of the heterogeneous spatial autoregressive models is considered when the response variables are missing under a different missingness mechanism; and (iii) in the penalized estimation, we can also penalize the spatial parameter just like regression coefficients, which can help us directly judge whether the analyzed data have a spatial structure. These are all topics of interest and worthy of further study.

Author Contributions: Conceptualization, R.T. and D.X.; methodology, R.T. and D.X.; software, M.X. and D.X.; data curation, R.T. and D.X.; formal analysis, M.X. and D.X.; writing—original draft, R.T. and D.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Statistical Science Research Project (2021LY061), the Startup Foundation for Talents at Hangzhou Normal University (2019QDL039), and the Statistical Science Research Project of Zhejiang Province (in 2022).

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. Proof of Theorems

To prove the theorems in the paper, we require the following regular conditions:

- C1. The $\{e_i\}, i = 1, \dots, n$ are independent with $E(e_i) = 0$ and $var(e_i) = \sigma_i^2$. The moment $E(|e_i|^{4+v})$ exists for some $v > 0$.
- C2. The elements $w_{n,ij} = O(1/h_n)$ and $w_{n,ii} = 0, i, j = 1, \dots, n$ are in W , where $h_n/n \rightarrow 0$ as $n \rightarrow \infty$.
- C3. The matrix S is a nonsingular matrix.
- C4. The sequences of matrices $\{W\}$ and $\{S^{-1}\}$ are uniformly bounded in both row and column sums.
- C5. The $\lim_{n \rightarrow \infty} n^{-1} X^T \Sigma^{-1} X$ and $\lim_{n \rightarrow \infty} n^{-1} Z^T Z$ exist and are nonsingular. The elements of X and Z are uniformly bounded constants for all n .
- C6. $S^{-1}(\rho)$ are uniformly bounded in row or column sums, uniformly in ρ in a closed subset Λ of $(-1, 1)$. The true ρ_0 is an interior point of Λ .
- C7. The $\lim_{n \rightarrow \infty} (X, GX\beta_0)^T \Sigma^{-1} (X, GX\beta_0)$ exists and is a nonsingular matrix.
- C8. The $\lim_{n \rightarrow \infty} E(n^{-1}(\partial^2 \ell_n(\theta_0))/\partial \theta \partial \theta^T)$ exists.
- C9. The third derivatives $(\partial^3 \mathcal{L}(\theta)/\partial \theta_j \partial \theta_l \partial \theta_m)$ exist for all θ in an open set Θ that contains the true parameter point θ_0 . Furthermore, there exist functions M_{jlm} such that $|n^{-1} \partial^3 \ell_n(\theta)/\partial \theta_j \partial \theta_l \partial \theta_m| \leq M_{jlm}$ for all $\theta \in \Theta$, where $E(M_{jlm}) < \infty$ for j, l, m .
- C10. The penalty function satisfies

$$\liminf_{n \rightarrow \infty} \liminf_{\theta_j \rightarrow 0^+} \lambda_{nj}^{-1} p'_{\lambda_{nj}}(|\theta_j|) > 0, \quad j = 2, \dots, s.$$

Remarks: Conditions 1–8 are sufficient conditions for the correctness of the global identification, asymptotic normality, and the consistent estimation of the QMLE of the model (3), which are similar to the conditions provided by Lee [27] and Liu et al. [5]. Concretely, Condition 1 is applied in the use of the central-limit theorem of Kelejian and Prucha [31]. Condition 2 shows the features of the weight matrix. If $\{h_n\}$ is a bounded sequence, Condition 2 holds. In Case's model (Case [32]), Condition 2 is still satisfied although h_n might diverge to infinity. Condition 3 is used to guarantee the existence of mean and variance of independent variable. Condition 4 implies that the variance of Y is bounded when n tends to infinity, see Kelejian and Prucha [31] as well as Lee [27]. Condition 5 can exclude the multicollinearity of the regressors of X and Z . For convenient analysis, we assume that the regressors are uniformly bounded. If not, they can be replaced by stochastic regressors with certain finite-moment conditions (Lee [27]). Condition 6 is meaningful to deal with the nonlinearity of $\ln|S(\rho)|$ in a log-likelihood function. Conditions 7–8 are used for the asymptotic normality of the QMLE. Condition 9 plays an important role in the Taylor expansion of related functions, which is similar to the condition (C) provided by Fan and Li [22]. Condition 10 is an assumption on the penalty function.

In order to prove Theorems 1 and 2, we need for the log-likelihood function $\ell_n(\rho, \beta, \gamma)$ to have several properties, which are stated in the following Lemmas.

Lemma A1. Suppose that Conditions 1–9 hold, then we can obtain

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta_0)}{\partial \theta} = O_p(1).$$

Proof of Lemma A1. By calculating the first-order partial derivatives of the log-likelihood function at θ_0 , we obtain

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta_0)}{\partial \rho} &= \frac{1}{\sqrt{n}} (G_n X \beta)^T \Sigma^{-1} e + \frac{1}{\sqrt{n}} e^T G_n \Sigma^{-1} e - \frac{1}{\sqrt{n}} \text{tr}(G_n), \\ \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta_0)}{\partial \beta} &= \frac{1}{\sqrt{n}} X^T \Sigma^{-1} e, \\ \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta_0)}{\partial \gamma} &= -\frac{1}{2\sqrt{n}} \sum_{i=1}^n \frac{e_i^2}{\sigma_i^2} Z_i, \end{aligned}$$

The variance of $\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta_0)}{\partial \rho}$ is

$$\begin{aligned} & \text{var}\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta_0)}{\partial \rho}\right) \\ &= \text{var}\left(\frac{1}{\sqrt{n}} (G_n X \beta)^T \Sigma^{-1} e + \left(\frac{1}{\sqrt{n}} e^T G_n \Sigma^{-1} e - \frac{1}{\sqrt{n}} \text{tr}(G_n)\right)\right) \\ &\leq \frac{2}{n} [\text{var}((G_n X \beta)^T \Sigma^{-1} e) + \text{var}(e^T G_n \Sigma^{-1} e - \text{tr}(G_n))] \\ &= \frac{2}{n} (G_n X \beta)^T (G_n X \beta) + \frac{2}{n} \text{var}(e^T G_n \Sigma^{-1} e) \\ &= O(1) + O\left(\frac{1}{h_n}\right) = O(1). \end{aligned}$$

Thus, $\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta_0)}{\partial \rho} = O_p(1)$. When the elements of X and Z are uniformly bounded for all n , it is obvious that $\frac{1}{\sqrt{n}} X^T \Sigma^{-1} e = O_p(1)$. In addition, by some elementary calculations, we have the variance of $\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta_0)}{\partial \gamma}$ is

$$\begin{aligned} \text{var}\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta_0)}{\partial \gamma}\right) &= \frac{1}{4n} \sum_{i=1}^n \text{var}(e_i^2) \frac{1}{\sigma_i^4} Z_i^T Z_i \\ &= \frac{1}{4n} \sum_{i=1}^n \frac{1}{\sigma_i^4} (\mu_{4i} - \sigma_i^4) = O(1), \end{aligned}$$

where $\mu_{4i} = E(e_i^4)$, therefore, $\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta_0)}{\partial \gamma} = O_p(1)$. Thus, the proof of the Lemma A1 is completed. $\square$

Lemma A2. If Conditions 1–8 hold, then

$$\frac{1}{n} \frac{\partial^2 \ell_n(\theta_0)}{\partial \theta \partial \theta^T} = E\left(\frac{1}{n} \frac{\partial^2 \ell_n(\theta_0)}{\partial \theta \partial \theta^T}\right) + o_p(1).$$

Proof of Lemma A2. The proof of Lemma A2 is similar to the proof of Theorem 3.2 (Lee [27]), so we omit the details. $\square$

Proof of Theorem 1. Let $\alpha_n = n^{-1/2} + a_n$. Similar to the proof of Theorem 1 in Fan and Li [22], we just need to prove that for any given $\varepsilon > 0$, there exists a large constant C , such that

$$P\left\{\sup_{\|u\|=C} \mathcal{L}(\theta_0 + \alpha_n u) < \mathcal{L}(\theta_0)\right\} \geq 1 - \varepsilon. \quad (\text{A1})$$

Note that $p_{\lambda_n}(0) = 0$. By using the Taylor expansion of the log-likelihood function, we have

$$\begin{aligned}
D_n(u) &= \mathcal{L}(\theta_0 + \alpha_n u) - \mathcal{L}(\theta_0) \\
&= \ell_n(\theta_0 + \alpha_n u) - \ell_n(\theta_0) - \left\{ n \sum_{j=2}^s p_{\lambda_{nj}}(|\theta_{j0} + \alpha_n u_j|) - n \sum_{j=2}^s p_{\lambda_{nj}}(|\theta_{j0}|) \right\} \\
&\leq \alpha_n \left(\frac{\partial \ell_n(\theta_0)}{\partial \theta} \right)^T u - \frac{1}{2} u^T \mathcal{I}_n(\theta_0) u n \alpha_n^2 \{1 + o_p(1)\} - \\
&\quad \sum_{j=2}^{s_1} [n \alpha_n p'_{\lambda_{nj}}(|\theta_{j0}|) \text{sgn}(\theta_{j0}) u_j + n \alpha_n^2 p''_{\lambda_{nj}}(|\theta_{j0}|) u_j^2 \{1 + o(1)\}] \\
&= A_1 - A_2 - A_3,
\end{aligned}$$

where s_1 is the number of dimensions of $\theta_0^{(1)}$. It follows from Lemma A1 that

$$\begin{aligned}
A_1 &= \alpha_n \left(\frac{\partial \ell_n(\theta_0)}{\partial \theta} \right)^T u \leq \left\| \alpha_n \left(\frac{\partial \ell_n(\theta_0)}{\partial \theta} \right)^T \right\| \cdot \|u\| \\
&= \|u\| \cdot O_p(n \alpha_n^2).
\end{aligned}$$

Under Condition 8,

$$A_2 = \frac{1}{2} u^T \mathcal{I}_n(\theta_0) u n \alpha_n^2 \{1 + o_p(1)\} = \|u\|^2 \cdot O_p(n \alpha_n^2).$$

In addition,

$$\begin{aligned}
A_3 &= \sum_{j=2}^{s_1} [n \alpha_n p'_{\lambda_{nj}}(|\theta_{j0}|) \text{sgn}(\theta_{j0}) u_j + n \alpha_n^2 p''_{\lambda_{nj}}(|\theta_{j0}|) u_j^2 \{1 + o(1)\}] \\
&\leq \sum_{j=2}^{s_1} [n \alpha_n |p'_{\lambda_{nj}}(|\theta_{j0}|)| \cdot |u_j| + n \alpha_n^2 p''_{\lambda_{nj}}(|\theta_{j0}|) u_j^2 \{1 + o(1)\}] \\
&\leq \sqrt{s_1} n \alpha_n a_n \|u\| + n \alpha_n^2 \max\{p''_{\lambda_{nj}}(|\theta_{j0}|) : \theta_{j0} \neq 0\} \|u\|^2 \{1 + o(1)\} \\
&= \|u\| \cdot O_p(n \alpha_n^2) + \|u\|^2 \cdot o_p(n \alpha_n^2).
\end{aligned}$$

Furthermore, by choosing a sufficiently large C , A_1 and A_3 are dominated by A_2 uniformly in $\|u\| = C$. Hence, (A1) holds, which implies the proof of Theorem 1 is completed. $\square$

Proof of Theorem 2. We first prove part (i). According to Theorem 1, it is sufficient to show that, for any $\theta^{(1)}$ that satisfies $\|\theta^{(1)} - \theta_0^{(1)}\| = O_p(n^{-1/2})$ and some given small $\epsilon_n = Cn^{-1/2}$, when $n \rightarrow \infty$, with probability tending to 1, we have

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} < 0, \quad \text{for } 0 < \theta_j < \epsilon_n, \quad j = s_1 + 1, \dots, s, \quad (\text{A2})$$

and

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} > 0, \quad \text{for } -\epsilon_n < \theta_j < 0, \quad j = s_1 + 1, \dots, s. \quad (\text{A3})$$

By the Taylor expansion, it is easy to prove that

$$\begin{aligned}
\frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} &= \frac{\partial \ell_n(\theta)}{\partial \theta_j} - n p'_{\lambda_{nj}}(|\theta_j|) \text{sgn}(\theta_j) \\
&= \frac{\partial \ell_n(\theta_0)}{\partial \theta_j} + \sum_{l=1}^s \frac{\partial^2 \ell_n(\theta_0)}{\partial \theta_j \partial \theta_l} (\theta_l - \theta_{l0}) \\
&\quad + \sum_{l=1}^s \sum_{m=1}^s \frac{\partial^3 \ell_n(\theta^*)}{\partial \theta_j \partial \theta_l \partial \theta_m} (\theta_l - \theta_{l0}) (\theta_m - \theta_{m0}) - n p'_{\lambda_{nj}}(|\theta_j|) \text{sgn}(\theta_j)
\end{aligned}$$

where θ^* lies between θ and θ_0 . From Lemmas A1–A2 and Condition 9, we obtain $\frac{1}{n} \frac{\partial \ell_n(\theta_0)}{\partial \theta_j} = O_p(n^{-1/2})$,

$$\begin{aligned} \frac{1}{n} \frac{\partial^2 \ell_n(\theta_0)}{\partial \theta \partial \theta^T} &= E \left(\frac{1}{n} \frac{\partial^2 \ell_n(\theta_0)}{\partial \theta \partial \theta^T} + o_p(1) \right), \\ \frac{1}{n} \frac{\partial^3 \ell_n(\theta^*)}{\partial \theta_j \partial \theta_l \partial \theta_m} &= O_p(1). \end{aligned}$$

When $\|\theta^{(1)} - \theta_0^{(1)}\| = O_p(n^{-1/2})$, we have

$$\begin{aligned} \frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} &= n\lambda_n \left\{ \lambda_n^{-1} \frac{1}{n} \frac{\partial \ell_n(\theta_0)}{\partial \theta_j} + \lambda_n^{-1} \sum_{l=1}^s \frac{1}{n} \frac{\partial^2 \ell_n(\theta_0)}{\partial \theta_j \partial \theta_l} (\theta_l - \theta_{l0}) \right. \\ &\quad \left. + \lambda_n^{-1} \sum_{l=1}^s \sum_{m=1}^s \frac{\partial^3 \ell_n(\theta^*)}{\partial \theta_j \partial \theta_l \partial \theta_m} (\theta_l - \theta_{l0}) (\theta_m - \theta_{m0}) - \lambda_n^{-1} p'_{\lambda_n}(|\theta_j|) \text{sgn}(\theta_j) \right\} \\ &= n\lambda_n [\lambda_n^{-1} O_p(n^{-1/2}) + \lambda_n^{-1} O_p(n^{-1/2}) + \lambda_n^{-1} O_p(n^{-1}) - \lambda_n^{-1} p'_{\lambda_n}(|\theta_j|) \text{sgn}(\theta_j)] \\ &= n\lambda_n [-\lambda_n^{-1} p'_{\lambda_n}(|\theta_j|) \text{sgn}(\theta_j) + O_p(n^{-1/2} \lambda_n^{-1})]. \end{aligned}$$

Note that $\liminf_{n \rightarrow \infty} \liminf_{\delta \rightarrow 0^+} \lambda_n^{-1} p'_{\lambda_n}(\delta) > 0$ and $\lim_{n \rightarrow \infty} n^{-1/2} \lambda_n^{-1} = 0$. The sign of the derivative is determined by that of θ_j with a sufficiently large n . This shows that (A2) and (A3) follow. This completes the proof of part (i).

Next, we prove part (ii). By Theorem 1, there is a $\sqrt{n}$ consistent local maximizer of $\mathcal{L}_n(\theta^{(1)T}, 0^T)$, denoted as $\hat{\theta}^{(1)}$, which satisfies

$$\left. \frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} \right|_{\theta = (\theta^{(1)T}, 0^T)^T} = 0, \quad \text{for } j = 1, \dots, s.$$

For $j \geq 2$, note that

$$\begin{aligned} \frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} &= \frac{\partial \ell_n(\theta)}{\partial \theta_j} - n p'_{\lambda_{nj}}(|\theta_j|) \text{sgn}(\theta_j) \\ &= \frac{\partial \ell_n(\theta_0)}{\partial \theta_j} + \sum_{l=1}^s \left\{ \frac{\partial^2 \ell_n(\theta_0)}{\partial \theta_j \partial \theta_l} + o_p(1) \right\} (\theta_l - \theta_{l0}) \\ &\quad - n [p'_{\lambda_{nj}}(|\theta_{j0}|) \text{sgn}(\theta_{j0}) + (p''_{\lambda_{nj}}(|\theta_{j0}|) + o_p(1)) (\theta_j - \theta_{j0})] \end{aligned}$$

Hence, we have

$$\begin{aligned} \frac{\partial \ell_n(\theta_0)}{\partial \theta_j} &= \sum_{l=1}^s \left\{ \frac{\partial^2 \ell_n(\theta_0)}{\partial \theta_j \partial \theta_l} + o_p(1) \right\} (\hat{\theta}_l - \theta_{l0}) \\ &\quad + n [p'_{\lambda_n}(|\theta_{j0}|) \text{sgn}(\theta_{j0}) + (p''_{\lambda_n}(|\theta_{j0}|) + o_p(1)) (\hat{\theta}_j - \theta_{j0})] \end{aligned}$$

where $\hat{\theta}_l$ is the l th element of $\hat{\theta}_1^{(1)}$.

Let $\Omega = \left(-\frac{1}{n} \frac{\partial^2 \ell_n(\theta_0)}{\partial \theta_j \partial \theta_l} \right)_{s_1 \times s_1}$, $B = [I_{s_1 \times s_1}, 0_{s_1 \times (s-s_1)}]$, then

$$\begin{aligned}
B \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta)}{\partial \theta} &= \Omega \sqrt{n}(\hat{\theta}^{(1)} - \theta_0^{(1)}) + A_n \sqrt{n}(\hat{\theta}^{(1)} - \theta_0^{(1)}) + \sqrt{n}b_n + o_p(1) \\
&= [\Omega - \mathcal{I}_n^{(1)}(\theta_0^{(1)})] \sqrt{n}(\hat{\theta}^{(1)} - \theta_0^{(1)}) + \mathcal{I}_n^{(1)}(\theta_0^{(1)}) \sqrt{n}(\hat{\theta}^{(1)} - \theta_0^{(1)}) \\
&\quad + A_n \sqrt{n}(\hat{\theta}^{(1)} - \theta_0^{(1)}) + \sqrt{n}b_n + o_p(1) \\
&= \mathcal{I}_n^{(1)}(\theta_0^{(1)}) \sqrt{n}(\hat{\theta}^{(1)} - \theta_0^{(1)}) + A_n \sqrt{n}(\hat{\theta}^{(1)} - \theta_0^{(1)}) + \sqrt{n}b_n + o_p(1) \\
&= \sqrt{n} \left[(\mathcal{I}_n^{(1)}(\theta_0^{(1)}) + A_n)(\hat{\theta}^{(1)} - \theta_0^{(1)}) + b_n \right] + o_p(1).
\end{aligned}$$

Furthermore, by using the central-limit theorem for the linear-quadratic forms of Kelejian and Prucha [31], it has

$$B \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta_0)}{\partial \theta} \xrightarrow{\mathcal{L}} N(0, (\mathcal{I}_1^{(1)}(\theta_0^{(1)}) + \mathcal{J}_1(\theta_0^{(1)})).$$

Therefore, according to Slutsky's theorem, we obtain

$$\sqrt{n} \left[(\mathcal{I}_n^{(1)}(\theta_0^{(1)}) + A_n)(\hat{\theta}^{(1)} - \theta_0^{(1)}) + B_n \right] \xrightarrow{\mathcal{L}} N(0, (\mathcal{I}_1^{(1)}(\theta_0^{(1)}) + \mathcal{J}_1(\theta_0^{(1)})).$$

The proof of Theorem 2 is, hence, completed. $\square$

References

1. Cliff, A.; Ord, J.K. *Spatial Autocorrelation*; Pion: London, UK, 1973.
2. Anselin, L. *Spatial Econometrics: Methods and Models*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1988.
3. Anselin, L.; Bera, A.K. Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics. In *Handbook of Applied Economics Statistics*; Ullah, A., Giles, D.E.A., Eds.; Marcel Dekker: New York, NY, USA, 1998.
4. Xu, X.B.; Lee, L.F. A spatial autoregressive model with a nonlinear transformation of the dependent variable. *J. Econom.* **2015**, *186*, 1–18. [\[CrossRef\]](#)
5. Liu, X.; Chen, J.B.; Cheng, S.L. A penalized quasi-maximum likelihood method for variable selection in the spatial autoregressive model. *Spat. Stat.* **2018**, *25*, 86–104. [\[CrossRef\]](#)
6. Xie, L.; Wang, X.R.; Cheng, W.H.; Tang, T. Variable selection for spatial autoregressive models. *Commun.-Stat. Methods* **2021**, *50*, 1325–1340. [\[CrossRef\]](#)
7. Xie, T.F.; Cao, R.Y.; Du, J. Variable selection for spatial autoregressive models with a diverging number of parameters. *Stat. Pap.* **2020**, *61*, 1125–1145. [\[CrossRef\]](#)
8. Su, L.J.; Jin, S.N. Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. *J. Econom.* **2010**, *157*, 18–33. [\[CrossRef\]](#)
9. Du, J.; Sun, X.Q.; Cao, R.Y.; Zhang, Z.Z. Statistical inference for partially linear additive spatial autoregressive models. *Spat. Stat.* **2018**, *25*, 52–67. [\[CrossRef\]](#)
10. Cheng, S.L.; Chen, J.B. Estimation of partially linear single-index spatial autoregressive model. *Stat. Pap.* **2021**, *62*, 485–531. [\[CrossRef\]](#)
11. Wei, C.H.; Guo, S.; Zhai, S.F. Statistical inference of partially linear varying coefficient spatial autoregressive models. *Econ. Model.* **2017**, *64*, 553–559. [\[CrossRef\]](#)
12. Hu, Y.P.; Wu, S.Y.; Feng, S.Y.; Jin, J.L. Estimation in Partial Functional Linear Spatial Autoregressive Model. *Mathematics* **2020**, *8*, 1680. [\[CrossRef\]](#)
13. Lin, X.; Lee, L.F. GMM estimation of spatial autoregressive models with unknown heteroskedasticity. *J. Econom.* **2010**, *157*, 34–52. [\[CrossRef\]](#)
14. Dai, X.W.; Jin, L.B.; Tian, M.Z.; Shi, L. Bayesian Local Influence for Spatial Autoregressive Models with Heteroscedasticity. *Stat. Pap.* **2019**, *60*, 1423–1446. [\[CrossRef\]](#)
15. Wu, L.C.; Li, H.Q. Variable selection for joint mean and dispersion models of the inverse Gaussian distribution. *Metrika* **2012**, *75*, 795–808. [\[CrossRef\]](#)
16. Xu, D.K.; Zhang, Z.Z. A semiparametric Bayesian approach to joint mean and variance models. *Stat. Probab. Lett.* **2013**, *83*, 1624–1631. [\[CrossRef\]](#)
17. Zhao, W.H.; Zhang, R.Q.; Lv, Y.Z.; Liu, J.C. Variable selection for varying dispersion beta regression model. *J. Appl. Stat.* **2014**, *41*, 95–108. [\[CrossRef\]](#)

18. Li, H.Q.; Wu, L.C.; Ma, T. Variable selection in joint location, scale and skewness models of the skew-normal distribution. *J. Syst. Sci. Complex.* **2017**, *30*, 694–709. [[CrossRef](#)]
19. Zhang, D.; Wu, L.C.; Ye, K.Y.; Wang, M. Bayesian quantile semiparametric mixed-effects double regression models. *Stat. Theory Relat. Fields* **2021**, *5*, 303–315. [[CrossRef](#)]
20. Tibshirani, R. Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [[CrossRef](#)]
21. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [[CrossRef](#)]
22. Fan, J.Q.; Li, R.Z. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [[CrossRef](#)]
23. Luo, G.; Wu, M. Variable selection for semiparametric varying-coefficient spatial autoregressive models with a diverging number of parameters. *Commun.-Stat. Methods* **2021**, *50*, 2062–2079. [[CrossRef](#)]
24. Li, R.; Liang, H. Variable selection in semiparametric regression modeling. *Ann. Stat.* **2008**, *36*, 261–286. [[CrossRef](#)] [[PubMed](#)]
25. Zhao, P.X.; Xue, L.G. Variable selection for semiparametric varying coefficient partially linear errors-in-variables models. *J. Multivar. Anal.* **2010**, *101*, 1872–1883. [[CrossRef](#)]
26. Tian, R.Q.; Xue, L.G.; Liu, C.L. Penalized quadratic inference functions for semiparametric varying coefficient partially linear models with longitudinal data. *J. Multivar. Anal.* **2014**, *132*, 94–110. [[CrossRef](#)]
27. Lee, L.F. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* **2004**, *72*, 1899–1925. [[CrossRef](#)]
28. Wang, H.; Li, R.; Tsai, C. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **2007**, *94*, 553–568. [[CrossRef](#)] [[PubMed](#)]
29. Pace, R.K.; Gilley, O.W. Using the spatial configuration of the data to improve estimation. *J. Real Estate Financ. Econ.* **1997**, *14*, 333–340. [[CrossRef](#)]
30. Su, L.; Yang, Z. *Instrumental Variable Quantile Estimation of Spatial Autoregressive Models*; Working Paper; Singapore Management University: Singapore, 2009.
31. 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](#)]
32. Case, A.C. Spatial patterns in household demand. *Econometrica* **1991**, *59*, 953–965. [[CrossRef](#)]