Annals of the Institute of Statistical Mathematics (2023) 75:949–977
https://doi.org/10.1007/s10463-023-00870-w

Robust variable selection with exponential squared loss for partially linear spatial autoregressive models

Xiuli Wang[1] · Jingchang Shao[1] · Jingjing Wu[2] · Qiang Zhao[1]

Received: 13 September 2022 / Revised: 12 January 2023 / Accepted: 24 February 2023 /
Published online: 3 May 2023

Abstract

In this paper, we consider variable selection for a class of semiparametric spatial autoregressive models based on exponential squared loss (ESL). Using the orthogonal projection technique, we propose a novel orthogonality-based variable selection procedure that enables simultaneous model selection and parameter estimation, and identifies the significance of spatial effects. Under appropriate conditions, we show that the proposed procedure is consistent and the resulting estimator has oracle properties. Furthermore, some simulation studies and an analysis of the Boston housing price data are also carried out to examine the finite-sample performance of the proposed method.

Keywords Orthogonal projection · Exponential squared loss · Semiparametric spatial autoregressive models · Oracle property · Variable selection

✉ Qiang Zhao
qzhao@sdnu.edu.cn

Xiuli Wang
wxlmath@163.com

Jingchang Shao
jcshao1998@163.com

Jingjing Wu
jinwu@ucalgary.ca

[1] School of Mathematics and Statistics, Shandong Normal University, No.1 University Road, Science Park, Changqing District, Jinan 250358, China

[2] Department of Mathematics and Statistics, University of Calgary, 2500 University Drive NW, Calgary Alberta T2N 1N4, Canada

1 Introduction

In recent years, spatial autoregressive models have received extensive attention in the literature of statistics and econometrics. These models can describe the spatial correlation effects of data, which extend traditional regression models by considering the spatially lagged terms of response variables. For parametric spatial autoregressive models, especially the linear spatial autoregressive models, due to their strong explanatory power and easier estimation, many researchers have proposed and studied different estimation methods, Ord (1975), Kelejian and Prucha (1998), Kelejian and Prucha (1999) and Lee (2004), among others. If the form of regression function can be specified correctly, the parametric spatial autoregressive model can provide more accurate and effective estimation and statistical inferences. However, sometimes the form of regression function may possibly be misspecified, which will result in inconsistent parameter estimation of the parametric spatial autoregressive model.

However, Basile (2009) suggested that the relationship between response variables and covariates in a spatial autoregressive model may not only be linear but also be possibly nonlinear. In order to capture the potential relationship between response variables and covariates, some semiparametric spatial autoregressive models have been proposed in recent literatures. For example, Su and Jin (2010) proposed a profile quasi-maximum likelihood estimation method for partially linear spatial autoregressive models. Using the approximation of nonparametric functions by basis functions, Du et al. (2018) proposed a generalized method of moment estimation for partially linear additive spatial autoregressive models. Cheng et al. (2019) studied the generalized moment estimation method for partial linear single-index spatial autoregressive models.

Furthermore, variable selection is an important topic for statistical inference of spatial autoregressive models. Including too many spurious covariates significantly reduces estimation efficiency, while ignoring any important covariates results in serious bias. For semiparametric spatial autoregressive models, there are many literatures on its variable selection problem. Li and Guo (2020) considered variable selection for a partially linear spatial autoregressive model based on a penalized quasi-maximum likelihood estimation procedure, which selects important explanatory variables and simultaneously estimates nonzero parameters. Luo and Wu (2021) considered variable selection for a class of semiparametric spatial autoregressive models and proposed a variable selection method based on a two-stage estimation. Li et al. (2020) proposed a penalized contour least squares method to address the variable selection problem in semiparametric spatial autoregressive models.

However, many classical variable selection methods are closely related to least squares. It is well known that the least squares method is very sensitive to outliers in finite samples, and thus many robust methods have been proposed. Wang et al. (2013) proposed a robust estimation method based on exponential squared loss (ESL) function $\phi_\gamma(t) = 1 - \exp(-t^2/\gamma)$. An obvious feature of this method is that it controls the robustness and effectiveness of an estimator by introducing an

adjustment parameter γ. Specifically, in the linear model $y_i = x_i^T\beta + \epsilon_i$, the parameter β can be estimated by minimizing

$$\sum_{i=1}^{n}\left\{1-\exp\left[-\left(y_i-x_i^T\beta\right)^2/\gamma\right]\right\}, \tag{1}$$

where $\gamma > 0$ controls the degree of robustness and efficiency. When γ is large, $1-\exp(-t^2/\gamma) \approx t^2/\gamma$ so that the proposed estimator is approximately the least squares estimator. When γ is small, observations with large $|t_i|$ values will have a smaller impact on the estimator, and thus smaller γ will reduce the influence of outliers on the estimation. Wang et al. (2013) proposed how to choose γ, and pointed out that compared with Huber's estimation, quantile regression estimator (Koenker and Bassett 1978) and composite quantile regression estimator (Zou and Yuan 2008), this ESL-based estimator has stronger robustness.

The ESL estimation method has been applied to some semiparametric models, such as Wang and Lin (2016), Jiang et al. (2017) and Jiang et al. (2019). Recently, Song et al. (2021) proposed a class of penalized robust regression estimators based on ESL for parametric spatial autoregressive models. However, as far as we know, no one has studied this method for semiparametric spatial autoregressive models. Therefore, we propose a variable selection method based on ESL by using QR decomposition technique. Among the variations of the QR decomposition technique, the one proposed by Zhao et al. (2021) can select important covariates in the parametric components independently without affecting the nonparametric components. Under appropriate regularization conditions, we show that the proposed procedure is consistent and the resulting estimators have oracle property. Finite numerical simulations show that our variable selection procedure can handle data with outliers and as what the theory indicates, the variable selection results are consistent with the oracle ones, which means that our proposed method is not only robust but also efficient.

The rest of this paper is organized as follows. In Sect. 2, we present a variable selection procedure for a partially linear semiparametric spatial autoregressive model. In Sect. 3, we derive the main asymptotic results of the estimation method under certain regularization conditions. Section 4 is devoted to developing algorithms for computation and tuning parameter selection. In Sect. 5, we conduct simulation studies to evaluate the performance of our proposed variable selection procedure. In Sect. 6, we illustrate the proposed method through the analysis of the Boston housing price data set. Section 7 is conclusion. Finally, we provide in Appendix the detailed technical proofs of all asymptotic results.

2 Methodology

Consider the following partially linear spatial autoregressive model

$$Y_i = \rho \sum_{j=1}^{n} w_{ij} Y_j + X_i^T \beta + g(U_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where Y_i denotes the dependent variable and X_i and U_i are exogenous covariates of the i-th individual, ρ is the scalar autoregressive parameter, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the vector of unknown parameters, w_{ij} is the specific spatial weight between subjects i and j, $g(\cdot)$ is an unknown nonparametric function, and the model error ϵ_i satisfies $E(\epsilon_i|X_i, U_i) = 0$ and $Var(\epsilon_i|X_i, U_i) = \sigma^2 < \infty$. In this paper, we assume that the parameter β is sparse, which implies that some components in X_i are insignificant covariates. In addition, to make sure model (1) is sensible, we also assume that not all covariates are insignificant covariates, which implies that some components in β are nonzero coefficients. The variable U_i is assumed to range over a nondegenerate compact interval and, without loss of generality, we assume U_i takes values in the unit interval $[0, 1]$. The spatial weights matrix W is constructed from geographical or economic information to characterize the spatial dependence in practice, which is often assumed a known matrix with zero diagonal elements and standardized rows, that is, $w_{ii} = 0$ and $\sum_{j=1}^{n} w_{ij} = 1$; see Su and Yang (2007) for more details.

Inspired by Song et al. (2021), we propose a robust and efficient variable selection estimation method using the ESL function for estimating $g(\cdot)$ and β by maximizing the objective function

$$\ell(g, \beta) = \sum_{i=1}^{n} \exp\left\{ -\left[Y_i - \rho \ddot{Y}_i - X_i^T \beta - g(U_i)\right]^2 \Big/ \gamma \right\} - n \sum_{j=1}^{p} p_{\lambda_j}\left(\left|\beta_j\right|\right), \tag{3}$$

where $\ddot{Y}_i = \sum_{j=1}^{n} w_{ij} Y_j$ and $\lambda_j > 0$ is the regularization parameter. The selection of tuning parameter γ will be discussed in Sect. 4.

Based on the B-spline approximation technique (see Schumaker 1981), we first use B-spline basis functions to approximate and replace $g(\cdot)$ in (1). We denote $B(u) = (B_1(u), B_2(u), \ldots, B_L(u))^T$ as the B-spline basis functions, where $L = K + \kappa + 1$, κ is the order of the B-spline basis functions and K is the number of interior knots. Then, the nonparametric function $g(u)$ can approximated by

$$g(u) \approx B(u)^T \eta, \tag{4}$$

where $\eta = (\eta_1, \eta_2, \ldots, \eta_L)^T$ is the vector of basis function coefficients. With this approximation, model (1) can be written as

$$Y_i \approx \rho \sum_{j=1}^{n} w_{ij} Y_j + X_i^T \beta + B(U_i)^T \eta + \epsilon_i, \quad i = 1, \ldots, n, \tag{5}$$

where $B(U_i) = ((B_1(U_i), B_2(U_i), \ldots, B_L(U_i))^T$. Denote $Y = (Y_1, Y_2, \ldots, Y_n)^T$, $X = (X_1, X_2, \ldots, X_n)^T$, $S = (B(U_1), B(U_2), \ldots, B(U_n))^T$ and $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$. Then, model (4) can be rewritten as in the matrix form

$$Y \approx \rho W Y + X\beta + S\eta + \epsilon. \tag{6}$$

Similar to Zhao et al. (2021), we assume that S is a $n \times L$ column full rank matrix. Then using QR decomposition method for processing, S can be decomposed as

$$S = Q \begin{pmatrix} R \\ 0 \end{pmatrix}, \tag{7}$$

where Q is a $n \times n$ orthogonal matrix, R is a $L \times L$ triangular matrix, and 0 is a $(n-L) \times L$ zero matrix. Let $Q = (q_1, q_2, \ldots, q_n)$, $Q_1 = (q_1, q_1, \ldots, q_L)$ and $Q_2 = (q_{L+1}, q_{L+2}, \ldots, q_n)$, then we have $S = Q_1 R$, $Q_2^T Q_1 = 0$ and further $Q_2^T S = Q_2^T Q_1 R = 0$. As a result, left multiplying both sides of model (5) by Q_2^T yields

$$Q_2^T Y \approx Q_2^T W Y \rho + Q_2^T X \beta + Q_2^T \epsilon. \tag{8}$$

Define $\tilde{Y}_i = q_i^T Y$, $\tilde{D}_i = q_i^T W Y$, $\tilde{X}_i = q_i^T X$ and $\tilde{\epsilon}_i = q_i^T \epsilon$ for $i = L+1, L+2, \ldots, n$, then model (7) can be rewritten as

$$\tilde{Y}_i \approx \tilde{D}_i \rho + \tilde{X}_i^T \beta + \tilde{\epsilon}_i, \quad i = L+1, L+2, \ldots, n. \tag{9}$$

Further let $\tilde{Y} = (\tilde{Y}_{L+1}, \tilde{Y}_{L+2}, \ldots, \tilde{Y}_n)^T$, $\tilde{Z}_i = (\tilde{D}_i, \tilde{X}_i^T)^T$, $\tilde{Z} = (\tilde{Z}_{L+1}, \tilde{Z}_{L+2}, \ldots, \tilde{Z}_n)^T$, $\tilde{\epsilon} = (\tilde{\epsilon}_{L+1}, \tilde{\epsilon}_{L+2}, \ldots, \tilde{\epsilon}_n)^T$ and $\theta = (\rho, \beta^T)^T$, then (8) can be rewritten as

$$\tilde{Y} \approx \tilde{Z}\theta + \tilde{\epsilon}. \tag{10}$$

Hence invoking model (9), the penalized ESL function can be rewritten as

$$\ell(\theta) = \sum_{i=L+1}^{n} \exp\left\{ -\left[\tilde{Y}_i - \tilde{Z}_i^T \theta\right]^2 \Big/ \gamma \right\} - (n-L) \sum_{j=1}^{p+1} p_{\lambda_j}(|\theta_j|), \tag{11}$$

where $p_{\lambda_j}(\ldots)$ is the adaptive LASSO (ALASSO) penalized function, and λ_j is the tuning parameter. Finally, the regularized estimator of $\theta = (\rho, \beta^T)^T$ is given by

$$\hat{\theta}_n = \arg\max_{\theta} \ell(\theta). \tag{12}$$

3 Large sample and oracle properties

For notational convenience and simplicity, let $\theta_0 = (\rho_0, \beta_0^T)^T$ be the true value of $\theta = (\rho, \beta^T)^T$ and γ_0 be the true value of γ. Without loss of generality, θ_0 can be partitioned as $\theta_0 = (\theta_{01}^T, \theta_{02}^T)^T$, where all the elements of θ_{01} are nonzeros while those of θ_{02} are all zeros. Similarly, θ can be partitioned accordingly as $\theta = (\theta_1^T, \theta_2^T)^T$ with $\theta_1 \in \mathbb{R}^s$ and $\theta_2 \in \mathbb{R}^{p+1-s}$. Accordingly, the estimator $\hat{\theta}_n$ given in (11) can be partitioned as $\hat{\theta}_n = \left(\hat{\theta}_{n1}^T, \hat{\theta}_{n2}^T\right)^T$. Let

$$a_n = \max\{p'_{\lambda_j}(|\theta_{0j}|) : \theta_{0j} \neq 0\},$$
$$b_n = \max\{p''_{\lambda_j}(|\theta_{0j}|) : \theta_{0j} \neq 0\}$$

and

$$I(\theta,\gamma)=\frac{2}{\gamma}\int \tilde{Z}\tilde{Z}^T\exp(-r^2/\gamma)\left(\frac{2r^2}{\gamma}-1\right)dF(\tilde{Z},\tilde{Y}),$$

where $r=\tilde{Y}-\tilde{Z}\theta$, $F(\tilde{Z},\tilde{Y})$ is the joint distribution function of $(\tilde{Y},\tilde{Z})$. Let $\tilde{Z}_i=(\tilde{Z}_{i1}^T,\tilde{Z}_{i2}^T)^T$, where $\tilde{Z}_{i1}$ and $\tilde{Z}_{i2}$ are the covariates corresponding to θ_1 and θ_2, respectively.

We assume the following regularity conditions:

Assumption 1. $\Sigma=E(\tilde{Z}\tilde{Z}^T)$ is positive definite and $E||\tilde{Z}||^3<\infty$.

Assumption 2. The matrix $I-\rho W$ is nonsingular with $|\rho|<1$.

Assumption 3. The sequences of matrices W and $I-\rho W$ are uniformly bounded in absolute value of both row and column sums.

Assumption 4. The nonparametric function $g(u)$ is v-th continuously differentiable on (0, 1) with $v\geq 2$. In addition, $g'(u)$ and $g''(u)$ are both bounded on (0, 1).

Assumption 5. There exists a constant C such that $\frac{\max\{h_i\}}{\min\{h_i\}}\leq C$ and $\max\{|h_{i+1}-h_i|\}=o(K^{-1})$, where K is the number of interior knots, $e_0=0$, $e_{K+1}=1$, $e_1,\ldots,e_K$ are the interior knots on [0, 1], and $h_i=e_i-e_{i-1}$.

Assumption 6. The tuning parameter λ_j satisfies $1/\min_{s+1\leq j\leq p+1}\lambda_j=o_p(1)$, and the penalty function satisfies $\liminf_{n\to\infty}\liminf_{t\to 0^+}\left\{\min_{s+1\leq j\leq p+1}\lambda_j^{-1}p'_{\lambda_j}(|t|)\right\}>0$.

Assumption 7. $\sqrt{n}a_n=o_p(1)$, $b_n=o_p(1)$.

Assumption 8. $\gamma_n-\gamma_0=o_p(1)$ for some $\gamma_0>0$.

Assumption 9. There exist constants C_1 and C_2 such that $\lambda_j\left|p''_{\lambda_j}(\theta_1)-p''_{\lambda_j}(\theta_2)\right|\leq C_2|\theta_1-\theta_2|$ for $\theta_1,\theta_2>C_1$ and $j=1,\ldots,p+1$.

Assumption 1 ensures that the main term dominates the remainder in the Taylor expansion (18). Assumptions 2 and 3 are required according to the literature on spatial autoregressive models. Assumptions 4 and 5 are common assumptions on the nonparametric function used in B-spline approximation techniques. Assumption 6 makes the penalty function singular at the origin, which makes the penalized estimator sparse. Assumptions 7 and 8 ensure the unbiasedness of large parameters and the existence of the $\sqrt{n}$-consistent penalized exponential square estimator, and ensure that the impact

of the penalty function on the penalized estimators is no greater than that of the least square function. Assumption 9 is the smoothing condition imposed on the non-concave penalty function.

Theorem 1 *Suppose that Assumptions* 1-9 *hold and the number of interior knots* K *satisfies* $K = O(n^{1/(2v+1)})$. *Then there exists a local maximizer* $\hat{\theta}_n$ *such that* $\|\hat{\theta}_n - \theta_0\| = O_p(n^{-1/2} + a_n)$.

Theorem 2 (*Oracle Property*) *Suppose that Assumptions* 1-9 *hold*, *the number of interior knots* K *satisfies* $K = O(n^{1/(2v+1)})$, *and* $I(\theta_0, \gamma_0)$ *is negative definite*. *If* $\sqrt{n}(\gamma_n - \gamma_0) = o_p(1)$ *for some* $\gamma_0 > 0$, *then* $\hat{\theta}_n = (\hat{\theta}_{n1}^T, \hat{\theta}_{n2}^T)^T$ *satisfy*

(a) *Sparsity*: $\hat{\theta}_{n2} = 0$ *with probability* 1;
(b) *Asymptotic normality*:

$$\sqrt{n-L}(I_1(\theta_{01}, \gamma_0) + \Sigma_1)\{\hat{\theta}_{n1} - \theta_{01} + (I_1(\theta_{01}, \gamma_0) + \Sigma_1)^{-1}\Delta\} \to N(\mathbf{0}, \Sigma_2),$$

where $\Sigma_1 = \text{diag}\{p''_{\lambda_j}(|\theta_{01}|), \ldots, p''_{\lambda_j}(|\theta_{0s}|)\}$, $\Sigma_2 = \text{Cov}(\exp(-r^2/\gamma_0)\frac{2r}{\gamma_0}\tilde{Z}_{i1})$, $\Delta = (p'_{\lambda_j}(|\theta_{01}|)\text{sign}(\theta_{01}), \ldots, p'_{\lambda_j}(|\theta_{0s}|)\text{sign}(\theta_{0s}))^T$, and $I_1(\theta_{01}, \gamma_0) = \frac{2}{\gamma_0}E[\exp(-r^2/\gamma_0)(\frac{2r^2}{\gamma_0} - 1)] \times (E\tilde{Z}_{i1}\tilde{Z}_{i1}^T)$.

Theorem 3 *Suppose that Assumptions* 1-9 *hold*, *and the number of interior knots* K *satisfies* $K = O(n^{1/(2v+1)})$. *Then we have*

$$\|\hat{g}(u) - g(u)\| = O_p(n^{-v/(2v+1)}),$$

where v *is defined in Assumption* 4, *and* $\| \ldots \|$ *denotes the* L_2 *norm*.

4 Algorithm and choice of tuning parameters

To facilitate the computation, we use a quadratic approximation to replace the loss function. Let

$$\ell^*(\theta) = \sum_{i=L+1}^{n} \exp\left\{-\left[\tilde{Y}_i - \tilde{Z}_i^T\theta\right]^2/\gamma\right\}. \tag{13}$$

Let $\tilde{\theta} = (\tilde{\rho}, \tilde{\beta}^T)^T$ denote an initial estimator, then the loss function is approximated as $\ell^*(\theta) \approx \ell^*(\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})^T\nabla^2\ell^*(\tilde{\theta})(\theta - \tilde{\theta})$. Thus, θ can be estimated by maximizing the penalized loss function approximation in quadratic form

$$\ell(\theta) \approx \frac{1}{2}(\theta - \tilde{\theta})^T\nabla^2\ell^*(\tilde{\theta})(\theta - \tilde{\theta}) - (n-L)\sum_{j=1}^{p+1} p_{\lambda_j}(\left|\theta_j\right|), \tag{14}$$

which leads to an approximated solution of (11).

To implement our proposed method, we need to choose the tuning parameters λ_j and γ. Since λ_j and γ are interdependent, it can be treated as a binary optimization problem. In this paper, we consider a simple selection method for λ_j and a data-driven procedure for γ. For the nonparametric component, we choose cubic B-splines to approximate the nonparametric functions and use equidistant knots where the number of internal knots is taken to be the integer part of $K = n^{1/(2\kappa+3)}$.

4.1 The choice of tuning parameter λ_j

In our simulation, we use the ALASSO penalty $p_{\lambda_j}(\left|\theta_j\right|) = \lambda_j|\theta_j|$, where $\lambda_j = \tau_j/|\tilde{\theta}_j|^k$ for some $k > 0$ and τ_j is the regularization parameter. We set $k = 1$ as suggested by Zou (2006) and then the ALASSO penalty can be rewritten as

$$\sum_{j=1}^{p+1} p_{\lambda_j}(\left|\theta_j\right|) = \sum_{j=1}^{p+1} \lambda_j \left|\theta_j\right|,$$

where $\lambda_j = \tau_j/|\tilde{\theta}_j|$. In general, many methods can be used to select λ_j, such as cross-validation, Akaike information criterion (AIC), and Bayesian information criterion (BIC). To reduce intensive computation and guarantee consistent variable selection, we choose the regularization parameter by minimizing a BIC-type objective function (see Wang et al. 2007)

$$\sum_{i=L+1}^{n} \left[1 - \exp\left\{-\left[\tilde{Y}_i - \tilde{Z}_i^T\theta)\right]^2/\gamma\right\}\right] + (n-L)\sum_{j=1}^{p+1} \lambda_j \left|\theta_j\right| - \sum_{j=1}^{p+1} \log(0.5(n-L)\lambda_j)\log(n-L),$$

which leads to $\lambda_j = \hat{\tau}_j/|\tilde{\theta}_j|$ with $\hat{\tau}_j = \frac{\log(n-L)}{n-L}$. It's easy to notice that this simple choice satisfies both $\sqrt{n}\lambda_j \longrightarrow 0$ for $j \le p_0$ and $\sqrt{n}\lambda_j \longrightarrow \infty$ for $j > p_0$, where p_0 is the number of nonzero elements in the true θ vector. Further, we can prove that our estimators are $\sqrt{n}$-consistent under this condition.

4.2 The choice of tuning parameter γ

The tuning parameter γ controls the degree of robustness and efficiency of the proposed robust regression parameter estimators. We use the procedure proposed by Wang et al. (2013) to handle γ. This procedure is data-driven and yields both high robustness and high efficiency simultaneously. At first, a set of adjustment parameters are determined such that the proposed penalized robust estimator has an asymptotic breakdown point of 1/2, and then tuning parameters are selected with maximum efficiency. The whole process is described in the following steps.

Step 1. Set the initial estimator.

Let $\hat{\theta} = (\hat{\rho}, \hat{\beta}^T)^T$ be an initial estimator, for which we can use the MM-estimator.

Step 2. Find the pseudo outlier set of the sample.

Let $D_{n-L} = \{(\tilde{Z}_{L+1}, \tilde{Y}_{L+1}), (\tilde{Z}_{L+2}, \tilde{Y}_{L+2}), (\tilde{Z}_n, \tilde{Y}_n)\}$, and calculate $r_i(\hat{\theta}) = \tilde{Y}_i - \tilde{Z}_i^T\hat{\theta}$, $i = L+1, L+2, \ldots, n$, and $S_{n-L} = 1.4826 \times \text{median}\{|r_i(\hat{\theta}) - \text{median}_j(r_j(\hat{\theta}))|, i = L+1, L+2, \ldots, n\}$. Then, take the pseudo outlier set

$$D_m = \{(\tilde{Z}_i, \tilde{Y}_i) : |r_i(\hat{\theta})| \geq 2.5S_{n-L}\},$$

where $m = \sharp\{1 \leq i \leq n : |r_i(\hat{\theta})| \geq 2.5S_{n-L}\}$. Let $D_{n-L-m} = D_{n-L}/D_m$.

Step 3. Chose the tuning parameter γ.

Define $\hat{V}(\gamma) = \{\hat{I}_1(\hat{\theta})\}^{-1}\tilde{\Sigma}_2\{\hat{I}_1(\hat{\theta})\}$, where

$$\hat{I}_1(\hat{\theta}) = \frac{2}{\gamma}\left\{\frac{1}{n-L}\sum_{i=L+1}^{n}\exp(-r_i^2(\hat{\theta})/\gamma)\left(\frac{2r_i^2(\hat{\theta})}{\gamma} - 1\right) \times \left(\frac{1}{n-L}\sum_{i=L+1}^{n}\tilde{Z}_i\tilde{Z}_i^T\right)\right\},$$

$$\tilde{\Sigma}_2 = \text{Cov}\left\{\exp(-r_{L+1}^2(\hat{\theta})/\gamma)\frac{2r_{L+1}^2(\hat{\theta})}{\gamma}\tilde{Z}_{L+1}, \ldots, \exp(-r_n^2(\hat{\theta})/\gamma)\frac{2r_n^2(\hat{\theta})}{\gamma}\tilde{Z}_n\right\}.$$

Let γ be the minimizer of $\det(\hat{V}(\gamma))$ in the set $G = \{\gamma : \zeta(\gamma) \in (0, 1]\}$, where $\zeta(\gamma) = \frac{2m}{n-L} + \frac{2}{n-L}\sum_{i=m+L+1}^{n}\phi_\gamma\{r_i(\hat{\theta})\}$, $\phi_\gamma(t) = 1 - \exp(-t^2/\gamma)$, and $\det(\ldots)$ denotes the determinant operator.

Step 4. Updata $\hat{\theta}$.

With the chosen regularization parameter $\lambda_j = \log(n-L)/((n-L)|\hat{\theta}_j|)$ and the selected γ in Step 3, update θ by maximizing (10). Go to Step 3 until convergence.

5 Simulation studies

In this section, we conduct simulation studies to assess the finite-sample performance of the proposed variable selection method. We simulate $N = 1000$ data sets from the partially linear spatial autoregressive model

$$Y_i = \rho\sum_{j=1}^{n} w_{ij}Y_j + X_i^T\beta + g(U_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{15}$$

where $g(u) = \sin(2\pi u) + \cos(2\pi u)$, the p-dimensional parameter vector $\beta = (1.5, 3, 2, 2.5, 0, \ldots, 0)^T$, and the covariates $X_i \sim N_p(0, I_p)$ and $U_i \sim U[0, 1]$. We consider four different settings for ϵ_i:

(i) $\frac{1}{\sigma}\epsilon_i \sim \chi^2(2)/2 - 1$,

(ii) $\frac{1}{\sigma}\epsilon_i \sim 0.99N(0, 0.5) + 0.01N(0, 100)$,

(iii) $\frac{\sqrt{3}}{\sigma}\epsilon_i \sim t(3)$,

(iv) $\frac{\sqrt{3}}{\sigma}\epsilon_i \sim C(1, 0)$,

where $C(1, 0)$ represents a Cauchy distribution with location parameter 0 and scale parameter 1. In order to examine the effect of the error term on the variable selection process, we consider three different σ^2 values $\sigma^2 = 0.2, 0.5, 0.8$. Following Lee (2004) and Case (1991), we assume there are R districts in W and each district has m members. More specifically, we assume the spatial weight matrix is $W = (w_{ij}) = I_R \otimes \Omega_m$, where $\Omega_m = \frac{1}{m-1}(\mathbf{1}_m \mathbf{1}_m^T - I_m)$, $\mathbf{1}_m$ is the m-dimensional vector with all elements 1, and $\otimes$ is Kronecker product. In our simulation, we take $m = 5$ and $n = R \times m$. For spatial scalar parameter ρ, we consider five different values $\rho = 0, 0.3, 0.5, 0.6, 0.9$.

We first assess the performance of the recognition ability in terms of the significance of spatial effects. In this simulation, we set the dimension of parameter β as $p = 10$, $\epsilon_i \sim N(0, 0.5)$, and the spatial scalar parameter ρ is taken from 0 to 0.2 at intervals of 0.01. Recognition accuracy is used to evaluate spatial effects at different levels. In addition, the penalty function is taken to be the ALASSO penalty. In our variable selection process, since we need to estimate some nonparametric functions at the same time, we need a large sample size. In the following simulation, we consider sample sizes $n = 250, 400, 550$. Similar to Zhao et al. (2021), the nonzero identification rate (NIR), which is used to assess the effectiveness of the spatial effect significance procedure, is defined as

$$\text{NIR} = \frac{NT}{TT}, \tag{16}$$

where NT represents the number of times the spatial effect parameter ρ is estimated to be nonzero among the total TT number of simulation runs. Obviously, $TT = 1000$ in our case.

Based on 1000 simulation runs, the simulation results are shown in Fig. 1. From Fig. 1, we have the following observations.

(1) Note that for $\rho = 0$, NIR provides the proportion of misidentified spatial effects when they are actually not significant. Figure 1 shows that the NIR values are

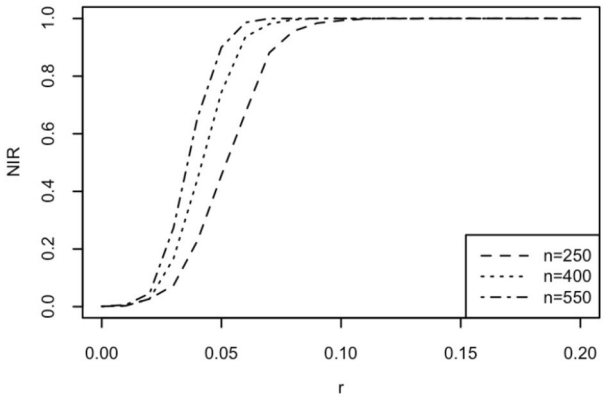

Fig. 1 The identification power for spatial effect significance with the ALASSO penalty

very small for different sample sizes. This indicates that when the spatial effect is not significant, the probability of misidentification by our proposed identification method is almost zero, i.e., the proposed method can correctly identify the insignificant spatial effect.

(2) Note that for $\rho \neq 0$, NIR provides the proportion of correctly identified spatial effects when they are actually significant. For any given sample size n, the NIR value increases rapidly and approaches 1 with the increase in ρ, indicating that when the spatial effect is significant, the proposed identification method can correctly identify the significance of the spatial effect.

Next, we evaluate the performance of the proposed covariate selection procedure. In the following simulation study, we compare the performance of the selection procedure under different errors. For comparison, we also calculate the estimates obtained by ESL-ALASSO and oracle. Again we simulate 1000 replicates each time for different n and ρ. Define the generalized mean square error (GMSE) as

$$\text{GMSE} = (\hat{\beta} - \beta)^T \left[\frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \right] (\hat{\beta} - \beta).$$

Let C denote the average number of true zero coefficients that are correctly estimated to be zero, I denote the average number of true nonzero coefficients that are erroneously estimated to be zero, and FSR denote the false selection rate, i.e., FSR $= IN/TN$, where IN is the number of insignificant covariates incorrectly set as important covariates and TN is the total number of significant covariates. The simulation results are reported in Tables 1, 2, 3 and 4.

From Tables 1, 2, 3 and 4 we can see that the error distribution has little effect on the results. Regardless of the four different error distributions, both FSR and GMSE decrease as the sample size n increases, and the C value always tends to be very close to the true number of zero coefficients, regardless of sample size. We also find that our selection procedure performs significantly better for larger sample sizes than for smaller sample sizes, due to the fact that consistent estimators are involved in the selection process. It is noted from Tables 1, 2, 3 and 4 that with the increase in sample size n or the decrease in variance σ^2, the performance of the ESL-ALASSO method is getting closer and closer to the oracle result. In addition, comparing the results under the five different ρ values, we can conclude that the spatial autoregressive parameter ρ has little effect on the finite-sample performance of the proposed variable selection method. This demonstrates that the variable selection method we proposed in this paper is

Table 1 Variable selection for the parametric components with model error of case (i)

		$\sigma^2 = 0.8$				$\sigma^2 = 0.5$				$\sigma^2 = 0.2$			
(n, ρ)	Methods	C	I	FSR	GMSE	C	I	FSR	GMSE	C	I	FSR	GMSE
(250,0)	ESL-ALASSO	5.992	0	0.002	0.171	5.994	0	0.002	0.095	5.991	0	0.002	0.040
	Oracle	6	0	0	0.159	6	0	0	0.092	6	0	0	0.039
(250,0.3)	ESL-ALASSO	5.991	0	0.002	0.173	5.995	0	0.001	0.091	5.994	0	0.002	0.042
	Oracle	6	0	0	0.159	6	0	0	0.088	6	0	0	0.040
(250,0.5)	ESL-ALASSO	5.989	0	0.003	0.177	5.990	0	0.003	0.090	5.990	0	0.003	0.042
	Oracle	6	0	0	0.166	6	0	0	0.086	6	0	0	0.039
(250,0.6)	ESL-ALASSO	5.988	0	0.003	0.176	5.995	0	0.001	0.095	5.996	0	0.001	0.040
	Oracle	6	0	0	0.166	6	0	0	0.092	6	0	0	0.039
(250,0.9)	ESL-ALASSO	5.986	0	0.004	0.190	5.990	0	0.003	0.103	5.989	0	0.003	0.046
	Oracle	6	0	0	0.181	6	0	0	0.100	6	0	0	0.044
(400,0)	ESL-ALASSO	5.990	0	0.003	0.061	5.998	0	0.001	0.036	5.993	0	0.002	0.016
	Oracle	6	0	0	0.060	6	0	0	0.036	6	0	0	0.015
(400,0.3)	ESL-ALASSO	5.993	0	0.002	0.064	5.996	0	0.001	0.038	5.997	0	0.001	0.016
	Oracle	6	0	0	0.063	6	0	0	0.038	6	0	0	0.015
(400,0.5)	ESL-ALASSO	5.990	0	0.003	0.065	5.997	0	0.001	0.038	5.995	0	0.001	0.016
	Oracle	6	0	0	0.063	6	0	0	0.037	6	0	0	0.016
(400,0.6)	ESL-ALASSO	5.992	0	0.002	0.065	5.994	0	0.002	0.039	5.995	0	0.001	0.017
	Oracle	6	0	0	0.064	6	0	0	0.039	6	0	0	0.017
(400,0.9)	ESL-ALASSO	5.992	0	0.002	0.075	5.998	0	0.001	0.039	5.996	0	0.001	0.018
	Oracle	6	0	0	0.073	6	0	0	0.038	6	0	0	0.017
(550,0)	ESL-ALASSO	5.996	0	0.001	0.041	5.996	0	0.001	0.023	5.994	0	0.002	0.009
	Oracle	6	0	0	0.040	6	0	0	0.023	6	0	0	0.009
(550,0.3)	ESL-ALASSO	5.995	0	0.001	0.039	5.999	0	0.000	0.023	5.998	0	0.001	0.010
	Oracle	6	0	0	0.039	6	0	0	0.023	6	0	0	0.010

Table 1 (continued)

		$\sigma^2 = 0.8$				$\sigma^2 = 0.5$				$\sigma^2 = 0.2$			
(n, ρ)	Methods	C	I	FSR	GMSE	C	I	FSR	GMSE	C	I	FSR	GMSE
(550,0.5)	ESL-ALASSO	5.996	0	0.001	0.040	5.998	0	0.001	0.025	5.991	0	0.002	0.010
	Oracle	6	0	0	0.040	6	0	0	0.025	6	0	0	0.009
(550,0.6)	ESL-ALASSO	5.995	0	0.001	0.039	5.999	0	0.000	0.024	5.998	0	0.001	0.010
	Oracle	6	0	0	0.039	6	0	0	0.024	6	0	0	0.010
(550,0.9)	ESL-ALASSO	5.995	0	0.001	0.042	5.998	0	0.001	0.025	5.997	0	0.001	0.010
	Oracle	6	0	0	0.042	6	0	0	0.025	6	0	0	0.010

Table 2 Variable selection for the parametric components with model error of case (ii)

		$\sigma^2 = 0.8$				$\sigma^2 = 0.5$				$\sigma^2 = 0.2$			
(n, ρ)	Methods	C	I	FSR	GMSE	C	I	FSR	GMSE	C	I	FSR	GMSE
(250,0)	ESL-ALASSO	5.989	0	0.003	0.230	5.992	0	0.002	0.101	5.993	0	0.002	0.044
	Oracle	6	0	0	0.229	6	0	0	0.100	6	0	0	0.044
(250,0.3)	ESL-ALASSO	5.987	0	0.003	0.236	5.996	0	0.001	0.107	5.991	0	0.002	0.044
	Oracle	6	0	0	0.229	6	0	0	0.102	6	0	0	0.044
(250,0.5)	ESL-ALASSO	5.992	0	0.002	0.237	5.997	0	0.001	0.110	5.988	0	0.003	0.043
	Oracle	6	0	0	0.234	6	0	0	0.107	6	0	0	0.043
(250,0.6)	ESL-ALASSO	5.993	0	0.002	0.239	5.996	0	0.001	0.106	5.994	0	0.002	0.045
	Oracle	6	0	0	0.236	6	0	0	0.106	6	0	0	0.045
(250,0.9)	ESL-ALASSO	5.988	0	0.003	0.263	5.998	0	0.001	0.113	5.996	0	0.001	0.046
	Oracle	6	0	0	0.257	6	0	0	0.111	6	0	0	0.046
(400,0)	ESL-ALASSO	5.992	0	0.002	0.129	5.998	0	0.001	0.059	5.997	0	0.001	0.027
	Oracle	6	0	0	0.127	6	0	0	0.058	6	0	0	0.027
(400,0.3)	ESL-ALASSO	5.995	0	0.001	0.137	5.998	0	0.001	0.056	5.994	0	0.002	0.026
	Oracle	6	0	0	0.135	6	0	0	0.055	6	0	0	0.026
(400,0.5)	ESL-ALASSO	5.994	0	0.002	0.138	5.996	0	0.001	0.058	5.995	0	0.001	0.027
	Oracle	6	0	0	0.136	6	0	0	0.057	6	0	0	0.027
(400,0.6)	ESL-ALASSO	5.997	0	0.001	0.136	5.996	0	0.001	0.061	5.995	0	0.001	0.027
	Oracle	6	0	0	0.135	6	0	0	0.060	6	0	0	0.027
(400,0.9)	ESL-ALASSO	5.995	0	0.001	0.151	5.996	0	0.001	0.067	5.995	0	0.001	0.029
	Oracle	6	0	0	0.149	6	0	0	0.065	6	0	0	0.029
(550,0)	ESL-ALASSO	5.996	0	0.001	0.098	5.998	0	0.001	0.040	5.998	0	0.001	0.019
	Oracle	6	0	0	0.097	6	0	0	0.040	6	0	0	0.019
(550,0.3)	ESL-ALASSO	5.996	0	0.001	0.094	5.999	0	0.000	0.041	5.996	0	0.001	0.019
	Oracle	6	0	0	0.094	6	0	0	0.041	6	0	0	0.019

Table 2 (continued)

		$\sigma^2 = 0.8$				$\sigma^2 = 0.5$				$\sigma^2 = 0.2$			
(n, ρ)	Methods	C	I	FSR	GMSE	C	I	FSR	GMSE	C	I	FSR	GMSE
(550,0.5)	ESL-ALASSO	5.996	0	0.001	0.099	5.997	0	0.001	0.042	5.996	0	0.001	0.020
	Oracle	6	0	0	0.098	6	0	0	0.042	6	0	0	0.020
(550,0.6)	ESL-ALASSO	5.996	0	0.001	0.096	5.999	0	0.000	0.040	5.997	0	0.001	0.019
	Oracle	6	0	0	0.095	6	0	0	0.040	6	0	0	0.019
(550,0.9)	ESL-ALASSO	5.999	0	0.000	0.109	6	0	0.000	0.045	5.996	0	0.001	0.020
	Oracle	6	0	0	0.107	6	0	0	0.045	6	0	0	0.020

Table 3 Variable selection for the parametric components with model error of case (iii)

		$\sigma^2 = 0.8$				$\sigma^2 = 0.5$				$\sigma^2 = 0.2$			
(n, ρ)	Methods	C	I	FSR	GMSE	C	I	FSR	GMSE	C	I	FSR	GMSE
(250,0)	ESL-ALASSO	5.994	0	0.002	0.140	5.996	0	0.001	0.069	5.998	0	0.001	0.030
	Oracle	6	0	0	0.139	6	0	0	0.069	6	0	0	0.030
(250,0.3)	ESL-ALASSO	5.991	0	0.002	0.142	5.996	0	0.001	0.070	5.990	0	0.003	0.030
	Oracle	6	0	0	0.141	6	0	0	0.070	6	0	0	0.030
(250,0.5)	ESL-ALASSO	5.987	0	0.003	0.141	5.994	0	0.002	0.071	5.991	0	0.002	0.030
	Oracle	6	0	0	0.138	6	0	0	0.071	6	0	0	0.030
(250,0.6)	ESL-ALASSO	5.992	0	0.002	0.148	5.995	0	0.001	0.075	5.994	0	0.002	0.029
	Oracle	6	0	0	0.145	6	0	0	0.075	6	0	0	0.029
(250,0.9)	ESL-ALASSO	5.989	0	0.003	0.150	5.997	0	0.001	0.075	5.991	0	0.002	0.033
	Oracle	6	0	0	0.147	6	0	0	0.075	6	0	0	0.033
(400,0)	ESL-ALASSO	5.995	0	0.001	0.082	5.999	0	0.000	0.039	5.994	0	0.002	0.017
	Oracle	6	0	0	0.081	6	0	0	0.039	6	0	0	0.017
(400,0.3)	ESL-ALASSO	5.997	0	0.001	0.080	5.999	0	0.000	0.039	5.993	0	0.002	0.017
	Oracle	6	0	0	0.079	6	0	0	0.039	6	0	0	0.017
(400,0.5)	ESL-ALASSO	5.993	0	0.002	0.085	5.992	0	0.002	0.041	5.999	0	0.000	0.017
	Oracle	6	0	0	0.084	6	0	0	0.040	6	0	0	0.017
(400,0.6)	ESL-ALASSO	5.996	0	0.001	0.083	5.997	0	0.001	0.041	5.997	0	0.001	0.018
	Oracle	6	0	0	0.083	6	0	0	0.041	6	0	0	0.018
(400,0.9)	ESL-ALASSO	5.994	0	0.002	0.086	5.997	0	0.001	0.045	5.992	0	0.002	0.017
	Oracle	6	0	0	0.086	6	0	0	0.045	6	0	0	0.017
(550,0)	ESL-ALASSO	5.996	0	0.001	0.053	5.999	0	0.000	0.027	6	0	0	0.011
	Oracle	6	0	0	0.052	6	0	0	0.027	6	0	0	0.011
(550,0.3)	ESL-ALASSO	5.995	0	0.001	0.055	5.997	0	0.001	0.027	5.991	0	0.002	0.012
	Oracle	6	0	0	0.055	6	0	0	0.027	6	0	0	0.012

Table 3 (continued)

		$\sigma^2 = 0.8$				$\sigma^2 = 0.5$				$\sigma^2 = 0.2$			
(n, ρ)	Methods	C	I	FSR	GMSE	C	I	FSR	GMSE	C	I	FSR	GMSE
(550,0.5)	ESL-ALASSO	5.994	0	0.002	0.057	5.999	0	0.000	0.027	5.995	0	0.001	0.012
	Oracle	6	0	0	0.057	6	0	0	0.027	6	0	0	0.012
(550,0.6)	ESL-ALASSO	5.996	0	0.001	0.057	5.998	0	0.001	0.027	5.999	0	0.000	0.012
	Oracle	6	0	0	0.057	6	0	0	0.027	6	0	0	0.012
(550,0.9)	ESL-ALASSO	5.995	0	0.001	0.061	6	0	0	0.030	5.998	0	0.001	0.013
	Oracle	6	0	0	0.061	6	0	0	0.030	6	0	0	0.013

Table 4 Variable selection for the parametric components with model error of case (iv)

		$\sigma^2 = 0.8$				$\sigma^2 = 0.5$				$\sigma^2 = 0.2$			
(n, ρ)	Methods	C	I	FSR	GMSE	C	I	FSR	GMSE	C	I	FSR	GMSE
(250,0)	ESL-ALASSO	5.990	0	0.003	0.271	5.991	0	0.002	0.182	5.996	0	0.001	0.080
	Oracle	6	0	0	0.266	6	0	0	0.180	6	0	0	0.078
(250,0.3)	ESL-ALASSO	5.991	0	0.002	0.297	5.991	0	0.002	0.185	5.996	0	0.001	0.096
	Oracle	6	0	0	0.292	6	0	0	0.182	6	0	0	0.093
(250,0.5)	ESL-ALASSO	5.989	0	0.003	0.281	5.994	0	0.002	0.199	5.997	0	0.001	0.087
	Oracle	6	0	0	0.278	6	0	0	0.196	6	0	0	0.086
(250,0.6)	ESL-ALASSO	5.989	0	0.003	0.278	5.993	0	0.002	0.201	5.996	0	0.001	0.093
	Oracle	6	0	0	0.276	6	0	0	0.197	6	0	0	0.091
(250,0.9)	ESL-ALASSO	5.992	0	0.002	0.321	5.994	0	0.002	0.210	5.994	0	0.002	0.100
	Oracle	6	0	0	0.317	6	0	0	0.208	6	0	0	0.100
(400,0)	ESL-ALASSO	5.998	0	0.001	0.121	5.997	0	0.001	0.080	5.997	0	0.001	0.038
	Oracle	6	0	0	0.120	6	0	0	0.079	6	0	0	0.038
(400,0.3)	ESL-ALASSO	5.995	0	0.001	0.128	5.997	0	0.001	0.089	5.996	0	0.001	0.036
	Oracle	6	0	0	0.127	6	0	0	0.088	6	0	0	0.036
(400,0.5)	ESL-ALASSO	5.996	0	0.001	0.129	5.997	0	0.001	0.081	5.996	0	0.001	0.038
	Oracle	6	0	0	0.129	6	0	0	0.081	6	0	0	0.038
(400,0.6)	ESL-ALASSO	5.993	0	0.002	0.131	5.995	0	0.001	0.086	5.998	0	0.001	0.037
	Oracle	6	0	0	0.130	6	0	0	0.086	6	0	0	0.037
(400,0.9)	ESL-ALASSO	5.995	0	0.001	0.139	5.996	0	0.001	0.094	6	0	0	0.042
	Oracle	6	0	0	0.139	6	0	0	0.093	6	0	0	0.042
(550,0)	ESL-ALASSO	5.992	0	0.002	0.078	5.997	0	0.001	0.048	5.997	0	0.001	0.021
	Oracle	6	0	0	0.078	6	0	0	0.048	6	0	0	0.021
(550,0.3)	ESL-ALASSO	5.995	0	0.001	0.078	5.992	0	0.002	0.050	5.997	0	0.001	0.023
	Oracle	6	0	0	0.078	6	0	0	0.050	6	0	0	0.023

Table 4 (continued)

		$\sigma^2 = 0.8$				$\sigma^2 = 0.5$				$\sigma^2 = 0.2$			
(n, ρ)	Methods	C	I	FSR	GMSE	C	I	FSR	GMSE	C	I	FSR	GMSE
(550,0.5)	ESL-ALASSO	5.997	0	0.001	0.076	6	0	0	0.054	5.998	0	0.001	0.023
	Oracle	6	0	0	0.076	6	0	0	0.054	6	0	0	0.023
(550,0.6)	ESL-ALASSO	5.997	0	0.001	0.079	5.997	0	0.001	0.055	5.998	0	0.001	0.023
	Oracle	6	0	0	0.079	6	0	0	0.054	6	0	0	0.023
(550,0.9)	ESL-ALASSO	5.998	0	0.001	0.089	5.999	0	0.000	0.056	5.999	0	0.000	0.025
	Oracle	6	0	0	0.088	6	0	0	0.056	6	0	0	0.025

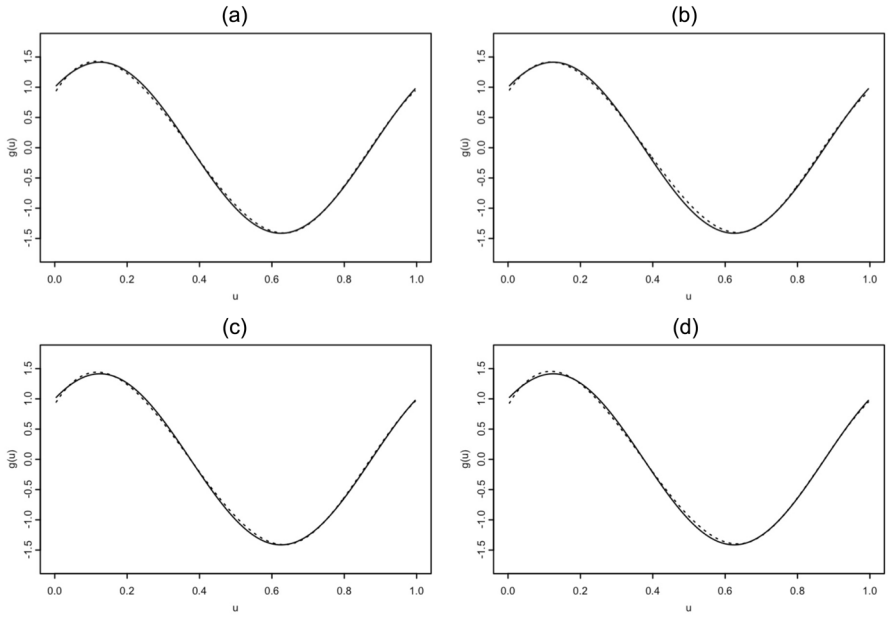

Fig. 2 The estimated (dashed) and true (solid) curves of $g(u)$ with the error distribution from case (i) to case (iv)

effective regardless of whether there is a spatial effect in model (1) and regardless of the size of the spatial autoregressive parameter ρ.

In Fig. 2, we display the true and the estimated nonparametric component $g(u)$ by our method, when $\beta = (1.5, 3, 2, 2.5, 0, 0, 0, 0, 0, 0)^T$, $\rho = 0.5$, $\sigma = 1$ and $n = 250$. The images given in (a)-(d) are for cases (i)-(iv) of the model error, respectively. We can see from Fig. 2 that the estimated curve is very close to the true curve, regardless of the error distribution. This not only indicates that our estimation procedure performs well for the nonparametric component as well, but also demonstrates that our estimation procedure is insensitive to error distribution.

6 A real data example

In this section, we use the Boston House Price Dataset as a real data example to show the application of the proposed variable selection method. The set contains 506 observations of the owner-occupied homes in 506 census tracts in the Boston Standard Metropolitan Statistical Area in 1970, and is now freely available through the GeoDa Center for Geospatial Analysis and Computation. Following Li and Guo (2020), we take MEDV(the median value of owner-occupied homes) as response variable, and consider 13 variables as explanatory variables that may explain changes in house prices. These 13 variables are as follows:

- CRIM, per capita crime rate per tract;

Table 5 The selection and estimated coefficient parameters for the Boston housing price data

$\hat{\rho}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$
0.1377	–0.0551	0	0	0	0	0.1250	0	–0.0158	0	–0.0235	–0.0337	0.0367

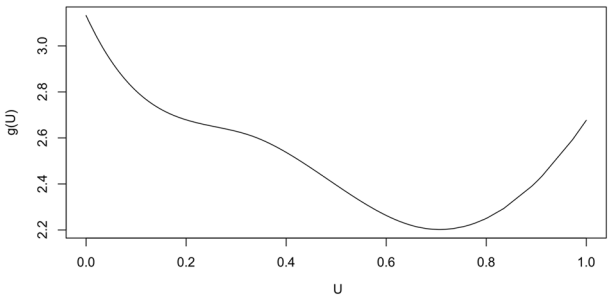

Fig. 3 The estimated curve of $g(u)$ for Boston housing price data

- ZN, proportion of a town's residential land zoned for lots greater than 25,000 square feet;
- INDUS, proportion of non-retail business acres per town;
- CHAS, Charles River dummy variable (1 if tract bounds the Charles River; 0 otherwise);
- NOX, nitrogen oxide concentration in pphm per town;
- RM, average number of rooms per dwelling;
- AGE, proportion of owner-occupied homes built prior to 1940 per tract;
- DIS, weighted average of distances of a tract to five employment centers in the Boston region;
- RAD, index of a town's accessibility to radial highways;
- TAX, full value property tax rate per $10,000 per town;
- PTRATIO, pupil-teacher ratio by town school district;
- Bk, proportion of blacks per tract;
- LSTAT, proportion of population that is in the lower status.

We consider the following partially linear spatial autoregressive model

$$Y_i = \rho \sum_{j \neq i} w_{ij} Y_j + X_i^T \beta + g(U_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{17}$$

where $n = 506$, Y_i is the observation of $\ln(MEDV)$, U_i is the observation vector of variable LSTAT, and we denote covariates CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO and Bk as $X_1, \ldots, X_{12}$. To facilitate the analysis, similar to Li and Guo (2020), we convert the explanatory variables in the nonparametric components to values in the interval (0, 1), and the explanatory variables in the parametric components are all standardized, which also approximates the normal

distribution of their marginal distribution. Following Su and Yang (2007), we take the element w_{ij} of W to be $w_{ij} = \frac{\tilde{w}_{ij}}{\sum_{j \neq i} \tilde{w}_{ij}}$, where $\tilde{w}_{ij} = \max(1 - \frac{d_{ij}}{d_0}, 0)$ when $i \neq j$, and $\tilde{w}_{ij} = 0$ otherwise, with d_{ij} the Euclidean distance calculated in terms of the longitude and latitude coordinates of census tract and d_0 a threshold distance used to control the degree of spatial dependence of the response variable. Here, we take $d_0 = 0.05$ as in Su and Yang (2007), which leads to a spatial weight matrix with 19.08% nonzero elements. We apply our proposed variable selection method to identify significant explanatory variables in the parametric component of model (16) and simultaneously estimate nonzero parameters. The selection and estimated coefficient parameters are summarized in Table 5. The estimation of the nonparametric function is shown in Fig. 3.

From Table 5, we see that the scalar autoregressive parameter estimator $\hat{\rho} > 0$, which means that there is a substantial spatial relationship between the responses. In addition, our method selects CRIM, RM, DIS, TAX, PTRATIO and Bk as significant explanatory variables, and ZN, INDUS, CHAS, NOX, AGE and RAD as insignificant variables. Among the significant explanatory variables, CRIM, DIS, TAX and PTRATIO have negative effects on the housing price, while the influence of RM and Bk is positive. These observations are clearly in line with the implications of their impact on house prices. One noteworthy point is that Bk has a significantly positive coefficient. Although this may be inconsistent with our common sense, this conclusion is consistent with Harrison and Rubinfeld (1978) and Kong and Xia (2012), with a detailed explanation given in Kong and Xia (2012). In addition, from Fig. 3, we see that the estimated nonparametric function shows a nonlinear downward trend with the increase in the explanatory variable LSTAT, which is also consistent with the observation in Li and Guo (2020).

7 Conclusion

In this paper, we propose a robust variable selection procedure based on ESL, for partially linear spatial autoregressive models. By using the technique of orthogonal projection, we are able to select important covariates in the parametric components without affecting the nonparametric components. Under appropriate regularization conditions, we show that the proposed procedure is consistent and the resulting estimators have oracle property. Simulation studies show that the proposed method can handle data with outliers and can identify small spatial effects. We examine the finite-sample performance of our proposed estimation method under different error distributions and find that both the parametric and nonparametric parts can be well estimated regardless of error distribution. Even when $\rho = 0$, i.e. there is no spatial effect, our method can still efficiently accomplish both variable selection and

parameter estimation tasks. The real data analysis also indicates that our proposed method performs well.

Our work mainly focuses on the study of partial linear spatial autoregressive models with complete data. In reality, the data in many cases is not necessarily complete, and thus one of our near future works will deal with spatial autoregressive models with missing data.

Acknowledgements The research is supported by NSF projects (ZR2021MA077 and ZR2021MA048) of Shandong Province of China.

Appendix

Proof of Theorem 1 Let $\xi = n^{-1/2} + a_n$. Similar to Fan and Li (2001), we first prove that for any given $\epsilon > 0$, there exists a constant C such that

$$P\left\{ \sup_{\|\mathbf{u}\|=C} \ell(\theta_0 + \xi\mathbf{u}) < \ell(\theta_0) \right\} \geq 1 - \epsilon, \tag{18}$$

where $\mathbf{u}$ is a $(p+1)$-dimensional vector such that $\|\mathbf{u}\| = C$, and C is a large enough constant. This means that the probability that there exists a local maximum in the sphere $\{\theta_0 + \xi\mathbf{u} : \|\mathbf{u}\| \leq C\}$ is at least $1 - \epsilon$. Hence, we prove that there exists a local maximizer $\hat{\theta}_n$ such that $\|\hat{\theta}_n - \theta_0\| = O_p(\xi)$. Let

$$D(\theta, \gamma) = \sum_{i=L+1}^{n} \exp\left\{-(\tilde{Y}_i - \tilde{Z}_i^T\theta)^2/\gamma\right\} \frac{2(\tilde{Y}_i - \tilde{Z}_i^T\theta)}{\gamma} \tilde{Z}_i.$$

Since $p_{\lambda_j}(0) = 0$ for $j = 1, \ldots, p+1$ and $\gamma_n - \gamma_0 = o_p(1)$, by Taylor's expansion we have

$$
\begin{aligned}
&\ell(\theta_0+\xi\mathbf{u})-\ell(\theta_0)\\
&\quad=\sum_{i=L+1}^{n}\exp\left\{-\frac{(\tilde{Y}_i-\tilde{Z}_i^T(\theta_0+\xi\mathbf{u}))^2}{\gamma_n}\right\}-\sum_{i=L+1}^{n}\exp\left\{-\frac{(\tilde{Y}_i-\tilde{Z}_i^T\theta_0)^2}{\gamma_n}\right\}\\
&\qquad-(n-L)\sum_{j=1}^{p+1}\{p_{\lambda_j}(|\theta_{0j}+\xi_j|)-p_{\lambda_j}(|\theta_{0j}|)\}\\
&\quad\le\sum_{i=L+1}^{n}\exp\left\{-\frac{(\tilde{Y}_i-\tilde{Z}_i^T(\theta_0+\xi\mathbf{u}))^2}{\gamma_n}\right\}-\sum_{i=L+1}^{n}\exp\left\{-\frac{(\tilde{Y}_i-\tilde{Z}_i^T\theta_0)^2}{\gamma_n}\right\}\\
&\qquad-(n-L)\sum_{j=1}^{s}\{p_{\lambda_j}(|\theta_{0j}+\xi_j|)-p_{\lambda_j}(|\theta_{0j}|)\}\\
&\quad=\xi D(\theta_0,\gamma_n)^T\mathbf{u}-\frac{1}{2}\mathbf{u}^T[-I(\theta_0,\gamma_n)]\mathbf{u}(n-L)\xi^2\{1+o(1)\}\\
&\qquad-\sum_{j=1}^{s}[(n-L)\xi p'_{\lambda_j}(|\theta_{0j}|)\mathrm{sign}(\theta_{0j})u_j+(n-L)\xi^2p''_{\lambda_j}(|\theta_{0j}|)u_j^2\{1+o(1)\}]\\
&\quad=\xi\{D(\theta_0,\gamma_0)+o_p(\sqrt{n})\}^T\mathbf{u}-\frac{1}{2}\mathbf{u}^T[-I(\theta_0,\gamma_0)+o(1)]\mathbf{u}(n-L)\xi^2\{1+o(1)\}\\
&\qquad-\sum_{j=1}^{s}[(n-L)\xi p'_{\lambda_j}(|\theta_{0j}|)\mathrm{sign}(\theta_{0j})u_j+(n-L)\xi^2p''_{\lambda_j}(|\theta_{0j}|)u_j^2\{1+o(1)\}]\\
&\quad\le\xi\{D(\theta_0,\gamma_0)+o_p(\sqrt{n}))\}^T\mathbf{u}-\frac{1}{2}\mathbf{u}^T[-I(\theta_0,\gamma_0)+o(1)]\mathbf{u}(n-L)\xi^2\{1+o(1)\}\\
&\qquad-[\sqrt{s}(n-L)\xi a_n\|\mathbf{u}\|+(n-L)\xi^2b_n\|\mathbf{u}\|^2].
\end{aligned}
\tag{19}
$$

Note that $n^{-1/2}D(\theta_0,\gamma_0)=O_p(1)$. Therefore, the order of the first term on the right side of Eq. (18) is equal to $O_p(n^{1/2}\xi)=O_p(n\xi^2)$. By choosing a sufficiently large C, the second term dominates the first term uniformly in $\|\mathbf{u}\|=C$. Since $b_n=o_p(1)$, the third term is also dominated by the second term of (18). Therefore, (17) holds by choosing a sufficiently large C. The proof of Theorem 1 is completed. □

Proof of Theorem 2(a) We now prove the sparsity. We will prove that with probability 1, for any θ_1 satisfying $\theta_1-\theta_{01}=O_p(n^{-1/2})$, and for some small $\epsilon_n=Cn^{-1/2}$ and $j=s+1,\ldots,p+1$, we have $\partial\ell(\theta)/\partial\theta_j>0$, for $0<\theta_j<\epsilon_n$, and $\partial\ell(\theta)/\partial\theta_j<0$, for $-\epsilon_n<\theta_j<0$. Let

$$
Q_n(\theta,\gamma)=\sum_{i=L+1}^{n}\exp\left\{-(\tilde{Y}_i-\tilde{Z}_i^T\theta)^2/\gamma\right\}. \tag{20}
$$

By Talylor's expansion, we have

$$\begin{aligned}\frac{\partial \ell(\theta)}{\partial \theta_j} &= \frac{\partial Q_n(\theta,\gamma_n)}{\partial \theta_j} - (n-L)p'_{\lambda_j}(|\theta_j|)\text{sign}(\theta_j)\\ &= \frac{\partial Q_n(\theta_0,\gamma_n)}{\partial \theta_j} + \sum_{l=1}^{p+1}\frac{\partial^2 Q_n(\theta_0,\gamma_n)}{\partial \theta_j \partial \theta_l}(\theta_l-\theta_{0l})\\ &\quad + \sum_{l=1}^{p+1}\sum_{k=1}^{p+1}\frac{\partial^3 Q_n(\theta^*,\gamma_n)}{\partial \theta_j \partial \theta_l \partial \theta_k}(\theta_l-\theta_{0l})(\theta_k-\theta_{0k}) - (n-L)p'_{\lambda_j}(|\theta_j|)\text{sign}(\theta_j),\end{aligned}$$

where θ^* lies between θ and θ_0. Here we assume $\left|(n-L)^{-1}\frac{\partial^3 Q_n(\theta,\gamma_n)}{\partial \theta_j \partial \theta_l \partial \theta_k}\right| \le M_{jlk}$, where $E(M_{jlk}) < \infty$. Note that

$$\begin{aligned}(n-L)^{-1/2}\frac{\partial Q_n(\theta_0,\gamma_0)}{\partial \theta_j} &= O_p(1),\\ (n-L)^{-1}\frac{\partial^2 Q_n(\theta_0,\gamma_0)}{\partial \theta_j \partial \theta_l} &= E\left\{\frac{\partial^2 Q_n(\theta_0)}{\partial \theta_j \partial \theta_l}\right\} + o_p(1),\end{aligned}$$

and

$$(n-L)^{-1}\frac{\partial^3 Q_n(\theta^*,\gamma_n)}{\partial \theta_j \partial \theta_l \partial \theta_k} = O_p(1).$$

Since $b_n = o_p(1)$ and $\sqrt{n}a_n = o_p(1)$, we obtain $\theta - \theta_0 = O_p(n^{-1/2})$. By $\sqrt{n}(\gamma_n - \gamma_0) = o_p(1)$, we have

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = (n-L)\lambda_j\left\{-\lambda_j^{-1}p'_{\lambda_j}(|\theta_j|)\text{sign}(\theta_j) + O_p((n-L)^{-1/2}/\lambda_j)\right\}.$$

Since $\frac{1}{\min_{s+1\le j\le p+1}\sqrt{n}\lambda_j} = o_p(1)$ and $\liminf_{n\to\infty}\liminf_{t\to 0^+}\lambda^{-1}p'_\lambda(|t|) > 0$ with probability 1, the sign of the derivative is completely determined by that of θ_j. This completes the proof of Theorem 2(a).

Proof of Theorem 2(b) It can be shown easily that there exists a $\hat{\theta}_{n1}$ in Theorem 1 that is a $\sqrt{n}$-consistent local maximizer of $\ell\{(\theta_1, 0)\}$, satisfying that

$$\frac{\partial \ell\{(\hat{\theta}_{n1}, 0)\}}{\partial \theta_j} = 0, \quad \text{for } j = 1, \ldots, s. \tag{21}$$

Note that $\hat{\theta}_{n1}$ is a consistent estimator,

$$\begin{aligned}
&\frac{\partial Q_n\{(\hat{\theta}_{n1}, 0), \gamma_n\}}{\partial \theta_j} - (n-L)p'_{\lambda_j}(|\theta_j|)\text{sign}(\theta_j) \\
&= \frac{\partial Q_n(\theta_0, \gamma_n)}{\partial \theta_j} + \sum_{l=1}^{s} \left\{ \frac{\partial^2 Q_n(\theta_0, \gamma_n)}{\partial \theta_j \partial \theta_l} + o_p(1) \right\} (\hat{\theta}_l - \theta_{0l}) \\
&\quad - (n-L)\Big[p'_{\lambda_j}(|\theta_{0j}|)\text{sign}(\theta_{0j}) + \left\{ p''_{\lambda_j}(|\theta_{0j}|) + o_p(1) \right\} (\hat{\theta}_j - \theta_{0j}) \Big] = 0.
\end{aligned}$$

The above equation can be rewritten as follows

$$\begin{aligned}
\frac{\partial Q_n(\theta_0, \gamma_n)}{\partial \theta_j} = \sum_{l=1}^{s} \left\{ E\left\{ -\frac{\partial^2 Q_n(\theta_0, \gamma_n)}{\partial \theta_j \partial \theta_l} \right\} + o_p(1) \right\} (n-L)(\hat{\theta}_l - \theta_{0l}) \\
+ (n-L)\Delta + (n-L)(\Sigma_1 + O_p(1))(\hat{\theta}_{n1} - \theta_{01}),
\end{aligned}$$

and

$$\begin{aligned}
&(n-L)I_1(\theta_{01}, \gamma_0)(\hat{\theta}_{n1} - \theta_{01}) + (n-L)\Delta + (n-L)(\Sigma_1 + O_p(1))(\hat{\theta}_{n1} - \theta_{01}) \\
&= (n-L)(I_1(\theta_{01}, \gamma_0) + \Sigma_1)(\hat{\theta}_{n1} - \theta_{01}) + (n-L)\Delta \\
&= (n-L)(I_1(\theta_{01}, \gamma_0) + \Sigma_1)\{(\hat{\theta}_{n1} - \theta_{01}) + (I_1(\theta_{01}, \gamma_0) + \Sigma_1)^{-1}\Delta\} \\
&= \frac{\partial Q_n(\theta_0, \gamma_n)}{\partial \theta_j} + o_p(1).
\end{aligned}$$

Since $\sqrt{n}(\gamma_n - \gamma_0) = o_p(1)$, invoking the Slutsky's lemma and the Lindeberg-Feller central limit theorem, we have

$$\sqrt{n-L}(I_1(\theta_{01}, \gamma_0) + \Sigma_1)\{(\hat{\theta}_{n1} - \theta_{01}) + (I_1(\theta_{01}, \gamma_0) + \Sigma_1)^{-1}\Delta\} \to N(\mathbf{0}, \Sigma_2),$$

where $\Sigma_1 = \text{diag}\{p''_{\lambda_j}(|\theta_{01}|), \ldots, p''_{\lambda_j}(|\theta_{0s}|)\}$, $\Sigma_2 = \text{Cov}(\exp(-r^2/\gamma_0)\frac{2r}{\gamma_0}\tilde{Z}_{i1})$, $\Delta = (p'_{\lambda_j}(|\theta_{01}|)\text{sign}(\theta_{01}), \ldots, p'_{\lambda_j}(|\theta_{0s}|)\text{sign}(\theta_{0s}))^T$, and $I_1(\theta_{01}, \gamma_0) = \frac{2}{\gamma_0}E[\exp(-r^2/\gamma_0)(\frac{2r^2}{\gamma_0} - 1)] \times (E\tilde{Z}_{i1}\tilde{Z}_{i1}^T)$. Then the proof of Theorem 2(b) is completed. □

Proof of Theorem 3 Let that $R(U_i) = g(U_i) - B(U_i)^T\eta$ and $R(U) = (R(U_1), \ldots, R(U_n))^T$. To facilitate expression, we set $Z = (WY, X), g(U) = (g(U_1), \ldots, g(U_n))^T$. Similar to Zhao et al. (2021), a simple calculation gives that

$$\begin{aligned}
\hat{\eta} - \eta &= (S^TS)^{-1}S^T(Y - Z\hat{\theta}_n) - \eta \\
&= (S^TS)^{-1}S^T(Z\theta_0 + g(U) + \epsilon - Z\hat{\theta}_n) - \eta \\
&= (S^TS)^{-1}S^T(Z\theta_0 + g(U) - S\eta + S\eta + \epsilon - Z\hat{\theta}_n) - \eta \\
&= (S^TS)^{-1}S^T(Z\theta_0 + R(U) + S\eta + \epsilon - Z\hat{\theta}_n) - \eta \\
&= (S^TS)^{-1}S^T(Z(\theta_0 - \hat{\theta}_n) + R(U) + \epsilon + S\eta) - \eta \\
&= (S^TS)^{-1}S^TR(U) + (S^TS)^{-1}S^T\epsilon + (S^TS)^{-1}S^TZ(\theta_0 - \hat{\theta}_n) \\
&= R_1 + R_2 + R_3.
\end{aligned} \tag{22}$$

By $\|R(u)\| = O(n^{-v/(2v+1)})$, we have

$$\|R_1\| \le \left(\frac{1}{n}\sum_{i=1}^{n}\|B(U_i)B(U_i)^T\|\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\|B(U_i)R(U_i)\| = O_p(n^{-v/(2v+1)}).$$

Note that $E\{B(U_i)\epsilon|X_i, U_i\} = 0$, then by the central limit theorem we have $n^{-1/2}\sum_{i=1}^{n} B(U_i)\epsilon_i = O_p(1)$. Therefore, we have

$$\|R_2\| \le \frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\|B(U_i)B(U_i)^T\|\right)^{-1}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}B(U_i)\epsilon_i\right\| = O_p(n^{-1/2}).$$

By Theorem 1, we have $\sqrt{n}(\theta - \hat{\theta}_n) = O_p(1)$. Similar to the above proof, we can obtain $\|R_3\| = O_p(n^{-1/2})$. Hence, we have

$$\|\hat{\eta} - \eta\| = O_p(n^{-v/(2v+1)} + n^{-1/2}) = O_p(n^{-v/(2v+1)}). \tag{23}$$

Therefore,

$$\begin{aligned}
\|\hat{g}(u) - g(u)\|^2 &= \int_0^1 \{\hat{g}(u) - g(u)\}^2 du \\
&= \int_0^1 \{B^T(u)\hat{\eta} - B^T(u)\eta + R(u)\}^2 du \\
&\le 2\int_0^1 \{B^T(u)\hat{\eta} - B^T(u)\eta\}^2 du + 2\int_0^1 R(u)^2 du \\
&= 2(\hat{\eta} - \eta)^T\int_0^1 B(u)B(u)^T du(\hat{\eta} - \eta) + 2\int_0^1 R(u)^2 du,
\end{aligned} \tag{24}$$

Note that $\|\int_0^1 B(u)B(u)^T du\| = O(1)$, and thus invoking (22) gives

$$(\hat{\eta}-\eta)^T \int_0^1 B(u)B(u)^T du(\hat{\eta}-\eta) = O_p(n^{-2v/(2v+1)}).$$

From $\|R(u)\| = O(n^{-v/(2v+1)})$, we have

$$\int_0^1 R(u)^2 du = O_p(n^{-2v/(2v+1)}).$$

As a result, $\|\hat{g}(u) - g(u)\|^2 = O_p(n^{-2v/(2v+1)})$. This completes the proof of Theorem 3. □

References

Basile, R. (2009). Productivity polarization across regions in Europe: The role of nonlinearities and spatial dependence. *International Regional Science Review, 32*(1), 92–115.

Case, A. C. (1991). Spatial patterns in household demand. *Econometrica, 59*(4), 953–965.

Cheng, S., Chen, J., Liu, X. (2019). GMM estimation of partially linear single-index spatial autoregressive model. *Spatial Statistics, 31*, 100354.

Du, J., Sun, X., Cao, R., et al. (2018). Statistical inference for partially linear additive spatial autoregressive models. *Spatial Statistics, 25*, 52–67.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.

Harrison, D., Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management, 5*(1), 81–102.

Jiang, Y., Ji, Q., Xie, B. (2017). Robust estimation for the varying coefficient partially nonlinear models. *Journal of Computational and Applied Mathematics, 326*, 31–43.

Jiang, Y., Tian, G. L., Fei, Y. (2019). A robust and efficient estimation method for partially nonlinear models via a new MM algorithm. *Statistical Papers, 60*(6), 2063–2085.

Kelejian, H. H., Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics, 17*(1), 99–121.

Kelejian, H. H., Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review, 40*(2), 509–533.

Koenker, R., Bassett, G., Jr. (1978). Regression quantiles. *Econometrica, 46*(1), 33–50.

Kong, E., Xia, Y. (2012). A single-index quantile regression model and its estimation. *Econometric Theory, 28*(4), 730–768.

Lee, L. F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica, 72*(6), 1899–1925.

Li, T., Guo, Y. (2020). Penalized profile quasi-maximum likelihood method of partially linear spatial autoregressive model. *Journal of Statistical Computation and Simulation, 90*(15), 2705–2740.

Li, T., Yin, Q., Peng, J. (2020). Variable selection of partially linear varying coefficient spatial autoregressive model. *Journal of Statistical Computation and Simulation, 90*(15), 2681–2704.

Luo, G., Wu, M. (2021). Variable selection for semiparametric varying-coefficient spatial autore-gressive models with a diverging number of parameters. *Communications in Statistics-Theory and Methods, 50*(9), 2062–2079.

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association, 70*(349), 120–126.

Schumaker, L. (1981). *Spline functions: Basic theory*. New York: Wiley.

Song, Y., Liang, X., Zhu, Y., et al. (2021). Robust variable selection with exponential squared loss for the spatial autoregressive model. *Computational Statistics and Data Analysis, 155*, 107094.

Su, L., Jin, S. (2010). Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. *Journal of Econometrics, 157*(1), 18–33.

Su, L., Yang, Z. (2007). Instrumental variable quantile estimation of spatial autoregressive models. In Development economics working papers 22476, East Asian Bureau of Economic Research. https://ideas.repec.org/p/eab/develo/22476.html.

Wang, H., Li, G., Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business & Economic Statistics, 25*(3), 347–355.

Wang, K., Lin, L. (2016). Robust structure identification and variable selection in partial linear varying coefficient models. *Journal of Statistical Planning and Inference, 174*, 153–168.

Wang, X., Jiang, Y., Huang, M., et al. (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association, 108*(502), 632–643.

Zhao, P., Gan, H., Cheng, S., et al. (2021). Orthogonality based penalized GMM estimation for variable selection in partially linear spatial autoregressive models. *Communications in Statistics-Theory and Methods, 52,* 1676–1691.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.

Zou, H., Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics, 36*(3), 1108–1126.