AIMS Mathematics, 11(4): 10133–10158.
DOI:10.3934/math.2026418
Received: 24 January 2026
Revised: 27 March 2026
Accepted: 02 April 2026
Published: 14 April 2026

https://www.aimspress.com/journal/Math

Research article

Variable selection for semi-parametric varying coefficient spatial error models

Yu Liu and Zengchao Xu*

Department of Mathematics, Shanghai Normal University, 100 Guilin RD, Shanghai 200234, China

* **Correspondence:** Email: xuzengchao@shnu.edu.cn.

Abstract: This paper develops an efficient variable selection method for semi-parametric varying coefficient spatial error models (SVCSEM) by integrating profile likelihood with adaptive LASSO penalization. The proposed method performs variable selection and model estimation simultaneously for the SVCSEM. The asymptotic normality and selection consistency of the resulting estimators are established under mild regularity conditions. Extensive numerical simulations demonstrate that the proposed method identifies the true effects accurately while delivering precise estimation for both parametric and nonparametric components under finite samples. Notably, simulation results highlight the superior performance of the adaptive LASSO, which consistently outperforms both the standard LASSO and the smoothly clipped absolute deviation (SCAD) penalty in terms of selection accuracy and estimation efficiency in the SVCSEM framework. In the analysis of China's outward foreign direct investment (OFDI) across 51 Belt and Road Initiative (BRI) countries, our findings reveal that institutional quality exerts significant nonlinear moderating effects on the relationship between gross domestic product (GDP) and OFDI, while also identifying the key determinants of investment. Further robustness analyses confirm the reliability of our methodology under alternative specifications of spatial weight matrices, affirming its broad applicability and effectiveness in empirical research.

Keywords: spatial error model; varying coefficient; adaptive LASSO; shrinkage estimator; Belt and Road initiative
Mathematics Subject Classification: 62H11, 62J07

1. Introduction

The increasing sophistication of spatial econometrics has highlighted the necessity of integrating both parametric and nonparametric elements into regression modeling, especially for capturing complex spatial dependencies. Traditional spatial models, which often rely on fixed coefficients, frequently fail to adequately represent heterogeneous spatial effects across different locations.

Consequently, research has progressively shifted towards semi-parametric spatial regression frameworks. Among these, semi-parametric varying coefficient spatial regression models (SVC-SRMs) have gained attention for allowing coefficients to vary smoothly over space while maintaining interpretability [1–3]. Despite these advantages, conducting variable selection within such models presents challenges, particularly under conditions of spatial autocorrelation and with finite sample sizes, where accurately distinguishing relevant predictors from irrelevant ones remains crucial.

To address these issues, penalization methods such as LASSO, adaptive LASSO, and smoothly clipped absolute deviation (SCAD) [4–6] have been extensively utilized in spatial modeling. These methods provide efficient and robust variable selection by shrinking the coefficients of irrelevant variables to zero. For instance, [7] introduced an adaptive LASSO penalized estimator for the spatial autoregressive (SAR) model, demonstrating its statistical consistency and asymptotic normality. In a similar vein, [8] incorporated the SCAD penalty into the SAR log-likelihood framework, developing a penalized quasi-maximum likelihood estimator that exhibits oracle properties. Furthermore, in high-dimensional settings, [9] proposed a novel penalized estimation approach by combining SCAD regularization with instrumental variables.

Although penalization techniques are well-studied in parametric spatial models, their application to semi-parametric spatial regression remains limited. Recent efforts have sought to fill this gap. For instance, [10] employed B-spline approximations and a two-stage least squares approach combined with SCAD for variable selection in semi-parametric varying coefficient models. More recently, [11] and [12] developed adaptive LASSO penalized estimators via profile maximum likelihood for semi-parametric varying-coefficient spatial panel models with random and fixed effects, respectively, establishing consistency and asymptotic normality. However, these studies have primarily focused on the spatial autoregressive (SAR) specification, where spatial dependence is incorporated through the dependent variable. In contrast, the spatial error model (SEM), which accounts for spatial autocorrelation in the error term and is of significant relevance in many economic and environmental studies, has received limited attention within the variable selection literature.

Recently, there has been growing interest in extending these ideas to accommodate temporal heterogeneity, leading to the development of space–time-varying coefficient models. [13] proposed a semi-parametric space–time-varying coefficient model using penalized spline decomposition, and [14] developed a Bayesian alternative via Gaussian process priors that relaxes functional form assumptions on the space–time interaction. These works represent significant advances in modeling complex spatiotemporal dynamics. However, the present paper addresses a distinct and complementary problem: variable selection in a cross-sectional spatial error model with varying coefficients. Unlike space–time models, our framework focuses on the case where coefficients vary smoothly over a univariate index (e.g., institutional quality), and spatial dependence is captured through the error term–a common structure in many economic and environmental applications.

Parallel to these developments, several advanced spatial variable selection methods have been proposed for models with specific structures. For instance, [15] developed a spatially weighted LASSO for high-dimensional spatial additive models, incorporating spatial correlation directly into the penalty. [16] introduced a geographically weighted group LASSO (GWGPL) that performs variable selection while accommodating spatial heterogeneity through shared structures across locations. [17] proposed a Bayesian group LASSO for confounded spatial data within the spatial generalized linear mixed model framework. Although these methods represent significant progress, they are tailored to

different model specifications–such as spatial additive models, geographically weighted regression, or spatial generalized linear mixed models–and are not directly applicable to the spatial error model with varying coefficients considered in this paper. We acknowledge their importance and discuss potential comparisons in future work (see Section 6).

To perform variable selection within the SEM framework, it is essential to choose an appropriate penalization method that ensures both reliable selection accuracy and computational efficiency during the iterative estimation for spatial models. Among the available methods, standard LASSO tends to introduce estimation bias due to its uniform shrinkage nature. In contrast, the SCAD method, despite having oracle properties, involves nonconvex optimization, which can lead to convergence issues. Given these considerations, our investigation employs adaptive LASSO. This method retains the convexity of the optimization problem and achieves oracle properties by utilizing data-driven weights, thereby enhancing stability in variable selection and computational efficiency. These attributes are particularly advantageous in the SEM context, where accurate identification and estimation of variables amidst potential unobserved spatial shocks are crucial.

This paper introduces the first systematic investigation of variable selection within semi-parametric varying coefficient spatial error models (SVCSEM) through the adaptive LASSO method. Our contributions can be summarized into three aspects: (1) We develop a novel variable selection procedure that integrates adaptive LASSO for the profile likelihood of SVCSEM, simultaneously achieving model estimation and variable selection. (2) Under mild regularity conditions, we establish asymptotic normality and selection consistency of the proposed estimator. (3) Through extensive Monte Carlo simulations, we assess the finite-sample performance of our method across diverse spatial structures and dimensions. Despite employing a standard estimation approach, its application to SVCSEM presents a novel contribution, providing empirical researchers with a computationally efficient and theoretically justified instrument for the analysis of spatially heterogeneous datasets.

The remainder of this paper is organized as follows. In Section 2, we introduce the SVCSEM framework and develop an adaptive, LASSO-penalized profile maximum likelihood estimator along with a practical computational algorithm for its implementation. Section 3 presents the asymptotic properties of the proposed estimator, and Section 4 provides extensive Monte Carlo simulation studies across various scenarios. In Section 5, we illustrate the application of our method through an analysis of China's investment patterns in countries participating in the Belt and Road Initiative. Section 6 concludes the paper, and technical proofs are presented in the Appendix.

2. Methodology

2.1. The semi-parametric varying coefficient spatial error model

Let $(\boldsymbol{x}_{n,i}, y_{n,i})$ be the observation collected from the ith ($i = 1, 2, \cdots, n$) subject of $(\boldsymbol{X}_n, \boldsymbol{Y}_n)$, where $y_{n,i}$ is the ith observation of response variable, and $\boldsymbol{x}_{n,i} = (x_{n,i1}, x_{n,i2}, \cdots, x_{n,ip})^\top$ acts as the p-dimensional regressors. A typical spatial error model assumes that

$$\begin{cases} y_{n,i} = \boldsymbol{x}_{n,i}^\top \boldsymbol{\beta}_0 + v_{n,i}, \\ v_{n,i} = \rho_0 \sum_{j=1}^{n} w_{n,ij} v_{n,j} + \varepsilon_{n,i}, \end{cases}$$

where $w_{n,ij}$ is a specified constant reflecting the strength of spatial connectivity between i and j, ρ_0 is the true spatial autoregressive coefficient of the error terms, $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the vector of true regression coefficients, and $\varepsilon_{n,i}$ $(i = 1, 2, \cdots, n)$ are independent and identically distributed (i.i.d.) random variables following $N(0, \sigma_0^2)$.

To capture underlying nonlinear associations among variables, we introduce a varying coefficient function into the spatial error model, thus developing a more flexible and generalized semi-parametric varying coefficient spatial error model (SVCSEM):

$$\begin{cases} y_{n,i} = \boldsymbol{x}_{n,i}^{\top}\boldsymbol{\beta}_0 + \boldsymbol{z}_{n,i}^{\top}\boldsymbol{\alpha}_0(u_{n,i}) + v_{n,i}, \\ v_{n,i} = \rho_0 \sum_{j=1}^{n} w_{n,ij} v_{n,j} + \varepsilon_{n,i}, \end{cases} \tag{2.1}$$

where $\boldsymbol{z}_{n,i} = (z_{n,i1}, z_{n,i2}, \cdots, z_{n,iq})^{\top}$ is the q-dimensional regressors, and $\boldsymbol{\alpha}_0(u_{n,i})$ is the true q-dimensional varying coefficient function vector. To avoid the curse of dimensionality, $u_{n,i}$ is assumed to be scalar. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^{\top}, \rho_0, \sigma_0^2)^{\top}$ be the true finite dimensional parameter vector; by suppressing the vector subscript n, denote $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)^{\top}$, $\boldsymbol{Y} = (y_1, y_2, \cdots, y_n)^{\top}$, $\boldsymbol{W} = (w_{ij})_{1\le i,j\le n}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^{\top}$, $\boldsymbol{v} = (v_1, v_2, \cdots, v_n)^{\top}$, $\boldsymbol{M}_0 = (\boldsymbol{z}_1^{\top}\alpha_0(u_1), \boldsymbol{z}_2^{\top}\alpha_0(u_2), \cdots, \boldsymbol{z}_n^{\top}\alpha_0(u_n))^{\top}$. The SVCSEM (2.1) can be concisely written as

$$\begin{cases} \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{M}_0 + \boldsymbol{v}, \\ \boldsymbol{v} = \rho_0 \boldsymbol{W}\boldsymbol{v} + \boldsymbol{\varepsilon}. \end{cases} \tag{2.2}$$

2.2. Model estimation

Throughout this paper, let $\boldsymbol{I}$ be an identity matrix of size n, $\boldsymbol{S}(\rho) = \boldsymbol{I} - \rho\boldsymbol{W}$ for any value of ρ, and $\boldsymbol{\varepsilon}(\boldsymbol{\varsigma}) = \boldsymbol{S}(\rho)(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{M})$, where $\boldsymbol{\varsigma} = (\boldsymbol{\beta}^{\top}, \rho)^{\top}$. Under the setup above, the profile log-likelihood function for Model (2.2) has the form

$$\log L(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) + \log|\boldsymbol{S}(\rho)| - \frac{1}{2\sigma^2}\boldsymbol{\varepsilon}(\boldsymbol{\varsigma})^{\top}\boldsymbol{\varepsilon}(\boldsymbol{\varsigma}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\top}, \rho, \sigma^2)^{\top}$. To estimate the varying coefficient component $\boldsymbol{M}$, we adopt the local linear approximation approach of [18] for $\alpha(u_i)$. The estimation is carried out in four steps:

Step 1 *(Local log-likelihood specification):* For each location u, define the local log-likelihood:

$$\begin{aligned} \log\tilde{L}(\boldsymbol{\theta}) = &-\frac{n}{2}\log(2\pi\sigma^2) + \log|\boldsymbol{S}(\rho)| \\ &- \frac{1}{2\sigma^2}\{\boldsymbol{Y}^* - \boldsymbol{P}(u)\boldsymbol{\delta}(u)\}^{\top}\boldsymbol{K}^{1/2}(u)\boldsymbol{\Psi}^{-1}(\rho)\boldsymbol{K}^{1/2}(u)\{\boldsymbol{Y}^* - \boldsymbol{P}(u)\boldsymbol{\delta}(u)\}, \end{aligned} \tag{2.3}$$

where $\boldsymbol{\delta}(u) = (\alpha_1(u), \cdots, \alpha_q(u), h\alpha_1'(u), \cdots, h\alpha_q'(u))^{\top}$, $\boldsymbol{Y}^* = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$, $\boldsymbol{\Psi}(\rho) = (\boldsymbol{S}(\rho)^{\top}\boldsymbol{S}(\rho))^{-1}$, $\boldsymbol{K}(u)$ is an n-dimensional diagonal matrix with its (i,i)-entry being $k_h(u_i - u) = h^{-1}k((u_i - u)/h)$, $k((u_i - u)/h)$ is the kernel function, h represents the bandwidth that can be determined using the rule of thumb, and $\boldsymbol{P}(u)$ is an $n \times 2q$ matrix with the ith row being $(\boldsymbol{z}_i^{\top}, (u_i - u)\boldsymbol{z}_i^{\top}/h)$.

Step 2 *(Local polynomial estimation):* Given (ρ, σ^2), maximize (2.3) with respect to $\boldsymbol{\delta}(u)$. We adopt a working independence approximation to simplify the local estimation of the varying-coefficient functions. As established in [18] and [19], this approach yields $\sqrt{n}$-consistent estimators for

the parametric components and consistent estimates for the nonparametric components under standard regularity conditions. The validity of this simplification in our context rests on a two-stage strategy: the spatial dependence captured by the error structure is temporarily set aside during this local smoothing step but is fully accounted for in the subsequent profile likelihood estimation of ρ and $\boldsymbol{\beta}$. Therefore, this computationally convenient approximation does not compromise the consistency of the final estimators. Following this approach, we use:

$$\hat{\boldsymbol{\delta}}(u) = \arg\min_{\boldsymbol{\delta}(u)} \left[\boldsymbol{Y}^* - \boldsymbol{P}(u)\boldsymbol{\delta}(u)\right]^\top \boldsymbol{K}(u) \left[\boldsymbol{Y}^* - \boldsymbol{P}(u)\boldsymbol{\delta}(u)\right],$$

which has the closed-form solution $\hat{\boldsymbol{\delta}}(u) = \left[\boldsymbol{P}^\top(u)\boldsymbol{K}(u)\boldsymbol{P}(u)\right]^{-1} \boldsymbol{P}^\top(u)\boldsymbol{K}(u)\boldsymbol{Y}^*$. Therefore, the varying coefficient functions are estimated as

$$\hat{\boldsymbol{\alpha}}(u) = \boldsymbol{e}_0^\top \hat{\boldsymbol{\delta}}(u), \quad \boldsymbol{e}_0 = (\boldsymbol{I}_q, \boldsymbol{0}_q)^\top,$$

and $\boldsymbol{M}$ is approximated by $\hat{\boldsymbol{M}} = \boldsymbol{A}\boldsymbol{Y}^*$, where $\boldsymbol{A}$ is the $n \times n$ smoothing matrix with ith row being $[z_i^\top \boldsymbol{0}_q][\boldsymbol{P}^\top(u_i)\boldsymbol{K}(u_i)\boldsymbol{P}(u_i)]^{-1}\boldsymbol{P}^\top(u_i)\boldsymbol{K}(u_i)$.

Step 3 *(Profile maximum log-likelihood estimation):* Given $\hat{\boldsymbol{M}}$, the estimated profile log-likelihood becomes

$$\log \hat{L}(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) + \log|\boldsymbol{S}(\rho)| - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{I} - \boldsymbol{A})^\top \boldsymbol{\Psi}^{-1}(\rho)(\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}). \quad (2.4)$$

Maximizing (2.4) yields closed-form expressions for $\boldsymbol{\beta}$ and σ^2 conditional on ρ:

$$\hat{\boldsymbol{\beta}}(\rho) = \{\tilde{\boldsymbol{X}}^\top \boldsymbol{\Psi}^{-1}(\rho)\tilde{\boldsymbol{X}}\}^{-1}\tilde{\boldsymbol{X}}^\top \boldsymbol{\Psi}^{-1}(\rho)\tilde{\boldsymbol{Y}}, \quad (2.5)$$

$$\hat{\sigma}^2(\rho) = \frac{1}{n}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}(\rho))^\top \boldsymbol{\Psi}^{-1}(\rho)(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}(\rho)), \quad (2.6)$$

where $\tilde{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{Y}$, $\tilde{\boldsymbol{X}} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{X}$.

Step 4 *(Spatial parameter estimation):* Substitute (2.5) and (2.6) into (2.4) to obtain the concentrated log-likelihood for ρ:

$$\log \hat{L}(\rho) = -\frac{n}{2}(\log 2\pi + 1) + \log|\boldsymbol{S}(\rho)| - \frac{n}{2}\log(\hat{\sigma}^2(\rho)). \quad (2.7)$$

The spatial parameter estimate $\hat{\rho}$ can be estimated via maximizing (2.7), thereby providing the final estimates $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$, and $\hat{\boldsymbol{\alpha}}(u)$.

2.3. The adaptive LASSO procedure

In practice, high-dimensional variables are common. Improper handling may reduce the model's explanatory power and lead to biased estimation. For the semi-parametric varying coefficient spatial error model proposed above, conducting variable selection alongside model estimation is essential when dealing with high-dimensional X.

Adaptive LASSO is particularly well-suited for the SVCSEM framework due to its unique combination of convexity, oracle properties, and computational efficiency. First, by retaining the convex optimization structure of the standard LASSO, it ensures numerical stability and global

convergence–a crucial advantage given the iterative nature of profile likelihood estimation in spatial models. Second, through data-driven weights that are inversely proportional to initial consistent estimates, adaptive LASSO attains the oracle property [6]: it identifies the true nonzero coefficients with probability approaching one and estimates them as efficiently as if the true model were known. This property is especially valuable in spatial error models, where residual spatial dependence can obscure the identification of relevant predictors. Third, unlike SCAD, which relies on nonconvex optimization and can be sensitive to starting values, adaptive LASSO offers a computationally efficient path to sparse solutions without compromising asymptotic optimality. These features make it an ideal tool for variable selection in settings characterized by both spatial autocorrelation and varying coefficients.

Therefore, we plan to employ adaptive LASSO to construct the variable selection process. First, we substitute the $\hat{\rho}$ into Eq (2.4). Then, adapting the penalized likelihood framework for variable selection from classical linear models to our spatial context, we obtain the adaptive LASSO estimator $\tilde{\boldsymbol{\beta}}$ by solving the following optimization problem:

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})^{\top}\boldsymbol{\Psi}^{-1}(\hat{\rho})(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_j|\beta_j|\}, \tag{2.8}$$

where the terms $\lambda_j|\beta_j|(j = 1, 2, \cdots, p)$ represent adaptive LASSO penalty functions. The tuning parameters $\lambda_j(j = 1, 2, \cdots, p)$ assign distinct weights to each coefficient β_j. Selecting p distinct shrinkage parameters simultaneously is challenging. To overcome this difficulty, we adopt [6]'s approach by setting $\lambda_j = \lambda/\hat{\beta}_j^{init}$, where $\hat{\beta}_j^{init}$ is an initial consistent estimator (e.g., from unpenalized estimation). Obviously, the first term in (2.8) derives from (2.4) by omitting components independent of $\boldsymbol{\beta}$.

2.4. Algorithm implementation

Equation (2.8) represents a standard LASSO-type problem. For computational efficiency, we apply the local quadratic approximation algorithm [5] to approximate the penalty function locally. Specifically, let $\tilde{\boldsymbol{\beta}}^{(k)}$ is the shrinkage estimate of $\boldsymbol{\beta}$ in the kth iteration. The iteration is given by

$$\begin{aligned}\tilde{\boldsymbol{\beta}}^{(k+1)} &= \arg\min_{\boldsymbol{\beta}} \left\{(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})^{\top}\boldsymbol{\Psi}^{-1}(\hat{\rho})(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_j \frac{\beta_j^2}{|\tilde{\beta}_j^{(k)}|}\right\} \\ &= \{\tilde{\boldsymbol{X}}^{\top}\boldsymbol{\Psi}^{-1}(\hat{\rho})\tilde{\boldsymbol{X}} + \boldsymbol{D}^{(k)}\}^{-1}\tilde{\boldsymbol{X}}^{\top}\boldsymbol{\Psi}^{-1}(\hat{\rho})\tilde{\boldsymbol{Y}},\end{aligned} \tag{2.9}$$

where $\boldsymbol{D}^{(k)} = \text{diag}\{\lambda_j/|\tilde{\beta}_j^{(k)}|\}$ is a $p \times p$ diagonal matrix. To ensure stable adaptive LASSO variable selection, we implement the following iterative algorithm.

Step 1: Initialize $\tilde{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the nonpenalized estimate from Section 2.2. Set $k = 0$ and $tol = 10^{-6}$.

Step 2: Prune insignificant variables: if $|\tilde{\beta}_j^{(k)}| < tol$, set $\tilde{\beta}_j^{(k+1)} = 0$, remove the jth column from $\tilde{\boldsymbol{X}}$ and the jth diagonal element from $\boldsymbol{D}^{(k)}$.

Step 3: Update coefficients using fixed $\hat{\rho}$:

$$\tilde{\boldsymbol{\beta}}^{(k+1)} = \left[\tilde{\boldsymbol{X}}^{(k)\top}\boldsymbol{\Psi}^{-1}(\hat{\rho})\tilde{\boldsymbol{X}}^{(k)} + \boldsymbol{D}^{(k)}\right]^{-1}\tilde{\boldsymbol{X}}^{(k)\top}\boldsymbol{\Psi}^{-1}(\hat{\rho})\tilde{\boldsymbol{Y}},$$

where $\boldsymbol{D}^{(k)}$ has diagonal elements $\lambda_j/|\tilde{\beta}_j^{(k)}|$ for retained coefficients.

Step 4: Update parameters and select λ:

(i) Compute $\tilde{\sigma}^2$ via (2.6)

(ii) Solve (2.7) for $\tilde{\rho}$

(iii) Select λ by minimizing BIC:

$$\text{BIC} = n\log(\tilde{\sigma}^2) - 2\log|\boldsymbol{S}(\tilde{\rho})| + \log(n)\cdot\|\tilde{\boldsymbol{\beta}}^{(k+1)}\|_0,$$

where $\|\tilde{\boldsymbol{\beta}}^{(k+1)}\|_0$ is the L_0 norm of $\tilde{\boldsymbol{\beta}}$, which denotes the number of nonzero elements in the vector $\tilde{\boldsymbol{\beta}}$.

Step 5: Terminate if $\max_j|\tilde{\beta}_j^{(k+1)} - \tilde{\beta}_j^{(k)}| < tol$; else increase k and return to Step 2.

Remark: The tuning parameter λ controls the trade-off between model fit and sparsity. In our implementation, we select λ by minimizing the Bayesian information criterion (BIC) defined in Step 4(iii). This BIC incorporates the Jacobian term $-2\log|S(\tilde{\rho})|$, which accounts for the spatial dependence by reflecting the transformation from spatially correlated errors $\boldsymbol{\nu}$ to independent errors $\boldsymbol{\varepsilon}$ in the likelihood (see Eq (2.4)). Including this term is standard practice in spatial model selection, and this BIC-type criterion has been shown to be consistent for model selection in spatial econometric models [7, 8]. Although there is no closed-form theoretically optimal λ for finite samples, the BIC-based selection achieves the oracle property asymptotically under the regularity conditions in Section 3, ensuring that the selected model converges to the true sparse model as $n \to \infty$. In practice, we search over a grid of λ values and choose the one that minimizes BIC, which balances goodness-of-fit and model complexity effectively.

3. Asymptotic properties

Firstly, we introduce the following quantities:

$$\boldsymbol{G}(\rho) = \boldsymbol{W}\boldsymbol{S}^{-1}(\rho),\ \boldsymbol{G} = \boldsymbol{W}\boldsymbol{S}^{-1},\ \boldsymbol{\Phi} = \lim_{n\to\infty}\frac{1}{n}E(\boldsymbol{X}^\top\boldsymbol{\Delta}\boldsymbol{X}),$$
$$\boldsymbol{\Delta} = (\boldsymbol{I}-\boldsymbol{A})^\top\boldsymbol{S}^\top\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{A}),\ \boldsymbol{\Omega} = \lim_{n\to\infty}\frac{1}{n}E(\boldsymbol{X}^\top\boldsymbol{\Delta}\boldsymbol{\Psi}\boldsymbol{\Delta}\boldsymbol{X}),$$
$$\tau_1 = \lim_{n\to\infty}\frac{1}{n}\left[\text{tr}(\boldsymbol{G}^2) + \text{tr}(\boldsymbol{G}\boldsymbol{G}^\top)\right],\ \tau_2 = \lim_{n\to\infty}\frac{1}{n}\text{tr}(\boldsymbol{G}).$$

Let $p_0 < p$ denote the number of nonzero elements in the true coefficient vector $\boldsymbol{\beta}_0$. Denote $\mathcal{A} = \{j : \beta_{0j} \neq 0\}$. Without loss of generality, assume $\mathcal{A} = \{1,\ldots,p_0\}$, and consequently, $\mathcal{A}^c = \{p_0+1,\ldots,p\}$. Let $a_n = \max\{\lambda_j : j\in\mathcal{A}\}$ and $b_n = \min\{\lambda_j : j\in\mathcal{A}^c\}$ denote the maximal and minimal amounts of penalties applied to relevant and irrelevant coefficients, respectively. Furthermore, denote $\mathcal{A}_n = \{j : \tilde{\beta}_j \neq 0\}$, $\boldsymbol{X} = (\boldsymbol{X}_{\mathcal{A}}, \boldsymbol{X}_{\mathcal{A}^c})$, and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0\mathcal{A}}^\top, \boldsymbol{\beta}_{0\mathcal{A}^c}^\top)^\top$. The matrix $\boldsymbol{\Phi}$ has the partitioned form $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_{11}, \boldsymbol{\Phi}_{12}; \boldsymbol{\Phi}_{21}, \boldsymbol{\Phi}_{22}]$, where $\boldsymbol{\Phi}_{11}$ is the $p_0\times p_0$ submatrix corresponding to the active set $\mathcal{A}$.

The asymptotic normality and selection consistency of the proposed estimators are established under the following regularity conditions.

Assumption 1. *(i) $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{z}_i\}_{i=1}^n$ are i.i.d. with compact support. $\{\varepsilon_i\}_{i=1}^n$ are i.i.d., independent of $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}$, satisfying $E(\varepsilon_i \mid \boldsymbol{x}_j, \boldsymbol{z}_j) = 0$ and $Var(\varepsilon_i \mid \boldsymbol{x}_j, \boldsymbol{z}_j) = \sigma^2 < \infty$. The matrix $E(\boldsymbol{x}_1\boldsymbol{x}_1^\top)$ is positive definite. For some $s > 2$, $E\|\boldsymbol{x}_1\|^{2s} < \infty$, $E\|\boldsymbol{z}_1\|^{2s} < \infty$, and $n^{2\zeta-1}h \to \infty$ with $\zeta < 2 - s^{-1}$. (ii) The sequence $\{u_i\}$ is i.i.d. with twice continuously differentiable density $f(u)$ satisfying*

$0 < \inf_u f(u) \leq \sup_u f(u) < \infty$ *on its support. (iii) The coefficient functions* $\alpha_j(u)$ $(j = 1, \ldots, q)$ *are twice continuously differentiable and uniformly bounded.*

Assumption 2. *(i) Elements* w_{ij} *of* $\boldsymbol{W}$ *satisfy* $\max_{i,j} |w_{ij}| = O(l_n^{-1})$ *uniformly, where* $l_n/n \to 0$ *as* $n \to \infty$. *(ii) Matrices* $\boldsymbol{W}$ *and* $\boldsymbol{S}^{-1}$ *are uniformly bounded in row and column sums. (iii) For* ρ *in compact parameter space* Λ *(with* ρ_0 *interior),* $\boldsymbol{S}^{-1}(\rho)$ *is uniformly bounded in either row or column sums.*

Assumption 3. *(i) The kernel* $k(\cdot)$ *is a symmetric, non-negative density with* $\mu_\nu = \int u^\nu k(u)du$ *and* $\nu_\nu = \int u^\nu k^2(u)du$ *finite for* $\nu = 0, 1, 2$. *(ii) The bandwidth* h *satisfies* $nh^8 \to 0$, $nh^2/(\log n)^2 \to \infty$, *and* $\sqrt{n}h^2 \to 0$.

Assumption 4. *(i)* $\Phi > 0$. *(ii)* $\lim_{n\to\infty} \left\{ \frac{1}{n} \mathrm{tr}(\boldsymbol{G}^2 + \boldsymbol{G}\boldsymbol{G}^\top) - \frac{2}{n^2}[\mathrm{tr}(\boldsymbol{G})]^2 \right\} > 0$.

Theorem 1. *Under Assumptions 1–4, if* $a_n/\sqrt{n} \to 0$ *and* $b_n/\sqrt{n} \to \infty$,

(i) (asymptotic normality) $\sqrt{n}(\tilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) \xrightarrow{d} \boldsymbol{\Phi}_{11}^{-1}\boldsymbol{\xi}_{\mathcal{A}}$, *where* $\boldsymbol{\xi}_{\mathcal{A}}$ *is a* p_0*-dimensional subvector of* $\boldsymbol{\xi}$, *which follows a normal distribution of* $N(\boldsymbol{0}, \sigma_0^2\boldsymbol{\Omega})$.

(ii) (selection consistency) $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$.

The proof of Theorem 1 is delayed in the supplementary document. Here, we make some remarks on the technical assumptions.

Remark 1. Assumption 1 establishes conditions on model variables, Assumption 2 constrains spatial dependence structures, and Assumption 3 specifies kernel smoothing requirements. The conditions in Assumption 1 and Assumption 3 are standard for semi-parametric varying coefficient models [18], whereas those in Assumption 2 conform to spatial econometric conventions [20].

Remark 2. Assumption 4 ensures the concentrated likelihood for ρ to have a unique maximizer, which is essential for the $\sqrt{n}$-convergence rate of the adaptive LASSO estimator.

4. Simulation studies

This section presents comprehensive Monte Carlo simulations to evaluate the performance of the proposed variable selection method for semi-parametric varying coefficient spatial error models.

4.1. Simulation settings

The data are generated from one of the following two model specifications:

(1) Model 1 (Single varying coefficients): $z_i \sim$ Uniform$(-2, 2)$, and the varying coefficient function is $\alpha(u) = \sin(2\pi u) + 2u$.

(2) Model 2 (Two varying coefficients): $z_{1i} \sim N(0, 1)$, $z_{2i} \sim$ Uniform$(-2, 2)$, and the varying coefficient functions are specified as

$$\alpha_1(u) = \sin(2\pi u) + 2u,$$
$$\alpha_2(u) = 3.5\left(\exp(-(4u-1)^2) + \exp(-(4u-3)^2)\right) - 1.5.$$

The spatial weight matrix is taken from one of the following configurations:

(1) Rook contiguity: $w_{ij} = \mathbb{I}\{\text{units } i \text{ and } j \text{ are adjacent}\}$, $n \in \{49, 64, 81, 100, 225, 400\}$

(2) Case structure [21]: $\boldsymbol{W} = \boldsymbol{I}_R \otimes \boldsymbol{B}_m$, where $\boldsymbol{B}_m = (m-1)^{-1}(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)$ with R districts and m units per district, $R \in \{5, 10\}$, $m \in \{10, 20, 30\}$

The remaining parameter settings are as follows: (1) The dimension of covariates, p, is taken from $\{6, 20, 50\}$, and the sparsity level, p_{zero}, is 4, 10, or 40. (2) $u_i \sim \text{Uniform}(0, 1)$, and the covariates $\boldsymbol{x}_i \in \mathbb{R}^p$ follow $N_p(\mathbf{0}, \Sigma)$ with $\Sigma = (0.5^{|i-j|})$ for $i, j = 1, \ldots, p$. The errors ε_i are i.i.d. from $N(0, \sigma_\varepsilon^2)$ with $\sigma_{\varepsilon 0}$ being 0.5 or 1. (3) Spatial autocorrelation ρ_0 is taken as $\{-0.8, -0.5, -0.2, 0, 0.2, 0.5, 0.8\}$.

The true parameter vector is specified as one of the following settings:

$$\begin{aligned}
\boldsymbol{\beta}_{01} &= (1, 1.5, \mathbf{0}_4)^\top \quad (\text{for } p = 6,\ p_{\text{zero}} = 4),\\
\boldsymbol{\beta}_{02} &= (\mathbf{1}_{10}^\top, \mathbf{0}_{10}^\top)^\top \quad (\text{for } p = 20,\ p_{\text{zero}} = 10),\\
\boldsymbol{\beta}_{03} &= (\mathbf{1}_{10}^\top, \mathbf{0}_{40}^\top)^\top \quad (\text{for } p = 50,\ p_{\text{zero}} = 40).
\end{aligned}$$

4.2. Analysis of numerical results

We carry out extensive simulations with 500 replications for each combination of ρ_0, $\sigma_{\varepsilon_0}^2$, $\boldsymbol{\beta}_0$, and n. The performance metrics include three aspects:

(1) **Variable selection**:
   (a) C: Average count of correctly identified zero coefficients
   (b) IC: Average count of incorrectly shrunk nonzero coefficients
   (c) Correct (%): Proportion of correct model selections

(2) **Parametric estimation**: Bias and standard error (SE) for ρ and $\boldsymbol{\beta}$

(3) **Nonparametric estimation**: Root mean square error (RMSE) of $\hat{\alpha}(u)$ over 100 equidistant points:

$$\text{RMSE} = \sqrt{\frac{1}{100}\sum_{i=1}^{100}\left[\hat{\alpha}(u_i) - \alpha(u_i)\right]^2}.$$

The median of RMSE across replications is reported.

4.2.1. Baseline scenarios

Tables 1 and 2 present comparative results for Rook and Case weight matrices under the baseline configuration of model 1 ($p = 6$, $p_{\text{zero}} = 4$, $\boldsymbol{\beta}_{0,1} = (1, 1.5, \mathbf{0}_4)^\top$, $\sigma_{\varepsilon_0} = 0.5$, $\rho_0 \in \{0.2, 0.5, 0.8\}$). The simulations reveal three consistent patterns: First, variable selection performance improves markedly with sample size, evidenced by correct selection rates exceeding 95% for $n \geq 225$ in Table 1, whereas C approaches the true value of 4, and IC converges to 0 as n increases. Second, the precision of parametric estimation demonstrates expected behavior: we find that for the same ρ_0, standard errors of both $\tilde{\rho}$ and $\tilde{\sigma}_\varepsilon$ decrease systematically with larger n, and $\tilde{\sigma}_\varepsilon$'s standard error remains remarkably stable across different ρ_0 values at fixed sample sizes. However, we find that the standard errors of $\tilde{\rho}$ depend on the true value of ρ and sample size. Third, for nonparametric estimation, the RMSE decreases monotonically with increasing n. Moreover, when the total sample size $n = m \times R$ is held constant, the RMSE remains invariant to ρ_0. Both of these findings are visually confirmed in Figure 1. Collectively, these findings underscore the robustness of the proposed adaptive LASSO penalized estimator against variations in spatial weight structures and autocorrelation intensities.

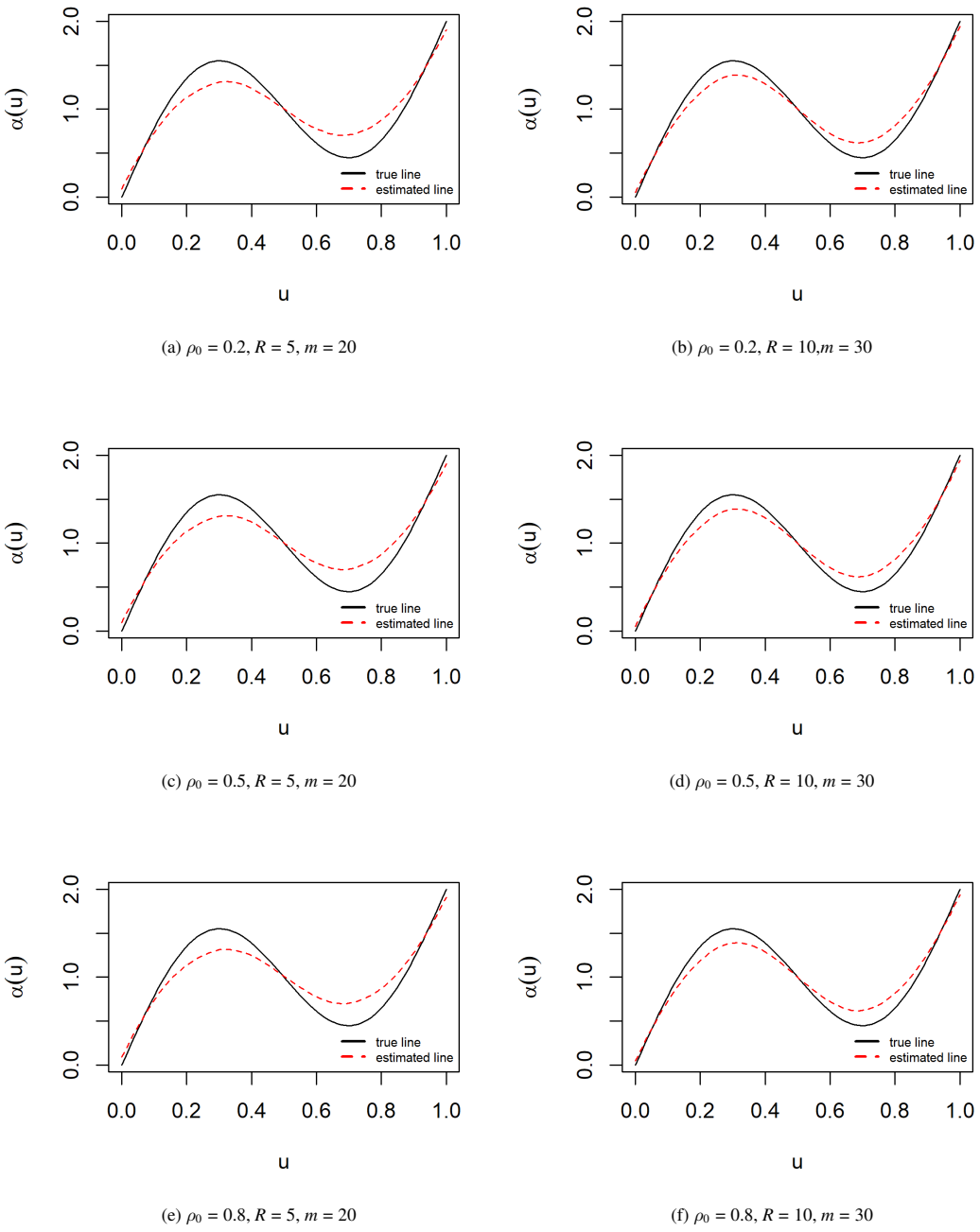
Table 1. Simulation results for Model 1, with Rook matrix, $p = 6$, $p_{zero} = 4$, and $\sigma_{\varepsilon 0} = 0.5$.

ρ_0	n	Correct(%)	C	IC	ρ	$\sigma_{\varepsilon 0}$	RMSE
0.2	49	82.2	3.736	0	0.012(0.180)	0.003(0.054)	0.252
	64	89.0	3.858	0	−0.016(0.153)	0.012(0.046)	0.223
	81	89.8	3.872	0	−0.017(0.141)	0.015(0.038)	0.206
	100	91.6	3.900	0	−0.017(0.128)	0.010(0.035)	0.192
	225	95.6	3.952	0	−0.014(0.091)	0.012(0.023)	0.143
	400	97.0	3.970	0	−0.009(0.066)	0.009(0.018)	0.117
0.5	49	81.4	3.728	0	−0.074(0.193)	0.013(0.059)	0.259
	64	89.2	3.864	0	−0.075(0.157)	0.021(0.048)	0.230
	81	90.4	3.888	0	−0.067(0.143)	0.023(0.040)	0.213
	100	90.6	3.886	0	−0.060(0.126)	0.017(0.037)	0.196
	225	96.0	3.956	0	−0.033(0.079)	0.017(0.024)	0.146
	400	97.0	3.970	0	−0.023(0.056)	0.011(0.019)	0.119
0.8	49	83.0	3.752	0	−0.097(0.138)	0.053(0.075)	0.294
	64	89.4	3.878	0	−0.082(0.107)	0.052(0.060)	0.257
	81	90.4	3.876	0	−0.072(0.095)	0.049(0.049)	0.235
	100	90.8	3.890	0	−0.064(0.084)	0.039(0.046)	0.211
	225	95.4	3.946	0	−0.035(0.051)	0.029(0.028)	0.159
	400	97.2	3.972	0	−0.023(0.035)	0.019(0.021)	0.128

Note: The penultimate column and the antepenultimate column represent bias and standard deviations in parentheses. This notation is consistently applied in all subsequent tables.

Table 2. Simulation results for Model 1, with Case matrix, $p = 6$, $p_{zero} = 4$, $\sigma_{\varepsilon 0} = 0.5$, and $n = mR$.

ρ_0	R	m	Correct(%)	C	IC	ρ	$\sigma_{\varepsilon 0}$	RMSE
0.2	5	10	85.4	3.784	0	−0.036(0.178)	0.009(0.050)	0.248
		20	91.8	3.902	0	−0.033(0.167)	0.010(0.035)	0.191
		30	94.8	3.944	0	−0.040(0.166)	0.014(0.029)	0.164
	10	10	91.4	3.894	0	−0.039(0.145)	0.010(0.035)	0.191
		20	96.0	3.958	0	−0.032(0.138)	0.014(0.023)	0.144
		30	95.8	3.954	0	−0.037(0.138)	0.010(0.021)	0.129
0.5	5	10	85.4	3.798	0	−0.116(0.204)	0.015(0.052)	0.252
		20	92.2	3.906	0	−0.101(0.194)	0.012(0.035)	0.193
		30	94.8	3.944	0	−0.102(0.194)	0.015(0.029)	0.164
	10	10	91.6	3.898	0	−0.073(0.145)	0.014(0.036)	0.195
		20	96.2	3.958	0	−0.056(0.134)	0.015(0.024)	0.145
		30	95.6	3.952	0	−0.057(0.133)	0.011(0.021)	0.129
0.8	5	10	87.6	3.828	0	−0.075(0.112)	0.052(0.065)	0.291
		20	92.0	3.904	0	−0.058(0.107)	0.022(0.038)	0.206
		30	94.8	3.940	0	−0.056(0.113)	0.021(0.030)	0.172
	10	10	92.8	3.914	0	−0.043(0.066)	0.037(0.044)	0.222
		20	96.4	3.958	0	−0.028(0.055)	0.022(0.026)	0.156
		30	96.2	3.960	0	−0.026(0.056)	0.014(0.021)	0.136



(a) $\rho_0 = 0.2$, $R = 5$, $m = 20$

(b) $\rho_0 = 0.2$, $R = 10$,$m = 30$

(c) $\rho_0 = 0.5$, $R = 5$, $m = 20$

(d) $\rho_0 = 0.5$, $R = 10$, $m = 30$

(e) $\rho_0 = 0.8$, $R = 5$, $m = 20$

(f) $\rho_0 = 0.8$, $R = 10$, $m = 30$

Figure 1. Fitted curves of $\alpha(U)$ for the settings in Table 2.

4.2.2. Impact of error variance

Tables 3–5 investigate higher-dimensional scenarios of Model 1 ($p = 20, 50$) with varying error variances ($\sigma_{\varepsilon_0} \in \{0.5, 1\}$). Under low-noise conditions ($\sigma_{\varepsilon_0} = 0.5$), the method maintains robust performance with correct selection rates exceeding 90% even at $p = 50$ (Table 5) while consistently achieving $C \approx p_{\text{zero}}$ and $IC \approx 0$. In contrast, high-noise settings ($\sigma_{\varepsilon_0} = 1$) induce a substantial

performance decline, where correct selection drops to approximately 60% at $p = 20$ (Table 4); however, this degradation proves recoverable through sample size augmentation, with all other metrics aligning with low-noise results. Notably, both C and IC converge to their theoretical values across all configurations, demonstrating the estimator's dimensional stability.

Table 3. Simulation results for Model 1 with Rook matrix, $p = 20$, $p_{zero} = 10$, $\sigma_{\varepsilon 0} = 0.5$, and $n = 225$.

ρ_0	Correct(%)	C	IC	ρ	$\sigma_{\varepsilon 0}$	RMSE
−0.8	93.2	9.914	0	0.025(0.050)	0.016(0.028)	0.159
−0.5	91.6	9.894	0	0.024(0.086)	0.005(0.024)	0.148
−0.2	90.6	9.886	0	0.014(0.104)	0.001(0.023)	0.145
0	91.8	9.900	0	0.009(0.113)	0.001(0.023)	0.144
0.2	91.6	9.896	0	0.003(0.105)	0.001(0.023)	0.144
0.5	91.2	9.886	0	−0.011(0.090)	0.004(0.025)	0.148
0.8	90.4	9.880	0	−0.019(0.052)	0.015(0.028)	0.162

Table 4. Simulation results for Model 1 with Rook matrix, $p = 20$, $p_{zero} = 10$, $\sigma_{\varepsilon 0} = 1$, and $n = 225$.

ρ_0	Correct(%)	C	IC	ρ	$\sigma_{\varepsilon 0}$	RMSE
−0.8	60.8	9.506	0	0.008(0.046)	−0.011(0.053)	0.224
−0.5	62.6	9.534	0	0.000(0.083)	−0.029(0.048)	0.187
−0.2	62.6	9.530	0	0.003(0.104)	−0.032(0.047)	0.178
0	64.6	9.564	0	0.010(0.113)	−0.033(0.047)	0.175
0.2	63.0	9.552	0	0.016(0.106)	−0.033(0.048)	0.179
0.5	62.6	9.530	0	0.013(0.085)	−0.031(0.049)	0.187
0.8	60.6	9.524	0	−0.002(0.047)	−0.013(0.052)	0.224

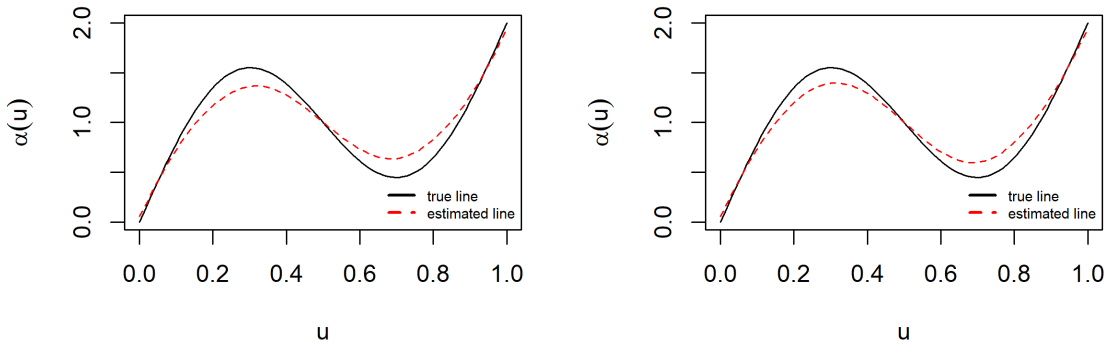
Table 5. Simulation results for Model 1, with Rook matrix, $p = 50$, $p_{zero} = 40$, $\sigma_{\varepsilon 0} = 0.5$, and $n = 400$.

ρ_0	Correct(%)	C	IC	ρ	$\sigma_{\varepsilon 0}$	RMSE
−0.8	93.6	39.906	0	−0.003(0.035)	0.010(0.021)	0.127
−0.5	93.0	39.894	0	−0.015(0.062)	0.004(0.018)	0.118
−0.2	93.4	39.916	0	−0.011(0.080)	0.003(0.017)	0.116
0	93.2	39.912	0	0.000(0.084)	0.003(0.017)	0.115
0.2	93.6	39.914	0	0.011(0.080)	0.003(0.017)	0.115
0.5	93.0	39.898	0	0.014(0.062)	0.004(0.018)	0.117
0.8	93.6	39.914	0	0.002(0.035)	0.010(0.020)	0.128

4.2.3. Nonparametric estimation accuracy

For the nonparametric function in Model 1, Figure 1 indicates that under the same sample size, the precision of the estimated curves remains similar across different values of ρ_0. However, for a

fixed ρ_0, the estimated curves converge closer to the true curve as the sample size increases. Figure 2 demonstrates that even as the variable dimension increases to $p = 20$ and $p = 50$, the estimation accuracy of $\alpha(u)$ remains robust. Moreover, under the same sample size, the fitted curves for these two dimensions are visually indistinguishable.



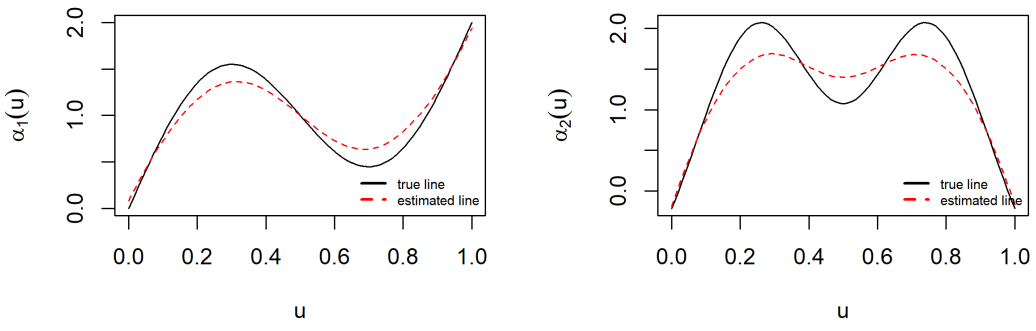
(a) $\rho_0 = 0.8$, $n = 225$, $p = 20$, $p_{zero} = 10$ (b) $\rho_0 = 0.8$, $n = 400$, $p = 50$, $p_{zero} = 40$

Figure 2. Fitted curves of $\alpha(U)$ for the settings in Tables 3 and 5.

Table 6 presents the variable selection results of Model 2. It can be observed that, compared to the results of Model 1 in Table 3, the three metrics—Correct(%), C, and IC—are relatively similar, demonstrating that the proposed method can select the correct variables with high accuracy. However, the estimation accuracy for parameters in Table 6 has decreased slightly compared to that in Table 3, as reflected by the increased standard deviation. Figure 3 displays the fitting results of the two nonparametric functions in Model 2. It can be observed that both $\alpha_1(u)$ and $\alpha_2(u)$ closely approximate the true curves.

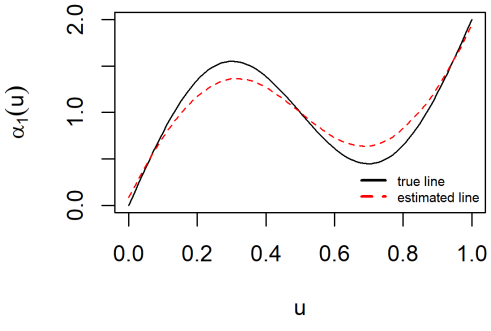
Table 6. Simulation results for Model 2, with Rook matrix, $p = 20$, $p_{zero} = 10$, $\sigma_{\varepsilon 0} = 0.5$, and $n = 225$.

ρ_0	Correct(%)	C	IC	ρ	$\sigma_{\varepsilon 0}$	RMSE$_1$	RMSE$_2$
−0.8	91.8	9.890	0	0.086(0.064)	0.086(0.033)	0.163	0.282
−0.5	92.8	9.918	0	0.090(0.094)	0.062(0.026)	0.150	0.270
−0.2	93.4	9.924	0	0.038(0.099)	0.052(0.025)	0.148	0.267
0	92.6	9.916	0	−0.004(0.107)	0.050(0.024)	0.148	0.267
0.2	91.2	9.900	0	−0.045(0.096)	0.052(0.024)	0.150	0.266
0.5	92.2	9.910	0	−0.096(0.095)	0.062(0.026)	0.153	0.267
0.8	91.4	9.900	0	−0.089(0.065)	0.086(0.031)	0.168	0.276

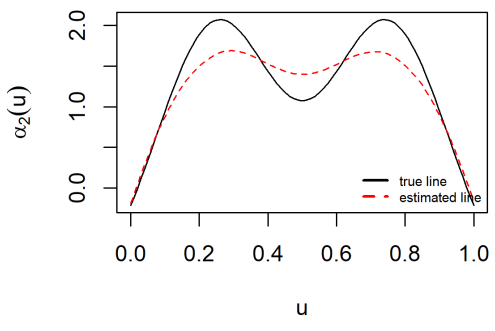


(a) $\rho_0 = 0.2$, $n = 225$, $p = 20$, $p_{zero} = 10$

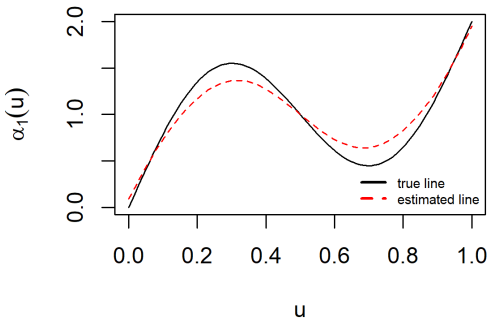
(b) $\rho_0 = 0.2$, $n = 225$, $p = 20$, $p_{zero} = 10$



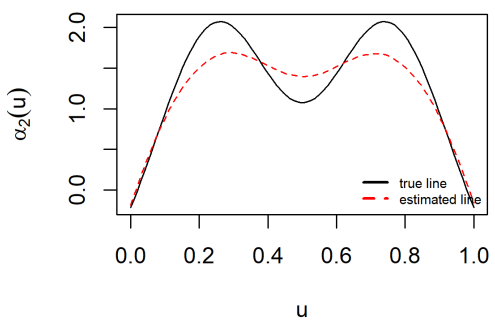
(c) $\rho_0 = 0.5$, $n = 225$, $p = 20$, $p_{zero} = 10$



(d) $\rho_0 = 0.5$, $n = 225$, $p = 20$, $p_{zero} = 10$



(e) $\rho_0 = 0.8$, $n = 225$, $p = 20$, $p_{zero} = 10$



(f) $\rho_0 = 0.8$, $n = 225$, $p = 20$, $p_{zero} = 10$

Figure 3. Fitted curves of $\alpha_1(U)$ and $\alpha_2(U)$ for the settings in Table 6.

4.2.4. Comparisons of different methods

Table 7 compares the variable selection performance of adaptive LASSO, SCAD, and LASSO under varying sample sizes n and levels of spatial dependence ρ_0. The results consistently demonstrate the superior efficacy of adaptive LASSO, which achieves the highest correct selection rate (approaching 96% for $n = 225$) and accurately identifies the true model across all settings. SCAD exhibits moderate and relatively stable performance, generally outperforming standard LASSO, particularly at smaller sample sizes. However, due to its nonconvex penalty function, SCAD can be sensitive to initial values and may converge to local optima, which explains why its advantage diminishes in some larger-sample scenarios (e.g., $n = 225$, $\rho_0 = 0.2, 0.5$), whereas LASSO achieves comparable selection accuracy. In contrast, LASSO demonstrates considerable sensitivity to both sample size and spatial correlation. Although its performance improves with larger n, it is substantially compromised under high spatial dependence ($\rho_0 = 0.8$), rendering it the least robust method in this study. Overall, the convex nature of adaptive LASSO, combined with data-driven adaptive weights, ensures reliable convergence and stable variable selection across all scenarios, a particularly valuable property for spatial error models requiring iterative estimation.

Table 7. Comparisons of different methods for Model 1 with Rook matrix, $p = 6$, $p_{zero} = 4$, $\sigma_{\varepsilon 0} = 0.5$, and different n and ρ_0 values.

		Adaptive LASSO			SCAD			LASSO		
n	ρ	Correct%	C	IC	Correct%	C	IC	Correct%	C	IC
64	0.2	89.0	3.858	0	79.8	3.616	0	57.4	3.508	0
	0.5	89.2	3.864	0	79.6	3.580	0	52.0	3.412	0
	0.8	89.4	3.878	0	73.8	3.460	0	38.6	3.126	0
81	0.2	89.8	3.872	0	79.4	3.578	0	65.2	3.600	0
	0.5	90.4	3.888	0	78.6	3.552	0	59.0	3.520	0
	0.8	90.4	3.876	0	74.2	3.452	0	38.2	3.142	0
100	0.2	91.6	3.900	0	82.0	3.626	0	76.0	3.724	0
	0.5	90.6	3.886	0	79.8	3.562	0	68.0	3.606	0
	0.8	90.8	3.890	0	77.2	3.510	0	44.8	3.270	0
225	0.2	95.6	3.952	0	82.0	3.636	0	93.8	3.936	0
	0.5	96.0	3.956	0	80.8	3.644	0	90.8	3.904	0
	0.8	95.4	3.946	0	77.6	3.584	0	75.8	3.732	0

4.2.5. Computational cost and scalability

To assess the algorithm's scalability, we examine its performance under larger sample sizes and higher dimensions in Table 3 ($n = 225$, $p = 20$) and Table 5 ($n = 400$, $p = 50$). The results show that the correct selection rate remains above 90% even at $p = 50$, and the estimation errors for both parametric and nonparametric components decrease with increasing n, indicating that the method scales well to moderately high dimensions. Regarding computational cost, on a standard desktop (Intel Core i5-10500 @ 3.10GHz, 8GB RAM), a single replication with $n = 225$ and $p = 20$ takes approximately 2 seconds, and $n = 400$, $p = 50$ takes about 7 seconds. The algorithm typically converges within 5–10 iterations under the tolerance $tol = 10^{-6}$, demonstrating efficient computation.

5. Real data analysis

In this part, we collect empirical data from 51 Belt and Road Initiative (BRI) partner countries in 2015 to validate the feasibility of the proposed variable selection method for SVCSEM in practice.

5.1. Background

In September and October 2013, China proposed the major initiatives of building the "Silk Road Economic Belt" and the "21st Century Maritime Silk Road", which garnered significant attention from the international community. The year 2015 marked the completion of the planning phase and the official launch of the Belt and Road Initiative (BRI), with many countries actively joining the cooperative framework. During this period, China's outward foreign direct investment (OFDI) in BRI countries intensified substantially, accompanied by growing scholarly interest in understanding the determinants and underlying mechanisms of these cross-border capital flows. However, Chinese investments have also faced considerable risks shaped by host countries' institutional environments, making institutional quality (IQ) a critical factor in investment decisions.

Existing literature (e.g., [22]) confirms that institutional quality exerts a nonlinear moderating effect on the relationship between host country GDP and China's OFDI in BRI countries. Although these findings provide valuable insights, the presence of such nonlinear dynamics, coupled with the challenge of identifying the most influential determinants among a diverse set of potential predictors, calls for more flexible and robust modeling approaches. Traditional econometric methods often struggle to simultaneously accommodate complex nonlinear structures and perform effective variable selection, particularly in settings with limited sample sizes.

To address these methodological challenges, we apply the semi-parametric varying coefficient spatial error model (SVCSEM) with coefficient shrinkage to the 2015 data on China's OFDI in 51 BRI countries. This approach is particularly well-suited for our research context, as it simultaneously addresses the dual challenges of modeling complex nonlinear spatiotemporal dynamics and mitigating overfitting in high-dimensional predictors, while performing variable selection and model estimation in one step. By doing so, we aim to both validate the nonlinear moderating effect of institutional quality and identify the core determinants that drive China's OFDI patterns across BRI countries during this pivotal early stage of the initiative.

5.2. Data description

This real data set includes 11 indicators collected from 51 BRI partner countries, sourced from the *Statistical Bulletin of China's OFDI*, *World Development Indicators*, *Worldwide Governance Indicators*, and the *Heritage Foundation*. To ensure comparability and mitigate heteroscedasticity, logarithmic transformations are applied to OFDI, GDP, ATV, DSB, PD, TFI, CDE, and CEH. The institutional quality (IQ) index is synthesized from six governance dimensions (voice and accountability, political stability, government effectiveness, regulatory quality, rule of law, and control of corruption) using min–max normalization followed by averaging. Detailed descriptions and notations of these indicators are provided in Table 8, and their summary statistics (minimum, quartiles, mean, maximum) are presented in Table 9.

Table 8. Variables description.

Variable	Notation	Description	Data Source
OFDI	Y	China's OFDI stock (in 10,000 USD) in the BRI partner countries	Statistical Bulletin of China's OFDI
GDP	Z	Host country market size, proxied by GDP	World Development Indicator Database
IQ	U	Institutional quality of host countries	World Governance Indicator Database
IM	X_1	Imports of goods and services as a share of GDP in the host countries	World Development Indicator Database
EX	X_2	Exports of goods and services as a share of GDP in the host countries	World Development Indicator Database
ATV	X_3	Air traffic volume (number of domestic and overseas departures of the registered carrier in the host countries)	World Development Indicator Database
DSB	X_4	Number of days required to start a business in the host country	World Development Indicator Database
TFI	X_5	Trade freedom index in the host countries	The Heritage Foundation
PD	X_6	Population density (people/km^2 of land area)	
CDE	X_7	Carbon dioxide emissions (metric tons per capita)	World Development Indicator Database
CEH	X_8	China's exports to the host countries	International Monetary Fund

Table 9. Descriptive statistics of variables ($n = 51$).

Variable	Min	Q_1	Median	Mean	Q_3	Max
$\ln Y$	15.450	20.930	21.850	21.760	23.640	25.880
$\ln Z$	22.530	24.140	25.520	25.330	26.330	28.460
U	0.270	0.360	0.460	0.483	0.590	0.870
X_1	0.100	0.270	0.370	0.455	0.575	1.780
X_2	0.150	0.285	0.450	0.477	0.605	1.510
$\ln X_3$	4.250	9.985	10.800	10.735	12.080	13.580
$\ln X_4$	−0.690	2.090	2.530	2.581	3.130	5.230
$\ln X_5$	3.720	4.310	4.360	4.345	4.465	4.500
$\ln X_6$	0.660	4.145	4.530	4.634	5.305	8.960
$\ln X_7$	−2.040	0.550	1.390	1.248	2.100	3.560
$\ln X_8$	19.630	22.820	23.740	23.990	25.260	27.030

5.3. Spatial model analysis

We first fit a linear regression, and compute Moran's I of residuals. The outcome shows that $Moran's\ I = 0.095$, $p-value = 0.093$, which manifests a pattern of spatial correlation in the residuals. Then, we use Lagrange multiplier tests to choose between the SAR model and SEM. According to Table 10, the test results for both LM_{lag} and LM_{error} are not significant. However, the robust LM_{lag} and LM_{error} results both pass the test, with LM_{error} showing higher significance. This indicates that we can choose the SEM. We subsequently specify the proposed SVCSEM to accommodate both nonlinear covariate effects and spatial dependence in the error structure:

$$\begin{cases} \ln Y_i = X_{i1} \times \beta_1 + X_{i2} \times \beta_2 + \sum_{p=3}^{8} \ln X_{ip} \times \beta_p + \ln Z_i \times \alpha(U_i) + v_i, \\ v_i = \rho \sum_{j=1}^{51} w_{ij} v_j + \varepsilon_i, \end{cases}$$

where w_{ij} is the element in the ith row and jth column of the spatial weight matrix W. Here, we calculate the geographical distances between 51 countries based on their latitude and longitude data, setting a geographical distance threshold of $d = 1500$ km. If the distance between country i and country j is less than d, then $w_{ij} = 1$; otherwise, $w_{ij} = 0$. This yields a 0-1 spatial weight matrix W based on geographical distance. We are interested in determining which variables among $X_1, X_2, \cdots, X_8$ influence China's OFDI. Additionally, we aim to explore how the host country's institutional quality moderates the effect of GDP on OFDI.

Table 10. Lagrange multiplier tests.

Type	Statistics	P-value
LM_{lag}	0.032	0.857
LM_{error}	0.232	0.630
RLM_{lag}	15.171	9.818×10^{-5}
RLM_{error}	15.371	8.834×10^{-5}

For comparison, we report the estimation results of the linear model (LM), spatial error model (SEM), semi-parametric varying coefficient spatial error model (SVCSEM), and the adaptive LASSO semi-parametric varying coefficient spatial error model (ALasso-SVCSEM), as shown in Table 11.

As indicated by the positive spatial error coefficients (ρ) in the last three columns of Table 11, significant positive spatial autocorrelation is present in the error terms. When incorporating the varying coefficient structure, we fitted the curve for $\alpha(U)$ using both the SVCSEM and ALasso-SVCSEM methods. The resulting curves are closely aligned, so only the ALasso-SVCSEM curve is presented here.

Table 11. Model estimation results ($n = 51$).

Variables	LM	SEM	SVCSEM	ALasso-SVCSEM
X_1	3.618	1.859	3.287	0.338
X_2	−1.882	1.587	−2.061	0
X_3	−0.262	−0.257	−0.331	−0.075
X_4	−0.042	0.058	0.227	0
X_5	−3.102	−2.199	−0.754	0
X_6	−0.640	−0.782	−0.658	−0.489
X_7	−0.418	−0.446	−0.376	0
X_8	0.893	0.144	0.860	0.512
ρ	/	0.630	0.278	0.409
σ^2	1.642	1.734	1.738	2.054

As is shown in Figure 4, the marginal effect of GDP on OFDI exhibits a U-shaped relationship with institutional quality (IQ). The U-shaped pattern indicates that GDP's marginal effect on OFDI is stronger at both low and high institutional quality levels, but weaker at intermediate levels. A plausible explanation: In low-quality institutions, Chinese investments may rely on relational contracting rather than formal institutions, making market size a dominant pull factor. In high-quality institutions, transparent and efficient markets amplify the effect of market size. At intermediate levels, institutional constraints may exist without full transactional efficiency, dampening investment responsiveness to market size. This interpretation aligns with the "institutional escapism" and "institutional advantage" perspectives in international business literature [23, 24].

Applying the proposed variable selection method (see last column of Table 11), the regression coefficients for X_2, X_4, X_5, and X_7 are shrunk to zero. The remaining significant coefficients indicate that imports (as a share of GDP) and China's exports to host countries exert promoting effects on China's OFDI. In contrast, air traffic volume and population density play a suppressing role.

As a robustness check, we assess the predictive performance of the competing models. Given the small sample size ($n = 51$) and model complexity, repeated 10-fold cross-validation provides a more stable estimate of prediction error than leave-one-out cross-validation (LOOCV) [25, 26]. We therefore conduct a 10-fold cross-validation repeated 10 times to evaluate the predictive performance of the competing models. For each model (the linear model (LM), spatial error model (SEM), semi-parametric varying coefficient spatial error model (SVCSEM), and the proposed ALasso-SVCSEM), we randomly partition the 51 countries into 10 folds, iteratively using 9 folds for training and the remaining fold for testing, and compute the mean absolute error (MAE) for the test set. The process is repeated 10 times, and the average MAE across all repetitions, together with its standard deviation, is reported.

Table 12 summarizes the results. The ALasso-SVCSEM achieves the smallest average MAE (1.760, s.d. = 0.039), followed by SVCSEM (2.406, s.d. = 0.054), SEM (2.580, s.d. = 0.101), and LM (2.839, s.d.= 0.128). These findings confirm that the proposed method not only performs well in variable selection and in-sample estimation but also exhibits superior predictive ability. This improvement underscores the benefit of jointly modeling spatial dependence and nonlinearity while performing variable selection.

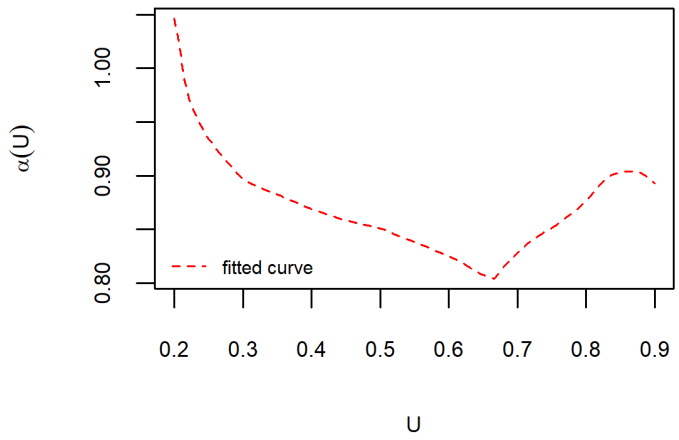


Figure 4. The fitted curve of $\alpha(U)$ for ALasso-SVCSEM.

Table 12. Cross-validated mean absolute error (MAE) for competing models.

Model	Average MAE	(Std. dev.)
Linear model (LM)	2.839	(0.128)
Spatial error model (SEM)	2.580	(0.101)
SVCSEM (Nonpenalized)	2.406	(0.054)
ALasso-SVCSEM (Proposed)	**1.760**	(0.039)

6. Conclusions

This paper addresses the problem of variable selection in semi-parametric varying coefficient spatial error models (SVCSEM), a flexible framework that accommodates nonlinear relationships and spatial dependence. The main contributions of this study are threefold:

(1) **Methodology.** We develop a penalized estimation procedure that integrates adaptive LASSO with profile likelihood, enabling simultaneous model estimation and variable selection for SVCSEM. The algorithm iteratively estimates the varying coefficients, spatial parameter, and regression coefficients, with the tuning parameter selected via a BIC-type criterion to ensure model parsimony.

(2) **Theory.** Under mild regularity conditions, we establish the asymptotic normality and selection consistency of the proposed adaptive LASSO estimator, demonstrating its oracle property in the spatial semiparametric context.

(3) **Empirical evidence.** Extensive simulations across various spatial structures (Rook and Case matrices), sample sizes (n = 49 to 400), and covariate dimensions (p = 6 to 50) confirm that the method achieves high selection accuracy (Correct% > 90% for large n) and low estimation errors. An application to China's OFDI in Belt and Road Initiative countries reveals a U-

shaped moderating effect of institutional quality on the GDP–OFDI relationship and identifies key determinants of investment flows.

Overall, this study enriches the spatial econometric literature by combining variable selection with semi-parametric modeling, providing empirical researchers with a robust and theoretically justified tool for analyzing spatially structured data in fields such as regional economics and environmental studies.

Despite its contributions, this study has several limitations that suggest directions for future research. First, our theoretical results rely on the assumption of normally distributed errors and a fixed number of covariates; extending the framework to accommodate non-Gaussian errors or diverging dimensions would broaden its applicability. Second, the current algorithm assumes a known spatial weight matrix, whereas data-driven estimation of W remains an open problem. Third, although our simulations cover a range of settings, real-world spatial structures may be more complex, warranting further robustness checks. Future work could explore variable selection in dynamic spatial panels or incorporate spatiotemporal structures, as in [13] and [14], while retaining the variable selection focus. In addition, developing post-selection inference methods for SVCSEM would provide more reliable statistical guarantees for empirical researchers. Furthermore, comparing the proposed method with emerging spatial variable selection techniques, such as spatially weighted LASSO [15] and geographically weighted group LASSO [16], would offer deeper insights into the relative merits of different approaches for handling spatial dependence and heterogeneity.

Author contributions

Yu Liu: Methodology, Formal analysis, Data curation, Software, Visualization, Writing–original draft; Zengchao Xu: Software, Validation, Writing–review & editing.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare no potential conflict of interests.

References

1. Y. Zhang, D. Shen, Estimation of semi-parametric varying-coefficient spatial panel data models with random-effects, *J. Stat. Plan. Infer.*, **159** (2015), 64–80. https://doi.org/10.1016/j.jspi.2014.11.001
2. E. Malikov, Y. Sun, Semiparametric estimation and testing of smooth coefficient spatial autoregressive models, *J. Econometrics*, **199** (2017), 12–34. https://doi.org/10.1016/j.jeconom.2017.02.005
3. X. Liang, J. Gao, X. Gong, Semiparametric spatial autoregressive panel data model with fixed effects and time-varying coefficients, *Journal of Business & Economic Statistics*, **40** (2022), 1784–1802. https://doi.org/10.1080/07350015.2021.1979564

4. R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Roy. Stat. Soc. B*, **58** (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
5. J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Stat. Assoc.*, **96** (2001), 1348–1360. https://doi.org/10.1198/016214501753382273
6. H. Zou, The adaptive lasso and its oracle properties, *J. Amer. Stat. Assoc.*, **101** (2006), 1418–1429. https://doi.org/10.1198/016214506000000735
7. Y. Wu, Y. Sun, Shrinkage estimation of the linear model with spatial interaction, *Metrika*, **80** (2017), 51–68. https://doi.org/10.1007/s00184-016-0590-z
8. X. Liu, J. Chen, S. Cheng, A penalized quasi-maximum likelihood method for variable selection in the spatial autoregressive model, *Spat. Stat.*, **25** (2018), 86–104. https://doi.org/10.1016/j.spasta.2018.05.001
9. T. Xie, R. Cao, J. Du, Variable selection for spatial autoregressive models with a diverging number of parameters, *Stat. Papers*, **61** (2020), 1125–1145. https://doi.org/10.1007/s00362-018-0984-2
10. G. Luo, M. Wu, Variable selection for semiparametric varying-coefficient spatial autoregressive models with a diverging number of parameters, *Commun. Stat.-Theor. Meth.*, **50** (2021), 2062–2079. https://doi.org/10.1080/03610926.2019.1659367
11. Y. Liu, X. Zhuang, Shrinkage estimation of semi-parametric spatial autoregressive panel data model with fixed effects, *Stat. Probabil. Lett.*, **194** (2023), 109746. https://doi.org/10.1016/j.spl.2022.109746
12. Y. Liu, Adaptive lasso variable selection method for semiparametric spatial autoregressive panel data model with random effects, *Commun. Stat.-Theor. Method.*, **53** (2024), 2122–2140. https://doi.org/10.1080/03610926.2022.2119088
13. N. Serban, A space–time varying coefficient model: the equity of service accessibility, *Ann. Appl. Stat.*, **5** (2011), 2024–2051. https://doi.org/10.1214/11-AOAS473
14. G. Goh, J. Yu, M. Kim, J. Tack, Bayesian space–time varying coefficient modeling for climate econometrics: a spatial–temporal Gaussian process approach, *J. Econometrics*, in press. https://doi.org/10.1016/j.jeconom.2025.106075
15. S. Nandy, C. Y. Lim, T. Maiti, Additive model building for spatial regression, *J. Roy. Stat. Soc. B*, **79** (2017), 779–800. https://doi.org/10.1111/rssb.12195
16. T. Zheng, R. Li, M. Wu, C. Ma, Accommodating spatial heterogeneity in geographically weighted regression with group penalty, *Stat. Med.*, **44** (2025), e70226. https://doi.org/10.1002/sim.70226
17. T. J. Hefley, M. B. Hooten, E. M. Hanks, R. E. Russell, D. P. Walsh, The Bayesian group Lasso for confounded spatial data, *JABES*, **22** (2017), 42–59. https://doi.org/10.1007/s13253-016-0274-1
18. J. Fan, T. Huang, Profile likelihood inferences on semiparametric varying-coefficient partially linear models, *Bernoulli*, **11** (2005), 1031–1057. https://doi.org/10.3150/bj/1137421639
19. X. Lin, R. J. Carroll, Nonparametric function estimation for clustered data when the predictor is measured without/with error, *J. Amer. Stat. Assoc.*, **95** (2000), 520–534. https://doi.org/10.2307/2669396
20. L.-F. Lee, Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models, *Econometrica*, **72** (2004), 1899–1925. https://doi.org/10.1111/j.1468-0262.2004.00558.x

21. A. C. Case, Spatial patterns in household demand, *Econometrica*, **59** (1991), 953–965. https://doi.org/10.2307/2938168
22. Y. Liu, L. Tang, M. Jin, Host countries' institutional environment and china's OFDI: a semi-parametric spatial panel approach, (Chinese), *Statistical Research*, **40** (2023), 85–99. https://doi.org/10.19343/j.cnki.11-1302/c.2023.03.007
23. M. A. Witt, A. Y. Lewin, Outward foreign direct investment as escape response to home country institutional constraints, *J. Int. Bus. Stud.*, **38** (2007), 579–594. https://doi.org/10.1057/palgrave.jibs.8400285
24. C. Jones, Y. Temouri, K. Kirollos, J. Du, Tax havens and emerging market multinationals: the role of property rights protection and economic freedom, *J. Bus. Res.*, **155** (2023), 113373. https://doi.org/10.1016/j.jbusres.2022.113373
25. S. Borra, A. Di Ciaccio, Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods, *Comput. Stat. Data Anal.*, **54** (2010), 2976–2989. https://doi.org/10.1016/j.csda.2010.03.004
26. J.-H. Kim, Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap, *Comput. Stat. Data Anal.*, **53** (2009), 3735–3745. https://doi.org/10.1016/j.csda.2009.04.009
27. J. Chen, N. Qiao, Estimation of semi-parametric varying coefficient spatial error model, (Chinese), *The Journal of Quantitative & Technical Economics*, **34** (2017), 129–146.

Appendix

The following fundamental lemma is useful in proof of Theorem 1.

Lemma 1. *Under Assumptions 1–3, if there exists* $\hat{\boldsymbol{\theta}} = \arg\max \log \hat{L}(\boldsymbol{\theta})$*, then* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

Lemma 1 follows from standard results on the consistency of profile likelihood estimators in spatial semi-parametric models; see, for example, Theorem 1 of [27] and the quasi-maximum likelihood (QML) theory in [20]. Its proof is omitted here for brevity.

Furthermore, under Assumptions 1–4, $\hat{\rho}$ is $\sqrt{n}$-consistent and asymptotically normal. This result is essential for the expansion of $\boldsymbol{S}(\hat{\rho})^\top \boldsymbol{S}(\hat{\rho})$ in the proof of Theorem 1 and follows from standard QML estimation theory for spatial models [20].

Proof of Theorem 1. Let $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \boldsymbol{\eta}/\sqrt{n}$,

$$\boldsymbol{\Psi}(\boldsymbol{\eta}) = \left[\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}(\boldsymbol{\beta}_0 + \frac{\boldsymbol{\eta}}{\sqrt{n}})\right]^\top \boldsymbol{S}^\top(\hat{\rho})\boldsymbol{S}(\hat{\rho})\left[\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}(\boldsymbol{\beta}_0 + \frac{\boldsymbol{\eta}}{\sqrt{n}})\right] + \sum_{j=1}^{p} \lambda_j \left|\beta_{0j} + \frac{\eta_j}{\sqrt{n}}\right|.$$

Define $\hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} \boldsymbol{\Psi}(\boldsymbol{\eta})$, then $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \hat{\boldsymbol{\eta}}/\sqrt{n}$. By straightforward calculations, we have

$$\begin{aligned}\boldsymbol{\Psi}(\boldsymbol{\eta}) - \boldsymbol{\Psi}(\mathbf{0}) = &\frac{\boldsymbol{\eta}^\top \tilde{\boldsymbol{X}}^\top \boldsymbol{S}^\top(\hat{\rho})\boldsymbol{S}(\hat{\rho})\tilde{\boldsymbol{X}}\boldsymbol{\eta}}{n} - 2\frac{(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_0)^\top \boldsymbol{S}^\top(\hat{\rho})\boldsymbol{S}(\hat{\rho})\tilde{\boldsymbol{X}}\boldsymbol{\eta}}{\sqrt{n}} \\ &+ \sum_{j=1}^{p} \lambda_j \left(\left|\beta_{0j} + \frac{\eta_j}{\sqrt{n}}\right| - \left|\beta_{0j}\right|\right).\end{aligned} \tag{6.1}$$

In Eq (6.1), $\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_0 = (\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{M}_0 + \boldsymbol{S}^{-1}\boldsymbol{\varepsilon})$. Under Assumption 3, $(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{M}_0 = o_p(1)$, so then we have

$$\Big[\frac{(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_0)^\top \boldsymbol{S}^\top(\hat{\rho})\boldsymbol{S}(\hat{\rho})\tilde{\boldsymbol{X}}}{\sqrt{n}}\Big]^\top = \frac{\boldsymbol{\varepsilon}^\top(\boldsymbol{S}^{-1})^\top(\boldsymbol{I} - \boldsymbol{A})^\top \boldsymbol{S}^\top(\hat{\rho})\boldsymbol{S}(\hat{\rho})\tilde{\boldsymbol{X}}}{\sqrt{n}} + o_p(1).$$

From Lemma 1, we know that $\hat{\rho} \xrightarrow{p} \rho_0$. Because $\hat{\rho}$ is $\sqrt{n}$-consistent, we can expand $\boldsymbol{S}(\hat{\rho})^\top \boldsymbol{S}(\hat{\rho})$ around ρ_0:

$$\boldsymbol{S}(\hat{\rho})^\top \boldsymbol{S}(\hat{\rho}) = \boldsymbol{S}^\top \boldsymbol{S} + (\hat{\rho} - \rho_0)\boldsymbol{H} + o_p(1/\sqrt{n}),$$

where $\boldsymbol{H} = \frac{\partial}{\partial \rho}(\boldsymbol{S}(\rho)^\top \boldsymbol{S}(\rho))\Big|_{\rho=\rho_0} = -(\boldsymbol{W}^\top \boldsymbol{S} + \boldsymbol{S}^\top \boldsymbol{W})$. Substituting this expansion into the term involving $\hat{\rho}$ yields

$$\begin{aligned}&\frac{(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_0)^\top \boldsymbol{S}^\top(\hat{\rho})\boldsymbol{S}(\hat{\rho})\tilde{\boldsymbol{X}}}{\sqrt{n}} \\ &= \frac{(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_0)^\top \boldsymbol{S}^\top \boldsymbol{S}\tilde{\boldsymbol{X}}}{\sqrt{n}} + (\hat{\rho} - \rho_0)\frac{(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_0)^\top \boldsymbol{H}\tilde{\boldsymbol{X}}}{\sqrt{n}} + o_p(1).\end{aligned}$$

The first term will be treated below. For the second term, note that $(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_0)^\top \boldsymbol{H}\tilde{\boldsymbol{X}}/\sqrt{n} = O_p(1)$ (by the same argument as for the leading term), and $(\hat{\rho} - \rho_0) = O_p(1/\sqrt{n})$, so their product is $O_p(1/\sqrt{n}) = o_p(1)$. Hence, the second term is negligible, and we may replace $\boldsymbol{S}(\hat{\rho})^\top \boldsymbol{S}(\hat{\rho})$ by $\boldsymbol{S}^\top \boldsymbol{S}$ in the asymptotic distribution.

Now, consider the first term. Using $\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_0 = (\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{M}_0 + \boldsymbol{S}^{-1}\boldsymbol{\varepsilon})$ and the fact that $(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{M}_0 = o_p(1)$ by Assumption 3, we have

$$\begin{aligned}\frac{(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_0)^\top \boldsymbol{S}^\top \boldsymbol{S}\tilde{\boldsymbol{X}}}{\sqrt{n}} &= \frac{\boldsymbol{\varepsilon}^\top(\boldsymbol{S}^{-1})^\top(\boldsymbol{I} - \boldsymbol{A})^\top \boldsymbol{S}^\top \boldsymbol{S}(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{X}}{\sqrt{n}} + o_p(1) \\ &= \frac{1}{\sqrt{n}}\boldsymbol{\varepsilon}^\top(\boldsymbol{S}^{-1})^\top \boldsymbol{\Delta}\boldsymbol{X}.\end{aligned}$$

Let $\boldsymbol{v}_i$ denote the ith row of the matrix $(\boldsymbol{S}^{-1})^\top \boldsymbol{\Delta}\boldsymbol{X}$; then, this term can be written as $(\sum_{i=1}^n \varepsilon_i \boldsymbol{v}_i^\top)/\sqrt{n}$, where $\boldsymbol{v}_i$ are nonrandom conditional on the covariates. Under Assumption 1, ε_i are i.i.d. with mean zero and variance σ_0^2, and the Lindeberg–Feller central limit theorem applies because $\max_i \|\boldsymbol{v}_i\|^2 / \sum_i \|\boldsymbol{v}_i\|^2 \to 0$ and $(\sum_i \boldsymbol{v}_i \boldsymbol{v}_i^\top)/n \to \boldsymbol{\Omega}$ by the definition of $\boldsymbol{\Omega}$. Consequently,

$$\frac{1}{\sqrt{n}}\boldsymbol{\varepsilon}^\top(\boldsymbol{S}^{-1})^\top \boldsymbol{\Delta}\boldsymbol{X} \xrightarrow{d} N(\boldsymbol{0}, \sigma_0^2 \boldsymbol{\Omega}),$$

where $\boldsymbol{\Omega} = \lim_{n\to\infty} \frac{1}{n} E(\boldsymbol{X}^\top \boldsymbol{\Delta}\boldsymbol{\Psi}\boldsymbol{\Delta}\boldsymbol{X})$. Denote this limiting random vector by $\boldsymbol{\xi}$.

Now, we deal with the last term in (6.1). If $\beta_{0j} \neq 0$, then

$$\left|\lambda_j\left(\left|\beta_{0j} + \frac{\eta_j}{\sqrt{n}}\right| - \left|\beta_{0j}\right|\right)\right| = \frac{\left|\eta_j\right|}{\sqrt{n}}\lambda_j \leq a_n n^{-1/2}\left|\eta_j\right| \to 0,$$

under the assumption $a_n n^{-1/2} \to 0$ as n goes to infinity. If $\beta_{0j} = 0$, under the assumption that $b_n n^{-1/2} \to \infty$, then we have

$$\lambda_j\left(\left|\beta_0 j + \frac{\eta_j}{\sqrt{n}}\right| - |\beta_0 j|\right) = \frac{\left|\eta_j\right|}{\sqrt{n}}\lambda_j \geq b_n n^{-1/2}\left|\eta_j\right| \to \begin{cases} \infty, & \eta_j \neq 0 \\ 0, & \eta_j = 0 \end{cases}$$

Therefore, as $n \to \infty$,

$$\sum_{j=1}^{p} \lambda_j \left(\left| \beta_{0j} + \frac{\eta_j}{\sqrt{n}} \right| - |\beta_{0j}| \right) \to \begin{cases} 0, & \eta_j = 0 \quad \text{and} \quad \forall j \notin \mathcal{A} \\ \infty, & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_{\mathcal{A}} \\ \boldsymbol{\eta}_{\mathcal{A}^c} \end{pmatrix}$, and $\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\xi}_{\mathcal{A}} \\ \boldsymbol{\xi}_{\mathcal{A}^c} \end{pmatrix}$. Combined with $\boldsymbol{X}_{\mathcal{A}}^{\top} \boldsymbol{\Delta} \boldsymbol{X}_{\mathcal{A}} / n \xrightarrow{p} \boldsymbol{\Phi}_{11}$, we have

$$\boldsymbol{G}(\boldsymbol{\eta}) = \boldsymbol{\Psi}(\boldsymbol{\eta}) - \boldsymbol{\Psi}(\boldsymbol{0}) \xrightarrow{d} \begin{cases} \boldsymbol{\eta}_{\mathcal{A}}^{\top} \boldsymbol{\Phi}_{11} \boldsymbol{\eta}_{\mathcal{A}} - 2 \boldsymbol{\xi}_{\mathcal{A}}^{\top} \boldsymbol{\eta}_{\mathcal{A}}, & \eta_j = 0 \quad \text{and} \quad \forall j \in \mathcal{A}^c \\ \infty, & \text{otherwise.} \end{cases}$$

We see that at $(\boldsymbol{\xi}_{\mathcal{A}}^{\top} \boldsymbol{\Phi}_{11}^{-1}, \boldsymbol{0}_{1 \times (p-p_0)})^{\top}$, $\boldsymbol{G}(\boldsymbol{\eta})$ reaches the minimum. We then get $\hat{\boldsymbol{\eta}}_{\mathcal{A}} \xrightarrow{d} \boldsymbol{\Phi}_{11}^{-1} \boldsymbol{\xi}_{\mathcal{A}}$ and $\hat{\boldsymbol{\eta}}_{\mathcal{A}^c} \xrightarrow{d} \boldsymbol{0}$. That implies $\sqrt{n}(\tilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) \xrightarrow{d} \boldsymbol{\Phi}_{11}^{-1} \boldsymbol{\xi}_{\mathcal{A}}$, and **Theorem 1(i)** is verified.

Now, we prove the consistency part. For each $j \in \mathcal{A}$, by the result of Theorem 1(i), we know that $\tilde{\beta}_j \xrightarrow{p} \beta_{0j}$ when $n \to \infty$. Therefore, $P(\tilde{\beta}_j \neq 0) = P(j \in \mathcal{A}_n) \to 1$ as $n \to \infty$. That is,

$$P(\mathcal{A} \subseteq \mathcal{A}_n) \to 1, \quad n \to \infty.$$

We need to prove the opposite side, $P(\mathcal{A}_n \subseteq \mathcal{A}) \to 1$. It suffices to show that $\forall j' \notin \mathcal{A}$, $P(j' \in \mathcal{A}_n) \to 0$. Consider the event $j' \in \mathcal{A}_n$; by the Karush–Kuhn–Tucker (KKT) optimality condition, we know that

$$2 \tilde{\boldsymbol{X}}_{j'}^{\top} \boldsymbol{S}^{\top}(\hat{\rho}) \boldsymbol{S}(\hat{\rho}) (\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \tilde{\boldsymbol{\beta}}) = \lambda_{j'} \operatorname{sgn}(\tilde{\beta}_{j'}),$$

where $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_p)$, $\tilde{\boldsymbol{X}}_j$ is the jth column of $\tilde{\boldsymbol{X}}$. Because $\tilde{\boldsymbol{\beta}}$ is consistent, we have $\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \tilde{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{Y} - \boldsymbol{X} \tilde{\boldsymbol{\beta}}) = (\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{S}^{-1} \boldsymbol{\varepsilon} + o_p(1))$, and thus,

$$\frac{\tilde{\boldsymbol{X}}_{j'}^{\top} \boldsymbol{S}^{\top}(\hat{\rho}) \boldsymbol{S}(\hat{\rho}) (\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \tilde{\boldsymbol{\beta}})}{\sqrt{n}} = \frac{\tilde{\boldsymbol{X}}_{j'}^{\top} \boldsymbol{S}^{\top}(\hat{\rho}) \boldsymbol{S}(\hat{\rho}) \boldsymbol{S}^{-1} \boldsymbol{\varepsilon}}{\sqrt{n}} + o_p(1).$$

It is apparent that as $n \to \infty$, $\tilde{\boldsymbol{X}}_{j'}^{\top} \boldsymbol{S}^{\top}(\hat{\rho}) \boldsymbol{S}(\hat{\rho}) \boldsymbol{S}^{-1} \boldsymbol{\varepsilon} / \sqrt{n}$ converges in distribution to a normal random variable (by the same CLT argument used for $\boldsymbol{\xi}$). Therefore, for any $j' \in \mathcal{A}^c$, we have $\lambda_{j'} / \sqrt{n} \to \infty$ and then

$$P(j' \in \mathcal{A}_n) = P\left(2 \frac{\tilde{\boldsymbol{X}}_{j'}^{\top} \boldsymbol{S}^{\top}(\hat{\rho}) \boldsymbol{S}(\hat{\rho}) (\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \tilde{\boldsymbol{\beta}})}{\sqrt{n}} = \frac{\lambda_{j'} \operatorname{sgn}(\tilde{\beta}_{j'})}{\sqrt{n}} \right) \to 0 \quad (n \to \infty),$$

which implies $P(\mathcal{A}^c \subseteq \mathcal{A}_n^c) \to 1$, namely $P(\mathcal{A} \supseteq \mathcal{A}_n) \to 1$. Combined with the previous result, we obtain $\lim_{n \to \infty} P(\mathcal{A}_n = \mathcal{A}) = 1$, which is **Theorem 1(ii)**. □