Ann Inst Stat Math (2018) 70:323–351
https://doi.org/10.1007/s10463-016-0589-2

Variable selection for spatial semivarying coefficient models

Kangning Wang[1,2]

Received: 7 January 2016 / Revised: 25 June 2016 / Published online: 20 December 2016
© The Institute of Statistical Mathematics, Tokyo 2016

Abstract Spatial semiparametric varying coefficient models are a useful extension of spatial linear model. Nevertheless, how to conduct variable selection for it has not been well investigated. In this paper, by basis spline approximation together with a general M-type loss function to treat mean, median, quantile and robust mean regressions in one setting, we propose a novel partially adaptive group $L_r (r \geq 1)$ penalized M-type estimator, which can select variables and estimate coefficients simultaneously. Under mild conditions, the selection consistency and oracle property in estimation are established. The new method has several distinctive features: (1) it achieves robustness against outliers and heavy-tail distributions; (2) it is more flexible to accommodate heterogeneity and allows the set of relevant variables to vary across quantiles; (3) it can keep balance between efficiency and robustness. Simulation studies and real data analysis are included to illustrate our approach.

Keywords Geostatistics · Variable selection · Robustness · Heterogeneity · Penalized M-type estimator · Oracle property

The research was supported by NNSF project (11171188, 11071145, 11221061, 11231005 and 11601283), the Scientific and Technological Research Program of Chongqing Municipal Education Commission (No. KJ1501109) and Research Project of Chongqing University of Arts and Sciences (No. Y2014SC35).

✉ Kangning Wang
wkn1986@126.com

[1] School of Statistics, Shandong Technology and Business University, Yantai 264005, China

[2] Institute for Financial Studies, Shandong University, Jinan 250100, China

1 Introduction

Regression models are widely used in analysis of geostatistical data that arise often in environmental and ecological studies (Hallin et al. 2004, 2009; Tang and Cheng 2009; Tang 2014). In this paper, we focus on variable selection in spatial semiparametric varying coefficient regression models. This problem has not been well developed in both theory and methodology.

Shrinkage methods have emerged as a popular approach for variable selection in traditional regression settings. The examples include but not limited to the LASSO (Tibshirani 1996), the bridge regression (Fu 1998), the SCAD (Fan and Li 2001). Recently, many works have been done to extend such methods to nonparametric and semiparametric settings. Fan and Li (2001) for partially linear models; Wang et al. (2008), Wang and Xia (2009) for varying coefficient regression models; Xue (2009), Huang et al. (2010) and Wang et al. (2014) for additive models; Kai et al. (2011), Hohsuk et al. (2012) and Tang et al. (2013) discussed variable selection issue for quantile varying coefficient models. Wang and Lin (2014) and Zhao et al. (2014) further discussed robust variable selection methods for partially linear varying coefficient models. An excellent discussion of group selection can be found in Huang et al. (2012).

For geostatistical data, variable selection method is mainly limited to linear regressions. Hoeting et al. (2006) used Akaikes information criterion (AIC) for variable selection. Wang and Zhu (2009) proposed a penalized least squares (LS) which can construct simultaneous variable selection and parameter estimation for spatial linear regressions. Zhu et al. (2010) and Chu et al. (2011) proposed two penalized maximum likelihood estimation (PMLE) methods for spatial linear regression models with certain error distribution assumption.

However, variable selection for the spatial nonparametric or semiparametric models has not yet been well investigated to the best of our knowledge. In a linear regression setup, it has been very well understood that ignoring any important predictor can lead to seriously biased results, whereas including spurious covariates can degrade the estimation efficiency substantially. Thus, variable selection is important for any regression problems. Furthermore, due to the complex spatial correlation, outliers and heterogeneity in the spatial data may be possible, we need to develop a genuine variable selection method to automatically adapt to these environments. In this paper, this issue is studied in a unified framework. Using a general M-type loss function as a unified method to treat mean regression, median regression, quantile regression and robust mean regression in one setting and B-spline approximation, we propose a partially adaptive group $L_r (r \geq 1)$ penalized M-type estimation, which can construct variable selection and coefficients estimation simultaneously. Theoretical analysis shows that the new method with $r = 1$ or $r = 2$ enjoys the favorable asymptotic properties: the variable selection procedure is consistent, and estimators enjoy the oracle property. Here, the oracle property means that the estimators of the nonparametric components achieve the optimal convergence rate, and the estimators of the parametric components have the same asymptotic distribution as that obtained under the true model. Theoretical investigation also indicates that the asymptotic properties can be extended to more general L_r penalty with $r \geq 1$.

Our new method offers the following progresses. (I) The existing methods are limited to spatial parameter regression. Our work develops the method to spatial semiparametric regression. (II) Compared with the variable selection methods in spatial parametric regression (Wang and Zhu 2009; Zhu et al. 2010; Chu et al. 2011), the new method has the following distinct features: firstly, it can achieve the robustness against outliers and heavy-tail distributions; secondly, it can accommodate heterogeneity and allows the sets of relevant covariates that may differ when we consider different quantiles, and thus the new method enables us to explore the entire conditional distribution of the spatial response; thirdly, our method contains some other methods such as Wang and Zhu (2009) as special cases in some sense.

It is remarkable that investigating the statistical theory of penalty method for spatial semiparametric models is not trivial. Firstly, for spatial linear model, it involves only one type of penalized parameters (i.e., the tuning parameters); nevertheless, for spatial semiparametric models, three types of regularization parameters: the smoothing parameters, tuning parameters for parametric part and tuning parameters for nonparametric part, are involved. Due to their interaction, what are the right convergence rates for the tuning parameters and smoothing parameters is much less well understood. Furthermore, the main technical challenge of our work is to establish the oracle property in the complex spatial dependence semiparametric setting for the penalized general M-type loss function, which includes both smooth case (e.g., the mean regression) and nonsmooth case (e.g., the quantile regression).

The rest of this paper is organized as follows. Section 2 introduces our new method, investigates its theoretical properties and discusses the implementation issues. Numerical studies and real data analysis are reported in Sect. 3. All the technical proofs are provided in the Appendix.

2 Partially adaptive group penalized M-type estimator

2.1 Spatial semiparametric varying coefficient models

For a wide application, we consider the spatial data in a general context. Let $\mathbb{Z}^N$, $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$, $N \geq 1$, denote the integer lattice points in the N-dimensional Euclidean space. A point $\mathbb{i} = (i_1, \ldots, i_N) \in \mathbb{Z}^N$ is referred to as a site. Spatial data are modeled as finite realizations of vector stochastic processes indexed by $\mathbb{i} \in \mathbb{Z}^N$: random fields. In this paper, we will consider strictly stationary $(p+q+2)$ dimensional real random fields, of the form

$$\left\{(Y_{\mathbb{i}}, \boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, U_{\mathbb{i}}), \mathbb{i} \in \mathbb{Z}^N\right\}, \tag{1}$$

where $Y_{\mathbb{i}} \in \mathbb{R}^1$, $\boldsymbol{X}_{\mathbb{i}} \in \mathbb{R}^p$, $\boldsymbol{Z}_{\mathbb{i}} \in \mathbb{R}^q$, and $U_{\mathbb{i}} \in [U_0, U_1]$ are defined over some probability space $(\Omega, \mathscr{A}, P)$. Let $\mathcal{S}, \mathcal{S}' \subset \mathbb{Z}^N$ be two sets of sites, and the generated Borel σ-fields are defined as $\mathcal{B}(\mathcal{S}) := \mathcal{B}\left((Y_{\mathbb{i}}, \boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, U_{\mathbb{i}}) : \mathbb{i} \in \mathcal{S}\right)$ and $\mathcal{B}(\mathcal{S}') := \mathcal{B}\left((Y_{\mathbb{i}}, \boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, U_{\mathbb{i}}) : \mathbb{i} \in \mathcal{S}'\right)$, respectively. For each couple $\mathcal{S}$ and $\mathcal{S}'$, $d(\mathcal{S}, \mathcal{S}') := \min\left\{\|\mathbb{i} - \mathbb{i}'\|_2 : \mathbb{i} \in \mathcal{S}, \mathbb{i}' \in \mathcal{S}'\right\}$ denotes the distance between $\mathcal{S}$ and $\mathcal{S}'$, where $\|\cdot$

$\|_2$ stands for the Euclidean norm. We will assume that $\left\{(Y_{\mathbb{i}}, \boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, U_{\mathbb{i}}), \mathbb{i} \in \mathbb{Z}^N\right\}$ satisfies the following mixing condition: there exist two functions, $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\varphi(t) \downarrow 0$ as $t \to \infty$, and $\psi : \mathbb{N}^2 \to \mathbb{R}^+$ is a symmetric positive function nondecreasing in each variable, such that whenever $\mathcal{S}, \mathcal{S}' \subset \mathbb{Z}^N$,

$$\begin{aligned} \alpha\left(\mathcal{B}(\mathcal{S}), \mathcal{B}(\mathcal{S}')\right) &:= \sup\left\{|P(AB) - P(A)P(B)| : A \in \mathcal{B}(\mathcal{S}), B \in \mathcal{B}(\mathcal{S}')\right\} \\ &\leq \psi\left(\text{Card}(\mathcal{S}), \text{Card}(\mathcal{S}')\right) \varphi\left(d(\mathcal{S}, \mathcal{S}')\right), \end{aligned} \tag{2}$$

where Card$(\cdot)$ denotes cardinality. $\psi \equiv 1$ corresponds to strongly mixing.

Throughout, we assume that the random field (1) is observed over a rectangular region of the form $\mathcal{I}_{\boldsymbol{n}} := \left\{\mathbb{i} = (i_1, \ldots, i_N) \in \mathbb{Z}^N : 1 \leq i_k \leq n_k, k = 1, \ldots, N\right\}$, for $\boldsymbol{n} = (n_1, \ldots, n_N)$ with strictly positive coordinates $n_1, \ldots, n_N$. The total sample size is $\widehat{\boldsymbol{n}} = \prod_{k=1}^N n_k$. As in Tran (1990), we write $\boldsymbol{n} \to \infty$ if $\min_{1 \leq k \leq N}\{n_k\} \to \infty$ and, moreover, $n_i/n_j < C$, $1 \leq i, j \leq N$ for some $0 < C < \infty$.

Remark 1 In fact, the α-mixing condition (2.2) is quite general, it generalizes the classical time series ($N = 1$) mixing concepts and means that process is asymptotically independent. Recently, much work has been done for the spatial process with such mixing condition, e.g., Hallin et al. (2004) gave a local linear spatial regression, Hallin et al. (2009) studied the spatial nonparametric estimation, Tang (2014) and Lu et al. (2014) investigated the estimation for spatial varying coefficient models, Lu and Tjøstheim (2014) considered nonparametric estimation of probability density functions for irregularly observed spatial data. It was shown in Hallin et al. (2004) that spatial process of the form $X_{\boldsymbol{n}} = \sum_{\mathbb{i} \in \mathbb{Z}^N} a_{\mathbb{i}} Z_{\boldsymbol{n}-\mathbb{i}}$ can satisfy the mixing condition (2.2), if $Z_{\mathbb{i}}$'s are independent random variables, $a_{\mathbb{i}} \to 0$ exponentially fast and the probability density function of $Z_{\mathbb{i}}$ exists (such as normal, cauchy, exponential, and uniform distributions). For more detailed introduction about this mixing condition (2.2), one can see Tran (1990, 1993) and Fan and Yao (2003).

As discussed in (Gao et al. 2006; Lu et al. 2007, 2014), the issue of avoiding the curse of dimensionality is particularly important in spatial nonparametric regression analysis. Thus, for $\{(Y_{\mathbb{i}}, \boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, U_{\mathbb{i}}), \mathbb{i} \in \mathcal{I}_{\boldsymbol{n}}\}$, we consider the following spatial semiparametric varying coefficient regression models

$$Y_{\mathbb{i}} = \boldsymbol{X}_{\mathbb{i}}^T \boldsymbol{\alpha}(U_{\mathbb{i}}) + \boldsymbol{Z}_{\mathbb{i}}^T \boldsymbol{\beta} + \epsilon_{\mathbb{i}}, \tag{3}$$

where $\epsilon_{\mathbb{i}} \in \mathbb{R}^1$ is a random error, $\boldsymbol{\alpha}(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_p(\cdot))^T \in \mathbb{R}^p$ is an unknown smooth function vector, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ is the constant coefficient vector, and their true values are $\boldsymbol{\alpha}_0(\cdot)$ and $\boldsymbol{\beta}_0$, respectively. However, the true error distribution and spatial dependence structure among the spatial response are not assumed accurately.

Model (3) is a useful extension of the spatial linear regression model (Wang and Zhu 2009; Zhu et al. 2010; Chu et al. 2011), but much less has been done about its variable selection issue.

2.2 The main results

To estimate $\alpha_k(u), k = 1, \ldots, p$, we consider B-spline approximations. Denote by $\mathbb{B}^{\hbar}_{K_{\boldsymbol{n}}}(u)$ the set of spline functions of order $\hbar + 1$ with knots $\overline{\mathbb{K}} = \{U_0 = \tau_0 < \tau_1 < \cdots < \tau_{K_n} < \tau_{K_n+1} = U_1\}$. $B(u) \in \mathbb{B}^{\hbar}_{K_{\boldsymbol{n}}}(u)$ if and only if $B(u) \in C^{\hbar-1}[U_0, U_1]$, and its restriction to each $[\tau_k, \tau_{k+1})$ is a polynomial of degree at most $\hbar$. A piecewise constant function, linear spline, quadratic spline and cubic spline corresponds to $\hbar = 0, 1, 2, 3$, respectively. Let

$$B_k(u) = (\tau_k - \tau_{k-\hbar-1})[\tau_{k-\hbar-1}, \ldots, \tau_k](z-u)^{\hbar}_+, k = 1, \ldots, q_{\boldsymbol{n}},$$

where $q_{\boldsymbol{n}} = K_{\boldsymbol{n}} + \hbar + 1$, $[\tau_{k-\hbar-1}, \ldots, \tau_k](z-u)^{\hbar}_+$ denotes the $(\hbar+1)$th-order divided difference of the function $(z-u)^{\hbar}_+$, $\tau_k = U_0$, when $k = -\hbar, \ldots, -1$, and $\tau_k = U_1$, when $k = K_{\boldsymbol{n}} + 2, \ldots, K_{\boldsymbol{n}} + \hbar + 1$. Then, $\boldsymbol{\pi}(u) = (B_1(u), \ldots, B_{q_{\boldsymbol{n}}}(u))^T$ forms a basis for $\mathbb{B}^{\hbar}_{K_{\boldsymbol{n}}}(u)$. For more details about spline function, see Schumaker (1981). Thus, $\alpha_k(\cdot)$ can be approximated as:

$$\alpha_k(u) \approx \sum_{s=1}^{q_{\boldsymbol{n}}} B_s(u)\theta_{k,s} = \boldsymbol{\pi}(u)^T \boldsymbol{\theta}_k, \tag{4}$$

where $\{\boldsymbol{\theta}_k = (\theta_{k,1}, \ldots, \theta_{k,q_{\boldsymbol{n}}})^T \in \mathbb{R}^{q_{\boldsymbol{n}}}\}_{k=1}^p$ are spline coefficient vectors, and model (3) can be approximated as:

$$Y_{\mathbb{i}} \approx \sum_{k=1}^{p}\sum_{s=1}^{q_{\boldsymbol{n}}} X_{\mathbb{i}k} B_s(U_{\mathbb{i}})\theta_{k,s} + \boldsymbol{Z}_{\mathbb{i}}^T\boldsymbol{\beta} + \epsilon_{\mathbb{i}} = \boldsymbol{\Pi}_{\mathbb{i}}^T\boldsymbol{\Theta} + \boldsymbol{Z}_{\mathbb{i}}^T\boldsymbol{\beta} + \epsilon_{\mathbb{i}}, \tag{5}$$

where $\boldsymbol{\Pi}_{\mathbb{i}} = (X_{\mathbb{i}1}\boldsymbol{\pi}_{\mathbb{i}}^T, \ldots, X_{\mathbb{i}p}\boldsymbol{\pi}_{\mathbb{i}}^T)^T \in \mathbb{R}^{pq_{\boldsymbol{n}}}$, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_p^T)^T \in \mathbb{R}^{pq_{\boldsymbol{n}}}$, $\boldsymbol{\pi}_{\mathbb{i}} = \boldsymbol{\pi}(U_{\mathbb{i}})$.

The main interest is selecting the relevant variables in the parametric and nonparametric parts simultaneously for model (3). For the nonparametric part, we say, e.g., X_k is irrelevant for some $k \in \{1, \ldots, p\}$, if and only if $\alpha_k(\cdot) \equiv 0$, or equivalently $\boldsymbol{\theta}_k = \mathbf{0}$. Thus, we treat $\boldsymbol{\theta}_k, k = 1, \ldots, p$ as groups, and by selecting groups with nonzero $\boldsymbol{\theta}_k$, we can identify the relevant variables. Similarly, for the parametric part, Z_k is regarded as irrelevant, if and only if $\beta_k = 0$. Without loss of generality, assume that $\{\beta_{0k} \neq 0\}_{k=1}^c$ and $\{\beta_{0k} = 0\}_{k=c+1}^q$. Let $\boldsymbol{\beta}^* = (\beta_1, \ldots, \beta_c)^T$ with true value $\boldsymbol{\beta}_0^*$. Similarly, assume $\{\alpha_{0l}(u)\}_{l=1}^v$ are nonzero components of $\boldsymbol{\alpha}_0(u)$ and $\{\alpha_{0l}(u) \equiv 0\}_{l=v+1}^p$.

Thus, for a general loss function $\rho(\cdot)$, the partially adaptive group $L_r (r \geq 1)$ norm penalized M-type estimator $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}})$ is obtained by minimizing

$$Q(\boldsymbol{\Theta}, \boldsymbol{\beta}) := \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \rho\left(Y_{\mathbb{i}} - \boldsymbol{\Pi}_{\mathbb{i}}^T\boldsymbol{\Theta} - \boldsymbol{Z}_{\mathbb{i}}^T\boldsymbol{\beta}\right) + \sum_{k=1}^{p} \lambda^*_{\boldsymbol{n}k}\|\boldsymbol{\theta}_k\|_r + \sum_{k=1}^{q} \lambda^{**}_{\boldsymbol{n}k}|\beta_k|, \tag{6}$$

where $\|\boldsymbol{\theta}_k\|_r = \{\sum_{s=1}^{q_n} |\theta_{k,s}|^r\}^{1/r}$, $\{\lambda_{nl}^*\}_{l=1}^p$ and $\{\lambda_{nk}^{**}\}_{k=1}^q$ are tuning parameters that control model complexity of nonparametric and parametric components, respectively. In this paper, we mainly focus on the cases of $r = 1$ and $r = 2$ for simplicity. Then, $\alpha_k(u)$ can be estimated as: $\widehat{\alpha}_k(u) = \boldsymbol{\pi}(u)^T \widehat{\boldsymbol{\theta}}_k, k = 1, \ldots, p$.

Typical choices for $\rho(\cdot)$ are convex. One well known case is the LS loss $\rho(u) = \frac{1}{2}u^2$. But it is well known that the LS method can be adversely influenced by outliers or heavy-tail distributions. To achieve the robustness, we can consider the robust loss functions such as least absolute deviation (LAD) loss $\rho(u) = |u|$, or more generally, Huber's loss with bounded derivative $\rho'(u) = \max\{-k, \min\{u, k\}\}$, where $k > 0$. The LS and LAD can be regarded as two extremes of the Huber loss for $k = 0$ and $k = \infty$, respectively. For more details, see Huber (1981). Furthermore, if the data are heterogeneous, then the set of active variables may be different when modeling different conditional quantiles. Thus, the quantile regression with $\rho(u) = u(\tau - \mathrm{I}_{(u<0)})$, $\tau \in (0, 1)$, provides a more complete picture of the conditional distribution of the spatial response and is more natural and effective for analyzing spatial data with heteroscedasticity. Thus, by choosing appropriate M-type loss function $\rho(u)$, the new method not only keeps balance between efficiency and robustness but also can be more flexible to accommodate heterogeneity.

Remark 2 The penalized M-type estimator considered here does not take into account the spatial dependence structure in spatial data, because the exact correlation structure is usually unknown in advance. For spatial parametric regression, a method for dealing with dependence structure is to assume that the errors follow a Gaussian process with a parametric covariance (Zhu et al. 2010; Chu et al. 2011), then penalized maximum likelihood can be implemented. But these assumptions do not exactly hold, as is usually the case in practice; furthermore, ignoring spatial dependence will not affect the consistency of variable selection (see Theorems 1 and 2 below).

Next, we will discuss the asymptotic properties of the proposed new method. Let $\Psi(u)$ be the derivative of $\rho(u)$. To establish the asymptotic properties, we first introduce a definition and some regularity conditions.

Definition 1 Define $\mathscr{H}_\gamma$ as the collection of all functions on $[U_0, U_1]$ whose dth order derivative satisfies the Hölder condition of order ν with $\gamma \equiv d + \nu$. That is, for any $h \in \mathscr{H}_\gamma$, there exists a constant $c \in (0, \infty)$ such that for each $h \in \mathscr{H}_\gamma$, $|h^{(d)}(s) - h^{(d)}(t)| \le c|s - t|^\nu$, for any $U_0 \le s, t \le U_1$.

C1. $\{\alpha_k(u) \in \mathscr{H}_\gamma\}_{k=1}^v$ for some $\gamma > N + 1/2$.

C2. The random field $\{(Y_{\mathbb{i}}, \boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, U_{\mathbb{i}}), \mathbb{i} \in \mathbb{Z}^N\}$ is strictly stationary. For all distinct $\mathbb{i}$ and $\mathbb{j}$ in $\mathbb{Z}^N$, $U_{\mathbb{i}}$ and $U_{\mathbb{j}}$ admit a joint density $f_{\mathbb{i},\mathbb{j}}$ satisfying $|f_{\mathbb{i},\mathbb{j}}(u_1, u_2) - f(u_1)f(u_2)| \le c_1$ for all $u_1, u_2 \in [U_0, U_1]$, where $c_1 > 0$ is some constant, and f denotes the marginal density of $U_{\mathbb{i}}$.

C3. For all $\mathbb{i} \in \mathbb{Z}^N$, the random design vectors $\boldsymbol{X}_{\mathbb{i}}$ and $\boldsymbol{Z}_{\mathbb{i}}$ are bounded in probability, and the eigenvalues of $E(\boldsymbol{X}_{\mathbb{i}}\boldsymbol{X}_{\mathbb{i}}^T \mid U_{\mathbb{i}} = u)$, $u \in [U_0, U_1]$ are bounded away from 0 and infinity uniformly.

C4. $\rho(u)$ is convex and $E(\Psi(\epsilon_{\mathbb{i}}) \mid \boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, U_{\mathbb{i}}) = 0$, for all $\mathbb{i} \in \mathbb{Z}^N$. Furthermore, for some $\delta > 0$, $\sup_{\mathbb{i} \in \mathbb{Z}^N} E\left(|\Psi(\epsilon_{\mathbb{i}})|^{2+\delta} \mid \boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, U_{\mathbb{i}}\right) < \infty$, and there exist positive numbers $b_{\mathbb{i}}$ with $0 < \inf b_{\mathbb{i}} \le \sup b_{\mathbb{i}} < \infty$ such that $\sup_{\mathbb{i} \in \mathbb{Z}^N} |E(\Psi(\epsilon_{\mathbb{i}} + s) \mid \boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, U_{\mathbb{i}}) - b_{\mathbb{i}}s| = O_p(s^2)$ as $s \to 0$.

C5. There exist constants $0 < c_2, c_3 < \infty$ such that $\sup_{\mathbb{i}\in\mathbb{Z}^N} E\{[\Psi(\epsilon_\mathbb{i}+s)-\Psi(\epsilon_\mathbb{i})]^2 \mid \boldsymbol{X}_\mathbb{i}, \boldsymbol{Z}_\mathbb{i}, U_\mathbb{i}\} \le c_3|s|$, as $s \to 0$, and $|\Psi(v+s)-\Psi(v)| \le c_3$ for any $|s| \le c_2$ and $v \in \mathbb{R}^1$.

C6. The eigenvalues of $\frac{1}{\widehat{\boldsymbol{n}}}\boldsymbol{\Lambda}_{\boldsymbol{n}}^*$ and $\frac{1}{\widehat{\boldsymbol{n}}}\boldsymbol{\Xi}_{\boldsymbol{n}}^*$ are bounded away from infinity and zero for sufficiently large $\widehat{\boldsymbol{n}}$. Where $\boldsymbol{\Lambda}_{\boldsymbol{n}}^*$ and $\boldsymbol{\Xi}_{\boldsymbol{n}}^*$ are given in the following.

C7. Tuning parameters satisfy: $\max\{a_{\boldsymbol{n}}^*, a_{\boldsymbol{n}}^{**}\}K_{\boldsymbol{n}}^{1/2}\widehat{\boldsymbol{n}}^{-1/2} \to 0$, $b_{\boldsymbol{n}}^*\widehat{\boldsymbol{n}}^{-1/2} \to \infty$ and $b_{\boldsymbol{n}}^{**}(\widehat{\boldsymbol{n}}K_{\boldsymbol{n}})^{-1/2} \to \infty$. Where $a_{\boldsymbol{n}}^* = \sup\{\lambda_{\boldsymbol{n}k}^*, k = 1, \ldots, v\}$, $b_{\boldsymbol{n}}^* = \inf\{\lambda_{\boldsymbol{n}k}^*, k = v+1, \ldots, p\}$ and $a_{\boldsymbol{n}}^{**} = \sup\{\lambda_{\boldsymbol{n}k}^{**}, k = 1, \ldots, c\}$, $b_{\boldsymbol{n}}^{**} = \inf\{\lambda_{\boldsymbol{n}k}^{**}, k = c+1, \ldots, q\}$.

Condition C1 is a smoothness condition on function coefficients determining the rate of convergence of the spline estimator. C2 is a standard condition in this context; it has been used, for example, by Tran (1990), Hallin et al. (2004) and Tang (2014) in the spatial setting. C3 imposed in Huang et al. (2002) is a technical condition to derive the optimal convergence rate of the estimators in functional coefficient setting. Conditions C4 and C5 are only on score function $\Psi(u)$ and the random error $\epsilon_\mathbb{i}$; same conditions were also used in Tang (2014) and Lu et al. (2014). C6 is used to represent the asymptotic covariance matrix of the nonzero parametric part. C7 is the convergence rate of tuning parameters and their data-adaptive version will be given in Sect. 2.4. Obviously, these conditions do not contradict the condition (2), this is because (2) only characterizes the dependence structure of the spatial process, while these regularity conditions do not involve the dependence structure.

Theorem 1 (Estimation Sparsity) *Suppose the regularity conditions C1-C7 hold,* $K_n = O(n^{1/(2\gamma+1)})$ *and* $\varphi(t) = O(\exp(-\kappa t))$ *in* (2) *for some* $\kappa > 0$. *Then, for* $r = 1$ *and* $r = 2$, $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{\boldsymbol{\alpha}}(u)$ *satisfy*

(I) $\widehat{\alpha}_l(u) \equiv 0, l = v+1, \ldots, p$ *holds with probability tending to 1;*
(II) $\widehat{\beta}_l = 0, l = c+1, \ldots, q$ *holds with probability tending to 1.*

By Theorem 1 we know that, as long as (6) is used to obtain $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}(u)$, sparse solutions can be consistently produced. To establish the oracle property in estimation, we first present the following definition and some notations.

Definition 2 The function $g(x^*, u)$ is said to belong to the varying coefficient class of functions $\mathfrak{F}$ if (I) $g(x^*, u) = x^{*T}h(u) = \sum_{k=1}^v x_k h_k(u)$; (II) $\sum_{k=1}^v E[x_k h_k(u)]^2 < \infty$, where x_k and $h_k(u) \in \mathscr{H}_\gamma$ are the kth coordinates of x^* and $h(u)$, respectively, $k = 1, \ldots, v$.

Let $\boldsymbol{X}_\mathbb{i}^* = (X_{\mathbb{i}1}, \ldots, X_{\mathbb{i}v})^T$ and $\boldsymbol{Z}_\mathbb{i}^* = (Z_{\mathbb{i}1}, \ldots, Z_{\mathbb{i}c})^T$. To obtain the asymptotic distribution of $\widehat{\boldsymbol{\beta}}^*$, we first need to adjust for the dependence of $\boldsymbol{Z}_\mathbb{i}^*$ and $(\boldsymbol{X}_\mathbb{i}^*, U_\mathbb{i})$, which is common in semiparametric models. Denote $g_k(\boldsymbol{X}_\mathbb{i}^*, U_\mathbb{i}) = \sum_{l=1}^v X_{\mathbb{i}l} g_{kl}(U_\mathbb{i})$,

$$g_k^*(\boldsymbol{X}_\mathbb{i}^*, U_\mathbb{i}) = \arg\inf_{g_k(\boldsymbol{X}_\mathbb{i}^*, U_\mathbb{i})\in\mathfrak{F}} E\{b_\mathbb{i}[Z_{\mathbb{i}k} - g_k(\boldsymbol{X}_\mathbb{i}^*, U_\mathbb{i})]^2\},$$

and $\varsigma_k(\boldsymbol{X}_\mathbb{i}^*, U_\mathbb{i}) = E(Z_{\mathbb{i}k}|\boldsymbol{X}_\mathbb{i}^*, U_\mathbb{i})$. Note that,

$$\begin{aligned} &E\{b_\mathbb{i}[Z_{\mathbb{i}k} - g_k(\boldsymbol{X}_\mathbb{i}^*, U_\mathbb{i})]^2\} \\ &\quad = E\{b_\mathbb{i}[Z_{\mathbb{i}k} - \varsigma_k(\boldsymbol{X}_\mathbb{i}^*, U_\mathbb{i})]^2\} + E\{b_\mathbb{i}[\varsigma_k(\boldsymbol{X}_\mathbb{i}^*, U_\mathbb{i}) - g_k(\boldsymbol{X}_\mathbb{i}^*, U_\mathbb{i})]^2\}, \end{aligned}$$

therefore, $g_k^*(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}})$ are the projections of $\varsigma_k(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}})$ onto the varying coefficient functional space $\mathfrak{F}$ (under the L_2 norm). In other words, $g_k^*(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}})$ is an element that belongs to $\mathfrak{F}$ and it is the closest function to $\varsigma_k(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}})$ among all the functions in $\mathfrak{F}$, for any $k = 1, \ldots, c$. We define $g^*(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}}) = (g_1^*(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}}), \ldots, g_c^*(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}}))^T$, $\varsigma(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}}) = (\varsigma_1^*(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}}), \ldots, \varsigma_c^*(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}}))^T$ and

$$\boldsymbol{\Xi}_{\boldsymbol{n}}^* = \sum_{\mathbb{i}} [\boldsymbol{Z}_{\mathbb{i}}^* - g^*(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}})]^T b_{\mathbb{i}} [\boldsymbol{Z}_{\mathbb{i}}^* - g^*(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}})],$$

$$\boldsymbol{\Lambda}_{\boldsymbol{n}}^* = \sum_{\mathbb{i}} [\boldsymbol{Z}_{\mathbb{i}}^* - g^*(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}})]^T E[\Psi^2(\epsilon_{\mathbb{i}})][\boldsymbol{Z}_{\mathbb{i}}^* - g^*(\boldsymbol{X}_{\mathbb{i}}^*, U_{\mathbb{i}})].$$

Then, the following theorem gives the oracle property of the estimator.

Theorem 2 (Oracle Property) *Under the same conditions used in Theorem 1, we have*

(I) $\frac{1}{\widehat{\boldsymbol{n}}} \sum_{\mathbb{i} \in \mathcal{I}_{\boldsymbol{n}}} \{\widehat{\alpha}_l(U_{\mathbb{i}}) - \alpha_{0l}(U_{\mathbb{i}})\}^2 = O_p\left(\widehat{\boldsymbol{n}}^{-2\gamma/(2\gamma+1)}\right), l = 1, \ldots, v;$

(II) $\boldsymbol{\Lambda}_{\boldsymbol{n}}^{*-1/2} \boldsymbol{\Xi}_{\boldsymbol{n}}^* \left(\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0^*\right) \to_d N(\boldsymbol{0}, \boldsymbol{I}_c).$

Theorem 2 means that the penalized M-estimators have the oracle property, which implies that the estimators of the nonparametric components achieve the optimal convergence rate, and the estimators of the parametric components have the same asymptotic distribution as that obtained under the correct submodel.

As commented by the reviewer, our method only focuses on the estimation and variable selection, while the inference results based on the estimation are not involved. Actually, it is an interesting topic, and we will investigate it in the future.

Remark 3 Let $g_{\mathbb{i}}(\cdot)$ be the density of $\epsilon_{\mathbb{i}}$ conditional on $(\boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, U_{\mathbb{i}})$. In this remark, we will show that, if $g_{\mathbb{i}}(\cdot)$ satisfies some conditions, then the general M-type loss function including mean regression, median regression, quantiles regression and robust mean regression can satisfy C4 and C5. Thus, the conclusions in Theorems 1 and 2 hold for them uniformly.

(I) For $\rho(u) = u(\tau - \mathrm{I}_{(u<0)})$, if $g_{\mathbb{i}}(\cdot)$ is bounded away from zero and infinite, satisfies $\int_{-\infty}^{0} g_{\mathbb{i}}(s)ds = \tau$, and has a bounded first derivative in the neighborhood of zero, then $E\Psi(\epsilon_{\mathbb{i}}) = 0$ and C5 is satisfied, and C4 holds with $b_{\mathbb{i}} = g_{\mathbb{i}}(0)$.

(II) For $\Psi(u) = \max\{-k, \min\{u, k\}\}$, $k > 0$, if $g_{\mathbb{i}}(\cdot)$ is bounded and symmetric about 0, then C5 and $E\Psi(\epsilon_{\mathbb{i}}) = 0$ hold automatically, C4 holds with $b_{\mathbb{i}} = \int_{-k}^{k} g_{\mathbb{i}}(s)ds$.

(III) Loss functions $\rho(u) = |u|$ and $\rho(u) = \frac{1}{2}u^2$ satisfy the two conditions automatically.

2.3 Issues in practical implementation

In this section, we discuss the algorithm for the new method. If $\rho(\cdot)$ is two-order differentiable everywhere, by local quadratic approximation (LQA, Fan and Li (2001); Hunter and Li 2005), a general algorithm for (6) can be realized by a modified Newton–Raphson algorithm. More specifically, when $r = 2$, let $(\boldsymbol{\Theta}^{(0)}, \boldsymbol{\beta}^{(0)})$ be the initial value chosen as the nonpenalized estimator, i.e.,

$$(\boldsymbol{\Theta}^{(0)}, \boldsymbol{\beta}^{(0)}) = \arg\min_{\boldsymbol{\Theta},\boldsymbol{\beta}} \Big\{ \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \rho(Y_{\mathbb{i}} - \boldsymbol{\Pi}_{\mathbb{i}}^T \boldsymbol{\Theta} - \boldsymbol{Z}_{\mathbb{i}}^T \boldsymbol{\beta}) \Big\}.$$

Then, the $(t+1)$-step estimate $(\boldsymbol{\Theta}^{(t+1)}, \boldsymbol{\beta}^{(t+1)})$ can be obtained by minimizing

$$\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \rho\left(Y_{\mathbb{i}} - \boldsymbol{\Pi}_{\mathbb{i}}^T \boldsymbol{\Theta} - \boldsymbol{Z}_{\mathbb{i}}^T \boldsymbol{\beta}\right) + \sum_{k=1}^{p} \lambda_{\boldsymbol{n}k}^{*} \frac{\|\boldsymbol{\theta}_k\|_2^2}{\|\boldsymbol{\theta}_k^{(t)}\|_2} + \sum_{k=1}^{q} \lambda_{\boldsymbol{n}k}^{**} \frac{\beta_k^2}{|\beta_k^{(t)}|}. \tag{7}$$

For $r = 1$, a similar procedure can be implemented by replacing the second term in (7) with $\sum_{k=1}^{p} \lambda_{\boldsymbol{n}k} \sum_{s=1}^{q_{\boldsymbol{n}}} \theta_{k,s}^2 / |\theta_{k,s}^{(t)}|$. Then, $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}})$ is the limit value of $(\boldsymbol{\Theta}^{(t+1)}, \boldsymbol{\beta}^{(t+1)})$.

While, an alternative algorithm that does not make use of the derivative of $\Psi(\cdot)$ is the following re-weighted least squares algorithm. Let $\boldsymbol{M}_{\mathbb{i}} = (\boldsymbol{\Pi}_{\mathbb{i}}^T, \boldsymbol{Z}_{\mathbb{i}}^T)^T$, $r_{\mathbb{i}} = Y_{\mathbb{i}} - \boldsymbol{\Pi}_{\mathbb{i}}^T \boldsymbol{\Theta}^{(t)} - \boldsymbol{Z}_{\mathbb{i}}^T \boldsymbol{\beta}^{(t)}$ and $W_{\mathbb{i}} = \Psi(r_{\mathbb{i}})/r_{\mathbb{i}}$. Then, the next iteration gives

$$(\boldsymbol{\Theta}^{(t+1)^T}, \boldsymbol{\beta}^{(t+1)^T})^T = \left(\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} W_{\mathbb{i}} \boldsymbol{M}_{\mathbb{i}} \boldsymbol{M}_{\mathbb{i}}^T + \boldsymbol{\Lambda}^{(t)} \right)^{-1} \left(\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} W_{\mathbb{i}} \boldsymbol{M}_{\mathbb{i}} Y_{\mathbb{i}} \right), \tag{8}$$

where

$$\boldsymbol{\Lambda}^{(t)} = \operatorname{diag}\left(\underbrace{\frac{\lambda_{\boldsymbol{n}1}^{*}}{\|\boldsymbol{\theta}_1^{(t)}\|_2}, \cdots, \frac{\lambda_{\boldsymbol{n}1}^{*}}{\|\boldsymbol{\theta}_1^{(t)}\|_2}}_{q_{\boldsymbol{n}}}, \cdots, \underbrace{\frac{\lambda_{\boldsymbol{n}p}^{*}}{\|\boldsymbol{\theta}_p^{(t)}\|_2}, \cdots, \frac{\lambda_{\boldsymbol{n}p}^{*}}{\|\boldsymbol{\theta}_p^{(t)}\|_2}}_{q_{\boldsymbol{n}}}, \frac{\lambda_{\boldsymbol{n}1}^{**}}{|\beta_1^{(t)}|}, \cdots, \frac{\lambda_{\boldsymbol{n}q}^{**}}{|\beta_q^{(t)}|} \right),$$

for the L_2 penalty, while for the L_1 penalty

$$\boldsymbol{\Lambda}^{(t)} = \operatorname{diag}\left(\frac{\lambda_{\boldsymbol{n}1}^{*}}{|\theta_{1,1}^{(t)}|}, \cdots, \frac{\lambda_{\boldsymbol{n}1}^{*}}{|\theta_{1,q_{\boldsymbol{n}}}^{(t)}|}, \cdots, \frac{\lambda_{\boldsymbol{n}p}^{*}}{|\theta_{p,1}^{(t)}|}, \cdots, \frac{\lambda_{\boldsymbol{n}p}^{*}}{|\theta_{p,q_{\boldsymbol{n}}}^{(t)}|}, \frac{\lambda_{\boldsymbol{n}1}^{**}}{|\beta_1^{(t)}|}, \cdots, \frac{\lambda_{\boldsymbol{n}q}^{**}}{|\beta_q^{(t)}|} \right).$$

The re-weighted least squares algorithm usually converges quickly. But when some of the residuals are close to 0, these points receive too large weight. We adopt the modification (Fan and Li 2001) to replace the weight $W_{\mathbb{i}}$ by $\Psi(r_{\mathbb{i}})/(r_{\mathbb{i}} + a_{\boldsymbol{n}})$, where $a_{\boldsymbol{n}}$ is the $2\widehat{\boldsymbol{n}}^{-1/2}$ quantile of the absolute residuals $\{|r_{\mathbb{i}}|, \mathbb{i} \in \mathcal{I}_{\boldsymbol{n}}\}$.

Furthermore, if $\rho(u) = \rho_\tau(u) = u(\tau - \mathrm{I}_{(u<0)})$, $\tau \in (0,1)$, a convenient linear programming algorithm can be directly applied to solve problem (6). More specifically, for $r = 1$, (6) can be reformulated as the following:

$$\begin{aligned} &\arg\min \Big\{ \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} [\tau \eta_{\mathbb{i}}^{+} + (1-\tau)\eta_{\mathbb{i}}^{-}] + \sum_{k=1}^{p} \lambda_{\boldsymbol{n}k}^{*} \sum_{s=1}^{q_{\boldsymbol{n}}} (\theta_{k,s}^{+} + \theta_{k,s}^{-}) + \sum_{k=1}^{q} \lambda_{\boldsymbol{n}k}^{**} (\beta_k^{+} + \beta_k^{-}) \Big\}, \\ &\text{s.t. } \eta_{\mathbb{i}}^{+} - \eta_{\mathbb{i}}^{-} + \boldsymbol{\Pi}_{\mathbb{i}}^T (\boldsymbol{\Theta}^{+} - \boldsymbol{\Theta}^{-}) + \boldsymbol{Z}_{\mathbb{i}}^T (\boldsymbol{\beta}^{+} - \boldsymbol{\beta}^{-}) = Y_{\mathbb{i}}, \mathbb{i} \in \mathcal{I}_{\boldsymbol{n}}, \\ &\eta_{\mathbb{i}}^{+} \geq 0, \eta_{\mathbb{i}}^{-} \geq 0, \theta_{k,s}^{+} \geq 0, \theta_{k,s}^{-} \geq 0, \beta_k^{+} \geq 0, \beta_k^{-} \geq 0. \end{aligned} \tag{9}$$

where $\boldsymbol{\beta}^+ = (\beta_1^+, \ldots, \beta_q^+)^T \in \mathbb{R}_+^q$, $\boldsymbol{\beta}^- = (\beta_1^-, \ldots, \beta_q^-)^T \in \mathbb{R}_+^q$ and $\boldsymbol{\Theta}^+ = (\theta_{k,s}^+) \in \mathbb{R}_+^{pq_n}$, $\boldsymbol{\Theta}^- = (\theta_{k,s}^-) \in \mathbb{R}_+^{pq_n}$. Thus, interior-point method can be used to solve it. For $r = 2$, by local linear approximation (LLA) Zou and Li (2008), the solution can be obtained by an iterative linear programming algorithm. More specifically, choose an initial value $\boldsymbol{\Theta}^{(0)}$, $\|\boldsymbol{\theta}_k\|_2$ can be approximated as:

$$\|\boldsymbol{\theta}_k\|_2 \approx \|\boldsymbol{\theta}_k^{(0)}\|_2 + \|\boldsymbol{\theta}_k^{(0)}\|_2^{-1} |\boldsymbol{\theta}_k^{(0)}|^T (|\boldsymbol{\theta}_k| - |\boldsymbol{\theta}_k^{(0)}|), \tag{10}$$

where $|\boldsymbol{\theta}_k| = (|\theta_{k,1}|, \ldots, |\theta_{k,q_n}|)^T$. Then, update the estimate repeatedly until convergence, in which the $(t+1)$-step solution is given by minimizing

$$\sum_{\mathbb{i} \in \mathcal{I}_n} \rho_\tau \left(Y_\mathbb{i} - \boldsymbol{\Pi}_\mathbb{i}^T \boldsymbol{\Theta} - \boldsymbol{Z}_\mathbb{i}^T \boldsymbol{\beta} \right) + \sum_{k=1}^{p} \lambda_{\boldsymbol{n}k}^* \sum_{s=1}^{q_n} \frac{|\theta_{k,s}^{(t)}|}{\|\boldsymbol{\theta}_k^{(t)}\|_2} |\theta_{k,s}| + \sum_{k=1}^{q} \lambda_{\boldsymbol{n}k}^{**} |\beta_k|. \tag{11}$$

Here, (11) can be computed by (9) directly.

To implement the above procedures, tuning parameters $\{\lambda_{\boldsymbol{n}l}^*\}_{l=1}^p$, $\{\lambda_{\boldsymbol{n}k}^{**}\}_{k=1}^q$ and $K_{\boldsymbol{n}}$ need to be determined. In practice, we take $\lambda_{\boldsymbol{n}j}^* = \lambda_{\boldsymbol{n}}^* / \|\boldsymbol{\theta}_j^{(0)}\|_r$, $1 \le j \le p$ and $\lambda_{\boldsymbol{n}j}^{**} = \lambda_{\boldsymbol{n}}^{**} / |\beta_j^{(0)}|$, $1 \le j \le q$. By this selection, we can verify that as long as $\widehat{\boldsymbol{n}}^{-1/2} K_{\boldsymbol{n}}^{1/2} \max\{\lambda_{\boldsymbol{n}}^*, \lambda_{\boldsymbol{n}}^{**}\} \to 0$ and $K_{\boldsymbol{n}}^{-3/2} \min\{\lambda_{\boldsymbol{n}}^*, \lambda_{\boldsymbol{n}}^{**}\} \to \infty$, C7 can be satisfied. Thus, we only need to select tuning parameters $(\lambda_{\boldsymbol{n}}^*, \lambda_{\boldsymbol{n}}^{**})$ and $K_{\boldsymbol{n}}$.

For $(\boldsymbol{\Theta}^{(0)}, \boldsymbol{\beta}^{(0)})$, we also need to choose the $K_{\boldsymbol{n}}$. Similar to Wang et al. (2009), we choose $K_{\boldsymbol{n}}$ as the minimizer of the following Schwarz-type information criterion,

$$\mathrm{SIC}(K_{\boldsymbol{n}}) = \log \left\{ \sum_{\mathbb{i} \in \mathcal{I}_n} \rho_\tau \left(Y_\mathbb{i} - \boldsymbol{\Pi}_\mathbb{i}^T \boldsymbol{\Theta}^{(0)} - \boldsymbol{Z}_\mathbb{i}^T \boldsymbol{\beta}^{(0)} \right) \right\} + \frac{\log \widehat{\boldsymbol{n}}}{2\widehat{\boldsymbol{n}}} \{ p(K_{\boldsymbol{n}} + \hbar + 1) + q \}. \tag{12}$$

Moreover, in this work we restrict our attention to the spline with $\hbar = 3$. For the proposed estimator $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}})$, we also need to choose $K_{\boldsymbol{n}}$ again, and for the simplicity of implementation, we use the same $K_{\boldsymbol{n}}$ as used in the procedure of constructing $(\boldsymbol{\Theta}^{(0)}, \boldsymbol{\beta}^{(0)})$. Then, the optimal $(\lambda_{\boldsymbol{n}}^*, \lambda_{\boldsymbol{n}}^{**})$ can be chosen as the minimizer of the following BIC type criterion

$$\begin{aligned} \mathrm{BIC}(\lambda_{\boldsymbol{n}}^*, \lambda_{\boldsymbol{n}}^{**}) = \log & \left\{ \frac{1}{\widehat{\boldsymbol{n}}} \sum_{\mathbb{i} \in \mathcal{I}_n} \rho_\tau \left(Y_\mathbb{i} - \boldsymbol{\Pi}_\mathbb{i}^T \widehat{\boldsymbol{\Theta}} - \boldsymbol{Z}_\mathbb{i}^T \widehat{\boldsymbol{\beta}} \right) \right\} \\ & + \mathrm{DF}_v \frac{\log(\widehat{\boldsymbol{n}}/K_{\boldsymbol{n}})}{(\widehat{\boldsymbol{n}}/K_{\boldsymbol{n}})} + \mathrm{DF}_c \frac{\log \widehat{\boldsymbol{n}}}{\widehat{\boldsymbol{n}}}, \end{aligned} \tag{13}$$

where $0 \le \mathrm{DF}_v \le p$ and $0 \le \mathrm{DF}_c \le q$ are simply the numbers of nonzero nonparametric and parametric components, respectively.

3 Simulation studies and real data analysis

3.1 Simulation studies

In this section, we use two simulation experiments to investigate the finite sample performance of the new method. Experiment 1 is homoscedastic, and Experiment 2 is heterogeneous data. We consider the spatial semiparametric varying coefficient regression models in a two-dimensional space ($N = 2$). For the sake of simplicity, we write (i, j) instead of (i_1, i_2), $X_{(i,j)k}$ and $Z_{(i,j)k}$ instead of $X_{\mathbb{i}k}$ and $Z_{\mathbb{i}k}$, respectively, for the sites $i \in \mathbb{Z}^2$.

Experiment 1: (homoscedastic model) Denote by $\{\epsilon_{(i,j)}^{(\alpha)} : (i, j) \in \mathbb{Z}^2\}$ a i.i.d $\alpha N(0, 1) + (1 - \alpha)N(0, 15)$ white-noise processes, where $\alpha \in [0, 1]$. Let

$$Y_{(i,j)} = \beta_1(U_{(i,j)}) + \sum_{k=2}^{12} X_{(i,j)k}\alpha_k(U_{(i,j)}) + \sum_{k=1}^{10} Z_{(i,j)k}\beta_k + \epsilon_{(i,j)}. \tag{14}$$

The relevant variables $\{X_{(i,j)k}\}_{k=2}^{4}$ and $\{Z_{(i,j)k}\}_{k=1}^{3}$ are generated by

$$\begin{aligned}
&X_{(i,j)2} = \frac{U_{(i-1,j)} + U_{(i+1,j)}}{2}, X_{(i,j)3} = \frac{U_{(i,j-1)} + U_{(i,j+1)}}{2},\\
&X_{(i,j)4} = U_{(i-1,j)} + U_{(i+1,j)} + U_{(i,j-1)} + U_{(i,j+1)},\\
&Z_{(i,j)1} = \frac{U_{(i-1,j)} - U_{(i+1,j)}}{2} + e_{(i,j)}^{(0)}, Z_{(i,j)2} = \frac{U_{(i,j-1)} - U_{(i,j+1)}}{2} + e_{(i,j)}^{(0)}\\
&Z_{(i,j)3} = U_{(i-1,j)} + U_{(i+1,j)} + U_{(i,j-1)} + U_{(i,j+1)} + e_{(i,j)}^{(0)},
\end{aligned}$$

and the redundant variables $\{X_{(i,j)k}\}_{k=5}^{12}$ and $\{Z_{(i,j)k}\}_{k=4}^{10}$, index variable $\{U_{(i,j)} : (i, j) \in \mathbb{Z}^2\}$ and random error $\{\epsilon_{(i,j)} : (i, j) \in \mathbb{Z}^2\}$ are generated by the following spatial autoregression:

$$\begin{aligned}
&X_{(i,j)k} = \sinh(X_{(i-1,j)k} + X_{(i+1,j)k} + X_{(i,j-1)k} + X_{(i,j+1)k}) + e_{(i,j)}^{(k)}, k = 5, \ldots, 12,\\
&Z_{(i,j)k} = \cos(Z_{(i-1,j)k} + Z_{(i+1,j)k} + Z_{(i,j-1)k} + Z_{(i,j+1)k}) + e_{(i,j)}^{(0)}, k = 4, \ldots, 10,\\
&U_{(i,j)} = \sin(U_{(i-1,j)} + U_{(i+1,j)} + U_{(i,j-1)} + U_{(i,j+1)}) + e_{(i,j)}^{(0)},\\
&\epsilon_{(i,j)} = \frac{\epsilon_{(i-1,j)} + \epsilon_{(i+1,j)} + \epsilon_{(i,j-1)} + \epsilon_{(i,j+1)}}{6} + \epsilon_{(i,j)}^{(\alpha)},
\end{aligned}$$

where $\{e_{(i,j)}^{(0)} : (i, j) \in \mathbb{Z}^2\}$ and $\{e_{(i,j)}^{(k)} : (i, j) \in \mathbb{Z}^2\}, k = 5, \ldots, 12$, are mutually independent i.i.d $N(0, 1)$ white-noise processes, $\sinh(u) = (e^u - e^{-u})/(e^u + e^{-u})$. The coefficients $\alpha_k(u), k = 1, \ldots, 12$ and $\beta_k, k = 1, \ldots, 10$ are given by $\alpha_1(u) = 0.8\exp(u/2) + 0.5\exp(-u/2)$, $\alpha_2(u) = 0.6u(1 + 0.3u^2)$, $\alpha_3(u) = 6 - 6u$, $\alpha_4(u) = 5 + 10\sin(\pi u/3)$, $\alpha_k(u) \equiv 0, k = 5, \ldots, 12$, $\beta_1 = 1, \beta_2 = -1, \beta_3 = 2$ and $\beta_4 = \cdots = \beta_{10} = 0$. This is an extension of the model used in Hallin et al. (2004). In terms of data sets, we consider three cases: (1) $\alpha = 1$, (2) $\alpha = 0.9$ and (3) an outlier case, in which 5% of the values of response is shifted with shift value $c = 3$. In each

case, by the iterative steps of Hallin et al. (2004), spatial data are generated from (14) over a rectangular domain of $m \times n$ sites, more specially, over a grid of the form $\{(i, j) : 81 \le i \le 81+m, 81 \le j \le 81+n\}$, for various values of m and n. Through 500 repetitions, 500 data sets are obtained.

For comparison, three types of loss functions are chosen as: $\rho_1(u) = \frac{1}{2}u^2$ (LS); $\rho_2(u) = |u|$ (LAD); Huber's loss: $\rho_3(u) = \frac{1}{2}u^2$ if $|u| \le 1.345$ and $\rho_3(u) = 1.345|u| - 0.5(1.345)^2$ otherwise. For the three loss functions with the L_r norm penalty, the corresponding estimators are abbreviated as LS(r), LAD(r) and Huber(r), respectively.

To measure the estimation accuracy of $\widehat{\boldsymbol{\alpha}}(u)$, we use the average squared error (ASE) and the average absolute error (ADE) from the true curves, which are defined, respectively, as:

$$\text{ASE} = \frac{1}{T}\sum_{j=1}^{T}\sum_{k=1}^{p}\left[\widehat{\alpha}_k(u_j) - \alpha_{0k}(u_j)\right]^2, \text{ADE} = \frac{1}{T}\sum_{j=1}^{T}\sum_{k=1}^{p}\left|\widehat{\alpha}_k(u_j) - \alpha_{0k}(u_j)\right|,$$

where $u_j = u_{([2.5\%mn])} + j(u_{([97.5\%mn])} - u_{([2.5\%mn])})/T$, $j = 1, \ldots, T$, are grid points ($T = 200$ in our simulation) and $u_{(1)} \le u_{(1)} \le \cdots \le u_{(mn)}$ is a permutation of $\{U_{(i,j)} : 80 \le i \le 80+m, 80 \le j \le 80+n\}$. The means of the ASE and ADE are denoted by μ(ASE) and μ(ADE), respectively. While the estimation accuracy of $\widehat{\boldsymbol{\beta}}$ is measured by the mean squared error (MSE). To indicate the variable selection results, the percentage of correctly fitting CF% is used, which represents the percentage of selecting true relevant variables. Furthermore, we also compare the average numbers of coefficients that are correctly shrunk to zero, which is denoted by C.

Tables 1 and 2 report the simulation results for this experiment. We can see that the results for L_1 and L_2 norm penalties are similar. Furthermore, for $\alpha = 1$, we find that LS(r), Huber(r) and LAD(r) perform similarly in terms of C and CF% values, implying that they perform similarly for variable selection; all they can select the true relevant variables in the nonparametric and parametric components simultaneously with high probability. While by the values of μ(ASE), μ(ADE) and MSE, we find that LS(r) performs better than LAD(r) and Huber(r). This is not a surprise, since in this case the LAD(r) and Huber(r) will lose efficiency in some degree. Furthermore, when $\alpha = 0.9$ or outliers exist in the dataset, LAD(r) and Huber(r) have better performance than LS(r) does, that they can increase the estimation accuracy noticeably and increase CF% and have better performance in terms of C. All of these results affirm that this new method works well in the spatial semiparametric setting.

Experiment 2: (heterogeneous model) In this experiment, $\{Z_{(i,j)k}\}_{k=1}^{10}$ and $\{U_{(i,j)} : (i, j) \in \mathbb{Z}^2\}$ are generated in the same as in Experiment 1, while $\{(X_{(i,j)2}, \ldots, X_{(i,j)12}) : (i, j) \in \mathbb{Z}^2\}$ are generated as follows. First, $\{(\widetilde{X}_{(i,j)2}, \ldots, \widetilde{X}_{(i,j)12}) : (i, j) \in \mathbb{Z}^2\}$ are generated through the following spatial autoregression:

$$\widetilde{X}_{(i,j)2} = \cos(\widetilde{X}_{(i-1,j)2} + \widetilde{X}_{(i+1,j)2} + \widetilde{X}_{(i,j-1)2} + \widetilde{X}_{(i,j+1)2}) + e^{(2)}_{(i,j)},$$
$$\widetilde{X}_{(i,j)3} = \sin(\widetilde{X}_{(i-1,j)3} + \widetilde{X}_{(i+1,j)3} + \widetilde{X}_{(i,j-1)3} + \widetilde{X}_{(i,j+1)3}) + e^{(3)}_{(i,j)},$$

Table 1 Simulation results for the nonparametric components in Experiment 1

	$m \times n = 15 \times 15$				$m \times n = 20 \times 20$			
	C	CF%	μ(ASE)	μ(ADE)	C	CF%	μ(ASE)	μ(ADE)
$\alpha = 1$								
LAD(1)	7.852	87.20	2.247×10^{-3}	0.1001	8.000	100.00	4.217×10^{-4}	0.0457
LAD(2)	7.848	86.80	2.301×10^{-3}	0.1012	7.995	99.60	4.169×10^{-4}	0.0460
LS(1)	7.860	88.00	1.902×10^{-3}	0.0823	8.000	100.00	3.811×10^{-4}	0.0336
LS(2)	7.842	86.00	1.673×10^{-3}	0.0787	8.000	100.00	3.512×10^{-4}	0.0309
Huber(1)	7.840	85.40	2.013×10^{-3}	0.2264	8.000	100.00	3.718×10^{-4}	0.0511
Huber(2)	7.844	86.00	2.190×10^{-3}	0.2371	8.000	100.00	4.069×10^{-4}	0.0439
$\alpha = 0.9$								
LAD(1)	7.802	80.60	4.076×10^{-3}	0.1409	7.980	98.60	8.083×10^{-4}	0.0560
LAD(2)	7.810	81.00	3.759×10^{-3}	0.1657	7.994	99.40	8.131×10^{-4}	0.0574
LS(1)	4.550	50.20	2.936×10^{-2}	0.2542	7.632	70.20	5.682×10^{-3}	0.0793
LS(2)	4.546	48.80	2.789×10^{-2}	0.2613	7.630	69.40	5.591×10^{-3}	0.0801
Huber(1)	7.788	80.00	2.089×10^{-2}	0.2178	7.980	98.80	3.713×10^{-3}	0.0701
Huber(2)	7.786	79.80	1.897×10^{-2}	0.2069	7.985	99.00	4.016×10^{-3}	0.0813
Outlier case								
LAD(1)	7.678	71.80	3.412×10^{-2}	0.2963	7.968	95.50	3.039×10^{-3}	0.0756
LAD(2)	7.664	70.00	3.357×10^{-2}	0.2872	7.964	94.80	2.018×10^{-3}	0.0760
LS(1)	4.486	44.80	8.963×10^{-2}	0.3847	5.668	57.60	4.108×10^{-2}	0.0931
LS(2)	4.502	46.00	9.030×10^{-2}	0.3727	5.672	60.40	3.013×10^{-2}	0.0907
Huber(1)	7.665	70.60	2.777×10^{-2}	0.2354	7.970	96.00	7.215×10^{-3}	0.0713
Huber(2)	7.602	69.80	2.679×10^{-2}	0.2401	7.962	94.60	8.135×10^{-3}	0.0697

$$\widetilde{X}_{(i,j)4} = \sinh(\widetilde{X}_{(i-1,j)4} + \widetilde{X}_{(i+1,j)4} + \widetilde{X}_{(i,j-1)4} + \widetilde{X}_{(i,j+1)4}) + e^{(4)}_{(i,j)},$$
$$\widetilde{X}_{(i,j)k} = \frac{1}{4}(\widetilde{X}_{(i-1,j)k} + \widetilde{X}_{(i+1,j)k} + \widetilde{X}_{(i,j-1)k} + \widetilde{X}_{(i,j+1)k}) + e^{(k)}_{(i,j)},\ k = 5, \ldots, 12,$$

where $\{e^{(k)}_{(i,j)} : (i, j) \in \mathbb{Z}^2\}, k = 2, \ldots, 12$, are i.i.d $N(0, 1)$ white-noise processes. Then, set $X_{(i,j)2} = \Phi(\widetilde{X}_{(i,j)2})$ and $X_{(i,j)k} = \widetilde{X}_{(i,j)k}$ for $k = 3, \ldots, 12$. The spatial response $\{Y_{(i,j)} : (i, j) \in \mathbb{Z}^2\}$ is generated according to the following heteroscedastic model:

$$Y_{(i,j)} = \alpha_1(U_{(i,j)}) + X_{(i,j)3}\alpha_3(U_{(i,j)}) + X_{(i,j)4}\alpha_4(U_{(i,j)}) + Z_{(i,j)1} - 2Z_{(i,j)2} + \epsilon_{(i,j)}, \tag{15}$$

where $\epsilon_{(i,j)} = \alpha_2(U_{(i,j)})X_{(i,j)2}\widetilde{\epsilon}_{(i,j)}$, $\{\widetilde{\epsilon}_{(i,j)} : (i, j) \in \mathbb{Z}^2\}$ is an $N(0, 1)$ white-noise process, and $\alpha_1(u) = 0.6u(1 + 0.3u^2)$, $\alpha_2(u) = 0.5\sin^2(u + 0.4)$, $\alpha_3(u) = 6 - 0.2u$, $\alpha_4(u) = 2.5\sin(0.95u + 0.5)$. In this experiment, $X_{(i,j)2}$ plays an essential role in the conditional distribution of $Y_{(i,j)}$ given the covariates, but it does not directly

Table 2 Simulation results for the parametric components in Experiment 1

	$m \times n = 15 \times 15$			$m \times n = 20 \times 20$		
	C	CF%	MSE	C	CF%	MSE
$\alpha = 1$						
LAD(1)	6.858	89.40	1.298×10^{-2}	7.000	100.00	6.361×10^{-3}
LAD(2)	6.863	90.00	1.276×10^{-2}	7.000	100.00	6.093×10^{-3}
LS(1)	6.870	91.00	1.108×10^{-2}	7.000	100.00	5.339×10^{-3}
LS(2)	6.868	91.60	1.112×10^{-2}	7.000	100.00	4.988×10^{-3}
Huber(1)	6.849	88.40	1.264×10^{-2}	7.000	100.00	6.175×10^{-3}
Huber(2)	6.850	89.60	1.301×10^{-2}	7.000	100.00	6.118×10^{-3}
$\alpha = 0.9$						
LAD(1)	6.797	89.00	1.509×10^{-2}	6.987	98.80	6.911×10^{-3}
LAD(2)	6.789	88.60	1.612×10^{-2}	6.992	99.60	6.836×10^{-3}
LS(1)	5.539	50.40	2.739×10^{-1}	6.109	68.20	2.093×10^{-1}
LS(2)	5.528	49.80	2.819×10^{-1}	6.116	69.00	1.998×10^{-1}
Huber(1)	6.807	87.80	1.316×10^{-2}	6.979	98.80	6.570×10^{-3}
Huber(2)	6.819	89.80	1.357×10^{-2}	6.981	99.00	6.768×10^{-3}
Outlier case						
LAD(1)	6.776	83.60	4.387×10^{-2}	6.976	96.60	8.873×10^{-3}
LAD(2)	6.761	82.00	4.366×10^{-2}	6.969	95.80	9.017×10^{-3}
LS(1)	4.132	45.20	4.916×10^{-1}	5.008	57.60	6.678×10^{-1}
LS(2)	4.096	46.40	5.009×10^{-1}	4.972	60.40	7.019×10^{-1}
Huber(1)	6.795	80.80	1.712×10^{-2}	6.979	96.80	1.035×10^{-2}
Huber(2)	6.801	81.60	1.809×10^{-2}	6.988	97.20	9.867×10^{-3}

influence the center (mean or median) of the conditional distribution, because the conditional mean and median of $\widetilde{\epsilon}_{(i,j)}$ are zero. Similar to Experiment 1, 500 simulation spatial data sets are independently generated from (15) over a rectangular domain $\{(i, j) : 81 \leq i \leq 81 + m, 81 \leq j \leq 81 + n\}$, for various values of m and n.

For comparison, we further consider $\rho(u) = u(\tau - \mathrm{I}_{(u<0)})$ as loss function, for different τ and r, the corresponding estimators are abbreviated as $\mathrm{Q}(\tau, r)$, here we choose $\tau = 0.25$, $\tau = 0.5$ and $\tau = 0.75$. In this experiment, to assess the variable selection results, we further use P% to denote the proportion of selecting X_2 as a relevant variable in the nonparametric component.

Tables 3 and 4 list the simulation results for the nonparametric and parametric parts, respectively. Similar to Experiment 1, in terms of variable selection, we find that Q(0.5,r), LS(r) and Huber(r) perform similarly, while due to heteroscedastic error, Q(0.5,r) and Huber(r) perform better than LS(r) does in terms of parameter estimation accuracy. Furthermore, from the values of P% in Table 3, we find that the active variable X_2 is often missed by LS(r), $\mathrm{Q}(0.5, r)$ and Huber(r) methods. However, with high probability, X_2 can be selected into the model when $\mathrm{Q}(0.75, r)$

Table 3 Simulation results for the nonparametric components in Experiment 2

	$m \times n = 15 \times 15$				$m \times n = 20 \times 20$			
	P%	CF%	μ(ASE)	μ(ADE)	P%	CF%	μ(ASE)	μ(ADE)
Q(0.75, 1)	100.00	80.60	6.362×10^{-3}	0.4062	100.00	99.00	7.218×10^{-4}	0.0502
Q(0.75, 2)	100.00	82.00	5.962×10^{-3}	0.4007	100.00	100.00	8.034×10^{-4}	0.0470
Q(0.5, 1)	6.00	77.80	1.384×10^{-3}	0.1972	0.00	100.00	5.181×10^{-4}	0.0259
Q(0.5, 2)	6.50	78.00	1.575×10^{-3}	0.1859	1.00	98.00	5.167×10^{-4}	0.0260
Q(0.25, 1)	100.00	81.80	6.310×10^{-3}	0.4072	100.00	100.00	7.203×10^{-4}	0.0498
Q(0.25, 2)	100.00	81.60	6.171×10^{-3}	0.4101	100.00	100.00	7.368×10^{-4}	0.0479
LS(1)	4.00	81.80	4.915×10^{-2}	0.2808	0.00	100.00	4.902×10^{-3}	0.0652
LS(2)	5.80	83.00	4.730×10^{-2}	0.3012	0.00	99.40	4.512×10^{-3}	0.0709
Huber(1)	7.00	82.80	2.259×10^{-2}	0.2264	0.00	100.00	3.118×10^{-3}	0.0602
Huber(2)	6.50	83.00	2.298×10^{-2}	0.2371	0.00	100.00	4.012×10^{-3}	0.0579

Table 4 Simulation results for the parametric components in Experiment 2

	$m \times n = 15 \times 15$			$m \times n = 20 \times 20$		
	C	CF%	MSE	C	CF%	MSE
Q(0.75, 1)	7.868	87.80	2.011×10^{-2}	8.000	100.00	6.673×10^{-3}
Q(0.75, 2)	7.876	89.00	1.983×10^{-2}	8.000	100.00	6.597×10^{-3}
Q(0.5, 1)	7.859	86.40	2.109×10^{-2}	8.000	100.00	7.018×10^{-3}
Q(0.5, 2)	7.868	87.00	2.133×10^{-2}	8.000	100.00	6.886×10^{-3}
Q(0.25, 1)	7.883	90.00	1.997×10^{-2}	8.000	100.00	6.639×10^{-3}
Q(0.25, 2)	7.871	88.60	2.116×10^{-2}	8.000	100.00	7.102×10^{-3}
LS(1)	7.839	85.20	2.919×10^{-2}	8.000	100.00	8.018×10^{-3}
LS(2)	7.833	84.90	2.863×10^{-2}	8.000	100.00	7.961×10^{-3}
Huber(1)	7.898	90.40	1.971×10^{-2}	8.000	100.00	6.871×10^{-3}
Huber(2)	7.902	91.00	1.983×10^{-2}	8.000	100.00	6.907×10^{-3}

and Q(0.25, r) are used. This is not surprising, since the conditional mean and median of $\widetilde{\epsilon}_{(i,j)}$ are zero, while its 25 and 75% quantiles are not zero. These demonstrate that by considering several different quantile positions, it is likely to gain a more complete picture of the underlying structure of the conditional distribution. Furthermore, by values of CF%, with high probability, all the methods can correctly select the relevant variables in the nonparametric and parametric parts simultaneously.

3.2 Real data analysis

To further illustrate the usefulness of the proposed method, we apply the new method to a real data. It was known from the simulation study that the adaptive group L_1

and L_2 penalties lead to similar results. Thus, we only consider the L_1 penalty for simplicity in the real data analysis.

The data set studied here comes from the Boston Standard Metropolitan Statistical Area with 506 observations (1 observation per census tract). This data set has been studied in several literatures. For example, Harrison and Rubinfeld (1978) investigated various methodological issues related to the use of housing data to estimate the demand for clean air, Pace and Gilley (1997) demonstrated the substantial benefits obtained by modeling the spatial dependence of the errors, Tang (2014) discussed the estimation issue for spatial functional coefficient model.

Following Harrison and Rubinfeld (1978) and Pace and Gilley (1997), we take ln (Price) as the response, NOX (levels of nitrogen oxides) as the index variable, and the following indictors as the covariates: CRIM (crime rate), RM (average number of rooms per dwelling), AGE (proportion of structures built before 1940), TAX (property tax rate), PTRATIO (pupil-teacher ratio), B (black population proportion). For the variables AGE, TAX, PTRATIO and LSTAT, we make the same modification as Tang (2014). Then, we consider the following model:

$$\begin{aligned}\ln(\text{Price}) = {} & \beta_0(\text{NOX}) + \beta_1(\text{NOX})\text{CRIM} + \beta_2(\text{NOX})\text{RM} + \beta_3(\text{NOX})\text{AGE} \\ & + \beta_4(\text{NOX})\text{TAX} + \beta_5(\text{NOX})\text{PTRATIO} + \beta_6(\text{NOX})\text{B} + \text{error}. \end{aligned} \tag{16}$$

Figure 1 presents the boxplot of response variable ln(Price), which indicates that there are outliers. Thus, we choose the robust loss function $\rho(u) = |u|$. Then, the resulting LAD(1) estimate suggests that RM, AGE, TAX, PTRATIO and B are relevant variables, whereas CRIM is inactive. Figure 2 gives the estimated coefficient functions for $\beta_0(\cdot)$, $\beta_2(\cdot)$, $\beta_3(\cdot)$, $\beta_4(\cdot)$, $\beta_5(\cdot)$ and $\beta_6(\cdot)$, respectively. Obviously, Fig. 2 shows that they are unlikely to be constant zero, because none of them is significantly close to zero. To confirm whether the eliminated variable (i.e., CRIM) is truly irrelevant, we consider non-penalized LAD estimates, which is shown in Fig. 3. We can see that it is always close to zero over the entire range of the index variable NOX; thus, Fig. 3 further confirms that CRIM is unlikely to be relevant. Therefore, the new method works well in variable selection.

For comparison, we carry out an AIC based all-subset selection procedure to select among $2^7 = 128$ candidate models. Specifically, we fit each candidate model, and calculate its AIC value, and then select the model with the smallest AIC. To better understand the performance of our penalization-based LAD(1) method and the all-subset selection procedure, in the following, we compare the two methods by assessing their selection stability and computation time.

To compare the selection stability, we apply them to 200 bootstrap samples, each of which is obtained by resampling 300 observations without replacement. Table 5 summarizes the average computation time and the frequencies that the covariates are selected in the observed data and in the bootstrap data among 200 bootstraps. We can see that both the proposed LAD(1) method and the all-subset selection method are stable in variable selection. However, the proposed LAD(1) method is much faster in computation than the all-subset selection method. This is because that all-subset approach requires searching over all possible combinations of covariates.

Fig. 1 Boxplot of ln(Price)

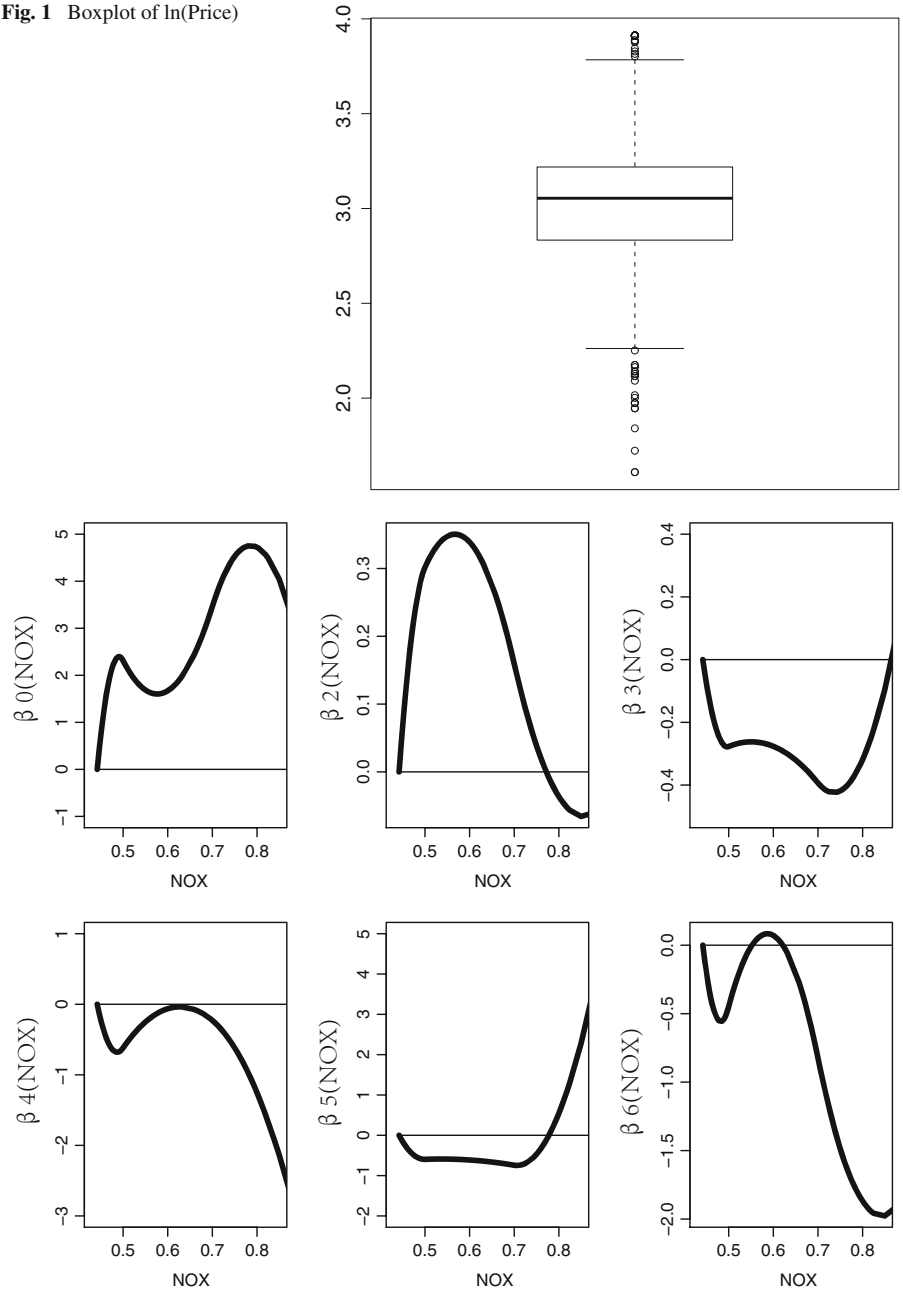

Fig. 2 Estimates of coefficient functions of relevant variables, where the *horizontal thin lines* are $y = 0$

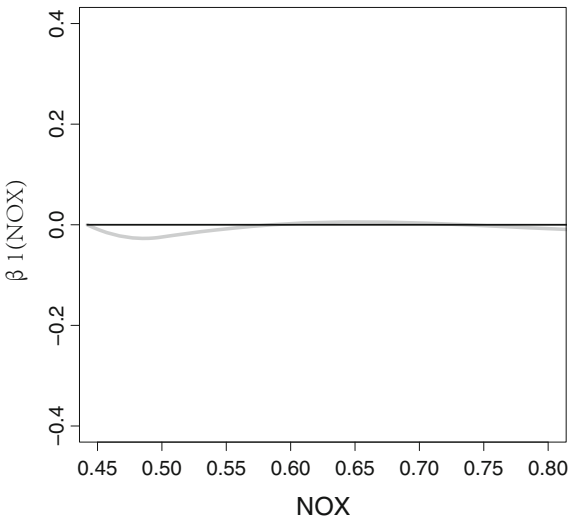

Fig. 3 Non-penalized LAD estimator for coefficient function of variable CRIM; the *horizontal black thin line* is $y = 0$

Table 5 Frequencies that the covariates are selected in the observed sample and in the bootstrap samples. Av.Time: the average computation time of the methods

	$\beta_0(\cdot)$	$\beta_1(\cdot)$	$\beta_2(\cdot)$	$\beta_3(\cdot)$	$\beta_4(\cdot)$	$\beta_5(\cdot)$	$\beta_6(\cdot)$	Av.Time(sd)
All-subset								
Observed	1	0	1	1	1	0	1	/
Bootstrap	200	15	192	190	187	23	165	18.32(2.58)
LAD(1)								
Observed	1	0	1	1	1	1	1	/
Bootstrap	200	2	198	200	177	190	200	3.11(1.06)

4 Appendix

Let C denote some positive constants not depending on $\boldsymbol{n}$, but which may assume different values at each appearance. We first list some notations used in the following, define $\boldsymbol{Z} = (\boldsymbol{Z}_{\mathbb{i}}, \mathbb{i} \in \mathcal{I}_{\boldsymbol{n}})^T$, $\boldsymbol{\Pi} = (\boldsymbol{\Pi}_{\mathbb{i}}, \mathbb{i} \in \mathcal{I}_{\boldsymbol{n}})^T$, $\boldsymbol{B} = \text{diag}(b_{\mathbb{i}}, \mathbb{i} \in \mathcal{I}_{\boldsymbol{n}})$, $\boldsymbol{\Delta} = \text{diag}(E[\Psi^2(\epsilon_{\mathbb{i}})], \mathbb{i} \in \mathcal{I}_{\boldsymbol{n}})$, $\boldsymbol{P} = \boldsymbol{\Pi}(\boldsymbol{\Pi}^T \boldsymbol{B} \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}^T \boldsymbol{B}$, $\widehat{\boldsymbol{Z}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Z}$ and $\widehat{\boldsymbol{\Xi}}_{\boldsymbol{n}} = \widehat{\boldsymbol{Z}}^T \boldsymbol{B} \widehat{\boldsymbol{Z}}$, $\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{n}} = \widehat{\boldsymbol{Z}}^T \boldsymbol{\Delta} \widehat{\boldsymbol{Z}}$. Let $R_{\boldsymbol{n}\mathbb{i}k} = \alpha_{0k}(U_{\mathbb{i}}) - \boldsymbol{\pi}(U_{\mathbb{i}})^T \boldsymbol{\theta}_k^0$ and $R_{\boldsymbol{n}\mathbb{i}} = \boldsymbol{X}_{\mathbb{i}}^T \boldsymbol{\alpha}_0(U_{\mathbb{i}}) - \boldsymbol{\Pi}_{\mathbb{i}}^T \boldsymbol{\Theta}^0$.

Lemma 1 *Suppose that C1-C4 hold. There exists a vector* $\boldsymbol{\Theta}^0 = ({\boldsymbol{\theta}_1^0}^T, \ldots, {\boldsymbol{\theta}_p^0}^T)^T$ *which satisfies*

(I) $\|\boldsymbol{\theta}_k^0\|_1 \neq 0, k \in \{1, \ldots, v\}$; $\|\boldsymbol{\theta}_k^0\|_1 = 0, k \in \{v+1, \ldots, p\}$;
(II) $\sup_{u \in [U_0, U_1]} |\alpha_{0k}(u) - \boldsymbol{\pi}(u)^T \boldsymbol{\theta}_k^0| = O(K_{\boldsymbol{n}}^{-\gamma}), k = 1, \ldots, p.$

Lemma 1 *follows directly from Corollary* 6.21 *of (Schumaker 1981, Chapter 6). So by Lemma* 1 *and condition C*3*, we have*

$$\sup_{\mathbb{i},k} |R_{\boldsymbol{n}\mathbb{i}k}| = O(K_{\boldsymbol{n}}^{-\gamma}) \text{ and } \sup_{\mathbb{i}} |R_{\boldsymbol{n}\mathbb{i}}| = O(K_{\boldsymbol{n}}^{-\gamma}).$$

Lemma 2 *Suppose Equation* (2) *holds. Let* $\mathcal{L}_p(\mathcal{F})$ *denote the class of* $\mathcal{F}$*-measurable random variable* ξ *satisfying* $\|\xi\|_p = (E|\xi|^p)^{1/p} < \infty$*. Let* $\xi \in \mathcal{L}_p(\mathcal{B}(\mathcal{S}))$ *and* $\eta \in \mathcal{L}_p(\mathcal{B}(\mathcal{S}'))$*. Then, for any* $1 \le p, h, q < \infty$ *such that* $\frac{1}{p} + \frac{1}{q} + \frac{1}{h} = 1$*,*

$$|E(\xi\eta) - E\xi E\eta| \le C\|\xi\|_p \|\eta\|_q [\alpha\left(\mathcal{B}(\mathcal{S}), \mathcal{B}(\mathcal{S}')\right)]^{1/h}.$$

The proof of Lemma 2 *can be found in Lemma 5.1 of Hallin et al. (2004). The following Lemma* 3 *can be founded in Lee et al. (2004).*

Lemma 3 *Let* $(Z_{\mathbb{i}} : \mathbb{i} \in \mathbb{Z}^N)$ *be a zero-mean real valued random fields such that* $\sup_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \|Z_{\mathbb{i}}\|_2 \le b_0 < \infty$*. Then, for each* $\boldsymbol{q} = (q_1, \ldots, q_N)$ *with integer-valued coordinates* $q_k \in [1, n_k/2]$ *and for each* $\varepsilon > 0$ *we have*

$$P\left(\left|\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} Z_{\mathbb{i}}\right| \ge \widehat{\boldsymbol{n}}\varepsilon\right) \le 2^N \left\{2\exp\left(-\frac{\varepsilon^2 \widehat{\boldsymbol{q}}}{2^{N+1} v^2(\boldsymbol{q})}\right) + \frac{4b_0}{\varepsilon}\varphi\left(\min_{1\le k\le N} p_k\right)\right\},$$

where $\widehat{\boldsymbol{q}} = \prod_{k=1}^N q_k$*,* $p_k = n_k/(2q_k)$ *and* $v^2(\boldsymbol{q}) = 2^{N+1}\sigma^2(\boldsymbol{q})/\widehat{\boldsymbol{p}}^2 + b_0\varepsilon$ *with* $\widehat{\boldsymbol{p}} = \prod_{k=1}^N p_k$*,* $\sigma^2(\boldsymbol{q}) = \min_{\mathbb{i},\mathbb{j}} E(\sum_{\mathbb{k}\in\mathbb{A}_{\mathbb{ij}}} Z_{\mathbb{k}})^2$ *and* $\mathbb{A}_{\mathbb{ij}} = \prod_{k=1}^N ((i_k + 2j_k)p_k, (i_k + 2j_k + 1)p_k]$*. The minimization in the defining equation for* $\sigma^2(\boldsymbol{q})$ *is taken over all pairs of* N*-tuple indices* i *and* j *with* $i_k = 0, 1$ *and* $i_k = 0, 1, \ldots, q_k - 1$*.*

Proof (I) of Theorem 2 For any given $\boldsymbol{\Theta} \in \mathbb{R}^{pq_{\boldsymbol{n}}}$ and $\boldsymbol{\beta} \in \mathbb{R}^q$, let

$$\boldsymbol{\varsigma}(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \begin{pmatrix} \boldsymbol{\varsigma}_1 \\ \boldsymbol{\varsigma}_2 \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{n}}^{-1/2} \widehat{\boldsymbol{\Xi}}_{\boldsymbol{n}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\ K_{\boldsymbol{n}}^{-1/2} \boldsymbol{H}_{\boldsymbol{n}}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^0) + K_{\boldsymbol{n}}^{1/2} \boldsymbol{H}_{\boldsymbol{n}}^{-1} \boldsymbol{\Pi}^T \boldsymbol{B}\boldsymbol{Z}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \end{pmatrix}, \tag{17}$$

where $\boldsymbol{H}_{\boldsymbol{n}}^2 = K_{\boldsymbol{n}} \boldsymbol{\Pi}^T \boldsymbol{B} \boldsymbol{\Pi}$, and we standardize $\widetilde{\boldsymbol{Z}}_{\mathbb{i}} = \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{n}}^{1/2} \widehat{\boldsymbol{\Xi}}_{\boldsymbol{n}}^{-1} \widehat{\boldsymbol{Z}}_{\mathbb{i}}$, $\widetilde{\boldsymbol{\Pi}}_{\mathbb{i}} = K_{\boldsymbol{n}}^{1/2} \boldsymbol{H}_{\boldsymbol{n}}^{-1} \boldsymbol{\Pi}_{\mathbb{i}}$. So we have

$$\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \rho(Y_{\mathbb{i}} - \boldsymbol{\Pi}_{\mathbb{i}}^T \boldsymbol{\Theta} - \boldsymbol{Z}_{\mathbb{i}}^T \boldsymbol{\beta}) = \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \rho(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} - \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}). \tag{18}$$

Let

$$\Phi_{\boldsymbol{n}}(\boldsymbol{\varsigma}) = \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \rho(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} - \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) - \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \rho(\epsilon_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}})$$

$$
\begin{aligned}
&= \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \{E[\rho(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} - \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) - \rho(\epsilon_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}})|\boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, \boldsymbol{U}_{\mathbb{i}}] \\
&\quad - [\boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} + \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}]\Psi(\epsilon_{\mathbb{i}})\} \\
&\quad + \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} [\rho(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} - \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) - \rho(\epsilon_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) + [\boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} \\
&\quad + \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}]\Psi(\epsilon_{\mathbb{i}}) - E\{\rho(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} - \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) - \rho(\epsilon_{\mathbb{i}} \\
&\quad + R_{\boldsymbol{n}\mathbb{i}})|\boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, \boldsymbol{U}_{\mathbb{i}}\}] \\
&= \Phi_{\boldsymbol{n}1}(\boldsymbol{\varsigma}) + \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} [\Gamma_{\mathbb{i}}(\boldsymbol{\varsigma}) - E\{\rho(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} - \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) \\
&\quad - \rho(\epsilon_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}})|\boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, \boldsymbol{U}_{\mathbb{i}}\}] \\
&= \Phi_{\boldsymbol{n}1}(\boldsymbol{\varsigma}) + \Phi_{\boldsymbol{n}2}(\boldsymbol{\varsigma}),
\end{aligned} \tag{19}
$$

where $\Gamma_{\mathbb{i}}(\boldsymbol{\varsigma}) = \rho(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} - \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) - \rho(\epsilon_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) + [\boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} + \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}]\Psi(\epsilon_{\mathbb{i}})$. By the definition of $Q(\boldsymbol{\Theta}, \boldsymbol{\beta})$, we have

$$
Q(\boldsymbol{\Theta}, \boldsymbol{\beta}) - Q(\boldsymbol{\Theta}^0, \boldsymbol{\beta}_0) = \Phi_{\boldsymbol{n}}(\boldsymbol{\varsigma}) + \sum_{k=1}^{p} \lambda_{\boldsymbol{n}k}^{*}(\|\boldsymbol{\theta}_k\|_r - \|\boldsymbol{\theta}_k^0\|_r) + \sum_{k=1}^{q} \lambda_{\boldsymbol{n}k}^{**}(|\beta_k| - |\beta_{0k}|). \tag{20}
$$

Then, a sufficient condition for proving this part is for any $\varepsilon > 0$ and $r \geq 1$; there exists some C, such that

$$
\Pr\left(\inf_{\|\boldsymbol{\varsigma}\|_2 = C K_{\boldsymbol{n}}^{1/2}} Q(\boldsymbol{\Theta}, \boldsymbol{\beta}) > Q(\boldsymbol{\Theta}^0, \boldsymbol{\beta}_0) \right) > 1 - \varepsilon, \tag{21}
$$

as $\boldsymbol{n} \to \infty$. Specially, using the fact that $Q(\boldsymbol{\Theta}, \boldsymbol{\beta})$ is minimized at $\widehat{\boldsymbol{\varsigma}}$. By the convexity of $Q(\cdot)$ and Corollary 2.5 of Eggleston (1958, Chapter 3), (21) can lead to $\Pr\left(\|\widehat{\boldsymbol{\varsigma}}\|_2 \leq C K_{\boldsymbol{n}}^{1/2}\right) > 1 - \varepsilon$, as $\boldsymbol{n} \to \infty$, and thus $\|\widehat{\boldsymbol{\varsigma}}\|_2 = O_p(K_{\boldsymbol{n}}^{1/2})$. Then, by Lemma 1, we can obtain

$$
\begin{aligned}
\widehat{\boldsymbol{n}}^{-1} \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \{\widehat{\alpha}_l(U_{\mathbb{i}}) - \alpha_{0l}(U_{\mathbb{i}})\}^2 &\leq 2\widehat{\boldsymbol{n}}^{-1} \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \{\boldsymbol{\pi}_{\mathbb{i}}^T (\widehat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l^0)\}^2 + 2C K_{\boldsymbol{n}}^{-2\gamma} \\
&\leq O_p(\widehat{\boldsymbol{n}}^{-1} \|\widehat{\boldsymbol{\varsigma}}_2\|_2^2) + O_p(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2) + O(K_{\boldsymbol{n}}^{-2\gamma}) \\
&= O_p(K_{\boldsymbol{n}}^{-2\gamma}) = O_p\left(\widehat{\boldsymbol{n}}^{-2\gamma/(2\gamma+1)}\right).
\end{aligned} \tag{22}
$$

Next, we will prove (21). Firstly, by the convex property of $\rho(\cdot)$, we have

$$
\begin{aligned}
&|\rho(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} - \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) - \rho(\epsilon_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) + [\boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} + \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}]\Psi(\epsilon_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}})| \\
&\leq |\boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} + \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}| \cdot |\Psi(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} - \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) - \Psi(\epsilon_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}})|,
\end{aligned} \tag{23}
$$

thus

$$
\begin{aligned}
&|\Gamma_{\mathbb{i}}(\boldsymbol{\varsigma})| \\
&\le |\boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} + \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}| \cdot (|\Psi(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} - \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) - \Psi(\epsilon_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}})| + |\Psi(\epsilon_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) - \Psi(\epsilon_{\mathbb{i}})|) \\
&\le |\boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} + \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}| \cdot (|\Psi(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} - \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) - \Psi(\epsilon_{\mathbb{i}})| + 2|\Psi(\epsilon_{\mathbb{i}} + R_{\boldsymbol{n}\mathbb{i}}) - \Psi(\epsilon_{\mathbb{i}})|).
\end{aligned} \tag{24}
$$

Thus, when $\|\boldsymbol{\varsigma}\|_2 \le c K_{\boldsymbol{n}}^{1/2}$, by condition C5 and note that $E\left[\Psi(\epsilon_{\mathbb{i}}) \mid \boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, U_{\mathbb{i}}\right] = 0$, we have that

$$
\max_{\mathbb{i} \in \mathcal{I}_{\boldsymbol{n}}} |\Phi_{2\mathbb{i}}(\boldsymbol{\varsigma})| = \max_{\mathbb{i} \in \mathcal{I}_{\boldsymbol{n}}} |\Gamma_{\mathbb{i}}(\boldsymbol{\varsigma}) - E\left(\Gamma_{\mathbb{i}}(\boldsymbol{\varsigma}) \mid \boldsymbol{X}_{\mathbb{i}}, \boldsymbol{Z}_{\mathbb{i}}, U_{\mathbb{i}}\right)| \le \frac{C \cdot c \cdot K_{\boldsymbol{n}}}{\widehat{\boldsymbol{n}}^{1/2}}. \tag{25}
$$

Furthermore, by regularity conditions C1 and C5,

$$
\begin{aligned}
\sigma^2(\boldsymbol{q}) &= \min_{\mathbb{i},\mathbb{j}} E\left(\sum_{\boldsymbol{k} \in \mathbb{A}_{\mathbb{i}\mathbb{j}}} \Phi_{2\boldsymbol{k}}(\boldsymbol{\varsigma})\right)^2 \\
&= E\left(\sum_{\boldsymbol{k} \in \mathbb{A}_{00}} \Phi_{2\boldsymbol{k}}(\boldsymbol{\varsigma})\right)^2 \\
&= \sum_{\boldsymbol{k} \in \mathbb{A}_{00}} E\Phi_{2\boldsymbol{k}}^2(\boldsymbol{\varsigma}) + \sum_{\mathbb{i},\mathbb{j} \in \mathbb{A}_{00}} \sum_{\mathbb{i} \ne \mathbb{j}} E\Phi_{2\mathbb{i}}(\boldsymbol{\varsigma})\Phi_{2\mathbb{j}}(\boldsymbol{\varsigma}),
\end{aligned} \tag{26}
$$

where $\mathbb{A}_{00} = \prod_{k=1}^{N}(0,\, p_k]$. For the first part of formula (26), one can verify

$$
\begin{aligned}
\sum_{\boldsymbol{k} \in \mathbb{A}_{00}} E\Phi_{2\boldsymbol{k}}^2(\boldsymbol{\varsigma}) &\le \sum_{\boldsymbol{k} \in \mathbb{A}_{00}} E\Gamma_{\boldsymbol{k}}^2(\boldsymbol{\varsigma}) \\
&\le \sum_{\boldsymbol{k} \in \mathbb{A}_{00}} C K_{\boldsymbol{n}} E[(\boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} + \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}})^2 (K_{\boldsymbol{n}}^{1/2} |\boldsymbol{\varsigma}_1^T \widetilde{\boldsymbol{Z}}_{\mathbb{i}} + \boldsymbol{\varsigma}_2^T \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}| + R_{\boldsymbol{n}\mathbb{i}})] \\
&\le C K_{\boldsymbol{n}} \widehat{\boldsymbol{p}} \|\boldsymbol{\varsigma}\|_2^2 \frac{\|\boldsymbol{\varsigma}\|_2 K_{\boldsymbol{n}} \widehat{\boldsymbol{n}}^{-1/2} + K_{\boldsymbol{n}}^{-\gamma}}{\widehat{\boldsymbol{n}}}.
\end{aligned} \tag{27}
$$

Note that

$$
\sum_{\mathbb{i},\mathbb{j} \in \mathbb{A}_{00}} \sum_{\mathbb{i} \ne \mathbb{j}} E\Phi_{2\mathbb{i}}(\boldsymbol{\varsigma})\Phi_{2\mathbb{j}}(\boldsymbol{\varsigma}) = \sum_{\mathbb{i},\mathbb{j} \in \mathbb{S}_1} E\Phi_{2\mathbb{i}}(\boldsymbol{\varsigma})\Phi_{2\mathbb{j}}(\boldsymbol{\varsigma}) + \sum_{\mathbb{i},\mathbb{j} \in \mathbb{S}_2} E\Phi_{2\mathbb{i}}(\boldsymbol{\varsigma})\Phi_{2\mathbb{j}}(\boldsymbol{\varsigma}), \tag{28}
$$

where

$$
\begin{aligned}
\mathbb{S}_1 &= \{\mathbb{i} \ne \mathbb{j} \in \mathbb{A}_{00} : |j_k - i_k| \le c_{\boldsymbol{n}k}, k = 1, \ldots, N\}, \\
\mathbb{S}_2 &= \{\mathbb{i}, \mathbb{j} \in \mathbb{A}_{00} : |j_k - i_k| > c_{\boldsymbol{n}k}, k = 1, \ldots, N\},
\end{aligned}
$$

$c_{\boldsymbol{n}k} = [K_{\boldsymbol{n}}^{\delta/(2+\delta)a}]$, for $k = 1, \ldots, N$, and constant $a > (4+\delta)N/(2+\delta)$. According to formula (27), we can obtain that

$$\begin{aligned}\sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_1} |E\Phi_{2\mathbb{i}}(\boldsymbol{\varsigma})\Phi_{2\mathbb{j}}(\boldsymbol{\varsigma})| &\le \sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_1} [E\Phi_{2\mathbb{i}}^2(\boldsymbol{\varsigma})]^{1/2}[E\Phi_{2\mathbb{j}}^2(\boldsymbol{\varsigma})]^{1/2} \\ &\le CK_{\boldsymbol{n}}\widehat{\boldsymbol{p}}\prod_{k=1}^{N} c_{\boldsymbol{n}k}\|\boldsymbol{\varsigma}\|_2^2 \frac{\|\boldsymbol{\varsigma}\|_2 K_{\boldsymbol{n}}\widehat{\boldsymbol{n}}^{-1/2} + K_{\boldsymbol{n}}^{-\gamma}}{\widehat{\boldsymbol{n}}} \\ &= O\left(\frac{\widehat{\boldsymbol{p}}K_{\boldsymbol{n}}^{2+\delta N/(2+\delta)a}}{\widehat{\boldsymbol{n}}^{3/2}}\right). \end{aligned} \tag{29}$$

By regularity conditions C2, C3 and C5, and let $p = h = 2+\delta$, $q = (2+\delta)/\delta$ in Lemma 2, we can obtain that

$$\begin{aligned}&\sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_2} E\Phi_{2\mathbb{i}}(\boldsymbol{\varsigma})\Phi_{2\mathbb{j}}(\boldsymbol{\varsigma}) \\ &\le \sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_2} C[E(|\Phi_{2\mathbb{i}}(\boldsymbol{\varsigma})|^{2+\delta})]^{2/(2+\delta)}[\varphi(\|\mathbb{i}-\mathbb{j}\|_2)]^{\delta/(2+\delta)} \\ &\le \sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_2} C[E(|\Gamma_{\mathbb{i}}(\boldsymbol{\varsigma})|^{2+\delta})]^{2/(2+\delta)}[\varphi(\|\mathbb{i}-\mathbb{j}\|_2)]^{\delta/(2+\delta)} \\ &\le \sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_2} C\{E[|\boldsymbol{\varsigma}_1^T\widetilde{\boldsymbol{Z}}_{\mathbb{i}} + \boldsymbol{\varsigma}_2^T\widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}|^{2+\delta}(|\boldsymbol{\varsigma}_1^T\widetilde{\boldsymbol{Z}}_{\mathbb{i}} + \boldsymbol{\varsigma}_2^T\widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}| + |R_{\boldsymbol{n}\mathbb{i}}|)]\}^{2/(2+\delta)}[\varphi(\|\mathbb{i}-\mathbb{j}\|_2)]^{\delta/(2+\delta)} \\ &\le \sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_2} CK_{\boldsymbol{n}}^2\widehat{\boldsymbol{n}}^{-(3+\delta)/(2+\delta)}\|\boldsymbol{\varsigma}\|_2^{2+2/(2+\delta)}[\varphi(\|\mathbb{i}-\mathbb{j}\|_2)]^{\delta/(2+\delta)}. \end{aligned} \tag{30}$$

Note that $\varphi(t) = O(\exp(-\kappa t))$, by the similar arguments used in Lemma 5.2 in Hallin et al. (2004),

$$\begin{aligned}\sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_2} [\varphi(\|\mathbb{i}-\mathbb{j}\|_2)]^{2/(2+\delta)} &\le C\widehat{\boldsymbol{p}}\sum_{k=1}^{N}\sum_{t=c_{\boldsymbol{n}k}}^{\|\boldsymbol{p}\|_2} t^{N-1}[\varphi(t)]^{\delta/(2+\delta)} \\ &\le C\widehat{\boldsymbol{p}}\exp\left(\frac{-\zeta_1 K_{\boldsymbol{n}}^{\delta/(2+\delta)a}\delta}{2+\delta}\right), \end{aligned} \tag{31}$$

where constant $\zeta_1 \in (0, \kappa)$. So $\sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_2} E\Phi_{2\mathbb{i}}(\boldsymbol{\varsigma})\Phi_{2\mathbb{j}}(\boldsymbol{\varsigma}) = O_p(\widehat{\boldsymbol{p}}K_{\boldsymbol{n}}^{2+\delta N/(2+\delta)a}\widehat{\boldsymbol{n}}^{\frac{-3}{2}})$, furthermore, by formulas (26)–(31), we can get that

$$\sigma^2(\boldsymbol{q}) = O\left(\frac{K_{\boldsymbol{n}}^{2+\delta N/(2+\delta)a}\widehat{\boldsymbol{p}}}{\widehat{\boldsymbol{n}}^{3/2}}\right). \tag{32}$$

Thus, combining (26), (32) and Lemma 3, for any $\varepsilon > 0$, we can obtain

$$\begin{aligned}
&\Pr\left\{\sup_{\|\boldsymbol{\varsigma}\|_2 \le C K_{\boldsymbol{n}}^{1/2}} \left|\frac{1}{K_{\boldsymbol{n}}}\Phi_{\boldsymbol{n}2}(\boldsymbol{\varsigma})\right| > \varepsilon\right\} \\
&\le 2^N \left(\frac{4C K_{\boldsymbol{n}} \widehat{\boldsymbol{n}}^{1/2} c}{\varepsilon}\right)^{pK_{\boldsymbol{n}}} \left\{2\exp\left(-\frac{K_{\boldsymbol{n}}^2 \varepsilon^2 \widehat{\boldsymbol{q}}}{2^{N+2}(2^{N+2}\widehat{\boldsymbol{n}}^2\sigma^2(\boldsymbol{q})/\widehat{\boldsymbol{p}}^2 + b_0 \widehat{\boldsymbol{n}} K_{\boldsymbol{n}} \varepsilon)}\right)\right. \\
&\quad \left. + \frac{8\widehat{\boldsymbol{n}} b_0}{K_{\boldsymbol{n}}\varepsilon}\varphi(\min_{1\le k\le N} p_k)\right\} \\
&\le 2^{N+1}\left\{\exp\left(pK_{\boldsymbol{n}}\log\widehat{\boldsymbol{n}}\left[1 - \frac{K_{\boldsymbol{n}}\varepsilon^2\widehat{\boldsymbol{q}}}{2^{N+2}p\log\widehat{\boldsymbol{n}}(2^{N+2}\widehat{\boldsymbol{n}}^2\sigma^2(\boldsymbol{q})/\widehat{\boldsymbol{p}}^2 + b_0\widehat{\boldsymbol{n}}K_{\boldsymbol{n}}\varepsilon)}\right]\right)\right. \\
&\quad \left. + \exp\left(pK_{\boldsymbol{n}}\log\widehat{\boldsymbol{n}}\left[1 - \kappa \min_{1\le k\le N}\frac{p_k}{pK_{\boldsymbol{n}}\log\widehat{\boldsymbol{n}}}\right]\right)\right\} \\
&= o(1). \qquad (33)
\end{aligned}$$

What is more, for arbitrary $\varepsilon > 0$, we can prove that, when $\boldsymbol{n} \to \infty$, there exists large constant C, such that

$$\Pr\left\{\inf_{\|\boldsymbol{\varsigma}\|_2 = C_\varepsilon K_{\boldsymbol{n}}^{1/2}} \frac{1}{K_{\boldsymbol{n}}}\Phi_{\boldsymbol{n}1}(\boldsymbol{\varsigma}) > 0\right\} > 1 - \varepsilon. \qquad (34)$$

Thus, by formulas (33) and (34), we have that, for some large enough positive constant C,

$$\lim_{\boldsymbol{n}\to\infty}\Pr\left(\inf_{\|\boldsymbol{\varsigma}\|_2 = C K_{\boldsymbol{n}}^{1/2}} \frac{1}{K_{\boldsymbol{n}}}\Phi_{\boldsymbol{n}}(\boldsymbol{\varsigma}) > 0\right) = 1. \qquad (35)$$

Then note, for any m-dimension vector ζ and $r \ge 1$,

$$\|\zeta\|_r \le \|\zeta\|_1 \le m^{1/2}\|\zeta\|_2, \qquad (36)$$

holds. So for $r \ge 1$,

$$\left|\|\boldsymbol{\theta}_k\|_r - \|\boldsymbol{\theta}_k^0\|_r\right| \le \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^0\|_r \le \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^0\|_1 \le q_{\boldsymbol{n}}^{1/2}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^0\|_2. \qquad (37)$$

When $\|\boldsymbol{\varsigma}\|_2 = C K_{\boldsymbol{n}}^{1/2}$, by (37), we have

$$\begin{aligned}
\sum_{k=1}^{p}\lambda_{\boldsymbol{n}k}^*(\|\boldsymbol{\theta}_k\|_r - \|\boldsymbol{\theta}_k^0\|_r) &\ge \sum_{k=1}^{s_0}\lambda_{\boldsymbol{n}k}^*(\|\boldsymbol{\theta}_k\|_r - \|\boldsymbol{\theta}_k^0\|_r) \\
&\ge a_{\boldsymbol{n}1}^* \sum_{k=1}^{s_0}(\|\boldsymbol{\theta}_k\|_r - \|\boldsymbol{\theta}_k^0\|_r)
\end{aligned}$$

$$\begin{aligned} &\geq -vCa^{*}_{\boldsymbol{n}1}q_{\boldsymbol{n}}^{1/2}K_{\boldsymbol{n}}\widehat{\boldsymbol{n}}^{-1/2} \\ &= o(K_{\boldsymbol{n}}), \end{aligned} \tag{38}$$

and similarly,

$$\sum_{k=1}^{q}\lambda^{**}_{\boldsymbol{n}k}(|\beta_k| - |\beta_{0k}|) \geq -c(a^{**}_{\boldsymbol{n}}CK_{\boldsymbol{n}}^{1/2}\boldsymbol{n}^{-1/2}) = o_p(1). \tag{39}$$

Thus, for any $\varepsilon > 0$, there exists some constant C, such that

$$\begin{aligned} &\lim_{\boldsymbol{n}\to\infty}\Pr\left(\inf_{\|\boldsymbol{\varsigma}\|_2=CK_{\boldsymbol{n}}^{1/2}}\Phi_{\boldsymbol{n}}(\boldsymbol{\varsigma}) + \sum_{k=1}^{p}\lambda^{*}_{\boldsymbol{n}k}(\|\boldsymbol{\theta}_k\|_r - \|\boldsymbol{\theta}^0_k\|_r) + \sum_{k=1}^{q}\lambda^{**}_{\boldsymbol{n}k}(|\beta_k| - |\beta_{0k}|) > 0\right) \\ &> 1 - \varepsilon. \end{aligned} \tag{40}$$

□

Proof (I) of Theorem 1 Now, we prove $\widehat{\boldsymbol{\theta}}_k = \mathbf{0}, k = v+1, \ldots, p$ with probability tending to one. A sufficient condition is that

$$\begin{cases} \frac{\partial Q(\boldsymbol{\Theta},\boldsymbol{\beta})}{\partial\theta_{k,s}} < 0, & \theta_{k,s} < 0, \\ \frac{\partial Q(\boldsymbol{\Theta},\boldsymbol{\beta})}{\partial\theta_{k,s}} > 0, & \theta_{k,s} > 0 \end{cases}$$

holds for $k = v+1, \ldots, p$ and $s = 1, \ldots, q_{\boldsymbol{n}}$. This suffices to prove that

$$\left\|\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}}\Psi\left(Y_{\mathbb{i}} - \boldsymbol{\Pi}_{\mathbb{i}}^T\widehat{\boldsymbol{\Theta}} - \boldsymbol{Z}_{\mathbb{i}}^T\widehat{\boldsymbol{\beta}}\right)X_{\mathbb{i}k}\boldsymbol{\pi}_{\mathbb{i}}\right\|_2 \leq \lambda_{\boldsymbol{n}k}, k = v+1, \ldots, p, \tag{41}$$

for $r = 1, 2$.

Let $S_{\boldsymbol{n}}(\boldsymbol{\Theta}, \boldsymbol{\beta}) = \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}}\Psi(Y_{\mathbb{i}} - \boldsymbol{\Pi}_{\mathbb{i}}^T\boldsymbol{\Theta} - \boldsymbol{Z}_{\mathbb{i}}^T\boldsymbol{\beta})\boldsymbol{\Pi}_{\mathbb{i}}$, $S^0_{\boldsymbol{n}} = \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}}\Psi(\epsilon_{\mathbb{i}})\boldsymbol{\Pi}_{\mathbb{i}}$, and $\Lambda_{\mathbb{i}} = \Psi^2(\epsilon_{\mathbb{i}})\boldsymbol{\Pi}_{\mathbb{i}}^T\boldsymbol{\Pi}_{\mathbb{i}}$, $\Lambda_{\mathbb{i}\mathbb{j}} = \Psi(\epsilon_{\mathbb{i}})\Psi(\epsilon_{\mathbb{j}})\boldsymbol{\Pi}_{\mathbb{i}}^T\boldsymbol{\Pi}_{\mathbb{j}}$, then

$$E\left\{\left[\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}}\Psi(\epsilon_{\mathbb{i}})\boldsymbol{\Pi}_{\mathbb{i}}\right]^T\left[\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}}\Psi(\epsilon_{\mathbb{i}})\boldsymbol{\Pi}_{\mathbb{i}}\right]\right\} = \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}}E\Lambda_{\mathbb{i}} + \sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_1}E\Lambda_{\mathbb{i}\mathbb{j}} + \sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_2}E\Lambda_{\mathbb{i}\mathbb{j}}. \tag{42}$$

Thus, by the regularity condition C4, we have

$$\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}}E\Lambda_{\mathbb{i}} = \sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}}E\left(E(\Psi^2(\epsilon_{\mathbb{i}})|\boldsymbol{X}_{\mathbb{i}}, U_{\mathbb{i}})\boldsymbol{\Pi}_{\mathbb{i}}^T\boldsymbol{\Pi}_{\mathbb{i}}\right) \leq C\widehat{\boldsymbol{n}}. \tag{43}$$

Furthermore, let $L_{\boldsymbol{n}} = K_{\boldsymbol{n}}^{2/(4+\delta)}$, $\boldsymbol{\Pi}_{\mathbb{i}1} = \boldsymbol{\Pi}_{\mathbb{i}}\Psi(\epsilon_{\mathbb{i}})\mathrm{I}_{|\Psi(\epsilon_{\mathbb{i}})|\le L_{\boldsymbol{n}}}$ and $\boldsymbol{\Pi}_{\mathbb{i}2} = \boldsymbol{\Pi}_{\mathbb{i}}\Psi(\epsilon_{\mathbb{i}})\mathrm{I}_{|\Psi(\epsilon_{\mathbb{i}})|>L_{\boldsymbol{n}}}$. Using arguments similar to those used in the proof of Theorem 1 of Tang and Cheng (2009), one can verify

$$\sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_1} E\Lambda_{\mathbb{i}\mathbb{j}} \le CK_{\boldsymbol{n}}(L_{\boldsymbol{n}}^{-\delta/2} + L_{\boldsymbol{n}}^2 K_{\boldsymbol{n}}^{-1}) \prod_{k=1}^{N} c_{\boldsymbol{n}k} = o(\widehat{\boldsymbol{n}}). \tag{44}$$

Using arguments similar to those used in the proof of Lemma 5.2 of Hallin et al. (2004), and noting that $\varphi(u) = O(\exp(-\kappa u))$, we have

$$\begin{aligned}
\sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_2} E\Lambda_{\mathbb{i}\mathbb{j}} &\le CK_{\boldsymbol{n}}^{\delta/(2+\delta)} \sum_{\mathbb{i},\mathbb{j}\in\mathbb{S}_2} (\varphi(\|\mathbb{i}-\mathbb{j}\|_2))^{\delta/(2+\delta)} \\
&\le C\widehat{\boldsymbol{p}} \sum_{k=1}^{N} \sum_{u=c_{\boldsymbol{n}k}}^{\|\boldsymbol{p}\|_2} u^{N-1}(\varphi(u))^{\delta/(2+\delta)} \\
&\le C\widehat{\boldsymbol{p}} \exp\left(\frac{-\kappa(K_{\boldsymbol{n}}^{\delta/(2+\delta)a})\delta}{2+\delta}\right) \\
&= o(\widehat{\boldsymbol{n}}).
\end{aligned} \tag{45}$$

Thus, $\|S_{\boldsymbol{n}}^0\|_2 = O_p(\widehat{\boldsymbol{n}}^{1/2})$, and similarly,

$$\begin{aligned}
\left\|E\{S_{\boldsymbol{n}}(\boldsymbol{\Theta}^0,\boldsymbol{\beta}_0) - S_{\boldsymbol{n}}^0\}\right\|_2 &= \left\|\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \boldsymbol{\Pi}_{\mathbb{i}} E\{\Psi(\epsilon_{\mathbb{i}} - R_{\boldsymbol{n}\mathbb{i}}) - \Psi(\epsilon_{\mathbb{i}})\}\right\|_2 \\
&= O_p\left(\left\|\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} \boldsymbol{\Pi}_{\mathbb{i}} b_{\mathbb{i}} R_{\boldsymbol{n}\mathbb{i}}\right\|_2\right) \\
&= O_p(K_{\boldsymbol{n}}^{-\gamma}\widehat{\boldsymbol{n}}^{1/2}),
\end{aligned} \tag{46}$$

$$\begin{aligned}
&E\left\{\left[\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} (\Psi(\epsilon_{\mathbb{i}} - R_{\boldsymbol{n}\mathbb{i}}) - \Psi(\epsilon_{\mathbb{i}}))\boldsymbol{\Pi}_{\mathbb{i}}\right]^T \left[\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} (\Psi(\epsilon_{\mathbb{i}} - R_{\boldsymbol{n}\mathbb{i}}) - \Psi(\epsilon_{\mathbb{i}}))\boldsymbol{\Pi}_{\mathbb{i}}\right]\right\} \\
&= O_p\left(\frac{\widehat{\boldsymbol{n}}^{1/2}}{K_{\boldsymbol{n}}^{\gamma/2}}\right).
\end{aligned} \tag{47}$$

Note,

$$\|S_{\boldsymbol{n}}(\boldsymbol{\Theta}^0,\boldsymbol{\beta}_0) - S_{\boldsymbol{n}}^0\|_2 = O\left(\|E\{S_{\boldsymbol{n}}(\boldsymbol{\Theta}^0,\boldsymbol{\beta}_0) - S_{\boldsymbol{n}}^0\}\|_2 + \{E\|S_{\boldsymbol{n}}(\boldsymbol{\Theta}^0,\boldsymbol{\beta}_0) - S_{\boldsymbol{n}}^0\|_2^2\}^{1/2}\right). \tag{48}$$

Thus, by (46) and (47), $\|S_{\boldsymbol{n}}(\boldsymbol{\Theta}^0, \boldsymbol{\beta}_0) - S_{\boldsymbol{n}}^0\|_2 = O_p(K_{\boldsymbol{n}}^{-\gamma}\widehat{\boldsymbol{n}}^{1/2})$; furthermore, by Lemma 3 and the similar arguments as used in the proofs of Lemmas 8.4 and 8.5 in Wei and He (2006), and note $K_{\boldsymbol{n}} = o(\widehat{\boldsymbol{n}}^{1/4})$, we can verify that $\|S_{\boldsymbol{n}}(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}}) - S_{\boldsymbol{n}}(\boldsymbol{\Theta}^0, \boldsymbol{\beta}_0)\|_2 = o_p(\widehat{\boldsymbol{n}}^{1/2})$ holds. So, by the above discussion, we have

$$\begin{aligned}&\|S_{\boldsymbol{n}}(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}})\|_2\\ &\quad\le \|S_{\boldsymbol{n}}(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}}) - S_{\boldsymbol{n}}(\boldsymbol{\Theta}^0, \boldsymbol{\beta}_0)\|_2 + \|S_{\boldsymbol{n}}(\boldsymbol{\Theta}^0, \boldsymbol{\beta}_0) - S_{\boldsymbol{n}}^0\|_2 + \|S_{\boldsymbol{n}}^0\|_2\\ &\quad= O_p(\widehat{\boldsymbol{n}}^{1/2}).\end{aligned} \tag{49}$$

Furthermore, according to C7,

$$\lambda_{\boldsymbol{n}k}^*/\widehat{\boldsymbol{n}}^{1/2} \ge a_{\boldsymbol{n}2}^*/\widehat{\boldsymbol{n}}^{1/2} \to \infty, \tag{50}$$

as $\boldsymbol{n} \to \infty$. Thus, the proof is completed. □

Proof (II) of Theorem 1 Similar to the proof of (I) in Theorem 1, we can prove that

$$\left\| \sum_{\mathbb{i} \in \mathcal{I}_{\boldsymbol{n}}} \Psi(Y_{\mathbb{i}} - \boldsymbol{\Pi}_{\mathbb{i}}^T \widehat{\boldsymbol{\Theta}} - \boldsymbol{Z}_{\mathbb{i}}^T \widehat{\boldsymbol{\beta}}) Z_{\mathbb{i}k} \right\|_2 = O_p(\sqrt{\boldsymbol{n} K_{\boldsymbol{n}}}), \tag{51}$$

for $k = c+1, \ldots, q$. Furthermore, by C8, $\lambda_{\boldsymbol{n}k}^{**}/\sqrt{\boldsymbol{n} K_{\boldsymbol{n}}} \ge b_{\boldsymbol{n}}^{**}/\sqrt{\boldsymbol{n} K_{\boldsymbol{n}}} \to \infty$, for $l = c+1, \ldots, q$. So, this imply

$$\lim_{\boldsymbol{n}\to\infty} \Pr\left(\lambda_{\boldsymbol{n}k}^{**} > \left\| \sum_{\mathbb{i} \in \mathcal{I}_{\boldsymbol{n}}} \Psi(Y_{\mathbb{i}} - \boldsymbol{\Pi}_{\mathbb{i}}^T \widehat{\boldsymbol{\Theta}} - \boldsymbol{Z}_{\mathbb{i}}^T \widehat{\boldsymbol{\beta}}) Z_{\mathbb{i}k} \right\|_2 \right) = 1, k = c+1, \ldots, q. \tag{52}$$

Thus, the proof is completed. □

Proof (II) of Theorem 2 Let

$$\boldsymbol{\varsigma}^*\left(\boldsymbol{\beta}^*, \boldsymbol{\Theta}_*\right) = \begin{pmatrix} \boldsymbol{\varsigma}_1^* \\ \boldsymbol{\varsigma}_2^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Lambda}_{\boldsymbol{n}}^{*-1/2} \boldsymbol{\Xi}_{\boldsymbol{n}}^* (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0^*) \\ K_{\boldsymbol{n}}^{-1/2} \boldsymbol{H}_{\boldsymbol{n}}^* (\boldsymbol{\Theta}_* - \boldsymbol{\Theta}_*^0) + K_{\boldsymbol{n}}^{1/2} \boldsymbol{H}_{\boldsymbol{n}}^{*-1} \boldsymbol{\Pi}^{*T} \boldsymbol{B} \boldsymbol{Z}^* (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0^*) \end{pmatrix}, \tag{53}$$

where $\boldsymbol{H}_{\boldsymbol{n}}^{*2} = K_{\boldsymbol{n}} \boldsymbol{\Pi}^{*T} \boldsymbol{B} \boldsymbol{\Pi}^*$, $\boldsymbol{\Theta}_* = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_v^T)^T$ and $\boldsymbol{\Theta}_*^0 = \left(\boldsymbol{\theta}_1^{0T}, \ldots, \boldsymbol{\theta}_v^{0T}\right)^T$, $\boldsymbol{Z}^* = (\boldsymbol{Z}_{\mathbb{i}}^*, \mathbb{i} \in \mathcal{I}_{\boldsymbol{n}})^T$, $\boldsymbol{\Pi}^* = (\boldsymbol{\Pi}_{\mathbb{i}}^*, \mathbb{i} \in \mathcal{I}_{\boldsymbol{n}})^T$. By theorem 1, we know that, with probability tending to one, $\boldsymbol{\varsigma}^*(\widehat{\boldsymbol{\beta}}^*, \widehat{\boldsymbol{\Theta}}_*)$ is the minimizer of

$$Q(\boldsymbol{\varsigma}_1^*, \boldsymbol{\varsigma}_2^*) = \sum_{\mathbb{i} \in \mathcal{I}_{\boldsymbol{n}}} \rho(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^{*T} \widetilde{\boldsymbol{Z}}_{\mathbb{i}}^* - \boldsymbol{\varsigma}_2^{*T} \widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}^* - R_{\boldsymbol{n}\mathbb{i}}) + \sum_{l=1}^{v} \lambda_{nl}^* \|\boldsymbol{\theta}_l\|_r + \sum_{k=1}^{c} \lambda_{nk}^{**} |\beta_k|, \tag{54}$$

where $\widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}^{*} = K_{\boldsymbol{n}}^{1/2}\boldsymbol{H}_{\boldsymbol{n}}^{*-1}\boldsymbol{\Pi}_{\mathbb{i}}^{*}$, $\widetilde{\boldsymbol{Z}}_{\mathbb{i}}^{*} = \boldsymbol{\Lambda}_{n}^{*1/2}\boldsymbol{\Xi}_{n}^{*-1}\boldsymbol{Z}_{\mathbb{i}}^{**}$, $\boldsymbol{Z}_{\mathbb{i}}^{**}$ is defined similarly to $\widehat{\boldsymbol{Z}}_{\mathbb{i}}$.

Firstly, define $\widehat{\boldsymbol{\varsigma}}_1^{**} = \boldsymbol{\Lambda}_{\boldsymbol{n}}^{*-1/2}\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}}\boldsymbol{Z}_{i}^{**T}\Psi(\epsilon_{\mathbb{i}})$. Then, using arguments similar to those used in the proof of Lemma 6 of Tang and Cheng (2009), we have that $\widehat{\boldsymbol{\varsigma}}_1^{**}$ converges to normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}_c$. So, in order to prove this theorem, we only need to verify $\|\widehat{\boldsymbol{\varsigma}}_1^{**} - \widehat{\boldsymbol{\varsigma}}_1^{*}\|_2 = o_p(1)$, or equivalently, for any $\delta > 0$, $\Pr(\|\widehat{\boldsymbol{\varsigma}}_1^{**} - \widehat{\boldsymbol{\varsigma}}_1^{*}\|_2 < \delta) \to 1$.

By the definition of $\widehat{\boldsymbol{\varsigma}}_1^{**}$ and $\widehat{\boldsymbol{\varsigma}}_1^{*}$, there exists some constant $C > 0$, such that $\lim_{\boldsymbol{n}\to\infty}\Pr(\|\widehat{\boldsymbol{\varsigma}}_1^{**}\|_2 < C) = 1$ and $\lim_{\boldsymbol{n}\to\infty}\Pr(\|\widehat{\boldsymbol{\varsigma}}_1^{*}\|_2 < CK_{\boldsymbol{n}}^{1/2}) = 1$. Let

$$V_{\mathbb{i}}(\boldsymbol{\varsigma}_1^{*}, \widehat{\boldsymbol{\varsigma}}_1^{**}) = \rho(\epsilon_{\mathbb{i}} - \boldsymbol{\varsigma}_1^{*T}\widetilde{\boldsymbol{Z}}_{\mathbb{i}}^{*} - \widehat{\boldsymbol{\varsigma}}_2^{*T}\widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}^{*} - R_{\boldsymbol{n}\mathbb{i}}) - \rho(\epsilon_{\mathbb{i}} - \widehat{\boldsymbol{\varsigma}}_1^{**T}\widetilde{\boldsymbol{Z}}_{\mathbb{i}}^{*} - \widehat{\boldsymbol{\varsigma}}_2^{*T}\widetilde{\boldsymbol{\Pi}}_{\mathbb{i}}^{*} - R_{\boldsymbol{n}\mathbb{i}}), \tag{55}$$

thus by the convexity of loss function $\rho(\cdot)$, it is sufficient to prove that

$$\Pr\left\{\inf_{\|\boldsymbol{\varsigma}_1^{*} - \widehat{\boldsymbol{\varsigma}}_1^{**}\|_2 = \delta}\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} V_{\mathbb{i}}(\boldsymbol{\varsigma}_1^{*}, \widehat{\boldsymbol{\varsigma}}_1^{**}) > 0\right\} \to 1. \tag{56}$$

By arguments similar to those used in the proof of Theorem 3.1 of Tang (2014), for any given $\delta > 0$,

$$\sup_{\|\boldsymbol{\varsigma}_1^{*} - \widehat{\boldsymbol{\varsigma}}_1^{**}\|_2 \le \delta}\left|\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}}\{V_{\mathbb{i}}(\boldsymbol{\varsigma}_1^{*}, \widehat{\boldsymbol{\varsigma}}_1^{**}) - (\boldsymbol{\varsigma}_1^{*} - \widehat{\boldsymbol{\varsigma}}_1^{**})^{T}\widetilde{\boldsymbol{Z}}_{\mathbb{i}}^{*}\Psi(\epsilon_{\mathbb{i}}) - EV_{\mathbb{i}}(\boldsymbol{\varsigma}_1^{*}, \widehat{\boldsymbol{\varsigma}}_1^{**})\}\right| = o_p(1), \tag{57}$$

$$\sup_{\|\boldsymbol{\varsigma}_1^{*} - \widehat{\boldsymbol{\varsigma}}_1^{**}\|_2 \le \delta}\left|\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} EV_{\mathbb{i}}(\boldsymbol{\varsigma}_1^{*}, \widehat{\boldsymbol{\varsigma}}_1^{**}) - \frac{1}{2}(\boldsymbol{\varsigma}_1^{*\top}\boldsymbol{\Lambda}_{\boldsymbol{n}}^{*\frac{1}{2}}\boldsymbol{\Xi}_{\boldsymbol{n}}^{*-1}\boldsymbol{\Lambda}_{\boldsymbol{n}}^{*\frac{1}{2}}\boldsymbol{\varsigma}_1^{*} - \widehat{\boldsymbol{\varsigma}}_1^{**\top}\boldsymbol{\Lambda}_{\boldsymbol{n}}^{*\frac{1}{2}}\boldsymbol{\Xi}_{\boldsymbol{n}}^{*-1}\boldsymbol{\Lambda}_{\boldsymbol{n}}^{*\frac{1}{2}}\widehat{\boldsymbol{\varsigma}}_1^{**})\right| = o_p(1). \tag{58}$$

Combining (57) and (58) yields

$$\sup_{\|\boldsymbol{\varsigma}_1^{*} - \widehat{\boldsymbol{\varsigma}}_1^{**}\|_2 \le \delta}\left|\sum_{\mathbb{i}\in\mathcal{I}_{\boldsymbol{n}}} V_{\mathbb{i}}(\boldsymbol{\varsigma}_1^{*}, \widehat{\boldsymbol{\varsigma}}_1^{**}) - \frac{1}{2}(\boldsymbol{\varsigma}_1^{*} - \widehat{\boldsymbol{\varsigma}}_1^{**})^{\top}\boldsymbol{\Lambda}_{\boldsymbol{n}}^{*\frac{1}{2}}\boldsymbol{\Xi}_{\boldsymbol{n}}^{*-1}\boldsymbol{\Lambda}_{\boldsymbol{n}}^{*\frac{1}{2}}(\boldsymbol{\varsigma}_1^{*} - \widehat{\boldsymbol{\varsigma}}_1^{**})\right| = o_p(1). \tag{59}$$

What is more, for constant $C > 0$, when $\|\boldsymbol{\varsigma}_1^{*} - \widehat{\boldsymbol{\varsigma}}_1^{**}\|_2 \le CK_{\boldsymbol{n}}^{1/2}$, $\sum_{k=1}^{c}\lambda_{\boldsymbol{n}k}^{**}\{|\beta_k| - |\widehat{\beta}_k^{**}|\} \ge -a_{\boldsymbol{n}}^{**}\|\boldsymbol{\Xi}_{\boldsymbol{n}}^{*-1}\boldsymbol{\Lambda}_{\boldsymbol{n}}^{*\frac{1}{2}}(\boldsymbol{\varsigma}_1^{*} - \widehat{\boldsymbol{\varsigma}}_1^{**})\|_1 = o_p(1)$, where $\widehat{\boldsymbol{\beta}}^{**} = \boldsymbol{\Xi}_{\boldsymbol{n}}^{*-1}\boldsymbol{\Lambda}_{\boldsymbol{n}}^{*\frac{1}{2}}\widehat{\boldsymbol{\varsigma}}_1^{**} + \boldsymbol{\beta}_0^{*}$. So, based on the above discussion, as $\boldsymbol{n}$ tends to infinite, (56) holds. Thus, for any $\delta > 0$, $\lim_{\boldsymbol{n}\to\infty}\Pr(\|\widehat{\boldsymbol{\varsigma}}_1^{**} - \widehat{\boldsymbol{\varsigma}}_1^{*}\|_2 > \delta) = 0$. The proof is completed. □

References

Chu, T., Zhu, J., Wang, H. (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *The Annals of Statistics*, *39*, 2607–2625.

Eggleston, H. (1958). *Convexity (Cambridge Tracts in Mathematics)*. Cambridge: Cambridge University Press.
Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.
Fan, J., Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, *99*, 710–723.
Fan, J., Yao, Q. (2003). *Nonlinear time series*: *nonparametric and parametric methods*. New York: Springer.
Fu, W. (1998). Penalized regression: the bridge versus the LASSO. *Journal of Computational and Graphical Statistics*, *7*, 397–416.
Gao, J., Lu, Z., Tjtheim, D. (2006). Estimation in semiparametric spatial regression. *The Annals of Statistics*, *34*, 1395–1435.
Hallin, M., Lu, Z., Tran, L. (2004). Local linear spatial regression. *The Annals of Statistics*, *32*, 2469–2500.
Hallin, M., Lu, Z., Yu, K. (2009). Local linear spatial quantile regression. *Bernoulli*, *15*, 659–686.
Harrison, D., Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*, 81–102.
Hoeting, J., Davis, R., Merton, A., Thompson, S. (2006). Model selection for geostatistical models. *Ecological Applications*, *16*, 87–98.
Hohsuk, N., Kwanghun, C., Ingrid, V. (2012). Variable selection of varying coefficient models in quantile regression. *Electronic Journal of Statistics*, *6*, 1220–1238.
Huang, J., Wu, C., Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, *89*, 111–128.
Huang, J., Horowitz, J., Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, *38*, 2282–2313.
Huang, J., Breheny, P., Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, *27*, 481–499.
Huber, P. (1981). *Robust estimation*. New York: Wiley.
Hunter, D., Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics*, *33*, 1617–1642.
Kai, B., Li, R., Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics*, *39*, 305–332.
Lee, Y., Choi, H., Park, B., Yu, K. (2004). Local likelihood density estimation on random fields. *Statistics and Probability Letters*, *68*, 347–357.
Lu, Z., Tjøstheim, D. (2014). Nonparametric estimation of probability density functions for irregularly observed spatial data. *Journal of the American Statistical Association*, *109*, 1546–1564.
Lu, Z., Lundervold, A., Tjøstheim, D., Yao, Q. (2007). Exploring spatial nonlinearity using additive approximation. *Bernoulli*, *13*, 447–472.
Lu, Z., Tang, Q., Cheng, L. (2014). Estimating spatial quantile regression with functional coefficients: a robust semiparametric framework. *Bernoulli*, *20*, 164–189.
Pace, R., Gilley, O. (1997). Using the spatial configuration of the data to improve estimation. *Journal of Real Estate Finance and Economics*, *14*, 333–340.
Schumaker, L. (1981). *Spline functions*: *basic theory*. New York: Wiley.
Tang, Q. (2014). Robust estimation for functional coefficient regression models with spatial data. *Statistics*, *48*, 388–404.
Tang, Q., Cheng, L. (2009). B-spline estimation for varying coefficient regression with spatial data. *Science in China Series A*: *Mathematics*, *52*, 2321–2340.
Tang, Y., Wang, H., Zhu, Z. (2013). Variable selection in quantile varying coefficient models with longitudinal data. *Computational Statistics and Data Analysis*, *57*, 435–449.
Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*: *Series B*, *58*, 267–288.
Tran, L. (1990). Kernel density estimation on random field. *Journal of Multivariate Analysis*, *34*, 37–53.
Tran, L. (1993). Nearest neighbor estimators for random fields. *Journal of Multivariate Analysis*, *44*, 23–46.
Wang, H., Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, *104*, 747–757.
Wang, H., Zhu, J. (2009). Variable selection in spatial regression via penalized least squares. *The Canadian Journal of Statistics*, *37*, 607–624.
Wang, K., Lin, L. (2014). Variable selection in robust semiparametric modeling for longitudinal data. *Journal of the Korean Statistical Society*, *43*, 303–314.

Wang, L., Li, H., Huang, J. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, *103*, 1556–1569.
Wang, H., Zhu, Z., Zhou, J. (2009). Quantile regression in partially linear varying coe.cient models. *The Annals of Statistics*, *37*, 3841–3866.
Wang, L., Xue, L., Qu, A., Liang, H. (2014). Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *The Annals of Statistics*, *42*, 592–624.
Wei, Y., He, X. (2006). Conditional growth charts (with discussion). *The Annals of Statistics*, *19*, 801–817.
Xue, L. (2009). Variable selection in additive models. *Statistica Sinica*, *19*, 1281–1296.
Zhao, W., Zhang, R., Liu, J., Lv, Y. (2014). Robust and efficient variable selection for semiparametric partially linear varying coefficient model based on modal regression. *Annals of the Institute of Statistical Mathematics*, *66*, 165–191.
Zhu, J., Huang, H., Reyes, P. (2010). On selection of spatial linear models for lattice data. *Journal of the Royal Statistical Society*: *Series B*, *72*, 389–402.
Zou, H., Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, *36*, 1509–1533.