Metrika (2015) 78:85–118
DOI 10.1007/s00184-014-0491-y

Robust spline-based variable selection in varying coefficient model

Long Feng · Changliang Zou · Zhaojun Wang · Xianwu Wei · Bin Chen

Received: 3 August 2013 / Published online: 14 May 2014

Abstract The varying coefficient model is widely used as an extension of the linear regression model. Many procedures have been developed for the model estimation, and recently efficient variable selection procedures for the varying coefficient model have been proposed as well. However, those variable selection approaches are mainly built on the least-squares (LS) type method. Although the LS method is a successful and standard choice in the varying coefficient model fitting and variable selection, it may suffer when the errors follow a heavy-tailed distribution or in the presence of outliers. To overcome this issue, we start by developing a novel robust estimator, termed rank-based spline estimator, which combines the ideas of rank inference and polynomial spline. Furthermore, we propose a robust variable selection method, incorporating the smoothly clipped absolute deviation penalty into the rank-based spline loss function. Under mild conditions, we theoretically show that the proposed rank-based spline estimator is highly efficient across a wide spectrum of distributions. Its asymptotic relative efficiency with respect to the LS-based method is closely related to that of the signed-rank Wilcoxon test with respect to the *t* test. Moreover, the proposed variable selection method can identify the true model consistently, and the resulting estimator can be as efficient as the oracle estimator. Simulation studies show that our

Electronic supplementary material The online version of this article (doi:10.1007/s00184-014-0491-y) contains supplementary material, which is available to authorized users.

L. Feng · C. Zou (✉) · Z. Wang (✉) · X. Wei
LPMC and Institute of Statistics, Nankai University, Tianjin 300071, China
e-mail: nk.chlzou@gmail.com

Z. Wang
e-mail: zjwang@nankai.edu.cn

B. Chen
School of Mathematics and Statistics, Jiangsu Normal University, Xuzhou, Jiangsu, China

procedure has better performance than the LS-based method when the errors deviate from normality.

Keywords KLASSO · Oracle property · Polynomial spline · Rank regression · Robust estimation · Robust model selection · SCAD

1 Introduction

Consider the varying coefficient model

$$Y = \boldsymbol{X}^T(U)\boldsymbol{\beta}(U) + \varepsilon, \tag{1}$$

where Y is the response variable, U and $\boldsymbol{X}$ are the covariates, and $\boldsymbol{\beta}(U)$ are some unknown smooth functions. The random error ε is independent of $\boldsymbol{X}$ and U, and has probability density function $h(\cdot)$ which has finite Fisher information. In this paper, it is assumed that U is a scalar and $\boldsymbol{X}$ is a p-dimensional vector which may depend on U. Since introduced by Hastie and Tibshirani (1993), the varying coefficient model has been widely applied in many scientific areas, such as economics, finance, politics, epidemiology, medical science, ecology, and so on.

Due to its flexibility and interpretability, in the past ten years, it has experienced rapid developments in both theory and methodology; see Fan and Zhang (2008) for a comprehensive survey. In general, there are at least three common ways to estimate this model. One is the kernel-based local polynomial smoothing, see for instance, Wu et al. (1998), Hoover et al. (1998), Fan and Zhang (1999), Kauermann and Tutz (1999); One is the polynomial spline, see Huang et al. (2002, 2004) and Huang and Shen (2004); The last one is the smoothing spline, see Hastie and Tibshirani (1993), Hoover et al. (1998) and Chiang et al. (2001). Recently, efficient variable selection procedures for the varying coefficient model have been proposed as well. In a typical 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 problem. In a traditional linear regression setting, many selection criteria, e.g., Akaike information criterion (AIC) and Bayesian information criterion (BIC) have been extensively used in practice. Recently, various shrinkage methods have been developed, which include but are not limited to the least absolute shrinkage and selection operator (LASSO; c.f., Tibshirani 1996; Zou 2006) and the smoothly clipped absolute deviation (SCAD; Fan and Li 2001). These regularized estimation procedures were developed for varying coefficient models. Among others, Lin and Zhang (2006) develop COSSO for component selection and smoothing in smoothing spline ANOVA. Wang et al. (2007) propose to use group SCAD method for varying-coefficient model selection. Wang et al. (2008) extend the application of the SCAD penalty to varying coefficient models with longitudinal data. Li and Liang (2008) study variable selection for partially linear varying coefficient models, where the parametric components are identified via the SCAD but the nonparametric components are selected via a generalized likelihood ratio test, instead of a shrinkage method. Leng (2009) proposes

a penalized likelihood method in the framework of the smoothing spline ANOVA models. Wang and Xia (2009) develop a shrinkage method, called KLASSO (Kernel-based LASSO), which combines the ideas of the local polynomial smoothing and LASSO. Tang et al. (2012) develop a unified variable selection approach for both least squares regression and quantile regression models with possibly varying coefficients. Their method is carried out by using a two-step iterative procedure based on basis expansion and an adaptive-LASSO-type penalty.

The estimation and variable selection procedures in the aforementioned papers are mainly built on least-squares (LS) type methods. Although the LS methods are successful and standard choice in varying coefficient model fitting, they may suffer when the errors follow a heavy-tailed distribution or in the presence of outliers. Thus, some efforts have been devoted to construct robust estimators for the varying coefficient models. Kim (2007) develops a quantile regression procedure for varying coefficient models when the random errors are assumed to have a certain quantile equal to zero. Wang et al. (2009) recently develop a local rank estimation procedure, which integrates the rank regression (Hettmansperger and McKean 1998) and local polynomial smoothing. In traditional linear regression settings, some also draw much attention to robust variable selection. Wang et al. (2007) propose a LASSO-based procedure using the least absolute deviation regression. Zou and Yuan (2008) propose the composite quantile regression (CQR) estimator by averaging K quantile regressions. They have shown that CQR is selection consistent and can be more robust in various circumstances. Wang and Li (2009) and Leng (2010) independently propose two efficient shrinkage estimators, using the idea of rank regression. However, to the best of our knowledge, there has hitherto been no existing appropriate robust variable selection procedure for the varying coefficient model, which is the focus of this paper.

In this paper, we aim to propose an efficient robust variable selection method for varying coefficient models. Motivated by the local rank inference (Wang et al. 2009), we start by developing a robust rank-based spline estimator. Under some mild conditions, we establish the asymptotic representation of the proposed estimator and further prove its asymptotic normality. We derive the formula of the asymptotic relative efficiency (ARE) of the rank-based spline estimator relative to the LS-based estimator, which has an expression that is closely related to that of the signed-rank Wilcoxon test in comparison with the t test. Further, we extend the application of the SCAD penalty to the rank-based spline estimator. Theoretical analysis reveals that our procedure is consistent in variable selection; that is, the probability that it correctly selects the true model tends to one. Also, we show that our procedure has the so-called oracle property; that is, the asymptotic distribution of an estimated coefficient function is the same as that when it is known a priori which variables are in the model. Simulation studies show that our procedure has better performance than KLASSO (Wang and Xia 2009) and LSSCAD (Wang et al. 2008) when the errors deviate from normality. Even in the most favorable case for KLASSO and LSSCAD, i.e., normal distribution, our procedure does not lose much, which coincides with our theoretical analysis.

This article is organized as follows. Section 2 presents the rank-based spline procedure for estimating the varying coefficient model, and some theoretical properties are provided. In Sect. 3, with the help of the rank-based spline procedure, we propose a new robust variable selection method and study its theoretical properties. Its

numerical performance is investigated in Sect. 4. Several remarks draw the paper to its conclusion in Sect. 5. The technical details are provided in the "Appendix". Some other simulation results are provided in another appendix, which is available online as supplementary material.

2 Methodology

To develop an efficient scheme for variable selection, we choose to consider a polynomial spline smoothing method rather than a local polynomial smoother. The reason is that using the former the varying coefficient model can be re-formulated as a traditional multiple regression model and thus it serves the variable selection purpose more naturally (Wang et al. 2008). In contrast, although works also very well, using local polynomial smoothers requires more sophisticated approximation and techniques in the selection procedures and proofs of oracle properties (Wang and Xia 2009). Therefore, in this section, we develop a rank-based spline method for estimating $\boldsymbol{\beta}(\cdot)$, which can be regarded as a parallel to the local rank estimator proposed by Wang et al. (2009).

2.1 The estimation procedure

Suppose that $\{U_i, \boldsymbol{X}_i, Y_i\}_{i=1}^n$ is a random sample from the model (1). Write $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ and $\boldsymbol{\beta}(U) = (\beta_1(U), \ldots, \beta_p(U))^T$. Suppose that each $\beta_l(U), l = 1, \ldots, p$, can be approximated by some spline functions, that is

$$\beta_l(U) \approx \sum_{k=1}^{K_l} \gamma_{lk} B_{lk}(U), \tag{2}$$

where each $\{B_{lk}(\cdot), k = 1, \ldots, K_l\}$ is a basis for a linear space $\mathbb{G}_l$ of spline functions with a fixed degree and knot sequence. In our applications we use the B-spline basis for its good numerical properties. Following (1) and (2), we have

$$Y_i \approx \sum_{l=1}^{p} \sum_{k=1}^{K_l} X_{il} B_{lk}(U_i) \gamma_{lk} + \varepsilon_i.$$

Define $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T, \mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T, \boldsymbol{\gamma}_l = (\gamma_{l1}, \ldots, \gamma_{lK_l})^T, \boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_p^T)^T$,

$$\mathbf{B}(u) = \begin{pmatrix} B_{11}(u) & \cdots & B_{1K_1}(u) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & B_{p1}(u) & \cdots & B_{pK_p}(u) \end{pmatrix},$$

$\boldsymbol{Z}_i = \boldsymbol{X}_i^T \mathbf{B}(U_i)$, and $\mathbf{Z} = (\boldsymbol{Z}_1^T, \ldots, \boldsymbol{Z}_n^T)^T$. Based on the above approximation, we obtain the residual at U_i

$$e_i = Y_i - \boldsymbol{Z}_i \boldsymbol{\gamma}.$$

Motivated by the rank regression (Jaeckel 1972; Hettmansperger and McKean 1998), we define the rank-based spline objective (loss) function

$$Q_n(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i<j}|e_i - e_j|. \tag{3}$$

An estimator of $\beta_l(u)$ is obtained by $\hat{\beta}_l(u) = \sum_{k=1}^{K_l}\hat{\gamma}_{lk}B_{lk}(u)$, where the $\hat{\gamma}_{lk}$'s are the minimizers of (3). We term it as rank-based spline estimator because the objective (loss) function is equivalent to the classic rank loss function in linear models based on Wilcoxon scores (Hettmansperger and McKean 1998).

2.2 Asymptotic properties

In this subsection, we establish the asymptotic properties of the rank-based spline estimator. The main challenge comes from the nonsmoothness of the objective function $Q_n(\boldsymbol{\gamma})$. To overcome this difficulty, we first derive an asymptotic representation of $\hat{\boldsymbol{\gamma}}$ via a quadratic approximation of $Q_n(\boldsymbol{\gamma})$, which holds uniformly in a local neighborhood of the true parameter values. Throughout this manuscript, we will use the following notation for ease of exposition. Let $|\boldsymbol{a}|$ denote the Euclidean norm of a real valued vector $\boldsymbol{a}$. For a real-valued function g, $||g||_\infty = \sup_u |g(u)|$. For a vector-valued function $\boldsymbol{g} = (g_1, \ldots, g_p)^T$, denote $||\boldsymbol{g}||_{L_2} = \sum_{1\le l\le p}||g_l||_{L_2}^2$ and $||\boldsymbol{g}||_\infty = \max_l ||g_l||_\infty$. Define $K_n = \max_{1\le l\le p} K_l$, $\rho_n = \max_{1\le l\le p}\inf_{g\in\mathbb{G}_l}||\beta_l - g||_\infty$. Let $\boldsymbol{g}^* = (g_1^*, \ldots, g_p^*) \in \mathbb{G}$ be such that $||\boldsymbol{g}^* - \boldsymbol{\beta}||_\infty = \rho_n$, where $\mathbb{G} = \mathbb{G}_1 \times \cdots \times \mathbb{G}_p$ and $\boldsymbol{\beta}$ is the real varying-coefficient function. Then there exists $\boldsymbol{\gamma}_0$, such that $\boldsymbol{g}^* = \mathbf{B}(u)\boldsymbol{\gamma}_0$.

Define $\theta_n = \sqrt{K_n/n}$, $\boldsymbol{\gamma}^* = \theta_n^{-1}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)$, and $\Delta_i = \boldsymbol{X}_i^T\boldsymbol{\beta}(U_i) - \boldsymbol{Z}_i\boldsymbol{\gamma}_0$. Let $\widehat{\boldsymbol{\gamma}^*}$ be the value of $\boldsymbol{\gamma}^*$ that minimizes the following reparametrized function

$$Q_n^*(\boldsymbol{\gamma}^*) = \frac{1}{n}\sum_{i<j}|\varepsilon_i + \Delta_i - \varepsilon_j - \Delta_j - \theta_n(\boldsymbol{Z}_i - \boldsymbol{Z}_j)\boldsymbol{\gamma}^*|.$$

Then it can be easily seen that

$$\widehat{\boldsymbol{\gamma}^*} = \theta_n^{-1}\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\right).$$

We use $\boldsymbol{S}_n(\boldsymbol{\gamma}^*)$ to denote the gradient function of $Q_n^*(\boldsymbol{\gamma}^*)$. More specifically,

$$\boldsymbol{S}_n(\boldsymbol{\gamma}^*) = -\frac{\theta_n}{n}\sum_{i<j}\left(\boldsymbol{Z}_i^T - \boldsymbol{Z}_j^T\right)\operatorname{sgn}\left(\varepsilon_i + \Delta_i - \varepsilon_j - \Delta_j - \theta_n\left(\boldsymbol{Z}_i - \boldsymbol{Z}_j\right)\boldsymbol{\gamma}^*\right),$$

where $\operatorname{sgn}(\cdot)$ is the sign function. Furthermore, we consider the following quadratic function of $\boldsymbol{\gamma}^*$

$$A_n(\boldsymbol{\gamma}^*) = \tau\theta_n^2\boldsymbol{\gamma}^{*T}\mathbf{Z}^T\mathbf{Z}\boldsymbol{\gamma}^* + \boldsymbol{\gamma}^{*T}\boldsymbol{S}_n(\mathbf{0}) + Q_n^*(\mathbf{0}),$$

where $\tau = \int h^2(t)dt$ is the well-known Wilcoxon constant, and $h(\cdot)$ is the density function of the random error ε.

For the asymptotic analysis, we need the following regularity conditions.

(C1) The distribution of U_i has a Lebesgue density $f(u)$ which is bounded away from 0 and infinity.

(C2) $E(\boldsymbol{X}_i(u)|u = U_i) = \mathbf{0}$, and the eigenvalues $\lambda_1(u) \leq \cdots \leq \lambda_p(u)$ of $\Sigma(u) = E[\boldsymbol{X}_i(u)\boldsymbol{X}_i(u)^T]$ are bounded away from 0 and infinity uniformly; that is, there are positive constants W_1 and W_2 such that $W_1 \leq \lambda_1(u) \leq \cdots \leq \lambda_p(u) \leq W_2$ for all u.

(C3) There is a positive constant M_1 such that $|X_{il}(u)| \leq M_1$ for all u and $l = 1, \ldots, p$, $i = 1, \ldots, n$.

(C4) $\limsup_n(\max_l K_l / \min_l K_l) < \infty$.

(C5) The error ε has finite Fisher information, i.e., $\int [h'(x)]^2/h(x)dx < \infty$.

Remark 1 Conditions (C1)–(C4) are the same as those in Huang et al. (2004). The assumption on the random errors in (C5) is a standard condition for rank analysis in multiple linear regression (Hettmansperger and McKean 1998). These conditions are mild and can be satisfied in many practical situations.

Lemma 1 *Suppose Conditions (C1)–(C5) all hold, then for any* $\epsilon > 0$ *and* $c > 0$,

$$P\left(\sup_{\sqrt{1/K_n}|\boldsymbol{\gamma}^*|\leq c} |Q_n^*(\boldsymbol{\gamma}^*) - A_n(\boldsymbol{\gamma}^*)| \geq \epsilon\right) \to 0.$$

Lemma 1 implies that the nonsmooth objective function $Q_n^*(\boldsymbol{\gamma}^*)$ can be uniformly approximated by a quadratic function $A_n(\boldsymbol{\gamma}^*)$ in a neighborhood of $\mathbf{0}$. It is also shown that the minimizer of $A_n(\boldsymbol{\gamma}^*)$ is asymptotic within $o(\sqrt{K_n})$ neighborhood of $\widehat{\boldsymbol{\gamma}^*}$, say $|\widehat{\boldsymbol{\gamma}^*} - (2\tau)^{-1}(\theta_n^2\mathbf{Z}^T\mathbf{Z})^{-1}\boldsymbol{S}_n(\mathbf{0})| = o_p(\sqrt{K_n})$ (see "Appendix"). This further allows us to derive the asymptotic distribution.

Let $\breve{\beta}_l(u) = E[\hat{\beta}_l(u) \mid \mathscr{X}]$ be the mean of $\hat{\beta}_l(u)$ conditioning on $\mathscr{X} = \{(\boldsymbol{X}_i, U_i)\}_{i=1}^n$. It is useful to consider the decomposition $\hat{\beta}_l(u) - \beta_l(u) = \hat{\beta}_l(u) - \breve{\beta}_l(u) + \breve{\beta}_l(u) - \beta_l(u)$, where $\hat{\beta}_l(u) - \breve{\beta}_l(u)$ and $\breve{\beta}_l(u) - \beta_l(u)$ contribute to the variance and bias terms, respectively. Denote $\breve{\boldsymbol{\beta}}(u) = (\breve{\beta}_1(u), \ldots, \breve{\beta}_p(u))$. The following two theorems establish the consistency and asymptotic normality of the rank-based spline estimator, respectively.

Theorem 1 *Suppose Conditions (C1)–(C5) hold. If* $K_n \log K_n/n \to 0$, *then* $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_{L_2}^2 = O_p(\rho_n^2 + K_n/n)$; *consequently, if* $\rho_n \to 0$, *then* $\hat{\beta}_l, l = 1, \ldots, p$ *are consistent.*

Theorem 2 *Suppose Conditions (C1)–(C5) hold. If* $K_n \log K_n/n \to 0$, *then*

$$\left\{\text{var}[\hat{\boldsymbol{\beta}}(u)]\right\}^{-1/2}\left(\hat{\boldsymbol{\beta}}(u) - \breve{\boldsymbol{\beta}}(u)\right) \xrightarrow{d} N\left(0, \mathbf{I}_p\right).$$

The above two theorems are parallel to those in Huang et al. (2004). Theorem 1 implies that the magnitude of the bias is bounded in probability by the best approximation rates

by the spaces $\mathbb{G}_l$. Theorem 2 provides the asymptotic normality and can thus be used to construct confidence intervals.

Next, we study the ARE of the rank-based spline estimator with respect to the polynomial spline estimator [denoted by $\hat{\boldsymbol{\beta}}_P(u)$]) for estimating $\boldsymbol{\beta}(u)$ in the varying coefficient model, say ARE$(\hat{\boldsymbol{\beta}}(u), \hat{\boldsymbol{\beta}}_P(u))$. Unlike the ARE study in Wang et al. (2009) in which the theoretical optimal bandwidth of local polynomial estimators is used, it seems difficult to plug in theoretical optimal K_l's in evaluating ARE$(\hat{\boldsymbol{\beta}}(u), \hat{\boldsymbol{\beta}}_P(u))$ because the closed-form optimal K_l's are not available. Thus, in the following analysis, we consider a common choice of the smoothing parameters for both two spline estimators $\hat{\boldsymbol{\beta}}(u)$ and $\hat{\boldsymbol{\beta}}_P(u)$.

According to Huang et al. (2004), we know that

$$\text{var}\left[\hat{\boldsymbol{\beta}}_P(u) \mid \mathscr{X}\right] = \sigma^2 \mathbf{B}(u)\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{B}^T(u),$$

where σ^2 is the variance of ε. Now, we give the conditioned variance of $\hat{\boldsymbol{\beta}}(u)$. From the Proof of Theorem 1 shown in the Appendix, we have

$$\begin{aligned}\text{var}\left[\hat{\boldsymbol{\beta}}(u) \mid \mathscr{X}\right] &= \mathbf{B}(u)\text{var}(\hat{\boldsymbol{\gamma}})\mathbf{B}^T(u)\\ &= \frac{1}{4\tau^2}\theta_n^2\mathbf{B}(u)(\theta_n^2\mathbf{Z}^T\mathbf{Z})^{-1}\text{var}\left[\boldsymbol{S}_n(\mathbf{0})|\mathscr{X}\right]\left(\theta_n^2\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{B}(u),\end{aligned}$$

where

$$\begin{aligned}\text{var}[\boldsymbol{S}_n(\mathbf{0}) \mid \mathscr{X}] &= n^{-2}\theta_n^2 E\left\{\sum_{i<j}(\mathbf{Z}_i^T - \mathbf{Z}_j^T)\text{sgn}\left((Y_i - Y_j) - (\mathbf{Z}_i - \mathbf{Z}_j)\boldsymbol{\gamma}_0\right)\right\}^2\\ &= n^{-2}\theta_n^2\left\{\sum_{i<j}(\mathbf{Z}_i^T - \mathbf{Z}_j^T)\right\}^2 E\left(2H(\varepsilon) - 1\right)^2 + o(1)\\ &= \frac{1}{3}\theta_n^2\mathbf{Z}^T\mathbf{Z} + o(1),\end{aligned}$$

and $H(\cdot)$ denotes the distribution of ε and the second equation follows from the independence of ε and $\boldsymbol{X}, U$. Thus,

$$\text{var}\left[\hat{\boldsymbol{\beta}}(u) \mid \mathscr{X}\right] = \frac{1}{12\tau^2}\mathbf{B}(u)\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{B}^T(u) + o(1).$$

It immediately follows from Theorem 2 that the ARE of $\hat{\boldsymbol{\beta}}(u)$ with respect to $\hat{\boldsymbol{\beta}}_P(u)$ is

$$\text{ARE}\left(\hat{\boldsymbol{\beta}}(u), \hat{\boldsymbol{\beta}}_P(u)\right) = 12\sigma^2\tau^2.$$

Remark 2 This asymptotic relative efficiency is the same as that of the signed-rank Wilcoxon test with respect to the t test. It is well known in the literature of rank analysis that the ARE is as high as 0.955 for normal error distribution, and can be significantly higher than one for many heavier-tailed distributions. For instance, this quantity is 1.5 for the double exponential distribution, and 1.9 for the t distribution with three degrees of freedom. For symmetric error distributions with finite Fisher information, this asymptotic relative efficiency is known to have a lower bound equal to 0.864.

2.3 Automatic selection of smoothing parameters

Smoothing parameters, such as the degrees of splines, the numbers and locations of knots, play an important role in nonparametric models. However, due to the computational complexity, automatically selecting those three smoothing parameters is difficult in practice. In this paper, we select only $D = D_l$, the numbers of knots for $\beta_l(\cdot)$'s, using the data. The location of knots are equally spaced and the degrees of splines are fixed. We use "leave-one-out" cross-validation to choose D. To be more specific, let $\hat{\boldsymbol{\beta}}^{(i)}(u)$ be the spline estimator obtained by deleting the i-th sample. The cross-validation procedure minimizes the target function

$$CV(D) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}^{(i)}(U_i)\right)^2.$$

In practice, some other criteria, such as the GCV, fivefold CV, BIC and AIC can also be used. Our simulation studies show that those procedures are also quite effective but the variable selection results are hardly affected by the choice of selection procedure for D_l. Moreover, in this paper we restrict our attention to the spline with $d = 3$ degrees. This works well for the applications we considered. It might be worthwhile to investigate using the data to decide the knot positions (free-knot splines), which merits definitely some future research. Also, we may not use the same number of knots and degree of splines for each coefficient function because each coefficient function may have different features.

3 Rank-based variable selection and estimation

In this section, in order to conduct variable selection for the varying coefficient model in a computationally efficient manner, we incorporate the SCAD penalty function into the objective function (3) to implement nonparametric estimation and variable selection simultaneously.

3.1 The SCAD-penalty method

Now, suppose that some variables are not relevant in the regression model, so that the corresponding coefficient functions are zero functions. Let $\mathbf{R}_k = (r_{ij})_{K_k\times K_k}$ be

a matrix with entries $r_{ij} = \int B_{ki}(t)B_{kj}(t)dt$. Then, we define $||\boldsymbol{\gamma}_k||^2_{R_k} \equiv \boldsymbol{\gamma}_k^T \mathbf{R}_k \boldsymbol{\gamma}_k$. The penalized rank-based loss function is then defined as

$$PL_n(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i<j}|e_i - e_j| + n\sum_{k=1}^{p} p_{\lambda_n}\left(||\boldsymbol{\gamma}_k||_{R_k}\right), \tag{4}$$

where λ_n is the tuning parameter and $p_\lambda(\cdot)$ is chosen as the SCAD penalty function of Fan and Li (2001), defined as

$$p_\lambda(u) = \begin{cases} \lambda u, & 0 \le u \le \lambda, \\ -\frac{u^2 - 2a\lambda u + \lambda^2}{2(a-1)}, & \lambda < u < a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & u \ge a\lambda. \end{cases}$$

where a is another tuning parameter. Here we adopted $a = 3.7$ as suggested by Fan and Li (2001). This penalized loss function takes a similar form to that of Wang et al. (2008) except that the rank-based loss function is used instead of LS-based functions. An estimator of $\beta_l(u)$ is obtained by $\bar{\beta}_l(u) = \sum_{k=1}^{K_l} \bar{\gamma}_{lk} B_{lk}(u)$, where the $\bar{\gamma}_{lk}$ are minimizers of (4). In practice, one can also use the adaptive LASSO penalty to replace SCAD in (4) and we can expect that the resulting procedure will have similar asymptotic properties and comparable finite-sample performance (Zou 2006).

3.2 Computational algorithm

Because of nondifferentiability of the penalized loss (4), the commonly used gradient-based optimization method is not applicable here. In this section we develop an iterative algorithm using local quadratic approximation of the rank-based objective function $\sum_{i<j}|e_i - e_j|$ and the nonconvex penalty function $p_{\lambda_n}(||\boldsymbol{\gamma}_k||_{R_k})$. Denote that $R(e_i)$ is the rank of e_i among $\{e_i\}_{i=1}^n$. Following Sievers and Abebe (2004), the objective function is approximated by

$$\sum_{i<j}|e_i - e_j| \approx \sum_{i=1}^{n} w_i\,(e_i - \zeta)^2 \tag{5}$$

where ζ is the median of $\{e_i\}_{i=1}^n$ and

$$w_i = \begin{cases} \frac{\frac{R(e_i)}{n+1} - \frac{1}{2}}{e_i - \zeta}, & e_i \ne \zeta, \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, following Fan and Li (2001), in the neighborhood of a given positive $u_0 \in R^+$,

$$p_\lambda(u) \approx p_\lambda(u_0) + \frac{1}{2}\left[p'_\lambda(u_0)/u_0\right]\left(u^2 - u_0^2\right).$$

Then, given an initial value, $\boldsymbol{\gamma}_k^{(0)}$, with $||\boldsymbol{\gamma}_k^{(0)}||_{R_k} > 0$, the corresponding weights $w_i^{(0)}$ and the median of residuals, $\zeta^{(0)}$, can be obtained. Consequently, the penalized loss function (4) can be approximated by a quadratic form

$$PL_n(\boldsymbol{\gamma}) \approx \frac{1}{n}\sum_{i=1}^{n} w_i^{(0)}(e_i - \zeta^{(0)})^2 + p_{\lambda_n}(||\boldsymbol{\gamma}_k^{(0)}||_{R_k}) + \frac{1}{2}\left\{\frac{p'_{\lambda_n}(||\boldsymbol{\gamma}_k^{(0)}||_{R_k})}{||\boldsymbol{\gamma}_k^{(0)}||_{R_k}}\right\}\left\{\boldsymbol{\gamma}_k^T\mathbf{R}_k\boldsymbol{\gamma}_k - (\boldsymbol{\gamma}_k^{(0)})^T\mathbf{R}_k\boldsymbol{\gamma}_k^{(0)}\right\}.$$

Consequently, removing an irrelevant constant, the above quadratic form becomes

$$PL_n(\boldsymbol{\gamma}) \approx \frac{1}{n}\left(\boldsymbol{S}^{(0)} - \mathbf{Z}\boldsymbol{\gamma}\right)^T \mathbf{W}^{(0)}(\boldsymbol{S}^{(0)} - \mathbf{Z}\boldsymbol{\gamma}) + \frac{1}{2}\boldsymbol{\gamma}^T\boldsymbol{\Omega}_{\lambda_n}(\boldsymbol{\gamma}^{(0)})\boldsymbol{\gamma},$$

where $\boldsymbol{S}^{(0)} = \boldsymbol{Y} - \zeta^{(0)}$, and $\mathbf{W}^{(0)}$ and $\boldsymbol{\Omega}_{\lambda_n}(\boldsymbol{\gamma}^{(0)})$ are diagonal weight matrices with w_i, and $p'_{\lambda_n}(||\boldsymbol{\gamma}_k^{(0)}||_{R_k})/||\boldsymbol{\gamma}_k^{(0)}||_{R_k}\mathbf{R}_k$ on the diagonals, respectively. This is a quadratic form with a minimizer satisfying

$$\left\{\mathbf{Z}^T\mathbf{W}^{(0)}\mathbf{Z} + \frac{n}{2}\boldsymbol{\Omega}_{\lambda_n}(\boldsymbol{\gamma}^{(0)})\right\}\boldsymbol{\gamma} = \mathbf{Z}^T\mathbf{W}^{(0)}\boldsymbol{S}^{(0)}. \tag{6}$$

The foregoing discussion leads to the following algorithm:

Step 1: Initialize $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(0)}$.
Step 2: Given $\boldsymbol{\gamma}^{(m)}$, update $\boldsymbol{\gamma}$ to $\boldsymbol{\gamma}^{(m+1)}$ by solving (6), where $\boldsymbol{\gamma}^{(0)}$ and the $\boldsymbol{\gamma}^{(0)}$ in $\mathbf{W}^{(0)}$, $\boldsymbol{S}^{(0)}$ and $\boldsymbol{\Omega}_{\lambda_n}(\boldsymbol{\gamma}^{(0)})$ are all set to be $\boldsymbol{\gamma}^{(m)}$.
Step 3: Iterate Step 2 until convergence of $\boldsymbol{\gamma}$ is achieved.

Due to the use of nonconvex penalty SCAD, the global minimizer cannot be achieved in general and only some local minimizers can be obtained (Fan and Li 2001). In the literature, all the penalized methods based on nonconvex penalties would suffer from the same drawback as that of SCAD. Thus, a suitable initial value is usually required for fast convergence. The initial estimator of $\boldsymbol{\gamma}$ in Step 1 can be chosen as the unpenalized estimator, which can be solved by fitting a L_1 regression on $n(n-1)/2$ pseudo observations $\{(\mathbf{Z}_i - \mathbf{Z}_j, Y_i - Y_j)\}_{i<j}$. In our numerical studies, we use the function *rq* in the R package *quantreg*. From our numerical experience, our algorithm converges fast with the unpenalized estimator, and the resulting solution is reasonably good as demonstrated in our simulation study.

Note that the iterated algorithm will be instable when the weights in (5) are too large. As suggested by Sievers and Abebe (2004), the algorithm should be modified so it removes those residuals with very large weights from the iteration and reinstates them in subsequent iterations when their contribution to the sum $\sum_{i=1}^{n} w_i(e_i - \zeta)^2$ becomes significant. Such an algorithm is quite efficient and reliable in practice. The R code for implementing the proposed scheme is available from the authors upon request. It is worth noting that we are doing an iterative approximation for both the original target function and the penalty function. Our numerical experience shows that

such an algorithm is usually completed in less than ten iterations and never fails to converge. For example, it takes <1 s per iteration in R using an Inter Core 2.2 MHz CPU for a $n = 200$, $p = 7$ case and the entire procedure is generally completed in <10 s. Theoretical investigation of the convergence property of the proposed algorithm definitely deserves future research.

3.3 Asymptotic properties

Without loss of generality, let $\beta_k(u)$, $k = 1, \ldots, s$, be the nonzero coefficient functions and let $\beta_k(u) \equiv 0$, for $k = s+1, \ldots, p$.

Theorem 3 *Suppose Conditions (C1)–(C5) hold. If* $K_n \log K_n/n \to 0$, $\rho_n \to 0$, $\lambda_n \to 0$, *and* $\lambda_n / \max\left\{\sqrt{K_n/n}, \rho_n\right\} \to \infty$, *we have the following:*

(i) $\bar{\beta}_k = 0$, $k = s+1, \ldots, p$, *with probability approaching 1.*

(ii) $||\bar{\beta}_k - \beta_k||_{L_2} = O_p\left(\max\left\{\sqrt{\frac{K_n}{n}}, \rho_n\right\}\right), k = 1, \ldots, s.$

Part (i) of Theorem 3 says that the proposed penalized rank-based method is consistent in variable selection; that is, it can identify the zero coefficient functions with probability tending to one. The second part provides the rate of convergence in estimating the nonzero coefficient functions.

Now we consider the asymptotic variance of the proposed estimate. Let $\boldsymbol{\beta}^{(1)} = (\beta_1, \ldots, \beta_s)^T$ denote the vector of nonzero coefficient functions, and let $\bar{\boldsymbol{\beta}}^{(1)} = (\bar{\beta}_1, \ldots, \bar{\beta}_s)^T$ denote its estimate obtained by minimizing (4). Let $\bar{\boldsymbol{\gamma}}^{(1)} = (\bar{\boldsymbol{\gamma}}_1^T, \ldots, \bar{\boldsymbol{\gamma}}_s^T)^T$ and $\mathbf{Z}^{(1)}$ denote the selected columns of $\mathbf{Z}$ corresponding to $\boldsymbol{\beta}^{(1)}$. By using Lemma 1 and the quadratic approximation stated in the above subsection, we obtain another approximated loss function

$$PL_n^{'}(\boldsymbol{\gamma}) = A_n\left(\theta_n^{-1}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)\right) + \frac{1}{2}\boldsymbol{\gamma}^T \boldsymbol{\Omega}_{\lambda_n}(\boldsymbol{\gamma})\boldsymbol{\gamma}. \tag{7}$$

Similarly, let $\boldsymbol{\Omega}_{\lambda}^{(1)}$ denote the selected diagonal blocks of $\boldsymbol{\Omega}_{\lambda}$, and $\boldsymbol{S}_n^{(1)}(\mathbf{0})$ denote the selected entries corresponding to $\boldsymbol{\beta}^{(1)}$. Thus, the minimizer of (7) yields

$$\bar{\boldsymbol{\gamma}}^{(1)} = \left\{2\tau(\mathbf{Z}^{(1)})^T\mathbf{Z}^{(1)} + \frac{n}{2}\boldsymbol{\Omega}_{\lambda_n}^{(1)}\bar{\boldsymbol{\gamma}}^{(1)}\right\}^{-1}\theta_n^{-1}\boldsymbol{S}_n^{(1)}(\mathbf{0}) + \boldsymbol{\gamma}_0.$$

Denote $\mathbf{H}^{(1)} = 2\tau(\mathbf{Z}^{(1)})^T\mathbf{Z}^{(1)} + \frac{n}{2}\boldsymbol{\Omega}_{\lambda_n}^{(1)}\bar{\boldsymbol{\gamma}}^{(1)}$, so the asymptotic variance of $\bar{\boldsymbol{\gamma}}^{(1)}$ is

$$\mathrm{avar}(\bar{\boldsymbol{\gamma}}^{(1)}) = \theta_n^{-2}(\mathbf{H}^{(1)})^{-1}\mathrm{var}(\boldsymbol{S}_n^{(1)}(\mathbf{0}))(\mathbf{H}^{(1)})^{-1}.$$

Since $\bar{\boldsymbol{\beta}}^{(1)} = (\mathbf{B}^{(1)})^T\bar{\boldsymbol{\gamma}}^{(1)}$, where $\mathbf{B}^{(1)}$ is the first s rows of $\mathbf{B}(u)$, we have $\mathrm{avar}(\bar{\boldsymbol{\beta}}^{(1)}) = (\mathbf{B}^{(1)})^T\mathrm{avar}(\bar{\boldsymbol{\gamma}}^{(1)})\mathbf{B}^{(1)}$. Let $\mathrm{var}^*(\bar{\boldsymbol{\beta}}(u))$ denote a modification of $\mathrm{avar}(\bar{\boldsymbol{\beta}}^{(1)})$ by replacing $\boldsymbol{\Omega}_{\lambda_n}^{(1)}$ with 0, that is

$$\text{var}^*\left(\bar{\boldsymbol{\beta}}(u)\right) = (4\tau^2)^{-1}\theta_n^{-2}\left(\mathbf{B}^{(1)}\right)^T\left((\mathbf{Z}^{(1)})^T\mathbf{Z}^{(1)}\right)^{-1}\text{var}\left(\boldsymbol{S}_n^{(1)}(\mathbf{0})\right)$$
$$\times\left((\mathbf{Z}^{(1)})^T\mathbf{Z}^{(1)}\right)^{-1}\mathbf{B}^{(1)}.$$

Accordingly, the diagonal elements of $\text{var}^*(\bar{\boldsymbol{\beta}}(u))$ can be employed as the asymptotic variances of $\bar{\beta}_k(u)$'s, i.e., $\text{avar}(\bar{\beta}_k(u))$, $k = 1, \ldots, s$.

Theorem 4 *Suppose Conditions (C1)–(C5) hold.* $K_n \log K_n/n \to 0$, $\rho_n \to 0$, $\lambda_n \to 0$, *and* $\lambda_n/\max\left\{\sqrt{K_n/n}, \rho_n\right\} \to \infty$. *Then, as* $n \to \infty$,

$$\left\{\text{var}^*\left(\bar{\boldsymbol{\beta}}(u)\right)\right\}^{-1/2}\left(\bar{\boldsymbol{\beta}}(u) - \breve{\boldsymbol{\beta}}(u)\right) \xrightarrow{d} N(0, \mathbf{I}_s),$$

where $\breve{\boldsymbol{\beta}}(u) = E[\bar{\boldsymbol{\beta}}(u) \mid \mathscr{X}]$ *and in particular,*

$$\left\{\text{var}^*(\bar{\beta}_k(u)\right\}^{-1/2}\left(\bar{\beta}_k(u) - \breve{\beta}_k(u)\right) \xrightarrow{d} N(0, 1), \quad k = 1, \ldots, s.$$

Here $\text{var}^*(\bar{\boldsymbol{\beta}}(u))$ is exactly the same asymptotic variance of nonpenalized rank-based estimate using only those covariates corresponding to nonzero coefficient functions (See Sect. 2). Theorem 4 implies that our penalized rank-based estimate has the oracle property in the sense that the asymptotic distribution of an estimated coefficient function is the same as that when it is known a priori which variables are in the model.

3.4 Selection of tuning parameters

The tuning parameter λ controls the model complexity and plays a critical role in the above procedure. It is desirable to select λ automatically by a data-driven method. Motivated by the Wilcoxon-type generalized BIC of Wang (2009) in which the multiple linear regression model is considered, we propose to select λ by minimizing

$$\begin{aligned}\text{BIC}_\lambda &= 12\hat{\tau}n^{-2}\sum_{i<j}\left|\left(Y_i - \mathbf{Z}_i\bar{\boldsymbol{\gamma}}_\lambda\right) - \left(Y_j - \mathbf{Z}_j\bar{\boldsymbol{\gamma}}_\lambda\right)\right| + df_\lambda \log(n/K_n)/(n/K_n)\\ &= 12\hat{\tau}n^{-2}\sum_{i<j}\left|\left(Y_i - \boldsymbol{X}_i^T\bar{\boldsymbol{\beta}}_\lambda\right) - \left(Y_j - \boldsymbol{X}_j^T\bar{\boldsymbol{\beta}}_\lambda\right)\right| + df_\lambda \log(n/K_n)/(n/K_n),\end{aligned} \tag{8}$$

where $\bar{\boldsymbol{\gamma}}_\lambda$ is the penalized local rank spline estimator with tuning parameter λ, df_λ is the number of nonzero components in $\bar{\boldsymbol{\beta}}_\lambda = \mathbf{B}\bar{\boldsymbol{\gamma}}_\lambda$, and $\hat{\tau}$ is an estimate of the Wilcoxon constant τ. The $\hat{\tau}$ can be robustly estimated by using the approach given in Hettmansperger and McKean (1998) and easily be calculated by the function *wilcoxontau* in the R package (Terpstra and McKean 2005) with the unpenalized estimates. We refer to this approach as the BIC-selector, and denote the selected λ by $\hat{\lambda}_{BIC}$. Similar to the BIC in Wang (2009), the first term in (8) can be viewed as an "artificial" likelihood as it shares some essential properties of a parametric log-likelihood.

Note that due to the use of spline smoothing, the effective sample size would be n/K_n rather than the original sample size n. This is because the classic parametric estimation methods is $\sqrt{n}$-consistent, while the convergence rate of spline methods is $\sqrt{n/K_n}$. Say, for each u, the spline estimator performs similarly to a parametric estimator as if a sample of size $\sqrt{n/K_n}$ from the model (1) with $\beta(u)$ were available. Therefore, the BIC_λ in (8) replaces $\log(n)/n$ in the Wang's (2009) Wilcoxon-type generalized BIC by $\log(n/K_n)/(n/K_n)$. It can be seen from the proof of Theorem 5, the BIC cannot achieve consistency without this modification.

Let S_T denote the true model and S_F denote the full model, and S_λ denote the set of the indices of the covariates selection by our robust variable selection method with tuning parameter λ. For a given candidate model S, let $\boldsymbol{\beta}_S$ be a vector of parameters and its ith coordinate is set to be zero, if $i \notin S$. Further, define $L_n^S = n^{-2}\sum_{i<j} |(Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}_S) - (Y_j - \boldsymbol{X}_j^T\hat{\boldsymbol{\beta}}_S)|$, where $\hat{\boldsymbol{\beta}}_S$ is the unpenalized robust estimator, i.e., the rank-based spline estimator for model S. We make the following same assumptions as those of Wang and Li (2009):

(1) for any $S \subset S_F$, $L_n^S \xrightarrow{p} L^S$ for some $L^S > 0$;
(2) for any $S \not\supseteq S_T$, we have $L^S > L^{S_T}$.

The next theorem indicates that $\hat{\lambda}_{BIC}$ leads to a penalized rank-based estimator which consistently yields the true model.

Theorem 5 *Suppose the assumptions above and Conditions (C1)–(C5) hold, then we have*

$$P(S_{\hat{\lambda}_{BIC}} = S_T) \to 1.$$

4 Numerical studies

4.1 Simulation

We study the finite-sample performance of the proposed rank-based spline SCAD (abbreviated by RSSCAD hereafter) method in this section. Wang and Xia (2009) have shown that the KLASSO is an efficient procedure in finite-sample cases and Wang et al. (2008) also proposed an efficient procedure based on least-squares and SCAD (abbreviated by LSSCAD hereafter). Thus, KLASSO and LSSCAD should be two ideal benchmarks in our comparison. For a clear comparison, we adopt the settings used in Wang and Xia (2009) for the following two models:

$$\text{(I)}: Y_i = 2\sin(2\pi U_i)X_{i1} + 4U_i(1-U_i)X_{i2} + \theta\varepsilon_i,$$
$$\text{(II)}: Y_i = \exp(2U_i - 1)X_{i1} + 8U_i(1-U_i)X_{i2} + 2\cos^2(2\pi U_i)X_{i3} + \theta\varepsilon_i,$$

where for the first model, $X_{i1} = 1$ and $(X_{i2}, \ldots, X_{i7})^T$ are generated from a multivariate normal distribution with $\mathrm{cov}(X_{ij_1}, X_{ij_2}) = \rho^{|j_1 - j_2|}$ for any $2 \le j_1, j_2 \le 7$, while for the second model, $X_{i1}, \ldots, X_{i7}$ are generated from a multivariate normal distribution with $\mathrm{cov}(X_{ij_1}, X_{ij_2}) = \rho^{|j_1 - j_2|}$ for any $1 \le j_1, j_2 \le 7$. Three cases of the

correlation between the covariates are considered, $\rho = 0.3, 0.5$, and 0.8. The index variable is simulated from Uniform(0, 1). The value of θ is fixed as 1.5. The following model, which is similar to the one used in Wang et al. (2008), is also included in the comparison:

$$\text{(III)} : Y_i = \beta_0(U_i) + \sum_{k=1}^{23} \beta_k(U_i) X_{ik} + \varsigma \varepsilon_i,$$

where

$$\beta_0(U) = 15 + 20\sin(0.5\pi U), \quad \beta_1(U) = 2 - 3\cos\left(\frac{\pi(6U-5)}{3}\right),$$
$$\beta_2(U) = 6 - 6U, \quad \beta_3(U) = -4 + \frac{1}{2}(2-3U)^3,$$

and the remaining coefficients are vanish. The index variable is still simulated from Uniform(0, 1). In this model, $\boldsymbol{X}$ depends on U in the following way. The first three variables are the true relevant covariates: X_{i1} is sampled uniformly from $[3U_i, 2+3U_i]$ at any given index U_i; X_{i2}, conditioning on X_{i1}, is Gaussian with mean 0 and variance $(1 + X_{i1})/(2 + X_{i1})$; and X_{i3} independent of X_{i1}, X_{i2}, is a Bernouli random variable with success rate 0.6. The other irrelevant variables are generated from a multivariate normal distribution with $\text{cov}(X_{ij_1}, X_{ij_2}) = 4\exp(-|j_1 - j_2|)$ for any $4 \le j_1, j_2 \le 23$. The parameter ς, which controls the model's signal-to-noise ratio, is set to 5. For all these three models, four error distributions are considered: $N(0, 1)$, $t(3)$ (Student's t-distribution with three degrees of freedom), Tukey contaminated normal $T(0.10; 5)$ (with the cumulative distribution function $F(x) = 0.9\Phi(x) + 0.1\Phi(x/5)$ where $\Phi(x)$ is the distribution function of a standard normal distribution) and Lognormal. In addition, an outlier case is considered, in which the responses of 10 % generated samples are shifted with a constant c. We use $c = 5$ and 25 for the first two models and the third model, respectively.

Throughout this section we use the B-spline and 1,000 replications for each considered example. For every simulated data, we firstly fit an unpenalized varying coefficient estimate $\hat{\boldsymbol{\beta}}(U_i)$, for which the number of knots, D, is selected via the method in Sect. 2.3. Then, the same D is used for RSSCAD, where the tuning parameter λ in the penalty function is chosen by the BIC (8). We report the average numbers of correct 0's (the average numbers of the true zero coefficients that are correctly estimated to be zero), the average number of incorrect 0's (the average number of the non-zero coefficients that are incorrectly estimated to be zero). Moreover, we also report the proportion of under-fitted models (at least one of the non-zero coefficients is incorrectly estimated to be zero), correctly fitted models (all the coefficients are selected correctly) and over-fitted models (all the non-zero coefficients are selected but at least one of the zero coefficient is estimated incorrectly to be non-zero). In addition, the performance of estimators in terms of estimation accuracy is assessed via the following two estimated errors which are defined by

Table 1 The simulation results of variable selection for model (I) with $\rho = 0.5$

		Number of zeros		Percentage of models		
		Correct	Incorrect	Under fitted	Correct	Over fitted
Error	Method					
$n = 200$						
Normal	RSSCAD	$4.998_{(0.0014)}$	$0.037_{(0.0060)}$	$0.037_{(0.0060)}$	$0.961_{(0.0061)}$	$0.002_{(0.0014)}$
	KLASSO	$4.970_{(0.0055)}$	$0.020_{(0.0044)}$	$0.030_{(0.0054)}$	$0.950_{(0.0069)}$	$0.020_{(0.0044)}$
	LSSCAD	$4.852_{(0.0120)}$	$0.003_{(0.0017)}$	$0.003_{(0.0017)}$	$0.872_{(0.0106)}$	$0.125_{(0.0105)}$
$t(3)$	RSSCAD	$4.994_{(0.0024)}$	$0.248_{(0.0137)}$	$0.248_{(0.0137)}$	$0.752_{(0.0137)}$	$0.000_{(0.0000)}$
	KLASSO	$4.267_{(0.0250)}$	$0.043_{(0.0064)}$	$0.043_{(0.0064)}$	$0.458_{(0.0157)}$	$0.499_{(0.0158)}$
	LSSCAD	$4.113_{(0.0270)}$	$0.060_{(0.0075)}$	$0.060_{(0.0075)}$	$0.402_{(0.0155)}$	$0.538_{(0.0158)}$
Lognormal	RSSCAD	$4.996_{(0.0020)}$	$0.029_{(0.0053)}$	$0.029_{(0.0053)}$	$0.968_{(0.0056)}$	$0.003_{(0.0017)}$
	KLASSO	$3.690_{(0.0311)}$	$0.060_{(0.0075)}$	$0.060_{(0.0075)}$	$0.270_{(0.0140)}$	$0.670_{(0.0149)}$
	LSSCAD	$3.465_{(0.0326)}$	$0.084_{(0.0088)}$	$0.084_{(0.0088)}$	$0.220_{(0.0131)}$	$0.696_{(0.0145)}$
$T(0.10, 5)$	RSSCAD	$4.996_{(0.0020)}$	$0.185_{(0.0123)}$	$0.185_{(0.0123)}$	$0.813_{(0.0123)}$	$0.002_{(0.0014)}$
	KLASSO	$2.268_{(0.0352)}$	$0.193_{(0.0125)}$	$0.178_{(0.0121)}$	$0.038_{(0.0060)}$	$0.784_{(0.0130)}$
	LSSCAD	$3.893_{(0.0294)}$	$0.078_{(0.0085)}$	$0.078_{(0.0085)}$	$0.320_{(0.0148)}$	$0.602_{(0.0155)}$
Outlier	RSSCAD	$4.995_{(0.0022)}$	$0.348_{(0.0151)}$	$0.348_{(0.0151)}$	$0.651_{(0.0151)}$	$0.001_{(0.0010)}$
	KLASSO	$3.700_{(0.0310)}$	$0.010_{(0.0031)}$	$0.010_{(0.0031)}$	$0.200_{(0.0126)}$	$0.790_{(0.0129)}$
	LSSCAD	$4.428_{(0.0225)}$	$0.041_{(0.0063)}$	$0.041_{(0.0063)}$	$0.562_{(0.0157)}$	$0.397_{(0.0155)}$
$n = 400$						
Normal	RSSCAD	$5.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$1.000_{(0.0000)}$	$0.000_{(0.0000)}$
	KLASSO	$5.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$1.000_{(0.0000)}$	$0.000_{(0.0000)}$
	LSSCAD	$4.967_{(0.0057)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.971_{(0.0053)}$	$0.029_{(0.0053)}$
$t(3)$	RSSCAD	$4.995_{(0.0041)}$	$0.015_{(0.0038)}$	$0.015_{(0.0038)}$	$0.984_{(0.0040)}$	$0.001_{(0.0010)}$
	KLASSO	$4.713_{(0.0164)}$	$0.008_{(0.0028)}$	$0.008_{(0.0028)}$	$0.767_{(0.0134)}$	$0.225_{(0.0132)}$
	LSSCAD	$4.633_{(0.0184)}$	$0.013_{(0.0036)}$	$0.013_{(0.0036)}$	$0.710_{(0.0143)}$	$0.277_{(0.0142)}$
Lognormal	RSSCAD	$5.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$1.000_{(0.0000)}$	$0.000_{(0.0000)}$
	KLASSO	$4.170_{(0.0263)}$	$0.017_{(0.0041)}$	$0.017_{(0.0041)}$	$0.436_{(0.0157)}$	$0.547_{(0.0157)}$
	LSSCAD	$4.354_{(0.0237)}$	$0.043_{(0.0064)}$	$0.043_{(0.0064)}$	$0.536_{(0.0158)}$	$0.421_{(0.0156)}$
$T(0.10, 5)$	RSSCAD	$4.997_{(0.0030)}$	$0.002_{(0.0014)}$	$0.002_{(0.0014)}$	$0.997_{(0.0017)}$	$0.001_{(0.0010)}$
	KLASSO	$2.546_{(0.0353)}$	$0.050_{(0.0069)}$	$0.048_{(0.0068)}$	$0.048_{(0.0068)}$	$0.903_{(0.0094)}$
	LSSCAD	$4.458_{(0.0220)}$	$0.016_{(0.0040)}$	$0.016_{(0.0040)}$	$0.661_{(0.0150)}$	$0.323_{(0.0148)}$
Outlier	RSSCAD	$5.000_{(0.0000)}$	$0.087_{(0.0089)}$	$0.087_{(0.0089)}$	$0.913_{(0.0089)}$	$0.000_{(0.0000)}$
	KLASSO	$4.505_{(0.0211)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.595_{(0.0155)}$	$0.405_{(0.0155)}$
	LSSCAD	$4.883_{(0.0107)}$	$0.001_{(0.0001)}$	$0.001_{(0.0001)}$	$0.895_{(0.0097)}$	$0.104_{(0.0097)}$

Standard errors are given in parentheses

$$\mathrm{EE1}(\hat{\boldsymbol{\beta}}_a) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}|\hat{\beta}_{aj}(U_i) - \beta_{0j}(U_i)|,$$

Table 2 The simulation results of estimated errors of $\boldsymbol{\beta}$ for model (I) with $\rho = 0.5$

Error		LSS	RS	LSSCAD	RSSCAD
$n = 200$					
Normal	MEE1	$0.19_{(0.035)}$	$0.23_{(0.052)}$	$0.06_{(0.027)}$	$0.09_{(0.043)}$
	MEE2	$0.16_{(0.028)}$	$0.20_{(0.048)}$	$0.05_{(0.022)}$	$0.08_{(0.041)}$
$t(3)$	MEE1	$0.31_{(0.077)}$	$0.28_{(0.066)}$	$0.12_{(0.067)}$	$0.11_{(0.057)}$
	MEE2	$0.25_{(0.063)}$	$0.24_{(0.061)}$	$0.10_{(0.054)}$	$0.10_{(0.053)}$
Lognormal	MEE1	$0.38_{(0.118)}$	$0.23_{(0.061)}$	$0.17_{(0.095)}$	$0.08_{(0.038)}$
	MEE2	$0.31_{(0.097)}$	$0.19_{(0.055)}$	$0.14_{(0.077)}$	$0.08_{(0.037)}$
$T(0.10, 5)$	MEE1	$0.34_{(0.083)}$	$0.28_{(0.067)}$	$0.14_{(0.067)}$	$0.11_{(0.054)}$
	MEE2	$0.28_{(0.068)}$	$0.24_{(0.061)}$	$0.12_{(0.055)}$	$0.10_{(0.051)}$
Outlier	MEE1	$0.36_{(0.063)}$	$0.33_{(0.076)}$	$0.15_{(0.063)}$	$0.14_{(0.062)}$
	MEE2	$0.29_{(0.051)}$	$0.28_{(0.069)}$	$0.13_{(0.051)}$	$0.13_{(0.059)}$
$n = 400$					
Normal	MEE1	$0.13_{(0.023)}$	$0.17_{(0.041)}$	$0.04_{(0.012)}$	$0.08_{(0.035)}$
	MEE2	$0.11_{(0.018)}$	$0.15_{(0.039)}$	$0.04_{(0.010)}$	$0.07_{(0.034)}$
$t(3)$	MEE1	$0.22_{(0.052)}$	$0.21_{(0.048)}$	$0.07_{(0.038)}$	$0.09_{(0.040)}$
	MEE2	$0.18_{(0.043)}$	$0.18_{(0.045)}$	$0.07_{(0.031)}$	$0.08_{(0.039)}$
Lognormal	MEE1	$0.26_{(0.069)}$	$0.16_{(0.040)}$	$0.10_{(0.054)}$	$0.07_{(0.029)}$
	MEE2	$0.22_{(0.056)}$	$0.14_{(0.037)}$	$0.09_{(0.044)}$	$0.07_{(0.029)}$
$T(0.10, 5)$	MEE1	$0.23_{(0.051)}$	$0.20_{(0.047)}$	$0.08_{(0.041)}$	$0.08_{(0.038)}$
	MEE2	$0.19_{(0.041)}$	$0.17_{(0.043)}$	$0.07_{(0.033)}$	$0.08_{(0.037)}$
Outlier	MEE1	$0.27_{(0.046)}$	$0.23_{(0.053)}$	$0.11_{(0.041)}$	$0.09_{(0.045)}$
	MEE2	$0.23_{(0.037)}$	$0.20_{(0.048)}$	$0.09_{(0.033)}$	$0.09_{(0.043)}$

Standard errors are given in parentheses

$$\mathrm{EE2}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_a) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}|X_{ij}\hat{\beta}_{aj}(U_i) - X_{ij}\beta_{0j}(U_i)|,$$

where $\hat{\boldsymbol{\beta}}_{aj}(\cdot)$ is an estimator of $\boldsymbol{\beta}_{0j}(\cdot)$ which is the true coefficient function. The means (denoted as MEE1 and MEE2) and standard deviations (in parentheses) of EE1 and EE2 values, are summarized. It is worth noting that because the KLASSO and RSSCAD use different smoothing approaches, we choose not to tabulate their MEE results to avoid misleading conclusions. Moreover, we also include two more unpenalized methods in the comparison, namely the rank-based spline estimator (RS) and the least-squares spline estimator (LSS).

We summarize the simulation results for the models (I)-(II) with $\rho = 0.5$ and the model (III) in Tables 1, 2, 3, 4, 5, and 6, respectively. The simulation results for the models (I)–(II) with $\rho = 0.3$ or 0.8 are presented in Tables A.1–A.8 of the supplemental file. A few observations can be made from Tables 1, 2, 3, 4, 5, and 6. Firstly, the proposed RSSCAD method is highly efficient for all the distributions under consideration. In terms of the probability of selecting the true model, the RSSCAD

Table 3 The simulation results of variable selection for model (II) with $\rho = 0.5$

		Number of zeros		Percentage of models		
		Correct	Incorrect	Under fitted	Correct	Over fitted
Error	Method					
$n = 200$						
Normal	RSSCAD	$3.996_{(0.0032)}$	$0.041_{(0.0063)}$	$0.041_{(0.0063)}$	$0.957_{(0.0064)}$	$0.002_{(0.0014)}$
	KLASSO	$3.962_{(0.0061)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.955_{(0.0066)}$	$0.045_{(0.0065)}$
	LSSCAD	$3.873_{(0.0111)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.899_{(0.0095)}$	$0.101_{(0.0095)}$
$t(3)$	RSSCAD	$3.998_{(0.0014)}$	$0.242_{(0.0138)}$	$0.239_{(0.0135)}$	$0.760_{(0.0135)}$	$0.001_{(0.0010)}$
	KLASSO	$3.790_{(0.0141)}$	$0.080_{(0.0086)}$	$0.070_{(0.0081)}$	$0.640_{(0.0152)}$	$0.290_{(0.0143)}$
	LSSCAD	$3.273_{(0.0244)}$	$0.064_{(0.0077)}$	$0.062_{(0.0076)}$	$0.466_{(0.0158)}$	$0.472_{(0.0158)}$
Lognormal	RSSCAD	$3.993_{(0.0044)}$	$0.074_{(0.0083)}$	$0.074_{(0.0083)}$	$0.923_{(0.0084)}$	$0.003_{(0.0017)}$
	KLASSO	$3.320_{(0.0238)}$	$0.160_{(0.0116)}$	$0.150_{(0.0113)}$	$0.480_{(0.0158)}$	$0.370_{(0.0153)}$
	LSSCAD	$2.094_{(0.0316)}$	$0.120_{(0.0103)}$	$0.118_{(0.0102)}$	$0.106_{(0.0097)}$	$0.776_{(0.0132)}$
$T(0.10, 5)$	RSSCAD	$3.999_{(0.0010)}$	$0.184_{(0.0126)}$	$0.180_{(0.0122)}$	$0.819_{(0.0122)}$	$0.001_{(0.0010)}$
	KLASSO	$2.090_{(0.0316)}$	$0.330_{(0.0149)}$	$0.300_{(0.0145)}$	$0.060_{(0.0075)}$	$0.640_{(0.0152)}$
	LSSCAD	$3.096_{(0.0265)}$	$0.095_{(0.0093)}$	$0.094_{(0.0092)}$	$0.353_{(0.0151)}$	$0.553_{(0.0157)}$
Outlier	RSSCAD	$3.997_{(0.0022)}$	$0.453_{(0.0171)}$	$0.430_{(0.0157)}$	$0.570_{(0.0157)}$	$0.000_{(0.0000)}$
	KLASSO	$2.950_{(0.0278)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.310_{(0.0146)}$	$0.690_{(0.0146)}$
	LSSCAD	$3.465_{(0.0215)}$	$0.056_{(0.0073)}$	$0.055_{(0.0072)}$	$0.558_{(0.0157)}$	$0.387_{(0.0154)}$
$n = 400$						
Normal	RSSCAD	$3.994_{(0.0042)}$	$0.001_{(0.0010)}$	$0.001_{(0.0010)}$	$0.997_{(0.0017)}$	$0.002_{(0.0014)}$
	KLASSO	$4.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$1.000_{(0.0000)}$	$0.000_{(0.0000)}$
	LSSCAD	$3.994_{(0.0024)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.997_{(0.0017)}$	$0.003_{(0.0017)}$
$t(3)$	RSSCAD	$4.000_{(0.0000)}$	$0.015_{(0.0038)}$	$0.015_{(0.0038)}$	$0.985_{(0.0038)}$	$0.000_{(0.0000)}$
	KLASSO	$3.930_{(0.0083)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.936_{(0.0077)}$	$0.064_{(0.0077)}$
	LSSCAD	$3.738_{(0.0156)}$	$0.012_{(0.0034)}$	$0.012_{(0.0034)}$	$0.769_{(0.0133)}$	$0.219_{(0.0131)}$
Lognormal	RSSCAD	$3.997_{(0.0030)}$	$0.000_{(0.0022)}$	$0.000_{(0.0017)}$	$0.999_{(0.0010)}$	$0.001_{(0.0010)}$
	KLASSO	$3.810_{(0.0135)}$	$0.110_{(0.0099)}$	$0.110_{(0.0099)}$	$0.780_{(0.0131)}$	$0.110_{(0.0099)}$
	LSSCAD	$3.113_{(0.0263)}$	$0.072_{(0.0082)}$	$0.072_{(0.0082)}$	$0.392_{(0.0154)}$	$0.536_{(0.0158)}$
$T(0.10, 5)$	RSSCAD	$4.000_{(0.0000)}$	$0.006_{(0.0024)}$	$0.006_{(0.0024)}$	$0.994_{(0.0024)}$	$0.000_{(0.0000)}$
	KLASSO	$2.488_{(0.0307)}$	$0.240_{(0.0135)}$	$0.230_{(0.0133)}$	$0.220_{(0.0131)}$	$0.550_{(0.0157)}$
	LSSCAD	$3.638_{(0.0181)}$	$0.020_{(0.0044)}$	$0.020_{(0.0044)}$	$0.681_{(0.0147)}$	$0.299_{(0.0145)}$
Outlier	RSSCAD	$4.000_{(0.0000)}$	$0.069_{(0.0080)}$	$0.069_{(0.0080)}$	$0.931_{(0.0080)}$	$0.000_{(0.0000)}$
	KLASSO	$3.905_{(0.0096)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.955_{(0.0066)}$	$0.045_{(0.0066)}$
	LSSCAD	$3.881_{(0.0107)}$	$0.001_{(0.0010)}$	$0.001_{(0.0010)}$	$0.895_{(0.0097)}$	$0.104_{(0.0096)}$

Standard errors are given in parentheses

performs slightly worse than KLASSO and LSSCAD when the random error comes from the normal distribution as we can expect, but it performs significantly better than KLASSO and LSSCAD when the error distribution is nonnormal. For instance, when the errors come from the contaminated normal distribution, the KLASSO hardly

Table 4 The simulation results of estimated errors of $\boldsymbol{\beta}$ for model (II) with $\rho = 0.5$

Error		LSS	RS	LSSCAD	RSSCAD
$n = 200$					
Normal	MEE1	$0.28_{(0.037)}$	$0.28_{(0.039)}$	$0.16_{(0.030)}$	$0.17_{(0.029)}$
	MEE2	$0.22_{(0.030)}$	$0.22_{(0.031)}$	$0.13_{(0.024)}$	$0.13_{(0.023)}$
$t(3)$	MEE1	$0.39_{(0.085)}$	$0.33_{(0.050)}$	$0.23_{(0.070)}$	$0.20_{(0.049)}$
	MEE2	$0.31_{(0.069)}$	$0.26_{(0.041)}$	$0.18_{(0.056)}$	$0.16_{(0.039)}$
Lognormal	MEE1	$0.46_{(0.125)}$	$0.29_{(0.046)}$	$0.28_{(0.105)}$	$0.17_{(0.031)}$
	MEE2	$0.36_{(0.099)}$	$0.23_{(0.038)}$	$0.22_{(0.083)}$	$0.13_{(0.025)}$
$T(0.10, 5)$	MEE1	$0.41_{(0.079)}$	$0.32_{(0.051)}$	$0.25_{(0.066)}$	$0.19_{(0.040)}$
	MEE2	$0.33_{(0.064)}$	$0.25_{(0.041)}$	$0.20_{(0.053)}$	$0.15_{(0.032)}$
Outlier	MEE1	$0.50_{(0.086)}$	$0.38_{(0.064)}$	$0.30_{(0.075)}$	$0.24_{(0.070)}$
	MEE2	$0.39_{(0.068)}$	$0.30_{(0.051)}$	$0.24_{(0.059)}$	$0.19_{(0.055)}$
$n = 400$					
Normal	MEE1	$0.22_{(0.025)}$	$0.22_{(0.026)}$	$0.14_{(0.016)}$	$0.14_{(0.016)}$
	MEE2	$0.17_{(0.021)}$	$0.18_{(0.021)}$	$0.11_{(0.014)}$	$0.11_{(0.013)}$
$t(3)$	MEE1	$0.29_{(0.057)}$	$0.25_{(0.032)}$	$0.18_{(0.039)}$	$0.15_{(0.020)}$
	MEE2	$0.23_{(0.045)}$	$0.20_{(0.026)}$	$0.14_{(0.031)}$	$0.12_{(0.016)}$
Lognormal	MEE1	$0.34_{(0.068)}$	$0.22_{(0.027)}$	$0.21_{(0.057)}$	$0.14_{(0.016)}$
	MEE2	$0.27_{(0.054)}$	$0.17_{(0.022)}$	$0.16_{(0.045)}$	$0.11_{(0.013)}$
$T(0.10, 5)$	MEE1	$0.31_{(0.049)}$	$0.24_{(0.031)}$	$0.19_{(0.040)}$	$0.15_{(0.018)}$
	MEE2	$0.25_{(0.039)}$	$0.19_{(0.025)}$	$0.15_{(0.032)}$	$0.12_{(0.014)}$
Outlier	MEE1	$0.35_{(0.051)}$	$0.28_{(0.038)}$	$0.21_{(0.046)}$	$0.17_{(0.031)}$
	MEE2	$0.28_{(0.041)}$	$0.22_{(0.031)}$	$0.17_{(0.037)}$	$0.13_{(0.024)}$

Standard errors are given in parentheses

selects the true model even for $n = 400$, whereas the RSSCAD selects the true model with quite large probability. For the third model in which the covariate $\boldsymbol{X}$ depends on the index U, RSSCAD is still quite effective in selecting the true variables. Also, from these three Tables 1, 3, and 5, we can see that the proposed smoothing parameter selection method and the BIC (8) perform satisfactorily and conform to the asymptotic results shown in Sects. 3.3 and 3.4.

In the literature, it is well demonstrated that BIC tends to identify the true sparse model well but would result in certain under-fitting when the sample size is not sufficiently large (as in the cases of $n = 200$). As shown in Theorem 5, the BIC is still consistent for selecting the variables in the present problem. When the sample size is larger (such as $n = 400$), our method would select the correctly fitted models with a quite large probability, at least 0.9. From Tables 2, 4, and 6, we observe that the MEEs of those penalized methods are smaller than those corresponding unpenalized methods in all cases. It means that the variable selection procedure can evidently increase the efficiency of estimators. Furthermore, the rank-based methods (RS and RSSCAD) perform better than the corresponding least squares methods (LSS and LSSCAD) when the error deviates from a normal distribution. Even for normal, the

Table 5 The simulation results of variable selection for model (III)

		Number of zeros		Percentage of models		
		Correct	Incorrect	Under fitted	Correct	Over fitted
Error	Method					
$n = 200$						
Normal	RSSCAD	$19.955_{(0.0079)}$	$0.014_{(0.0037)}$	$0.014_{(0.0037)}$	$0.950_{(0.069)}$	$0.036_{(0.0059)}$
	KLASSO	$19.410_{(0.0239)}$	$0.110_{(0.0099)}$	$0.110_{(0.0099)}$	$0.490_{(0.0158)}$	$0.400_{(0.0155)}$
	LSSCAD	$19.915_{(0.0092)}$	$0.002_{(0.0014)}$	$0.002_{(0.0014)}$	$0.927_{(0.0082)}$	$0.071_{(0.0081)}$
$t(3)$	RSSCAD	$19.668_{(0.0229)}$	$0.040_{(0.0064)}$	$0.039_{(0.0061)}$	$0.737_{(0.0139)}$	$0.224_{(0.0132)}$
	KLASSO	$17.440_{(0.0472)}$	$0.130_{(0.0106)}$	$0.130_{(0.0106)}$	$0.090_{(0.0090)}$	$0.780_{(0.0130)}$
	LSSCAD	$17.947_{(0.0429)}$	$0.022_{(0.0046)}$	$0.020_{(0.0045)}$	$0.318_{(0.0147)}$	$0.662_{(0.0149)}$
Lognormal	RSSCAD	$19.663_{(0.0301)}$	$0.022_{(0.0046)}$	$0.022_{(0.0046)}$	$0.777_{(0.0132)}$	$0.201_{(0.0127)}$
	KLASSO	$15.500_{(0.0591)}$	$0.230_{(0.0133)}$	$0.220_{(0.0131)}$	$0.050_{(0.0069)}$	$0.730_{(0.0140)}$
	LSSCAD	$15.793_{(0.0576)}$	$0.054_{(0.0071)}$	$0.053_{(0.0071)}$	$0.147_{(0.0112)}$	$0.800_{(0.0126)}$
$T(0.10, 5)$	RSSCAD	$19.656_{(0.0265)}$	$0.025_{(0.0049)}$	$0.025_{(0.0049)}$	$0.768_{(0.0134)}$	$0.207_{(0.0128)}$
	KLASSO	$6.880_{(0.0672)}$	$0.420_{(0.0156)}$	$0.380_{(0.0153)}$	$0.000_{(0.0000)}$	$0.620_{(0.0153)}$
	LSSCAD	$17.194_{(0.0491)}$	$0.022_{(0.0046)}$	$0.022_{(0.0046)}$	$0.187_{(0.0123)}$	$0.791_{(0.0129)}$
Outlier	RSSCAD	$18.838_{(0.0533)}$	$0.092_{(0.0094)}$	$0.090_{(0.0091)}$	$0.436_{(0.0157)}$	$0.474_{(0.0158)}$
	KLASSO	$16.740_{(0.0522)}$	$0.250_{(0.0137)}$	$0.240_{(0.0135)}$	$0.040_{(0.0062)}$	$0.720_{(0.0142)}$
	LSSCAD	$17.464_{(0.0471)}$	$0.038_{(0.0060)}$	$0.038_{(0.0060)}$	$0.154_{(0.0114)}$	$0.808_{(0.0125)}$
$n = 400$						
Normal	RSSCAD	$20.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$1.000_{(0.0000)}$	$0.000_{(0.0000)}$
	KLASSO	$19.890_{(0.0105)}$	$0.010_{(0.0031)}$	$0.010_{(0.0031)}$	$0.920_{(0.0086)}$	$0.070_{(0.0081)}$
	LSSCAD	$20.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$1.000_{(0.0000)}$	$0.000_{(0.0000)}$
$t(3)$	RSSCAD	$20.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$1.000_{(0.0000)}$	$0.000_{(0.0000)}$
	KLASSO	$17.179_{(0.0492)}$	$0.030_{(0.0054)}$	$0.030_{(0.0054)}$	$0.430_{(0.0157)}$	$0.540_{(0.0158)}$
	LSSCAD	$19.943_{(0.0075)}$	$0.016_{(0.0040)}$	$0.016_{(0.0040)}$	$0.955_{(0.0066)}$	$0.029_{(0.0053)}$
Lognormal	RSSCAD	$19.997_{(0.0022)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.998_{(0.0014)}$	$0.002_{(0.0014)}$
	KLASSO	$14.340_{(0.0637)}$	$0.040_{(0.0062)}$	$0.040_{(0.0062)}$	$0.268_{(0.0140)}$	$0.692_{(0.0146)}$
	LSSCAD	$19.574_{(0.0204)}$	$0.021_{(0.0045)}$	$0.021_{(0.0045)}$	$0.781_{(0.0131)}$	$0.198_{(0.0126)}$
$T(0.10, 5)$	RSSCAD	$20.000_{(0.0000)}$	$0.002_{(0.0014)}$	$0.002_{(0.0014)}$	$0.998_{(0.0014)}$	$0.000_{(0.0000)}$
	KLASSO	$10.244_{(0.0707)}$	$0.230_{(0.0133)}$	$0.210_{(0.0128)}$	$0.237_{(0.0134)}$	$0.553_{(0.0157)}$
	LSSCAD	$19.905_{(0.0097)}$	$0.008_{(0.0028)}$	$0.008_{(0.0028)}$	$0.918_{(0.0087)}$	$0.074_{(0.0083)}$
Outlier	RSSCAD	$19.996_{(0.0020)}$	$0.007_{(0.0026)}$	$0.007_{(0.0026)}$	$0.989_{(0.0033)}$	$0.004_{(0.0020)}$
	KLASSO	$17.490_{(0.0469)}$	$0.050_{(0.0069)}$	$0.050_{(0.0069)}$	$0.640_{(0.0152)}$	$0.310_{(0.0146)}$
	LSSCAD	$19.975_{(0.0050)}$	$0.000_{(0.0000)}$	$0.000_{(0.0000)}$	$0.971_{(0.0053)}$	$0.029_{(0.0053)}$

Standard errors are given in parentheses

MEEs of rank-based methods are merely larger than those least squares methods. This again reflects the robustness of our rank-based method to distributional assumption. Moreover, when the correlation between the covariates increases (decreases), all the three penalized methods become worse (better) but the comparison conclu-

Table 6 The simulation results of estimated errors of $\boldsymbol{\beta}$ for model (III)

Error		LSS	RS	LSSCAD	RSSCAD
$n = 200$					
Normal	MEE1	$0.62_{(0.098)}$	$0.66_{(0.111)}$	$0.31_{(0.084)}$	$0.35_{(0.103)}$
	MEE2	$0.84_{(0.140)}$	$0.89_{(0.153)}$	$0.35_{(0.115)}$	$0.39_{(0.136)}$
$t(3)$	MEE1	$0.99_{(0.253)}$	$0.88_{(0.170)}$	$0.57_{(0.270)}$	$0.46_{(0.147)}$
	MEE2	$1.36_{(0.350)}$	$1.19_{(0.233)}$	$0.69_{(0.376)}$	$0.52_{(0.196)}$
Lognormal	MEE1	$1.22_{(0.360)}$	$0.79_{(0.163)}$	$0.81_{(0.411)}$	$0.41_{(0.138)}$
	MEE2	$1.67_{(0.510)}$	$1.07_{(0.228)}$	$1.03_{(0.593)}$	$0.47_{(0.185)}$
$T(0.10, 5)$	MEE1	$1.07_{(0.227)}$	$0.85_{(0.171)}$	$0.65_{(0.244)}$	$0.44_{(0.140)}$
	MEE2	$1.47_{(0.318)}$	$1.16_{(0.235)}$	$0.80_{(0.346)}$	$0.50_{(0.186)}$
Outlier	MEE1	$1.28_{(0.224)}$	$1.14_{(0.211)}$	$0.87_{(0.258)}$	$0.57_{(0.192)}$
	MEE2	$1.77_{(0.312)}$	$1.56_{(0.294)}$	$1.11_{(0.369)}$	$0.68_{(0.266)}$
$n = 400$					
Normal	MEE1	$0.40_{(0.055)}$	$0.41_{(0.088)}$	$0.21_{(0.046)}$	$0.23_{(0.066)}$
	MEE2	$0.53_{(0.078)}$	$0.55_{(0.114)}$	$0.24_{(0.062)}$	$0.26_{(0.084)}$
$t(3)$	MEE1	$0.60_{(0.105)}$	$0.51_{(0.084)}$	$0.30_{(0.085)}$	$0.27_{(0.095)}$
	MEE2	$0.81_{(0.153)}$	$0.68_{(0.104)}$	$0.34_{(0.114)}$	$0.30_{(0.112)}$
Lognormal	MEE1	$0.73_{(0.181)}$	$0.43_{(0.073)}$	$0.37_{(0.161)}$	$0.24_{(0.061)}$
	MEE2	$1.00_{(0.254)}$	$0.57_{(0.095)}$	$0.43_{(0.222)}$	$0.27_{(0.070)}$
$T(0.10, 5)$	MEE1	$0.64_{(0.147)}$	$0.48_{(0.092)}$	$0.32_{(0.107)}$	$0.26_{(0.081)}$
	MEE2	$0.88_{(0.210)}$	$0.64_{(0.133)}$	$0.35_{(0.140)}$	$0.29_{(0.111)}$
Outlier	MEE1	$0.87_{(0.143)}$	$0.65_{(0.142)}$	$0.44_{(0.111)}$	$0.33_{(0.110)}$
	MEE2	$1.18_{(0.201)}$	$0.87_{(0.192)}$	$0.49_{(0.153)}$	$0.36_{(0.147)}$

Standard errors are given in parentheses

sion is similar to that of $\rho = 0.5$ (see Tables A.1–A.8 in the supplemental file). We also examine other error variance magnitudes for both models and the conclusion is similar.

To examine how well the method estimates the coefficient functions, Fig. 1 shows the estimates of the coefficients functions $\beta_1(\cdot)$ and $\beta_2(\cdot)$ for the model (I) with the normal and lognormal errors when $\rho = 0.5$ and $n = 200$. It can be seen that the estimates fit the true function well from the average viewpoint. The patterns of lower and upper confidence bands differ much from the true one at the right boundary, especially for $\beta_2(\cdot)$. This may be caused by the lack of data in that region. The curves for the other error distributions, which give similar pictures of the estimated functions, are shown in Figure A.1 of the supplemental file.

4.2 The Boston housing data

To further illustrate the usefulness of RSSCAD, we consider here the Boston Housing Data, which has been analyzed by Wang and Xia (2009) and is publicly available in the R package *mlbench*, (http://cran.r-project.org/). Following Wang and Xia (2009),

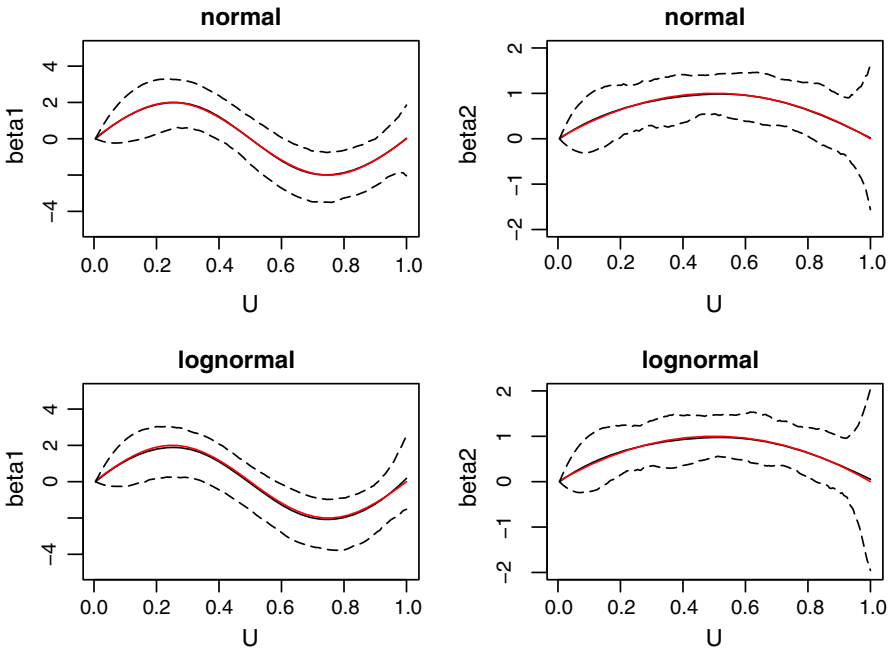

Fig. 1 Fitted regression coefficient functions of the Model (I) with the normal and lognormal errors when $\rho = 0.5$ and $n = 200$. The *red line* is the true coefficient function and the *black solid line* is the average of estimated coefficient function over 1,000 replications. The *lower* and *upper dash lines* form the 95 % confidence bands (color figure online)

we take MEDV [median value of owner-occupied homes in 1,000 United States dollar (USD)] as the response, LSTAT (the percentage of lower status of the population) as the index variable, and the following predictors as the X-variables: CRIM (per capita crime rate by town), RM (average number of rooms per dwelling), PTRATIO (pupil-teacher ratio by town), NOX (nitric oxides concentration parts per 10 million), TAX (full-value property-tax rate per 10,000 USD), and AGE (proportion of owner-occupied units built prior to 1940). Figure A.2 in the supplemental file shows the normal qq-plot of residuals obtained by using a standard local linear non-penalized varying coefficient estimation (Fan and Zhang 2008). This figure clearly indicates that the errors are not normal. In Wang and Xia (2009), the variables are firstly transformed so that their marginal distribution is approximately $N(0, 1)$. In our analysis, we do not take the transformation step since the RSSCAD is designed for robustness purpose. Similar to Wang and Xia (2009), the index variable, LSTAT, is transformed so that its marginal distribution is $U[0, 1]$.

A standard "leave-one-out" cross-validation method suggests an optimal number of knots $D = 5$. The RSSCAD method is then applied to the data with this number of knots. The optimal shrinkage parameter selected by the BIC criterion (8) is $\hat{\lambda} = 0.0284$. The resulting RSSCAD estimate suggests that NOX, RM, and PTRATIO are all relevant variables, whereas CRIM, TAX, and AGE seem not quite significant for predicting MEDV. To confirm whether the selected variables (NOX, RM, and

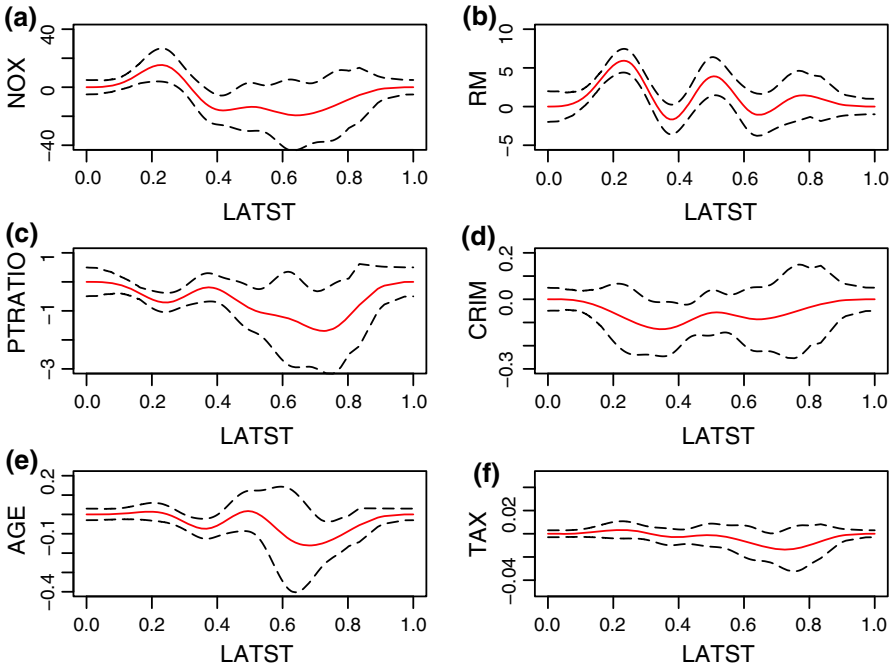

Fig. 2 **a–c** The RSSCAD estimates of the relevant coefficients NOX, RM, and PTRATIO; **d** the unpenalized estimates of the irrelevant coefficients

PTRATIO) are truly relevant, we provide in Fig. 2a–c their RSSCAD estimates with 95 % confidence bands. Obviously, they all suggest that these three coefficients are unlikely to be constant zero, because none of them is close to 0. The unpenalized estimates of the eliminated variables CRIM, TAX and AGE, are shown in Fig. 2d–f. We find that they are always close to zero over the entire range of the index variable LSTAT. Thus, Fig. 2 further confirms that those variables eliminated by RSSCAD are unlikely to be relevant. In contrast, without transformation of data, the KLASSO estimate suggests that all the variables are relevant except for AGE. Therefore, the proposed RSSCAD should be a reasonable alternative for variable selection in varying coefficient model by taking its efficiency, convenience and robustness into account.

5 Discussion

It is of interest to extend our proposed methodology to other more complex models, such as varying coefficient partially linear models (Li and Liang 2008; Zhao and Xue 2009). In fact, this amounts to adding further penalty terms into the rank-based loss function. Moreover, it is also of great interests to see whether RSSCAD and its oracle property are still valid in high-dimensional settings in which p diverges and even is larger than the sample size n. The consistency of the BIC criterion proposed in Sect. 3.4 deserves further study as well. Furthermore, our rank-based spline estimator could also

deal with the case that the distribution of the error term ε varies with time as well as the coefficient. For example, we can assume the following varying coefficient model $Y = \boldsymbol{X}^T(U)\boldsymbol{\beta}(U) + \sigma(U)\varepsilon$ where $\sigma(U)$ is a smooth function and the random error ε is independent of X and U. With certain modifications of the proof and conditions, we are able to establish the consistency of the rank-based methods.

Acknowledgments The authors thank the editor and two anonymous referees for their many helpful comments that have resulted in significant improvements in the article. This research was supported by the NNSF of China Grants Nos. 11131002, 11101306, 11371202, 71202087, 11271169, the RFDP of China Grant No. 20110031110002, Foundation for the Author of National Excellent Doctoral Dissertation of PR China, New Century Excellent Talents in University NCET-12-0276 and PAPD of Jiangsu Higher Education Institutions.

Appendix: Proofs of Theorems

In order to prove the theorems, we firstly state a few necessary lemmas. Throughout this appendix, $M_i, i = 1, \ldots, 11$ are all some positive constants which are independent of the samples.

Lemma 2 *Suppose Conditions (C1)–(C5) all hold and $\rho_n \to 0$, then*

$$\boldsymbol{S}_n(\boldsymbol{\gamma}^*) - \boldsymbol{S}_n(\mathbf{0}) = 2\tau\theta_n^2 \mathbf{Z}^T\mathbf{Z}\boldsymbol{\gamma}^* + o_p(1)\mathbf{1}_K,$$

where $\mathbf{1}_K$ is a K-dimension vector of ones, and $K = \sum_{i=1}^p K_i$.

Proof By $\Delta_i = O_p(\rho_n)$,

$$\begin{aligned}
&\boldsymbol{S}_n(\boldsymbol{\gamma}^*) - \boldsymbol{S}_n(\mathbf{0}) \\
&\quad = -\frac{\theta_n}{n}\sum_{i<j}(\mathbf{Z}_i^T - \mathbf{Z}_j^T)\left[\operatorname{sgn}\left(\varepsilon_i + \Delta_i - \varepsilon_j - \Delta_j - (\mathbf{Z}_i - \mathbf{Z}_j)\boldsymbol{\gamma}^*\right)\right. \\
&\qquad \left. - \operatorname{sgn}\left(\varepsilon_i + \Delta_i - \varepsilon_j - \Delta_j\right)\right] \\
&\quad = \frac{\theta_n}{n}\sum_{i=1}^n\sum_{j=1}^n \mathbf{Z}_i^T\left[\operatorname{sgn}\left(\varepsilon_i + \Delta_i - \varepsilon_j - \Delta_j\right)\right. \\
&\qquad \left. - \operatorname{sgn}\left(\varepsilon_i + \Delta_i - \varepsilon_j - \Delta_j - (\mathbf{Z}_i - \mathbf{Z}_j)\boldsymbol{\gamma}^*\right)\right] \\
&\quad = \frac{\theta_n}{n}\sum_{i=1}^n\sum_{j=1}^n \mathbf{Z}_i^T\left[\operatorname{sgn}\left(\varepsilon_i - \varepsilon_j\right) - \operatorname{sgn}\left(\varepsilon_i - \varepsilon_j - (\mathbf{Z}_i - \mathbf{Z}_j)\boldsymbol{\gamma}^*\right)\right]\left(1 + O_p(\rho_n)\right).
\end{aligned}$$

Fixing $(\boldsymbol{X}_i, U_i, \varepsilon_i)$, we define

$$W_i = \frac{1}{n}\sum_{j=1}^n\left[\operatorname{sgn}\left(\varepsilon_i - \varepsilon_j\right) - \operatorname{sgn}\left(\varepsilon_i - \varepsilon_j - \theta_n(\mathbf{Z}_i - \mathbf{Z}_j)\boldsymbol{\gamma}^*\right)\right].$$

Note that

$$\begin{aligned} E(W_i) &= \frac{1}{n}\sum_{j=1}^{n} E\left[\operatorname{sgn}\left(\varepsilon_i-\varepsilon_j\right)-\operatorname{sgn}\left(\varepsilon_i-\varepsilon_j-\theta_n(\mathbf{Z}_i-\mathbf{Z}_j)\boldsymbol{\gamma}^*\right)\right] \\ &= 2E\left[H(\varepsilon_i)-H(\varepsilon_i-\theta_n(\mathbf{Z}_i-\mathbf{Z}_j)\boldsymbol{\gamma}^*)\right] \\ &= 2h(\varepsilon_i)\theta_n\mathbf{Z}_i\boldsymbol{\gamma}^*(1+o_p(1)), \end{aligned}$$

and also

$$\begin{aligned} \operatorname{var}(W_i) &= E(W_i^2)-(E(W_i))^2 \\ &= \frac{1}{n^2}E\left\{\sum_{1\le l,k\le n}\left[\operatorname{sgn}\left(\varepsilon_i-\varepsilon_l\right)-\operatorname{sgn}\left(\varepsilon_i-\varepsilon_l-\theta_n(\mathbf{Z}_i-\mathbf{Z}_l)\boldsymbol{\gamma}^*\right)\right]\right. \\ &\quad \left.\times\left[\operatorname{sgn}\left(\varepsilon_i-\varepsilon_k\right)-\operatorname{sgn}\left(\varepsilon_i-\varepsilon_k-\theta_n(\mathbf{Z}_i-\mathbf{Z}_k)\boldsymbol{\gamma}^*\right)\right]\right\}-(E(W_i))^2 \\ &= -\frac{1}{n^2}(E(W_i))^2 \\ &\quad +\frac{1}{n^2}\sum_{j=1}^{n}E\left[\operatorname{sgn}\left(\varepsilon_i-\varepsilon_j\right)-\operatorname{sgn}\left(\varepsilon_i-\varepsilon_j-\theta_n(\mathbf{Z}_i-\mathbf{Z}_j)\boldsymbol{\gamma}^*\right)\right]^2 \\ &= -\frac{1}{n^2}(E(W_i))^2 \\ &\quad +\frac{2}{n^2}\sum_{j=1}^{n}E\left[\operatorname{sgn}\left(\varepsilon_i-\varepsilon_j\right)-\operatorname{sgn}\left(\varepsilon_i-\varepsilon_j-\theta_n(\mathbf{Z}_i-\mathbf{Z}_j)\boldsymbol{\gamma}^*\right)\right] \\ &= -\frac{1}{n^2}(E(W_i))^2+\frac{2}{n^2}E(W_i). \end{aligned}$$

so, we obtain $W_i = 2h(\varepsilon_i)\theta_n\mathbf{Z}_i\boldsymbol{\gamma}^*(1+o_p(1))$. Thus

$$\begin{aligned} \boldsymbol{S}_n(\boldsymbol{\gamma}^*)-\boldsymbol{S}_n(\mathbf{0}) &= \theta_n\sum_{i=1}^{n}\mathbf{Z}_i^T W_i(1+o_p(1)) = 2\theta_n^2\sum_{i=1}^{n}h(\varepsilon_i)\mathbf{Z}_i^T\mathbf{Z}_i\boldsymbol{\gamma}^*(1+o_p(1)) \\ &= 2\tau\theta_n^2\mathbf{Z}^T\mathbf{Z}\boldsymbol{\gamma}^*(1+o_p(1)). \end{aligned}$$

According to Lemma A.3 in Huang et al. (2004), $\theta_n^2\mathbf{Z}^T\mathbf{Z}\boldsymbol{\gamma}^*$ is bounded by a positive constant with probability tending to one, so

$$\theta_n^2\mathbf{Z}^T\mathbf{Z}\boldsymbol{\gamma}^*(1+o_p(1)) = \theta_n^2\mathbf{Z}^T\mathbf{Z}\boldsymbol{\gamma}^*+o_p(1)\mathbf{1}_K.$$

This completes the proof. □

Lemma 3 *Suppose Conditions (C1)–(C5) all hold. Then* $|\widehat{\boldsymbol{\gamma}^*}-\widetilde{\boldsymbol{\gamma}^*}|^2 = o_p(K_n)$, *where* $\widetilde{\boldsymbol{\gamma}^*} = \arg\min A_n(\boldsymbol{\gamma}^*)$.

Proof Define $Q_n^*(\boldsymbol{\gamma}^*) = A_n(\boldsymbol{\gamma}^*) + r_n(\boldsymbol{\gamma}^*)$ and for any constant $c > 0$,

$$T_n = \inf_{\sqrt{1/K}|\boldsymbol{\gamma}^* - \widetilde{\boldsymbol{\gamma}^*}| = c} |A_n(\boldsymbol{\gamma}^*) - A_n(\widetilde{\boldsymbol{\gamma}^*})|,$$
$$R_n = \sup_{\sqrt{1/K}|\boldsymbol{\gamma}^* - \widetilde{\boldsymbol{\gamma}^*}| \le c} |r_n(\boldsymbol{\gamma}^*)|.$$

If $\boldsymbol{\gamma}_1$ is outside the ball $\{\boldsymbol{\gamma} : \sqrt{1/K}|\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}^*}| \le c\}$, then $\boldsymbol{\gamma}_1 = \widetilde{\boldsymbol{\gamma}^*} + l\mathbf{1}$, where $\mathbf{1}$ is unit vector and $l > c$. Then,

$$\begin{aligned}
&\frac{c}{l}\left[Q_n^*(\boldsymbol{\gamma}_1) - Q_n^*(\widetilde{\boldsymbol{\gamma}})\right] \\
&= \frac{c}{l} Q_n^*(\boldsymbol{\gamma}_1) + \left(1 - \frac{c}{l}\right) Q_n\left(\widetilde{\boldsymbol{\gamma}^*}\right) - Q_n^*\left(\widetilde{\boldsymbol{\gamma}^*}\right) \\
&\ge Q_n^*\left(\frac{c}{l}\boldsymbol{\gamma}_1 + \left(1 - \frac{c}{l}\right)\widetilde{\boldsymbol{\gamma}^*}\right) - Q_n^*\left(\widetilde{\boldsymbol{\gamma}^*}\right) \\
&= A_n(\widetilde{\boldsymbol{\gamma}^*} + c\mathbf{1}) - A_n\left(\widetilde{\boldsymbol{\gamma}^*}\right) + \left(r_n(\widetilde{\boldsymbol{\gamma}^*} + c\mathbf{1}) - r_n\left(\widetilde{\boldsymbol{\gamma}^*}\right)\right) \\
&\ge T_n - 2R_n.
\end{aligned}$$

This implies if $R_n \le \frac{1}{2}T_n$ then the minimizer of Q_n^* must be inside the ball. Thus,

$$P(\sqrt{1/K}|\widehat{\boldsymbol{\gamma}^*} - \widetilde{\boldsymbol{\gamma}^*}| \ge c) \le P\left(R_n \ge \frac{1}{2}T_n\right).$$

Since $A_n(\boldsymbol{\gamma}^*)$ is a quadratic form, and $\widetilde{\boldsymbol{\gamma}^*}$ is its minimizer, so after some simple calculations, $A_n(\boldsymbol{\gamma}^*)$ can be rewritten as

$$A_n(\boldsymbol{\gamma}^*) = A_n\left(\widetilde{\boldsymbol{\gamma}^*}\right) + \theta_n^2\left(\boldsymbol{\gamma}^* - \widetilde{\boldsymbol{\gamma}}^*\right)^T \mathbf{Z}^T\mathbf{Z}\left(\boldsymbol{\gamma}^* - \widetilde{\boldsymbol{\gamma}}^*\right).$$

As a consequence, if $\boldsymbol{\gamma}^* = \widetilde{\boldsymbol{\gamma}}^* + c\mathbf{1}$,

$$A_n(\boldsymbol{\gamma}^*) - A_n(\widetilde{\boldsymbol{\gamma}^*}) = c^2\mathbf{1}^T\left(\frac{K}{n}\mathbf{Z}^T\mathbf{Z}\right)\mathbf{1} \ge M_5c^2,$$

where M_5 is the smallest eigenvalue of $\frac{K_n}{n}\mathbf{Z}^T\mathbf{Z}$ which is a positive constant with probability tending to one by Lemma A.3 in Huang et al. (2004). This implies that $T_n \ge \frac{1}{2}M_5c^2$. Hence,

$$P\left(R_n \ge \frac{1}{2}T_n\right) \le P\left(R_n \ge \frac{1}{2}M_5c^2\right).$$

On the other hand, according to Lemma 1, we obtain that $R_n \xrightarrow{p} 0$. Thus, by condition (C4), $|\widehat{\boldsymbol{\gamma}^*} - \widetilde{\boldsymbol{\gamma}^*}|^2 = o_p(K) = o_p(K_n)$. □

Lemma 4 *Suppose Conditions (C1)–(C5) all hold, then*

$$\boldsymbol{S}_n(\mathbf{0}) = O_p(1)\mathbf{1}_K.$$

Proof Write $\boldsymbol{S}_n(\boldsymbol{\gamma}) = (S_{n1}(\boldsymbol{\gamma}), \ldots, S_{nK}(\boldsymbol{\gamma}))^T$ and denote $\Delta_i = \boldsymbol{X}_i^T\boldsymbol{\beta}(U_i) - \boldsymbol{Z}_i\boldsymbol{\gamma}_0$. It suffices to show that $E(S_{nq}(\mathbf{0}))^2 = O(1), q = 1, \ldots, K$.

$$\begin{aligned}
&E\left(S_{n1}(\mathbf{0})\right)^2 \\
&= \frac{K_n}{4n^3} E\left\{\left(\sum_{m=1}^{n}\sum_{i=1}^{n}(Z_{m1} - Z_{i1})\,\mathrm{sgn}\left(\varepsilon_m + \Delta_m - \varepsilon_i - \Delta_i\right)\right)\right. \\
&\quad \left.\times\left(\sum_{k=1}^{n}\sum_{j=1}^{n}\left(Z_{k1} - Z_{j1}\right)\mathrm{sgn}\left(\varepsilon_k + \Delta_k - \varepsilon_j - \Delta_j\right)\right)\right\} \\
&= \frac{K_n}{4n^3}\sum_{k=1}^{n} E\left\{\sum_{i=1}^{n}(Z_{k1} - Z_{i1})\mathrm{sgn}\left(\varepsilon_k + \Delta_k - \varepsilon_i - \Delta_i\right)\right\}^2 \\
&\quad + \frac{K_n}{4n^3} E\left\{\sum_{m\neq k}\sum_{i=1}^{n}(Z_{m1} - Z_{i1})\,\mathrm{sgn}\left(\varepsilon_m + \Delta_m - \varepsilon_i - \Delta_i\right)\right. \\
&\quad \left.\times\sum_{j=1}^{n}\left(Z_{k1} - Z_{j1}\right)\mathrm{sgn}\left(\varepsilon_k + \Delta_k - \varepsilon_j - \Delta_j\right)\right\} := R_1 + R_2.
\end{aligned}$$

We next deal with R_1 and R_2. Firstly,

$$\begin{aligned}
R_1 &= \frac{K_n}{4n^3}\sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left\{\left(Z_{k1} - Z_{i1}\right)\left(Z_{k1} - Z_{j1}\right)\mathrm{sgn}\left(\varepsilon_k + \Delta_k - \varepsilon_j - \Delta_j\right)\right. \\
&\quad \left.\times\,\mathrm{sgn}\left(\varepsilon_k + \Delta_k - \varepsilon_i - \Delta_i\right)\right\} \\
&\le O\left(n^{-3}K_n\right)\sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left\{\left|\left(Z_{k1} - Z_{i1}\right)\left(Z_{k1} - Z_{j1}\right)\right|\right\} \\
&\le O\left(n^{-1}K_n\right)\sum_{k=1}^{n} E\left(Z_{k1}^2\right) + O\left(n^{-2}K_n\right)\sum_{k=1}^{n}\sum_{i=1}^{n} E\left(|Z_{k1}Z_{j1}|\right) \\
&\quad + O\left(n^{-2}K_n\right)\sum_{k=1}^{n}\sum_{i=1}^{n} E\left(|Z_{i1}Z_{k1}|\right) + O\left(n^{-2}K_n\right)\sum_{i=1}^{n} E\left(Z_{i1}^2\right) \\
&\quad + O\left(n^{-2}K_n\right)\sum_{i\neq j} E\left(|Z_{i1}Z_{j1}|\right).
\end{aligned}$$

By Lemma A.3 (Huang et al. 2004), there exists an interval $[M_3, M_4], 0 < M_3 < M_4 < \infty$, such that all of eigenvalues of $\frac{K_n}{n}\mathbf{Z}^T\mathbf{Z}$ fall into $[M_3, M_4]$ with probability tending to 1. We have $R_1 = O(1)$ immediately.

Note that $\Delta_i = O_p(\rho_n), i = 1, \ldots, n$. Thus,

$$\begin{aligned} R_2 = \frac{K_n}{4n^3} \sum_{m \neq k} E \Bigg\{ & \sum_{i=1}^{n} (Z_{m1} - Z_{i1}) \mathrm{sgn}(\varepsilon_m - \varepsilon_i + O_p(\rho_n)) \\ & \times \sum_{j=1}^{n} (Z_{k1} - Z_{j1}) \mathrm{sgn}(\varepsilon_k - \varepsilon_j + O_p(\rho_n)) \Bigg\} \\ = O(n^{-3} K_n) & \sum_{m \neq k} \sum_{1 \leq i, j \leq n} E \left\{ (Z_{m1} - Z_{i1})(Z_{k1} - Z_{j1}) \right\}. \end{aligned}$$

By taking the same procedure as R_1, we have $R_2 = O(1)$, which completes the proof. □

Proof of Theorem 1 Note that $\widetilde{\boldsymbol{\gamma}^*} = (2\tau)^{-1}(\theta_n^2 \mathbf{Z}^T\mathbf{Z})^{-1}\boldsymbol{S}_n(\mathbf{0})$, so

$$\begin{aligned} |\widetilde{\boldsymbol{\gamma}^*}|^2 &= \frac{1}{4\tau^2} \boldsymbol{S}_n^T(\mathbf{0})(\theta_n^2 \mathbf{Z}^T\mathbf{Z})^{-1}(\theta_n^2 \mathbf{Z}^T\mathbf{Z})^{-1}\boldsymbol{S}_n(\mathbf{0}) \\ &= O_p(1)\boldsymbol{S}_n^T(\mathbf{0})\boldsymbol{S}_n(\mathbf{0}) = O_p(1)\sum_{i=1}^{K} S_{ni}^2(\mathbf{0}) \\ &= O_p(K). \end{aligned}$$

By condition (C4), we obtain that $|\widetilde{\boldsymbol{\gamma}^*}|^2 = O_p(K_n)$. By the triangle inequality, we have

$$|\widehat{\boldsymbol{\gamma}^*}|^2 \leq |\widehat{\boldsymbol{\gamma}^*} - \widetilde{\boldsymbol{\gamma}^*}|^2 + |\widetilde{\boldsymbol{\gamma}^*}|^2,$$

and thus, by Lemma 3, $|\widehat{\boldsymbol{\gamma}^*}|^2 = O_p(K_n)$. Consequently, $|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0|^2 = O_p(n^{-1}K_n^2)$. By Lemma A.1 in Huang et al. (2004), $||\hat{\boldsymbol{\beta}} - \boldsymbol{g}^*||_{L_2}^2 = O_p(|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0|^2/K_n) = O_p(n^{-1}K_n)$. Finally, by the Cauchy-Schwarz inequality, $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_{L_2}^2 = O_p(\rho_n^2 + n^{-1}K_n)$. □

Lemma 5 *Suppose Conditions (C1)–(C5) all hold, then for any p-variate vector* $\boldsymbol{c}_n$ *whose components are not all zero,*

$$\boldsymbol{c}_n^T(\widetilde{\boldsymbol{\gamma}^*})/\sqrt{\mathrm{var}(\boldsymbol{c}_n^T(\widetilde{\boldsymbol{\gamma}^*}))} \xrightarrow{d} N(0, 1).$$

Proof By using $\widetilde{\boldsymbol{\gamma}^*} = (2\tau)^{-1}(\theta_n^2 \mathbf{Z}^T\mathbf{Z})^{-1}\boldsymbol{S}_n(\mathbf{0})$ again, it suffices to show that for any p-variate vector $\boldsymbol{b}_n$ whose components are not all zero, $\boldsymbol{b}_n^T\boldsymbol{S}_n(\mathbf{0})$ satisfies the Lindeberg–Feller condition. This can be easily verified by applying the dominated convergence theorem as briefly described below. Define

$$W_i = \frac{\sqrt{K_n}}{n^{3/2}} \sum_{j=1}^{n} \boldsymbol{b}_n^T \left(\boldsymbol{Z}_i^T - \boldsymbol{Z}_j^T \right) \mathrm{sgn}(\varepsilon_i - \varepsilon_j + \Delta_i - \Delta_j),$$

then we can write $\boldsymbol{b}_n^T \boldsymbol{S}_n(\boldsymbol{\gamma}_0) = \sum_{i=1}^n W_i$. Obviously, by applying Lemma 2, $EW_i = 0$, $\mathrm{var}(W_i) = \varsigma_i^2 < \infty$ and as $n \to \infty$

$$\max_{1 \le i \le n} \varsigma_i^2 \to 0,$$
$$\sum_{i=1}^{n} \varsigma_i^2 \to \varsigma^2, \quad 0 < \varsigma^2 < \infty.$$

We only need to check that

$$\lim_{n \to \infty} \sum_{i=1}^{n} E(W_i^2 I(|W_i| > \epsilon)) = 0, \tag{9}$$

for all $\epsilon > 0$, where $I(\cdot)$ is the indicator function. By applying Lemma 2 and the Cauchy–Schwarz inequality, we have $\sqrt{K_n/n^3} \sum_{j=1}^n \boldsymbol{b}_n^T (\boldsymbol{Z}_i^T - \boldsymbol{Z}_j^T) = o_p(1)$. Note that the random variable inside the expectation in (A.1) is bounded; hence, by dominated convergence we can interchange the limit and expectation. Since $I(|W_i| > \epsilon) \to 0$ the expectation goes to 0 and the assertion of the lemma follows from the central limited theorem. □

Proof of Theorem 2 By applying Lemmas 3 and 5, the theorem follows immediately from $\hat{\beta}_l(u) = \sum_{k=1}^{K_l} \hat{\gamma}_{lk} B_{lk}(u)$. □

Lemma 6 *Suppose Conditions (C1)–(C5) all hold. If* $\rho_n \to 0$, $\lambda_n \to 0$, $K_n/n \to 0$, *and* $\lambda_n/\rho_n \to \infty$ *as* $n \to \infty$, *then* $|\bar{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}| = O_p(K_n/\sqrt{n} + \sqrt{\lambda_n \rho_n K_n})$.

Proof Let $\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 = \delta_n K^{1/2} \boldsymbol{u}$, with $\boldsymbol{u}$ a vector satisfying $|\boldsymbol{u}| = 1$, $\delta_n > 0$ and $\boldsymbol{\gamma}_0 = ({\boldsymbol{\gamma}_1^0}^T, \ldots, {\boldsymbol{\gamma}_p^0}^T)^T$. We first show that $\delta_n = O_p(\theta_n + \lambda_n)$. Using the identity

$$|z - y| - |z| = -y\mathrm{sgn}(z) + 2(y - z)\{I(0 < z < y) - I(y < z < 0)\}$$

which holds for $z \neq 0$, we have

$$\begin{aligned} & Q_n(\bar{\boldsymbol{\gamma}}) - Q_n(\boldsymbol{\gamma}_0) \\ &= \frac{1}{n} \sum_{i<j} \left\{ |Y_i - Y_j - (\boldsymbol{Z}_i - \boldsymbol{Z}_j)\bar{\boldsymbol{\gamma}}| - |Y_i - Y_j - (\boldsymbol{Z}_i - \boldsymbol{Z}_j)\boldsymbol{\gamma}_0| \right\} \\ &= \frac{1}{n} \sum_{i<j} (\boldsymbol{Z}_i - \boldsymbol{Z}_j)(\boldsymbol{\gamma}_0 - \bar{\boldsymbol{\gamma}}) \mathrm{sgn}(Y_i - Y_j - (\boldsymbol{Z}_i - \boldsymbol{Z}_j)\boldsymbol{\gamma}_0) \\ &\quad + \frac{2}{n} \sum_{i<j} (-Y_i + Y_j + (\boldsymbol{Z}_i - \boldsymbol{Z}_j)\bar{\boldsymbol{\gamma}}) \end{aligned}$$

$$\times \left\{ I\Big(0 < Y_i - Y_j - (\mathbf{Z}_i - \mathbf{Z}_j)\boldsymbol{\gamma}_0 < (\mathbf{Z}_i - \mathbf{Z}_j)(\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\Big) \right.$$
$$\left. - I((\mathbf{Z}_i - \mathbf{Z}_j)(\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) < Y_i - Y_j - (\mathbf{Z}_i - \mathbf{Z}_j)\boldsymbol{\gamma}_0 < 0) \right\}$$
$$\doteq Q_1 + Q_2$$

According to Lemma 4, we can show that

$$\frac{1}{n}\sum_{i<j}(\mathbf{Z}_i - \mathbf{Z}_j)\mathrm{sgn}(Y_i - Y_j - (\mathbf{Z}_i - \mathbf{Z}_j)\boldsymbol{\gamma}_0) = \theta_n^{-1}\mathbf{1}_K.$$

And then $Q_1 \geq -M_6\delta_n n\theta_n$ for some positive constants M_6. Taking the same procedure as W_i in Lemma 2, we can also obtain that $Q_2 = \tau(\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^T\mathbf{Z}^T\mathbf{Z}(\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)(1 + o_p(1))$. Thus, according to Lemma A.3 (Huang et al. 2004), $Q_2 \geq M_3\delta_n^2 n$. Furthermore,

$$\sum_{k=1}^{p}(p_{\lambda_n}(||\bar{\boldsymbol{\gamma}}_k||_{R_k}) - p_{\lambda_n}(||\boldsymbol{\gamma}_k^0||_{R_k})) \geq -\sum_{k=1}^{p}\lambda_n||\bar{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k^0||_{R_k} \geq -M_7\lambda_n\delta_n.$$

Thus,

$$\begin{aligned} 0 &\geq PL_n(\bar{\boldsymbol{\gamma}}) - PL_n(\boldsymbol{\gamma}_0) \\ &= Q_n(\bar{\boldsymbol{\gamma}}) - Q_n(\boldsymbol{\gamma}_0) + n\sum_{k=1}^{p}(p_{\lambda_n}(||\bar{\boldsymbol{\gamma}}_k||_{R_k}) - p_{\lambda_n}(||\boldsymbol{\gamma}_k^0||_{R_k})) \\ &\geq \tau M_3\delta_n^2 n - M_6\delta_n n\theta_n - M_7 n\lambda_n\delta_n, \end{aligned}$$

which implies $\delta_n = O_p(\theta_n + \lambda_n)$.

Next we proceed to improve the obtained rate and show that $\delta_n = O_p(\theta_n + (\lambda_n\rho_n)^{1/2})$. For $k = 1, \ldots, p$, using properties of B-splines basis functions, we have

$$||\bar{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k^0||_{R_k}^2 \asymp K_k^{-1}|\bar{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k^0|^2,$$

where $A \asymp B$ means A/B is bounded. Thus, according to the Cauchy–Schwarz inequality, we have

$$|||\bar{\boldsymbol{\gamma}}_k||_{R_k} - ||\boldsymbol{\gamma}_k^0||_{R_k}| \leq ||\bar{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k^0||_{R_k} \asymp K_k^{-1/2}|\bar{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k^0| = o_p(1).$$

Note that

$$|||\boldsymbol{\gamma}_k^0||_{R_k} - ||\beta_k||_{L_2}| \leq |||g_k^*||_{L_2} - ||\beta_k||_{L_2}| \leq ||g_k^* - \beta_k||_{L_2} = O_p(\rho_n) = o_p(1).$$

It follows that $||\bar{\boldsymbol{\gamma}}_k||_{R_k} \to ||\beta_k||_{L_2}$ and $||\boldsymbol{\gamma}_k^0||_{R_k} \to ||\beta_k||_{L_2}$ with probability tending to one. Because $||\beta_k||_{L_2} > 0$ for $k = 1, \ldots, s$ and $\lambda_n \to 0$, we obtain that with probability tending to one,

$$||\bar{\boldsymbol{\gamma}}_k||_{R_k} > a\lambda_n, \quad ||\boldsymbol{\gamma}_k^0||_{R_k} > a\lambda_n \quad \text{for} \quad k = 1, \ldots, s.$$

On the other hand, $||\beta_k||_{L_2} = 0$, for $k = s+1, \ldots, p$, so $||\boldsymbol{\gamma}_k^0||_{R_k} = O_p(\rho_n)$. Because $\lambda_n/\rho_n \to \infty$, with probability tending to one,

$$||\boldsymbol{\gamma}_k^0||_{R_k} < \lambda_n \quad \text{for} \quad k = s+1, \ldots, p.$$

By the definition of $p_\lambda(\cdot)$,

$$\begin{aligned} &P\left\{p_{\lambda_n}\left(||\bar{\boldsymbol{\gamma}}_k||_{R_k}\right) = p_{\lambda_n}(||\boldsymbol{\gamma}_k^0||_{R_k})\right\} \to 1, \quad k = 1, \ldots, s, \\ &P\left\{p_{\lambda_n}\left(||\boldsymbol{\gamma}_k^0||_{R_k}\right) = \lambda_n||\boldsymbol{\gamma}_k^0||_{R_k}\right\} \to 1, \quad k = s+1, \ldots, p. \end{aligned} \tag{10}$$

Therefore,

$$\sum_{k=1}^{p}(p_{\lambda_n}(||\bar{\boldsymbol{\gamma}}_k||_{R_k}) - p_{\lambda_n}(||\boldsymbol{\gamma}_k^0||_{R_k})) \geq -\lambda_n \sum_{k=s+1}^{p} ||\boldsymbol{\gamma}_k^0||_{R_k} \geq -O_p(\lambda_n\rho_n).$$

So according to the first part, with probability tending to one,

$$0 \geq PL_n(\bar{\boldsymbol{\gamma}}) - PL_n(\boldsymbol{\gamma}_0) \geq \tau M_3\delta_n^2 n - M_6\delta_n n\theta_n - M_7 n\lambda_n\rho_n$$

which in turn implies that $\delta_n = O_p(\theta_n + (\lambda_n\rho_n)^{1/2})$. Then the lemma follows. □

Proof of Theorem 3 To prove the first part, we use the reduction to absurdity. Suppose that for sufficiently large n, there exist a constant $\xi > 0$ such that with probability at least ξ, there exist a $k_0 > s$ such that $\bar{\beta}_{k_0}(u) \neq 0$. Then $||\bar{\boldsymbol{\gamma}}_{k_0}||_{k_0} = ||\bar{\beta}_{k_0}(u)||_{L_2} > 0$. Let $\bar{\boldsymbol{\gamma}}^*$ be a vector constructed by replacing $\bar{\boldsymbol{\gamma}}_{k_0}$ with 0 in $\bar{\boldsymbol{\gamma}}$.

Taking the same procedure of Q_n as Lemma 6, we obtain

$$\begin{aligned} &PL_n(\bar{\boldsymbol{\gamma}}) - PL_n\left(\bar{\boldsymbol{\gamma}}^*\right) \\ &\quad = Q_n(\bar{\boldsymbol{\gamma}}) - Q_n\left(\bar{\boldsymbol{\gamma}}^*\right) + np_{\lambda_n}(||\bar{\boldsymbol{\gamma}}_{k_0}||_{k_0}) \\ &\quad = \tau\left(\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}^*\right)^T \mathbf{Z}^T\mathbf{Z}\left(\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}^*\right)(1 + o_p(1)) \\ &\qquad + \tau(\bar{\boldsymbol{\gamma}}^* - \boldsymbol{\gamma}_0)^T\mathbf{Z}^T\mathbf{Z}(\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}^*)(1 + o_p(1)) \\ &\qquad - \theta_n^{-1}\left(\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}^*\right)^T \boldsymbol{S}_n(\mathbf{0}) + np_{\lambda_n}(||\bar{\boldsymbol{\gamma}}_{k_0}||_{k_0}) \end{aligned}$$

According to Lemma A.3 in Huang et al. (2004) and Lemma 4, we obtain the following inequalities,

$$\begin{aligned} &(\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}^*)^T\mathbf{Z}^T\mathbf{Z}(\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}^*) \geq M_3\frac{n}{K_n}|\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}^*|^2 = M_3\frac{n}{K_n}|\bar{\boldsymbol{\gamma}}_{k_0}|^2, \\ &|(\bar{\boldsymbol{\gamma}}^* - \boldsymbol{\gamma}_0)^T\mathbf{Z}^T\mathbf{Z}(\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}^*)| \leq M_4\frac{n}{K_n}|\bar{\boldsymbol{\gamma}}^* - \boldsymbol{\gamma}_0||\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}^*| = M_4\frac{n}{K_n}|\bar{\boldsymbol{\gamma}}^* - \boldsymbol{\gamma}_0||\bar{\boldsymbol{\gamma}}_{k_0}|, \\ &|\theta_n^{-1}(\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}^*)^T\boldsymbol{S}_n(\mathbf{0})| \leq \theta_n^{-1}|\bar{\boldsymbol{\gamma}}_{k_0}||\boldsymbol{S}_n(\mathbf{0})| \leq M_8\sqrt{n}|\bar{\boldsymbol{\gamma}}_{k_0}|. \end{aligned}$$

Consequently, according to (10), with probability tending to one,

$$p_{\lambda_n}(||\bar{\boldsymbol{\gamma}}_{k_0}||_{k_0}) = \lambda_n ||\bar{\boldsymbol{\gamma}}_{k_0}||_{k_0} \asymp \lambda_n \frac{1}{\sqrt{K_n}} |\bar{\boldsymbol{\gamma}}_{k_0}|$$

Then,

$$\begin{aligned}
&PL_n(\bar{\boldsymbol{\gamma}}) - PL_n(\bar{\boldsymbol{\gamma}}^*) \\
&\geq \tau M_3 \frac{n}{K_n} |\bar{\boldsymbol{\gamma}}_{k_0}|^2 - 2\tau M_4 \frac{n}{K_n} |\bar{\boldsymbol{\gamma}}_{k_0}| |\bar{\boldsymbol{\gamma}}^* - \boldsymbol{\gamma}_0| - M_8 \sqrt{n} |\bar{\boldsymbol{\gamma}}_{k_0}| \\
&\quad + n\lambda_n \frac{1}{\sqrt{K_n}} |\bar{\boldsymbol{\gamma}}_{k_0}| + o_p(1) \\
&\geq \tau (M_3 - M_4) \frac{n}{K_n} |\bar{\boldsymbol{\gamma}}_{k_0}|^2 - 2\tau M_9 \frac{n}{K_n} \left(\frac{K_n}{\sqrt{n}} + \sqrt{\lambda_n \rho_n K_n} \right) |\bar{\boldsymbol{\gamma}}_{k_0}| + o_p(1) \\
&\quad - M_8 \sqrt{n} |\bar{\boldsymbol{\gamma}}_{k_0}| + n\lambda_n \frac{1}{\sqrt{K_n}} |\bar{\boldsymbol{\gamma}}_{k_0}| + o_p(1) \\
&\geq -2\tau M_9 \frac{n}{K_n} \left(\frac{K_n}{\sqrt{n}} + \sqrt{\lambda_n \rho_n K_n} \right) |\bar{\boldsymbol{\gamma}}_{k_0}| - M_8 \sqrt{n} |\bar{\boldsymbol{\gamma}}_{k_0}| \\
&\quad + n\lambda_n \frac{1}{\sqrt{K_n}} |\bar{\boldsymbol{\gamma}}_{k_0}| + o_p(1).
\end{aligned}$$

According to the conditions, the third term dominates both the first and second terms, which contradicts the fact that $P_n(\bar{\boldsymbol{\gamma}}) - P_n(\bar{\boldsymbol{\gamma}}^*) \leq 0$. We thus have proved part (i). Next, we will prove part (ii). Denote $\boldsymbol{\beta} = ((\boldsymbol{\beta}^{(1)})^T, (\boldsymbol{\beta}^{(2)})^T)^T$, where $\boldsymbol{\beta}^{(1)} = (\beta_1, \ldots, \beta_s)^T$ and $\boldsymbol{\beta}^{(2)} = (\beta_{s+1}, \ldots, \beta_p)^T$, and $\boldsymbol{\gamma} = ((\boldsymbol{\gamma}^{(1)})^T, (\boldsymbol{\gamma}^{(2)})^T)^T$, where $\boldsymbol{\gamma}^{(1)} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_s)^T$ and $\boldsymbol{\gamma}^{(2)} = (\boldsymbol{\gamma}_{s+1}, \ldots, \boldsymbol{\gamma}_p)^T$. Similarly, denote $\boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)}, \boldsymbol{Z}_i^{(2)})$. Define the oracle version of $\boldsymbol{\gamma}$,

$$\bar{\boldsymbol{\gamma}}_{oracle} = \arg \min_{\boldsymbol{\gamma} = ((\boldsymbol{\gamma}^{(1)})^T, \mathbf{0}^T)^T} Q_n(\boldsymbol{\gamma})$$

which is obtained as if the information of nonzero components were given; the corresponding vector of coefficient functions is designated $\bar{\boldsymbol{\beta}}_{oracle}$. By the above lemmas, we can easily obtain that $||\bar{\beta}_{k,oracle}||_{L_2} \to ||\beta_k||_{L_2}$, for $k = 1, \ldots, s$ and by the definition, $||\bar{\beta}_{k,oracle}||_{L_2} = 0$, for $k = s+1, \ldots, p$. By part (i) of the theorem, $\bar{\boldsymbol{\gamma}} = ((\bar{\boldsymbol{\gamma}}^{(1)})^T, \mathbf{0}^T)^T$, and

$$\sum_{k=1}^{p} p_{\lambda_n}(||\bar{\beta}_{k,oracle}||_{L_2}) = p_{\lambda_n}(||\bar{\beta}_k||_{L_2})$$

with probability tending to one. Let $\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}_{oracle} = \delta_n K_n^{1/2} \boldsymbol{v}$, with $\boldsymbol{v} = ((\boldsymbol{v}^{(1)})^T, \mathbf{0}^T)^T$, and $|\boldsymbol{v}| = 1$. Then $||\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}_{oracle}||_{L_2} \asymp K_n^{-1} |\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}_{oracle}| = \delta_n$. Similar to part (i),

$$0 \geq PL_n(\bar{\boldsymbol{\gamma}}) - PL_n(\bar{\boldsymbol{\gamma}}_{oracle})$$

$$= A_n(\theta_n^{-1}(\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)) - A_n(\theta_n^{-1}(\bar{\boldsymbol{\gamma}}_{oracle} - \boldsymbol{\gamma}_0)) + o_p(1)$$
$$\geq -M_{10}\theta_n\delta_n + M_{11}\delta_n^2 + o_p(1).$$

Thus $||\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}_{oracle}||_{L_2} \asymp \delta_n = O_p(\theta_n)$, which implies that $||\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}||_{L_2} = O_p(\rho_n + \theta_n)$. The desired result follows. □

Proof of Theorem 4 By Theorem 3, with probability tending to one, $\bar{\boldsymbol{\gamma}} = ((\bar{\boldsymbol{\gamma}}^{(1)})^T, \mathbf{0}^T)^T$ is a local minimizer of $PL_n(\boldsymbol{\gamma})$. Thus, by the definition of $PL_n(\boldsymbol{\gamma})$,

$$0 = \frac{\partial PL_n(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=((\bar{\boldsymbol{\gamma}}^{(1)})^T,\mathbf{0}^T)^T}$$
$$= \frac{\partial Q_n(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=((\bar{\boldsymbol{\gamma}}^{(1)})^T,\mathbf{0}^T)^T} + n\sum_{k=1}^{p} p_{\lambda_n}(||\boldsymbol{\gamma}_k||_{R_k})\bigg|_{\boldsymbol{\gamma}=((\bar{\boldsymbol{\gamma}}^{(1)})^T,\mathbf{0}^T)^T}$$

According to the proof of Lemma 6, $||\bar{\boldsymbol{\gamma}}_k||_{R_k} > a\lambda_n$, for $k = 1, \ldots, s$, so the second part of the above equation is 0. Thus, $\frac{\partial Q_n(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\gamma}=((\bar{\boldsymbol{\gamma}}^{(1)})^T,\mathbf{0}^T)^T} = 0$, which implies

$$\bar{\boldsymbol{\gamma}}^{(1)} = \arg\min_{\boldsymbol{\gamma}^{(1)}} \frac{1}{n}\sum_{i<j} |Y_i - Y_j - (\mathbf{Z}_i^{(1)} - \mathbf{Z}_j^{(1)})\boldsymbol{\gamma}^{(1)}|.$$

Applying Theorem 2, we can easily obtain the result. □

Proof of Theorem 5 Firstly, note that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$ by Theorem 1, so, taking the same procedure as the proof of Theorem 1 in Koul et al. (1987), $\hat{\tau}$ is a consistent estimator of τ. So we replace $\hat{\tau}$ with τ in the following proof. To establish the consistency of BIC, we first construct a sequence of reference tuning parameters, $\lambda_n = \log(n/K_n)/\sqrt{n/K_n}$. By Theorem 4, the penalty estimator $\bar{\boldsymbol{\beta}}_{\lambda_n}$ is exactly the same as the oracle estimator $\bar{\boldsymbol{\beta}}_{oracle}$. It follows immediately that $P(BIC_{\lambda_n} = BIC_{S_T}) \to 1$, which implies $BIC_{\lambda_n} \xrightarrow{p} \log(L^{S_T})$. Next, we verify that $P(\inf_{\lambda\in\Omega_-\cup\Omega_+} BIC_\lambda > BIC_{\lambda_n}) \to 1$, where Ω_- and Ω_+ denote the underfitting case and overfitting case, respectively.

Case 1: Underfitted model, i.e., the model misses at least one covariate in the true model. For any $\lambda \in \Omega_-$, similar to Wang and Li (2009), we have

$$BIC_\lambda \geq n^{-2}\sum_{i<j} |(Y_i - \mathbf{Z}_i\bar{\boldsymbol{\gamma}}_\lambda) - (Y_j - \mathbf{Z}_j\bar{\boldsymbol{\gamma}}_\lambda)| \geq \inf_{S\supsetneq S_T} L_n^S > L^{S_T}$$

Case 2: Overfitted model, i.e., the model contains all the covariates in the true model and at least one covariate that does not belong to the true model. For any $\lambda \in \Omega_+$, by Lemma 1, we have

$$
\begin{aligned}
&Q_n(\bar{\boldsymbol{\gamma}}_{S_T}) - Q_n(\bar{\boldsymbol{\gamma}}_\lambda) \\
&\quad = A_n(\theta_n^{-1}(\bar{\boldsymbol{\gamma}}_{S_T} - \boldsymbol{\gamma}_0)) - A_n(\theta_n^{-1}(\bar{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}_0)) + o_p(1) \\
&\quad = A_n(\widetilde{\boldsymbol{\gamma}^*}) - A_n(\theta_n^{-1}(\bar{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}_0)) + o_p(1) \\
&\quad = \tau\theta_n^2(\widetilde{\boldsymbol{\gamma}^*} - \theta_n^{-1}(\bar{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}_0))^T \mathbf{Z}^T \mathbf{Z}(\widetilde{\boldsymbol{\gamma}^*} - \theta_n^{-1}(\bar{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}_0)) + o_p(1).
\end{aligned}
$$

By Lemma A.3 in Huang et al. (2004) and Lemma 6, we obtain $Q_n(\bar{\boldsymbol{\gamma}}_{S_T}) - Q_n(\bar{\boldsymbol{\gamma}}_\lambda) = O_p(K_n)$. Thus, with probability tending to one,

$$
\begin{aligned}
n(BIC_\lambda - BIC_{\lambda_n}) &= 12\tau(Q_n(\bar{\boldsymbol{\gamma}}_\lambda) - Q_n(\bar{\boldsymbol{\gamma}}_{\lambda_n})) + (df_\lambda - df_{\lambda_n})K_n \log(n/K_n) \\
&\geq 12\tau[Q_n(\bar{\boldsymbol{\gamma}}_\lambda) - Q_n(\bar{\boldsymbol{\gamma}}_{S_T})] + o_p(1) + K_n \log(n/K_n).
\end{aligned}
$$

This implies that

$$
\begin{aligned}
\inf_{\lambda \in \Omega_+} n(BIC_\lambda - BIC_{\lambda_n}) \geq 12\tau \min_{S \supset S_T} [Q_n(\bar{\boldsymbol{\gamma}}_\lambda) - Q_n(\bar{\boldsymbol{\gamma}}_{S_T})] \\
+ o_p(1) + K_n \log(n/K_n).
\end{aligned}
$$

Since the last term dominates the first term and diverges to $+\infty$, we obtain $P(\inf_{\lambda \in \Omega_+}(BIC_\lambda - BIC_{\lambda_n}) > 0) \to 1$.

Thus, according to the above results, those λ's which fail to identify the true model cannot be selected by BIC asymptotically, because at least the true model identified by λ_n is a better choice. As a result, the optimal value $\hat{\lambda}_{BIC}$ can only be one of those λ's whose corresponding estimator yields the true model. Hence, the theorem follows immediately. □

References

Chiang CT, Rice JA, Wu CO (2001) Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. J Am Stat Assoc 96:605–619

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360

Fan J, Zhang W (1999) Statistical estimation in varying-coefficient models. Ann Stat 27:1491–1518

Fan J, Zhang W (2008) Statistical methods with varying coefficient models. Stat Interface 1:179–195

Hastie TJ, Tibshirani RJ (1993) Varying-coefficient models. J R Stat Soc Ser B 55:757–796 (with discussion)

Hettmansperger TP, McKean JW (1998) Robust nonparametric statistical methods. Arnold, London

Hoover DR, Rice JA, Wu CO, Yang L-P (1998) Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika 85:809–822

Huang JZ, Shen H (2004) Functional coefficient regression models for nonlinear time series: a polynomial spline approach. Scand J Statist 31:515–534

Huang JZ, Wu CO, Zhou L (2002) Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika 89:111–128

Huang JZ, Wu CO, Zhou L (2004) Polynomial spline estimation and inference for varying coefficient models with longitudinal data. Stat Sinica 14:763–788

Jaeckel LA (1972) Estimating regression coefficients by minimizing the dispersion of residuals. Ann Math Stat 43:1449–1458

Kauermann G, Tutz G (1999) On model diagnostics using varying coefficient models. Biometrika 86:119–128

Koul HL, Sievers GL, McKean JW (1987) An estimator of the scale parameter for the rank analysis of linear models under general score functions. Scand J Stat 14:131–141

Kim M-O (2007) Quantile regression with varying coefficients. Ann Stat 35:92–108
Leng C (2009) A simple approach for varying-coefficient model selection. J Stat Plan Inferface 139:2138–2146
Leng C (2010) Variable selection and coefficient estimation via regularized rank regression. Stat Sinica 20:167–181
Li R, Liang H (2008) Variable selection in semi-parametric regression model. Ann Stat 36:261–286
Lin Y, Zhang HH (2006) Component selection and smoothing in smoothing spline analysis of variance models. Ann Stat 34:2272–2297
Sievers G, Abebe A (2004) Rank estimation of regression coefficients using iterated reweighted least squares. J Stat Comput Simul 74:821–831
Tang Y, Wang H, Zhu Z, Song X (2012) A unified variable selection approach for varying coefficient models. Stat Sinica 22:601–628
Terpstra J, McKean J (2005) Rank-based analysis of linear models using R. J Stat Softw 14:1–26
Tibshirani RJ (1996) Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B 58:267–288
Wang L, Chen G, Li H (2007) Group scad regression analysis for microarray time course gene expression data. Bioinformatics 23:1486–1494
Wang L, Kai B, Li R (2009) Local rank inference for varying coefficient models. J Am Stat Assoc 104:1631–1645
Wang L, Li H, Huang J (2008) Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. J Am Stat Assoc 103:1556–1569
Wang L (2009) Wilcoxon-type generalized Bayesian information criterion. Biometrika 96:163–173
Wang L, Li R (2009) Weighted Wilcoxon-type smoothly clipped absolute deviation method. Biometrics 65:564–571
Wang H, Li G, Jiang G (2007) Robust regression shrinkage and consistent variable selection via the LAD-LASSO. J Bus Econ Stat 25:347–355
Wang H, Xia Y (2009) Shrinkage estimation of the varying coefficient model. J Am Stat Assoc 104:747–757
Wu CO, Chiang C-T, Hoover DR (1998) Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. J Am Stat Assoc 93:1388–1402
Zhao P, Xue L (2009) Variable selection for semiparametric varying coefficient partially linear models. Stat Prob lett 79:2148–2157
Zou H (2006) The adaptive LASSO and its oracle properties. J Am Stat Assoc 101:1418–1429
Zou H, Yuan M (2008) Composite quantile regression and the oracle model selection theory. Ann Stat 36:1108–1126