Stat Papers (2019) 60:1255–1281
https://doi.org/10.1007/s00362-017-0874-z

REGULAR ARTICLE

# Rank-based shrinkage estimation for identification in semiparametric additive models

**Jing Yang**[1] · **Hu Yang**[2] · **Fang Lu**[3]

Received: 17 November 2015 / Revised: 2 November 2016 / Published online: 10 February 2017


**Abstract** In this paper, we propose a novel and robust procedure for model identification in semiparametric additive models based on rank regression and spline approximation. Under some mild conditions, we establish the theoretical properties of the identified nonparametric functions and the linear parameters. Furthermore, we demonstrate that the proposed rank estimate has a great efficiency gain across a wide spectrum of non-normal error distributions and almost not lose any efficiency for the normal error compared with that of least square estimate. Even in the worst case scenarios, the asymptotic relative efficiency of the proposed rank estimate versus least squares estimate, which is show to have an expression closely related to that of the signed-rank Wilcoxon test in comparison with the t-test, has a lower bound equal to 0.864. Finally, an efficient algorithm is presented for computation and the selections of tuning parameters are discussed. Some simulation studies and a real data analysis are conducted to illustrate the finite sample performance of the proposed method.

**Keywords** Semiparametric additive model · Model identification · Rank regression · B-spline · Robustness · Asymptotic relative efficiency

✉ Jing Yang
yang2009jing@163.com

[1] College of Mathematics and Computer Science, Key Laboratory of High Performance Computing and Stochastic Information Processing (Ministry of Education of China), Hunan Normal University, Changsha 410081, China

[2] College of Mathematics and Statistics, Chongqing University, Chongqing 401331, China

[3] College of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067, China

## 1 Introduction

Consider the following additive regression model

$$Y_i = u + \sum_{j=1}^{p} f_{0j}(X_{ij}) + \varepsilon_i, \tag{1}$$

where $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ is a $p$-dimensional covariate, $\{f_{0j}(\cdot), j = 1, 2, \ldots, p\}$ are unknown smooth functions satisfying $\mathrm{E}\{f_{0j}(X_{ij})\} = 0$ for the sake of model identifiability, and $\varepsilon_i$ is the random error independent of $X_i$. There exist at least two benefits of such an additive approximation. First, the additive combination of univariate functions can be more interpretable and easier to fit than the joint multivariate nonparametric models. Second, the so-called "curse of dimensionality" that besets multivariate nonparametric regression is largely circumvented because every individual additive component can be estimated using a univariate smoother via an iterative manner. Therefore, large amounts of studies have been done under this model due to its superior characteristics, and we refer, for instance, to Yu and Lu (2004), Mammen and Park (2006), Yu et al. (2008), Xue (2009) and Lian (2012a, b).

Although model (1) owns some wonderful properties, Opsomer and Ruppert (1999) noticed that in practice some covariates may have linear or even no effects on the response variable while other covariates enter nonlinearly, and recommended the so-called semiparametric additive model (SPAM) with the form

$$Y_i = u + \sum_{j=1}^{p_0} f_{0j}(X_{ij}) + \sum_{j=p_0+1}^{p} X_{ij}\beta_{0j} + \varepsilon_i. \tag{2}$$

Statistically, the SPAM could be more parsimonious than the general additive model in some cases, and hence attracted considerable attention. For related literature, see Härdle et al. (2004), Deng and Liang (2010), Liu et al. (2011), Wei and Liu (2012), Wei et al. (2012) among others. Nevertheless, all these works for SPAM are based on the assumption that the linear and nonlinear part are known in advance, which is not always true in practice. If the structure is misspecified, it can not only increase complexity of model but also reduce the estimation accuracy. Since the optimal parametric estimation rate is $n^{-1/2}$ and the optimal nonparametric estimation rate is $n^{-2/5}$, treating a parametric component as a nonparametric component can over-fit the data and leads to efficiency loss. Therefore, model identification is important to model (1), and it is of great interest to develop some efficient methods to distinguish nonzero components as well as linear components from nonlinear ones.

In general, this goal could be achieved by conducting some hypothesis testing as done in Jiang et al. (2007), whereas it might be cumbersome to perform in practice when there are more than just a few predictors to test. Besides, the theoretical properties of such identifications based on hypothesis testing can be somewhat hard to analyze. To this end, Huang et al. (2010) presented a new type of usage for the SCAD penalty as well as its related methods and successfully applied it to nonparametric additive models for the purpose of identifying zero components and parametric components. Following a similar idea, Zhang et al. (2011) simultaneously identified the

zero and linear components of partially linear models by using two penalty functions through an elegant mathematical framework; Lian (2012a) provided a way to determine linear components of additive models based on least square (LS) regression; Lian (2012b) successfully identified nonzero and linear components of model (1) in conditional quantile regression; Wang and Song (2013) applied the SCAD penalty to identify the model structure in semiparametric varying coefficient partially linear models. Note that all these papers were built on either LS regression, which is very sensitive and has low efficiency with respect to many commonly used non-normal errors, or quantile regression, for which the efficiency is proportional to the density at the median. Hence, it would be highly desirable to develop an efficient and robust method that can simultaneously conduct model identification and estimation.

Recently, Wang et al. (2009) proposed a novel procedure for the varying coefficient model based on rank regression and demonstrated that the new method is highly efficient across a wide class of error distributions and possesses comparable efficiency in the worst case scenario compared with LS regression. Similar conclusions on rank regression have been further confirmed in Leng (2010), Sun and Lin (2014), Feng et al. (2015) and the references therein. To the best known of our knowledge, none of these approaches has been studied in SPAM. Therefore, motivated by these observations, this paper is devoted to extending the rank regression to SPAM for identifying nonzero components as well as linear components. Specifically, we firstly embed the SPAM into an additive model and use the spline method to approximate unknown functions. A two-fold SCAD penalty is then employed to discriminate the nonzero components as well as linear components from the nonlinear ones by penalizing both the coefficient functions and their second derivatives. Furthermore, the theoretical properties of the estimator are established, and based on the asymptotic theory of the linear components, we show that the proposed rank estimate has a great efficiency gain across a wide spectrum of non-normal error distributions and loses almost no efficiency for the normal error compared with that of the LS estimate. Even in the worst case scenarios, the asymptotic relative efficiency (ARE) of the proposed rank estimate versus LS estimate has a lower bound being 0.864. In addition, it is worth noting that the ARE of the proposed rank estimate versus LS has an expression which is closely related to that of the signed-rank Wilcoxon test in comparison with the t-test.

The rest of this paper is organized as follows. In Sect. 2, we introduce our new penalized rank regression method based on basis expansion and the SCAD penalty. In Sect. 3, the asymptotic properties are established under some suitable conditions. The selection of optimal tuning parameters are discussed in Sect. 4 along with a computational algorithm for implementation. Sect. 5 illustrates the finite sample performance of the proposed procedure via some simulation studies, and short concluding remarks are followed in Sect. 6. All the technical proofs are deferred to Appendix.

## 2 Rank-based shrinkage regression for additive models

Suppose that $\{X_i, Y_i\}_{i=1}^n$ is an independent and identically distributed sample from model (2). Without loss of generality, we assume that the distribution of $X_i$ is supported on [0,1]. As we do not know which covariates have linear effects in advance, all $p$

components are considered as nonparametric and the polynomial splines are applied to approximate the components. Let $0 = \xi_0 < \xi_1 < \cdots < \xi_{K_n} < \xi_{K_n+1} = 1$ be a partition of [0,1] into $K_n + 1$ subintervals $[\xi_k, \xi_{k+1})$, $k = 0, 1, \ldots, K_n$, where $K_n$ denotes the number of internal knots that increases with sample size $n$. A polynomial spline of order $q$ is a function whose restriction to each subinterval is a polynomial of degree $q - 1$ and globally $q - 2$ times continuously differentiable on [0,1]. The collection of splines with a fixed sequence of knots has a normalized B-spline basis $\{B_1(x), B_2(x), \ldots, B_{K'}(x)\}$ with $K' = K_n + q$.

Note that the constraint condition $\mathrm{E}\{f_{0j}(X_{ij})\} = 0$ is required for the sake of model identifiability, so we instead focus on the space of spline functions $S_j^0 := \{\hbar : \hbar = \sum_{k=1}^{K} \gamma_{jk} B_{jk}(x), \hbar = \sum_{i=1}^{K} \hbar(X_{ij}) = 0\}$ with centered basis $\{B_{jk}(x) = B_k(x) - \sum_{i=1}^{n} B_k(X_{ij})/n, k = 1, 2, \ldots, K = K' - 1\}$, where $K = K' - 1$ due to the empirical version of the constraint. Then the nonlinear functions in model (1) can be approximated by

$$f_{0j}(x) \approx \sum_{k=1}^{K} \gamma_{jk} B_{jk}(x), \quad j = 1, 2, \ldots, p. \tag{3}$$

For simplicity, we restrict our attention to equally spaced knots, although other regular knot sequences like quasi-uniform or data-driven choices can be considered. It is also possible to specify different values of $K_n$ for each component. However, our choice of the equally spaced knots and the same number of knots for each component allows for a much simpler exposition of our results, and as in most of the literature based on spline methods, it can be shown that similar asymptotic results still hold for different choices of $K_n$ and different knots for each component. Let $\gamma_j = (\gamma_{ji}, \gamma_{j2}, \ldots, \gamma_{jK})^T$ and $B_j(x) = \big(B_{j1}(x), B_{j2}(x), \ldots, B_{jK}(x)\big)^T$. Following the approximation (3), model (1) can be rewritten as

$$Y_i \approx u + \sum_{j=1}^{p} \sum_{k=1}^{K} \gamma_{jk} B_{jk}(X_{ij}) + \varepsilon_i = u + \sum_{j=1}^{p} B_j(X_{ij})^T \gamma_j + \varepsilon_i.$$

Accordingly, the residual for estimating $Y_i$ at $X_i$ is $e_i = Y_i - u - \sum_{j=1}^{p} B_j(X_{ij})^T \gamma_j$.

By applying the technique of rank regression method, we propose the following minimization problem

$$\check{\gamma} = \arg\min_{\gamma} L_n(\gamma) := \frac{1}{n} \sum_{i<j} |e_i - e_j|, \tag{4}$$

where $\gamma = \big(\gamma_1^T, \gamma_2^T, \ldots, \gamma_p^T\big)^T$. Thus the estimated component functions are $\check{f}_j(x) = B_j(x)^T \check{\gamma}_j$. Note that the loss function $L_n(\gamma)$ essentially belongs to a local version of Gini's mean difference, which is a classical measure of concentration or dispersion; see David (1998) for details. In addition, it is worth mentioning that the above rank-based loss function cannot generate the estimate of intercept $u$ because it is canceled

out in $e_i - e_j$, which is an unique feature of using this type of estimate in the present problem. As pointed out in Wang et al. (2009), it is essential to have additional location constraint on the random errors in order to make the intercept identifiable, and they adopted the commonly used constraint that $\varepsilon_i$ has median zero. So following the same constraint on $\varepsilon_i$, a reasonable estimate of $u$ can be derived by $\hat{u} = \sum_{i=1}^{n} Y_i/n$ at the rate of $1/\sqrt{n}$, which is faster than any rate of convergence for nonparametric function estimation. Thus for notational convenience, one can safely assume $u = 0$, just as we done in the sequel.

Recall that we are interested in finding the zero components and linear components of model (1). Empirically, the former can be done by shrinking the function $\|f_j\|$ to zero, and the latter can be achieved via shrinking the second derivative $\|f_j''\|$ to zero because a function is linear if and only if it has a second derivative identically zero. Therefore, instead of (4), we consider the following two-fold penalization procedure

$$\hat{\gamma} = \arg\min_{\gamma} L_n^{\lambda}(\gamma) := \frac{1}{n}\sum_{i<j}|e_i - e_j| + n\sum_{k=1}^{p} p_{\lambda_1}(\|f_k\|) + n\sum_{k=1}^{p} p_{\lambda_2}(\|f_k''\|), \quad (5)$$

where $p_\lambda(\cdot)$ is the SCAD penalty function defined by its first derivative

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

where $\lambda$ is the penalized parameter, $a > 2$ is some constant usually taken to be 3.7 as suggested in Fan and Li (2001). Note that the SCAD penalty is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$ but singular at 0, and that its derivative vanishes outside $[-a\lambda, a\lambda]$. These features of SCAD penalty result in a solution with three desirable properties including unbiasedness, sparsity and continuity, which were defined in Fan and Li (2001).

Note that $\|f_j(x)\|^2 = \|B_j(x)^T\gamma_j\|^2 = \int\left(\sum_{k=1}^{K}\gamma_{jk}B_{jk}(x)\right)\left(\sum_{k'=1}^{K}\gamma_{jk'}B_{jk'}(x)\right)dx$ and $\|f_j''(x)\|^2 = \int\left(\sum_{k=1}^{K}\gamma_{jk}B_{jk}''(x)\right)\left(\sum_{k'=1}^{K}\gamma_{jk'}B_{jk'}''(x)\right)dx$, so $\|f_j(x)\|$ and $\|f_j''(x)\|$ can be equivalently expressed as $\sqrt{\gamma_j^T D_j \gamma_j}$ and $\sqrt{\gamma_j^T E_j \gamma_j}$ respectively, where $D_j, E_j \in R^{K\times K}$ with its $(k, k')$ entry equaling to $\int B_{jk}(x)B_{jk'}(x)dx$ and $\int B_{jk}''(x)B_{jk'}''(x)dx$, respectively. Then, the above minimization problem (5) is equivalent to

$$\hat{\gamma} = \arg\min_{\gamma} L_n^{\lambda}(\gamma) := \frac{1}{n}\sum_{i<j}|e_i - e_j| + n\sum_{k=1}^{p} p_{\lambda_1}\left(\sqrt{\gamma_k^T D_k \gamma_k}\right) + n\sum_{k=1}^{p} p_{\lambda_2}\left(\sqrt{\gamma_k^T E_k \gamma_k}\right). \quad (6)$$

Consequently, the estimated component functions are given by $\hat{f}_j(x) = B_j(x)^T\hat{\gamma}_j$.

## 3 Theoretical properties

### 3.1 Asymptotic properties

Without loss of generality, we assume that $f_{0j}$ is truly nonparametric for $j = 1, 2, \ldots, p_0$, linear for $j = p_0 + 1, p_0 + 2, \ldots, s$ with the true slope parameters for the parametric components are denoted by $\beta_0 = (\beta_{0,p_0+1}, \beta_{0,p_0+2}, \ldots, \beta_{0,s})$, and zero for $j = s + 1, s + 2, \ldots, p$. The vectors $X^{(1)} = (X_1, X_2, \ldots, X_{p_0})^T$ and $X^{(2)} = (X_{p_0+1}, X_{p_0+2}, \ldots, X_s)^T$ correspond to the nonlinear and linear components. Denote as $\mathcal{A}$ the subspace of functions on $R^{p_0}$ with an additive form

$$\mathcal{A} := \{h(x^{(1)}) : h(x^{(1)}) = h_1(x_1) + h_2(x_2) + \ldots + h_{p_0}(x_{p_0}), E\big(h_j(X_j)\big) = 0 \text{ and } E\big(h_j(X_j)^2\big) < \infty\},$$

and $E_{\mathcal{A}}(M)$ the subspace projection of $M$ onto $\mathcal{A}$ in the sense that

$$E\{\big(M - E_{\mathcal{A}}(M)\big)\big(M - E_{\mathcal{A}}(M)\big)\} = \inf_{h\in\mathcal{A}} E\{\big(M - h(X^{(1)})\big)\big(M - h(X^{(1)})\big)\}.$$

Let $h(X^{(1)}) = E_{\mathcal{A}}(X^{(2)})$. Each component of $h(X^{(1)}) = \big(h_{(1)}(X^{(1)}), \ldots, h_{(p-p_0)}(X^{(1)})\big)^T$ can be written in the form $h_{(u)}(x) = \sum_{j=1}^{p_0} h_{(u)j}(x_j)$ for some $h_{(u)j}(x_j) \in S_j^0$. To facilitate our asymptotic analysis, we further make the following regularity assumptions.

(A1) The density function $f(x)$ of $X$ is absolutely continuous and compactly supported. Without loss of generality, assume that the support of $X$ is $[0, 1]^p$. Furthermore, there exist constants $0 < c_1 \le c_2 < \infty$ such that $c_1 \le f(x) \le c_2$ for all $x \in \mathcal{X}$.

(A2) For $g = f_{0j}, 1 \le j \le p_0$ or $g = h_{(u)j}, 1 \le u \le s, 1 \le j \le p_0$, $g$ satisfies a Lipschitz condition of order $r > 1/2$. That is, $|g^{(\lfloor r \rfloor)}(x_1) - g^{(\lfloor r \rfloor)}(x_2)| \le C|x_1 - x_2|^{r-\lfloor r \rfloor}$, where $C$ is a constant, $\lfloor r \rfloor$ denotes the biggest integer strictly smaller than $r$ and $g^{(\lfloor r \rfloor)}$ is the $\lfloor r \rfloor$th derivative of $g$. In addition, the order of the B-spline used satisfies $q \ge r + 2$.

(A3) The matrix $\Sigma = E\{(X^{(2)} - h(X^{(1)})(X^{(2)} - h(X^{(1)})^T\}$ is positive definite.

(A4) The errors $\varepsilon$ has a positive density function $h(x)$ satisfying $\int [h'(x)]^2/h(x)dx < \infty$, which means that $\varepsilon$ has finite Fisher information.

Assumptions (A1)–(A2) are common in the polynomial spline estimation literatures; see for example Huang et al. (2010), Wang and Song (2013), Tang (2015) and Li et al. (2015). It was shown in Li (2000) that the positive definiteness of $\Sigma$ in (A3) is necessary for the identifiability of the model in the case that linear components are specified. Assumption (A4) is a regular condition on the random errors which is the same as those used in works on rank regression such as Wang et al. (2009), Hettmansperger and McKean (2011), Sun and Lin (2014) and Feng et al. (2015).

**Theorem 1** *Suppose that assumptions* (A1)–(A4) *hold. If the number of knots* $K_n \asymp n^{1/(2r+1)}$*, then we have*

$$\|\check{f}_j - f_{0j}\|^2 = O_p\left(n^{\frac{-2r}{2r+1}}\right), \quad j = 1, 2, \ldots, p,$$

*where* $\check{f}_j = B_j^T \check{\gamma}_j$ *is the unpenalized estimate of component function* $f_{0j}$ *with* $\check{\gamma}$ *generated by solving* (4).

Theorem 1 indicates that the nonparametric estimates obtained by our proposed method attain the optimal convergence rates. The following theorem will show that if the tuning parameters $\lambda_1$ and $\lambda_2$ are appropriately specified, we can identify the zero parts and linear parts consistently.

**Theorem 2** *Under the same assumptions of Theorem 1, if* $\max\{\lambda_1, \lambda_2\} \to 0$ *and* $n^{r/(2r+1)} \min\{\lambda_1, \lambda_2\} \to \infty$*, then with probability tending to 1,*

(i) $\|\hat{f}_j - f_{0j}\|^2 = O_p\left(n^{\frac{-2r}{2r+1}}\right)$ *for* $j = 1, 2, \ldots, p$,
(ii) $\hat{f}_j$ *is a linear function for* $j = p_0 + 1, p_0 + 2, \ldots, s$,
(iii) $\hat{f}_j \equiv 0$ *for* $j = s + 1, s + 2, \ldots, p$,

*where* $\hat{f}_j = B_j^T \hat{\gamma}_j$ *is the penalized estimate of component function with* $\hat{\gamma}$ *generated by solving* (5).

Finally, for the linear components, we will show that the estimate of the slope parameter is asymptotically normal.

**Theorem 3** *Under the same assumptions of Theorem 2, we have*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N\left(0, \frac{1}{12\tau^2}\Sigma^{-1}\right), \tag{7}$$

*where* $\Sigma$ *is defined in assumption* (A3) *and* $\tau = \int h(x)^2 dx$.

*Remark 1* Based on the results of Theorem 2 and Theorem 3, we observe that the proposed estimate enjoys an oracle property in the sense that it is asymptotically the same as the oracle estimate which is obtained when the true model is known in advance.

### 3.2 Asymptotic relative efficiency

Denote by $\hat{\beta}_{LS}$ and $\hat{\beta}_{RR}$ the estimates of $\beta_0$ generated by LS regression in Lian (2012a) and our proposed rank regression, respectively. To measure the efficiency, we consider the asymptotic variance of the estimates $\hat{\beta}_{LS}$ and $\hat{\beta}_{RR}$ since they all asymptotically unbiased. Hence, based on the asymptotic distribution of $\beta_0$ presented by Theorem 3 in Lian (2012a) and (7) of Theorem 3, we obtain the following theorem.

**Theorem 4** *The* ARE *of the rank-based estimate* $\hat{\beta}_{RR}$ *to the LS estimate* $\hat{\beta}_{LS}$ *for linear parameter* $\beta_0$ *is*

$$\text{ARE}(\hat{\beta}_{RR}, \hat{\beta}_{LS}) = \frac{\text{Var}(\hat{\beta}_{LS})}{\text{Var}(\hat{\beta}_{RR})} = 12\sigma^2\tau^2,$$

*where* $\sigma^2 = \text{E}(\varepsilon^2)$. *This* ARE *has a lower bound of 0.864 for estimating the parameter component, which is attained at the random error density* $h(x) = \frac{3}{20\sqrt{5}}(5-x^2)I(|x| \leq 5)$.

Note that the above obtained ARE 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 the normal error distribution, and can be significantly higher than 1 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.

## 4 Algorithm implementation and tuning parameters selections

In this section we first present an iterative estimation procedure for computation by employing locally quadratic approximation (LQA, Fan and Li , 2001) to the rank-based objective function $L_n(\gamma)$ as well as the two penalty functions $p_{\lambda_1}(\cdot)$ and $p_{\lambda_2}(\cdot)$. Then we discuss the selections of extra parameters including the number of interior knots $K_n$ and the tuning parameters $\lambda_1$ and $\lambda_2$.

### 4.1 Algorithm implementation

It is worth noting that the commonly used gradient-based optimization technique is not feasible here for solving (6) due to its irregularity at the origin. According to Sievers and Abebe (2004), we approximate the unpenalized $L_n(\gamma)$ by

$$L_n(\gamma) \approx \frac{1}{n}\sum_{i=1}^{n} w_i(e_i - \varsigma)^2,$$

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

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

with $R(e_i)$ being the rank of $e_i$ among $\{e_i\}_{i=1}^n$.

On the other hand, following Fan and Li (2001), we apply LQA to the last two penalty terms. That is, for a given initial estimate $\hat{\gamma}_j^{(0)}$, the corresponding weights $w_i^{(0)}$

and the median of residual $\varsigma^{(0)}$ can be obtained. If $\hat{f}_j^{(0)}$ ($\hat{f}_j^{(0)\prime\prime}$) is very close to 0, then set $\hat{f}_j = 0$ ($\hat{f}_j'' = 0$). Otherwise, we have

$$p_{\lambda_1}(\|f_j\|) \approx p_{\lambda_1}(\|\gamma_j^{(0)}\|_{D_j}) + \frac{1}{2}\frac{p'_{\lambda_1}(\|\gamma_j^{(0)}\|_{D_j})}{\|\gamma_j^{(0)}\|_{D_j}}\{\|\gamma_j\|_{D_j}^2 - \|\gamma_j^{(0)}\|_{D_j}^2\},$$

and

$$p_{\lambda_2}(\|f_j''\|) \approx p_{\lambda_2}(\|\gamma_j^{(0)}\|_{E_j}) + \frac{1}{2}\frac{p'_{\lambda_2}(\|\gamma_j^{(0)}\|_{E_j})}{\|\gamma_j^{(0)}\|_{E_j}}\{\|\gamma_j\|_{E_j}^2 - \|\gamma_j^{(0)}\|_{E_j}^2\},$$

where $\|\gamma_j\|_{D_j} = \sqrt{\gamma_j D_j \gamma_j}$ and $\|\gamma_j\|_{E_j} = \sqrt{\gamma_j E_j \gamma_j}$. Ignoring the irrelevant constants, (6) is equivalent to minimize the following quadratic function

$$\begin{aligned} Q_n^{\lambda}(\gamma) := \frac{1}{n}\sum_{i=1}^{n} w_i(e_i - \varsigma)^2 + \frac{n}{2}\sum_{k=1}^{p}\frac{p'_{\lambda_1}(\|\gamma_k^{(0)}\|_{D_k})}{\|\gamma_k^{(0)}\|_{D_k}}\gamma_k D_k \gamma_k + \\ + \frac{n}{2}\sum_{k=1}^{p}\frac{p'_{\lambda_2}(\|\gamma_k^{(0)}\|_{E_k})}{\|\gamma_k^{(0)}\|_{E_k}}\gamma_k E_k \gamma_k. \end{aligned}$$

To make the expression convenient, we introduce the following notations

$$\tilde{Y}^{(m)} = Y - \varsigma^{(m)}, \quad W^{(m)} = \operatorname{diag}\left\{w_1^{(m)}, w_2^{(m)}, \ldots, w_n^{(m)}\right\},$$

$$\Sigma_{\lambda_1}(\gamma^{(m)}) = \operatorname{diag}\left\{\frac{p'_{\lambda_1}(\|\gamma_j^{(m)}\|_{D_1})}{\|\gamma_j^{(m)}\|_{D_1}}, \frac{p'_{\lambda_1}(\|\gamma_j^{(m)}\|_{D_2})}{\|\gamma_j^{(m)}\|_{D_2}}, \ldots, \frac{p'_{\lambda_1}(\|\gamma_j^{(m)}\|_{D_p})}{\|\gamma_j^{(m)}\|_{D_p}}\right\},$$

$$\Sigma_{\lambda_2}(\gamma^{(m)}) = \operatorname{diag}\left\{\frac{p'_{\lambda_2}(\|\gamma_j^{(m)}\|_{E_1})}{\|\gamma_j^{(m)}\|_{E_1}}, \frac{p'_{\lambda_2}(\|\gamma_j^{(m)}\|_{E_2})}{\|\gamma_j^{(m)}\|_{E_2}}, \ldots, \frac{p'_{\lambda_2}(\|\gamma_j^{(m)}\|_{E_p})}{\|\gamma_j^{(m)}\|_{E_p}}\right\}.$$

Therefore, the computational algorithm can be implemented as follows:

Step 0: Choose the unpenalized estimate $\check{\gamma}$ as the initial estimate $\hat{\gamma}^{(0)}$ and let $\hat{\gamma}^{(m)} = \hat{\gamma}^{(0)}$.
Step 1: Update $\gamma^{(m)}$to obtain $\gamma^{(m+1)}$ by

$$\gamma^{(m+1)} = \arg\min_{\gamma} Q_n^{\lambda}(\gamma) = \left\{Z^T W^{(m)} Z + \frac{n^2}{2}\Sigma_{\lambda_1}(\gamma^{(m)}) + \frac{n^2}{2}\Sigma_{\lambda_2}(\gamma^{(m)})\right\}^{-1} Z^T W^{(m)} \tilde{Y}^{(m)},$$

where $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^T$, $Z = (Z_1, \ldots, Z_n)^T$ with $Z_i = \left(B_1(X_{i1})^T, \ldots, B_p(X_{ip})^T\right)^T$.

Step 2: Set $m = m + 1$ and return back to Step 1.
Step 3: Iterate Step 1 and Step 2 until convergence.

*Remark 2* As a stopping rule to check the convergence of $\hat{\gamma}$ in above estimation procedure, we propose to stop the iteration when the change in $\hat{\gamma}$ *between the* $i$*-th* and $(i+1)$-th iteration is below a pre-specified threshold.

### 4.2 Extra parameters selections

To achieve good numerical performance, one needs to choose the number of interior knots $K_n$ and the tuning parameters $\lambda_1$ and $\lambda_2$ appropriately. Here we fix the spline order to be 4, which means that cubic splines are used in all our numerical implementations. Then we use 5-fold cross-validation (CV) to select $K_n$ as well as $\lambda = (\lambda_1, \lambda_2)^T$ simultaneously. To be more specific, we randomly divide the data into five roughly equal parts, denoted as $\{(X_i^T, Y_i)^T,\ i \in S(j)\}$ for $j = 1, 2, \ldots, 5$, where $S(j)$ is the set of subject indices corresponding to the $j$th part. For each $j$, we treat $\{(X_i^T, Y_i)^T,\ i \in S(j)\}$ as the validation data set, and the remaining four parts of data as the training data set. For any candidate $(K_n, \lambda^T)^T$, for each $i \in S(j)$, we apply local polynomial fitting to the training data set to estimate $\{f_{0k}(\cdot)\}_{k=1}^p$ by solving (5). After we get the estimates $\{\hat{f}_k(\cdot)\}_{k=1}^p$ for all $i \in S(j)$, we can calculate the corresponding prediction $\hat{Y}_i = \sum_{k=1}^p \hat{f}_k(X_{ik})$. Then the cross validation error corresponding to a fixed $(K_n, \lambda^T)^T$ is defined as

$$CV_5(K_n, \lambda) = \sum_{j=1}^{5} \sum_{i \in S(j)} \left\{ \frac{R(e_i(\hat{f}))}{n+1} - \frac{1}{2} \right\} e_i(\hat{f}), \tag{8}$$

where $e_i(\hat{f}) = Y_i - \sum_{j=1}^p \hat{f}_j(X_{ij})$ and $R(e_i(\hat{f}))$ represents the rank of $e_i(\hat{f})$ among $\{e_i(\hat{f})\}_{i=1}^n$. Finally, the optimal $K_n$ and $\lambda$ are selected by minimizing the cross validation error $CV_5(K_n, \lambda)$.

*Remark 3 As stated in* Feng et al. (2015), *the variable selection results are hardly affected by the choice of selection procedure for* $K_n$. *Therefore, to reduce the computation burden, one may firstly fit the additive model* (1) *without any penalization and use the above 5-fold cross validation to select an optimal* $K_n$, *and then fix the same* $K_n$ *in* (8) *to select the optimal* $\lambda$.

## 5 Numerical examples

### 5.1 Monte Carlo simulation

We generate our sample from the following additive model:

$$Y_i = \sum_{j=1}^{10} f_{0j}(X_{ij}) + 0.3\varepsilon_i, \tag{9}$$

**Table 1** Component selection results with $n = 100$

| Dist. | Method | NN | NNT | NL | NLT |
|---|---|---|---|---|---|
| N(0,1) | LS | 2.62 (1.193) | 2 (0) | 3.86 (1.423) | 3.655 (1.266) |
| | QR | 2.825 (1.416) | 2 (0) | 3.755 (1.642) | 3.54 (1.533) |
| | RR | 2.68 (1.253) | 2 (0) | 3.845 (1.495) | 3.645 (1.304) |
| | CQR | 2.66 (1.272) | 2 (0) | 3.85 (1.501) | 3.645 (1.309) |
| t(3) | LS | 3.185 (1.907) | 2 (0) | 3.43 (2.149) | 3.16 (1.954) |
| | QR | 2.875 (1.495) | 2 (0) | 3.71 (1.683) | 3.48 (1.576) |
| | RR | 2.635 (1.318) | 2 (0) | 3.835 (1.492) | 3.63 (1.365) |
| | CQR | 2.64 (1.337) | 2 (0) | 3.83 (1.512) | 3.62 (1.378) |
| MN | LS | 3.29 (2.176) | 1.97 (0.222) | 3.27 (2.314) | 2.985 (2.051) |
| | QR | 2.86 (1.576) | 2 (0) | 3.725 (1.670) | 3.485 (1.518) |
| | RR | 2.65 (1.353) | 2 (0) | 3.83 (1.541) | 3.61 (1.386) |
| | CQR | 2.645 (1.401) | 2 (0) | 3.82 (1.568) | 3.605 (1.394) |
| LN | LS | 3.315 (2.209) | 1.955 (0.253) | 3.115 (2.327) | 2.97 (2.104) |
| | QR | 2.91 (1.638) | 2 (0) | 3.715 (1.724) | 3.46 (1.613) |
| | RR | 2.65 (1.384) | 2 (0) | 3.825 (1.583) | 3.615 (1.407) |
| | CQR | 2.63 (1.392) | 2 (0) | 3.805 (1.609) | 3.61 (1.415) |
| Exp(1) | LS | 3.23 (2.064) | 1.985 (0.159) | 3.385 (2.230) | 3.095 (1.986) |
| | QR | 2.895 (1.612) | 2 (0) | 3.71 (1.687) | 3.475 (1.548) |
| | RR | 2.645 (1.330) | 2 (0) | 3.835 (1.517) | 3.62 (1.371) |
| | CQR | 2.645 (1.359) | 2 (0) | 3.825 (1.534) | 3.62 (1.383) |

Enclosed in parentheses are the corresponding standard errors

where $f_{01}(x) = \sin(2\pi x)$, $f_{02}(x) = 6x(1-x)$, $f_{03}(x) = 2x$, $f_{04}(x) = x$, $f_{05}(x) = -x$, $f_{06}(x) = -2x$ and $f_{0j}(x) \equiv 0$ for $j = 7, \ldots, 10$. Thus the number of nonparametric components is 2 and the number of nonzero linear components is 4. The covariates $X_i = (X_{i1}, X_{i2}, \ldots, X_{i10})^T$ are generated from the standard normal distribution with the correlation between $X_{ij_1}$ and $X_{ij_2}$ being $0.5^{|j_1 - j_2|}$. A similar model setting was also applied in Lian (2012a) without the last four zero functions because they only consider model identification for the linear components. Beforehand, we apply the cumulative distribution function of standard normal distribution to transform $X_{ij}$ to be marginally uniform on [0,1]. Finally, four different methods including Lian (2012a) (LS), Lian (2012b) with 0.5th quantile (QR), composite quantile regression (CQR) by Kai et al. (2010) with the number of quantile being 9 and our proposed rank regression (RR) are conducted in this example.

In order to examine the robustness and efficiency of our proposed method, five different error distributions are considered including standard normally distributed N(0,1), t(3) distribution which is heavy-tailed, the mixture of normals 0.9N(0,1) + 0.1N(0,10) (MN) which is used to generate the outliers and two asymmetric errors Log-normal (LN) and Exponential (Exp(1)) distributions. For all scenarios, 200 data sets are generated and the corresponding results with $n = 100$ and $n = 200$ are summarized in Tables 1, 2, 3 and 4. Table 1 and Table 2 report the average number of nonparametric

**Table 2** Component selection results with $n = 200$

| Dist. | Method | NN | NNT | NL | NLT |
|---|---|---|---|---|---|
| N(0,1) | LS | 2.21 (0.558) | 2 (0) | 3.95 (0.654) | 3.95 (0.654) |
| | QR | 2.315 (0.624) | 2 (0) | 3.925 (0.803) | 3.915 (0.753) |
| | RR | 2.215 (0.565) | 2 (0) | 3.945 (0.681) | 3.94 (0.675) |
| | CQR | 2.21 (0.578) | 2 (0) | 3.94 (0.694) | 3.94 (0.694) |
| t(3) | LS | 2.785 (1.289) | 2 (0) | 3.51 (1.508) | 3.32 (1.334) |
| | QR | 2.35 (0.656) | 2 (0) | 3.93 (0.827) | 3.915 (0.760) |
| | RR | 2.22 (0.604) | 2 (0) | 3.94 (0.741) | 3.94 (0.741) |
| | CQR | 2.245 (0.631) | 2 (0) | 3.94 (0.758) | 3.935 (0.736) |
| MN | LS | 2.845 (1.422) | 2 (0) | 3.435 (1.711) | 3.235 (1.463) |
| | QR | 2.37 (0.691) | 2 (0) | 3.925 (0.843) | 3.91 (0.808) |
| | RR | 2.265 (0.636) | 2 (0) | 3.94 (0.764) | 3.935 (0.741) |
| | CQR | 2.27 (0.654) | 2 (0) | 3.935 (0.768) | 3.93 (0.752) |
| LN | LS | 2.86 (1.438) | 1.995 (0.054) | 3.385 (1.753) | 3.185 (1.501) |
| | QR | 2.415 (0.734) | 2 (0) | 3.915 (0.859) | 3.90 (0.847) |
| | RR | 2.28 (0.651) | 2 (0) | 3.925 (0.774) | 3.92 (0.755) |
| | CQR | 2.30 (0.663) | 2 (0) | 3.92 (0.789) | 3.91 (0.768) |
| Exp(1) | LS | 2.795 (1.316) | 2 (0) | 3.48 (1.603) | 3.305 (1.418) |
| | QR | 2.355 (0.672) | 2 (0) | 3.925 (0.824) | 3.915 (0.794) |
| | RR | 2.225 (0.627) | 2 (0) | 3.935 (0.748) | 3.935 (0.748) |
| | CQR | 2.235 (0.634) | 2 (0) | 3.935 (0.760) | 3.93 (0.751) |

Enclosed in parentheses are the corresponding standard errors

components selected (NN), the average number of true nonlinear components selected (NNT), the average number of linear components selected (NL), and the average number of true linear components selected (NLT). Table 3 and Table 4 present the performance of estimates for the first six nonzero component functions by using root mean squared errors (RMSE) defined by $\mathrm{RMSE}_j = \left\{ \frac{1}{n_{grid}} \sum_{i=1}^{n_{grid}} (\hat{f}_j(u_i) - f_{0j}(u_i))^2 \right\}^{1/2}$, where $\{u_i, i = 1, 2, \ldots, n_{grid}\}$ are the grid points at which the function $f_j(\cdot)$ is evaluated.

We make several observations from the results of Tables 1, 2, 3 and 4: (1) Our proposed RR method performs similar to CQR method in most situations; (2) For the normal error, the RR and CQR estimators are comparable to the LS estimator in terms of model selection as well as estimation accuracy, and all above three estimators are much superior to that of QR estimator; (3) For the other four types of error, the performance of LS method is terrible, whereas the RR and CQR approaches possess a significantly higher efficiency than that of QR although they are all robust to error structures in comparison with the LS method; (4) The model identification performance and estimation accuracy of all considered methods improved as the sample size $n$ increasing, which corroborates the theoretical properties. All these conclusions reveal that the CQR method and RR procedure are highly efficient in estimating and

**Table 3** Root mean squared errors for $f_{01}, \ldots, f_{06}$ with $n = 100$

| Dist. | Method | $f_{01}$ | $f_{02}$ | $f_{03}$ | $f_{04}$ | $f_{05}$ | $f_{06}$ |
|---|---|---|---|---|---|---|---|
| N(0,1) | LS | 0.1256 (0.0454) | 0.1009 (0.0449) | 0.0726 (0.0484) | 0.0816 (0.0492) | 0.0735 (0.0479) | 0.0719 (0.0479) |
| | QR | 0.1443 (0.0545) | 0.1165 (0.0558) | 0.0828 (0.0563) | 0.0843 (0.0558) | 0.0860 (0.0565) | 0.0813 (0.0552) |
| | RR | 0.1324 (0.0502) | 0.1094 (0.0513) | 0.0752 (0.0493) | 0.0812 (0.0513) | 0.0784 (0.0519) | 0.0693 (0.0461) |
| | CQR | 0.1342 (0.0507) | 0.1109 (0.0526) | 0.0761 (0.0502) | 0.0827 (0.0526) | 0.0792 (0.0528) | 0.0699 (0.0472) |
| t(3) | LS | 0.2294 (0.1053) | 0.2014 (0.1004) | 0.1313 (0.0966) | 0.1317 (0.0998) | 0.1304 (0.0984) | 0.1376 (0.1048) |
| | QR | 0.1522 (0.0584) | 0.1238 (0.0651) | 0.0855 (0.0636) | 0.0881 (0.0677) | 0.0862 (0.0667) | 0.0873 (0.0634) |
| | RR | 0.1381 (0.0547) | 0.1168 (0.0546) | 0.0822 (0.0539) | 0.0857 (0.0573) | 0.0796 (0.0517) | 0.0791 (0.0568) |
| | CQR | 0.1384 (0.0558) | 0.1173 (0.0562) | 0.0837 (0.0541) | 0.0865 (0.0578) | 0.0803 (0.0525) | 0.0810 (0.0571) |
| MN | LS | 0.2596 (0.1219) | 0.2470 (0.1246) | 0.1479 (0.1085) | 0.1500 (0.1087) | 0.1480 (0.1068) | 0.1352 (0.1067) |
| | QR | 0.1561 (0.0629) | 0.1344 (0.0701) | 0.0890 (0.0722) | 0.0907 (0.0705) | 0.0949 (0.0724) | 0.0894 (0.0646) |
| | RR | 0.1380 (0.0553) | 0.1182 (0.0578) | 0.0859 (0.0603) | 0.0846 (0.0561) | 0.0834 (0.0534) | 0.0829 (0.0593) |
| | CQR | 0.1386 (0.0562) | 0.1197 (0.0591) | 0.0875 (0.0617) | 0.0852 (0.0573) | 0.0839 (0.0541) | 0.0833 (0.0586) |
| LN | LS | 0.2643 (0.1258) | 0.2496 (0.1277) | 0.1538 (0.1124) | 0.1541 (0.1105) | 0.1492 (0.1094) | 0.1387 (0.1090) |
| | QR | 0.1570 (0.0643) | 0.1352 (0.0729) | 0.0911 (0.0735) | 0.0928 (0.0727) | 0.0964 (0.0738) | 0.0915 (0.0672) |
| | RR | 0.1385 (0.0561) | 0.1190 (0.0593) | 0.0869 (0.0615) | 0.0861 (0.0592) | 0.0841 (0.0553) | 0.0835 (0.0607) |
| | CQR | 0.1391 (0.0565) | 0.1201 (0.0602) | 0.0884 (0.0627) | 0.0862 (0.0585) | 0.0849 (0.0561) | 0.0841 (0.0608) |
| Exp(1) | LS | 0.2359 (0.1134) | 0.2138 (0.1129) | 0.1376 (0.0997) | 0.1421 (0.1005) | 0.1382 (0.998) | 0.1357 (0.1061) |
| | QR | 0.1536 (0.0631) | 0.1261 (0.0678) | 0.0874 (0.0692) | 0.0897 (0.0689) | 0.0906 (0.0715) | 0.0882 (0.0649) |
| | RR | 0.1383 (0.0551) | 0.1179 (0.0574) | 0.0841 (0.0567) | 0.0853 (0.0587) | 0.0813 (0.0538) | 0.0811 (0.0594) |
| | CQR | 0.1387 (0.0564) | 0.1186 (0.0585) | 0.0852 (0.0574) | 0.0859 (0.0581) | 0.0820 (0.0536) | 0.0819 (0.0588) |

Enclosed in parentheses are the corresponding standard errors

**Table 4** Root mean squared errors for $f_{01}, \ldots, f_{06}$ with $n = 200$

| Dist. | Method | $f_{01}$ | $f_{02}$ | $f_{03}$ | $f_{04}$ | $f_{05}$ | $f_{06}$ |
|---|---|---|---|---|---|---|---|
| N(0,1) | LS | 0.0610 (0.0189) | 0.0467 (0.0186) | 0.0236 (0.0143) | 0.0240 (0.0100) | 0.0226 (0.0135) | 0.0228 (0.0158) |
| | QR | 0.0681 (0.0227) | 0.0549 (0.0211) | 0.0264 (0.0175) | 0.0267 (0.0185) | 0.0262 (0.0182) | 0.0263 (0.0182) |
| | RR | 0.0638 (0.0199) | 0.0457 (0.0178) | 0.0244 (0.0145) | 0.0248 (0.0129) | 0.0237 (0.0147) | 0.0234 (0.0151) |
| | CQR | 0.0643 (0.0205) | 0.0462 (0.0196) | 0.0257 (0.0163) | 0.0258 (0.0136) | 0.0244 (0.0159) | 0.0239 (0.0164) |
| t(3) | LS | 0.1506 (0.0605) | 0.1242 (0.0629) | 0.0831 (0.0597) | 0.0842 (0.0594) | 0.0845 (0.0607) | 0.0823 (0.0577) |
| | QR | 0.0741 (0.0251) | 0.0510 (0.0227) | 0.0239 (0.0206) | 0.0286 (0.0221) | 0.0280 (0.0226) | 0.0296 (0.0214) |
| | RR | 0.0690 (0.0218) | 0.0493 (0.0195) | 0.0251 (0.0172) | 0.0232 (0.0164) | 0.0271 (0.0169) | 0.0207 (0.0158) |
| | CQR | 0.0703 (0.0225) | 0.0491 (0.0203) | 0.0247 (0.0174) | 0.0238 (0.0173) | 0.0278 (0.0175) | 0.0214 (0.0163) |
| MN | LS | 0.1618 (0.0645) | 0.1261 (0.0623) | 0.0877 (0.0637) | 0.0851 (0.0635) | 0.0860 (0.0650) | 0.0875 (0.0674) |
| | QR | 0.0752 (0.0259) | 0.0531 (0.0250) | 0.0268 (0.0238) | 0.0299 (0.0251) | 0.0289 (0.0254) | 0.0290 (0.0237) |
| | RR | 0.0704 (0.0227) | 0.0497 (0.0193) | 0.0290 (0.0186) | 0.0264 (0.0169) | 0.0276 (0.0173) | 0.0272 (0.0180) |
| | CQR | 0.0711 (0.0232) | 0.0503 (0.0205) | 0.0302 (0.0191) | 0.0271 (0.0174) | 0.0281 (0.0185) | 0.0279 (0.0184) |
| LN | LS | 0.1683 (0.0671) | 0.1294 (0.0652) | 0.0879 (0.0685) | 0.0891 (0.0663) | 0.0907 (0.0681) | 0.0912 (0.0695) |
| | QR | 0.0761 (0.0284) | 0.0539 (0.0265) | 0.0298 (0.0249) | 0.0306 (0.0258) | 0.0294 (0.0271) | 0.0299 (0.0268) |
| | RR | 0.0709 (0.0235) | 0.0507 (0.0204) | 0.0312 (0.0197) | 0.0275 (0.0184) | 0.0283 (0.0198) | 0.0274 (0.0196) |
| | CQR | 0.0715 (0.0247) | 0.0515 (0.0213) | 0.0307 (0.0199) | 0.0278 (0.0192) | 0.0279 (0.0206) | 0.0282 (0.0201) |
| Exp(1) | LS | 0.1574 (0.0625) | 0.1258 (0.0634) | 0.0863 (0.0612) | 0.0852 (0.0609) | 0.0864 (0.0647) | 0.0871 (0.0619) |
| | QR | 0.0748 (0.0264) | 0.0510 (0.0238) | 0.0246 (0.0217) | 0.0293 (0.0245) | 0.0288 (0.0259) | 0.0294 (0.0230) |
| | RR | 0.0697 (0.0219) | 0.0502 (0.0223) | 0.0258 (0.0191) | 0.0245 (0.0176) | 0.0272 (0.0177) | 0.0243 (0.0172) |
| | CQR | 0.0705 (0.0226) | 0.0511 (0.0236) | 0.0265 (0.0197) | 0.0249 (0.0183) | 0.0277 (0.0184) | 0.0251 (0.0179) |

Enclosed in parentheses are the corresponding standard errors

**Table 5** Estimation and model identification results with LASSO, ALASSO and MCP

| Dist. | Method | NNT | NLT | PC | RMSE($f$) |
|---|---|---|---|---|---|
| t(3) | RR-LASSO | 1.38 (0.744) | 2.815 (1.674) | 0.524 (0.461) | 1.374 (0.753) |
| | RR-ALASSO | 2 (0) | 3.94 (0.738) | 0.98 (0.102) | 0.105 (0.029) |
| | RR-MCP | 2 (0) | 3.94 (0.746) | 0.98 (0.097) | 0.112 (0.032) |
| Exp(1) | RR-LASSO | 1.275 (0.838) | 2.69 (1.806) | 0.511 (0.478) | 1.532 (0.816) |
| | RR-ALASSO | 2 (0) | 3.935 (0.756) | 0.975 (0.104) | 0.119 (0.033) |
| | RR-MCP | 2 (0) | 3.935 (0.749) | 0.98 (0.107) | 0.106 (0.038) |

Enclosed in parentheses are the corresponding standard errors

**Table 6** Component selection results in Boston housing price data

| Variable | LS | QR | RR | Variable | LS | QR | RR |
|---|---|---|---|---|---|---|---|
| crim | 2 | 1 | 1 | dis | 2 | 2 | 1 |
| zn | 0 | 0 | 0 | rad | 2 | 0 | 0 |
| indus | 0 | 0 | 0 | tax | 2 | 2 | 2 |
| nox | 2 | 2 | 2 | ptratio | 2 | 1 | 2 |
| rm | 2 | 2 | 2 | black | 2 | 2 | 1 |
| age | 0 | 1 | 0 | lstat | 2 | 2 | 2 |

identifying nonzero components as well as simultaneously discriminating linear components from nonlinear ones, and they are robust and adaptive to different errors. However, it is worth noting that in contrast with the CQR method whose performance depends on the choice of the number of quantiles to combine, a meta parameter which plays a vital role in balancing the performance of LS and absolute deviation-based methods, our proposed RR procedure does not need to choose the meta parameter. This characteristic can reduce the burden of calculation.

Note that, according to the anonymous reviewers's valuable suggestions, we have added some simulations to evaluate the performance of our proposed RR method under the penalties of lasso, Adaptive-lasso and MCP. The results based on 200 samples are reported in Table 5, where RMSE($f$) stands for the root mean squared errors of $f$ with $f = \sum_{j=1}^{10} f_{0j}$. We can obtain from these results that the performances under Adaptive-lasso and MCP are similar, and they all have a significant superiority to the lasso penalty which has a bad performance. This is expected because Adaptive-lasso and MCP have been demonstrated to own consistency of model selection but lasso does not have. In addition, we have conducted some other simulations under a relatively heavier sparsity by choosing 21 functions, in which the first 6 functions are the same as in model (9) and the last 15 functions are 0. From our obtained results we observe that the performances in the case of heavier sparsity are similar to the case originally considered in model (9). Thus, we omit presenting the corresponding results although they are obtained so as to reduce the length of this paper.

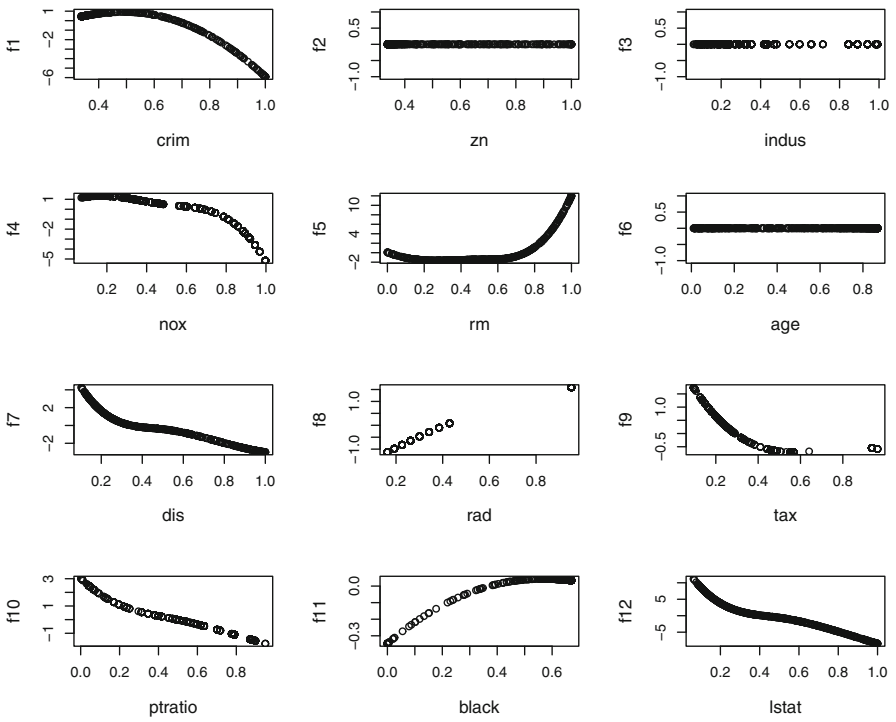



**Fig. 1** The selected components and their fits in for Boston housing price data based on LS method

### 5.2 Application to Boston housing price data

In this section, we consider an application of our proposed method to Boston housing price data, which has been analyzed by Yu and Lu (2004) and Xue (2009) among others. We take the median value of owner-occupied homes in $1000's (*medv*) as the response variable. The covariate variables include per capita crime rate by town (*crim*), proportion of residential land zoned for lots over 25,000 sq.ft (*zn*), proportion of non-retail business acres per town (*indus*), nitric oxides concentration per 10 million (*nox*), average number of rooms per dwelling (*rm*), proportion of owner-occupied units built prior to 1940 (*age*), weighted distances to five Boston employment centers (*dis*), index of accessibility to radial highways (*rad*), full-value property tax per $10,000 (*tax*), pupil-teacher ratio by town (*ptratio*), a parabolic function of the relative size of the Black population in the town (*black*), and percentage of lower status of the population (*lstat*). Beforehand, all the covariate variables are standardized so that they have mean zero and unit variance, and the cumulative distribution function of standard normal distribution is employed to transform the covariates to be marginally uniform on [0,1]. Then we apply LS, QR and RR methods to analyze the data set via an additive model stated as (1).

The component selection results are presented in Table 6, in which, 0, 1 and 2 denote the covariates selected as zero, linear and nonlinear components, respectively.

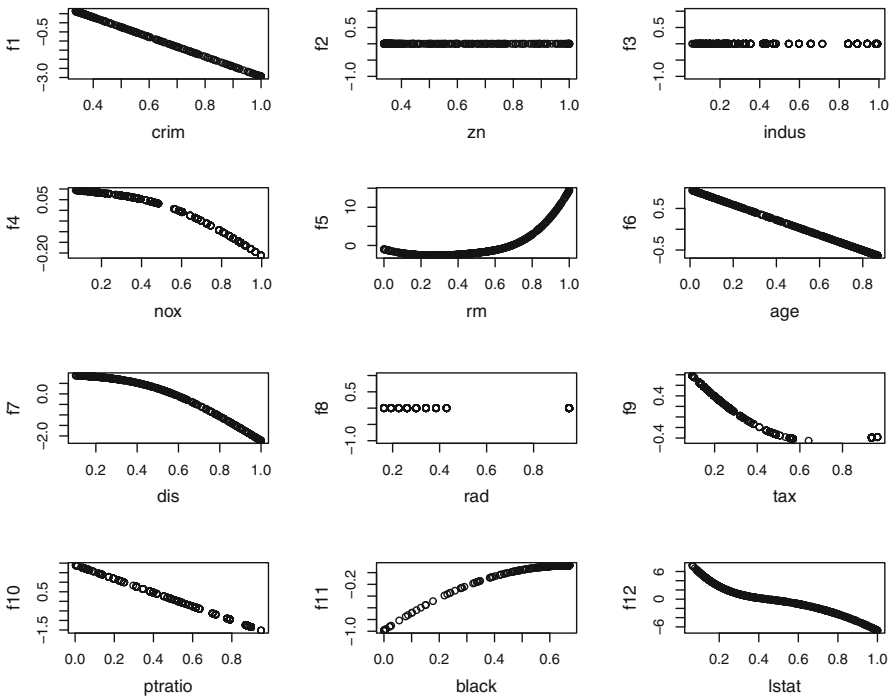



**Fig. 2** The selected components and their fits for the Boston housing price data based on QR method

As we can see from Table 6, all three methods reveal that *rm*, *rad* and *black* have nonnegative effects on house price, which are clearly coincide with the heuristics about their effects on house prices. In addition, compared with the LS approach that removes three covariates *zn*, *indus* and *age* out of the final model as unimportant covariates and identifies the remaining nine covariates as nonlinear components, QR method identified the three covariates *zn*, *indus* and *rad* as zero components, the three covariates *crim*, *age* and *ptratio* as linear components, and the remaining six covariates as nonlinear components. The RR method identified four covariates *zn*, *indus*, *age* and *rad* as zero components, the three covariates *crim*, *dis* and *black* as linear components, and the remaining five covariates as nonlinear components. Similar conclusions can also be derived by the corresponding fits for this data set presented in Figs. 1, 2 and 3. Evidently, our proposed rank approach generates the most parsimonious model among the three considered methods.

For a further study of the applicability of the RR method, we display the normal QQ-plot of the residuals resulted by RR procedure in Fig. 4a, from which we observe that the error term of Boston housing data probably come from a non-normal distribution. Moreover, to compare the performance of the proposed RR procedure with those of LS and QR methods, we give the boxplots of mean absolute prediction error (MAPE) in Fig. 4b, which is obtained based on 200 times simulation with each simulation randomly extract 400 samples. Obviously, RR method performs the best since it has the smallest mean value of MAPE and variance. Consequently, taking into account of

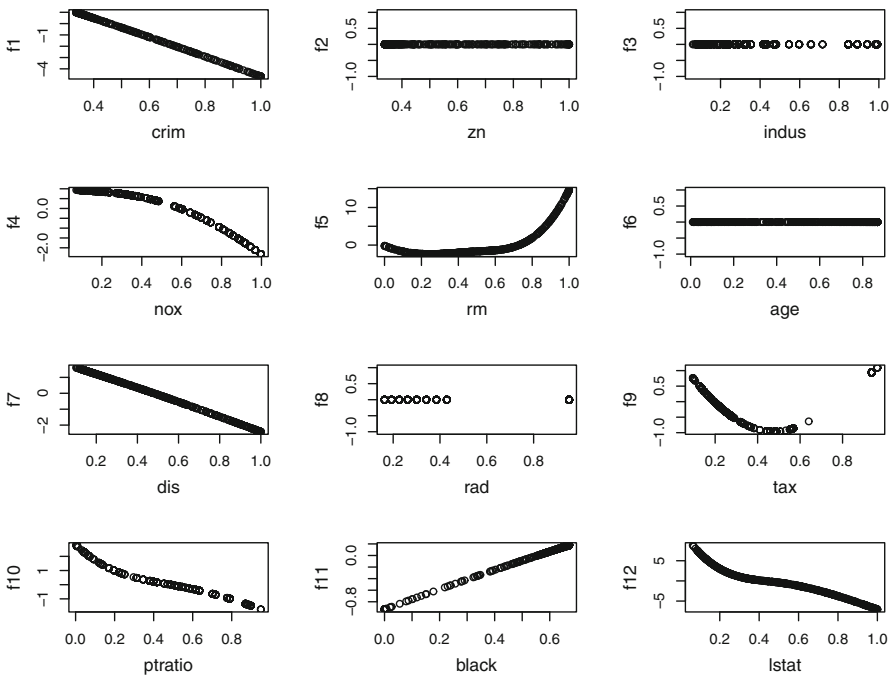
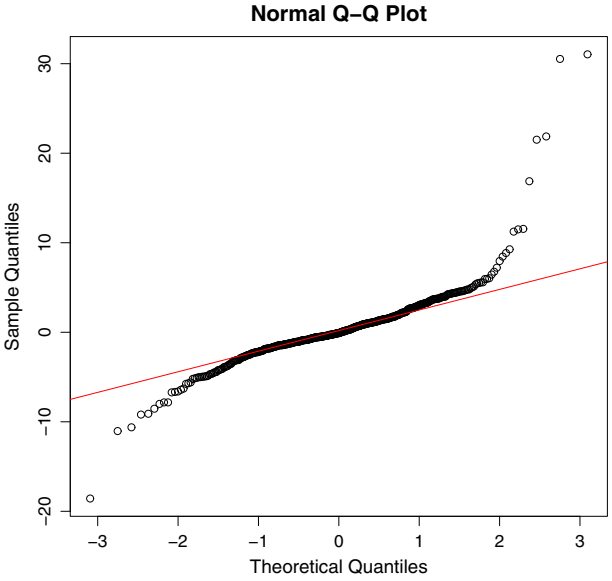



**Fig. 3** The selected components and their fits for the Boston housing price data based on RR method

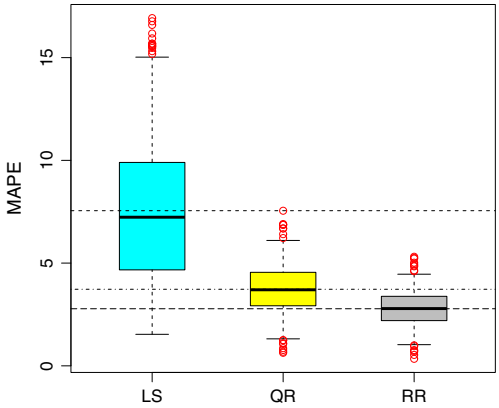
the complexity of model and the performance of prediction, our proposed rank-based regression is a preferred method for analyzing this data set.

## 6 Concluding remarks

In this paper, a novel and robust procedure based on rank regression and spline approximation was developed for model identification in semiparametric additive models. Via adding a two-fold SCAD penalty, the proposed method is able to simultaneously estimate and identify the nonzero components as well as the linear components. Theoretical properties of the estimators of both nonparametric parts and linear parameters were derived under some mild conditions. In addition, we show that the proposed rank estimator is highly efficient across a wide spectrum of error distributions; even in the worst case scenarios, the ARE of the proposed rank estimate versus least squares estimate, is show to have an expression closely related to that of the signed-rank Wilcoxon test in comparison with the t-test, which is equal to 0.864 for the linear parameters. Furthermore, we presented an efficient algorithm for computation and discussed the selections of tuning parameters. To extend our work to a generalized additive model or other nonparametric models seems a promising and useful project for practitioners; we leave it as a future work.




**(a)**




**(b)**

**Fig. 4** **a** is the normal QQ-plot of the residuals resulted by RR method. **b** is the boxplots of MAPE in Boston housing data, where the *dashed*, *dot-dashed* and *long-dashed horizontal lines* represent the average MAPEs based on LS, QR and RR methods, respectively

**Acknowledgements** The authors are grateful to the Editor, Associate Editor and two anonymous referees whose comments lead to a significant improvement of the paper. This work was supported in part by the National Natural Science Foundation of China (Grant No. 11671059).

## Appendix

In the proofs, $C$ denotes a generic constant that might assume different values at different places. Assume $\gamma_0 = (\gamma_{01}^T, \gamma_{02}^T, \ldots, \gamma_{0p}^T)^T$ be a $pK$-dimensional vector

satisfying $\|f_{0j} - B_j^T\gamma_{0j}\| = O_p(K^{-r})$ for $1 \le j \le p_0$ and $f_{0j} = B_j^T\gamma_{0j}$ for $p_0 < j \le p$. In order to prove the theoretical results, we first give some notations for convenience of expression. Let

$$\theta_n = \sqrt{K/n}, \quad \gamma^* = \theta_n^{-1}(\gamma - \gamma_0), \quad Z_i = \left(B_1(X_{i1})^T, \ldots, B_p(X_{ip})^T\right)^T,$$

$$Z_{ij} = Z_i - Z_j, \quad Z = (Z_1, \ldots, Z_n)^T, \quad \Delta_i = \sum_{l=1}^{p} f_{0l}(X_{il}) - Z_i^T\gamma_0,$$

$$\bar{K} = pK, \quad \text{and} \quad Q_n(\gamma^*) = \tau\theta_n^2 {\gamma^*}^T Z^T Z\gamma^* + {\gamma^*}^T S_n(0) + L_n(0).$$

Based on the notations, the objective function $L_n(\gamma)$ defined in (4) can be rewritten as

$$L_n^*(\gamma^*) = \frac{1}{n}\sum_{i<j} |(\varepsilon_i + \Delta_i) - (\varepsilon_j + \Delta_j) - \theta_n Z_{ij}^T\gamma^*|.$$

Further denote as $S_n(\gamma^*)$ the gradient function of $L_n(\gamma^*)$, that is,

$$S_n(\gamma^*) = \frac{\partial L_n^*(\gamma^*)}{\partial\gamma^*} = -\frac{\theta_n}{n}\sum_{i\neq j} \text{sgn}\{\varepsilon_i + \Delta_i - \varepsilon_j - \Delta_j - \theta_n Z_{ij}^T\gamma^*\}Z_{ij},$$

where $\text{sgn}(\cdot)$ denotes the sign function.

We first quote several necessary lemmas which are frequently used in the sequel, and the detailed proofs can be referred to Feng et al. (2015).

**Lemma 1** *Suppose that the assumptions* (A1)–(A4) *hold, then*

$$S_n(\gamma^*) - S_n(0) = 2\tau\theta_n^2 Z^T Z\gamma^* + o_p(1)\mathbf{1}_{\bar{K}},$$

*where* $\tau$ *is defined in Theorem* 3 *and* $\mathbf{1}_{\bar{K}}$ *is a* $\bar{K}$*-dimension vector of ones.*

**Lemma 2** *Let* $\hat{\gamma}^* = \arg\min L_n^*(\gamma^*)$ *and* $\tilde{\gamma}^* = \arg\min Q_n(\gamma^*)$. *Suppose that the assumptions* (A1)–(A4) *hold, then*

$$\|\hat{\gamma}^* - \tilde{\gamma}^*\|^2 = o_p(K).$$

**Lemma 3** *Suppose that the assumptions* (A1)–(A4) *hold, then*

$$S_n(0) = O_p(1)\mathbf{1}_{\bar{K}}.$$

*Proof of Theorem 1* By the definition of $A_n(\gamma^*)$, it follows from the convexity lemma in Pollard (1991) that

$$\tilde{\gamma}^* = -(2\tau\theta_n^2 Z^T Z)^{-1} S_n(0).$$

Note that, according to Lemma A.3 of Huang et al. (2004), there exists an interval $[C_1, C_2], 0 < C_1 < C_2 < \infty$, such that all the eigenvalues of $\frac{K}{n} Z^T Z$ fall into $[C_1, C_2]$ with probability tending to 1. Write $S_n(0) = (S_{n1}(0), \ldots, S_{n\bar{K}}(0))^T$, then we have

$$\begin{aligned}\|\tilde{\gamma}^*\|^2 &= \frac{1}{4\tau^2} S_n(0)^T \left(\frac{K}{n} Z^T Z\right)^{-1} \left(\frac{K}{n} Z^T Z\right)^{-1} S_n(0) \\ &= O_p(1) S_n(0)^T S_n(0) = O_p(1) \sum_{i=1}^{\bar{K}} S_{ni}(0)^2 = O_p(\bar{K}),\end{aligned}$$

where the last equality holds due to Lemma 3. As $\bar{K} = pK$, it follows that $|\tilde{\gamma}^*|^2 = O_p(K)$. Therefore, based on the triangle inequality and Lemma 2, we obtain

$$\|\check{\gamma}^*\|^2 = \|\check{\gamma}^* - \tilde{\gamma}^* + \tilde{\gamma}^*\|^2 \le \|\check{\gamma}^* - \tilde{\gamma}^*\|^2 + \|\tilde{\gamma}^*\|^2 = o_p(K) + O_p(K) = O_p(K).$$

This is equivalent to $\|\check{\gamma} - \gamma_0\|^2 = O_p(K^2/n)$ since $\check{\gamma}^* = \theta_n^{-1}(\check{\gamma} - \gamma_0)$ and $\theta_n = \sqrt{K/n}$.

In addition, by the properties of spline in De Boor (2001) that there exist some constants $C_3$ and $C_4$ satisfying

$$C_3 K \|\check{\gamma}_j^T B_j - \gamma_{0j}^T B_j\|^2 \le \|\check{\gamma}_j - \gamma_{0j}\|^2 \le C_4 K \|\check{\gamma}_j^T B_j - \gamma_{0j}^T B_j\|^2.$$

Thus, we can derive that $\|\check{\gamma}_j^T B_j - \gamma_{0j}^T B_j\|^2 = O_p(K/n)$. Consequently, by the fact that $\|f_{0j} - B_j^T \gamma_{0j}\| = O_p(K^{-r})$, we have

$$\begin{aligned}\|\check{f}_j - f_{0j}\|^2 &= \|\check{\gamma}_j^T B_j - f_{0j}\|^2 \le \|\check{\gamma}_j^T B_j - \gamma_{0j}^T B_j\|^2 + \|\gamma_{0j}^T B_j - f_{0j}\|^2 \\ &= O_p(K/n) + O_p(K^{-2r}) = O_p(n^{-2r/(2r+1)}),\end{aligned}$$

where the last equality holds due to the assumption that the number of knots $K = O_p\big(n^{1/(2r+1)}\big)$. This completes the proof. □

*Proof of Theorem 2* Firstly, we prove (i). Denote by $\delta_n = \theta_n + \lambda_1 + \lambda_2$, we first prove that $\|\hat{\gamma} - \gamma_0\| = O_p(\bar{K}^{1/2}\delta_n)$. Let $\gamma = \gamma_0 + \bar{K}^{1/2}\delta_n v$, where $v$ is a $\bar{K}$-dimensional vector. It is sufficient to show, for any given $\xi > 0$, there exists a large $C$ such that

$$P\left\{\inf_{\|v\|=C} L_n^{\lambda}(\gamma) > L_n^{\lambda}(\gamma_0)\right\} \ge 1 - \xi. \tag{10}$$

By virtue of the identity $|x - y| - |x| = -y\mathrm{sgn}(x) + 2(y - x)\{I(0 < x < y) - I(y < x < 0)\}$ and the definition of $L_n^{\lambda}(\gamma)$, it follows that

$$\begin{aligned}&L_n^{\lambda}(\gamma) - L_n^{\lambda}(\gamma_0) \\ &= \frac{1}{n} \sum_{i<j} \left\{|Y_{ij} - Z_{ij}^T \gamma| - |Y_{ij} - Z_{ij}^T \gamma_0|\right\} + n \sum_{k=1}^{p} \left\{p_{\lambda_1}(\sqrt{\gamma_k^T D_k \gamma_k})\right.\end{aligned}$$

$$
\begin{aligned}
&\quad - p_{\lambda_1}(\sqrt{\gamma_{0k}^T D_k \gamma_{0k}})\} + n\sum_{k=1}^{p}\left\{p_{\lambda_2}\left(\sqrt{\gamma_k^T E_k \gamma_k}\right) - p_{\lambda_2}\left(\sqrt{\gamma_{0k}^T E_k \gamma_{0k}}\right)\right\} \\
&= \frac{-1}{n}\sum_{i<j} Z_{ij}^T(\gamma-\gamma_0)\mathrm{sgn}(Y_{ij} - Z_{ij}^T\gamma_0) + \frac{2}{n}\sum_{i<j}(Z_{ij}^T\gamma - Y_{ij})\cdot \\
&\quad \left\{I(0 < Y_{ij} - Z_{ij}^T\gamma_0 < Z_{ij}^T(\gamma-\gamma_0)) - I(Z_{ij}^T(\gamma-\gamma_0) < Y_{ij} - Z_{ij}^T\gamma_0 < 0)\right\} \\
&\quad + n\sum_{k=1}^{p}\left\{p_{\lambda_1}(\sqrt{\gamma_k^T D_k \gamma_k}) - p_{\lambda_1}(\sqrt{\gamma_{0k}^T D_k \gamma_{0k}})\right\} \\
&\quad + n\sum_{k=1}^{p}\left\{p_{\lambda_2}(\sqrt{\gamma_k^T E_k \gamma_k}) - p_{\lambda_2}(\sqrt{\gamma_{0k}^T E_k \gamma_{0k}})\right\} \\
&\triangleq L_1 + L_2 + L_3 + L_4. \qquad (11)
\end{aligned}
$$

From Lemma 3, it is easy to verify that $\frac{-1}{n}\sum_{i<j}\mathrm{sgn}(Y_{ij} - Z_{ij}^T\gamma_0)Z_{ij} = \theta_n^{-1}\mathbf{1}_{\bar{K}}$, thus we have $L_1 = O_p(\delta_n\theta_n^{-1}\bar{K}^{1/2}\|v\|) = O_p(n^{1/2}\delta_n\|v\|) = o_p(n\delta_n^2\|v\|)$ due to the assumption $n^{r/(2r+1)}\min\{\lambda_1,\lambda_2\} \to \infty$. Moreover, taking the similar arguments as in the proof of Lemma 1, we can obtain that

$$L_2 = \tau(\gamma-\gamma_0)^T Z^T Z(\gamma-\gamma_0)(1+o_p(1)).$$

By applying Lemma A.3 of Huang et al. (2004) to $L_2$ yields $L_2 = O_p(n\delta_n^2\|v\|^2)$. Obviously, by choosing a sufficiently large $C$, $L_2$ dominates $L_1$ with probability tending to 1.

On the other hand, based on the well-known properties of B-spline that $D_k$ and $E_k$ are of rank $K-1$ and all their positive eigenvalues are of order $1/K$, then according to the inequality $p_\lambda(|x|) - p_\lambda(|y|) \le \lambda|x-y|$, we have

$$L_3 \le nC\lambda_1\sum_{k=1}^{p}\|\gamma_k - \gamma_{0k}\|/\sqrt{K} = O_p(n\lambda_1\delta_n\|v\|) = O_p(n\delta_n^2\|v\|).$$

Thus $L_3$ is dominated by $L_2$ if a sufficiently large $C$ is chosen. Similarly, it is easy to verify that $L_4$ is also dominated by $L_2$. Recall that $L_2 > 0$, so we have (10) holds, which means $\|\hat{\gamma} - \gamma_0\| = O_p(\bar{K}^{1/2}\delta_n)$.

Finally, we will show that the convergence rate can be further improved to $\|\hat{\gamma} - \gamma_0\| = O_p(\bar{K}^{1/2}\theta_n)$. In fact, as the model is fixed as $n \to \infty$, we can find a constant $C > 0$, such that $\gamma_{0k}^T D_k \gamma_{0k} > C$ for $k \le s$ and $\gamma_{0k}^T E_k \gamma_{0k} > C$ for $k \le p_0$. As $\|\hat{\gamma}-\gamma_0\|^2 = O_p(\bar{K}\delta_n^2) = o_p(\bar{K})$ from above result and $\lambda_k = o_p(1)$, $k = 1, 2$, we have

$$
\begin{aligned}
&P\left(p_{\lambda_1}\left(\sqrt{\gamma_{0k}^T D_k \gamma_{0k}}\right) = p_{\lambda_1}\left(\sqrt{\hat{\gamma}_k^T D_k \hat{\gamma}_k}\right)\right) \to 1, \quad j \le s, \\
&P\left(p_{\lambda_1}\left(\sqrt{\gamma_{0k}^T E_k \gamma_{0k}}\right) = p_{\lambda_1}\left(\sqrt{\hat{\gamma}_k^T E_k \hat{\gamma}_k}\right)\right) \to 1, \quad j \le p_0.
\end{aligned}
$$

These facts indicate that

$$P\left(n\sum_{k=1}^{p} p_{\lambda_1}\left(\sqrt{\hat{\gamma}_k^T D_k \hat{\gamma}_k}\right) - n\sum_{k=1}^{p} p_{\lambda_1}\left(\sqrt{\gamma_{0k}^T D_k \gamma_{0k}}\right) \geq 0\right) \to 1,$$
$$P\left(n\sum_{k=1}^{p} p_{\lambda_1}\left(\sqrt{\hat{\gamma}_k^T E_k \hat{\gamma}_k}\right) - n\sum_{k=1}^{p} p_{\lambda_1}\left(\sqrt{\gamma_{0k}^T E_k \gamma_{0k}}\right) \geq 0\right) \to 1.$$

Removing the regularizing terms $L_3$ and $L_4$ in (11), the rate can be improved to $\|\hat{\gamma} - \gamma_0\| = O_p(\bar{K}^{1/2}\theta_n)$ by the same reasoning as above. That is $\|\hat{\gamma} - \gamma_0\|^2 = O_p(\bar{K}\theta_n^2) = O_p(K^2/n)$. As a consequence, following the same approach in the proof of the second part of Theorem 1, we obtain that $\|\hat{f}_j - f_{0j}\|^2 = O_p(n^{-2r/(2r+1)})$, this completes the proof.

In the next, we put our main attention on proving part (ii) as an illustration and part (iii) can be similarly proved with its detailed proof omitted. Suppose that $B_j^T \hat{\gamma}_j$ does not represent a linear function for $p_0 + 1 \leq j \leq s$. Define $\bar{\gamma}$ to be the same as $\hat{\gamma}$ except that $\hat{\gamma}_j$ is replaced by its projection onto the subspace $\{\ \gamma_j : B_j^T \gamma_j$ stands for a linear function $\}$. Therefore, we have that

$$\begin{aligned}
0 &\geq L_n^{\lambda}(\hat{\gamma}) - L_n^{\lambda}(\bar{\gamma}) = (L_n^{\lambda}(\hat{\gamma}) - L_n^{\lambda}(\gamma_0)) - (L_n^{\lambda}(\bar{\gamma}) - L_n^{\lambda}(\gamma_0)) \\
&= \frac{1}{n}\sum_{i<j}\left\{|Y_{ij} - Z_{ij}^T\hat{\gamma}| - |Y_{ij} - Z_{ij}^T\gamma_0|\right\} - \frac{1}{n}\sum_{i<j}\left\{|Y_{ij} - Z_{ij}^T\bar{\gamma}| - |Y_{ij} - Z_{ij}^T\gamma_0|\right\} \\
&\quad + n\sum_{k=1}^{p}\left\{p_{\lambda_1}(\sqrt{\hat{\gamma}_k^T D_k \hat{\gamma}_k}) - p_{\lambda_1}(\sqrt{\bar{\gamma}_k^T D_k \bar{\gamma}_k})\right\} \\
&\quad + n\sum_{k=1}^{p}\left\{p_{\lambda_2}(\sqrt{\hat{\gamma}_k^T E_k \hat{\gamma}_k}) - p_{\lambda_2}(\sqrt{\bar{\gamma}_k^T E_k \bar{\gamma}_k})\right\} \\
&\triangleq M_1(\hat{\gamma}, \gamma_0) - M_2(\bar{\gamma}, \gamma_0) + M_3(\hat{\gamma}, \bar{\gamma}) + M_4(\hat{\gamma}, \bar{\gamma}). \qquad (12)
\end{aligned}$$

Note that, by the same arguments to the derivation of (11), it is not difficult to verify that

$$M_1(\hat{\gamma}, \gamma_0) = \tau(\hat{\gamma} - \gamma_0)^T Z^T Z(\hat{\gamma} - \gamma_0)(1 + o_p(1)) + \theta_n^{-1}(\hat{\gamma} - \gamma_0)^T S_n(0)$$

and

$$M_2(\bar{\gamma}, \gamma_0) = \tau(\bar{\gamma} - \gamma_0)^T Z^T Z(\bar{\gamma} - \gamma_0)(1 + o_p(1)) + \theta_n^{-1}(\bar{\gamma} - \gamma_0)^T S_n(0).$$

Therefore, we can show that

$$\begin{aligned}
&M_1(\hat{\gamma}, \gamma_0) - M_2(\bar{\gamma}, \gamma_0) \\
&\quad = \tau\{(\hat{\gamma} - \bar{\gamma} + \bar{\gamma} - \gamma_0)^T Z^T Z(\hat{\gamma} - \bar{\gamma} + \bar{\gamma} - \gamma_0) \\
&\qquad -(\bar{\gamma} - \gamma_0)^T Z^T Z(\bar{\gamma} - \gamma_0)\}(1 + o_p(1)) + \theta_n^{-1}(\hat{\gamma} - \bar{\gamma})^T S_n(0)
\end{aligned}$$

$$= \tau(\hat{\gamma} - \bar{\gamma})^T Z^T Z(\hat{\gamma} - \bar{\gamma}) + 2\tau(\bar{\gamma} - \gamma_0)^T Z^T Z(\hat{\gamma} - \bar{\gamma}) + \theta_n^{-1}(\hat{\gamma} - \bar{\gamma})^T S_n(0)$$
$$\geq 2\tau(\bar{\gamma} - \gamma_0)^T Z^T Z(\hat{\gamma} - \bar{\gamma}) + \theta_n^{-1}(\hat{\gamma} - \bar{\gamma})^T S_n(0) \triangleq N_1 + N_2.$$

Recall that $\bar{\gamma}_k$ is the projection of $\hat{\gamma}_k$ onto $\{\gamma_k : \gamma_k^T E_k \gamma_k = 0\}$, then $\hat{\gamma}_k - \bar{\gamma}_k$ is orthogonal to the space. Furthermore, the space $\{\gamma_k : \gamma_k^T E_k \gamma_k = 0\}$ is just the eigenspace of $E_k$ corresponding to the zero eigenvalue. Consequently, based on the characterization of eigenvalues in terms of Rayleigh quotient, $(\hat{\gamma}_k - \bar{\gamma}_k)^T E_k(\hat{\gamma}_k - \bar{\gamma}_k)/\|\hat{\gamma}_k - \bar{\gamma}_k\|^2$ lies between the minimum and the maximum positive eigenvalues of $E_k$, which is of order $1/K$. Taking into account of the fact that $\hat{\gamma}_k^T E_k \hat{\gamma}_k = (\hat{\gamma}_k - \bar{\gamma}_k)^T E_k(\hat{\gamma}_k - \bar{\gamma}_k)$ since $\bar{\gamma}_k^T E_k \bar{\gamma}_k = 0$, we derive $\|\hat{\gamma}_k - \bar{\gamma}_k\| = O_p(\sqrt{K \hat{\gamma}_k^T E_k \hat{\gamma}_k})$. According to Lemma 3, Lemma A.3 of Huang et al. (2004) and the result $\|\bar{\gamma} - \gamma_0\| = O_p(K/\sqrt{n})$ from part (i), it follows that

$$\|N_1\| \leq O_p\left(\frac{n}{K}\|\bar{\gamma} - \gamma_0\| \cdot \|\hat{\gamma} - \bar{\gamma}\|\right) = O_p\left(\sqrt{nK}\sum_{k=1}^{p}\sqrt{\hat{\gamma}_k^T E_k \hat{\gamma}_k}\right),$$
$$\|N_2\| \leq O_p\left(\theta_n^{-1}\|\hat{\gamma} - \bar{\gamma}\| \cdot \|S_n(0)\|\right) = O_p\left(\sqrt{nK}\sum_{k=1}^{p}\sqrt{\hat{\gamma}_k^T E_k \hat{\gamma}_k}\right).$$

These facts leads to

$$M_1(\hat{\gamma}, \gamma_0) - M_2(\bar{\gamma}, \gamma_0) \geq -O_p\left(\sqrt{nK}\sum_{k=1}^{p}\sqrt{\hat{\gamma}_k^T E_k \hat{\gamma}_k}\right). \tag{13}$$

On the other hand, according to the proof of (i), we have $P\big(p_{\lambda_1}(\sqrt{\hat{\gamma}_k^T D_k \hat{\gamma}_k}) = p_{\lambda_1}(\sqrt{\bar{\gamma}_k^T D_k \bar{\gamma}_k})\big) \to 1$ and $P\big(\bar{\gamma}_k^T E_k \bar{\gamma}_k = 0\big) \to 1$. Substituting these results into (12) yields

$$P\left(M_1(\hat{\gamma}, \gamma_0) - M_2(\bar{\gamma}, \gamma_0) + n\sum_{k=1}^{p} p_{\lambda_2}(\sqrt{\hat{\gamma}_k^T E_k \hat{\gamma}_k}) \leq 0\right) \to 1. \tag{14}$$

In addition, based on the result of (i) and the condition $n^{r/(2r+1)} \min\{\lambda_1, \lambda_2\} \to \infty$, it is easy to verify that

$$\sqrt{\hat{\gamma}_k^T E_j \hat{\gamma}_k} = \sqrt{(\hat{\gamma}_k - \gamma_{0k})^T E_k(\hat{\gamma}_k - \gamma_{0k})} = O_p(\sqrt{K/n}) = o_p(\lambda_2).$$

Hence, we have $P\big(p_{\lambda_2}(\sqrt{\hat{\gamma}_k^T E_k \hat{\gamma}_k}) = \lambda_2\sqrt{\hat{\gamma}_k^T E_k \hat{\gamma}_k}\big) \to 1$ by the definition of SCAD penalty function.

As a consequence, if $\hat{\gamma}_k^T E_k \hat{\gamma}_k > 0$, we have

$$n \sum_{k=1}^{p} p_{\lambda_2}\left(\sqrt{\hat{\gamma}_k^T E_k \hat{\gamma}_k}\right) = O_p\left(n\lambda_2 \sum_{k=1}^{p} \sqrt{\hat{\gamma}_k^T E_k \hat{\gamma}_k}\right). \tag{15}$$

Combining (13) and (15) along with the condition $n^{r/(2r+1)} \min\{\lambda_1, \lambda_2\} \to \infty$, it follows that

$$M_1(\hat{\gamma}, \gamma_0) - M_2(\bar{\gamma}, \gamma_0) + n \sum_{k=1}^{p} p_{\lambda_2}\left(\sqrt{\hat{\gamma}_k^T E_k \hat{\gamma}_k}\right) > 0,$$

which is contradictory to (14). Then we complete the proof of Theorem 2. □

*Proof of Theorem 3* Note that, by the results of Theorem 2, we only need to consider a correctly specified partially linear additive model as (2) without regularization terms. Specifically, the corresponding objective function is

$$\Phi_n(\alpha, \beta) = \frac{1}{n} \sum_{i<j} |Y_{ij} - V_{ij}^T \alpha - X_{ij}^{(2)^T} \beta|,$$

where $V_i = \left(B_1(X_{i1})^T, \ldots, B_p(X_{ip_0})^T\right)^T$, $X_i^{(2)} = (X_{i(p_0+1)}, \ldots, X_{is})^T$ and $\alpha = (\gamma_1, \ldots, \gamma_{p_0})^T$ is the corresponding coefficient vector of the spline approximation. Let $(\hat{\alpha}^T, \hat{\beta}^T)^T = \arg\min \Phi_n(\alpha, \beta)$, $\tilde{\Delta}_i = \sum_{l=1}^{p_0} f_{0l}(X_{il}) - V_i^T \hat{\alpha}$, $\delta_n = n^{-1/2}$ and $\beta^* = \delta_n^{-1}(\beta - \beta_0)$. Then, $\hat{\beta}^*$ must be the minimizer of the following function

$$\Phi_n^*(\beta^*) = \frac{1}{n} \sum_{i<j} |(\varepsilon_i + \tilde{\Delta}_i) - (\varepsilon_j + \tilde{\Delta}_j) - \delta_n X_{ij}^{(2)^T} \beta^*|.$$

Denote by $S_n^*(\beta^*)$ the gradient function of $\Phi_n^*(\beta^*)$, that is

$$S_n^*(\beta^*) = \frac{\partial \Phi_n^*(\beta^*)}{\partial \beta^*} = -\frac{\delta_n}{n} \sum_{i \neq j} \operatorname{sgn}\{(\varepsilon_i + \tilde{\Delta}_i) - (\varepsilon_j + \tilde{\Delta}_j) - \delta_n X_{ij}^{(2)^T} \beta^*\} X_{ij}^{(2)}.$$

Then, we can show that

$$\begin{aligned} S_n^*(\beta^*) - S_n^*(0) = &-\frac{\delta_n}{n} \sum_{i \neq j} \operatorname{sgn}\left((\varepsilon_i + \tilde{\Delta}_i) - (\varepsilon_j + \tilde{\Delta}_j) - \delta_n X_{ij}^{(2)^T} \beta^*\right) X_{ij}^{(2)} \\ &+ \frac{\delta_n}{n} \sum_{i \neq j} \operatorname{sgn}\left((\varepsilon_i + \tilde{\Delta}_i) - (\varepsilon_j + \tilde{\Delta}_j)\right) X_{ij}^{(2)}. \end{aligned}$$

Taking into consideration of the results obtained in Theorem 2, we have $\tilde{\Delta}_i = O_p(K^{-r}) = o_p(1)$ as $n \to \infty$. Hence, following the similar proof of Lemma 1,

it is not difficult to obtain

$$S_n^*(\beta^*) - S_n^*(0) = 2\tau\delta_n^2\Sigma\beta^*, \tag{16}$$

where $\Sigma$ is defined in assumption (A3). Further let $B_n(\beta^*) = \tau\delta_n^2\beta^{*T}\Sigma\beta^* + \beta^{*T}S_n^*(0) + \Phi_n^*(0)$ and its minimizer denoted by $\tilde{\beta}^*$. Then it is not difficult to verify that $\tilde{\beta}^* = -(2\tau)^{-1}(\delta_n^2\Sigma)^{-1}S_n^*(0)$. Based on Equation (16) and a similar arguments of Lemma 2, it follows that

$$\hat{\beta}^* = \tilde{\beta}^* + o_p(1) = -(2\tau)^{-1}(\delta_n^2\Sigma)^{-1}S_n^*(0) + o_p(1). \tag{17}$$

In addition, by the assumption that $\varepsilon_i$ is the random error independent of $X_i$, combined with some calculations, we have

$$\delta_n^{-2}S_n^*(0) \xrightarrow{d} N\big(0, E\big\{(2H(\varepsilon) - 1)^2\big\}\Sigma\big), \tag{18}$$

where $H(\cdot)$ stands for the cumulative distribution function of $\varepsilon$. Furthermore, it can be shown that

$$\begin{aligned} E\{(2H(\varepsilon) - 1)^2\} &= \int (2H(\varepsilon) - 1)^2 h(\varepsilon)d\varepsilon \\ &= \int 4H(\varepsilon)^2 h(\varepsilon)d\varepsilon - 4\int H(\varepsilon)h(\varepsilon)d\varepsilon + \int h(\varepsilon)d\varepsilon \\ &= \int 4H(\varepsilon)^2 dH(\varepsilon) - 4\int H(\varepsilon)dH(\varepsilon) + 1 = 1/3. \end{aligned} \tag{19}$$

Therefore, substituting (18) and (19) into (17), we complete the proof. □

*Proof of Theorem 4* Based on the asymptotic results of Theorem 3 and the least square B-spline estimate given in Theorem 3 of Lian (2012a), we immediately obtain $\mathrm{ARE}(\hat{\beta}_{RR}, \hat{\beta}_{LS}) = 12\tau^2\sigma^2$. In addition, a result of Hodges and Lehmann (1956) indicates that the ARE has a lower bound 0.864, with this lower bound being obtained at the density $h(x) = \frac{3}{20\sqrt{5}}(5 - x^2)I(|x| \le 5)$. This completes the proof. □

## References

David HA (1998) Early sample measures of variability. Stat Sci 13:368–377

De Boor C (2001) A practical guide to splines, revised edn. Springer, New York

Deng G, Liang H (2010) Model averaging for semiparametric additive partial linear models. Sci China Math 53:1363–1376

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

Feng L, Zou C, Wang Z, Wei X, Chen B (2015) Robust spline-based variable selection in varying coefficient model. Metrika 78:85–118

Härdle W, Huet S, Mammen E, Sperlich S (2004) Bootstrap inference in semiparametric generalized additive models. Econ Theory 20:265–300

Hettmansperger TP, McKean JW (2011) Robust nonparametric statistical methods, 2nd edn. Chapman and Hall, Boca Raton
Hodges JL, Lehmann EL (1956) The efficiency of some nonparametric competitors of the t-test. Ann Math Stat 27:324–335
Huang J, Horowitz JL, Wei F (2010) Variable selection in nonparametric additive models. Ann Stat 38:2282–2313
Huang JZ, Wu CO, Zhou L (2004) Polynomial spline estimation and inference for varying coefficient models with longitudinal data. Stat Sin 14:763–788
Jiang J, Zhou H, Jiang X, Peng J (2007) Generalized likelihood ratio tests for the structure of semiparametric additive models. Can J Stat 35:381–398
Kai B, Li R, Zou H (2010) Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. J R Stat Soc Ser B 72:49–69
Leng C (2010) Variable selection and coefficient estimation via regularized rank regression. Stat Sin 20:167–181
Li Q (2000) Efficient estimation of additive partially linear models. Int Econ Rev 41:1073–1092
Li J, Li Y, Zhang R (2015) B spline variable selection for the single index models. Stat Pap. doi:10.1007/s00362-015-0721-z
Lian H (2012a) Shrinkage estimation for identification of linear components in additive models. Stat Probab Lett 82:225–231
Lian H (2012b) Semiparametric estimation of additive quantile regression models by two-fold penalty. J Bus Econ Stat 30:337–350
Liu X, Wang L, Liang H (2011) Estimation and Variable selection for semiparametric additive partial linear models. Stat Sin 21:1225–1248
Mammen E, Park B (2006) A simple smooth backfitting method for additive models. Ann Stat 34:2252–2271
Opsomer JD, Ruppert D (1999) A root-n consistent backfitting estimator for semiparametric additive modeling. J Comput Graph Stat 8:715–732
Pollard D (1991) Asymptotics for least absolute deviation regression estimators. Econ Theory 7:186–199
Sievers GL, Abebe A (2004) Rank estimation of regression coefficients using iterated reweighted least squares. J Stat Comput Simul 74:821–831
Sun J, Lin L (2014) Local rank estimation and related test for varying-coefficient partially linear models. J Nonparametr Stat 26:187–206
Tang Q (2015) Robust estimation for spatial semiparametric varying coefficient partially linear regression. Stat Pap 56:1137–1161
Wang L, Kai B, Li R (2009) Local rank inference for varying coefficient models. J Am Stat Assoc 488:1631–1645
Wang M, Song L (2013) Identification for semiparametric varying coefficient partially linear models. Stat Probab Lett 83:1311–1320
Wei C, Liu C (2012) Statistical inference on semi-parametric partial linear additive models. J Nonparametr Stat 24:809–823
Wei C, Luo Y, Wu X (2012) Empirical likelihood for partially linear additive errors-in-variables models. Stat Pap 53:485–496
Xue L (2009) Consistent variable selection in additive models. Stat Sin 19:1281–1296
Yu K, Lu Z (2004) Local linear additive quantile regression. Scand J Stat 31:333–346
Yu K, Park B, Mammen E (2008) Smooth backfitting in generalized additive models. Ann Stat 36:228–260
Zhang HH, Cheng G, Liu Y (2011) Linear or nonlinear? Automatic structure discovery for partially linear models. J Am Stat Assoc 106:1099–1112