Stat Papers (2020) 61:1911–1937
https://doi.org/10.1007/s00362-018-1013-1

REGULAR ARTICLE

A robust and efficient estimation and variable selection method for partially linear models with large-dimensional covariates

Hu Yang[1] · **Ning Li**[1] · **Jing Yang**[2]

Received: 19 December 2016 / Revised: 16 April 2018 / Published online: 21 May 2018

Abstract In this paper, a new robust and efficient estimation approach based on local modal regression is proposed for partially linear models with large-dimensional covariates. We show that the resulting estimators for both parametric and nonparametric components are more efficient in the presence of outliers or heavy-tail error distribution, and as asymptotically efficient as the corresponding least squares estimators when there are no outliers and the error distribution is normal. We also establish the asymptotic properties of proposed estimators when the covariate dimension diverges at the rate of $o\left(\sqrt{n}\right)$. To achieve sparsity and enhance interpretability, we develop a variable selection procedure based on SCAD penalty to select significant parametric covariates and show that the method enjoys the oracle property under mild regularity conditions. Moreover, we propose a practical modified MEM algorithm for the proposed procedures. Some Monte Carlo simulations and a real data are conducted

This work is supported by the National Natural Science Foundation of China (Grant No. 11671059).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00362-018-1013-1) contains supplementary material, which is available to authorized users.

✉ Ning Li
ninglicqu@163.com

Hu Yang
yh@cqu.edu.cn

Jing Yang
yang2009jing@163.com

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

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

to illustrate the finite sample performance of the proposed estimators. Finally, based on the idea of sure independence screening procedure proposed by Fan and Lv (J R Stat Soc 70:849–911, 2008), a robust two-step approach is introduced to deal with ultra-high dimensional data.

Keywords Partially linear models · Robust estimation · Variable selection · Oracle property

1 Introduction

Consider the partially linear models (PLM)

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

where $X = (x_1, \ldots, x_{p_n})^T \in \mathbb{R}^{p_n}$ and $Z = (z_1, \ldots, z_q)^T \in \mathbb{R}^q$ are the covariates in the parametric and nonparametric components, $\beta = (\beta_1, \ldots, \beta_{p_n})^T$ is a p_n-dimensional vector of unknown parameters, $f(\cdot)$ is an unknown smooth function, and the random error ε satisfies $E(\varepsilon\,|X, Z) = 0$. Ever since first introduced by Engle et al. (1986), the PLM have been extensively studied in the literature. For example, see Robinson (1988), Speckman (1988), Zeger and Diggle (1994), Severini and Staniswalis (1994) and Hardle et al. (2000).

In practice, large amounts of variables are usually included in regression model to reduce the possible modeling biases. However, inclusion of too many irrelevant variables can degrade the estimation accuracy and model interpretability. Classical variable selection procedures such as AIC (Akaike 1973), BIC (Schwarz 1978), Mallows' Cp (Mallows 1973) and k-fold Cross-Validation (Breiman 1995) all suffered from the problems of unstability and intensive computation. To address these deficiencies, Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) penalty to perform simultaneous estimation and variable selection. But just as Fan and Li (2001) conjectured, the oracle property does not hold for the LASSO penalty. Later, the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001) and adaptive LASSO penalty (Zou 2006) had been proposed to possess the oracle property. In recent years, these penalized methods have been widely used for variable selection in PLM. For instance, Li and Liang (2008) proposed two classes of penalized procedures for variable selection in PLM with measurement errors. Ni et al. (2009) proposed a new type of double-penalized method for PLM with a divergent number of covariates. Xie and Huang (2009) introduced SCAD-penalized regression in high-dimensional PLM. Zhou et al. (2010) proposed nonconcave penalized procedure for fixed-effects PLM with errors in variables. And, Chen et al. (2012) combined the ideas of profiling and adaptive Elastic-net for variable selection in PLM with large-dimensional covariates. It is important to note that all these works were built on the least squares (LS) regression, which is highly sensitive to outliers and their efficiency may be significantly decreased for many commonly used non-normal errors. To this end, researchers began to study the robust estimation and variable selection for PLM in the framework of the least absolute deviation (LAD) method

(Wang et al. 2007), which is particularly suited to the heavy-tailed error distributions. To the best of our knowledge, Zhu et al. (2013) proposed a class of penalized LAD approach in PLM with large-dimensional covariates. However, the LAD method may lose some efficiency when there are no outliers and the error distribution is normal. Hence, it is highly desirable to develop a robust and efficient method that can simultaneously conduct estimation and variable selection in PLM for different error distributions.

More recently, Yao et al. (2012) investigated a new estimation method based on a local modal regression (LMR) in a nonparametric model. They demonstrated that the LMR estimator 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 the LS estimator. Similar conclusions have also been confirmed in Zhang et al. (2013), Yang and Yang (2014), Yao and Li (2014) and Zhao et al. (2014). This fact motivates us to extend the local modal regression to PLM. The main goal of this paper is to develop a robust and efficient estimation and variable selection procedure for PLM, in which the covariate dimension diverges at the rate of $o(\sqrt{n})$. We show that the resulting estimators for both parametric and nonparametric components are more efficient in the case of outliers or heavy-tail error distribution, and as asymptotically efficient as the corresponding LS estimators when there are no outliers and the error distribution is normal. The main contributions of this paper are threefold. Firstly, the proposed LMR estimators for both parametric and nonparametric components are robust and efficient in PLM with large-dimensional covariates. Secondly, we develop a variable selection procedure based on SCAD penalty to identify significant covariates in the parametric component and prove that the method enjoys the oracle property under mild regularity conditions. Finally, a two-step robust procedure based on sure independence screening and penalized LMR is proposed to deal with ultra-high dimensional cases.

The rest of this paper is organized as follows. In Sect. 2, following the idea of LMR, we propose a new estimation method for PLM with large-dimensional covariates, and establish the theoretical properties of the resulting LMR estimators for both parametric and nonparametric components. A robust and efficient variable selection procedure via SCAD penalty is developed to select significant parametric covariates and its oracle property is also established in Sect. 3. In Sect. 4, we discuss the details of bandwidth selection and BIC criterion is suggested to select the regularization parameter. Moreover, we introduce a modified MEM algorithm for implementation. In Sect. 5, some Monte Carlo simulations as well as a real data example are conducted to show the finite sample performance of the proposed estimators. In Sect. 6, a two-step robust procedure based on sure independence screening and penalized LMR is proposed to deal with ultra-high dimensional cases and a simulation study is also presented in this section. We conclude with a few remarks in Sect. 7. All the technical proofs are given in the "Appendix".

2 Robust estimation procedure

2.1 Robust LMR for PLM

In this subsection, our strategy is that we first use the profile least squares approach (Speckman 1988) to transform the semiparametric model to the classic linear model, and then develop a robust estimation procedure for PLM.

Suppose that $\{(X_i, Y_i, Z_i), i = 1, 2, \ldots, n\}$ is an independent identically distributed sample from model (1). It follows from the profile least squares approach that

$$Y_i - E(Y_i|Z_i) = \{X_i - E(X_i|Z_i)\}^T \beta + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{2}$$

which is a standard linear model if $E(X_i|Z_i)$ and $E(Y_i|Z_i)$ are known. However, both $E(X_i|Z_i)$ and $E(Y_i|Z_i)$ in model (2) are not observed in practice. Thus, we first need to estimate $E(X_i|Z_i)$ and $E(Y_i|Z_i)$. This can be done through kernel smoothing or local linear approximation (Speckman 1988; Fan and Gijbels 1996). For example, we can estimate $E(X_i|Z_i)$ and $E(Y_i|Z_i)$ by

$$\widehat{E}(X_i|Z_i) = \frac{\sum_{j=1}^n K\left(\frac{Z_j - Z_i}{d_1}\right) X_j}{\sum_{j=1}^n K\left(\frac{Z_j - Z_i}{d_1}\right)},$$

and

$$\widehat{E}(Y_i|Z_i) = \frac{\sum_{j=1}^n K\left(\frac{Z_j - Z_i}{d_2}\right) Y_j}{\sum_{j=1}^n K\left(\frac{Z_j - Z_i}{d_2}\right)}, \tag{3}$$

respectively, where $K(\cdot)$ is a q-dimensional kernel function, d_1 and d_2 are the bandwidths. In what follows, we denote $m_X(Z) = E(X|Z)$, $m_Y(Z) = E(Y|Z)$, $\widetilde{X} = X - E(X|Z)$, $\widetilde{Y} = Y - E(Y|Z)$, $\widehat{\widetilde{X}} = X - \widehat{E}(X|Z)$, and $\widehat{\widetilde{Y}} = Y - \widehat{E}(Y|Z)$.

After profiling, we can focus on the general estimation procedures in the context of classic linear model. First, we introduce the commonly used LS method. Specifically, we construct the LS estimator by minimizing the following objective function

$$\sum_{i=1}^n \left(\widehat{\widetilde{Y}}_i - \widehat{\widetilde{X}}_i^T \beta\right)^2 \tag{4}$$

with respect to β.

It is well known that the LS method is very sensitive to outliers in the dataset. Then we can consider the outliers-resistant loss functions such as L_1 or, more generally, Huber's ψ function (Huber 1981). Without loss of generality, we only introduce the L_1 loss function to obtain the LAD estimator (Wang et al. 2007). One can also refer to Zhu et al. (2013) for a detailed discussion of this class of estimator. Therefore, instead of minimizing (4), we minimize

$$\sum_{i=1}^{n} \left| \widehat{\widetilde{Y}}_i - \widehat{\widetilde{X}}_i^T \beta \right| \tag{5}$$

with respect to β.

However, the LAD method may lose some efficiency when there are no outliers and the error distribution is normal. Hence, it is highly desirable to develop a robust and efficient method that can conduct robust estimation in PLM for different error distributions. In general, the mode is insensitive to the outliers in dataset or the heavy-tail error distributions. Moreover, the modal regression provides more meaningful point prediction and larger coverage probability for prediction than the mean regression when the error density is skewed (Zhang et al. 2013). Motivated by this fact, we devote to extending the local modal regression to PLM.

In this paper, we propose the LMR estimator $\widehat{\beta}$ by maximizing the following objective function

$$\frac{1}{n} \sum_{i=1}^{n} \phi_{h_1} \left(\widehat{\widetilde{Y}}_i - \widehat{\widetilde{X}}_i^T \beta \right), \tag{6}$$

with respect to β, where $\phi_{h_1}(t) = h_1^{-1} \phi(t/h_1)$, $\phi(t)$ is a kernel density function, and h_1 plays the role of the bandwidth, which determines the degree of robustness of the LMR estimator. For the ease of computation, we use the standard normal density for $\phi(\cdot)$ throughout this paper. Similar idea can be seen in Yao et al. (2012), Zhang et al. (2013), and Yang and Yang (2014).

After obtaining the LMR estimator of β, we can further estimate the unknown smooth function $f(\cdot)$. In this paper, we adopt local linear approximation (Fan and Gijbels 1996) to approximate $f(\cdot)$. That is to say, for any fixed $Z = z \in \mathbb{R}^q$, we approximate $f(z)$ by a linear function

$$f(z_0) \approx f(z) + (z_0 - z)^T f'(z) = a + (z_0 - z)^T b \tag{7}$$

for z_0 in a neighborhood of z. As a consequence, we turn to estimate the intercept term a. To this end, we propose the LMR estimator $\widehat{f}(z) = \widehat{a}$ by maximizing

$$\sum_{i=1}^{n} \phi_{h_3} \Big(Y_i - X_i^T \widehat{\beta} - a - (Z_i - Z)^T b \Big) K \left(\frac{Z_i - Z}{h_2} \right), \tag{8}$$

with respect to a and b.

It is noteworthy that the bandwidths h_1, h_2 and h_3 are selected by data-driven method so that the resulting estimators can be adaptively robust, and the detailed choices of the these bandwidths will be discussed in Sect. 4.

2.2 Theoretical properties

In this subsection, we first will study the theoretical properties of the proposed LMR estimators. For simplicity, we denote $U = (X, Z)$, and β_0 as the true value of β. Let $F(u, h) = E\left\{\phi_h''(\varepsilon) | U = u\right\}$, $G(u, h) = E\left\{\phi_h'(\varepsilon)^2 | U = u\right\}$.

To obtain the theoretical properties of the LMR estimator $\widehat{\beta}$, we assume the following regularity conditions:

(A1) The matrix $\mathrm{cov}(\widetilde{X})$ is positive-definite, $E\left(|\varepsilon|^3 \,|X, Z\right) < \infty$, and $\sup_z E\left(\|X\|^3 \,|Z = z\right) < \infty$.

(A2) The bandwidth d_k in Eq. (3) satisfies $nd_k^8 \to 0$ and $nd_k^{2q} \to \infty$ for $k = 1, 2$.

(A3) The kernel function $K(\cdot)$ is a symmetric density function with compact support and satisfies $\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} K(\mathbf{t}) dt_1 \cdots dt_q = 1$, $\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \mathbf{t} K(\mathbf{t}) dt_1 \cdots dt_q = 0$, where $\mathbf{t} = \left(t_1, \ldots, t_q\right)^T$.

(A4) $f(\cdot)$ and $m_X(\cdot)$ are continuous on their support sets.

(A5) $F(u, h)$, $G(u, h)$ are continuous with respect to u.

(A6) $F(u, h) < 0$ for any $h > 0$.

(A7) $E\left(\phi_h'(\varepsilon) \,|U = u\right) = 0$, $E\left(\phi_h'(\varepsilon)^3 \,|U = u\right)$, $E\left(\phi_h''(\varepsilon)^2 \,|U = u\right)$, and $E\left(\phi_h'''(\varepsilon) \,|U = u\right)$ are continuous with respect to u.

Remark 1 The conditions (A1)–(A4) are standard in the semiparametric regression literature. Condition (A1) is a necessary moment condition. Condition (A2) ensures that undersmoothing is not needed in order to obtain the root-n/p_n consistency and asymptotic normality. In practice, we can use data-driven approach to select the bandwidths d_k, $k = 1, 2$. Condition (A3) is a common condition on the kernel function. Condition (A4), together with conditions (A1)–(A3), ensures that the consistency of the kernel estimation. A detailed discussion of these conditions can be found in Hardle et al. (2000). The conditions (A5)–(A7) are necessary conditions used in local modal nonparametric regression in Yao et al. (2012). Condition (A5) is a basic assumption. Condition (A6) ensures that there exists a local maximizer in the objective function (6), while condition (A7) is used to control the magnitude of the remainder in a third-order Taylor expansion of this objective function. In particular, the condition $E\left(\phi_h'(\varepsilon) \,|U = u\right) = 0$ is satisfied if the error density is symmetric about zero, which ensures that the proposed LMR estimator is consistent.

Theorem 1 *Under the regularity conditions (A1)–(A7), if $p_n^2/n \to 0$ as $n \to \infty$, and h_1 is a constant and does not depend on n, then we have*

$$\|\widehat{\beta} - \beta_0\| = \mathrm{O}_p(\sqrt{p_n/n}),$$

where $\|\cdot\|$ stands for the Euclidean norm.

Theorem 1 indicates that the LMR estimator $\widehat{\beta}$ is root-n/p_n consistent. Meanwhile, the following theorem states the asymptotic normality of the LMR estimator $\widehat{\beta}$ when p_n diverges at the rate of $o(n^{1/2})$.

Theorem 2 *Under the regularity conditions (A1)–(A7), if $p_n^2/n \to 0$ as $n \to \infty$, and h_1 is a constant and does not depend on n, then we have*

$$\sqrt{n}(\widehat{\beta} - \beta_0) \to N(\mathbf{0}, \Sigma_1^{-1}\Sigma_2\Sigma_1^{-1})$$

in distribution, where $\Sigma_1 = E\{F(u, h_1)\widetilde{X}\widetilde{X}^T\}$, and $\Sigma_2 = Var\{\widetilde{X}\phi_{h_1}'(\varepsilon)\}$.

In this paper, we also provide the asymptotic normality of the LMR estimator $\widehat{f}(z)$ as follows.

Throughout, let $\rho(z)$ be the density function of Z and $g(\bar{y}|z)$ be the conditional density function of $\overline{Y} = Y - X^T\beta_0$ given $Z = z$ with respect to a measure μ. With a given constant h that does not depend on n, we let $\varphi_h(t|z) = E\{-\phi_h(\varepsilon + t)|Z = z\}$, and use $\varphi_h'(t\,|z)$ and $\varphi_h''(t\,|z)$ to denote $\partial(\varphi_h(t\,|z))/\partial t$ and $\partial^2(\varphi_h(t\,|z))/\partial^2 t$, respectively. In addition to the conditions in Theorem 1, we further assume the following regularity conditions:

(B1) The smooth function $f(\cdot)$ has a continuous second derivative.

(B2) The density function $\rho(\cdot)$ is continuous and positive on its support.

(B3) Assume that $\varphi_h(t|z_n)$, $\varphi_h'(t|z_n)$ and $\varphi_h''(t|z_n)$ as functions of z_n are bounded and continuous in a neighborhood of z for all small t and that $\varphi_h(0|z_n) \neq 0$. $\varphi_h''(t|z_n)$ as a function of t is continuous in a neighborhood of point 0, uniformly for z_n in a neighborhood of z.

(B4) The conditional density function $g(\bar{y}|z)$ is continuous in z for each $\bar{y}$. Moreover, there exist positive constants ϵ, σ and a positive function $G(\bar{y}|z)$ such that

$$\sup_{|z_n - z| \le \epsilon} g(\bar{y}|z_n) \le G(\bar{y}|z),$$
$$\int |\phi_h'(\varepsilon)|^{2+\sigma} G(\bar{y}|z) d\mu(\bar{y}) < \infty,$$

and

$$\int \{\phi_h(\bar{y} - t) - \phi_h(\bar{y}) - \phi_h'(\bar{y})t\}^2 G(\bar{y}|z) d\mu(\bar{y}) = o(t^2), \quad \text{as} \quad t \to 0.$$

Remark 2 The conditions (B1)–(B4) follow from the adaptations of the condition A of Fan et al. (1994) and can be easily ensured or verified. Conditions (B1) and (B2) are necessary conditions to ensure that the bias and the variance of $\widehat{f}(z)$ have the right rate of convergence, respectively. Condition (B3) ensures the uniqueness of the solution to the objective function (8). Conditions (B4) is required by the dominated convergence theorem and moment calculation in the proof of the asymptotic normality.

Theorem 3 *In addition to the conditions in Theorem 1, the regularity conditions (B1)–(B4) are satisfied. If $h_2 \to 0$, $nh_2^q \to \infty$ as $n \to \infty$, and h_3 is a constant and does not depend on n, then for any fixed $Z = z \in \mathbb{R}^q$,*

$$(nh_2^q)^{1/2}\{\widehat{f}(z) - f(z) - bias\} \to N(0, \tau^2(z)),$$

in distribution, where

$$bias = (1/2) f''(z) h_2^2 \int t^2 K(t) dt,$$
$$\tau^2(z) = \frac{\int K^2(t)dt}{\rho(z)} \frac{\int [\phi_{h_3}'(\varepsilon)]^2 g(\bar{y}|z) d\mu(\bar{y})}{[\varphi_{h_3}''(0|z)]^2}.$$

3 Variable selection procedure

3.1 Penalized LMR for PLM

In this subsection, we aim to develop a variable selection procedure to select significant parametric covariates for PLM. To this end, we consider the following penalized function based on LMR

$$Q_\lambda(\beta) = \sum_{i=1}^{n} \phi_{h_1}\left(\widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_i^T \beta\right) - n \sum_{j=1}^{p_n} p_\lambda(|\beta_j|), \tag{9}$$

where $p_\lambda(\cdot)$ is a penalty function with regularization parameter λ. One of the most commonly used penalty is the SCAD penalty, which is defined as follows

$$p_\lambda(t) = \begin{cases} \lambda|t|, & |t| \le \lambda, \\ \frac{(a^2-1)\lambda^2-(|t|-a\lambda)^2}{2(a-1)}, & \lambda_n < |t| \le a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & |t| > a\lambda, \end{cases}$$

where a is some constant usually taken to be 3.7 as suggested in Fan and Li (2001). For given regularization parameter λ, we can get a sparse estimator $\widehat{\beta}^\lambda$ of β and then conduct the variable selection procedure.

However, as the SCAD penalty function is singular at 0, it is difficult to maximize the objective function (9). Following Fan and Li (2001), we then apply the local quadratic approximation (LQA) algorithm to the SCAD penalty function for fixed λ. Suppose that the initial value $\widehat{\beta}^{\lambda(0)}$ is very close to the maximizer of the objective function (9). If $\widehat{\beta}_j^{\lambda(0)}$ is very close to 0, then we set $\widehat{\beta}_j^\lambda = 0$. Otherwise, $p_\lambda(\beta_j)$ can be locally approximated as

$$p_\lambda(\beta_j) \approx p_\lambda\left(\widehat{\beta}_j^{\lambda(0)}\right) + \frac{1}{2}\left\{p'_\lambda\left(\widehat{\beta}_j^{\lambda(0)}\right)\Big/\left|\widehat{\beta}_j^{\lambda(0)}\right|\right\}\left\{\beta_j^2 - \left(\widehat{\beta}_j^{\lambda(0)}\right)^2\right\}, \quad \text{for} \quad \beta_j \approx \widehat{\beta}_j^{\lambda(0)}.$$

As a consequence, we can obtain the penalized LMR estimator $\widehat{\beta}^\lambda$ by maximizing the following objective function

$$Q_\lambda(\beta) = \sum_{i=1}^{n} \phi_{h_1}\left(\widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_i^T \beta\right) - \frac{n}{2}\left\{p'_\lambda\left(\widehat{\beta}_j^{\lambda(0)}\right)\Big/\left|\widehat{\beta}_j^{\lambda(0)}\right|\right\}\left\{\beta_j^2 - \left(\widehat{\beta}_j^{\lambda(0)}\right)^2\right\} \tag{10}$$

with respect to β.

Then, similar as the objective function (8), we obtain the LMR estimator $\widehat{f}^\lambda(z) = \widehat{a}^\lambda$ by maximizing

$$\sum_{i=1}^{n} \phi_{h_3}(Y_i - X_i{}^T \widehat{\beta}^\lambda - a - (Z_i - Z)^T b) K\left(\frac{Z_i - Z}{h_2}\right), \tag{11}$$

with respect to a and b.

3.2 Theoretical properties

In this subsection, we will study the theoretical properties of the penalized LMR estimator $\widehat{\beta}^{\lambda}$. We first introduce some notations. Without loss of generality, we decompose the true parameter vector β_0 as $\beta_0 = (\beta_{0a}^T, \beta_{0b}^T)^T$, where $\beta_{0a} = (\beta_{01}, \ldots, \beta_{0k_n})^T$ is the vector corresponding to all the nonzero coefficients and $\beta_{0b} = (\beta_{0,k_n+1}, \ldots, \beta_{0p_n})^T$ is the vector corresponding to all the zero coefficients. In the same way, we also decompose $\widetilde{X} = (\widetilde{X}_a^T, \widetilde{X}_b^T)^T$ and $\widehat{\beta}^{\lambda} = ((\widehat{\beta}_a^{\lambda})^T, (\widehat{\beta}_b^{\lambda})^T)^T$. Denote

$$a_n = \max_{1\le j\le p_n} \left\{ \left| p_\lambda'(|\beta_{0j}|) \right|, \beta_{0j} \neq 0 \right\}, \quad \text{and} \quad b_n = \max_{1\le j\le p_n} \left\{ \left| p_\lambda''(|\beta_{0j}|) \right|, \beta_{0j} \neq 0 \right\},$$

$$\mathbf{s}_n = \left\{ p_\lambda'(|\beta_{01}|)\mathrm{sgn}(\beta_{01}), \ldots, p_\lambda'(|\beta_{0k_n}|)\mathrm{sgn}(\beta_{0k_n}) \right\}^T,$$
$$\text{and } \Psi_\lambda = \mathrm{diag}\left\{ p_\lambda''(|\beta_{01}|, \ldots, p_\lambda''(|\beta_{0k_n}|) \right\}.$$

Theorem 4 *Under the regularity conditions (A1)–(A7), if $p_n^2/n \to 0$ as $n \to \infty$, h_1 is a constant and does not depend on n, and the penalty function $p_\lambda(t)$ satisfies $a_n = O(n^{-1/2})$, $b_n \to 0$ as $n \to \infty$, and there are constants M_1 and M_2 such that $\left| p_\lambda''(t) - p_\lambda''(t) \right| \le M_2 \left| t_1 - t_2 \right|$ for any $t_1, t_2 > M_1\lambda$, then we have*

$$\left\| \widehat{\beta}^{\lambda} - \beta_0 \right\| = \mathrm{O}_p \left\{ p_n^{1/2}(n^{-1/2} + a_n) \right\},$$

where $\|\cdot\|$ stands for the Euclidean norm.

Theorem 4 indicates that the penalized LMR estimator $\widehat{\beta}^{\lambda}$ is root-n/p_n consistent with suitable penalty function. Furthermore, the following theorem states the oracle property of the penalized LMR estimator $\widehat{\beta}^{\lambda}$.

Theorem 5 *Under the same conditions as in Theorem 4, if $\lambda \to 0$, $(n/p_n)^{1/2}\lambda \to \infty$ as $n \to \infty$, and the penalty function $p_\lambda(t)$ satisfies*

$$\liminf_{n\to\infty} \liminf_{t\to 0^+} p_\lambda'(t)/\lambda > 0.$$

Then, with probability tending to 1, the root-n/p_n consistent estimator $\widehat{\beta}^{\lambda} = ((\widehat{\beta}_a^{\lambda})^T, (\widehat{\beta}_b^{\lambda})^T)^T$ in Theorem 4 satisfies:

(a) *Sparsity: $\widehat{\beta}_b^{\lambda} = 0$.*
(b) *Asymptotic normality*:

$$\sqrt{n}\left(\Sigma_1^{(1)} + \Psi_\lambda \right) \left\{ \widehat{\beta}_a^{\lambda} - \beta_{0a} + \left(\Sigma_1^{(1)} + \Psi_\lambda \right)^{-1} \mathbf{s}_n \right\} \to N\left(\mathbf{0}, \Sigma_2^{(1)} \right)$$

in distribution, where $\Sigma_1^{(1)}$, $\Sigma_2^{(1)}$ are the submatrices of Σ_1 and Σ_2 corresponding to β_{0a}.

Built upon the results on the penalized LMR estimator $\widehat{\beta}^{\lambda}$, we provide the asymptotic normality of the LMR estimator $\widehat{f}^{\lambda}(z)$ as follows:

Theorem 6 *In addition to the conditions in Theorem 5, the regularity conditions (B1)–(B4) are satisfied. If* $h_2 \to 0$, $nh_2^q \to \infty$ *as* $n \to \infty$, *and* h_3 *is a constant and does not depend on n, then for any fixed* $Z = z \in \mathbb{R}^q$,

$$\left(nh_2^q\right)^{1/2} \left\{\widehat{f}^{\lambda}(z) - f(z) - bias\right\} \to N(0, \tau^2(z)),$$

in distribution, where

$$bias = (1/2) f''(z) h_2^2 \int t^2 K(t) dt,$$
$$\tau^2(z) = \frac{\int K^2(t)dt}{\rho(z)} \frac{\int [\phi'_{h_3}(\varepsilon)]^2 g(\bar{y}|z) d\mu(\bar{y})}{[\varphi''_{h_3}(0|z)]^2}.$$

4 Bandwidth selection and estimation algorithm

In this section, we first discuss the selection of bandwidths both in theory and in practice. Then, BIC criterion is suggested to select the regularization parameter. Finally, we introduce a modified MEM algorithm to obtain our proposed estimators.

4.1 Asymptotic optimal bandwidth

Based on Theorem 2 and the asymptotic variance of the LS estimator given in Zhu et al. (2013), we can that the ratio of the asymptotic variance of the LMR estimator $\widehat{\beta}$ to that of the corresponding LS estimator is given by

$$R(u, h_1) = \frac{G(u, h_1) F^{-2}(u, h_1)}{\sigma^2(u)},$$

where $Var(\varepsilon\,|U = u) = \sigma^2(u)$.

We can see that the ratio $R(u, h_1)$ depends only on u and h_1, and it plays an important role in efficiency and robustness of the LMR estimator $\widehat{\beta}$. Therefore, the asymptotic optimal bandwidth of h_1 can be chosen as

$$h_{1,\text{opt}} = \arg\min_{h_1} R(u, h_1) = \arg\min_{h_1} G(u, h_1) F^{-2}(u, h_1),$$

which indicates that $h_{1,\text{opt}}$ does not depends on n and only depends on the conditional error distribution of ε given U. It follows from Theorem 2.4 in Yao et al. (2012) that $\inf_{h_1} R(u, h_1) \leq 1$ for any error distribution, and if the error term follows normal distribution, $R(u, h_1) > 1$ and $\inf_{h_1} R(u, h_1) = 1$. That is to say, the LMR estimator $\widehat{\beta}$ is more efficient than the corresponding LS estimator when there exist outliers and heavy-tail error distribution and does not lose efficiency under normal error distribution.

Similarly, the asymptotic optimal bandwidth of h_3 is given by

$$h_{3,\text{opt}} = \arg\min_{h_3} R(u, h_3) = \arg\min_{h_3} G(u, h_3) F^{-2}(u, h_3).$$

Moreover, according to Theorem 3, we can see that the optimal bandwidth for h_2 is of order $n^{-1/(4+q)}$. If $q = 1$, the order of the optimal bandwidth for h_2 is $n^{-1/5}$, which is a common requirement for semiparametric models in the kernel literature.

4.2 Bandwidth selection in practice

In this subsection, we will address how to select the bandwidths for the LMR estimators in practice. For simplicity, we further assume that ε is independent of U. Thus, we can estimate $F(u, h_1)$ and $G(u, h_1)$ by

$$\widehat{F}(h_1) = \frac{1}{n} \sum_{i=1}^{n} \phi''_{h_1}(\widehat{\varepsilon}_i) \quad \text{and} \quad \widehat{G}(h_1) = \frac{1}{n} \sum_{i=1}^{n} \{\phi'_{h_1}(\widehat{\varepsilon}_i)\}^2,$$

respectively, where $\widehat{\varepsilon}_i = \widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_i^T \widehat{\beta}$.

Then, we can estimate $R(h_1)$ by $\widehat{R}(h_1) = \widehat{G}(h_1)\widehat{F}(h_1)^{-2}/\widehat{\sigma}^2$, where $\widehat{\sigma}$ is estimated based on the pilot estimator. In this paper, we use the grid search method to find $h_{1,\text{opt}}$ to minimize $\widehat{R}(h_1)$. According to the suggestion of Yao et al. (2012), the possible grids points for h_1 can be $0.5\widehat{\sigma} \times 1.02^j$, $j = 0, 1, \ldots, k$, for some fixed k, such as $k = 50$ or $k = 100$. Similarly, we can find the optimal bandwidth for h_3.

Finally, we suggest to use k-fold Cross-Validation (Breiman 1995) or generalized Cross-Validation (Tibshirani 1996) to select the optimal bandwidth for h_2.

4.3 Selection of regularization parameter

To produce a sparse estimator of β_0, it remains to select the regularization parameter λ. In this paper, we propose a modified BIC-type criterion to choose the optimal λ, which minimizes the following objective function

$$\text{BIC}(\lambda) = -\log\left\{\frac{1}{n}\sum_{\text{i=1}}^{\text{n}} \phi_{h_1}\left(\widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_i^T \widehat{\beta}^{\lambda}\right)\right\} + \frac{\log(n)}{n} df_{\lambda}, \tag{12}$$

where df_{λ} is the number of non-zero coefficients of $\widehat{\beta}^{\lambda}$.

4.4 Estimation algorithm

In this subsection, we propose a modified modal expectation-maximization (MEM) algorithm, proposed by Li et al. (2007), for the proposed estimators. The algorithm has two iterative steps similar to the expectation and the maximization steps in EM (Dempster et al. 1977).

Step 1 We first calculate the LMR estimator $\widehat{\beta}$. Let $\widehat{\beta}^{(0)} = (\widehat{\beta}_1^{(0)}, \ldots, \widehat{\beta}_{p_n}^{(0)})^T$ be the initial value and set $k = 0$.
(E-step): Update $\pi_1(j|\widehat{\beta}^{(k)})$ by

$$\pi_1(j|\widehat{\beta}^{(k)}) = \frac{\phi_{h_1}\left(\widehat{\widehat{Y}}_j - \widehat{\widehat{X}}_j^T \widehat{\beta}^{(k)}\right)}{\sum_{i=1}^n \phi_{h_1}\left(\widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_j^T \widehat{\beta}^{(k)}\right)}, \quad j = 1, 2, \ldots, n.$$

item (M-step): Update β to obtain $\widehat{\beta}^{(k+1)}$ by

$$\begin{aligned}\widehat{\beta}^{(k+1)} &= \arg\max_{\beta} \sum_{i=1}^n \left\{\pi_1\left(i|\widehat{\beta}^{(k)}\right) \log\phi_{h_1}\left(\widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_j^T \beta\right)\right\} \\ &= \left(\widehat{\widehat{\mathbf{X}}}^T W_1^{(k)} \widehat{\widehat{\mathbf{X}}}\right)^{-1} \widehat{\widehat{\mathbf{X}}}^T W_1^{(k)} \widehat{\widehat{\mathbf{Y}}},\end{aligned}$$

where $\widehat{\widehat{\mathbf{X}}} = (\widehat{\widehat{X}}_1, \ldots, \widehat{\widehat{X}}_n)^T, \widehat{\widehat{\mathbf{Y}}} = (\widehat{\widehat{Y}}_1, \ldots, \widehat{\widehat{Y}}_n)^T$ and $W_1^{(k)} = \text{diag}\{\pi_1(1|\widehat{\beta}^{(k)}), \ldots, \pi_1(n|\widehat{\beta}^{(k)})\}$. Iterate the E-step and M-step until convergence.

Remark 3 When $\phi(t)$ is the standard normal density, the M-step has a unique maximum. Similar as in the EM algorithm, it is usually much easier to maximize $\sum_{i=1}^n \{\pi_1(i|\beta)\log\phi_{h_1}(\widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_i \beta)\}$ than the original objective function (6).

Step 2 Similarly, we then calculate the LMR estimator $\widehat{f}(z)$. Let $\widehat{\theta}^{(0)} = (\widehat{a}^{(0)T}, \widehat{b}^{(0)T})^T$ be the initial value and set $k = 0$.
(E-step): Update $\pi_2(j|\widehat{\theta}^{(k)})$ by

$$\pi_2\left(j|\widehat{\theta}^{(k)}\right) = \frac{K\left(\frac{Z_j - Z}{h_2}\right) \phi_{h_3}\left(Y_j^* - X_j^{*T} \widehat{\theta}^{(k)}\right)}{\sum_{i=1}^n K\left(\frac{Z_i - Z}{h_2}\right) \phi_{h_3}\left(Y_i^* - X_i^{*T} \widehat{\theta}^{(k)}\right)}, \quad j = 1, 2, \ldots, n,$$

where $Y_i^* = Y_i - X_i^T \widehat{\beta}$, and $X_i^* = \{1, (Z_i - Z)^T\}^T$.
(M-step): Update θ to obtain $\widehat{\theta}^{(k+1)}$ by

$$\begin{aligned}\widehat{\theta}^{(k+1)} &= \arg\max_{\beta} \sum_{i=1}^n \left\{\pi_2\left(i|\widehat{\theta}^{(k)}\right) \log\phi_{h_3}\left(Y_i^* - X_i^{*T} \theta\right)\right\} \\ &= \left(\mathbf{X}^{*T} W_2^{(k)} \mathbf{X}^*\right)^{-1} \mathbf{X}^{*T} W_2^{(k)} \mathbf{Y}^*,\end{aligned}$$

where $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)^T$, $\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*)^T$ and $W_2^{(k)} = \text{diag}\{\pi_2(1|\widehat{\theta}^{(k)}), \ldots, \pi_2(n|\widehat{\theta}^{(k)})\}$. Iterate the E-step and M-step until convergence.

With the aid of the LQA algorithm, we can obtain the penalized LMR estimator $\widehat{\beta}^{\lambda}$ by a revision of Step 1 as follows:

Step 1' Let $\widehat{\beta}^{\lambda(0)} = (\widehat{\beta}_1^{\lambda(0)}, \ldots, \widehat{\beta}_{p_n}^{\lambda(0)})^T$, and set $k = 0$.
(E-step): Update $\pi_1(j|\widehat{\beta}^{\lambda(k)})$ by

$$\pi_1\left(j|\widehat{\beta}^{\lambda(k)}\right) = \frac{\phi_{h_1}\left(\widehat{\widehat{Y}}_j - \widehat{\widehat{X}}_j^T \widehat{\beta}^{\lambda(k)}\right)}{\sum_{i=1}^n \phi_{h_1}\left(\widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_j^T \widehat{\beta}^{\lambda(k)}\right)}, \quad j = 1, 2, \ldots, n.$$

(M-step): Update β to obtain $\widehat{\beta}^{\lambda(k+1)}$ by

$$\begin{aligned}\widehat{\beta}^{\lambda(k+1)} &= \arg\max_{\beta} \sum_{i=1}^n \Bigg\{ \pi_1\left(i|\widehat{\beta}^{\lambda(k)}\right) \log\phi_{h_1}\left(\widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_j^T \widehat{\beta}^{\lambda(k)}\right) \\ &\quad - \frac{n}{2}\sum_{j=1}^{p_n}\left\{p'_\lambda\left(\widehat{\beta}_j^{\lambda(k)}\right)/|\widehat{\beta}_j^{\lambda(k)}|\right\}\beta_j^2\Bigg\} \\ &= \left\{\widehat{\widehat{\mathbf{X}}}^T W_1^{(k)} \widehat{\widehat{\mathbf{X}}} + n\Sigma_\lambda\left(\widehat{\beta}^{\lambda(k)}\right)\right\}^{-1} \widehat{\widehat{\mathbf{X}}}^T W_1^{(k)} \widehat{\widehat{\mathbf{Y}}},\end{aligned}$$

where $\Sigma_\lambda(\beta) = \mathrm{diag}\left\{p'_\lambda(|\beta_1|), \ldots, p'_\lambda(|\beta_{p_n}|)\right\}$, $\widehat{\widehat{\mathbf{X}}}$, $\widehat{\widehat{\mathbf{Y}}}$, and $W_1^{(k)}$ are the same as in Step 1. Iterate the E-step and M-step until convergence.

5 Simulation studies

5.1 Monte Carlo simulations

In this subsection, we will provide some Monte Carlo simulations to evaluate the finite-sample performance of the proposed estimators in terms of robust estimation and variable selection. Without loss of generality, we consider following four models:

(I): $Y = X^T\beta + 2\sin(\gamma^T Z) + \varepsilon$;
(II): $Y = X^T\beta + \left|(\gamma^T Z) + 1\right| + \varepsilon$;
(III): $Y = X^T\beta + \exp\{(\gamma^T Z)/2\}/2 + \varepsilon$;
(IV): $Y = X^T\beta + 2(\gamma^T Z) + \varepsilon$.

In each model, we generate $Z = (z_1, z_2)^T$ from a two-dimensional normal distribution with mean zero and identity covariance matrix, and X from $X_j = \gamma^T Z + 2e_j$ for $j = 1, 2, \ldots, p_n$, where $\gamma = (1/\sqrt{2}, 1/\sqrt{2})^T$ and e_j's are independently generated from standard normal distribution. Similar models were also considered in Zhu et al. (2013). We choose $\beta = (\mathbf{1}_{k_n}, \mathbf{0}_{p_n - k_n})^T$, where $k_n = p_n/4$, indicating that the size of significant parametric covariates is also diverging with the sample size. To examine the robustness and efficiency of our proposed LMR estimators, we compare the simulation results with corresponding LS estimators and LAD estimators. In our simulations, we considered the following error distributions: $N(0, 1)$ distribution; $t(3)$ distribution which is used to produce heavy-tailed error distribution; mixed normal distribution $0.9N(0, 1)+0.1N(0, 10)$ which is used to produce outliers. Similar error distributions were also considered in Yao et al. (2012), Zhang et al. (2013), and Zhao et al. (2015). The simulations are repeated 200 times with sample size $n = 400$ and dimension $p_n = 2n^{1/2} = 40$.

For sake of evaluation, the estimation accuracy of the parametric estimators is measured by the mean and median of the mean squared errors (MeanMSE and MedianMSE) over the 200 simulated datasets. Meanwhile, the performance of the nonparametric estimators is assessed by the median absolute prediction error (MAPE), which is defined by

$$\text{MAPE} = median \left\{ \left| \widehat{f}(Z_i) - f(Z_i) \right|, i = 1, 2, \ldots, n \right\},$$

and the sample mean and standard deviation (SD) of the MAPE's are presented in the last columns of Tables 1 and 2. In addition, we adopt the notation (C, IC) to identify the performance of the variable selection in Table 3. Here C means the average number of zero regression coefficients that are correctly estimated as zero, IC presents the average number of non-zero regression coefficients that are incorrectly set to zero.

From Table 1, we can see that the LMR estimator $\widehat{\beta}$ performs best in the case of non-normal error distributions, and as asymptotically efficient as the corresponding LS estimator when the error is normal distribution. Meanwhile, the LMR estimator $\widehat{f}(z)$ seems to perform no worse than the corresponding LAD estimator for t_3 error distribution and 10% outliers, and is comparable to the corresponding LS estimator when the error is drawn from the normal distribution. Similar conclusion can be draw from Table 2 for the penalized LMR procedure. Furthermore, the penalized LMR estimator $\widehat{\beta}^{\lambda}$ outperforms the other estimators in terms of C and IC. Finally, we provide the plots of the LMR estimator $\widehat{f}(z)$. For illustration, we only present the figures when the robust LMR procedure is applied for model (I). From Figs. 1, 2, and 3, we find the LMR estimator $\widehat{f}(z)$ performs favorably compared to other estimators.

5.2 A real data example

As an illustration, we apply the proposed procedures to analyze an automobile dataset (Johnson 2003), which had been analyzed by Zhu et al. (2012) and Zhu et al. (2013). We intend to investigate how the manufactures suggested retail price (MSRP) of vehicles depends upon different factors. Thus, it is appropriate to treat the MSRP as the response variable (Y). This dataset contains 412 available observations in total, after removing sixteen observations with missing values. There are seven important factors which possibly affect the MSRP of vehicles, such as engine size (x_1), number of cylinders (x_2), horsepower (x_3), weight in pounds (x_4), wheel base in inches (x_5), average city miles per gallon (z_1) and average highway miles per gallon (z_2). Similar as Zhu et al. (2013), we choose $Z = (z_1, z_2)^T$ as the nonparametric covariate. In the subsequent

Table 1 Simulation results for robust estimation procedure

Model	Error	Method	MeanMSE	MedianMSE	MAPE
(I)	$N(0, 1)$	LS	0.1193	0.1173	0.2507(0.0923)
		LAD	0.1825	0.1799	0.2837(0.0824)
		LMR	0.1211	0.1190	0.2517(0.0925)
	$t(3)$	LS	0.3349	0.3133	0.4443(0.1981)
		LAD	0.2589	0.2530	0.4315(0.2038)
		LMR	0.2381	0.2182	0.4343(0.2071)
	$0.9N(0, 1) + 0.1N(0, 10)$	LS	0.2270	0.2223	0.3284(0.1171)
		LAD	0.2324	0.2273	0.3265(0.1103)
		LMR	0.1775	0.1657	0.3141(0.1171)
(II)	$N(0, 1)$	LS	0.1216	0.1204	0.2611(0.1199)
		LAD	0.1875	0.1834	0.2923(0.1120)
		LMR	0.1278	0.1237	0.2618(0.1199)
	$t(3)$	LS	0.3553	0.3309	0.3957(0.1502)
		LAD	0.2686	0.2547	0.3745(0.1521)
		LMR	0.2371	0.2155	0.3700(0.1550)
	$0.9N(0, 1) + 0.1N(0, 10)$	LS	0.2451	0.2334	0.3121(0.1131)
		LAD	0.2293	0.2261	0.3019(0.1040)
		LMR	0.1754	0.1645	0.2943(0.1108)
(III)	$N(0, 1)$	LS	0.1217	0.1202	0.2622(0.1231)
		LAD	0.1870	0.1860	0.2930(0.1146)
		LMR	0.1265	0.1238	0.2627(0.1227)
	$t(3)$	LS	0.3280	0.3107	0.3921(0.1514)
		LAD	0.2465	0.2360	0.3654(0.1626)
		LMR	0.2181	0.2000	0.3673(0.1566)
	$0.9N(0, 1) + 0.1N(0, 10)$	LS	0.2165	0.2159	0.3805(0.1321)
		LAD	0.2175	0.2114	0.3815(0.1313)
		LMR	0.1677	0.1593	0.3728(0.1363)
(IV)	$N(0, 1)$	LS	0.1221	0.1185	0.2350(0.0778)
		LAD	0.1918	0.1866	0.2709(0.0720)
		LMR	0.1264	0.1217	0.2358(0.0780)
	$t(3)$	LS	0.3337	0.2973	0.3875(0.1203)
		LAD	0.2520	0.2506	0.3605(0.1297)
		LMR	0.2390	0.2238	0.3622(0.1173)
	$0.9N(0, 1) + 0.1N(0, 10)$	LS	0.2340	0.2289	0.3263(0.1302)
		LAD	0.2332	0.2251	0.3208(0.1210)
		LMR	0.1742	0.1705	0.3105(0.1315)

Table 2 Simulation results for penalized estimation procedure

Model	Error	Method	MeanMSE	MedianMSE	MAPE
(I)	$N(0, 1)$	LS	0.0293	0.0266	0.2056(0.0427)
		LAD	0.0704	0.0594	0.2426(0.0435)
		LMR	0.0452	0.0274	0.2058(0.0429)
	$t(3)$	LS	0.0847	0.0776	0.3166(0.0624)
		LAD	0.0842	0.0763	0.3014(0.0581)
		LMR	0.0560	0.0589	0.2978(0.0608)
	$0.9N(0, 1) + 0.1N(0, 10)$	LS	0.0564	0.0491	0.2689(0.0607)
		LAD	0.0845	0.0687	0.2717(0.0532)
		LMR	0.0407	0.0369	0.2506(0.0598)
(II)	$N(0, 1)$	LS	0.0308	0.0299	0.2045(0.0425)
		LAD	0.0689	0.0578	0.2372(0.0380)
		LMR	0.0307	0.0293	0.2049(0.0419)
	$t(3)$	LS	0.0913	0.0838	0.3107(0.0641)
		LAD	0.0869	0.0713	0.2865(0.0550)
		LMR	0.0526	0.0476	0.2827(0.0589)
	$0.9N(0, 1) + 0.1N(0, 10)$	LS	0.0591	0.0570	0.2614(0.0500)
		LAD	0.0829	0.0725	0.2616(0.0531)
		LMR	0.0417	0.0377	0.2452(0.0497)
(III)	$N(0, 1)$	LS	0.0308	0.0298	0.2050(0.0439)
		LAD	0.0702	0.0596	0.2382(0.0396)
		LMR	0.0308	0.0297	0.2051(0.0427)
	$t(3)$	LS	0.0813	0.0736	0.3069(0.0625)
		LAD	0.0788	0.0612	0.2842(0.0497)
		LMR	0.0444	0.0412	0.2848(0.0582)
	$0.9N(0, 1) + 0.1N(0, 10)$	LS	0.0524	0.0492	0.3074(0.0483)
		LAD	0.0779	0.0655	0.3236(0.0507)
		LMR	0.0378	0.0355	0.3006(0.0490)
(IV)	$N(0, 1)$	LS	0.0285	0.0262	0.2079(0.0416)
		LAD	0.0830	0.0622	0.2512(0.0498)
		LMR	0.0292	0.0268	0.2076(0.0417)
	$t(3)$	LS	0.0769	0.0722	0.3017(0.0576)
		LAD	0.0839	0.0589	0.2817(0.0552)
		LMR	0.0482	0.0448	0.2783(0.0498)
	$0.9N(0, 1) + 0.1N(0, 10)$	LS	0.0554	0.0516	0.2667(0.0531)
		LAD	0.0792	0.0634	0.2695(0.0462)
		LMR	0.0397	0.0388	0.2517(0.0513)

Table 3 Simulation results for variable selection

Model	Error	Method	C	IC
(I)	$N(0, 1)$	LS	29.9000	0
		LAD	29.3200	0
		LMR	29.9800	0
	$t(3)$	LS	29.9600	0
		LAD	29.2800	0
		LMR	30.0000	0
	$0.9N(0, 1) + 0.1N(0, 10)$	LS	29.9700	0
		LAD	29.3000	0
		LMR	30.0000	0
(II)	$N(0, 1)$	LS	29.8800	0
		LAD	29.3100	0
		LMR	29.9700	0
	$t(3)$	LS	29.7700	0
		LAD	29.5400	0
		LMR	30.0000	0
	$0.9N(0, 1) + 0.1N(0, 10)$	LS	29.8400	0
		LAD	29.1600	0
		LMR	29.9800	0
(III)	$N(0, 1)$	LS	29.8800	0
		LAD	29.3300	0
		LMR	29.9700	0
	$t(3)$	LS	29.7800	0
		LAD	29.4000	0
		LMR	30.0000	0
	$0.9N(0, 1) + 0.1N(0, 10)$	LS	29.8900	0
		LAD	29.3200	0
		LMR	30.0000	0
(IV)	$N(0, 1)$	LS	29.9200	0
		LAD	29.4200	0
		LMR	29.9600	0
	$t(3)$	LS	29.9400	0
		LAD	29.4600	0
		LMR	29.9800	0
	$0.9N(0, 1) + 0.1N(0, 10)$	LS	29.8700	0
		LAD	29.5100	0
		LMR	30.0000	0

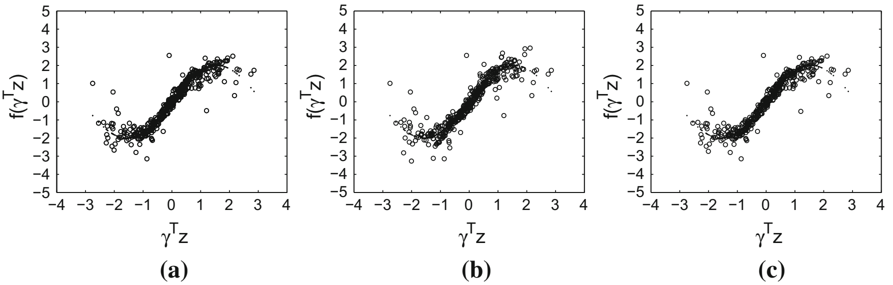

Fig. 1 Plots for the nonparametric estimators ($N(0, 1)$ error distribution): the true curve (dotted line), the estimated values (small circles) **a** LS, **b** LAD, **c** LMR

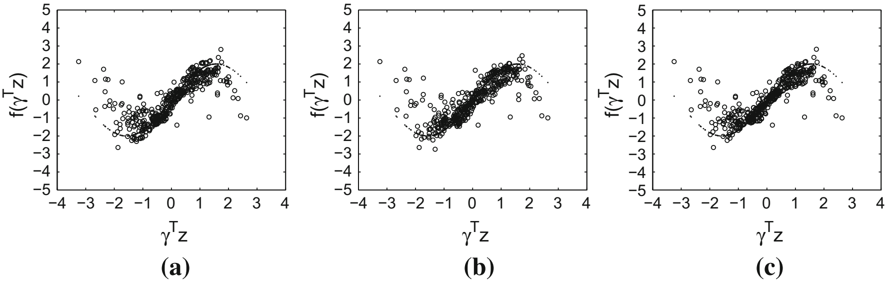

Fig. 2 Plots for the nonparametric estimators ($t(3)$ error distribution): the true curve (dotted line), the estimated values (small circles) **a** LS, **b** LAD, **c** LMR

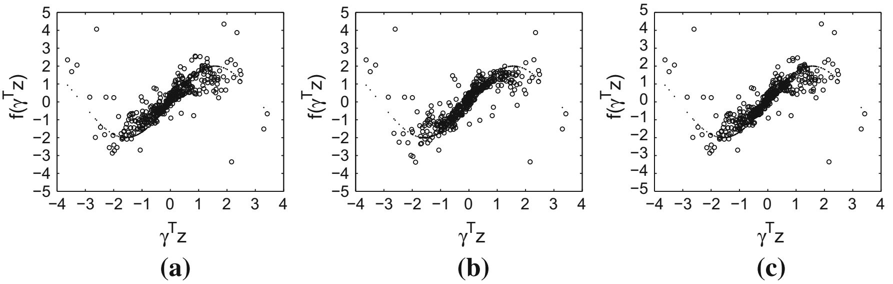

Fig. 3 Plots for the nonparametric estimators ($0.9N(0, 1) + 0.1N(0, 10)$ error distribution): the true curve (dotted line), the estimated values (small circles) **a** LS, **b** LAD, **c** LMR

analysis, we first standardize the response variable Y and the parametric covariate vector $X = (x_1, \ldots, x_5)^T$, respectively.

From the histogram and boxplot of the standardized Y presented in Figs. 4 and 5, we can see that the distribution of the standardized Y is highly skewed and that there exist a number of outliers in the standardized Y. Then, we apply the three different estimation procedures (LS, LAD, LMR) to analyze the dataset by a partially linear model stated as (1). The prediction performance is measured by the median absolute prediction error (MAPE), which is the median of $\{|Y_i - \widehat{Y_i}|, i = 1, 2, \ldots, 412\}$. The corresponding estimation results are summarized in Table 4, from which we can see

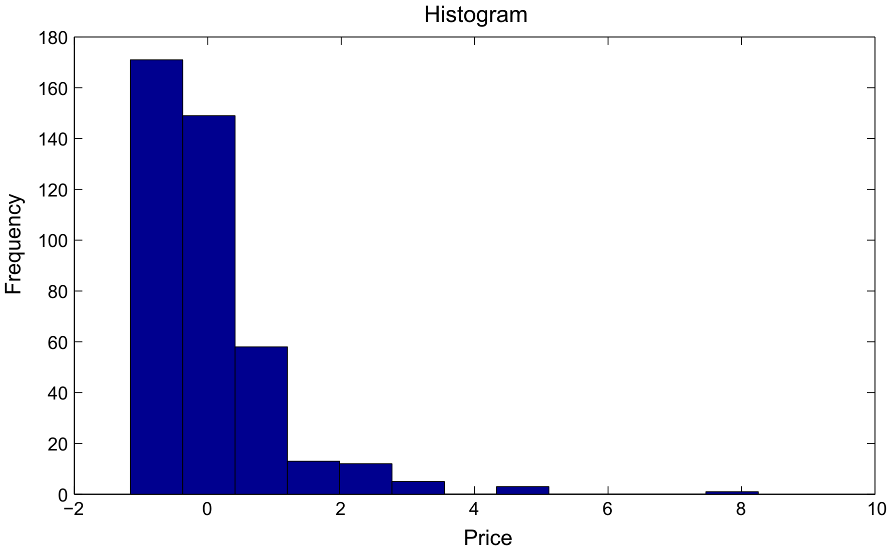

Fig. 4 Histogram of the standardized Y for real data

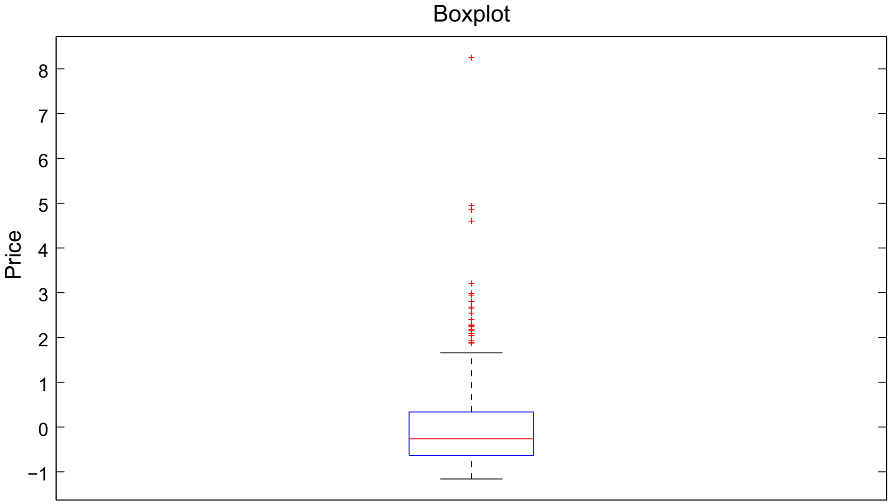

Fig. 5 Boxplot of the standardized Y for real data

that the LMR method has smaller MAPE than the other methods in the presence of outliers. Based on the previous analysis by Zhu et al. (2013), we further include other seven binary variables as auxiliary covariates, which have little contributions to the MSRP of vehicles. These auxiliary covariates are sport car (x_6), sport utility vehicle (x_7), wagon (x_8), mini-van (x_9), pickup (x_{10}), all-wheel drive (x_{11}), and rear-wheel drive (x_{12}). In this way, the dimension of the covariate vector X becomes 12, and then we consider the problem of variable selection. From Table 5, we can see that

Table 4 Estimation for real data

	LS	LAD	LMR
x_1	−0.2770	−0.2293	−0.2915
x_2	0.1709	0.1169	0.0871
x_3	0.8556	0.5957	0.5066
x_4	0.2501	0.3171	0.4087
x_5	−0.2390	−0.1731	−0.1760
MAPE	0.2743	0.2198	0.1949

Table 5 Variable selection for real data

	LS	LAD	LMR
x_1	0	0	−0.1970
x_2	0	0	0
x_3	0.8362	0.5300	0.6532
x_4	0.2368	0.4051	0.4176
x_5	−0.2775	−0.2462	−0.1732
x_6	0	0.0256	0
x_7	0	0	0
x_8	0	0	0
x_9	0	0	0
x_{10}	0	0	0
x_{11}	0	0	0
x_{12}	0	0.1031	0
MAPE	0.2668	0.2029	0.2114

all three penalized procedures identify x_3, x_4, and x_5 as important variables, which coincides with Zhu et al. (2012) that x_3 is seems to be the most important factor that affects the MSRP, followed by x_4. In addition, the penalized LMR procedure selects x_1 as important variable, which has negative connection with the MSRP. However, the LAD penalized procedure selects two auxiliary covariates, x_6 and x_{12}. To conclude, the penalized LMR procedure is better than the penalized LS procedure in terms of MAPE, and is sparse than the penalized LAD procedure by shrinking all the auxiliary covariates to zeros.

6 Extension

In this section, we further discuss how the penalized LMR procedure can be applied to ultra-high dimensional data in which $p_n > n$. Based on correlation learning, Fan and Lv (2008) proposed sure independence screening (SIS) to reduce dimensionality from high to a moderate scale that is below the sample size. They further established the sure screening property for SIS. Inspired by the results of Fan and Lv (2008), we propose a two-stage approach combined SIS and penalized LMR to deal with ultra-high dimensional data. We first apply SIS to reduce the model dimensions to $d_n = o\left(\sqrt{n}\right)$

Table 6 Simulation results for ultra-high dimensional data based on 100 replications

Error	Method	MeanMSE	MedianMSE	MAPE	C	IC
$N(0,1)$	SIS + LS	0.0340	0.0329	0.2035(0.0448)	989.3000	0.2200
	SIS + LAD	0.0828	0.0638	0.2443(0.0469)	988.8200	0.2200
	SIS + LMR	0.0356	0.0309	0.2041(0.0450)	989.6600	0.2200
$t(3)$	SIS + LS	0.1140	0.0718	0.2936(0.0571)	989.0400	0.2600
	SIS + LAD	0.0827	0.0669	0.2743(0.0470)	988.6000	0.2800
	SIS + LMR	0.0545	0.0452	0.2665(0.0526)	989.6200	0.3000
$0.9N(0,1) + 0.1N(0,10)$	SIS + LS	0.0602	0.0506	0.2771(0.1144)	989.5400	0.1500
	SIS + LAD	0.0827	0.0698	0.2746(0.0987)	989.0700	0.1500
	SIS + LMR	0.0415	0.0352	0.2587(0.1102)	989.8400	0.1500

and then fit the data using the penalized LMR to obtain the final estimation. We call this two-step procedure SIS + LMR.

To demonstrate SIS + LMR, we consider the commonly used model (I) in Sect. 5.1, except that $p_n = 1000$. From Table 6, we see that when the error is normally distributed, SIS + LMR is comparable to SIS + LS. However, in the case of 10% outliers or t_3 error distribution, SIS + LMR outperforms the other procedures in terms of parameter estimation and variable selection.

7 Concluding discussion

In this paper, we adopt the local modal regression for robust estimation in partially linear models with large-dimensional covariates. We show that the resulting estimators for both parametric and nonparametric components are more efficient in the case of outliers or heavy-tail error distribution, and as asymptotically efficient as the corresponding least squares estimators when there are no outliers and the error is normal distribution. We also develop the variable selection procedure to select significant parametric covariates and and establish its oracle property under mild regularity conditions.

In some applications, the parametric covariate X may be in much higher dimension than $o\left(\sqrt{n}\right)$. To this end, we introduce a robust two-step approach based on the idea of sure independence screening procedure to deal with ultra-high dimensional data. However, how to apply the proposed procedure directly in such high dimensional scenario without feature screening is of both theoretical and practical importance.

Furthermore, the proposed procedures may encounter the curse of dimensionality when the dimension of nonparametric covariate Z is large, although we complete the theoretical results when Z is allowed to be multivariate. In practice, we may consider the partially linear single-index model to solve this issue. But such an extension is by no means of trivial and needs additional investigations in the future.

Acknowledgements This work is supported by the National Natural Science Foundation of China (Grant No. 11671059).

Appendix: Proofs of Theorems

Proof of Theorem 1 We want to show that for any given $\delta > 0$, there exists a large constant C such that

$$P\left\{\sup_{\|\mathbf{v}\|=\mathbf{C}} R(\beta_0 + \mu_n \mathbf{v}) < R(\beta_0)\right\} \geq 1 - \delta, \tag{13}$$

where $R(\beta) = \frac{1}{n}\sum\limits_{i=1}^{n} \phi_{h_1}(\widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_i^T \beta)$, $\mu_n = p_n{}^{1/2} n^{-1/2}$. For simplicity, we define $D_n(\mathbf{v}) = R(\beta_0 + \mu_n \mathbf{v}) - R(\beta_0)$ and then obtain that

$$\begin{aligned}
D_n(\mathbf{u}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\phi_{h_1}\left(\widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_i^T (\beta_0 + \mu_n \mathbf{v})\right) - \phi_{h_1}\left(\widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_i^T \beta_0\right)\right\}\\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\phi_{h_1}\left(\varepsilon_i - \sigma_i - \mu_n \mathbf{v}^T \widehat{\widehat{X}}_i\right) - \phi_{h_1}(\varepsilon_i - \sigma_i)\right\}\\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{-\phi'_{h_1}(\varepsilon_i - \sigma_i)\left(\mu_n \mathbf{v}^T \widehat{\widehat{X}}_i\right) + \frac{1}{2}\phi''_{h_1}(\varepsilon_i - \sigma_i)\left(\mu_n \mathbf{v}^T \widehat{\widehat{X}}_i\right)^2\right.\\
&\quad \left. -\frac{1}{6}\phi'''_{h_1}(t_i)\left(\mu_n \mathbf{v}^T \widehat{\widehat{X}}_i\right)^3\right\}\\
&= I_1 + I_2 + I_3,
\end{aligned}$$

where $\sigma_i = \widetilde{Y}_i - \widehat{\widehat{Y}}_i - (\widetilde{X}_i - \widehat{\widehat{X}}_i)^T \beta_0$, and t_i is between $\varepsilon_i - \sigma_i$ and $\varepsilon_i - \sigma_i - \mu_n \mathbf{v}^T \widehat{\widehat{X}}_i$. By the regularity conditions (A1)–(A4), the consistence of the kernel estimation implies that $\max_{1 \leq i \leq n} \|\sigma_i\| = o_p(1)$ almost surely, which will repeatedly be used in our proof. A detailed discussion on this argument can be found in the Lemma 3.5.1 and Lemma A.1 of Hardle et al. (2000).

Based on the fact $\xi = E(\xi) + \mathrm{O}_p(\sqrt{Var(\xi)})$, the regularity condition (A7) and the uniform convergence of $\mathbf{v}^T \widehat{\widehat{X}}_i$ entail that $I_1 = \mathrm{O}_p(\frac{C\mu_n}{\sqrt{n}})$. One can refer to Rao (1983) and Zhu and Fang (1996) for these technical details. Similarly, we have $I_3 = \mathrm{O}_p(C^3 \mu_n{}^3)$. For I_2, we have $I_2 = \frac{1}{2}\mu_n{}^2 \mathbf{v}^T \Sigma_1 \mathbf{v}(1 + o_p(1))$, where $\Sigma_1 = E\{F(U, h_1)\widetilde{X}\widetilde{X}^T\}$. By the condition $p_n^2/n \to 0$ as $n \to \infty$, we can show that $I_1 = o_p\,(I_2)$, and $I_3 = o_p\,(I_2)$. Similar practice can been in Li et al. (2011).

By the regularity condition (A6), $F(\mathbf{v}, h_1) < 0$; hence, Σ_1 is a negative matrix. Noting $\|\mathbf{v}\| = C$, we can get C large enough such that I_2 dominates both I_1 and I_3 with a probability of at least $1 - \delta$. It follows that Eq. (13) holds. Hence, $\widehat{\beta}$ is a root-n/p_n consistent estimator of β. □

Proof of Theorem 2 Let $\widehat{\gamma}_i = \widehat{\widehat{X}}_i^T(\widehat{\beta}-\beta_0)$. If $\widehat{\beta}$ maximizes Eq. (6), then $\widehat{\beta}$ satisfies the following equation:

$$\begin{aligned} 0 &= \sum_{i=1}^n \widehat{\widehat{X}}_i \phi'_{h_1}\left(\widehat{\widehat{Y}}_i - \widehat{\widehat{X}}_i^T \widehat{\beta}\right) = \sum_{i=1}^n \widehat{\widehat{X}}_i \phi'_{h_1}(\varepsilon_i - \sigma_i - \widehat{\gamma}_i) \\ &= \sum_{i=1}^n \widehat{\widehat{X}}_i \left\{ \phi'_{h_1}(\varepsilon_i) - \phi''_{h_1}(\varepsilon_i)(\sigma_i + \widehat{\gamma}_i) + \frac{1}{2}\phi'''_{h_1}(\varepsilon_i^*)(\sigma_i + \widehat{\gamma}_i)^2 \right\} \\ &= I_4 + I_5 + I_6, \end{aligned}$$

where ε_i^* is between ε_i and $\varepsilon_i - \sigma_i - \widehat{\gamma}_i$.
For I_5, we have

$$\begin{aligned} -\sum_{i=1}^n \widehat{\widehat{X}}_i \phi''_{h_1}(\varepsilon_i)(\sigma_i + \widehat{\gamma}_i) &= -\sum_{i=1}^n \phi''_{h_1}(\varepsilon_i)\left\{ \widehat{\widehat{X}}_i \widehat{\widehat{X}}_i^T (\widehat{\beta} - \beta_0) + o_p(1) \right\} \\ &= -n\Sigma_1(\widehat{\beta} - \beta_0) + o_p(1), \end{aligned}$$

where $\Sigma_1 = E\{F(u, h_1)\widetilde{X}\widetilde{X}^T\}$, and the last equality is derived from the regularity conditions (A5) and (A7).
Based on $|\widehat{\gamma}_i|^2 = \mathrm{O}_p(\|\widehat{\beta} - \beta_0\|^2)$ and $p_n^2/n \to 0$ as $n \to \infty$, we have $I_6 = o_p(I_5)$. It can be shown, by easy calculation, that $\sqrt{n}(\widehat{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\Sigma_1^{-1}\sum_{i=1}^n \widehat{\widehat{X}}_i \phi'_{h_1}(\varepsilon_i) + o_p(1)$.
Note that $E\left(\phi'_h(\varepsilon)\,|U = u\right) = 0$, and by the central limit theorem, we have

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}),$$

where $\Sigma_2 = Var\{\widetilde{X}\phi'_{h_1}(\varepsilon)\}$. This completes the proof. □

Proof of Theorem 3 Since Theorem 3 is parallel to Theorem 6, we will only present detailed proof for Theorem 6. □

Proof of Theorem 4 It is sufficient to show that for any given $\delta > 0$, there exists a large constant C such that

$$P\left\{ \sup_{\|\mathbf{v}\|=C} Q_\lambda(\beta_0 + \omega_n \mathbf{v}) < Q_\lambda(\beta_0) \right\} \geq 1 - \delta, \tag{14}$$

where $\omega_n = p_n^{1/2}(n^{-1/2} + a_n)$.
Let $I_7 = -\sum_{j=1}^{k_n}\{p_\lambda(|\beta_{0j} + \omega_n v_j|) - p_\lambda(|\beta_{0j}|)\}$, where k_n is the number of components of β_{0a}. Note that $p_\lambda(0) = 0$ and $p_\lambda(|\beta_j|) \geq 0$ for all β_j. By the proof of Theorem 1, we have

$$\frac{1}{n}\{Q_\lambda(\beta_0 + \omega_n \mathbf{v}) - Q_\lambda(\beta_0)\} \leq \bar{I}_1 + \bar{I}_2 + \bar{I}_3 + I_7, \tag{15}$$

where $\bar{I}_1$, $\bar{I}_2$, and $\bar{I}_3$ are the same as I_1, I_2 and I_3 except the factor μ_n replaced by ω_n.

By the Taylor expansion and the Cauchy–Schwarz inequality, I_7 is bounded by

$$\sqrt{k_n}\omega_n a_n \|\mathbf{v}\| + \omega_n^2 b_n \|\mathbf{v}\|^2.$$

Consequently, as $b_n \to 0$, I_7 is dominated by $\bar{I}_2 = \frac{1}{2}\omega_n^2 \mathbf{v}^T \Sigma_1 \mathbf{v}(1+o_p(1))$, provided C is taken to be sufficiently large. Hence, for large C, $\bar{I}_2$ dominates all other three terms in Eq. (15). Based on the fact $\bar{I}_2 < 0$, Eq. (14) holds. Consequently, the result in Theorem 4 holds.

To prove Theorem 5, we need the following lemma. □

Lemma 1 *Under the conditions in Theorem 5, with probability tending to* 1*, for any given* β_a *satisfying* $\|\beta_a - \beta_{0a}\| = \mathrm{O}_p(\sqrt{p_n/n})$ *and any constant C, we have*

$$Q_\lambda\left\{\begin{pmatrix}\beta_a\\ 0\end{pmatrix}\right\} = \max_{\|\beta_b\| \le C(p_n/n)^{1/2}} Q_\lambda\left\{\begin{pmatrix}\beta_a\\ \beta_b\end{pmatrix}\right\}, \tag{16}$$

Proof of Lemma 1 From the proof of Theorem 2, we have

$$R_j^{'}(\beta) = \frac{\partial R(\beta)}{\partial \beta_j} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \widehat{\widehat{X}}_i \phi_{h_1}^{\prime}(\varepsilon_i) - \Sigma_1(\beta - \beta_0) + o(\omega_n).$$

It can be shown that $\frac{1}{n}\sum_{i=1}^{n} \widehat{\widehat{X}}_i \phi_{h_1}^{\prime}(\varepsilon_i) = O_p(n^{-1/2})$. By the assumption that $\|\beta_a - \beta_{0a}\| = O_p(\sqrt{p_n/n})$, then we have $R_j^{\prime}(\beta) = O_p(\sqrt{p_n/n})$. Therefore, for $\beta_j \neq 0$ and $j = k_n + 1, \ldots, p_n$,

$$\begin{aligned}\frac{\partial Q_\lambda(\beta)}{\partial \beta_j} &= nR_j^{'}(\beta) - np_\lambda^{\prime}(|\beta_j|)\mathrm{sgn}(\beta_j)\\ &= -n\lambda\left\{\lambda^{-1}p_\lambda^{\prime}(|\beta_j|)\mathrm{sgn}(\beta_j) + O_p\left(\frac{\sqrt{p_n/n}}{\lambda}\right)\right\}.\end{aligned}$$

Since $\liminf_{n\to\infty}\liminf_{t\to 0^+} p_\lambda^{\prime}(t)/\lambda > 0$ and $(n/p_n)^{1/2}\lambda \to \infty$, then the sign of the derivative for $\beta_j \in (-C\sqrt{p_n/n}, C\sqrt{p_n/n})$ is completely determined by that of β_j. Therefore, Eq. (16) holds. □

Proof of Theorem 5 From Lemma 1, it follows that $\widehat{\beta}_b^\lambda = 0$. We will next show the asymptotic normality of $\widehat{\beta}_a^\lambda$. By Theorem 4, it can be shown easily that there exists a $\widehat{\beta}_a^\lambda$ that is a root-n/p_n consistent local maximizer of $Q_\lambda\{(\beta_a^T, 0)^T\}$, which satisfies the following equations:

$$\frac{\partial Q_\lambda(\beta)}{\partial \beta_j}\Big|_{\beta=((\widehat{\beta}_a^\lambda)^T, 0)^T} = 0, \quad \text{for} \quad j = 1, 2, \ldots, k_n.$$

Therefore,

$$nR_j^{\prime}(\widehat{\beta}^\lambda) - np_\lambda^{\prime}(|\widehat{\beta}^\lambda|)\mathrm{sgn}(\widehat{\beta}^\lambda)$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \widehat{\widehat{X}}_i \left\{ \phi'_{h_1}(\varepsilon_i) - \phi''_{h_1}(\varepsilon_i)(\sigma_i + \widehat{\gamma}_i) + \frac{1}{2}\phi'''_{h_1}\left(\varepsilon_i^*\right)(\sigma_i + \widehat{\gamma}_i)^2 \right\} \\
&\quad - n\left\{ p'_\lambda(\left|\beta_{0j}\right|)\mathrm{sgn}(\beta_{0j}) + \left(p''_\lambda(\left|\beta_{0j}\right|) + o_p(1)\right)(\widehat{\beta}^\lambda - \beta_j) \right\},
\end{aligned}
$$

where ε_i^* is between ε_i and $\varepsilon_i - \widehat{\gamma}_i$.

By the similar proof in Theorem 2, it follows by the central limit theorem and the Slutsky's theorem that

$$
\sqrt{n}(\Sigma_1^{(1)} + \Psi_\lambda)\{\widehat{\beta}_a^\lambda - \beta_{0a} + (\Sigma_1^{(1)} + \Psi_\lambda)^{-1}\mathbf{s_n}\} \to N(\mathbf{0}, \Sigma_2^{(1)})
$$

in distribution, where $\Sigma_1^{(1)}$, $\Sigma_2^{(1)}$ are the submatrices of Σ_1 and Σ_2 corresponding to β_{0a}. □

Proof of Theorem 6 For notational clarity, we let $K_i = K(\frac{Z_i - Z}{h_2})$ and $l(r) = -\phi_{h_3}(r)$. Then, Eq. (11) can be rewritten as

$$
(\widehat{a}^\lambda, \widehat{b}^\lambda) = \arg\min_{a,b} \sum_{i=1}^{n} l(Y_i - X_i^T \widehat{\beta}^\lambda - a - (Z_i - Z)b)K_i.
$$

Let $\theta = \left(nh_2^q\right)^{1/2}[a - f(Z), h_2(b - f'(Z))]$, $z_i^* = [1, (Z_i - Z)^T/h_2]^T$, $s_i = X_i^T(\beta_0 - \widehat{\beta}^\lambda)$, $\delta_i = Y_i - X_i^T\widehat{\beta}^\lambda - f(Z) - f'(Z)(Z_i - Z)$, $\delta_i^* = Y_i - X_i^T\beta_0 - f(Z) - f'(Z)(Z_i - Z)$ and $f_i = f(Z_i) - f(Z) - f'(Z)(Z_i - Z)$. Then, $\theta_n = \left(nh_2^q\right)^{1/2}[\widehat{a}^\lambda - f(Z), h_2(\widehat{b}^\lambda - f'(Z))]$ minimizes the function

$$
\begin{aligned}
J_n(\theta) &= \sum_{i=1}^{n} \left\{ l(Y_i - X_i^T\widehat{\beta}^\lambda - a - b(Z_i - Z)) - l(\delta_i) \right\} K_i \\
&= \sum_{i=1}^{n} \{ l(\delta_i - (nh_2^q)^{-1/2}(\theta^T z_i^*)) - l(\delta_i) \} K_i.
\end{aligned}
$$

Since the function $J_n(\theta)$ is convex in θ, it is sufficient to prove that $J_n(\theta)$ converges pointwise to its conditional expectation (Pollard 1991).

Given $\mathbf{X} = (X_1, \ldots, X_n)^T$ and $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$, we can obtain that

$$
\begin{aligned}
E\left(J_n(\theta)\,|\mathbf{Z}\right) &= -\,(nh_2^q)^{-1/2} \sum_{i=1}^{n} \varphi'_{h_3}(f_i + s_i|Z_i)(\theta^T z_i^*)K_i \\
&\quad + \frac{1}{2}(nh_2^q)^{-1} \sum_{i=1}^{n} \varphi''_{h_3}(f_i + s_i|Z_i)(\theta^T z_i^*)^2 K_i(1 + o_p(1)) \\
&= -\,(nh_2^q)^{-1/2} \sum_{i=1}^{n} \varphi'_{h_3}(f_i|Z_i)(\theta^T z_i^*)K_i
\end{aligned}
$$

$$+\frac{1}{2}(nh_2^q)^{-1}\sum_{i=1}^{n}\varphi''_{h_3}(0|Z_i)(\theta^T z_i^*)^2 K_i + o_p\{(nh_2^q)^{-1}\},$$

where the last equality is derived from the regularity condition (B3). Similar arguments can be also seen in (D.2) and (D.3) of Zhu et al. (2013).

Then, we can obtain that

$$\begin{aligned} J_n(\theta) &= (nh_2^q)^{-1/2}\sum_{i=1}^{n}\phi'_{h_3}(f_i+\varepsilon_i)(\theta^T z_i^*)K_i \\ &\quad + \frac{1}{2}(nh_2^q)^{-1}\sum_{i=1}^{n}\varphi''_{h_3}(0|Z_i)(\theta^T z_i^*)^2 K_i + o_p\{(nh_2^q)^{1/2}\} \\ &= (nh_2^q)^{-1/2}\sum_{i=1}^{n}\phi'_{h_3}(\delta_i^*)(\theta^T z_i^*)K_i \\ &\quad + \frac{1}{2}(nh_2^q)^{-1}\sum_{i=1}^{n}\varphi''_{h_3}(0|Z_i)(\theta^T z_i^*)^2 K_i + o_p\{(nh_2^q)^{1/2}\} \end{aligned}$$

which is parallel to (4.6) of Fan et al. (1994). The rest of the proof follows literally from Fan et al. (1994) by treating the dimension of Z as fixed, so the detail is omitted here. □

References

Akaike H (1973) Maximum likelihood Identification of Gaussian autoregressive moving average models. Biometrika 60:255–265

Breiman L (1995) Better subset regression using the nonnegative garrote. Technometrics 37:373–384

Chen B, Yu Y, Zou H, Liang H (2012) Profiled adaptive Elastic-Net procedure for partially linear models with high-dimensional covariates. J Stat Plann Inference 142:1733–1745

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39:1–21

Engle R, Granger C, Rice J, Weiss A (1986) Semiparametric estimates of the relation between weather and electricity sales. J Am Stat Assoc 81:310–320

Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman and Hall, London

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

Fan J, Lv J (2008) Sure independence screening for ultra-high-dimensional feature space. J R Stat Soc 70:849–911

Fan J, Hu TC, Truong YK (1994) Robust nonparametric function estimation. Scand J Stat 21:433–446

Hardle W, Liang H, Gao JT (2000) Partial linear models. Springer, New York

Huber PJ (1981) Robust estimation. Wiley, New York

Johnson RW (2003) Kiplingers personal finance. J Stat Educ 57:104–123

Li R, Liang H (2008) Variable selection in semiparametric regression modeling. Ann Stat 36:261–286

Li J, Ray S, Lindsay B (2007) A nonparametric statistical approach to clustering via mode identification. J Mach Learn Res 8:1687–1723

Li GR, Peng H, Zhu LX (2011) Nonconcave penalized M-estimation with diverging number of parameters. Stat Sin 21:391–420

Mallows CL (1973) Some comments on Cp. Technometrics 15:661–675

Ni X, Zhang HH, Zhang D (2009) Automatic model selection for partially linear models. J Multivar Anal 100:2100–2111

Pollard D (1991) Asymptotics for least absolute deviation regression estimators. Econom Theory 7:186–199

Rao BLSP (1983) Nonparametric functional estimation. Academic Press, Orlando

Robinson PM (1988) Root *n*-consistent semiparametric regression. Econometrica 56:931–954

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Severini TA, Staniswalis JG (1994) Quasi-likelihood estimation in semiparametric models. J Am Stat Assoc 89:501–511

Speckman PE (1988) Kernel smoothing in partial linear models. J R Stat Soc 50:413–436

Tibshirani R (1996) Regression shrinkage and selection via the LASSO. J R Stat Soc 58:267–288

Wang H, Li G, Jiang G (2007) Robust regression shrinkage and consistent variable selection through the LAD-Lasso. J Bus Econ Stat 25:347–355

Xie H, Huang J (2009) SCAD-penalized regression in high-dimensional partially linear models. Ann Stat 37:673–696

Yang H, Yang J (2014) A robust and efficient estimation and variable selection method for partially linear single-index models. J Multivar Anal 129:227–242

Yao W, Li L (2014) A new regression model: modal linear regression. Scand J Stat 41:656–671

Yao W, Lindsay B, Li R (2012) Local modal regression. J Nonparametr Stat 24:647–663

Zeger S, Diggle P (1994) Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. Biometrics 50:689–699

Zhang R, Zhao W, Liu J (2013) Robust estimation and variable selection for semiparametric partially linear varying coefficient model based on modal regression. J Nonparametr Stat 25:523–544

Zhao W, Zhang R, Liu J, Lv Y (2014) Robust and efficient variable selection for semiparametric partially linear varying coefficient model based on modal regression. Ann Inst Stat Math 66:165–191

Zhao W, Zhang R, Liu Y, Liu J (2015) Empirical likelihood based modal regression. Stat Papers 56:411–430

Zhou H, You J, Zhou B (2010) Statistical inference for fixed-effects partially linear regression models with errors in variables. Stat Pap 51:629–650

Zhu LX, Fang KT (1996) Asymptotics for kernel estimation of sliced inverse regression. Ann Stat 3:1053–1068

Zhu L, Huang M, Li R (2012) Semiparametric quantile regression with high-dimensional covariates. Stat Sin 22:1379–1401

Zhu L, Li R, Cui H (2013) Robust estimation for partially linear models with large-dimensional covariates. Sci China Math 56:2069–2088

Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418–1429