SCIENCE CHINA
Mathematics

• **ARTICLES** •

October 2013 Vol. 56 No. 10: 2069–2088
doi: 10.1007/s11425-013-4675-0

Robust estimation for partially linear models with large-dimensional covariates

ZHU LiPing[1,2], LI RunZe[3] & CUI HengJian[4,*]

[1] *School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai* 200433*, China;*
[2] *The Key Laboratory of Mathematical Economics* (*SUFE*)*, Ministry of Education, Shanghai* 200433*, China;*
[3] *Department of Statistics and The Methodology Center, The Pennsylvania State University,*
University Park, PA 16802*, USA;*
[4] *School of Mathematical Science, Capital Normal University, Beijing* 100037*, China*

Email: zhu.liping@mail.shufe.edu.cn, rli@stat.psu.edu, hjcui@bnu.edu.cn

Received May 25, 2012; accepted July 30, 2012; published online August 30, 2013

Abstract We are concerned with robust estimation procedures to estimate the parameters in partially linear models with large-dimensional covariates. To enhance the interpretability, we suggest implementing a nonconcave regularization method in the robust estimation procedure to select important covariates from the linear component. We establish the consistency for both the linear and the nonlinear components when the covariate dimension diverges at the rate of $o(\sqrt{n})$, where n is the sample size. We show that the robust estimate of linear component performs asymptotically as well as its oracle counterpart which assumes the baseline function and the unimportant covariates were known a priori. With a consistent estimator of the linear component, we estimate the nonparametric component by a robust local linear regression. It is proved that the robust estimate of nonlinear component performs asymptotically as well as if the linear component were known in advance. Comprehensive simulation studies are carried out and an application is presented to examine the finite-sample performance of the proposed procedures.

Keywords partially linear models, robust model selection, smoothly clipped absolute deviation (SCAD), semiparametric models

MSC(2010) 62F35, 62G35

Citation: Zhu L P, Li R Z, Cui H J. Robust estimation for partially linear models with large-dimensional covariates. Sci China Math, 2013, 56: 2069–2088, doi: 10.1007/s11425-013-4675-0

1 Introduction

Let Y be the response variable, $\boldsymbol{x} = (X_1, \ldots, X_{p_n})^\top \in \mathbb{R}^{p_n}$ and $\boldsymbol{z} = (Z_1, \ldots, Z_d)^\top \in \mathbb{R}^d$ be the associated covariate vectors. Consider the partially linear model

$$Y = \boldsymbol{\beta}_{n0}^\top \boldsymbol{x} + \nu(\boldsymbol{z}) + \varepsilon, \tag{1.1}$$

where the random error ε satisfies $E(\varepsilon|\boldsymbol{x}, \boldsymbol{z}) = 0$. The subscript n in $\boldsymbol{\beta}_{n0}$ indicates that the dimension p_n of $\boldsymbol{\beta}_{n0}$ possibly depends on n. Model (1.1) contains both a linear component $\boldsymbol{\beta}_{n0}^\top \boldsymbol{x}$ and a nonparametric baseline function $\nu(\boldsymbol{z})$. It combines the flexibility of nonparametric regression and the parsimony and interpretability of linear regression.

[*] Corresponding author

As one of the most commonly used semiparametric regression models, the partially linear model has received considerable attention in the last two decades. The existing literatures often assume a univariate $\boldsymbol{z}$ and a finite-dimensional $\boldsymbol{x}$ (i.e., $d = 1$ and p_n is fixed as $n \to \infty$). Examples for estimating $\boldsymbol{\beta}_{n0}$ include the partial spline estimator [6, 15], and the partial residual estimator [5, 25, 26]. [26] and [12] proposed respectively kernel regression and local linear regression to estimate the parameters and systematically investigated the theoretical properties of the resulting estimators. For a comprehensive review on the study of (1.1), one can refer to the monograph by [13] and references therein. In these studies, however, either the dimension of the covariate vector $\boldsymbol{x}$ was fixed or the problem of variable selection in $\boldsymbol{x}$ via penalization was not considered. In addition, these studies did not consider the issue of robust estimation.

In the presence of a large number of covariates in model (1.1), [10] considered variable selection in the context of longitudinal data analysis, and [21] proposed penalized profile least squares to estimate the parameters in partially linear models with measurement errors. Both assume a framework with a fixed set of covariates as n increases. [32] established the consistency of penalized profile least squares estimate when p_n diverges at the rate of $o(\sqrt{n})$.

To handle non-normal or heavy-tailed errors in (1.1), [3] proposed a robust profile likelihood estimation, [14] and [23] extended the robust likelihood procedure to longitudinal data. [21] considered penalized quantile regression for partially linear models when the covariates are measured with additive errors. These robust estimation procedures are applicable for low-dimensional covariates, but they become infeasible when the dimension of covariates is comparably large due to the curse of dimensionality.

In this paper, we propose several general robust procedures to estimate the parameters in model (1.1). We allow the dimension p_n of $\boldsymbol{x}$ to diverge with the sample size n. This strategy makes model (1.1) more useful in usual practice [11, 32]. To estimate $\boldsymbol{\beta}_{n0}$ in (1.1), we follow the idea of "partial-residual" estimation [5, 25, 26] and transform (1.1) into a linear model. We build a general robust estimation procedure upon the transformed linear model. To enhance the interpretability, we then implement a nonconcave regularization approach to select the important covariates from $\boldsymbol{x}$. It is shown that the penalized robust estimate of $\boldsymbol{\beta}_{n0}$ performs asymptotically as well as an oracle procedure when $p_n = o(\sqrt{n})$. This oracle property also extends the terminology of [9] in that the oracle procedure for (1.1) assumes both the effects of $\boldsymbol{z}$ on both $\boldsymbol{x}$ and Y and the unimportant covariates amongst $\boldsymbol{x}$ were known in advance.

With a consistent estimate of $\boldsymbol{\beta}_{n0}$, we estimate the nonlinear component $\nu(\boldsymbol{z})$ in (1.1) by a robust local linear estimation. We show that the robust local linear regression estimate of $\nu(\boldsymbol{z})$ based upon the consistent estimate of $\boldsymbol{\beta}_{n0}$ has the same asymptotic bias and variance as the robust local linear regression based upon the true value $\boldsymbol{\beta}_{n0}$. In other words, the robust local linear estimation of the nonlinear component also has an oracle property asymptotically.

The rest of this paper is organized as follows. In Section 2, we propose a general robust estimation procedure to estimate the linear and the nonlinear components for (1.1). To select important covariates from the linear component, in Section 3 we implement a nonconcave penalty in the robust estimation procedure. Some practical issues, including the optimization and tuning parameter selection, are also discussed in this section. Comprehensive simulations are conducted in Section 4 to examine the performance of the proposed procedures. The simulations demonstrate that the proposed procedures with moderate sample size perform almost as well as the oracle estimators. In this section we also illustrate our proposal through an empirical analysis of a real-world dataset. This article is concluded with a brief discussion in Section 5. Regularity conditions and technical proofs are given in Appendix.

2 A robust estimation procedure

In this section, we suggest a general robust estimation procedure to estimate the linear component and the nonparametric baseline function.

2.1 A robust estimate of $\boldsymbol{\beta}_{n0}$

We first discuss how to estimate $\boldsymbol{\beta}_{n0}$ in (1.1) following the idea of "partial-residual" estimation [5, 25, 26]. For notational clarity, we write $\widetilde{\boldsymbol{x}} = \boldsymbol{x} - E(\boldsymbol{x}|\boldsymbol{z})$ and $\widetilde{Y} = Y - E(Y|\boldsymbol{z})$. We note that model (1.1) implies that

$$\widetilde{Y} = \boldsymbol{\beta}_{n0}^{\top}\widetilde{\boldsymbol{x}} + \varepsilon. \tag{2.1}$$

The transformed model (2.1) enables us to estimate $\boldsymbol{\beta}_{n0}$ in the context of classical linear model. Specifically, we suppose that $\{(\boldsymbol{x}_i, \boldsymbol{z}_i, Y_i), i = 1, \ldots, n\}$ is a random sample from (1.1). To get some insights into the general estimation procedure, for now we assume that $E(\boldsymbol{x}_i|\boldsymbol{z}_i)$ and $E(Y_i|\boldsymbol{z}_i)$, for $i = 1, \ldots, n$, are observable, which in turn implies that both $\widetilde{\boldsymbol{x}}_i$'s and $\widetilde{Y}_i$'s are observable. We will discuss how to estimate $E(\boldsymbol{x}_i|\boldsymbol{z}_i)$ and $E(Y_i|\boldsymbol{z}_i)$ later. With $\{(\widetilde{\boldsymbol{x}}_i, \widetilde{Y}_i), i = 1, \ldots, n\}$, we can estimate $\boldsymbol{\beta}_{n0}$ by minimizing

$$\sum_{i=1}^{n} \rho(\widetilde{Y}_i - \widetilde{\boldsymbol{x}}_i^{\top}\boldsymbol{\beta}_n) \tag{2.2}$$

for a suitable choice of loss function $\rho(\cdot)$. In general, $\rho(\cdot)$ can be any convex function. Some important examples include Huber's estimate with $\rho(r) = \frac{r^2}{2}\mathbf{1}(|r| \leqslant c) + (c|r| - \frac{c^2}{2})\mathbf{1}(|r| > c)$ for some positive constant c; the ℓ_q-regression estimate with $\rho(r) = |r|^q$ for some $1 \leqslant q \leqslant 2$; the regression quantile with $\rho(r) = \alpha r^+ + (1-\alpha)(-r)^+$, for $0 < \alpha < 1$, where $r^+ = \max(r, 0)$. When $q = 1$ or $\alpha = 1/2$, the minimizer of (2.2) is called the least absolute deviation (LAD) estimate, which is a robust estimation; when $c = +\infty$ or $q = 2$, it corresponds to the ordinary least squares estimate.

In practice, we must estimate $E(Y|\boldsymbol{z})$ and $E(\boldsymbol{x}|\boldsymbol{z})$. This can be done through nonparametric regression techniques such as kernel smoothing or local linear regression [7]. Denote the resulting estimates by $\widehat{m}_{\boldsymbol{x}}(\boldsymbol{z}_i)$ and $\widehat{m}_Y(\boldsymbol{z}_i)$ respectively. For example, we define

$$\widehat{m}_{\boldsymbol{x}}(\boldsymbol{z}_i) := \frac{\sum_{j\neq i}^{n} K_{h_1}(\boldsymbol{z}_j - \boldsymbol{z}_i)\boldsymbol{x}_j}{\sum_{j\neq i}^{n} K_{h_1}(\boldsymbol{z}_j - \boldsymbol{z}_i)},$$

and

$$\widehat{m}_Y(\boldsymbol{z}_i) := \frac{\sum_{j\neq i}^{n} K_{h_2}(\boldsymbol{z}_j - \boldsymbol{z}_i)Y_j}{\sum_{j\neq i}^{n} K_{h_2}(\boldsymbol{z}_j - \boldsymbol{z}_i)}, \tag{2.3}$$

where $K_{h_k}(\cdot) = K(\cdot/h_k)/h_k$ is a d-dimensional kernel function and h_k is the bandwidth for $k = 1, 2$. Note that $\widehat{m}_Y(\cdot)$ is not a robust estimate of $m_Y(\cdot)$ when ε is heavy-tailed. However, it will not affect the final estimate of $\boldsymbol{\beta}_{n0}$ because the issue of the presence of heavy-tailed errors will be taken into account when we implement a robust procedure to estimate $\boldsymbol{\beta}_{n0}$. Let $\widehat{\widetilde{\boldsymbol{x}}}_i = \boldsymbol{x}_i - \widehat{m}_{\boldsymbol{x}}(\boldsymbol{z}_i)$ and $\widehat{\widetilde{Y}}_i = Y_i - \widehat{m}_Y(\boldsymbol{z}_i)$. Denote

$$\widehat{\boldsymbol{\beta}}_{n0} := \underset{\boldsymbol{\beta}_n}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(\widehat{\widetilde{Y}}_i - \widehat{\widetilde{\boldsymbol{x}}}_i^{\top}\boldsymbol{\beta}_n), \tag{2.4}$$

which is the final robust estimate of $\boldsymbol{\beta}_{n0}$ at the sample level.

The following theorem states the convergence rate of the robust estimate.

Theorem 1. *Suppose that* $\rho(\cdot)$ *and* $\widetilde{\boldsymbol{x}}$ *satisfy Conditions* (C1)–(C2) *and* (C3)(i) *in Appendix A. If* $p_n^2/n \to 0$, *then there exists a robust estimator* $\widehat{\boldsymbol{\beta}}_{n0}$ *such that*

$$\|\widehat{\boldsymbol{\beta}}_{n0} - \boldsymbol{\beta}_{n0}\| = O_P(\sqrt{p_n/n}),$$

where $\widehat{\boldsymbol{\beta}}_{n0}$ *is defined in* (2.4), *and* $\|\cdot\|$ *stands for the Euclidean norm.*

The following theorem presents the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{n0}$ when p_n diverges at the rate of $o(n^{1/2})$. It improves the rate of $p_n = o(n^{1/3})$ obtained by [17].

Theorem 2. *Suppose Conditions* (C1)–(C2) *and* (C3)(i) *in Appendix A hold. If* $p_n^2/n \to 0$, *then*

$$n^{1/2}\boldsymbol{A}_n\boldsymbol{S}_n^{1/2}(\widehat{\boldsymbol{\beta}}_{n0} - \boldsymbol{\beta}_{n0}) \overset{\mathcal{D}}{\to} N(0, \gamma^{-2}\sigma_0^2\boldsymbol{G})$$

where " $\overset{\mathcal{D}}{\to}$ *" stands for "convergence in distribution";* $\boldsymbol{S}_n = \mathrm{var}(\widetilde{\boldsymbol{x}})$ *and* $\widetilde{\boldsymbol{x}} = \boldsymbol{x} - E(\boldsymbol{x}|\boldsymbol{z})$; $\boldsymbol{A}_n$ *is a* $q \times p_n$ *matrix such that* $\boldsymbol{A}_n\boldsymbol{A}_n^{\top} \to \boldsymbol{G}$ *as* $n \to \infty$, *and* $\boldsymbol{G}$ *is a* $q \times q$ *nonnegative symmetric matrix;* $\sigma_0^2 = E\{\psi^2(\varepsilon)\}$ *where* $\psi(\cdot)$ *is defined in Condition* (C1); γ *is a constant defined in Condition* (C2).

This theorem implies that the general robust estimate of $\boldsymbol{\beta}_{n0}$, denoted by $\widehat{\boldsymbol{\beta}}_{n0}$, performs asymptotically as well as if $\{(\widetilde{\boldsymbol{x}}_i, \widetilde{Y}_i), i = 1, \ldots, n\}$ were observed a priori. In other words, even when the dimensionality p_n diverges at the rate of $p_n = o(n^{1/2})$, the robust estimate $\widehat{\boldsymbol{\beta}}_{n0}$ performs asymptotically as well as its oracle counterpart which assumes the nonlinear effect of $\boldsymbol{z}$ on both $\boldsymbol{x}$ and Y were known in advance.

If ρ happens to be the least squares loss function, then $\gamma^{-2}\sigma_0^2 = E(\varepsilon^2)$. Thus, for a general loss function ρ, the asymptotic relative efficiency (ARE) of the resulting robust estimator to the ordinary least squares has the form of

$$\mathrm{ARE} = \frac{\gamma^2 E(\varepsilon^2)}{E\{\psi^2(\varepsilon)\}}.$$

2.2 A robust estimate of $\nu(\cdot)$

Next, we discuss how to estimate $\nu(\boldsymbol{z}_0)$ for any fixed $\boldsymbol{z}_0 \in \mathbb{R}^d$ when a consistent estimate of $\boldsymbol{\beta}_{n0}$ is available. This can be done through using the idea of the local linear fit which approximates the unknown function $\nu(\boldsymbol{z}_0)$ by a linear function $\nu(\boldsymbol{z}) \approx \nu(\boldsymbol{z}_0) + (\boldsymbol{z} - \boldsymbol{z}_0)^\top \nu'(\boldsymbol{z}_0) =: \mathrm{a} + (\boldsymbol{z} - \boldsymbol{z}_0)^\top \boldsymbol{b}$ for $\boldsymbol{z}$ in a neighborhood of $\boldsymbol{z}_0$. Locally, estimating $\nu(\boldsymbol{z}_0)$ is equivalent to estimating the intercept term a. Therefore, we are motivated to define an estimator $\widehat{\nu}(\boldsymbol{z}_0) = \widehat{\mathrm{a}}$, where

$$(\widehat{\mathrm{a}}, \widehat{\boldsymbol{b}}) =: \operatorname*{argmin}_{\mathrm{a},\boldsymbol{b}} \sum_{i=1}^{n} \rho\{Y_i - \widehat{\boldsymbol{\beta}}_{n0}^\top \boldsymbol{x}_i - \mathrm{a} - (\boldsymbol{z} - \boldsymbol{z}_0)^\top \boldsymbol{b}\} K\left(\frac{\boldsymbol{z}_i - \boldsymbol{z}_0}{h_n}\right), \tag{2.5}$$

where $K(\cdot)$ is a d-dimensional kernel function. One can also refer to [8] for a motivation and a discussion of this estimate.

The following theorem presents the asymptotic normality of $\widehat{\nu}(\boldsymbol{z}_0)$.

Theorem 3. *In addition to conditions in Theorem* 2*, we assume that Conditions* (C4)–(C7) *hold. If* $h_n \to 0$ *and* $nh_n^d \to \infty$*, then*

$$(nh_n^d)^{1/2}\left\{\widehat{\nu}(\boldsymbol{z}_0) - \nu(\boldsymbol{z}_0) - bias\right\} \xrightarrow{\mathcal{D}} N\left\{0, \frac{\int K^2(\boldsymbol{v})d\boldsymbol{v}}{f(\boldsymbol{z}_0)} \int G^2\left(y^* - \nu\left(\boldsymbol{z}_0\right)\right) g(y^*|\boldsymbol{z}_0) d\mu(y^*)\right\},$$

where $bias = h_n^2 \mathrm{tr}\left\{\nu''(\boldsymbol{z}_0)\right\} \int v^2 K(v)dv/2$; $f(\cdot)$ *denotes the density function of* $\boldsymbol{z}$; *and* $G(\cdot)$ *and* $g(\cdot)$ *are defined in Condition* (C6)*.*

The above theorem indicates that the robust local linear regression estimate of $\nu(\cdot)$ built upon the consistent estimate $\widehat{\boldsymbol{\beta}}_{n0}$ has the same asymptotic bias and variance as that built upon the true value $\boldsymbol{\beta}_{n0}$. This is not a very surprising result in that Theorem 2 states that the estimate of the linear component has a faster convergence rate than that of the nonlinear component.

3 A penalized robust estimation

In practice many covariates are often collected to attenuate modeling bias during the stage of data gathering. To enhance the interpretability of (1.1) when a large number of covariates are present, we consider penalized robust estimation procedures to select important covariates and to exclude unimportant covariates in the sequel.

3.1 Variable selection in robust estimation

To select important covariates from $\boldsymbol{x}$ in (1.1), we consider equivalently the transformed linear model (2.1). Over the past years, much effort has been devoted to establishing the consistency of various penalized robust estimations in the context of classical linear models [20, 28, 30, 34]. In the transformed linear model (2.1), however, both $\widetilde{\boldsymbol{x}}$ and $\widetilde{Y}$ are not observed. Thus we must replace them with their consistent estimators. This introduces some new difficulties in selecting important covariates from $\boldsymbol{x}$.

We follow the idea of (2.4) and propose to estimate $\boldsymbol{\beta}_{n0}$ by minimizing

$$\sum_{i=1}^{n}\rho(\widehat{\widetilde{Y}}_i-\widehat{\widetilde{\boldsymbol{x}}}_i^{\top}\boldsymbol{\beta}_n)+n\sum_{j=1}^{p_n}p_{\lambda}(|\beta_{nj}|), \tag{3.1}$$

where β_{nj} is the j-th coordinate of $\boldsymbol{\beta}_n$, $p_\lambda(\cdot)$ is a penalty function and λ is a regularization parameter. [9] studied the choice of penalty functions in depth. They proposed a unified approach via non-concave penalized likelihood to automatically select important variables and simultaneously estimate the coefficients of covariates. One of the most commonly used nonconcave penalty is the SCAD penalty, which is defined by

$$p_\lambda(|\beta_{nj}|)=\begin{cases}\lambda|\beta_{nj}|, & \text{if } 0\leqslant|\beta_{nj}|<\lambda,\\ \dfrac{(a^2-1)\lambda^2-(|\beta_{nj}|-a\lambda)^2}{2(a-1)}, & \text{if } \lambda\leqslant|\beta_{nj}|<a\lambda,\\ \dfrac{(a+1)^2\lambda^2}{2}, & \text{if } |\beta_{nj}|\geqslant a\lambda,\end{cases}$$

where a is often chosen as 3.7 as suggested in [9].

In practical implementation, we may employ the perturbed LQA algorithm proposed by [18] to minimize (3.1) for a fixed λ. Specifically, we updated $\boldsymbol{\beta}_n^{(k+1)}$ from $\boldsymbol{\beta}_n^{(k)}$ by using

$$\boldsymbol{\beta}_n^{(k+1)}=\underset{\boldsymbol{\beta}_n}{\operatorname{argmin}}\bigg\{\sum_{i=1}^{n}\rho(\widehat{\widetilde{Y}}_i-\widehat{\widetilde{\boldsymbol{x}}}_i^{\top}\boldsymbol{\beta}_n)+n\sum_{j=1}^{p_n}\frac{p_\lambda'(|\beta_{nj}^{(k)}|)}{2(|\beta_{nj}^{(k)}|+\tau)}\beta_{nj}^2\bigg\}, \tag{3.2}$$

where $\tau=\tau_0\frac{\min\{|\beta_{nj}^{(k)}|:\beta_{nj}^{(k)}\neq0\}}{2n\lambda_n}$ for a prespecified tolerance τ_0. In our subsequent implementations, we update $\boldsymbol{\beta}_n^{(k)}$ using (3.2) to obtain an estimate of $\boldsymbol{\beta}_{n0}$. This algorithm was implemented in our numerical study.

To produce a sparse robust estimate of $\boldsymbol{\beta}_{n0}$, it remains to select the regularization parameter λ. For any fixed λ, we update $\boldsymbol{\beta}_n^{(k+1)}$ from $\boldsymbol{\beta}_n^{(k)}$ iteratively using (3.2) until the algorithm converges. We denote the resulting estimate by $\widehat{\boldsymbol{\beta}}_n(\lambda)$. The regularization parameter λ can be chosen through a data-driven algorithm. We adopt the BIC-type tuning parameter selector. To be specific, we choose the optimal λ which minimizes the following BIC type criterion:

$$\mathrm{BIC}(\lambda)=n\log\bigg[\sum_{i=1}^{n}\rho\{\widehat{\widetilde{Y}}_i-\widehat{\widetilde{\boldsymbol{x}}}_i^{\top}\widehat{\boldsymbol{\beta}}_n(\lambda)\}\bigg]+\log(n)\mathrm{DF}_\lambda,$$

where

$$\mathrm{DF}_\lambda=\mathrm{trace}[\widehat{\widetilde{\boldsymbol{X}}}(\widehat{\widetilde{\boldsymbol{X}}}^{\top}\widehat{\widetilde{\boldsymbol{X}}}+n\boldsymbol{D}\{\widehat{\boldsymbol{\beta}}_n(\lambda),\lambda\})^{-1}\widehat{\widetilde{\boldsymbol{X}}}^{\top}],$$

$\widehat{\widetilde{\boldsymbol{X}}}=(\widehat{\widetilde{\boldsymbol{x}}}_1,\ldots,\widehat{\widetilde{\boldsymbol{x}}}_n)^{\top}$ is an $n\times p_n$ matrix, and $\boldsymbol{D}\{\widehat{\boldsymbol{\beta}}_n(\lambda),\lambda\}$ is the diagonal matrix whose j-th diagonal element is

$$\frac{p_\lambda'(|\widehat{\beta}_{nj}(\lambda)|)}{2|\widehat{\beta}_{nj}(\lambda)|}.$$

We denote by $\widehat{\boldsymbol{\beta}}_n$ the final estimation $\widehat{\boldsymbol{\beta}}_n(\lambda)$ with an optimal λ selected by the above BIC type criterion.

Regarding p_n as a fixed number, [29] proved the selection consistency of the BIC type criterion under the least squares framework. [27] proposed a modified BIC type criterion to accommodate the diverging p_n scenario. They also established the selection consistency for the modified BIC type criterion. We further adapted [27]'s BIC type criterion to accommodate the robust regression scenario. Though the selection consistency of the modified BIC type criterion under the robust regression context remains unknown to us, it is an important issue and deserves our further study.

3.2 Asymptotic properties

In this section, we investigate the asymptotic properties of the nonconcave-penalized robust estimation $\widehat{\boldsymbol{\beta}}_n$. The following theorem states the convergence rate of $\widehat{\boldsymbol{\beta}}_n$.

Theorem 4. *Suppose that $\rho(\cdot)$ and $\widetilde{\boldsymbol{x}}$ satisfy Conditions* (C1)–(C2) *and* (C3)(ii) *and the penalty function $p_{\lambda_n}(\cdot)$ satisfies Conditions* (C8)–(C10). *If $p_n^2/n \to 0$ as $n \to \infty$, then there exists a nonconcave-penalized robust estimator $\widehat{\boldsymbol{\beta}}_n$ such that*

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| = O_P\{p_n^{1/2}(n^{-1/2} + a_n)\},$$

where a_n is defined in Condition (C8) *in Appendix A.*

Theorem 4 shows that the nonconcave-penalized robust estimator $\widehat{\boldsymbol{\beta}}_n$ is root-n/p_n consistent if $a_n = O(n^{-1/2})$. For the SCAD penalty, if those nonzero coefficients are larger than $a\lambda_n$, it can be easily seen that $a_n = 0$ when n is large enough, and hence the estimate of nonconcave-penalized robust estimator $\widehat{\boldsymbol{\beta}}_n$ is root-n/p_n consistent.

Next, we investigate the oracle property of the nonconcave-penalized robust estimate $\widehat{\boldsymbol{\beta}}_n$. To facilitate illustration, we assume without loss of generality that the first k_n elements in $\boldsymbol{\beta}_{n0}$ are nonzero and the last $p_n - k_n$ elements are zero. In other words, $\boldsymbol{\beta}_{n0} = (\boldsymbol{\beta}_{n\boldsymbol{I}}^\top, \boldsymbol{\beta}_{n\boldsymbol{II}}^\top)^\top$, where $\boldsymbol{\beta}_{n\boldsymbol{I}} = (\beta_{n0,1}, \ldots, \beta_{n0,k_n})^\top$ and $\boldsymbol{\beta}_{n\boldsymbol{II}} = (\beta_{n0,k_n+1}, \ldots, \beta_{n0,p_n})^\top$, $\beta_{n0,j} \neq 0$ for $1 \leqslant j \leqslant k_n$ and $\beta_{n0,j} = 0$ for $k_n + 1 \leqslant j \leqslant p_n$. Denote

$$\boldsymbol{b}_n = \{p'_{\lambda_n}(|\beta_{n0,1}|)\mathrm{sign}(\beta_{n0,1}), \ldots, p'_{\lambda_n}(|\beta_{n0,k_n}|)\mathrm{sign}(\beta_{n0,k_n})\}^\top,$$

and

$$\boldsymbol{\Sigma}_{\lambda_n} = \mathrm{diag}\{p''_{\lambda_n}(|\beta_{n0,1}|), \ldots, p''_{\lambda_n}(|\beta_{n0,k_n}|)\}.$$

Theorem 5. *Under Conditions* (C1)–(C2) *and* (C3)(ii) *and* (C8)–(C11), *if $\lambda_n \to 0$ with a proper rate, then with probability tending to* 1, *the root-n/p_n consistent nonconcave-penalized robust estimator $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}}^\top, \widehat{\boldsymbol{\beta}}_{n\boldsymbol{II}}^\top)^\top$ must satisfy*:

(i) *Sparsity*: $\widehat{\boldsymbol{\beta}}_{n\boldsymbol{II}} = \boldsymbol{0}$, *if $(n/p_n)^{1/2}\lambda_n \to \infty$ as $n \to \infty$*;

(ii) *Asymptotic normality*: *If $p_n^2/n \to 0$ and $(n/p_n)^{1/2}\lambda_n \to \infty$, then*

$$n^{1/2}\boldsymbol{A}_{n\boldsymbol{I}}\boldsymbol{S}_{n\boldsymbol{I}}^{-1/2}(\gamma\boldsymbol{S}_{n\boldsymbol{I}} + \boldsymbol{\Sigma}_{\lambda_n})\{(\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}} - \boldsymbol{\beta}_{n\boldsymbol{I}}) + (\gamma\boldsymbol{S}_{n\boldsymbol{I}} + \boldsymbol{\Sigma}_{\lambda_n})^{-1}\boldsymbol{b}_n\} \xrightarrow{\mathcal{D}} N(0, \sigma_0^2\boldsymbol{G}),$$

where $\boldsymbol{S}_{n\boldsymbol{I}} = \mathrm{var}(\widetilde{\boldsymbol{x}}_{\boldsymbol{I}})$ and $\boldsymbol{A}_{n\boldsymbol{I}}$ is a $q \times k_n$ matrix such that $\boldsymbol{A}_{n\boldsymbol{I}}\boldsymbol{A}_{n\boldsymbol{I}}^\top \to \boldsymbol{G}$ as $n \to \infty$, and $\boldsymbol{G}$ is a $q \times q$ nonnegative symmetric matrix.

Theorem 5 implies that the nonconcave-penalized robust estimator of the zero coefficients are exactly zero with high probability when n is large. The asymptotic normality in Theorem 5 in parallel to that in Theorem 2, both allow the dimensionality p_n diverges at the rate of $p_n = o(n^{1/2})$. When n is large enough and those nonzero valued coefficients are larger than $a\lambda_n$, $\boldsymbol{\Sigma}_{\lambda_n} = \boldsymbol{0}$ and $\boldsymbol{b} = \boldsymbol{0}$ for the SCAD penalty. Consequently, the asymptotic normality (ii) of Theorem 5 becomes

$$n^{1/2}\boldsymbol{A}_{n\boldsymbol{I}}\boldsymbol{S}_{n\boldsymbol{I}}^{1/2}(\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}} - \boldsymbol{\beta}_{n\boldsymbol{I}}) \xrightarrow{\mathcal{D}} N(0, \gamma^{-2}\sigma_0^2\boldsymbol{G}),$$

which has the same efficiency of the robust estimator of $\boldsymbol{\beta}_{n\boldsymbol{I}}$ based on the sub-model with $\boldsymbol{\beta}_{n\boldsymbol{II}}$ known in advance. In addition, our estimator achieves the same efficacy as if $\widetilde{\boldsymbol{x}}$ and $\widetilde{Y}$ were observed in advance, although $\widetilde{\boldsymbol{x}}$ and $\widetilde{Y}$ have to be estimated from the data because $E(\boldsymbol{x}|\boldsymbol{z})$ and $E(Y|\boldsymbol{z})$ usually remain unknown in practice. This demonstrates that, the penalized robust estimate is as efficient as the oracle estimator which assumes $\boldsymbol{\beta}_{n\boldsymbol{II}}$ and the effects of $\boldsymbol{z}$ on $\boldsymbol{x}$ and Y were known in advance.

With the penalized robust estimate $\widehat{\boldsymbol{\beta}}_n$, we follow the idea of Subsection 2.2 and estimate $\nu(\boldsymbol{z}_0)$ using the local linear approximation. To be specific, we define

$$(\widehat{\mathrm{a}}, \widehat{\boldsymbol{b}}) := \mathop{\mathrm{argmin}}_{\mathrm{a},\boldsymbol{b}} \sum_{i=1}^n \rho\{Y_i - \widehat{\boldsymbol{\beta}}_n^\top\boldsymbol{x}_i - \mathrm{a} - (\boldsymbol{z} - \boldsymbol{z}_0)^\top\boldsymbol{b}\}K\left(\frac{\boldsymbol{z}_i - \boldsymbol{z}_0}{h_n}\right).$$

We define without notational confusion that $\widehat{\nu}(\boldsymbol{z}_0) = \widehat{\mathrm{a}}$. The following theorem is parallel to Theorem 3, yet it allows the dimension of $\boldsymbol{x}$ diverges at the rate of $p_n = o(n^{1/2})$, thanks to the penalized robust estimation.

Theorem 6. *In addition to conditions in Theorem* 5*, we assume that Conditions* (C4)–(C7) *hold. If* $h_n \to 0$ *and* $nh_n^d \to \infty$*, then*

$$(nh_n^d)^{1/2}\left\{\widehat{\nu}(\boldsymbol{z}_0) - \nu(\boldsymbol{z}_0) - bias\right\} \xrightarrow{\mathcal{D}} N\left\{0, \frac{\int K^2(\boldsymbol{v})d\boldsymbol{v}}{f(\boldsymbol{z}_0)}\int G^2\left\{y^* - \nu(\boldsymbol{z}_0)\right\} g(y^*|\boldsymbol{z}_0)d\mu(y^*)\right\},$$

where $bias = h_n^2 \mathrm{tr}\left\{\nu''(\boldsymbol{z}_0)\right\}\int v^2 K(v)dv/2$.

4 Numerical studies

4.1 Simulations

We conduct simulations to investigate the performance of the newly proposed procedures. The following four models are adopted for comparison purposes:

$$\begin{aligned}
&\text{model (I)} : Y = \boldsymbol{\beta}_{n0}^\top \boldsymbol{x} + 2\sin(\boldsymbol{\gamma}^\top \boldsymbol{z}) + \varepsilon;\\
&\text{model (II)} : Y = \boldsymbol{\beta}_{n0}^\top \boldsymbol{x} + |(\boldsymbol{\gamma}^\top \boldsymbol{z}) + 1| + \varepsilon;\\
&\text{model (III)} : Y = \boldsymbol{\beta}_{n0}^\top \boldsymbol{x} + \exp\{(\boldsymbol{\gamma}^\top \boldsymbol{z})/2\}/2 + \varepsilon;\\
&\text{model (IV)} : Y = \boldsymbol{\beta}_{n0}^\top \boldsymbol{x} + 2(\boldsymbol{\gamma}^\top \boldsymbol{z}) + \varepsilon.
\end{aligned}$$

We choose these models based on the following considerations. The effects of $\boldsymbol{z}$ on the response Y are nonlinear in (I)–(III) and linear in (IV). The link function $\nu(\boldsymbol{z})$ is oscillating in (I)–(II), and monotonic in (III)–(IV). In these models, we generate $\boldsymbol{z} = (Z_1, Z_2)^\top$ from a multivariate normal distribution with mean zero and identity variance-covariance matrix. We generate $\boldsymbol{x} = (X_1, \ldots, X_{p_n})^\top$ from models of the form $X_i = \boldsymbol{\gamma}^\top \boldsymbol{z} + 2\epsilon_i$ for $i = 1, \ldots, p_n$, where $\boldsymbol{\gamma} = (0.707, 0.707)^\top$ and the error terms ϵ_i's are independently generated from standard normal population. We generate $\boldsymbol{x}$ in this way such that $\boldsymbol{x}$ and $\boldsymbol{z}$ are correlated. We choose $\boldsymbol{\beta}_{n0} = (1.0, 0.8, 1.0, -1.5, 0.5, 0, \ldots, 0)^\top$, indicating that only the first five covariates are important, and all remaining covariates of $\boldsymbol{x}$ are unimportant given $(X_1, \ldots, X_5)^\top$.

To verify the robustness of our proposals, we consider three scenarios for the error term ε: (a) ε is generated from standard normal distribution; (b) ε is generated from standard t-distribution with 2 degrees of freedom; and (c) ε is generated from the mixture normal distribution $0.8N(0,1) + 0.2N(0, 5^2)$.

To provide a consistent estimate of $\boldsymbol{\beta}_{n0}$ for model (1.1), we adopt three loss functions: (1) $\rho(r) = r^2$ (least squares estimation, LSE); (2) $\rho(r) = |r|$ (least absolute deviation, LAD); and (3) $\rho(r) = 0.5r^2$ if $|r| \leqslant 1.345$ and $\rho(r) = 1.345|r| - 1.345^2/2$ if $|r| > 1.345$ (Huber function, HUB).

The simulations are repeated 200 times each of sample size $n = 400$ and dimension $p_n = 2n^{1/2} = 40$.

Estimation of $\boldsymbol{\beta}_{n0}$

We evaluate the performance of $\widehat{\boldsymbol{\beta}}$ using the squared errors. To measure the estimation accuracy of $\widehat{\boldsymbol{\beta}}$, we adopt the squared error (SE) which is defined by

$$\mathrm{SE}(\widehat{\boldsymbol{\beta}}) = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{n0}\|^2. \tag{4.1}$$

Recall that, to estimate $\boldsymbol{\beta}_{n0}$, we introduced the general robust estimate $\widehat{\boldsymbol{\beta}}_{n0}$ in Section 2 and the non-concave penalized robust estimate $\widehat{\boldsymbol{\beta}}_n$ in Section 3.

We estimate the squared errors by Monte Carlo simulations. The averages and the standard deviations of SE values of $\widehat{\boldsymbol{\beta}}_{n0}$ are summarized in Table 1, from which it can be easily seen that the general robust estimate $\widehat{\boldsymbol{\beta}}_{n0}$ performs equally as well as its oracle version which assumes $\nu(\cdot)$ is known in advance. This confirms our theoretical investigation in Theorem 5. Though we estimate $m_Y(\boldsymbol{z}_i)$ through the usual kernel regression (2.3) when $\varepsilon \sim t(2)$, the squared errors $\mathrm{SE}(\widehat{\boldsymbol{\beta}}_{n0})$ of both LAD and HUB estimators

Table 1 The average (“aver.”) and standard deviation (“stdev.”) of SE($\widehat{\boldsymbol{\beta}}_{n0}$) values based on 200 repetitions

Error		Standard normal		$t(2)$ distribution		Mixture normal	
		aver.	stdev.	aver.	stdev.	aver.	stdev.
Model (I)							
oracle	LSE	0.32	(0.04)	0.94	(0.45)	0.80	(0.11)
	LAD	0.40	(0.05)	0.49	(0.07)	0.50	(0.07)
	HUB	0.33	(0.04)	0.48	(0.06)	0.47	(0.07)
robust	LSE	0.34	(0.04)	0.98	(0.47)	0.83	(0.12)
	LAD	0.42	(0.05)	0.57	(0.08)	0.55	(0.08)
	HUB	0.35	(0.04)	0.53	(0.07)	0.53	(0.08)
Model (II)							
oracle	LSE	0.33	(0.04)	1.00	(0.54)	0.78	(0.11)
	LAD	0.41	(0.04)	0.49	(0.06)	0.50	(0.06)
	HUB	0.33	(0.04)	0.49	(0.06)	0.47	(0.06)
robust	LSE	0.35	(0.04)	1.05	(0.57)	0.82	(0.12)
	LAD	0.43	(0.05)	0.58	(0.09)	0.56	(0.08)
	HUB	0.35	(0.04)	0.55	(0.08)	0.53	(0.07)
Model (III)							
oracle	LSE	0.33	(0.04)	0.93	(0.52)	0.79	(0.11)
	LAD	0.41	(0.05)	0.49	(0.07)	0.51	(0.07)
	HUB	0.33	(0.04)	0.47	(0.07)	0.48	(0.06)
robust	LSE	0.35	(0.04)	0.97	(0.52)	0.83	(0.12)
	LAD	0.43	(0.05)	0.57	(0.08)	0.56	(0.08)
	HUB	0.35	(0.04)	0.53	(0.07)	0.53	(0.08)
Model (IV)							
oracle	LSE	0.33	(0.04)	1.07	(1.34)	0.79	(0.11)
	LAD	0.42	(0.05)	0.50	(0.07)	0.51	(0.06)
	HUB	0.34	(0.04)	0.48	(0.07)	0.48	(0.06)
robust	LSE	0.35	(0.04)	1.12	(1.42)	0.82	(0.11)
	LAD	0.44	(0.05)	0.56	(0.10)	0.56	(0.07)
	HUB	0.36	(0.04)	0.53	(0.09)	0.52	(0.07)

are small, indicating that the robust estimates are asymptotically unbiased. This is because the issue of heavy tailed errors has been automatically taken into account in (2.4). The LSE performs the best when the error is standard normal, and the worst otherwise, which complies with our expectation. We can also see that different estimation procedures have similar performance in different models.

The averages and the standard deviations of SE values of $\widehat{\boldsymbol{\beta}}_n$ are displayed in Table 2. In comparison of $\widehat{\boldsymbol{\beta}}_{n0}$ with $\widehat{\boldsymbol{\beta}}_n$, we can see that the non-concave penalized estimate $\widehat{\boldsymbol{\beta}}_n$ has much better performance across all scenarios. The oracle estimate in Table 2 assumes both the nonlinear effects and the unimportant covariates are known a priori, which is not surprisingly better than the oracle estimate in Table 1.

Estimation of $\boldsymbol{\nu}(\cdot)$

Throughout our simulations we use the Epanechnikov kernel function and the generalized cross-validation to determine an optimal bandwidth. To evaluate the performance of estimating $\nu(\cdot)$, we adopt the median absolute deviation defined by

$$\mathrm{MAD} = \text{median}\left\{\left|\widehat{\nu}(\boldsymbol{z}_i) - \nu(\boldsymbol{z}_i)\right|, i = 1, \ldots, n\right\}. \tag{4.2}$$

Table 3 charted the averages and the standard deviations for three different estimation procedures: The oracle estimate which assumes $\boldsymbol{\beta}_{n0}$ is known in advance; the robust estimation which regresses $Y - \boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}_{n0}$ onto $\boldsymbol{z}$; and the penalized robust estimation which regresses $Y - \boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}_n$ onto $\boldsymbol{z}$. The three loss functions,

Table 2 The average (“aver.”) and standard deviation (“stdev.”) of SE($\widehat{\boldsymbol{\beta}}_n$) values based on 200 repetitions

Error		Standard normal		$t(2)$ distribution		Mixture normal	
		aver.	stdev.	aver.	stdev.	aver.	stdev.
Model (I)							
oracle	LSE	0.10	(0.03)	0.28	(0.15)	0.23	(0.08)
	LAD	0.12	(0.05)	0.15	(0.05)	0.15	(0.05)
	HUB	0.10	(0.04)	0.14	(0.05)	0.13	(0.05)
robust	LSE	0.14	(0.04)	0.46	(0.29)	0.36	(0.12)
	LAD	0.20	(0.09)	0.26	(0.17)	0.23	(0.11)
	HUB	0.19	(0.09)	0.26	(0.12)	0.25	(0.10)
Model (II)							
oracle	LSE	0.10	(0.03)	0.28	(0.18)	0.23	(0.09)
	LAD	0.12	(0.04)	0.14	(0.05)	0.14	(0.05)
	HUB	0.10	(0.03)	0.14	(0.05)	0.13	(0.05)
robust	LSE	0.14	(0.04)	0.49	(0.33)	0.36	(0.11)
	LAD	0.20	(0.09)	0.27	(0.23)	0.24	(0.10)
	HUB	0.18	(0.08)	0.27	(0.14)	0.25	(0.11)
Model (III)							
oracle	LSE	0.10	(0.03)	0.28	(0.19)	0.23	(0.09)
	LAD	0.12	(0.04)	0.14	(0.05)	0.14	(0.05)
	HUB	0.10	(0.03)	0.13	(0.05)	0.14	(0.05)
robust	LSE	0.14	(0.04)	0.45	(0.27)	0.36	(0.12)
	LAD	0.21	(0.08)	0.26	(0.25)	0.24	(0.10)
	HUB	0.19	(0.09)	0.26	(0.14)	0.25	(0.10)
Model (IV)							
oracle	LSE	0.10	(0.04)	0.31	(0.50)	0.23	(0.08)
	LAD	0.13	(0.04)	0.14	(0.05)	0.15	(0.05)
	HUB	0.11	(0.04)	0.13	(0.05)	0.13	(0.05)
robust	LSE	0.14	(0.04)	0.49	(0.39)	0.36	(0.11)
	LAD	0.20	(0.09)	0.28	(0.29)	0.23	(0.11)
	HUB	0.18	(0.08)	0.26	(0.21)	0.25	(0.10)

LSE, LAD and HUB, are considered here. It can be seen that the general robust estimation and the non-concave penalized robust estimation have quite similar performance, both of which are very close to the oracle estimates. The phenomenon, once again, verifies our theoretical observations in Theorems 3 and 6.

Variable selection

To measure the performance of the penalized robust estimation procedure in terms of variable selection, we calculated the average proportions of the zero coefficients, which are summarized in Table 4, in which the column labeled “C” presents the average proportion restricted only to the true zero coefficients, while the column labeled “IC” depicts the average of coefficients erroneously set to zero. All numbers in the “IC” columns are exactly zero, indicating that three methods can identify successfully all five important predictors. A closer inspection finds that the numbers of LSE in the “C” columns are smaller that those of the robust estimations such LAD and HUB, indicating that the penalized robust estimation procedures are better than the penalized least squares in terms of variable selection.

4.2 An application

Motor-car manufactures produce different types of vehicles with different levels of attributes such as miles per gallon and horsepower. Given these attributes, the manufacturers would like to know how they can

Table 3 The average (“aver.”) and standard deviation (“stdev.”) of median absolute deviation values obtained from nonconcave-penalized robust estimation procedure

Error		Standard normal		$t(2)$ distribution		Mixture normal	
		aver.	stdev.	aver.	stdev.	aver.	stdev.
Model (I)							
oracle	LSE	0.18	(0.03)	0.36	(0.06)	0.39	(0.06)
	LAD	0.22	(0.03)	0.27	(0.04)	0.28	(0.04)
	HUB	0.20	(0.03)	0.26	(0.04)	0.26	(0.04)
robust	LSE	0.25	(0.09)	0.70	(0.49)	0.62	(0.29)
	LAD	0.28	(0.09)	0.66	(0.49)	0.56	(0.27)
	HUB	0.26	(0.09)	0.65	(0.50)	0.55	(0.28)
penalized	LSE	0.24	(0.10)	0.50	(0.25)	0.46	(0.12)
	LAD	0.27	(0.09)	0.42	(0.24)	0.37	(0.11)
	HUB	0.25	(0.10)	0.41	(0.24)	0.35	(0.12)
Model (II)							
oracle	LSE	0.19	(0.03)	0.37	(0.07)	0.39	(0.06)
	LAD	0.23	(0.03)	0.27	(0.04)	0.28	(0.04)
	HUB	0.20	(0.03)	0.27	(0.04)	0.26	(0.04)
robust	LSE	0.27	(0.11)	0.72	(0.47)	0.58	(0.23)
	LAD	0.30	(0.11)	0.68	(0.46)	0.53	(0.25)
	HUB	0.28	(0.11)	0.66	(0.46)	0.51	(0.25)
penalized	LSE	0.23	(0.08)	0.51	(0.26)	0.45	(0.11)
	LAD	0.27	(0.08)	0.44	(0.27)	0.36	(0.10)
	HUB	0.24	(0.08)	0.43	(0.27)	0.35	(0.11)
Model (III)							
oracle	LSE	0.18	(0.02)	0.37	(0.06)	0.39	(0.06)
	LAD	0.23	(0.03)	0.27	(0.04)	0.28	(0.04)
	HUB	0.19	(0.02)	0.27	(0.04)	0.27	(0.04)
robust	LSE	0.27	(0.11)	0.68	(0.54)	0.59	(0.27)
	LAD	0.30	(0.10)	0.63	(0.55)	0.55	(0.29)
	HUB	0.28	(0.11)	0.63	(0.55)	0.53	(0.29)
penalized	LSE	0.24	(0.09)	0.49	(0.23)	0.44	(0.10)
	LAD	0.27	(0.08)	0.42	(0.22)	0.36	(0.09)
	HUB	0.25	(0.09)	0.41	(0.23)	0.35	(0.10)
Model (IV)							
oracle	LSE	0.18	(0.03)	0.36	(0.06)	0.38	(0.06)
	LAD	0.23	(0.03)	0.27	(0.04)	0.28	(0.04)
	HUB	0.20	(0.03)	0.26	(0.04)	0.26	(0.04)
robust	LSE	0.26	(0.11)	0.82	(0.98)	0.59	(0.26)
	LAD	0.29	(0.10)	0.77	(1.01)	0.54	(0.28)
	HUB	0.27	(0.11)	0.76	(1.03)	0.52	(0.28)
penalized	LSE	0.23	(0.09)	0.53	(0.29)	0.45	(0.11)
	LAD	0.27	(0.08)	0.45	(0.27)	0.36	(0.10)
	HUB	0.24	(0.09)	0.44	(0.27)	0.34	(0.11)

charge the highest price that consumers are willing to pay. It is then of natural interest to investigate how the prices of vehicles depend upon their attributes. We collect a data set which contains information about 428 new vehicles for the year 2004 [19]. Sixteen observations with missing values are removed from our subsequent analysis, leaving 412 data points.

Table 4 The average proportion of zero coefficients obtained by SCAD-penalized estimators, where LSE, LAD and HUB denote the penalized least squares estimation, the penalized least absolute deviation and the penalized Huber function estimation, respectively, "C" denotes the average proportion of the zero coefficients which are correctly estimated as zero, and "IC" denotes the average proportion of the nonzero coefficients which are erroneously set to zero

Error		Standard normal		$t(2)$ distribution		Mixture normal	
Model	Method	C	IC	C	IC	C	IC
(I)	LSE	0.90	0.00	0.90	0.05	0.91	0.03
	LAD	0.95	0.00	0.99	0.04	0.99	0.02
	HUB	0.97	0.00	0.97	0.02	0.97	0.01
(II)	LSE	0.91	0.00	0.90	0.06	0.89	0.02
	LAD	0.95	0.01	0.98	0.06	0.98	0.02
	HUB	0.97	0.00	0.97	0.03	0.97	0.01
(III)	LSE	0.91	0.00	0.90	0.06	0.91	0.03
	LAD	0.96	0.00	0.98	0.06	0.98	0.03
	HUB	0.97	0.00	0.97	0.03	0.97	0.02
(IV)	LSE	0.90	0.00	0.90	0.07	0.90	0.03
	LAD	0.95	0.00	0.99	0.07	0.98	0.02
	HUB	0.97	0.00	0.98	0.03	0.97	0.01

The manufacturer suggested retail price (MSRP) in U.S. dollars serves as the response variable. This price is what the manufacture thinks the vehicle is worth given the set of attributes, including adequate profit for the manufacturer and the dealer. There are twelve covariates which classify the vehicles. The first seven covariates are binary variables: The sport car (X_1), the sport utility vehicle (X_2), wagon (X_3), mini-van (X_4), pickup (X_5), all-wheel drive (X_6) and rear-wheel drive (X_7). Other five covariates are continuous: Engine size (X_8), number of cylinders (X_9), horsepower (X_{10}), weight (X_{11}) and wheel base (X_{12}). In addition, we choose $\boldsymbol{z} = (Z_1, Z_2)^\top$ where Z_1 denotes city miles per gallon (MPG) and Z_2 denotes highway MPG.

We first examine the empirical distribution of the MSRP values. The histogram of the standardized MSRP is presented in Figure 1(b). It is revealed that the distribution of the MSRP is highly skewed. We also examined the boxplot of the standardized MSRP in Figure 1(b), through which it can be seen that there exist a number of outliers. The presence of non-normality and outliers in the response imposes serious challenges for variable selection and subsequent inference. One may suggest to impose logarithm transformation on the response variable. However, our preliminary analysis indicates that the transformed response is still non-normal. Since the transformation may not remove the outliers and often brings additional issues for interpretability, we choose to analyze the response on its standardized scale.

We fit a partially linear model (1.1) and apply three different estimation procedures to this data

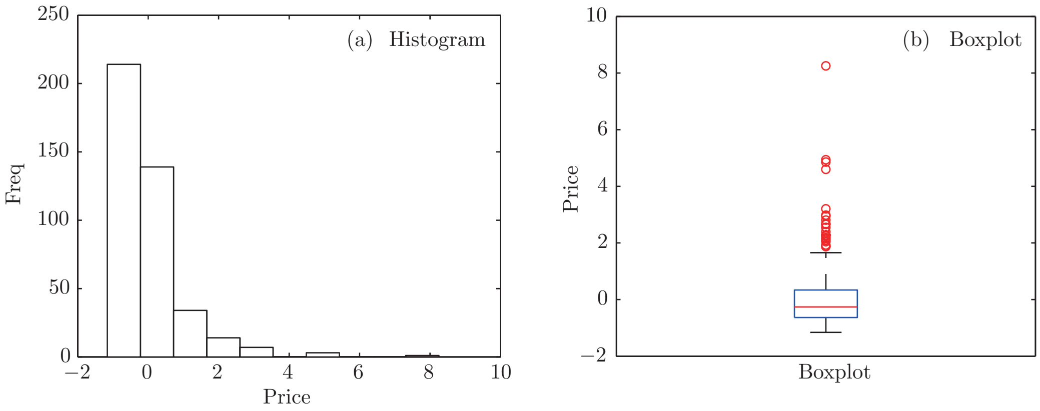

Figure 1 The empirical distribution of the MSRP values

set: LSE, LAD and HUB. The results are summarized in Table 5. The resulting estimators using all covariates are used as a benchmark for comparison. Next we apply penalized estimation procedures corresponding to LSE, LAD and HUB, respectively. The estimated coefficients and their standard errors are summarized in Table 6. Because all three penalized algorithms identify X_2, X_5, X_8, X_{10} and X_{12} as important variables, we re-run the above three un-penalized algorithms on this dataset. The results are again charted in Table 5. All above analysis conveys similar messages. For example, given other features, the sport utility vehicle (X_2) and pickup (X_5) are more expensive than other type of vehicles. The engine size (X_8), horsepower (X_{10}) and wheel base (X_{12}) are important factors affecting the MSRP, while others are not. It can also be seen that the penalized robust estimates (LAD and HUB) have smaller standard errors than the penalized least squares estimate (LSE), indicating that the penalized robust estimates are more accurate than the penalized least squares estimate when outliers and nonnormal errors are present.

Figure 2 presents the scatter plots of the observed MSRP values (on the vertical axis) versus the fitted values (on the horizontal axis) obtained from three estimation procedures, from which it can be seen that the penalized least squares estimate (LSE) is more influenced by the outliers than the other two procedures (LAD and HUB) because the fitted values are closer to the observed value on the right boundary.

5 Discussion

In this paper we study several robust estimation procedures to estimate the parameters in partially linear model. In many biomedical applications, however, the condition $p_n = o(n^{1/2})$ may be violated. In particular, in studies with microarray data as covariate measurements, the number of genes (covariates)

Table 5 The estimated coefficients of $\boldsymbol{x}$ in analysis of the new vehicle data, where LSE, LAD and HUB denote the least squares estimation, the least absolute deviation and the Huber function estimation, respectively. For each method, the first line utilizes all covariates $(X_1, \ldots, X_{12})^\top$, and the second line merely utilizes $(X_2, X_5, X_8, X_{10}, X_{12})^\top$

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
LSE												
all	−0.009	−0.534	−0.023	−0.124	−0.545	−0.026	0.091	−0.300	0.052	0.009	0.000	−0.024
subset		−0.518			−0.520			−0.214		0.010		−0.023
LAD												
all	−0.008	−0.434	−0.002	−0.172	−0.663	−0.014	0.181	−0.259	0.010	0.006	0.000	−0.012
subset		−0.376			−0.625			−0.252		0.007		−0.010
HUB												
all	0.034	−0.485	−0.036	−0.261	−0.674	−0.021	0.086	−0.328	0.060	0.006	0.000	−0.013
subset		−0.388			−0.571			−0.229		0.007		−0.014

Table 6 The estimated coefficients of $\boldsymbol{x}$ and their corresponding standard deviations in analysis of the new vehicle data, where LSE, LAD and HUB denote the penalized least squares estimation, the penalized least absolute deviation and the penalized Huber function estimation, respectively

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
LSE												
coef	0	−0.520	0	0	−0.544	0	0	−0.223	0	0.010	0	−0.021
stdev.		0.108			0.164			0.062		0.001		0.005
LAD												
coef	0	−0.316	0	0	−0.590	0	0	−0.075	0	0.006	0	−0.022
stdev.		0.037			0.097			0.009		0.001		0.003
HUB												
coef	0	−0.456	0	0	−0.592	0	0	−0.255	0	0.007	0	−0.017
stdev.		0.080			0.103			0.043		0.001		0.004

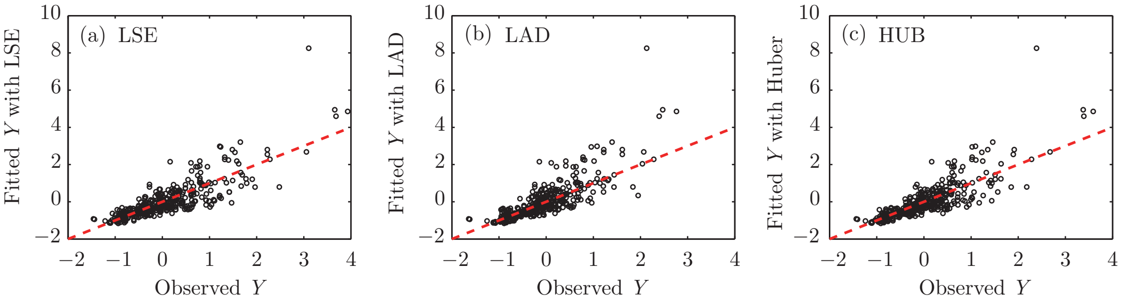

Figure 2 The scatter plots of the observed MSRP values (on the vertical axis) versus the fitted values (on the horizontal axis) obtained from three estimation procedures, where LSE, LAD and HUB denote the penalized least squares estimation, the penalized least absolute deviation and the penalized Huber function estimation, respectively.

is typically greater than the sample size. How to apply the penalized robust estimation to adapt those studies is of both theoretical and practical importance.

In our context, we allow the covariate vector $\boldsymbol{z}$ to be multivariate. It is remarkable here that, our estimation procedures may encounter the curse of dimensionality when the dimension of $\boldsymbol{z}$ is large. In that situation, we may consider partially linear single-index model [4] to tackle this issue. With an additional single-index structure, more elegant technical derivations are often expected. These issues are currently under investigation.

Acknowledgements This work was supported by National Institute on Drug Abuse (Grant Nos. R21-DA024260 and P50-DA10075), National Natural Science Foundation of China (Grant Nos. 11071077, 11371236, 11028103, 11071022 and 11028103), Innovation Program of Shanghai Municipal Education Commission, Pujiang Project of Science and Technology Commission of Shanghai Municipality (Grant No. 12PJ1403200) and Program for New Century Excellent Talents, Ministry of Education of China (Grant No. NCET-12-0901).

References

1 Bai Z D, Rao C R, Wu Y. *M*-estimation of multivariate linear regression parameters under a convex discrepancy function. Statist Sinica, 1992, 2: 237–254

2 Bai Z D, Wu Y. Limit behavior of *M*-estimators or regression coefficients in high dimensional linear models I: Scale-dependent case. J Multivariate Anal, 1994, 51: 211–239

3 Boente G, He X M, Zhou J H. Robust estimates in generalized partially linear models. Ann Statist, 2006, 34: 2856–2878

4 Carroll R J, Fan J, Gijbels I, et al. Generalized partially linear single-index models. J Amer Statist Assoc, 1997, 92: 477–489

5 Chen H. Convergence rates for parametric components in a partly linear model. Ann Statist, 1988, 16: 136–146

6 Engle R F, Granger C W J, Rice J, et al. Semiparametric estimates of the relation between weather and electricity sales. J Amer Statist Assoc, 1986, 81: 310–320

7 Fan J, Gijbels I. Local Polynomial Modeling and its Applications. New York: Chapman and Hall, 1996.

8 Fan J, Hu T C, Truong Y K. Robust nonparametric function estimation. Scandinavian J Statist, 1994, 21: 433–446

9 Fan J, Li R. Variable selection via nonconcave penalized likelihood and it oracle properties. J Amer Statist Assoc, 2001, 96: 1348–1360

10 Fan J, Li R. New estimation and model selection procedures for semi-parametric modeling in longitudinal data analysis. J Amer Statist Assoc, 2004, 99: 710–723

11 Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. Ann Statist, 2004, 32: 928–961

12 Hamilton S A, Truong Y K. Local linear estimation in partly linear models. J Multivariate Anal, 1997, 60: 1–19

13 Härdle W, Liang H, Gao J T. Partial Linear Models. New York: Springer-Verlag, 2000

14 He X M, Fung W K, Zhu Z Y. Robust estimation in generalized partial linear models for clustered data. J Amer Statist Assoc, 2005, 100: 1176–1184

15 Heckman N E. Spline smoothing in a partly linear model. J Royal Statist Soc Ser B, 1986, 48: 244–248

16 Huang J, Xie H. Asymptotic oracle properties of SCAD-penalized least square estimators. In: Institute of Mathematical Statistics Lecture Notes Monograph Seriess vol. 55. Asymptotics: Particles, Processes and Inverse Problems. Beachwood: IMS, 2007, 149–166

17 Huber P J. Robust regression: Asymptotics, conjectures and Monte Carlo. Ann Statist, 1973, 1: 799–821

18 Hunter D, Li R. Variable selection using MM algorithms. Ann Statist, 2005, 33: 1617–1642

19 Johnson R W. Kiplinger's personal finance. J Statist Education, 2003, 57: 104–123

20 Li G R, Peng H, Zhu L X. Nonconcave penalized M-estimation with diverging number of parameters. Statist Sinica, 2011, 21: 391–420

21 Liang H, Li R. Variable selection for partially linear models with measurement errors. J Amer Statist Assoc, 2009, 104: 234–248

22 Mammen E. Asymptotics with increasing dimension for robust regression with application to the bootstrap. Ann Statist, 1989, 17: 382–400

23 Qin G Y, Zhu Z Y. Robustified maximum likelihood estimation in generalized partial linear mixed model for longitudinal data. Biometrics, 2009, 65: 52–59

24 Rao B L S P. Nonparametric Functional Estimation. Orlando: Academic Press, 1983

25 Robinson P M. Root-n-consistent semiparametric regression. Econometrika, 1988, 56: 931–954

26 Speckman P. Kernel smoothing in partial linear models. J Royal Statist Soc Ser B, 1988, 50: 413–436

27 Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. J Royal Statist Soc Ser B, 2009, 71: 671–683

28 Wang H, Li G, Jiang G. Robust regression shrinkage and consistent variable selection via the LAD-LASSO. J Business Economics Statist, 2007, 25: 347–355

29 Wang H, Li R, Tsai C L. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika, 2007, 94: 553–568

30 Wang L, Li R. Weighted Wilcoxon-type smoothly clipped absolute deviation method. Biometrics, 2009, 65: 564–571

31 Wu W B. M-estimation of linear models with dependent errors. Ann Statist, 2007, 35: 495–521

32 Xie H L, Huang J. SCAD-penalized regression in high-dimensional partially linear models. Ann Statist, 2009, 37: 673–696

33 Zhu L X, Fang K T. Asymptotics for kernel estimation of sliced inverse regression. Ann Statist, 1996, 3: 1053–1068

34 Zou H, Yuan M. Composite quantile regression and the oracle model selection theory. Ann Statist, 2008, 36: 1108–1126

Appendix Proofs of theorems

Appendix A Some regularity conditions

To establish the asymptotics for the robust estimates, we present the following regularity conditions. These conditions are not the weakest possible conditions, but they are imposed to facilitate the technical derivations.

(C1) The loss function $\rho(\cdot)$ is convex on $\mathbb{R}^1$ with right and left derivative $\psi_+(\cdot)$ and $\psi_-(\cdot)$. Choose $\psi(\cdot)$ such that $\psi_-(\cdot) \leqslant \psi(\cdot) \leqslant \psi_+(\cdot)$, for all $t \in \mathbb{R}^1$. Let $\psi(\cdot)$ be any choice of the subgradient of $\rho(\cdot)$ and denote $\mathcal{S}$ as the set of discontinuity points of ψ, which is the same for all choices of $\psi(\cdot)$. We assume that $\psi(\cdot)$ satisfies the first order Lipschitz continuity. That is, there exist positive constants κ and C such that $|\psi(x+t) - \psi(x)| \leqslant C|t|$, for all $x \in \mathbb{R}^1$ and $|t| \leqslant \kappa$.

(C2) The common distribution function F of ε_i satisfies $F(\mathcal{S}) = 0$. In addition, we assume that $E\{\psi(\varepsilon_1)\} = 0$, $0 < E\{\psi^2(\varepsilon_1)\} = \sigma^2 < \infty$, and $G(t) =: E\{\psi(\varepsilon_1 + t)\} = \gamma t + o(|t|)$, as $t \to 0$, where γ is a positive constant.

(C3) Suppose $\max_{1\leqslant k\leqslant p} E(X_k^4) < \infty$. In addition, there exist N_0 and constants b and B such that for $n \geqslant N_0$,

(i) $0 < b \leqslant \rho_1(\mathbf{\Sigma}_n) \leqslant \rho_{p_n}(\mathbf{\Sigma}_n) \leqslant B$, where $\mathbf{\Sigma}_n = \mathrm{cov}(\widetilde{\boldsymbol{x}})$, or

(ii) $0 < b \leqslant \rho_1(\mathbf{\Sigma}_{n\boldsymbol{I}}) \leqslant \rho_{p_n}(\mathbf{\Sigma}_{n\boldsymbol{I}}) \leqslant B$, where $\mathbf{\Sigma}_{n\boldsymbol{I}} = \mathrm{cov}(\widetilde{\boldsymbol{x}}_{\boldsymbol{I}})$.

The subscript $\boldsymbol{I}$ denotes the index set of important predictors in $\boldsymbol{x}$. Conditions (C1)–(C2) are often imposed in the M-estimation theory of linear model. Condition (C3) is often assumed to study problems with diverging number of covariates. See [1] and [31] for more discussions about these conditions.

(C4) The kernel function $K(\cdot)$ used in (2.3) has a compact support $[-1,1]$. It satisfies $\int_{-1}^1 K(v)dv = 1$, $\int_{-1}^1 vK(v)dv = 0$. The bandwidths h_k in (2.3) satisfy $nh_k^8 \to 0$ and $nh_k^{2d} \to \infty$, here d is the dimension of $\boldsymbol{z}$.

(C5) The density function $f(\cdot)$ of $\boldsymbol{z}$ is continuous, and $f(\boldsymbol{z}) > 0$.

(C6) The conditional density function $g(y^*|\boldsymbol{z})$ given $\boldsymbol{z}$ is continuous in $\boldsymbol{z}$ for each $y^* = Y - \boldsymbol{x}^\top \boldsymbol{\beta}_{n0}$. More-over, there exist positive constants ε and δ and a positive function $G(y^*|\boldsymbol{z})$ such that $\sup_{|\boldsymbol{z}_n-\boldsymbol{z}|\leqslant\varepsilon} g(y^*|\boldsymbol{z}_n) \leqslant G_0(y^*|\boldsymbol{z})$ and that

$$\int |G\{y^* - \nu(\boldsymbol{x})\}|^{2+\delta} G_0(y^*|\boldsymbol{z}) d\mu(y^*) < \infty,$$

and

$$\int |\nu(y^* - t) - \nu(y^*) - \nu'(y^*)t|^2 G_0(y^*|\boldsymbol{z}) d\mu(y^*) = o(t^2), \quad \text{as } t \to 0.$$

(C7) The function $\nu(\boldsymbol{z})$ has a continuous second derivative.

Let $a_n = \max\{|p'_{\lambda_n}(|\beta_{n0,j}|)| : \beta_{n0,j} \neq 0\}$ and $b_n = \max\{|p''_{\lambda_n}(|\beta_{n0,j}|)| : \beta_{n0,j} \neq 0\}$, where we write λ as λ_n to emphasize that λ_n depends on the sample size n. Then the conditions are as follows:

(C8) $\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$.

(C9) $a_n = O(n^{-1/2})$.

(C10) $b_n \to 0$ as $n \to \infty$.

(C11) There are constants C and D such that $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \leqslant D|\theta_1 - \theta_2|$ for any $\theta_1, \theta_2 > C\lambda_n$.

Because Theorems 1–3 are respectively parallel to Theorems 4–6. In the sequel, we will only prove Theorems 4–6 and point out how the technical derivations can be adapted to prove Theorems 1–3.

Appendix B Proof of Theorem 4

In the sequel we will only sketch some key procedures because the major derivations are very standard in the nonconcave-penalized likelihood literature. One can refer to [9] and [11] for more technical details.

Let $\alpha_n = {p_n}^{1/2}(n^{-1/2} + a_n)$. We will show that, for any given $\epsilon > 0$, there exists a large constant C such that

$$\Pr\Big\{ \inf_{\|\boldsymbol{u}\|=C} Q_n(\boldsymbol{\beta}_{n0} + \alpha_n \boldsymbol{u}) > Q_n(\boldsymbol{\beta}_{n0}) \Big\} \geqslant 1 - \epsilon. \tag{B.1}$$

Denote

$$J_n(\boldsymbol{u}) = \sum_{i=1}^n \rho\{\widehat{\widetilde{Y}}_i - \widehat{\widetilde{\boldsymbol{x}}}_i^\top (\boldsymbol{\beta}_{n0} + \alpha_n \boldsymbol{u})\} - \sum_{i=1}^n \rho(\widehat{\widetilde{Y}}_i - \widehat{\widetilde{\boldsymbol{x}}}_i^\top \boldsymbol{\beta}_{n0}).$$

Let $s = (\widehat{\widetilde{Y}}_i - \widetilde{Y}_i) - (\widehat{\widetilde{\boldsymbol{x}}}_i - \widetilde{\boldsymbol{x}}_i)^\top \boldsymbol{\beta}_{n0}$. Then by Condition (C1) $J_n(\boldsymbol{u})$ can be expressed as follows,

$$\sum_{i=1}^n \int_0^{-\alpha_n \widehat{\widetilde{\boldsymbol{x}}}_i^\top \boldsymbol{u}} \{\psi(\varepsilon_i + t + s) - \psi(\varepsilon_i)\} dt - \alpha_n \sum_{i=1}^n \psi(\varepsilon_i) \widehat{\widetilde{\boldsymbol{x}}}_i^\top \boldsymbol{u} =: J_{n1}(\boldsymbol{u}) + J_{n2}(\boldsymbol{u}).$$

Invoking Condition (C2), for $J_{n1}(\boldsymbol{u})$ we have

$$\begin{aligned} E\{J_{n1}(\boldsymbol{u})\} &= \sum_{i=1}^n E\bigg[\int_0^{-\alpha_n \widehat{\widetilde{\boldsymbol{x}}}_i^\top \boldsymbol{u}} \{\gamma(t+s) + o(|t+s|)\} dt \bigg] \\ &= \frac{\gamma\{1+o(1)\}}{2} \sum_{i=1}^n E[(\alpha_n \widehat{\widetilde{\boldsymbol{x}}}_i^\top \boldsymbol{u})^2 - 2\{(\widehat{\widetilde{Y}}_i - \widetilde{Y}_i) - (\widehat{\widetilde{\boldsymbol{x}}}_i - \widetilde{\boldsymbol{x}}_i)^\top \boldsymbol{\beta}_{n0}\} \alpha_n \widehat{\widetilde{\boldsymbol{x}}}_i^\top \boldsymbol{u}] \\ &=: J_{n11}(\boldsymbol{u}) + J_{n12}(\boldsymbol{u}). \end{aligned}$$

It is easy to see that $J_{n11}(\boldsymbol{u})$ is positive if γ is positive. In addition, Condition (C3) and the uniform convergence of $\widehat{\widetilde{\boldsymbol{x}}}_i^\top \boldsymbol{u}$ [24, 33] entail that $J_{n11}(\boldsymbol{u}) = O(n\alpha_n^2 \|\boldsymbol{u}\|^2)$. Following similar arguments, we can have,

$$\|J_{n12}(\boldsymbol{u})\| = \bigg\| E\bigg[\sum_{i=1}^n \{(\widehat{\widetilde{Y}}_i - \widetilde{Y}_i) - (\widehat{\widetilde{\boldsymbol{x}}}_i - \widetilde{\boldsymbol{x}}_i)^\top \boldsymbol{\beta}_{n0}\} \alpha_n \widehat{\widetilde{\boldsymbol{x}}}_i^\top \boldsymbol{u} \bigg] \bigg\| = o(1) O(n^{1/2} \alpha_n \|\boldsymbol{u}\|).$$

Recall the definition of α_n. We can see that $E\{J_{n1}(\boldsymbol{u})\}$ is dominated by $J_{n11}(\boldsymbol{u})$ because $J_{n12}(\boldsymbol{u}) = o(n\alpha_n^2 \|\boldsymbol{u}\|^2)$. That is,

$$E\{J_{n1}(\boldsymbol{u})\} = O(n\alpha_n^2 \|\boldsymbol{u}\|^2). \tag{B.2}$$

Recall the definition of $\widetilde{\boldsymbol{x}}$. We have $E(\alpha_n\widetilde{\boldsymbol{x}}^\top\boldsymbol{u})=0$. By invoking Condition (C3), we have $\mathrm{var}(\alpha_n\widetilde{\boldsymbol{x}}^\top\boldsymbol{u})=O(\alpha_n^2\|\boldsymbol{u}\|^2)$. Thus, $\alpha_n\widetilde{\boldsymbol{x}}^\top\boldsymbol{u}=O_P(\alpha_n\|\boldsymbol{u}\|)$, which tends to zero as $n\to\infty$ because $\|\boldsymbol{u}\|=C$ and $p_n\log(n)/n\to 0$ as $n\to\infty$.

Next we prove that

$$\mathrm{var}\left\{J_{n1}(\boldsymbol{u})\right\}=O(np_n^2\alpha_n^4\|\boldsymbol{u}\|^4),\qquad \text{as } n\to\infty. \tag{B.3}$$

To prove (B.3), we note that Condition (C1) yields that

$$\begin{aligned}\mathrm{var}\{J_{n1}(\boldsymbol{u})\}&=\mathrm{var}\left[\sum_{i=1}^n\int_0^{-\alpha_n\widehat{\widetilde{\boldsymbol{x}}}_i^\top\boldsymbol{u}}\{\psi(\varepsilon_i+t+s)-\psi(\varepsilon_i)\}dt\right]\\&\leqslant\sum_{i=1}^n C^2\mathrm{var}\{(\alpha_n\widetilde{\boldsymbol{x}}_i^\top\boldsymbol{u})^2\}\{1+o(1)\}=O(n\alpha_n^4p_n^2\|\boldsymbol{u}\|^4).\end{aligned}$$

The last inequality follows again due to the uniform convergence of $\widehat{\widetilde{\boldsymbol{x}}}_i^\top\boldsymbol{u}$, and the last equation holds by invoking Condition (C3).

By combining (B.2) and (B.3), it can be seen that

$$J_{n1}(\boldsymbol{u})=J_{n1}(\boldsymbol{u})-E\left\{J_{n1}(\boldsymbol{u})\right\}+E\left\{J_{n1}(\boldsymbol{u})\right\}=O_P(n^{1/2}p_n\alpha_n^2\|\boldsymbol{u}\|^2)+O(n\alpha_n^2\|\boldsymbol{u}\|^2), \tag{B.4}$$

which is positive when $\|\boldsymbol{u}\|$ is sufficiently large, because $p_n^2/n\to 0$ as $n\to\infty$. In addition, following similar arguments in [20], we can show that $J_{n1}(\boldsymbol{u})$ dominates $J_{n2}(\boldsymbol{u})$ by taking a sufficiently large constant C. We remark here that this proves Theorem 1.

Recall the equation defined in (B.1). We write that

$$\begin{aligned}D_n(\boldsymbol{u})&:=Q_n(\boldsymbol{\beta}_{n0}+\alpha_n\boldsymbol{u})-Q_n(\boldsymbol{\beta}_{n0})\\&\geqslant J_{n1}(\boldsymbol{u})+J_{n2}(\boldsymbol{u})+n\sum_{j=1}^{k_n}\{p_{\lambda_n}(|\beta_{n0,j}+\alpha_nu_j|)-p_{\lambda_n}(|\beta_{n0,j}|)\},\end{aligned} \tag{B.5}$$

where k_n is the dimension of $\boldsymbol{\beta}_{n\boldsymbol{I}}$. As shown in [20], the third term in (B.5) is bounded by $J_{n2}(\boldsymbol{u})$ by invoking the assumptions (C8)–(C10). Therefore, by taking C large enough, $J_{n1}(\boldsymbol{u})$ dominates $J_{n2}(\boldsymbol{u})$ and the third term in (B.5). Recall that $J_{n1}(\boldsymbol{u})$ is positive, which is shown in the statement about (B.4). This proves (B.1), and hence completes the proof of Theorem 4. □

Appendix C Proof of Theorem 5

To enhance the readability, we split the proof of Theorem 5 into two parts. We first prove the sparsity, and then turn to the asymptotic normality part.

Proof of sparsity. We now prove the sparsity. Theorem 4 shows that $\|\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_{n0}\|=O_P(\alpha_n)$, where $\alpha_n=p_n^{1/2}(n^{-1/2}+a_n)$. Thus, it suffices to show that, for any $j=k_n+1,\dots,p_n$, $\partial Q_n(\widehat{\boldsymbol{\beta}}_n)/\partial\widehat{\beta}_{nj}>0$ for $0<\widehat{\beta}_{nj}<\varepsilon=C\alpha_n$, and $\partial Q_n(\widehat{\boldsymbol{\beta}}_n)/\partial\widehat{\beta}_{nj}<0$ for $-\varepsilon<\widehat{\beta}_{nj}<0$, where $\widehat{\beta}_{nj}$ is the j-th component in $\widehat{\boldsymbol{\beta}}_n$. Let

$$\Lambda_i(\boldsymbol{\theta})=\int_0^{\widetilde{\boldsymbol{x}}_i^\top\boldsymbol{\theta}}\left\{\psi\left(\varepsilon_i+t\right)-\psi\left(\varepsilon_i\right)\right\}dt,\qquad\text{and}\qquad\Lambda_i^*(\boldsymbol{\theta})=\int_0^{\widetilde{\boldsymbol{x}}_i^\top\boldsymbol{\theta}}\{\psi(\varepsilon_i+t)-\psi(\varepsilon_i-\alpha_n\widetilde{\boldsymbol{x}}_i^\top\boldsymbol{u})\}dt.$$

It follows from [2, (3.2) in p. 219 and (3.30) in p. 227] that, for any fixed $\kappa>0$,

$$\sup_{\|\boldsymbol{\theta}\|\leqslant\kappa\alpha_n}\left|\alpha_n^{-1}\pi^{-1}(\boldsymbol{\theta})\sum_{i=1}^n\left[\Lambda_i(\boldsymbol{\theta})-E\left\{\Lambda_i(\boldsymbol{\theta})\right\}\right]\right|=o_P(1), \tag{C.1}$$

and

$$\sup_{\|\boldsymbol{\theta}\|\leqslant\kappa\alpha_n}\sup_{\|\boldsymbol{u}\|\leqslant C}\left|\alpha_n^{-1}\pi^{-1}(\boldsymbol{\theta})\sum_{i=1}^n\left[\Lambda_i^*(\boldsymbol{\theta})-E\left\{\Lambda_i^*(\boldsymbol{\theta})\right\}\right]\right|=o_P(1), \tag{C.2}$$

where $\pi(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\| \vee 1$. Then, subtracting (C.2) from (C.1), we obtain that

$$\begin{aligned}\sup_{\|\boldsymbol{\theta}\|\leqslant\kappa\alpha_n}\sup_{\|\boldsymbol{u}\|\leqslant C}\bigg|\alpha_n^{-1}\pi^{-1}(\boldsymbol{\theta})\sum_{i=1}^n([\{\psi(\varepsilon_i-\alpha_n\widetilde{\boldsymbol{x}}_i^\top\boldsymbol{u})-\psi(\varepsilon_i)\}\widetilde{\boldsymbol{x}}_i^\top\boldsymbol{\theta}]&\\ -E[\{\psi(\varepsilon_i-\alpha_n\widetilde{\boldsymbol{x}}^\top\boldsymbol{u})-\psi(\varepsilon_i)\}\widetilde{\boldsymbol{x}}^\top\boldsymbol{\theta}])\bigg|&=o_P(1),\end{aligned}$$

from which it can be derived that,

$$\begin{aligned}\sup_{\|\boldsymbol{\theta}\|=\kappa\alpha_n}\sup_{\|\boldsymbol{u}\|\leqslant C}\bigg|\alpha_n^{-1}\sum_{i=1}^n([\{\psi(\varepsilon_i-\alpha_n\widetilde{\boldsymbol{x}}_i^\top\boldsymbol{u})-\psi(\varepsilon_i)\}\widetilde{\boldsymbol{x}}_i^\top\boldsymbol{\theta}]&\\ -E[\{\psi(\varepsilon_i-\alpha_n\widetilde{\boldsymbol{x}}^\top\boldsymbol{u})-\psi(\varepsilon_i)\}\widetilde{\boldsymbol{x}}^\top\boldsymbol{\theta}])\bigg|&=o_P(1).\end{aligned}$$

This is equivalent to

$$\sup_{\|\boldsymbol{u}\|\leqslant C}\bigg\|\alpha_n^{-1}\sum_{i=1}^n([\{\psi(\varepsilon_i-\alpha_n\widetilde{\boldsymbol{x}}_i^\top\boldsymbol{u})-\psi(\varepsilon_i)\}\widetilde{\boldsymbol{x}}_i]-E[\{\psi(\varepsilon_i-\alpha_n\widetilde{\boldsymbol{x}}^\top\boldsymbol{u})-\psi(\varepsilon_i)\}\widetilde{\boldsymbol{x}}])\bigg\|=o_P(1). \tag{C.3}$$

By Condition (C2), we can show without much difficulty that

$$\begin{aligned}E[\{\psi(\varepsilon_i-\alpha_n\widetilde{\boldsymbol{x}}^\top\boldsymbol{u})-\psi(\varepsilon_i)\}\widetilde{\boldsymbol{x}}]&=E(E[\{\psi(\varepsilon_i-\alpha_n\widetilde{\boldsymbol{x}}^\top\boldsymbol{u})-\psi(\varepsilon_i)\}|\boldsymbol{x},\boldsymbol{z}]\widetilde{\boldsymbol{x}})\\&=-\gamma E[\widetilde{\boldsymbol{x}}\{\alpha_n\widetilde{\boldsymbol{x}}^\top\boldsymbol{u}+o(\alpha_n\widetilde{\boldsymbol{x}}^\top\boldsymbol{u})\}]\\&=-\gamma\alpha_n\mathrm{cov}(\widetilde{\boldsymbol{x}})\boldsymbol{u}+o(\alpha_n),\end{aligned}$$

which, together with (C.3), entails that

$$\sup_{\|\boldsymbol{u}\|\leqslant C}\bigg\|\sum_{i=1}^n([\{\psi(\varepsilon_i-\alpha_n\widetilde{\boldsymbol{x}}_i^\top\boldsymbol{u})-\psi(\varepsilon_i)\}\widetilde{\boldsymbol{x}}_i]+\gamma\alpha_n\mathrm{cov}(\widetilde{\boldsymbol{x}})\boldsymbol{u})\bigg\|=o_P(1). \tag{C.4}$$

Let $\delta_i=\widehat{\widetilde{Y}}_i-\widetilde{Y}_i-(\widehat{\widetilde{\boldsymbol{x}}}_i-\widetilde{\boldsymbol{x}}_i)^\top\boldsymbol{\beta}_{n0}$. The uniform convergence implies that $\max_{1\leqslant i\leqslant n}\|\delta_i\|=o_p(1)$ almost surely. Therefore,

$$\sup_{\|\boldsymbol{u}\|\leqslant C}\bigg\|\sum_{i=1}^n([\{\psi(\varepsilon_i+\delta_i-\alpha_n\widetilde{\boldsymbol{x}}_i^\top\boldsymbol{u})-\psi(\varepsilon_i)\}\widetilde{\boldsymbol{x}}_i]+\gamma\alpha_n\mathrm{cov}(\widetilde{\boldsymbol{x}})\boldsymbol{u})\bigg\|=o_P(1).$$

By invoking Theorem 4 that $\|\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_{n0}\|=\alpha_n\|\boldsymbol{u}\|$, it follows that

$$\sum_{i=1}^n\psi(\widehat{\widetilde{Y}}_i-\widehat{\widetilde{\boldsymbol{x}}}_i^\top\widehat{\boldsymbol{\beta}}_n)\widetilde{\boldsymbol{x}}_i=\sum_{i=1}^n\psi(\varepsilon_i)\widetilde{\boldsymbol{x}}_i-n\gamma\mathrm{cov}(\widetilde{\boldsymbol{x}})(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_{n0})+o_P(1).$$

Use again Condition (C3). It is easy to check that

$$E\bigg\|\sum_{i=1}^n\psi(\varepsilon_i)\widetilde{\boldsymbol{x}}_i\bigg\|^2=\sum_{j=1}^{p_n}\sum_{i=1}^n\sum_{l=1}^n E\{\widetilde{X}_{ij}\widetilde{X}_{lj}\psi(\varepsilon_i)\psi(\varepsilon_l)\}=\sum_{j=1}^{p_n}\sum_{i=1}^n E(\widetilde{X}_{ij}^2)E\{\psi^2(\varepsilon_i)\}=O(np_n).$$

Consequently, $\sum_{i=1}^n\psi(\widehat{\widetilde{Y}}_i-\widehat{\widetilde{\boldsymbol{x}}}_i^\top\widehat{\boldsymbol{\beta}}_n)\widehat{\widetilde{X}}_{ij}=O_P\{(np_n)^{1/2}\}$. Recall the condition that $(p_n/n)^{1/2}/\lambda_n\to 0$. By using Condition (C8) and the fact that

$$\begin{aligned}\frac{\partial Q_n(\widehat{\boldsymbol{\beta}}_n)}{\partial\widehat{\beta}_{nj}}&=-\sum_{i=1}^n\psi(\widehat{\widetilde{Y}}_i-\widehat{\widetilde{\boldsymbol{x}}}_i^\top\widehat{\boldsymbol{\beta}}_n)\widehat{\widetilde{X}}_{ij}+np'_{\lambda_n}(|\widehat{\beta}_{n,j}|)\mathrm{sign}(\widehat{\beta}_{n,j})\\&=n\lambda_n\bigg[-O_P\{(p_n/n)^{1/2}/\lambda_n\}+\frac{p'_{\lambda_n}(|\widehat{\beta}_{n,j}|)}{\lambda_n}\mathrm{sign}(\widehat{\beta}_{n,j})\bigg],\end{aligned}$$

we can see that the sign of $\widehat{\beta}_{n,j}$ completely determines the sign of $\frac{\partial Q_n(\widehat{\boldsymbol{\beta}}_n)}{\partial \widehat{\beta}_{n,j}}$, i.e.,

$$\frac{\partial Q_n(\widehat{\boldsymbol{\beta}}_n)}{\partial \widehat{\beta}_{n,j}} = \begin{cases} > 0, & \text{for } 0 < \widehat{\beta}_{n,j} < \varepsilon, \\ < 0, & \text{for } -\varepsilon < \widehat{\beta}_{n,j} < 0, \end{cases}$$

where $j = k_n + 1, \ldots, p$. This completes the proof of sparsity part.

Proof of asymptotic normality. Next, we prove the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$. Theorem 4 proves that $\widehat{\boldsymbol{\beta}}_n$ is root-n/p_n consistent. Thus, each component of $\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}}$ stays away from zero for n sufficiently large because $\boldsymbol{\beta}_{n\boldsymbol{I}}$ is away from zero. Therefore, the partial derivatives exist for the first k_n components. As a consequence, the estimate $\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}}$ based on the penalized robust estimation are necessarily the solution of the following estimation equation

$$-\sum_{i=1}^{n} \widehat{\widetilde{\boldsymbol{x}}}_i \psi(\widehat{\widetilde{Y}}_i - \widehat{\widetilde{\boldsymbol{x}}}_i^{\top} \widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}}) + nP'_{\lambda_n}(|\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}}|) = 0, \tag{C.5}$$

where $P'_{\lambda_n}(|\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}}|)$ is a $k_n \times 1$ vector whose j-th element is $p'_{\lambda_n}(|\widehat{\beta}_{n\boldsymbol{I},j}|)\text{sign}(|\widehat{\beta}_{n\boldsymbol{I},j}|)$. Let $\widehat{\varepsilon}_i = \widehat{\widetilde{Y}}_i - \widehat{\widetilde{\boldsymbol{x}}}_i^{\top} \boldsymbol{\beta}_n$. By using Taylor expansion for (C.5) and re-arranging the resulting terms, we have

$$\begin{aligned} & n\{(\gamma \boldsymbol{S}_{1n} + \boldsymbol{\Sigma}_{\lambda_n})(\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}} - \boldsymbol{\beta}_{n\boldsymbol{I}}) + \boldsymbol{b}_n\} \\ & \quad = \sum_{i=1}^{n} \psi(\varepsilon_i) \widehat{\widetilde{\boldsymbol{x}}}_{\boldsymbol{I}i} \sum_{i=1}^{n} \left[-\{\psi'(\varepsilon_i) - \gamma\} \widehat{\widetilde{\boldsymbol{x}}}_{\boldsymbol{I}i} \widehat{\widetilde{\boldsymbol{x}}}_{\boldsymbol{I}i}^{\top} (\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}} - \boldsymbol{\beta}_{n\boldsymbol{I}}) + \frac{1}{2} \psi''(\widehat{\widetilde{Y}}_i - \widehat{\widetilde{\boldsymbol{x}}}_i^{\top} \boldsymbol{\beta}_n^*) \{\widehat{\widetilde{\boldsymbol{x}}}_{\boldsymbol{I}i}^{\top} (\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}} - \boldsymbol{\beta}_{n\boldsymbol{I}})\}^2 \widehat{\widetilde{\boldsymbol{x}}}_{\boldsymbol{I}i} \right], \end{aligned}$$

where $\boldsymbol{\beta}_n^*$ is a vector between $\boldsymbol{\beta}_{n\boldsymbol{I}}$ and $\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}}$. Multiply both sides of the above equation by $n^{-1/2} \boldsymbol{A}_{n\boldsymbol{I}} \boldsymbol{S}_{n\boldsymbol{I}}^{-1/2}$, where $\boldsymbol{S}_{n\boldsymbol{I}} = \text{var}\,(\widetilde{\boldsymbol{x}}_{\boldsymbol{I}i})$. We can obtain that

$$n^{1/2} \boldsymbol{A}_{n\boldsymbol{I}} \boldsymbol{S}_{n\boldsymbol{I}}^{-1/2} (\gamma \boldsymbol{S}_{n\boldsymbol{I}} + \boldsymbol{\Sigma}_{\lambda_n}) \{(\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}} - \boldsymbol{\beta}_{n\boldsymbol{I}}) + (\gamma \boldsymbol{S}_{n\boldsymbol{I}} + \boldsymbol{\Sigma}_{\lambda_n})^{-1} \boldsymbol{b}_n\} =: \boldsymbol{w}_1 - \boldsymbol{w}_2 + \boldsymbol{w}_3/2,$$

where

$$\begin{aligned} \boldsymbol{w}_1 &=: n^{-1/2} \boldsymbol{A}_{n\boldsymbol{I}} \boldsymbol{S}_{n\boldsymbol{I}}^{-1/2} \sum_{i=1}^{n} \psi(\varepsilon_i) \widehat{\widetilde{\boldsymbol{x}}}_{\boldsymbol{I}i}, \\ \boldsymbol{w}_2 &=: n^{-1/2} \boldsymbol{A}_{n\boldsymbol{I}} \boldsymbol{S}_{n\boldsymbol{I}}^{-1/2} \sum_{i=1}^{n} \{\psi'(\varepsilon_i) - \gamma\} \widehat{\widetilde{\boldsymbol{x}}}_{\boldsymbol{I}i} \widehat{\widetilde{\boldsymbol{x}}}_{\boldsymbol{I}i}^{\top} (\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}} - \boldsymbol{\beta}_{n\boldsymbol{I}}), \\ \boldsymbol{w}_3 &=: n^{-1/2} \boldsymbol{A}_{n\boldsymbol{I}} \boldsymbol{S}_{n\boldsymbol{I}}^{-1/2} \sum_{i=1}^{n} \psi''(\widehat{\widetilde{Y}}_i - \widehat{\widetilde{\boldsymbol{x}}}_i^{\top} \boldsymbol{\beta}_n^*) \{\widehat{\widetilde{\boldsymbol{x}}}_{\boldsymbol{I}i}^{\top} (\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}} - \boldsymbol{\beta}_{n\boldsymbol{I}})\}^2 \widehat{\widetilde{\boldsymbol{x}}}_{\boldsymbol{I}i}. \end{aligned}$$

In the sequel, we will show respectively that $\boldsymbol{w}_1$ satisfies the conditions of Lindeberg-Feller central limit theorem, $\boldsymbol{w}_2 = o_P(1)$ and $\boldsymbol{w}_3 = o_P(1)$.

Using similar arguments to the proof of Lemma 3 in [22], we can show that

$$\left\| \sum_{i=1}^{n} \{\psi'(\varepsilon_i) - \gamma\} \widetilde{\boldsymbol{x}}_{\boldsymbol{I}i} \widetilde{\boldsymbol{x}}_{\boldsymbol{I}i}^{\top} \right\| = o_P(1), \tag{C.6}$$

which, together with the consistency of $\widehat{\boldsymbol{\beta}}_n$ obtained Theorem 4, Condition (C3) and the Cauchy-Schwartz inequality, entails that

$$\|\boldsymbol{w}_2\| \leqslant n^{-1/2} \rho_q^{1/2}(\boldsymbol{A}_{n\boldsymbol{I}} \boldsymbol{A}_{n\boldsymbol{I}}^{\top}) \rho_1^{-1/2}(\boldsymbol{S}_{n\boldsymbol{I}}) o_P(1) \|\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}} - \boldsymbol{\beta}_{n\boldsymbol{I}}\| = O_P(p_n^{1/2}/n) = o_P(1). \tag{C.7}$$

Since $\|\boldsymbol{w}_3\|^2 = \text{tr}(\boldsymbol{w}_3 \boldsymbol{w}_3^{\top})$ and $p_n \log(n)/n \to 0$, we have

$$E\{\|\boldsymbol{w}_3\|^2\} \leqslant \rho_q(\boldsymbol{A}_{n\boldsymbol{I}} \boldsymbol{A}_{n\boldsymbol{I}}^{\top}) B/bE \{\psi''(\varepsilon_i)\}^2 E\{\widetilde{\boldsymbol{x}}_{\boldsymbol{I}i}^{\top} (\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}} - \boldsymbol{\beta}_{n\boldsymbol{I}})\}^4 = O\left(p_n^4/n^2\right) = o(1), \tag{C.8}$$

as long as $p_n^2/n \to 0$. From (C.7)–(C.8), we obtain

$$n^{1/2}\boldsymbol{A}_{n\boldsymbol{I}}\boldsymbol{S}_{n\boldsymbol{I}}^{-1/2}\left(\gamma\boldsymbol{S}_{n\boldsymbol{I}}+\boldsymbol{\Sigma}_{\lambda_n}\right)\{(\widehat{\boldsymbol{\beta}}_{n\boldsymbol{I}}-\boldsymbol{\beta}_{n\boldsymbol{I}})+\left(\gamma\boldsymbol{S}_{n\boldsymbol{I}}+\boldsymbol{\Sigma}_{\lambda_n}\right)^{-1}\boldsymbol{b}_n\}=\boldsymbol{w}_1+o_P(1). \tag{C.9}$$

Next, we verify that $\boldsymbol{w}_1$ satisfies the conditions of the Lindeberg-Feller central limit theorem. Let $\boldsymbol{\omega}_{ni}=n^{-1/2}\boldsymbol{A}_{n\boldsymbol{I}}\boldsymbol{S}_{n\boldsymbol{I}}^{-1/2}\psi(\varepsilon_i)\widetilde{\boldsymbol{x}}_{\boldsymbol{I}i}, i=1,\dots,n$. We note that $E(\boldsymbol{\omega}_{ni})=\boldsymbol{0}$ and

$$\operatorname{var}\left(\sum_{i=1}^n\boldsymbol{\omega}_{ni}\right)=E\{\psi^2\left(\varepsilon_i\right)\}\boldsymbol{A}_{n\boldsymbol{I}}\boldsymbol{S}_{n\boldsymbol{I}}^{-1/2}E\left(\frac{1}{n}\sum_{i=1}^n\widetilde{\boldsymbol{x}}_{\boldsymbol{I}i}\widetilde{\boldsymbol{x}}_{\boldsymbol{I}i}^\top\right)\boldsymbol{S}_{n\boldsymbol{I}}^{-1/2}\boldsymbol{A}_{n\boldsymbol{I}}^\top\to\sigma_0^2\boldsymbol{G}, \tag{C.10}$$

since $\boldsymbol{A}_{n\boldsymbol{I}}\boldsymbol{A}_{n\boldsymbol{I}}^\top\to\boldsymbol{G}$. For any $\varepsilon>0$, it follows that

$$\sum_{i=1}^n E\{\|\boldsymbol{\omega}_{ni}\|^2\mathbf{1}\left(\|\boldsymbol{\omega}_{ni}\|\geqslant\varepsilon\right)\}=nE\{\|\boldsymbol{\omega}_{ni}\|^2\mathbf{1}\left(\|\boldsymbol{\omega}_{ni}\|\geqslant\varepsilon\right)\}\leqslant nE(\|\boldsymbol{\omega}_{ni}\|^4)^{1/2}\left\{P\left(\|\boldsymbol{\omega}_{ni}\|\geqslant\varepsilon\right)\right\}^{1/2}. \tag{C.11}$$

By Condition (C3), $\boldsymbol{A}_{n\boldsymbol{I}}\boldsymbol{A}_{n\boldsymbol{I}}^\top\to\boldsymbol{G}$ and $\sigma_0^2=E\{\psi^2(\varepsilon_i)\}$, we have

$$P(\|\boldsymbol{\omega}_{ni}\|>\varepsilon)\leqslant\frac{E\|\boldsymbol{\omega}_{ni}\|^2}{\varepsilon^2}\leqslant\frac{\sigma_0^2\rho_q(\boldsymbol{A}_{n\boldsymbol{I}}\boldsymbol{A}_{n\boldsymbol{I}}^\top)E(\widetilde{\boldsymbol{x}}_{\boldsymbol{I}i}^\top\boldsymbol{S}_{n\boldsymbol{I}}^{-1}\widetilde{\boldsymbol{x}}_{\boldsymbol{I}i})}{n\varepsilon^2}=O(n^{-1}). \tag{C.12}$$

Let $\sigma_4=E\{\psi^4\left(\varepsilon_i\right)\}$. Similar to Theorem 6 in [16], we have

$$E(\|\boldsymbol{\omega}_{ni}\|^4)\leqslant\frac{\sigma_4}{n^2}\rho_q^2(\boldsymbol{A}_{n\boldsymbol{I}}\boldsymbol{A}_{n\boldsymbol{I}}^\top)\rho_1^{-2}(\boldsymbol{S}_{n\boldsymbol{I}})E\{(\widetilde{\boldsymbol{x}}_{\boldsymbol{I}}^\top\widetilde{\boldsymbol{x}}_{\boldsymbol{I}})^2\}=O(p_n^2/n^2). \tag{C.13}$$

Thus, by (C.10)–(C.13), we have

$$\sum_{i=1}^n E\{\|\boldsymbol{\omega}_{ni}\|^2\mathbf{1}(\|\boldsymbol{\omega}_{ni}\|\geqslant\varepsilon)\}=O\left(n\frac{p_n}{n}\frac{1}{n^{1/2}}\right)=o(1).$$

Combining the above arguments, and invoking the Lindeberg-Feller central limit theorem, we complete the proof of the asymptotic normality part in Theorem 5. □

The proof of Theorem 3 is almost identical to that of Theorem 6. In the sequel, we will only prove Theorem 6.

Appendix D Proof of Theorem 6

For notational clarity, we let $K_i=K(\frac{\boldsymbol{z}_i-\boldsymbol{z}}{h_n})$. Recall that $\widehat{\nu}(\boldsymbol{z})=\widehat{a}_n$ and $(\widehat{a}_n,\widehat{\boldsymbol{b}}_n)$ minimizes

$$\sum\rho\{Y_i-\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_n-a_n-\boldsymbol{b}_n^\top(\boldsymbol{z}_i-\boldsymbol{z})\}K_i.$$

Let $\boldsymbol{\theta}_n=(nh_n^d)^{1/2}[\widehat{a}_n-\nu(\boldsymbol{z}),h_n\{\widehat{\boldsymbol{b}}_n-\nu'(\boldsymbol{z})\}]$, $s_i=\boldsymbol{x}_i^\top(\boldsymbol{\beta}_{n0}-\widehat{\boldsymbol{\beta}}_n)$, $\boldsymbol{z}_i^*=\{1,(\boldsymbol{z}_i-\boldsymbol{z})^\top/h_n\}^\top$, and $\delta_i=Y_i-\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_n-\nu(\boldsymbol{z})-\nu'(\boldsymbol{z})(\boldsymbol{z}_i-\boldsymbol{z})$. Following similar arguments for proving Theorem 4, we obtain that

$$\begin{aligned}D_n&=\sum_{i=1}^n[\rho\{Y_i-\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_n-a_n-b_n(\boldsymbol{z}_i-\boldsymbol{z})\}-\rho(\delta_i)]K_i\\&=\sum_{i=1}^n\left[\int_0^{(nh_n^d)^{-1/2}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)}\{\psi(\delta_i+t)-\psi(\delta_i)\}dt+(nh_n^d)^{-1/2}\psi(\delta_i)(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)\right]K_i.\end{aligned} \tag{D.1}$$

Given $\boldsymbol{X}=(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)^\top$ and $\boldsymbol{Z}=(\boldsymbol{z}_1,\dots,\boldsymbol{z}_n)^\top$, we can obtain that

$$\begin{aligned}E(D_n|\boldsymbol{X},\boldsymbol{Z})=&(nh_n^d)^{-1}\sum_{i=1}^n(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)^2K_i\{1+o_p(1)\}/2\\&+(nh_n^d)^{-1/2}\sum_{i=1}^n G\{\nu(\boldsymbol{z}_i)-\nu(\boldsymbol{z})-\nu'(\boldsymbol{z})(\boldsymbol{z}_i-\boldsymbol{z})+s\}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i,\end{aligned}$$

where $G(t)=E\{\psi(\varepsilon+t)\}$ is defined in Condition (C2). The above display follows from Condition (C2) which assumes that there exists a constant γ such that $G(t)=:E\{\psi(\varepsilon+t)\}=\gamma t+o(|t|)$. This also implies that

$$\begin{aligned}(nh_n^d)^{-1/2}&\sum_{i=1}^n G\{\nu(\boldsymbol{z}_i)-\nu(\boldsymbol{z})-\nu'(\boldsymbol{z})(\boldsymbol{z}_i-\boldsymbol{z})+s_i\}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i\\&-(nh_n^d)^{-1/2}\sum_{i=1}^n G\{\nu(\boldsymbol{z}_i)-\nu(\boldsymbol{z})-\nu'(\boldsymbol{z})(\boldsymbol{z}_i-\boldsymbol{z})\}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i\\&=\gamma(\boldsymbol{\beta}_{n0}-\widehat{\boldsymbol{\beta}}_n)^\top(nh_n^d)^{-1/2}\sum_{i=1}^n\boldsymbol{x}_i(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i\\&=\gamma(nh_n^d)^{-1/2}\sum_{i=1}^n\{(\boldsymbol{\beta}_{n\widehat{\boldsymbol{I}}0}-\widehat{\boldsymbol{\beta}}_{n\widehat{\boldsymbol{I}}})^\top\boldsymbol{x}_{\widehat{\boldsymbol{I}}i}-\boldsymbol{\beta}_{n\widehat{\boldsymbol{II}}}^\top\boldsymbol{x}_{\widehat{\boldsymbol{II}}i}\}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i,\end{aligned}$$

where $\widehat{\boldsymbol{I}}$ and $\widehat{\boldsymbol{II}}$ denote respectively the sets of estimated indices of important and unimportant predictors. Next, we show that

$$(nh_n^d)^{-1/2}\sum_{i=1}^n(\boldsymbol{\beta}_{n\widehat{\boldsymbol{I}}0}-\widehat{\boldsymbol{\beta}}_{n\widehat{\boldsymbol{I}}})^\top\boldsymbol{x}_{\widehat{\boldsymbol{I}}i}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i=o_p\{(nh_n^d)^{1/2}\}.\tag{D.2}$$

For positive constant c,

$$\begin{aligned}\Pr&\bigg\{(nh_n^d)^{-1}\sum_{i=1}^n(\boldsymbol{\beta}_{n\widehat{\boldsymbol{I}}0}-\widehat{\boldsymbol{\beta}}_{n\widehat{\boldsymbol{I}}})^\top\boldsymbol{x}_{\widehat{\boldsymbol{I}}i}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i\geqslant C\bigg\}\\&=\Pr\bigg\{(nh_n^d)^{-1}\sum_{i=1}^n(\boldsymbol{\beta}_{n\widehat{\boldsymbol{I}}0}-\widehat{\boldsymbol{\beta}}_{n\widehat{\boldsymbol{I}}})^\top\boldsymbol{x}_{\widehat{\boldsymbol{I}}i}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i\geqslant C\,\bigg|\,\widehat{\boldsymbol{I}}=\boldsymbol{I}\bigg\}\Pr(\widehat{\boldsymbol{I}}=\boldsymbol{I})\\&\quad+\Pr\bigg\{(nh_n^d)^{-1}\sum_{i=1}^n(\boldsymbol{\beta}_{n\widehat{\boldsymbol{I}}0}-\widehat{\boldsymbol{\beta}}_{n\widehat{\boldsymbol{I}}})^\top\boldsymbol{x}_{\widehat{\boldsymbol{I}}i}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i\geqslant C\,\bigg|\,\widehat{\boldsymbol{I}}\neq\boldsymbol{I}\bigg\}\Pr(\widehat{\boldsymbol{I}}\neq\boldsymbol{I}).\end{aligned}$$

It is easy to show that $(nh_n^d)^{-3/2}\sum_{i=1}^n\boldsymbol{x}_{\boldsymbol{I}i}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i$ converges in probability to a constant vector $\boldsymbol{A}_{n\boldsymbol{I}}$, which together with the asymptotic normality of $\boldsymbol{\beta}_n$ in Theorem 5 entails that $(nh_n^d)^{-1}\sum_{i=1}^n(\boldsymbol{\beta}_{n\widehat{\boldsymbol{I}}0}-\boldsymbol{\beta}_{n\widehat{\boldsymbol{I}}})^\top\boldsymbol{x}_{\widehat{\boldsymbol{I}}i}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i=O_p\{(nh_n^d)^{1/2}\}O_p(n^{-1/2})=o_p(1)$. Therefore, the first term in the above display converges to zero as $n\to\infty$, which together with that $\Pr(\widehat{\boldsymbol{I}}\neq\boldsymbol{I})\to0$ [29] proves (D.2). Similarly, we can show that

$$(nh_n^d)^{-1/2}\sum_{i=1}^n\boldsymbol{\beta}_{n\widehat{\boldsymbol{II}}}^\top\boldsymbol{x}_{\widehat{\boldsymbol{II}}i}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i=o_p\{(nh_n^d)^{1/2}\}.\tag{D.3}$$

The results of (D.2) and (D.3) yield that $E(D_n|\boldsymbol{X},\boldsymbol{Z})$ can be simplified as follows,

$$\begin{aligned}E(D_n|\boldsymbol{X},\boldsymbol{Z})&=(nh_n^d)^{-1}\sum_{i=1}^n(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)^2K_i\{1+o_p(1)\}/2\\&\quad+(nh_n^d)^{-1/2}\sum_{i=1}^n G\{\nu(\boldsymbol{z}_i)-\nu(\boldsymbol{z})-(\boldsymbol{z}_i-\boldsymbol{z})^\top\nu'(\boldsymbol{z})\}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i+o_p\{(nh_n^d)^{1/2}\}.\end{aligned}$$

Similar to the proof of Theorem 4, we can obtain that

$$\begin{aligned}D_n&=(nh_n^d)^{-1}\sum_{i=1}^n(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)^2K_i/2\\&\quad+(nh_n^d)^{-1/2}\sum_{i=1}^n G\{\nu(\boldsymbol{z}_i)-\nu(\boldsymbol{z})-(\boldsymbol{z}_i-\boldsymbol{z})^\top\nu'(\boldsymbol{z})\}(\boldsymbol{\theta}_n^\top\boldsymbol{z}_i^*)K_i+o_p\{(nh_n^d)^{1/2}\}.\end{aligned}$$

The rest of the proof follows literally from [8] by recalling that we treat the dimension of $\boldsymbol{z}$ as fixed. Thus we omit the details. □