Robust variable selection for nonlinear models with diverging number of parameters

Zhike Lv[a,*], Huiming Zhu[a], Keming Yu[b]

[a] College of Business Administration, Hunan University, Changsha, 410082, PR China

[b] Department of Mathematical Sciences, Brunel University, London UB8 3PH, UK

Abstract

We focus on the problem of simultaneous variable selection and estimation for nonlinear models based on modal regression (MR), when the number of coefficients diverges with sample size. With appropriate selection of the tuning parameters, the resulting estimator is shown to be consistent and to enjoy the oracle properties.

Keywords: Variable selection; Asymptotic normality; Oracle properties; Nonlinear models; Modal regression

1. Introduction

Mode, the most likely value of a distribution, has wide applications in biology, astronomy, economics and finance. For example, it is not uncommon in many fields to encounter data distributions that are skewed or have outliers. In those cases, the mean may not be an appropriate statistic to represent the center of location of the data. Alternative statistics with less bias are the median and the mode. The mean or median of two densities may be identical, while the shapes of the two densities are quite different. Mode preserves some of the important features, such as wiggles, of the underlying distribution function, whereas the mean or median tend to average out the data. In fact, as an important statistic, mode has been used in modern science to identify the most frequent or the most typical element in certain network systems (see, Hedges and Shah (2003)). Because of its advantages and wide applications, mode estimation has gained much attention in the statistics literature (e.g., Berlinet et al. (1998), Meyer (2001), Kemp and Santos Silva (2012)).

More recently, Yao and Li (2013) proposed a new regression model called modal linear regression (MODLR) that assumes that the mode of $f(y|x)$ is a linear function of the predictor x. A distinguishing characteristic of this method is that it introduces an additional tuning parameter which is automatically selected using the observed data to achieve both robustness and efficiency of the resulting estimate. Namely, their method is not only robust when there are outliers or the error distribution is heavy-tail, but as

*Corresponding author. Tel.: +86 731 88684461.
E-mail address: lzk2011@hnu.edu.cn(Z. Lv), zhuhuiming@hnu.edu.cn(H. Zhu).

asymptotically efficient as the ordinary least-square-based estimator when the data include no outliers and the error distribution is a Gaussian distribution. Then, Yao, Lindsay, and Li (2012) extended this new approach to the classical nonparametric model. Subsequently, Zhang et al. (2013) and Zhao et al. (2014) considered the semiparametric partially linear varying coefficient models based on the modal regression, they also develop a variable selection procedure to select significant parametric components for it. And Liu et al. (2013) studied the single-index model based on the local modal regression. Due to its nice properties, in this paper, we focus on the following nonlinear model

$$y_i = g(x_i; \beta) + \varepsilon_i, \tag{1.1}$$

where $g(\cdot;\cdot)$ is a known real-valued function, $\beta = (\beta_1, ..., \beta_p)^T$ is a p-dimensional unknown parameter vector and ε_i is random error with mean zero. The model (1.1) is a very flexible model, which contains many submodels of which linear models and generalized linear models with continuous responses are specific examples.

Variable selection is important for any regression problem in that ignoring important variables brings out seriously biased results, whereas including spurious variables leads to substantial loss in estimation efficiency. Traditional variable selection methods such as stepwise regression and best subset selection is computationally infeasible when the number of predictors is large. Thus, various shrinkage methods such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), the adaptive LASSO (Zou 2006) and the smoothly clipped absolute deviation (SCAD) (Fan & Li 2001) have gained much attention in recent years. However, the LASSO is known to be near mini-max optimal as well as consistent under certain regularity conditions, Zou (2006) showed that it falls short of attaining the oracle property. By this property, an estimator estimates a zero coefficient exactly as zero with probability approaching one, while still being asymptotically normal for the non-zero coefficients in large samples. In this respect, the LASSO is inferior to the SCAD estimator which possesses the oracle property. So in the present paper, we prefer the SCAD of Fan and Li (2001) since it enjoys the oracle properties. Previous research mainly focus on situations with fixed p. However, Fan and Peng (2004) and Lam and Fan (2008) advocated that, in most variable selection problems, the number of parameters should be large and grow with the sample size. Hence, in this paper, we study variable selection for the linear part in nonlinear model when the number of parameters p depends on the sample size n, then model (1.1) can be rewritten as

$$y_i = g(x_i; \beta_n) + \varepsilon_i, \tag{1.2}$$

where $\beta_n = (\beta_{n1}, ..., \beta_{np_n})^T$ is a p_n-dimensional unknown parameter vector, and ε_i is the random error with $E(\varepsilon_i|x_i) = 0$.

The main contributions of this paper are twofold. First, we propose a variable selection procedure for model (1.2) based on modal regression, when the number of parameters p depends on the sample size n. With proper choice of tuning parameters, we show that this variable selection procedure is consistent, and the estimators of regression coefficients have oracle property. Here, the oracle property means that the estimators of the parametric components have the same asymptotic distribution as that based on the

correct submodel. This indicates that the penalized estimators work as well as if the subset of true zero coefficients were already known. Second, we propose a modified modal expectation-maximisation (MEM) type algorithm to obtain the solutions for the target function.

The rest of the paper is organized as follows. In Section 2, we present the proposed variable selection method and study the asymptotic properties of the estimators. In Section 3, we describe the MEM type algorithm. We assess the finite sample performance of the proposed method through a simulation study in Section 4. We give concluding remarks in Section 5, and relegate the technical proofs to Appendix.

2. Methodology and asymptotic properties

2.1. Modal estimation and variable selection procedure

For the linear model $y_i = x_i^T\beta + \varepsilon_i$, Yao and Li (2013) proposed to estimate the modal regression parameter β by maximising

$$\frac{1}{n}\sum_{i=1}^{n}\phi_h(y_i - x_i^T\beta), \tag{2.1}$$

where $\phi_h(t) = h^{-1}\phi(t/h)$ and $\phi(t)$ is a kernel density function. Throughout this paper, we will assume that $\phi(t)$ is the standard normal density (for the simplicity of computation). Thus, based on the idea in Yao and Li (2013), the robust modal estimator β_n of model (1.2) is to maximise

$$\frac{1}{n}\sum_{i=1}^{n}\phi_h(y_i - g(x_i;\beta_n)), \tag{2.2}$$

over β_n.

It is well known that variable selection is a crucial step in high-dimensional regression modeling. However, (2.2) cannot directly be used to select variables, we introduce the following penalized estimation by maximising

$$Q_n(\beta_n) = \sum_{i=1}^{n}\phi_h(y_i - g(x_i;\beta_n)) - n\sum_{j=1}^{p_n}p_{\lambda_n}(|\beta_{nj}|), \tag{2.3}$$

where $p_{\lambda_n}(\cdot)$ is a penalty function and λ_n is a non-negative regularization parameter.

Remark 1. Notice that our method is in fact also a M-type estimator (The bandwidth h determines the degree of robustness of the estimator) and its first derivative of $\phi_h(\cdot)$ is bounded. This explains why the proposed method is robust. Formulation (2.3) includes many popular variable selection methods, for example, the Lasso (Tibshirani 1996) uses the L_1 penalty with $p_{\lambda_n}(\|\cdot\|) = \lambda_n\|\cdot\|$. Bridge regression (Frank and Friedman 1993) uses the L_q penalty with $p_{\lambda_n}(\|\cdot\|) = \lambda_n\|\cdot\|^q$. When $0 < q < 1$ the L_q penalty is concave over $(0,\infty)$ and nondifferentiable at zero. Fan and Li (2001) proposed the use of the SCAD penalty defined by its first derivative as

$$p'_\lambda(x) = \lambda\Big\{I(x \le \lambda) + \frac{(a\lambda - x)_+}{(a-\lambda)}I(x > \lambda)\Big\},$$

where a is some constant usually taken to be $a = 3.7$ and $p_\lambda(0) = 0$. As demonstrated in Fan and Li (2001), the SCAD is an improvement of the Lasso in terms of modeling bias and of the bridge regression with $q < 1$ in terms of stability. Therefore, we take $p_{\lambda_n}(\cdot)$ as the SCAD penalty function throughout this paper. The adaptive Lasso (Zou 2006) can also be used here and are expected to lead to similar consistency results, but this need further research.

2.2. Asymptotic properties

Define $\mathcal{A} = \{j : \beta_{nj} \neq 0\}$. Then without loss of generality, let the true value of β_n be $\beta_n^* = (\beta_{n\mathcal{A}}^*, \beta_{n\mathcal{A}^c}^*)^T$, where $\beta_{n\mathcal{A}}^* \in \mathbb{R}^{s_n}$ consists of all nonzero components with s_n means the number of nonzero components, while $\beta_{n\mathcal{A}^c}^* \in \mathbb{R}^{p_n - s_n}$ consists of all zero components.

Theorem 1 *Suppose that conditions (C1)-(C8) given in the appendix hold. If $p_n^3/n \to 0$ as $n \to \infty$, then there is a local maximizer $\hat{\beta}_n$ of $Q_n(\beta_n)$ in (2.3), such that*

$$\|\hat{\beta}_n - \beta_n^*\| = O_p(\sqrt{p_n/n}). \tag{2.4}$$

To present the oracle properties of the resulting estimators, we require further notations. Let $F(x,h) = \mathrm{E}[\phi_h''(\varepsilon)|x]$, $G(x,h) = \mathrm{E}[\phi_h'(\varepsilon)^2|x]$, and

$$\boldsymbol{b}_n = \{p'_{\lambda_n}(|\beta_{n1}^*|)\mathrm{sgn}(\beta_{n1}^*), ..., p'_{\lambda_n}(|\beta_{ns_n}^*|)\mathrm{sgn}(\beta_{ns_n}^*)\}^T, \quad \Sigma_{\lambda_n} = \mathrm{diag}\{p''_{\lambda_n}(\beta_{n1}^*), ..., p''_{\lambda_n}(\beta_{ns_n}^*)\}$$

and

$$\Xi = -F(x,h)\mathrm{E}[g'(x_i;\beta_n^*)g'(x_i;\beta_n^*)^T|x], \quad \Omega = G(x,h)\mathrm{E}[g'(x_i;\beta_n^*)g'(x_i;\beta_n^*)^T|x].$$

where $g'(;)$ is a $p_n \times 1$ vector.

Theorem 2 (Oracle property) *Suppose that conditions (C1)-(C8) given in the appendix hold. If $\lambda_n \to 0$, $\sqrt{n/p_n}\lambda_n \to \infty$ and $p_n^3/n \to 0$ as $n \to \infty$ with probability tending to 1, the $\sqrt{n/p_n}$ consistent local maximizer $\hat{\beta}_n = (\hat{\beta}_{n\mathcal{A}}^T, \hat{\beta}_{n\mathcal{A}^c}^T)^T$ in Theorem 1 satisfy:*

(a) *Sparsity* : $\hat{\beta}_{n\mathcal{A}^c} = 0$, (2.5)

(b) *Asymptotic normality* : $\sqrt{n}(\Xi_{\mathcal{A}} + \Sigma_{\lambda_n})\{\hat{\beta}_{n\mathcal{A}} - \beta_{n\mathcal{A}}^* + (\Xi_{\mathcal{A}} + \Sigma_{\lambda_n})^{-1}\boldsymbol{b}_n\} \xrightarrow{d} N(0, \Omega_{\mathcal{A}})$, (2.6)

where $\Xi_{\mathcal{A}}$ and $\Omega_{\mathcal{A}}$ consist of the first s_n rows and columns of Ξ and Ω.

Remark 2. Theorems 1 and 2 indicate that the penalized estimators have the oracle property. That is, the estimators of the parametric components have the same asymptotic distribution as that based on the correct submodel.

In the process of variable selection, the bandwidth h and the tuning parameters λ_n should be determined. First, we give the optimal bandwidth in theoretical. When n is large enough, and the regularity condition (C8) in the Appendix is satisfied by the model, we have $\Sigma_{\lambda_n} = 0$ and $\boldsymbol{b}_n = 0$ for the SCAD penalty. Thus, (2.6) becomes

$$\sqrt{n}(\hat{\beta}_{n\mathcal{A}} - \beta_{n\mathcal{A}}^*) \xrightarrow{d} N(0, \Xi_{\mathcal{A}}^{-1}\Omega_{\mathcal{A}}\Xi_{\mathcal{A}}^{-1}), \tag{2.7}$$

and it is easy to obtain the asymptotic variance of the estimator under the least squares loss of $\hat{\beta}_{n\mathcal{A}}$ is

$$\mathrm{Var}(\varepsilon|x)/\mathrm{E}[g'_{\mathcal{A}}(x_i;\beta_n^*)g'_{\mathcal{A}}(x_i;\beta_n^*)^T|x], \tag{2.8}$$

where $g'_{\mathcal{A}}(;)$ means the first derivative of $g_{\mathcal{A}}(;)$ with respect to β_n^*, and $g_{\mathcal{A}}(;)$ consist of the first s_n columns of $g(;)$. Then, based on Yao et al. (2012), the ratio of asymptotic variance of (2.7) to that of (2.8) is given by

$$R(x,h) = \frac{G(x,h)}{F^2(x,h)\mathrm{Var}(\varepsilon|x)}. \tag{2.9}$$

Thus, the ideal choice of h is

$$h_{opt} = \arg\min_h G(x,h)/F^2(x,h). \tag{2.10}$$

Next, we consider the selection of λ_n, various techniques can be used to select λ_n, such as cross-validation, AIC and BIC. To reduce intensive computation and guarantee consistent variable selection, we consider the regularization parameter by minimizing a BIC-type objective function (see Wang, Li, and Jiang 2007). That is, the optimal λ_n minimizes

$$BIC(\lambda_n) = -\frac{1}{n}\sum_{i=1}^{n}\phi_h(y - g(x_i;\hat{\beta}_n)) + \frac{\log(n)}{n}df(\lambda_n), \tag{2.11}$$

where $df(\lambda_n)$ is the total number of nonzero coefficients in $\hat{\beta}_{\lambda_n}$. For details, we refer the reader to Zou et al. (2007).

3. A modified modal expectation-maximization algorithm

In this section, we extend the MEM algorithm, proposed by Li et al. (2007) to maximise (2.3). Since the SCAD penalty is irregular at the origin, maximising (2.3) directly may be difficult. Here, we use an iterative algorithm based on the local quadratic approximation of the penalty function $p_\lambda(\cdot)$ as in Fan and Li (2001). More specifically, Suppose that we are given an initial value $\beta_n^{(0)}$ that is close to the minimizer of (2.3). If $\beta_{nj}^{(0)}$ is very close to 0, then set $\hat{\beta}_{nj} = 0$. Otherwise they can be locally approximated by a quadratic function as

$$p_\lambda(|\beta_{nj}|) \approx p_\lambda(|\beta_{nj}^{(0)}|) + \frac{1}{2}\frac{p'_\lambda(|\beta_{nj}^{(0)}|)}{|\beta_{nj}^{(0)}|}(\beta_{nj}^2 - \beta_{nj}^{(0)2}), \text{ for } \beta_{nj} \approx \beta_{nj}^{(0)}.$$

Then, we can use the modified EM algorithm to maximise (2.3). Let $\beta_n^{(0)}$ be the initial estimation and start with $k = 0$.

E-step: In this step, we calculate weights $\pi(j|\beta_n^{(k)})$, $j = 1, ..., n$ as

$$\pi(j|\beta_n^{(k)}) = \frac{\phi_h(y_i - g(x_i;\beta_n^{(k)}))}{\sum_{i=1}^{n}\phi_h(y_i - g(x_i;\beta_n^{(k)}))} \propto \phi_h(y_i - g(x_i;\beta_n^{(k)})).$$

M-step: Then, we update $\beta_n^{(k+1)}$ by

$$\begin{aligned}\beta_n^{(k+1)} &= \arg\max_{\beta_n}\Big(\sum_{i=1}^{n}\{\pi(j|\beta_n^{(k)})\log\phi_h(y_i - g(x_i;\beta_n^{(k+1)}))\} - n\sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|)\\ &\approx \arg\max_{\beta_n}\Big(\sum_{i=1}^{n}\{\pi(j|\beta_n^{(k)})\log\phi_h(y_{i,k} - g'(x_i;\beta_n^{(k)})\beta_n)\} - n\sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|)\\ &= (G^TWG + n\Sigma_\lambda(\beta_n^{(k)}))^{-1}G^TW\widetilde{Y},\end{aligned}$$

where $\widetilde{Y} = (y_{1,k}, ..., y_{n,k})^T$ with $y_{i,k} = y_i - g(x_i, \beta_n^{(k)}) + g'(x_i;\beta^{(k)})\beta^{(k)}$, $G = (g'(x_1;\beta_n^{(k)}), ..., g'(x_n;\beta_n^{(k)}))^T$, W is an $n \times n$ diagonal matrix with diagonal elements $\pi(j|\beta_n^{(k)})$s and

$$\Sigma_\lambda(\beta_n^{(k)})) = \mathrm{diag}\left\{\frac{p'_{\lambda_n}(|\beta_{n1}^{(k)}|)}{|\beta_{n1}^{(k)}|}, ..., \frac{p'_{\lambda_n}(|\beta_{np_n}^{(k)}|)}{|\beta_{np_n}^{(k)}|}\right\}.$$

Iterate the E-step and M-step until convergence. Note that in the M-step, we approximate $g(x_i;\beta_n^{(k+1)})$ in the neighborhood of $\beta_n^{(k)}$ by using first order approximation of Taylor expansion, that is

$$g(x_i;\beta_n^{(k+1)}) \approx g(x_i;\beta_n^k) + g'(x_i;\beta_n^k)(\beta_n^{(k+1)} - \beta_n^k).$$

4. Simulation studies

In this section, we first consider how to select the bandwidth h in practice, and then assess the performance of the proposed procedure by some simulation studies.

4.1. Bandwidth selection in practice

In this subsection, we present the details of bandwidth selection in our simulation studies. In our simulation setting, we assume the error ε and x are independent. Thus, we first need to estimate $F(h)$ and $G(h)$ to obtain the optimal bandwidth h_{opt} based on (2.10). Then, $F(h)$ and $G(h)$ can be estimated by

$$\hat{F}(h) = \frac{1}{n}\sum_{i=1}^{n}\phi_h''(\hat{\varepsilon}_i) \quad \text{and} \quad \hat{G}(h) = \frac{1}{n}\sum_{i=1}^{n}[\phi_h'(\hat{\varepsilon}_i)]^2, \tag{4.1}$$

where $\hat{\varepsilon} = y_i - g(x_i;\hat{\beta}_n)$, $\hat{\beta}_n$ is the traditional penalized least squares estimate (or a robust estimate if there are some outliers) of β_n. Therefore, we can estimate $R(h)$ by $\hat{R}(h) = \hat{G}(h)/\hat{F}^2(h)\widehat{\mathrm{Var}}(\varepsilon|x)$, where $\widehat{\mathrm{Var}}(\varepsilon|x)$ is estimated based on the pilot estimates, $\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n$ of the error term. Since there is no explicit solution for h, thus, we use the grid search method to obtain the optimal bandwidth h_{opt}. Yao et al. (2012) showed that the possible grids points for h can be $h = 0.5\widehat{\mathrm{Var}}(\varepsilon|x) \times 1.02^j$, $j = 0, 1, ..., k$, for some fixed k (such as k=70).

4.2. Simulation study

In this section, we conduct a Monte Carlo simulation study to assess the performance of our proposed approach under a finite sample size scenario. We generated independently and identically distributed sample $\{(y_i, x_i), i = 1, ..., n\}$ from the following exponential regression model

$$y_i = \frac{1}{2} + \exp(x_i^T \beta_n) + \varepsilon_i, \tag{4.2}$$

where $\beta_n = (1, 2/3, 0.5, ..., 0)^T$, and $x_i \sim N(0, 0.5 I_{p_n})$. Noise ε_i were generated from three different distributions: the standard normal, the mixture normal $0.9N(0, 1) + 0.1N(0, 9^2)$ and the standard t with three degrees of freedom.

In the simulations, we draw 1000 random samples of sizes 100 and 400 with $p_n = [2n^{1/4}] + 3$ from model (4.2), then the corresponding dimensions of the parameter vector β_n are 9 and 12, respectively. Furthermore, the selection of h and λ_n are based on Equations (2.10) and (2.11), respectively. In each simulation the "root of mean squared errors (RMSE)" for their average over simulations is reported in Tables 1. To examine the robustness and efficiency of the proposed procedure, we compare the simulation results with the penalized least-squares (PLS) estimator (Fan and Peng 2004) and the weighted composite quantile regression (WCQR) method (Jiang et al. 2012). The average number of zero coefficients is also reported in Table 1, Column "C" shows the average number of zero coefficients correctly estimated to be zero, and Column "IC" presents the average number of non-zero coefficients incorrectly estimated to be zero.

Based on Table 1, we can see that, as expected, the performance of Oracle procedure is best in all cases in term of model errors. The performances of MR is slight better than that of WCQR, and significantly better than that of PLS when the error distribution is non-normal. Especially, when the error follows a mixture normal, the superiority of MR become more and more obvious.

5. Conclusions

In this paper, we have proposed a variable selection method in nonlinear models based on modal regression, where the number of coefficients can diverges with sample size n. This approach is used to simultaneously estimate parameters and select important variables. Theoretically, we showed that our proposed method estimators enjoy the oracle properties, which is desirable as a variable selection procedure. And from a practical point of view, we illustrated through a simulation study when the error distribution are generated from three different distributions, the performances of the proposed method outperform the PLS and WCQR in terms of the consistency of the variable selection method and the efficiency of the estimation procedure.

Acknowledgements

We are very grateful to the two anonymous reviewers, the associate editor, and the editor for their constructive comments that greatly improved presentation of our work. This work was supported by the National Natural Science Foundation of China under grants 71171075, 71221001, 71031004 and 11261048.

Table1 :Simulation results

Error distribution	Method	(n, p_n)	RMSE	Aver. No. of zeros	
				C	IC
N(0, 1)	MR	(100, 9)	0.0441	5.778	0
	PLS		0.0355	5.815	0.001
	WCQR		0.0372	5.793	0
	MR oracle		0.0292	6	0
	MR	(400, 12)	0.0332	8.865	0
	PLS		0.0313	8.891	0
	WCQR		0.0305	8.869	0
	MR oracle		0.0277	9	0
t(3)	MR	(100, 9)	0.0343	5.833	0
	PLS		0.0560	5.690	0.005
	WCQR		0.0333	5.827	0
	MR oracle		0.0318	6	0
	MR	(400, 12)	0.0271	8.872	0
	PLS		0.0515	8.725	0.002
	WCQR		0.0296	8.835	0
	MR oracle		0.0240	9	0
0.9N(0, 1)+0.1N(0, 9^2)	MR	(100, 9)	0.0415	5.889	0
	PLS		0.0950	4.684	0.015
	WCQR		0.0533	5.826	0
	MR oracle		0.0383	6	0
	MR	(400, 12)	0.0378	8.902	0
	PLS		0.0918	7.695	0.006
	WCQR		0.0485	8.873	0
	MR oracle		0.0350	9	0

Appendix: Proofs

For simplicity, let C denote a positive constant that may be different at each appearance throughout this paper, and define $a_n = \max_{1\le j\le p_n}\{p'_{\lambda_n}(|\beta^*_{nj}|), \beta^*_{nj} \neq 0\}$ and $b_n = \max_{1\le j\le p_n}\{p''_{\lambda_n}(|\beta^*_{nj}|), \beta^*_{nj} \neq 0\}$. Before we prove our main theorems, we list some regularity conditions that are used in this paper.

(C1) $\liminf_{n\to\infty} \liminf_{\theta\to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0$.

(C2) $a_n = O_p(n^{-1/2})$, and $b_n \to 0$ as $n \to \infty$.

(C3) There are constants C_1 and C_2 such that, when θ_1, $\theta_2 > C_1\lambda_n$, then $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \le C_2|\theta_1 - \theta_2|$.

(C4) $F(x, h)$ and $G(x, h)$ are continuous with respect to u. Furthermore, $F(x, h) < 0$ for any $h > 0$, where the bandwidth h is a constant and does not depend on n.

(C5) $\mathrm{E}(\phi_h'(\varepsilon)|x) = 0$. $\mathrm{E}(\phi_h''(\varepsilon)^2|x)$, $\mathrm{E}(\phi_h'(\varepsilon)^3|x)$, and $\mathrm{E}(\phi_h'''(\varepsilon)|x)$ are continuous with respect to x.
(C6) There is a large enough open subset $\Theta_n \in \mathbb{R}^{p_n}$ that contains the true parameter point β_n^*, such that for all x_i the second derivative matrix $g''(x;\beta_n)$ of $g(x;\beta_n)$ with respect to β_n, satisfies

$$\|g''(x;\beta_{n1}) - g''(x;\beta_{n2})\| \le M(x_i)\|\beta_{n1} - \beta_{n2}\| \text{ and } \Big|\frac{\partial g(x;\beta_n)}{\partial \beta_{nj}\beta_{nk}}\Big| \le N_{jk}(x_i)$$

for all $\beta_n \in \Theta_n$, with $\mathrm{E}[\mathrm{M}^2(\mathrm{x_i})] < \infty$, $\mathrm{E}[\mathrm{N}_{\mathrm{jk}}^2(\mathrm{x_i})] < \mathrm{C} < \infty$ for all j, k.
(C7) Assume that $g(x;\beta_n)$ is a continuous function of β_n, The second derivatives of $g(x;\beta_n)$ with respect to β exist and are continuous. In addition, $n^{-1}\sum_{i=1}^n g'(x;\beta_n)g'(x;\beta_n)^T$ converges to a finite positive definite matrix $\Psi(\beta_n)$.
(C8) Let the values of β_{n1}^*, ..., $\beta_{ns_n}^*$ be nonzero and $\beta_{n(s_n+1)}^*$, ..., $\beta_{np_n}^*$ be zero. Then β_{n1}^*, ..., $\beta_{ns_n}^*$ such that

$$\min_{1\le j\le s_n} |\beta_{nj}^*|/\lambda_n \to \infty, \quad as \quad n \to \infty.$$

Remark 3. Conditions (C1)-(C3) are essentially the same as those in Fan and Peng (2004). Conditions (C4)-(C5) are assumed in Yao et al. (2012) for local modal nonparametric regression. The condition $\mathrm{E}(\phi_h'(\varepsilon)|x) = 0$ ensures that the proposed estimate is consistent and is satisfied if the error density is symmetric about zero. Conditions (C6)-(C7) are similar to the conditions (F)-(G) placed on the information matrix in Fan and Peng (2004). Condition (C8) is used to obtain the oracle property when using the SCAD penalty.

Proof of Theorem 1. Note that maximizing the objective function (2.3) is equivalent to minimizing $\mathscr{Q}_n(\beta_n) = -\sum_{i=1}^n \phi_h(y_i - g(x_i;\beta_n)) + n\sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|)$. Let $\delta_n = \sqrt{p_n}(n^{-1/2} + a_n)$, $\mathbf{v} = \delta_n^{-1}(\beta_n - \beta_n^*)$ and set $\|\mathbf{v}\| = C$. Let us first show that, for any given $\xi > 0$, there exists a large C such that

$$P\Big\{\inf_{\|\mathbf{v}\|=C} \mathscr{Q}_n(\beta_n^* + \delta_n\mathbf{v}) > \mathscr{Q}_n(\beta_n^*)\Big\} \ge 1 - \xi. \tag{A.1}$$

This implies that, with probability at least $1-\xi$, there exists a local minimizer in the ball $\{\beta_n^* + \delta_n\mathbf{v} : \|\mathbf{v}\| \le C\}$.

Let $D_n(\mathbf{v}) = \mathscr{Q}_n(\beta_n^* + \delta_n\mathbf{v}) - \mathscr{Q}_n(\beta_n^*)$. Then by definition of $\mathscr{Q}_n(\beta_n)$ in (2.3), we have

$$\begin{aligned}
D_n(\mathbf{v}) &\equiv \sum_{i=1}^n \big[-\phi_h(y_i - g(x_i;\beta_n^* + \delta_n\mathbf{v})) + \phi_h(y_i - g(x_i;\beta_n^*))\big] \\
&\quad + n\sum_{j=1}^{p_n}\big\{p_{\lambda_n}(|\beta_{nj}^* + \delta_n v_j|) - p_{\lambda_n}(|\beta_{nj}^*|)\big\} \\
&\ge \sum_{i=1}^n \big[-\phi_h(y_i - g(x_i;\beta_n^* + \delta_n\mathbf{v})) + \phi_h(y_i - g(x_i;\beta_{n0}))\big] \\
&\quad + n\sum_{j=1}^{s_n}\big\{p_{\lambda_n}(|\beta_{nj}^* + \delta_n v_j|) - p_{\lambda_n}(|\beta_{nj}^*|)\big\} \\
&\equiv: J_1 + J_2.
\end{aligned} \tag{A.2}$$

Using Taylor expanding $g(x_i;\beta_n)$ around β_n^*, on the basis of the boundness of $g''(\ ;\ )$ and $\|\beta_n-\beta_n^*\|\le C\delta_n$, then we have

$$g(x_i;\beta_n^*+\delta_n\mathbf{v})=g(x_i;\beta_n^*)+g'(x_i;\beta_n^*)^T\delta_n\mathbf{v}(1+o_p(1)). \tag{A.3}$$

For the first part J_1, by using the Taylor expansion and (A.3), we obtain that

$$\begin{aligned}J_1&=\sum_{i=1}^n\delta_n\phi_h'(\varepsilon_i)g'(x_i;\beta_{n0})^T\mathbf{v}-\sum_{i=1}^n\delta_n^2\phi_h''(\varepsilon_i)[g'(x_i;\beta_n^*)^T\mathbf{v}]^2+\sum_{i=1}^n\delta_n^3\phi_h''(\varepsilon_i^*)[g'(x_i;\beta_n^*)^T\mathbf{v}]^3\\&\equiv: J_{11}+J_{12}+J_{13},\end{aligned} \tag{A.4}$$

where ε_i^* lies in ε_i and $\varepsilon_i-\delta_n g'(x_i;\beta_n^*)^T\mathbf{v}$.

By directly calculating the mean and the variance, and the regularity condition (C5), we have $J_{11}=O(Cn\delta_n)$. Similarly, we can prove that $J_{13}=O(n\delta_n^2)$. As for J_{12}, we have

$$J_{12}=-\delta_n^2nF(x,h)\mathbf{v}^T\mathrm{E}[g'(x_i;\beta_n^*)g'(x_i;\beta_n^*)^T|x]\mathbf{v}(1+o_p(1)). \tag{A.5}$$

By the regularity condition (C4), $F(x,h)<0$ and $\mathrm{E}[g'(x_i;\beta_n^*)g'(x_i;\beta_n^*)^T|x]$ is a finite positive definite matrix by condition (C7). Hence, by choosing a sufficiently large C, J_{12} dominates both J_{11} and J_{13} in $\|\mathbf{v}\|=C$.

Next, we consider J_2, by invoking $p_{\lambda_n}(0)=0$, then by the standard argument of the Taylor expansion, we obtain that

$$\begin{aligned}J_2&\approx n\delta_n\sum_{j=1}^{s_n}\{p_{\lambda_n}'(|\beta_{nj}^*|)\mathrm{sgn}(\beta_{nj}^*)v_j+\frac{1}{2}\delta_np_{\lambda_n}''(|\beta_{nj}^*|)v_j^2\}\\&=n\sqrt{s_n}\delta_n^2C+nb_n\delta_n^2C^2.\end{aligned} \tag{A.6}$$

By the condition (C2), it is easy to show that J_2 is dominated by J_{12} uniformly in $\|\mathbf{v}\|=C$. Hence, by choosing a sufficiently large C, we have $D_n(\mathbf{v})>0$, which implies that with the probability at $1-\xi$, (A.1) holds and the proof of Theorem 1 is complete. □

To prove Theorem 2, we first show that the nonconcave penalized estimator possesses the sparsity property $\hat\beta_{n\mathcal{A}^c}=0$ by the following lemma.

Lemma A.1 *Under conditions (C1)-(C8). If $\lambda_n\to 0$, $\sqrt{p_n/n}\lambda_n\to\infty$ and $p_n^3/n\to 0$ as $n\to\infty$, then with probability tending to 1, for any given $\beta_{n\mathcal{A}}$ satisfying $\|\beta_{n\mathcal{A}}-\beta_{n\mathcal{A}}^*\|=O_p(\sqrt{p_n/n})$ and any constant C*

$$\mathscr{Q}_n\left\{\begin{pmatrix}\beta_{n\mathcal{A}}\\0\end{pmatrix}\right\}=\min_{\|\beta_{n\mathcal{A}^c}\|\le C\sqrt{p_n/n}}\mathscr{Q}_n\left\{\begin{pmatrix}\beta_{n\mathcal{A}}\\\beta_{n\mathcal{A}^c}\end{pmatrix}\right\}.$$

Proof of Lemma A.1. Let $\varsigma_n=C\sqrt{p_n/n}$, it is sufficient to prove that with probability tending to 1 as

$n \to \infty$, for any β_{n1} such that $\|\beta_{n1} - \beta_{n1}^*\| = O_p(\sqrt{p_n/n})$, we have,

$$\frac{\partial \mathscr{Q}_n(\beta_n)}{\partial \beta_{nj}} < 0 \quad for \quad 0 < \beta_{nj} < \varsigma_n, \quad j = s_n + 1, ..., p_n,$$
$$\frac{\partial \mathscr{Q}_n(\beta_n)}{\partial \beta_{nj}} > 0 \quad for \quad -\varsigma_n < \beta_{nj} < 0, \quad j = s_n + 1, ..., p_n.$$

By a similar proof of Theorem 1, we can obtain that

$$\begin{aligned}
\frac{\partial \mathscr{Q}_n(\beta_n)}{\partial \beta_{nj}} &= \sum_{i=1}^{n} \frac{\partial g(x_i; \beta_n^*)}{\partial \beta_{nj}} \phi_h'(\varepsilon_i - \delta_n g'(x_i; \beta_n^*)^T \mathbf{v}) + n p_{\lambda_n}'(|\beta_{nj}|)\mathrm{sgn}(\beta_{nj}) \\
&= \sum_{i=1}^{n} \frac{\partial g(x_i; \beta_n^*)}{\partial \beta_{nj}} \left\{ \phi_h'(\varepsilon_i) - \delta_n \phi_h''(\varepsilon_i) g'(x_i; \beta_n^*)^T \mathbf{v} + \delta_n^2 \phi_h'''(\varepsilon_i^{**}) [g'(x_i; \beta_n^*)^T \mathbf{v}]^2 \right\} \\
&\quad + n p_{\lambda_n}'(|\beta_{nj}|)\mathrm{sgn}(\beta_{nj}) \\
&= n\lambda_n \left\{ \lambda_n^{-1} p_{\lambda_n}'(|\beta_{nj}|)\mathrm{sgn}(\beta_{nj}) + O_p\left(\sqrt{\frac{p_n}{n}} \Big/ \lambda_n \right) \right\} \qquad \text{(A.7)}
\end{aligned}$$

Since $\sqrt{p_n/n}/\lambda_n \to 0$, and $\liminf_{n\to\infty} \inf_{t\to 0^+} p_{\lambda_n}'(t)/\lambda_n > 0$, then it is easy to see the sign of the derivative β_{nj} is completely determined by that of the sign of $\partial \mathscr{Q}_n(\beta_n)/\partial \beta_{nj}$. This completes the proof of Lemma A.1. □

Proof of Theorem 2. Part 1 of the theorem holds by Lemma A.1. We prove Part 2 of the theorem in the following. By Lemma A.1 and Theorem 1, there exists $\hat{\beta}_{n\mathcal{A}}$ satisfying the following equations:

$$\frac{\partial \mathscr{Q}_n(\beta_n)}{\partial \beta_{nj}}\Big|_{\beta_n = (\hat{\beta}_{n\mathcal{A}}^T, 0)^T} = 0, \quad j = 1, ..., s_n.$$

Then, by simple calculation, we have

$$\begin{aligned}
\frac{\partial \mathscr{Q}_n(\beta_n)}{\partial \beta_{nj}}\Big|_{\beta_n = (\hat{\beta}_{n\mathcal{A}}^T, 0)^T} &= \sum_{i=1}^{n} \frac{\partial g(x_i; \beta_n^*)}{\partial \beta_{nj}} \left\{ \phi_h'(\varepsilon_i) - \delta_n \phi_h''(\varepsilon_i) g_{\mathcal{A}}'(x_i; \beta_n^*)^T \mathbf{v}_{\mathcal{A}} + \delta_n^2 \phi_h'''(\varepsilon_i^{**}) [g_{\mathcal{A}}'(x_i; \beta_n^*)^T \mathbf{v}_{\mathcal{A}}]^2 \right\} \\
&\quad + n\{p_{\lambda_n}'(|\beta_{nj}^*|)\mathrm{sgn}(\beta_{nj}^*) + (p_{\lambda_n}''(|\beta_{nj}^*|) + o_p(1))(\hat{\beta}_{nj} - \beta_{nj}^*)\}, \qquad \text{(A.8)}
\end{aligned}$$

where $j = 1, ..., s_n$. Combining all these equations, we have

$$\begin{aligned}
0 = \sum_{i=1}^{n} g_{\mathcal{A}}'(x_i; \beta_n^*) &\left\{ \phi_h'(\varepsilon_i) - \phi_h''(\varepsilon_i) g_{\mathcal{A}}'(x_i; \beta_n^*)^T (\hat{\beta}_{n\mathcal{A}} - \beta_{n\mathcal{A}}^*) + \phi_h'''(\varepsilon_i^{**}) [g_{\mathcal{A}}'(x_i; \beta_n^*)^T (\hat{\beta}_{n\mathcal{A}} - \beta_{n\mathcal{A}}^*)]^2 \right\} \\
&+ n\{\boldsymbol{b}_n + (\Sigma_{\lambda_n} + o_p(1))(\hat{\beta}_{n\mathcal{A}} - \beta_{n\mathcal{A}}^*)\}. \qquad \text{(A.9)}
\end{aligned}$$

Then, it follows by Slutskys theorem and the central limit theorem that

$$\sqrt{n}(\Xi_{\mathcal{A}} + \Sigma_{\lambda_n})\{\hat{\beta}_{n\mathcal{A}} - \beta_{n\mathcal{A}}^* + (\Xi_{\mathcal{A}} + \Sigma_{\lambda_n})^{-1} \boldsymbol{b}_n\} \xrightarrow{d} N(0, \Omega_{\mathcal{A}}) \quad as\ n \to \infty, \qquad \text{(A.10)}$$

This completes the proof of Theorem 2. □

References

[1] Berlinet, A., Vajda, I., van der Meulen, E. C., 1998. About the asymptotic accuracy of Barron density estimates. Information Theory, IEEE Transactions on 44(3), 999-1009.

[2] Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Ann. Statist. 32, 928-961.

[3] Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96(456), 1348-1360.

[4] Frank, L. E., Friedman, J. H., 1993. A statistical view of some chemometrics regression tools. Technometrics 35(2), 109-135.

[5] Hedges, S. B., Shah, P., 2003. Comparison of mode estimation methods and application in molecular clock analysis. BMC bioinformatics 4(1), 31.

[6] Jiang, X., Jiang, J., Song, X., 2012. Oracle model selection for nonlinear models based on weighted composite quantile regression. Statist. Sin. 22(4), 1479-1506.

[7] Kemp, G. C., Santos Silva, J. M. C., 2012. Regression towards the mode. Journal of Econometrics, 170(1), 92-101.

[8] Meyer, M., 2001. An alternative unimodal density estimator with a consistent estimate of the mode. Statit. Sin. 11(4), 1159-1174.

[9] Lam, C., Fan, J., 2008. Profile-kernel likelihood inference with diverging number of parameters. Ann. Statist. 36(5), 2232-2260.

[10] Liu, J., Zhang, R., Zhao, W., Lv, Y., 2013. A robust and efficient estimation method for single index models. J. Multi. Analys. 122, 226-238.

[11] Li, J., Ray, S., Lindsay, B. G., 2007. A nonparametric statistical approach to clustering via mode identification. J. Machi. Learning Res. 8(8), 1687-1723.

[12] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Statist. Soci. Ser. B, 267-288.

[13] Wang, H., Li, G., Jiang, G., 2007. Robust regression shrinkage and consistent variable selection through the LAD-lasso. J. Busi. Econ. Statist. 25, 347-355.

[14] Yao, W., Li, L., 2013. A new regression model: modal linear regression. Scand. J. Statist., doi: 10.1111/sjos.12054.

[15] Yao, W., Lindsay, B. G., Li, R., 2012. Local modal regression. J. Nonpar. Stati. 24(3), 647-663.

[16] Zhang, R., Zhao, W., Liu, J., 2013. Robust estimation and variable selection for semiparametric partially linear varying coefficient model based on modal regression. J. Nonpar. Stati. 25(2), 523-544.

[17] 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. Insti. Statis. Math., 66(1), 165-191.

[18] Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101(476), 1418-1429.

[19] Zou, H., Hastie, T., Tibshirani, R., 2007. On the degree of freedom of the lasso. Ann. Statist. 35, 2173-2192.