应用概率统计　第三十一卷
第一期　2015年2月

Chinese Journal of Applied Probability
and Statistics Vol.31 No.1 Feb. 2015

《应用概率统计》版权所有

Variable Selection of Single-Index Quantile Regression [*]

Lu Yiqiang[1]　　Li Feng[2]　　Hu Bin[1]

([1] *The PLA Information Engineering University, Zhengzhou, 450000*)

([2] *School of Mathematics and Statistics, Zhengzhou University, Zhengzhou, 450001*)

Abstract

Nonparametric quantile regression with multivariate covariates is a difficult estimation. To reduce the dimensionality while still retaining the flexibility of nonparametric model, the single-index regression is often used to model the conditional quantile of a response variable. In this paper, we focus on the variable selection aspect of single-index quantile regression. Based on the minimized average loss estimation (MALE), the variable selection is done by minimizing the average loss with SCAD penalty. Under some mild conditions, we demonstrate the oracle properties about SCAD variable section of single-index quantile regression. Furthermore, the algorithm of the variable selection of SCAD penalized quantile regression is given. Some simulations are done to illustrate the performance of the proposed methods.

Keywords: Single-index model, quantile regression, SCAD, variable selection.

AMS Subject Classification: 62G05.

§1. Introduction

Least squares regression estimates the conditional mean, that is, the mean response as a function of regressors. Least absolute deviation regression estimates the conditional median function. Koenker and Bassett (1978) introduced quantile regression (QR), which can be used to estimate the conditional quantile function of the response. QR has attracted tremendous interest in the literature and been widely developed in the past decades. The linear quantile regression has been proved to be very useful (Koenker, 2005), but like linear regression, it is not flexible to capture complicated relations. Stone (1977) and Chaudhuri (1991) considered fully multivariate nonparametric quantile regression, which is flexible but usually unattractive in practice due to “curse of dimensionality”. The single-index approach has been proved to be an efficient way to cope with high-dimensional nonparametric estimation problems in not only conditional mean but also

[*] The project was supported by the National Natural Science Foundation of China (61272041, U1404104) and the Education Department of Henan Province (14A110015).

Received March 5, 2013. Revised February 19, 2014.

doi: 10.3969/j.issn.1001-4268.2015.01.003

quantile regression involving multivariate covariates (Wu, Yu and Yu, 2010 and Zhu, Huang and Li, 2012).

Suppose that Y is the response variable and X is a covariate vector. Single-index quantile regression (SIQR) is defined as

$$Y = m_0(X^{\mathsf{T}}\beta_0) + \epsilon, \qquad Q_\tau(\epsilon|X) = 0, \tag{1.1}$$

where $Q_\tau(\epsilon|X)$ is the τ-th conditional quantile of ϵ. $\beta_0 \in R^p$ with $\|\beta_0\| = 1$ and the first component $\beta_{01} > 0$, is an unknown parameter vector. $m_0(\cdot)$ is an unknown nonparametric function. The single-index model has many advantages such as flexibility, effectively reducing dimensionality, simple interpretation, and so on. Based on the local linear approach, Wu, Yu and Yu (2010) proposed the minimized average loss estimation for single-index quantile regression.

Variable selection plays an important role in the single-index model building process. As we know, when the dimension of covariates X is high, X maybe contain irrelevant regressors. Exclusion of irrelevant variables from a large number of candidate predictors becomes crucial since inclusion of irrelevant predictors may decrease the interpretative and predictive ability of the resultant model. To automatically select the variables with nonzero coefficients, many different types of penalties have been introduced in the literature. Compared with traditional estimation methods, the major advantage of penalized estimator is its simultaneous execution of both parameter estimation and variable selection. The L_1 penalty was used in the LASSO proposed by Tibshirani (1996). The smoothly clipped absolute deviation (SCAD) penalty function was proposed by Fan and Li (2001) and shown to possess the oracle properties of variable selection (consistent, sparse and efficient). Zou (2006) introduced the adaptive LASSO (aLASSO), which is slightly different from LASSO in that different amounts of shrinkage are used for different regression coefficient, and demonstrated its oracle properties.

In the literatures, most works of variable selection were put on the conditional mean regression. The selection criterion of AIC for single-index mean regression was studied by Naik and Tsai (2001). The main novel part of this paper is the inclusion of variable selection of single-index quantile regression by the methods of SCAD penalty. The rest of this paper is organized as follows. Section 2 describes estimation methodology of the minimized average loss estimation (MALE). The SCAD variable selection and its oracle properties are given in Section 3. Numerical studies are conducted to evaluate the finite sample performance of the proposed methods in Section 4. All technical proofs are relegated to the Appendix.

《应用概率统计》版权所有

《应用概率统计》版权所有

§2. The Minimized Average Loss Estimation

Suppose that the loss function is specified as

$$\rho_\tau(v) = v(\tau - I(v < 0)) = v(\tau I(v > 0) + (\tau - 1)I(v \le 0)), \tag{2.1}$$

where $0 < \tau < 1$ and $I(\cdot)$ is the identify function. Koenker and Bassett (1978) demonstrated that the τth conditional quantile function can be estimated by minimizing loss function defined by (2.1). Mathematically, the true model (1.1) solves the following minimizing problem

$$\arg\min \mathsf{E}[\rho_\tau(Y - m(X^\intercal\beta))], \tag{2.2}$$

with respect to $\beta \in \{\beta \in R^p : \|\beta\| = 1 \text{ and } \beta_1 > 0\}$ and $m(\cdot) \in L_1$. For $X^\intercal\beta$ "close" to u, $m(X^\intercal\beta)$ can be approximated by

$$m(X^\intercal\beta) \approx m(u) + m'(u)(X^\intercal\beta - u) = a + b(X^\intercal\beta - u), \tag{2.3}$$

where $a = m(u)$ and $b = m'(u)$. Suppose that $\{y_i, X_i\}$, $i = 1, 2, \ldots, n$ is a sample of size n from the model (1.1). The sample analog of (2.2) can be written as

$$\sum_{j=1}^{n}\sum_{i=1}^{n} \rho_\tau(y_i - a_j - b_j X_{ij}^\intercal\beta)W_{ij}, \tag{2.4}$$

where $X_{ij} = X_i - X_j$, $a_j = m(X_j^\intercal\beta)$, $b_j = m'(X_j^\intercal\beta)$,

$$W_{ij} = K\Big(\frac{\beta^\intercal X_{ij}}{h}\Big)\Big/\sum_{l=1}^{n} K\Big(\frac{\beta^\intercal X_{lj}}{h}\Big),$$

$K(\cdot)$ is the kernel function and h is the bandwidth. The estimators obtained by minimizing (2.4) with respect to β and (a_j, b_j) are said to be the minimized average loss estimators (MALE), which can refer to Wu, Yu and Yu (2010). When $\rho_\tau(\cdot)$ is replaced by least square loss function, which is used for conditional mean regression, the obtained estimator is called to be the minimized average variance estimator (MAVE, see Xia and Härdle, 2006).

Minimizing (2.4) can be decomposed to two typical quantile regression problems by fixing β and (a_j, b_j) alternatively. With β given,

$$(\widetilde{a}_j, \widetilde{b}_j)^\intercal = \arg\min_{a_j, b_j} \sum_{i=1}^{n} \rho_\tau(y_i - a_j - b_j X_{ij}^\intercal\beta)K(X_{ij}^\intercal\beta/h), \qquad j = 1, 2, \ldots, n. \tag{2.5}$$

With (a_j, b_j) given,

$$\widetilde{\beta} = \arg\min_{\beta} \sum_{j=1}^{n}\sum_{i=1}^{n} \rho_\tau(y_i - a_j - b_j X_{ij}^\intercal\beta)W_{ij}, \tag{2.6}$$

《应用概率统计》版权所有

where W_{ij} is evaluated at the previous estimate of β. Standardize $\widetilde{\beta}$ to $\widetilde{\beta} = s_1\widetilde{\beta}/\|\widetilde{\beta}\|$, where s_1 is the sign of the first entry in $\widetilde{\beta}$. Repeat the above two steps until convergence.

At last, obtain the estimate of $m(u)$ at any u, $\widetilde{m}(u; h, \widetilde{\beta}) = \widetilde{a}$, where

$$(\widetilde{a}, \widetilde{b}) = \arg\min_{a,b} \sum_{i=1}^{n} \rho_\tau(y_i - a - b(X_i^{\mathsf{T}}\widetilde{\beta} - u))K_h(X_i^{\mathsf{T}}\widetilde{\beta} - u) \tag{2.7}$$

in which $K_h(u) = (1/h)K(u/h)$. The initial estimate can be obtained by the average derivative estimation (Chaudhuri, Doksum and Samarov, 1997). Denote by $\widetilde{\beta}$ and $(\widetilde{a}_j^{\mathsf{T}}, \widetilde{b}_j^{\mathsf{T}})^{\mathsf{T}}$, $j = 1, 2, \ldots, n$, the obtained estimates.

Both (2.5) and (2.6) are simple linear quantile regression problem. Several efficient algorithms for linear quantile are available and see Koenker (2005). Under some regular conditions (see Appendix), Wu, Yu and Yu (2010) obtained the asymptotic properties of MALE $\widetilde{m}(u)$ and $\widetilde{\beta}$.

Lemma 2.1 Suppose that Assumptions A1−A4 in Appendix hold. If $n \to \infty$, $h \to 0$ and $nh \to \infty$, then for an interior point u,

$$\sqrt{nh}\left\{\widetilde{m}(u; h, \widetilde{\beta}) - m_0(u) - \frac{m_0''(u)\displaystyle\int v^2 K(v)\mathrm{d}v}{2}h^2\right\} \xrightarrow{w} \mathrm{N}(0, \alpha^2(u)),$$

where

$$\alpha^2(u) = \frac{\displaystyle\int K^2(v)\mathrm{d}v}{f_{U_0}(u)} \frac{\tau(1-\tau)}{[f_y(m_0(u))]^2},$$

$f_{U_0}(u)$ is the density of $U_0 = X^{\mathsf{T}}\beta_0$ and $f_y(\cdot)$ is the conditional density of y given $X^{\mathsf{T}}\beta = u$.

Lemma 2.2 Suppose that Assumptions A1−A4 in Appendix hold. If $n \to \infty$, $h \to 0$ and $nh \to \infty$, we have

$$\sqrt{n}(\widetilde{\beta} - \beta) \xrightarrow{w} \mathrm{N}(0, \tau(1-\tau)\Delta^{-1}\Sigma\Delta^{-1}),$$

where

$$\Sigma = \mathsf{E}\{m_0'(X^{\mathsf{T}}\beta_0)^2[X - \mathsf{E}(X|X^{\mathsf{T}}\beta_0)][X - \mathsf{E}(X|X^{\mathsf{T}}\beta_0)]^{\mathsf{T}}\}$$

and

$$\Delta = \mathsf{E}\{f_y(m_0(X^{\mathsf{T}}\beta_0))m_0'(X^{\mathsf{T}}\beta_0)^2[X - \mathsf{E}(X|X^{\mathsf{T}}\beta_0)][X - \mathsf{E}(X|X^{\mathsf{T}}\beta_0)]^{\mathsf{T}}\}.$$

Lemma 2.1 and 2.2 see Theorem 1 and 3 in Wu, Yu and Yu (2010).

《应用概率统计》版权所有

§3. The Variable Selection of SIQR

Denote by $\widetilde{\beta}$ and $\{\widetilde{a}_j, \widetilde{b}_j\}_{j=1}^n$ the above obtained nonpenalized estimators. When the dimension of X is large, the coefficients are usually sparse and variable selection is crucial. As for linear quantile regression, the variable selection has been considered in several papers, such as the penalized methods of Wu and Liu (2009) and Bayesian method of Alhamzawi and Yu (2012). For SIQR (1.1), to avoid over-fitting and improve generalization, we consider the penalized version of (2.4). The finial estimator of β is obtained by minimizing the average loss with penalty, that is,

$$\widehat{\beta} = \arg\min_{\beta} \sum_{j=1}^n \sum_{i=1}^n \rho_\tau(y_i - \widetilde{a}_j - \widetilde{b}_j X_{ij}^{\mathsf{T}}\beta) W_{ij} + \sum_{j=1}^p p_\lambda(|\beta_j|), \tag{3.1}$$

where $\lambda > 0$ is the regularization parameter and W_{ij} is evaluated at $\widetilde{\beta}$. Fan and Li (2001) argued that a good penalty should possess the following three properties in its estimator: unbiasedness, sparsity and continuity. In the recent literatures, the popular penalty includes SCAD (Fan and Li, 2001) and adaptive LASSO (Zou, 2006) penalty, both of which were proved to achieve these three desirable properties simultaneously. In this paper, we consider the variable selection of SIQR (1.1) via the SCAD penalty.

3.1 The SCAD Selection

The SCAD penalty is defined in term of its first order derivative and is symmetric around the origin. For $\theta > 0$, its first derive is given by

$$p'_\lambda(\theta) = \lambda\Big\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\Big\}, \tag{3.2}$$

where $a > 2$ and $\lambda > 0$ are tuning parameters. Particularly, $a = 3.7$ is often selected. The minimizer of (3.1) with SCAD penalty (3.2) is denoted by $\widehat{\beta}^{(S)}$. That is

$$\widehat{\beta}^{(S)} = \arg\min_{\beta} \sum_{j=1}^n \sum_{i=1}^n \rho_\tau(y_i - \widetilde{a}_j - \widetilde{b}_j X_{ij}^{\mathsf{T}}\beta) W_{ij} + n \sum_{j=1}^p p_{\lambda_n}(|\beta_j|), \tag{3.3}$$

where $p_{\lambda_n}(|\beta_j|)$ is defined in (3.2). The final estimator of single-index function, $\widehat{m}(u)$, is obtained from (2.7) through replacing $\widetilde{\beta}$ by $\widehat{\beta}$. Notice that the SCAD penalty function is symmetric non-convex on $[0, \infty)$ and singular at origin. The SCAD penalizes large coefficients equally. Hence, the SCAD results in unbiased penalized estimators for large coefficients. Fan and Li (2001) demonstrated that the SCAD penalty has the oracle properties of the variable selection in the penalized likelihood setting.

《应用概率统计》版权所有

Theorem 3.1 (Consistency) Under Assumptions A1 – A4 given in Appendix, if $\lambda_n \to 0$, there is a local minimizer $\widehat{\beta}^{(S)}$ defined by (3.3) such that $\widehat{\beta}^{(S)} - \beta_0 = O_p(n^{-1/2})$.

For convenience, let's denote $A = \{j : \beta_{0,j} \neq 0\}$, $\beta_A = (\beta_i : i \in A)$ and A^c be the complement of A. Define $M_A = (a_{ij}|i, j \in A)$ as the submatrix of $M = (a_{ij})$.

Theorem 3.2 (Oracle) Under the same conditions as in Theorem 3.1, if $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$, as $n \to \infty$, then

(a) Sparsity: $\mathsf{P}\{\widehat{\beta}^{(S)}_{A^c} = 0\} \to 1$.

(b) Asymptotic normality:

$$\sqrt{n}(\widehat{\beta}^{(S)}_A - \beta_{0,A}) \to \mathrm{N}(0, \tau(1-\tau)\Delta_A^{-1}\Sigma_A\Delta_A^{-1}),$$

where Δ_A and Σ_A are submatrix of Δ and Σ, the definitions of which see Lemma 2.2.

The proof of Theorem 3.1 and 3.2 sees Appendix.

3.2 Algorithm for SCAD Selection

Due to the SCAD penalty is non-convex, the corresponding minimization problem is hard to solve. In Fan and Li (2001), a unified least quadratic approximation (LQA) algorithm was proposed to solve the SCAD likelihood optimization problem. Hunter and Li (2005) studied LQA under a more general M-algorithm framework. To avoid a disturb parameter, Wu and Liu (2009) noticed that the SCAD penalty function can be decomposed as the difference of two convex function. That is, $p_\lambda(x) = p_{\lambda,1}(x) - p_{\lambda,2}(x)$, where both $p_{\lambda,1}(x)$ and $p_{\lambda,2}(x)$ are convex functions and their derivatives for $x > 0$ are given

$$\begin{cases} p'_{\lambda,1}(x) = \lambda; \\ p'_{\lambda,2}(x) = \lambda[1 - (a\lambda - x)_+/[(a-1)\lambda]]I(x > \lambda). \end{cases}$$

The above decomposition of SCAD penalty allows us to use the difference convex algorithm (DCA). More Specifically, the objective function (3.3) is decomposed as $Q_{\mathrm{vex}}(\beta) + Q_{\mathrm{cav}}(\beta)$, where

$$Q_{\mathrm{vex}}(\beta) = \sum_{j=1}^{n}\sum_{i=1}^{n} \rho_\tau(y_i - \widetilde{a}_j - \widetilde{b}_j X_{ij}^{\mathsf{T}}\beta)W_{ij} + n\sum_{j=1}^{p} p_{\lambda,1}(|\beta_j|), \quad Q_{\mathrm{cav}}(\beta) = -n\sum_{j=1}^{p} p_{\lambda,2}(|\beta_j|).$$

Repeat

$$\beta^{(k+1)} = \arg\min_{\beta}(Q_{\mathrm{vex}}(\beta) + \langle Q'_{\mathrm{cav}}(\beta^{(k)}), \beta - \beta^{(k)}\rangle)$$

until convergence. Notice that the derivative of the concave part is

$$Q'_{\mathrm{cav}}(\beta^{(k)}) = -n\big(p'_{\lambda,2}(|\beta_1^{(k)}|)\mathrm{sign}(\beta_1^{(k)}), p'_{\lambda,2}(|\beta_2^{(k)}|)\mathrm{sign}(\beta_2^{(k)}), \ldots, p'_{\lambda,2}(|\beta_p^{(k)}|)\mathrm{sign}(\beta_p^{(k)})\big).$$

《应用概率统计》版权所有

In the $(k+1)$-th iteration, DCA solves the following optimization problem:

$$\begin{aligned}\min_{\beta}\Big\{&\sum_{j=1}^{n}\sum_{i=1}^{n}\rho_\tau(y_i-\widetilde{a}_j-\widetilde{b}_jX_{ij}^{\mathrm{T}}\beta)W_{ij}+n\sum_{j=1}^{p}p_{\lambda,1}(|\beta_j|)\\&-n\sum_{j=1}^{p}p'_{\lambda,2}(|\beta_j^{(k)}|)\mathrm{sign}(\beta_j^{(k)})(\beta_j-\beta_j^{(k)})\Big\}.\end{aligned}\tag{3.4}$$

We can use the solution of non-penalized quatile regression as the initial value. By introducing some slack variables, we can recast the above minimization problem (3.4) into the following linear programming problem:

$$\begin{aligned}&\min \sum_{j=1}^{n}\sum_{i=1}^{n}[\tau\xi_{ij}+(1-\tau)\zeta_{ij}]W_{ij}+n\lambda_n\sum_{j=1}^{p}\nu_j-n\sum_{j=1}^{p}p'_{\lambda,2}(|\beta_j^{(k)}|)\mathrm{sign}(\beta_j^{(k)})(\beta_j-\beta_j^{(k)})\\&\text{s.t.}\ \begin{cases}\xi_{ij}\ge 0,\ \zeta_{ij}\ge 0,\ \xi_{ij}-\zeta_{ij}=y_i-\widetilde{a}_j-\widetilde{b}_jX_{ij}^{\mathrm{T}}\beta,\ i,j=1,2,\ldots,n;\\ \nu_j\ge\beta_j,\ \nu_j\ge-\beta_j,\ j=1,2,\ldots,p,\end{cases}\end{aligned}$$

which can be easily solved by many optimization softwares.

The proposed estimates depend on the appropriate specification of bandwidth h and penalty parameter λ_n. With β fixed, the bandwidth h is actually selected for a univariate local linear quantile regression. For local linear quantile regression, Yu and Jones (1998) derived an approximate optimal bandwidth under moderate assumptions and gave the following rule-of-thumb bandwidth h_τ:

$$h_\tau=h_m\{\tau(1-\tau)/\phi(\Phi^{-1}(\tau))^2\}^{1/5},\tag{3.5}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and the cumulative distribution function of the standard normal distribution, respectively. h_m is the optimal bandwidth used in least square mean regression. For h_m, there are many existing algorithms. In practice, the simple rule-of-thumb (Fan and Gijbels, 1996) often works well. That is,

$$h_m=1.364\widehat{\sigma}_0n^{-1/5}.$$

For the penalty parameter λ, BIC criterion can be defined as

$$\mathrm{BIC}(\lambda)=\log\Big(\sum_{j=1}^{n}\sum_{i=1}^{n}\rho_\tau(y_i-\widetilde{a}_j-\widetilde{b}_jX_{ij}^{\mathrm{T}}\widehat{\beta}_\lambda)W_{ij}\Big)+\frac{d_\lambda\log n}{n},$$

where d_λ is the number of nonzero coefficients in $\widehat{\beta}_\lambda$, a simple estimate for the degrees of freedom, The penalty parameter λ is selected as

$$\widehat{\lambda}=\arg\min_{\lambda}\mathrm{BIC}(\lambda).$$

《应用概率统计》版权所有

§4. Monte Carlo Study

We present some numerical studies to demonstrate finite sample performance. Consider the following model

$$y = 2\exp\{-3(X^{\mathsf{T}}\beta_0)^2\} + \sigma(\epsilon - Q_\tau(\epsilon)). \tag{4.1}$$

The components of $X = (X_1, X_2, \ldots, X_5)$ are standard normal. The correlation between any two components X_i and X_j is set to be $0.5^{|i-j|}$. X and ϵ are independent.

Here we present results for the case where $\beta_0 = (2/3, 1/3, -2/3, 0, 0)^{\mathsf{T}}$ and $\sigma = 0.1$. The parameters are estimated under different cases with quantiles and random errors. For each case, we simulate $N = 100$ random samples with $n = 200$. Table 1 reports the results of variable selection, including about the average model size (AMS) with standard deviations in its corresponding parentheses, the percentage of correct models identified (PCM), the average numbers of correct and wrong zero coefficients, and mean squared error of the parametric estimation among N runs

$$\mathrm{MSE}_\beta = \frac{1}{N}\sum_{i=1}^{N}\|\widehat{\beta}_i - \beta_0\|^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{p}(\widehat{\beta}_{ik} - \beta_{0,k})^2$$

and its standard deviations in its corresponding parentheses. Table 2 shows that the frequency of variable selection among 100 runs for the model (4.1).

From Table 1, we see that the average model size approximates true value 3 and the average number of correct zeros approximates true value 2. Table 1 also shows that the proposed variable selection methods are very efficient for $\tau = 0.25, 0.5$ whether under the error $\mathrm{N}(0,1)$, $\mathrm{t}(5)$, $\chi(3)$ or Cauchy. The efficiency of variable selection become a little low for $\tau = 0.9$ and Cauchy error, which maybe is since the quantile curve is more difficult to estimate for τ approximating 1 and existing too many outliers. Table 2 depicts that all most relevant predictors are selected although possibly including a few redundant predictors. Table 3 summarizes the average computing time in seconds used for estimating the index parameter and variable selection for one replication. It can be seen from Table 3 that the computing time may be related to the error distributions. The computation for the Cauchy error needed more time than other error distribution. In addition, we found that most of computing time is used in the MALE estimation while a little time in the variable selection.

For $\tau = 0.5$, Figure 1 gives the average estimates of single-index function $m(\cdot)$ over 100 simulations with sample size 200 and the corresponding 95% confidence bands. The difference between the true single-index function and the average fit is barely visible, which shows that there is little bias. Furthermore, the confidence bands are reasonably close to the true curve, showing small variation in the estimates. The boxplots in Figure 1 depicts the parameter estimates from single index quantile regression (with $\epsilon \sim \mathrm{t}(5)$). One can see that the distributions of estimates are centered around the true values and estimated well.

Table 1 Simulation results based on 100 replications for the model (4.1)

ϵ	τ	AMS(SD)	PCM	ANCZ	ANICZ	MSE(SD)
N(0, 1)	0.25	3(0)	1	2	0	0.159(0.026)
	0.50	3.03(0.223)	0.98	1.97	0	0.041(0.027)
	0.90	3.02(0.141)	0.98	1.98	0	0.173(0.132)
t(5)	0.25	3.06(0.239)	0.94	1.94	0	0.038(0.005)
	0.50	3.07(0.256)	0.93	1.93	0	0.037(0.007)
	0.90	3.03(0.332)	0.95	1.95	0.02	0.203(0.259)
$\chi(3)$	0.25	3.03(0.171)	0.97	1.96	0.01	0.165(0.065)
	0.50	3.09(0.321)	0.92	1.91	0	0.209(0.529)
	0.90	3.15(0.386)	0.83	1.83	0.02	0.205(0.326)
Cauchy	0.25	3.38(0.736)	0.78	1.74	0.08	0.279(0.655)
	0.50	3.15(0.626)	0.89	1.89	0.06	0.209(0.646)
	0.90	3.66(0.89)	0.73	1.62	0.12	0.308(0.674)

Note: ANCZ, average number of correct zeros (true value = 2).

Table 2 Frequency of each covariate appearing in the resultant models among 100 runs for the model (4.1) (True $\beta_0 = (2/3, 1/3, -2/3, 0, 0)^{\mathsf{T}}$)

ϵ	τ	X_1	X_2	X_3	X_4	X_5
N(0, 1)	0.25	100	100	100	0	0
	0.50	100	100	100	2	1
	0.90	100	100	100	2	0
t(5)	0.25	100	100	100	4	2
	0.50	100	100	100	3	4
	0.90	100	99	99	1	4
$\chi(3)$	0.25	100	99	100	3	1
	0.50	100	100	100	6	3
	0.90	100	99	99	10	7
Cauchy	0.25	99	95	98	14	12
	0.50	100	99	99	10	11
	0.90	98	95	95	23	25

《应用概率统计》版权所有

《应用概率统计》版权所有

Table 3　The averages of computing times (in seconds) for model (4.1) with $p = 5$

ϵ	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.90$
N(0, 1)	54.0	54.9	54.9
t(5)	60.2	59.9	56.0
$\chi^2(3)$	53.2	54.8	51.4
Cauchy	67.2	68.9	65.09

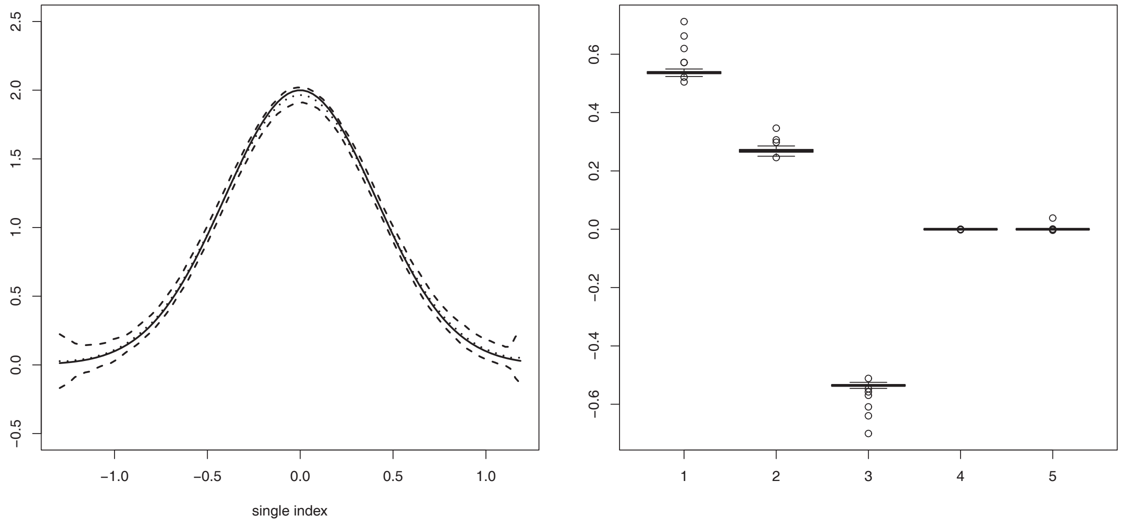

Figure 1　Estimates for single-index function and Boxplot of parameter estimates

In conclusion, by the use of the SCAD methods, most of the zero coefficients are estimated to be 0 and resultant model is easier to interpret. On the other hands, both the nonzero parameters and single-index function can also be estimated well. Hence, our proposed procedure is validated for the variable selection and estimation of single-index quantile regression.

Appendix

To prove the asymptotic properties, we need to lay out some basic assumptions.

A1　The density function of $\beta_0^{\mathsf{T}} X$ is continuous and bounded away from 0 and ∞ on its support. Further the density function of $\beta^{\mathsf{T}} X$ is continuous for β in a neighborhood of β_0.

A2　$K(\cdot)$ is a symmetric density function with bounded derivative and compact sup-

《应用概率统计》版权所有

port. It satisfies

$$\int u^2K(u)\mathrm{d}u<\infty \qquad \text{and} \qquad \Big|\int u^jK^2(u)\mathrm{d}u\Big|<\infty, \quad j=0,1,2.$$

A3 For each y, the conditional density function $f_y(y|X^{\mathsf{T}}\beta=u)$ is continuous in u. Furthermore, there exist positive constant δ_1,δ_2 and a positive function $G(y,u)$ such that

$$\sup_{|u'-u|\le\epsilon} f(y|u')\le G(y,u), \qquad \int|\rho_\tau(y-m(u))|^{2+\delta}G(y,u)\mathrm{d}y<\infty$$

and

$$\int(\rho_\tau(y-t)-\rho_\tau(y)-\rho_\tau'(y))^2G(y,u)\mathrm{d}y=o(t^2) \qquad \text{as } \ t\to 0.$$

A4 The single index function $m_0(\cdot)$ defined in SIQR (1.1) is bounded with continuous derivatives up to the second order.

The conditions above are commonly used in the literature. A1 guarantees that any ratio terms are meaningful when the density appears in the denominators; A2 requires that the kernel function is symmetric and has finite second moment, which is familiar to local estimation; A3 holds when $\rho_\tau'(\cdot)$ is Lipschitz continuous; A4 is a common requirement for a link function. These conditions were also assumed in Wu, Yu and Yu (2010).

Proof of Theorem 3.1 Notice that the SCAD estimator $\widehat{\beta}^{(S)}$ minimizes

$$\begin{aligned} Q_S(\beta)&=\sum_{j=1}^n\sum_{i=1}^n\{\rho_\tau(y_i-\widetilde{a}_j-\widetilde{b}_jX_{ij}^{\mathsf{T}}\beta)-\rho_\tau(Y_{ij})\}W_{ij}+n\sum_{j=1}^p p_{\lambda_n}(|\beta_j|)\\ &=\Phi_n(\beta)+n\sum_{j=1}^p p_{\lambda_n}(|\beta_j|), \end{aligned}\tag{5.1}$$

where

$$\Phi_n(\beta)=\sum_{j=1}^n\sum_{i=1}^n\{\rho_\tau(y_i-\widetilde{a}_j-\widetilde{b}_jX_{ij}^{\mathsf{T}}\beta)-\rho_\tau(Y_{ij})\}W_{ij} \quad \text{and} \quad Y_{ij}=Y_i-\widetilde{a}_j-\widetilde{b}_jX_{ij}^{\mathsf{T}}\beta_0.$$

To prove Theorem 3.1, it is enough to show that for any given $\delta>0$, there exists a large constant C such that

$$\mathsf{P}\Big\{\inf_{\|u\|=C}Q_s\Big(\beta_0+\frac{u}{\sqrt{n}}\Big)>Q_s(\beta_0)\Big\}\ge 1-\delta,\tag{5.2}$$

which implies that with probability at least $1-\delta$ there exists a local minimum in the ball $\{\beta_0+u/\sqrt{n}:\|u\|<C\}$. This in turn implies that there exists a local minimizer such that

$$|\widehat{\beta}^{(S)}-\beta_0|=O_p(n^{-1/2}).$$

《应用概率统计》版权所有

Let $u = \sqrt{n}(\beta-\beta_0)$ and $\widetilde{u} = \sqrt{n}(\widetilde{\beta}-\beta_0)$. It can be seen that $\widetilde{u}$ minimize the following

$$\Phi_n\Big(\beta_0+\frac{u}{\sqrt{n}}\Big) = \sum_{j=1}^{n}\sum_{i=1}^{n}\Big\{\rho_\tau\Big(Y_{ij}-\widetilde{a}_j-\frac{1}{\sqrt{n}}\widetilde{b}_j X_{ij}^{\top}u\Big)-\rho_\tau(Y_{ij})\Big\}W_{ij}.$$

From the proof of Theorem 3 in Wu, Yu and Yu (2010), we have

$$\Phi_n\Big(\beta_0+\frac{u}{\sqrt{n}}\Big) = \frac{1}{2}u^{\top}Su+V_n^{\top}u+o_p(1), \tag{5.3}$$

where $S=2\Delta$, $V_n=(4\tau(1-\tau))^{1/2}\Sigma^{1/2}Z_n$, $Z_n \xrightarrow{w} \mathrm{N}(0,I)$, and Σ and Δ are defined in Lemma 2.2.

From (5.1) and (5.3),

$$\begin{aligned}
&Q_s\Big(\beta_0+\frac{u}{\sqrt{n}}\Big)-Q_s(\beta_0)\\
\geq\ &\Phi_n\Big(\beta_0+\frac{u}{\sqrt{n}}\Big)+n\sum_{j\in A}\Big\{p_{\lambda_n}\Big(\Big|\beta_{0,j}+\frac{u}{\sqrt{n}}\Big|\Big)-p_{\lambda_n}(|\beta_{0,j}|)\Big\}\\
=\ &u^{\top}V_n+\frac{1}{2}u^{\top}Su+n\sum_{j\in A}\Big\{p_{\lambda_n}\Big(\Big|\beta_{0,j}+\frac{u}{\sqrt{n}}\Big|\Big)-p_{\lambda_n}(|\beta_{0,j}|)\Big\}+o_p(1).
\end{aligned}\tag{5.4}$$

Note that, for large n,

$$n\sum_{j\in A}\Big\{p_{\lambda_n}\Big(\Big|\beta_{j,0}+\frac{u_j}{\sqrt{n}}\Big|\Big)-p_{\lambda_n}(|\beta_{j,0}|)\Big\}=0 \tag{5.5}$$

uniformly in any compact set due to the facts that $|\beta_{j,0}|>0$ for $j\in A$, SCAD penalty is flat for coefficient magnitude larger than $a\lambda_n$ and $\lambda_n\to 0$.

Based on (5.4)$-$(5.5), $Q_s(\beta_0+u/\sqrt{n})-Q_s(\beta_0)$ is dominated by the quadratic term $(1/2)u^{\top}Su$ for $\|u\|$ equal to sufficiently large C. Hence S is positive definite matrix implies that (5.2) holds, as we have desired. □

Lemma 5.1 (Sparsity) If $\lambda_n\to 0$ and $\sqrt{n}\lambda_n\to\infty$, then with probability bending to one, for any β_A satisfying $\|\beta_A-\beta_{0,A}\|=O_p(n^{-1/2})$ and any constant C

$$Q_s((\beta_A^{\top},0^{\top})^{\top})=\min_{\|\beta_{A^c}\|\leq Cn^{-1/2}}Q_s((\beta_A^{\top},\beta_{A^c}^{\top})^{\top}).$$

Proof Notice that

$$(\beta_A^{\top},0)^{\top}=\beta_0+\frac{\sqrt{n}((\beta_A-\beta_{0,A})^{\top},0)}{\sqrt{n}}$$

and

$$(\beta_A^{\top},\beta_{A^c}^{\top})^{\top}=\beta_0+\frac{\sqrt{n}((\beta_A-\beta_{0,A})^{\top},\beta_{A^c}^{\top})}{\sqrt{n}}.$$

《应用概率统计》版权所有

By (5.1) and (5.3), we have

$$\begin{aligned}
&Q_S((\beta_A^{\top},0)^{\top})-Q_S((\beta_A^{\top},\beta_{A^c}^{\top})^{\top})\\
&=[Q_S((\beta_A^{\top},0)^{\top})-Q_S(\beta_0)]-[Q_S(\beta_A^{\top},\beta_{A^c}^{\top})-Q_S(\beta_0)]\\
&=\sqrt{n}((\beta_A-\beta_{0,A})^{\top},0^{\top})V_n+\frac{1}{2}\sqrt{n}((\beta_A-\beta_{0,A})^{\top},0)S\sqrt{n}((\beta_A-\beta_{0,A})^{\top},0)^{\top}\\
&\quad-\sqrt{n}((\beta_A-\beta_{0,A})^{\top},\beta_{A^c}^{\top})V_n-\frac{1}{2}\sqrt{n}((\beta_A-\beta_{0,A})^{\top},\beta_{A^c}^{\top})S\sqrt{n}((\beta_A-\beta_{0,A})^{\top},\beta_{A^c}^{\top})^{\top}\\
&\quad-n\sum_{j\in A^c}p_{\lambda_n}(|\beta_j|)+o_p(1). \qquad (5.6)
\end{aligned}$$

By the condition $|\beta_A-\beta_{0,A}|=O_p(n^{-1/2})$ and $|\beta_{A^c}|<cn^{-1/2}$, we have

$$\begin{aligned}
&\sqrt{n}((\beta_A-\beta_{0,A})^{\top},0^{\top})S\sqrt{n}((\beta_A-\beta_{0,A})^{\top},0)^{\top}=O_p(1),\\
&\sqrt{n}((\beta_A-\beta_{0,A})^{\top},\beta_{A^c}^{\top})S\sqrt{n}((\beta_A-\beta_{0,A})^{\top},\beta_{A^c})^{\top}=O_p(1)
\end{aligned}$$

and

$$\sqrt{n}((\beta_A-\beta_{0,A})^{\top},0^{\top})V_n-\sqrt{n}((\beta_A-\beta_{0,A})^{\top},\beta_{A^c}^{\top})V_n=-\sqrt{n}(0_A^{\top},\beta_{A^c}^{\top})V_n=O_p(1),$$

where the last step is based on the fact that $V_n\sim\mathrm{N}(0,4\tau(1-\tau)\Sigma)$ and the condition $|\beta_{A^c}|<cn^{-1/2}$. For any $\|\beta_A-\beta_{0,A}\|=O_p(n^{-1/2})$ and $0<\|\beta_{A^c}\|\le Cn^{-1/2}$, from (5.6) we have

$$Q_s((\beta_A^{\top},0_A^{\top})^{\top})-Q_s((\beta_A^{\top},\beta_{A^c}^{\top})^{\top})=O_p(1)-n\sum_{j=1}^{p}p_{\lambda}(|\beta_j|). \qquad (5.7)$$

Notice that

$$\begin{aligned}
n\sum_{j\in A^c}p_{\lambda_n}(|\beta_j|)&\ge n\sum_{j\in A^c}\Big[\lambda_n\lim_{\lambda_n\to 0}\inf\lim_{\beta\to 0+}\inf\frac{p'_{\lambda_n}(\beta)}{\lambda_n}\beta_j\mathrm{sign}(\beta_j)+o(|\beta_j|)\Big]\\
&=n\lambda_n\Big(\lim_{\lambda_n\to 0}\inf\lim_{\beta\to 0+}\inf\frac{p'_{\lambda_n}(\beta)}{\lambda_n}\Big)\Big(\sum_{j\in A^c}|\beta_j|(1+o(1))\Big)\\
&=n\lambda_n\sum_{j\in A^c}|\beta_j|(1+o(1)),
\end{aligned}$$

where the last step follows from the fact that

$$\lim_{\lambda_n\to 0}\inf\lim_{\beta\to 0+}\inf\frac{p'_{\lambda_n}(\beta)}{\lambda_n}=1.$$

Then $\sqrt{n}\lambda_n\to\infty$ implies in (5.7), the last term dominates in magnitude and as a result, $Q_s((\beta_A^{\top},0^{\top})^{\top})-Q_s((\beta_A^{\top},\beta_{A^c}^{\top})^{\top})<0$ for large n. □

Proof of Theorem 3.2 Part (a) holds simply due to Lemma 5.1. Next let's prove part (b). From the proof of Theorem 3.1, we see that $\sqrt{n}(\widehat{\beta}_A^{(S)} - \beta_{0,A})$ minimizes

$$\Phi_n\Big(\beta_0 + \frac{(u_A, 0)}{\sqrt{n}}\Big) + n\sum_{j\in A} p_\lambda\Big(\Big|\beta_{0,j} + \frac{u_j}{\sqrt{n}}\Big|\Big).$$

From (5.5), for large n,

$$n\sum_{j\in A} p_{\lambda_n}\Big(\Big|\beta_{0,j} + \frac{u_j}{\sqrt{n}}\Big|\Big) = n\sum_{j\in A} p_{\lambda_n}(|\beta_{0,j}|)$$

uniformly in any compact set of $R^{|A|}$. Hence we have

$$\Phi_n\Big(\beta_0 + \frac{(u_A, 0)}{\sqrt{n}}\Big) + n\sum_{j\in A} p_\lambda\Big(\Big|\beta_{0,j} + \frac{u_j}{\sqrt{n}}\Big|\Big) = u_A^{\mathsf{T}} V_A + \frac{1}{2} u_A^{\mathsf{T}} S_A u_A + n\sum_{j\in A} p_\lambda|\beta_{0,j}| + o_p(1).$$

We have the minimizer $\widehat{u}_A$ satisfies

$$\widehat{u}_A = -S_A^{-1} V_A + o_p(1).$$

Hence

$$\sqrt{n}(\widehat{\beta}_A - \beta_{0,A}) = \widehat{u}_A \to \mathrm{N}(0, \tau(1-\tau)\Delta_A^{-1}\Sigma_A\Delta_A^{-1}).$$

This completes the proof. □

References

[1] Alhamzawi, R. and Yu, K., Variable selection in quantile regression via Gibbs sampling, *Journal of Applied Statistics*, **39(4)**(2012), 799–813.

[2] Chaudhuri, P., Doksum, K. and Samarov, A., On average derivative quantile regression, *The Annals of Statistics*, **25(2)**(1997), 715–744.

[3] Chaudhuri, P., Nonparametric estimates of regression quantiles and their local Bahadur representation, *The Annals of Statistics*, **19(2)**(1991), 760–777.

[4] Fan, J. and Gijbels, I., *Local Polynomial Modeling and Its Applications*, Chapman & Hall, London, 1996.

[5] Fan, J. and Li, R., Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96(456)**(2001), 1348–1360.

[6] Hunter, D.R. and Li, R., Variable selection using MM algorithms, *The Annals of Statistics*, **33(4)** (2005), 1617–1642.

[7] Koenker, R., *Quantile Regression* (*Econometric Society Monographs*), Cambridge University Press, 2005.

[8] Koenker, R. and Bassett, G., Regression quantiles, *Econometrica*, **46(1)**(1978), 33–50.

[9] Naik, P.A. and Tsai, C., Single-index model selections, *Biometrika*, **88(3)**(2001), 821–832.

《应用概率统计》版权所有

[10] Stone, C.J., Consistent nonparametric regression, with discussion, *The Annals of Statistics*, **5(4)** (1977), 595–645.

[11] Tibshirani, R., Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B*, **58(1)**(1996), 267–288.

[12] Wu, Y. and Liu, Y., Variable selection in quantile regression, *Statistica Sinica*, **19(2)**(2009), 801–817.

[13] Wu, T.Z., Yu, K. and Yu, Y., Single-index quantile regression, *Journal of Multivariate Analysis*, **101(7)**(2010), 1607–1621.

[14] Xia, Y. and Härdle, W., Semi-parametric estimation of partially linear single-index models, *Journal of Multivariate Analysis*, **97(5)**(2006), 1162–1184.

[15] Yu, K. and Jones, M.C., Local linear quantile regression, *Journal of the American Statistical Association*, **93(441)**(1998), 228–237.

[16] Zhu, L., Huang, M. and Li, R., Semiparametric quantile regression with high-dimensional covariates, *Statistica Sinica*, **22(4)**(2012), 1379–1401.

[17] Zou, H., The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101(476)**(2006), 1418–1429.

《应用概率统计》版权所有

单指标分位数回归的变量选择

卢一强[1] 李锋[2] 胡斌[1]

([1]解放军信息工程大学, 郑州, 450000; [2]郑州大学数学与统计学院, 郑州, 450001)

多元非参数分位数回归常常是难于估计的, 为了降低维数同时保持非参数估计的灵活性, 人们常常用单指标的方法模拟响应变量的条件分位数. 本文主要研究单指标分位数回归的变量选择. 以最小化平均损失估计为基础, 我们通过最小化具有SCAD惩罚项的平均损失进行变量选择和参数估计. 在正则条件下, 得到了单指标分位数回归SCAD变量选择的Oracle性质, 给出了SCAD变量选择的计算方法, 并通过模拟研究说明了本文所提方法变量选择的样本性质.

关键词: 单指标模型, 分位数回归, SCAD, 变量选择.

学科分类号: O212.4.