DOI:10.3969/J.1SSN.100-5137.2015.03.005

Two-step variable selection in quantile regression models

FAN Yali

(College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: We propose a two-step variable selection procedure for high dimensional quantile regressions, in which the dimension of the covariates, p_n is much larger than the sample size n. In the first step, we perform ℓ_1 penalty, and we demonstrate that the first step penalized estimator with the LASSO penalty can reduce the model from an ultra-high dimensional to a model whose size has the same order as that of the true model, and the selected model can cover the true model. The second step excludes the remained irrelevant covariates by applying the adaptive LASSO penalty to the reduced model obtained from the first step. Under some regularity conditions, we show that our procedure enjoys the model selection consistency. We conduct a simulation study and a real data analysis to evaluate the finite sample performance of the proposed approach.

Key words: LASSO; adaptive LASSO; quantile regression; high dimensional

CLC number: O212.1 **Document code:** A **Article ID:** 1000-5137(2015)03-0270-14

MR (2000): 62H12 62H99.

1 Introduction

Quantile regression models have been widely applied in economics, medicine, survival analysis and many other different areas[1]. They are useful by providing much information about the conditional distribution of a response variable, and they are more robust against outliers compared to conditional mean regressions. In this paper, we are interested in variable selection and estimation in a quantile regression in high dimension settings.

Variable selection plays an important role in model building processes. It is common to include a large number of candidate predictor variables in the initial stage of modeling. However, it is undesirable to keep irrelevant predictors in the final model, since this makes it difficult to interpret the model and may decrease its predictive ability. In the last two decades, penalty based methods became popular for variable selection. For example, we saw the least absolute shrinkage and selection operator (LASSO,[2]), the smoothly clipped absolute deviation penalty (SCAD,[3]) and the adaptive LASSO (ALASSO, [4]), and their extensions in quantile regression, such as [5]. All these results focus on situations with fixed number of regressors. High or Ultra-high dimensional problem with number of regressors p_n being much larger than sample size n has attracted increased interest in recent years. Existing penalization methods for high-dimensional parametric least squares regression include Dantzig selector[6], the LASSO[7−8], the

Received date: 2014-09-03

Foundation item: National Natural Science Foundation of China (11401383).

Corresponding author: FAN Yali, lecturer, E-mail:yalifan@usst.edu.cn

SIS (the sure independence screening, [9]) etc.. [10] and [11] studied variable selection for high-dimensional linear quantile regression models via the LASSO and nonconvex penalties.

In this paper, we propose a two-step procedure for dimension reduction and variable selection in high-dimensional quantile regressions. In the first step, we perform ℓ_1 penalty and show that, the first step estimator with the LASSO penalty reduces the model from an ultra-high dimensional to the same size as that of the true model and having the sure screening property[9]. In the second step, the adaptive LASSO is applied to the reduced model for the data contained in that informative subset, to exclude the remained irrelevant covariates, and reduce the bias of ℓ_1-penalized estimator, leading to an estimator consistent in variable selection as well as relatively high efficiency.

The rest of the paper is organized as follows. In Section 2, we describe the two-step variable selection method, and provide the computational algorithm. In Section 3, we present the consistence of the proposed estimator in the first step and the variable selection consistency in second step. We assess the finite sample performance of the proposed method through an extensive simulation in Section 4, and through analysis of the real data set in Section 5. The proofs of the main results are provided in the Appendix.

2 Proposed variable selection method

Suppose that y_i is the ith response variable, $i = 1, \cdots, n$, at a given quantile level $0 < \tau < 1$. We consider the following quantile regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_0(\tau) + e_i, \tag{1}$$

where $(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \cdots, (y_n, \mathbf{x}_n)$ construct a random sample from $(y, \mathbf{x})$ for a response variable y and p_n-dimensional design vector $\mathbf{x}$ with support $\mathcal{R}_{\mathbf{x}} \subset \mathbb{R}^{p_n}$, $\boldsymbol{\beta}_0(\tau)$ is the unknown coefficient vector, and e_i is the random error whose τth conditional quantile given $\mathbf{x}_i$ equals zero. Throughout, we assume that e_i are independent of each other. Under this model, $\mathbf{x}_i^T \boldsymbol{\beta}_0(\tau)$ represents the τth conditional quantile of y_i given $\mathbf{x}_i$. We assume that there is sparsity in $\boldsymbol{\beta}_0(\tau)$, i.e., the number of truly relevant covariates is much smaller than p_n.

In variable selection problems, statistical efficiency criterion favors the use of the ℓ_0-penalty functions, but computational efficiency criterion favors the use of convex penalty functions, such as the LASSO. We choose the LASSO penalty because it is a convex function that is closest to the ℓ_0-penalty[12]. [10] studied variable selection for high-dimensional linear quantile regression and proposed a pivotal, data-driven choice of the regularization parameter. They showed that the LASSO estimator is consistent at a rate that is close to the oracle rate.

However, it is well known that applying the ℓ_1 penalty tends to overpenalize large coefficients and to include inactive variables in the selected model and to introduce bias (Wang, Wu and Li[11] (2012)). Although [10] showed that a post-ℓ_1-quantile regression procedure can further reduce the bias, post-L_1-quantile does not possess the oracle property in general.

To reduce the bias of the LASSO estimator and achieve variable selection consistence, we propose our two-step variable selection procedure as follows.

Step 1 To proceed dimension reduction, we propose to estimate $\boldsymbol{\beta}_0(\tau)$ by minimizing

$$L_{1,n}(\boldsymbol{\beta}(\tau)) = n^{-1} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}(\tau)) + \frac{\lambda_{1,n}}{n} \sum_{j=1}^{p_n} |\beta_j|, \tag{2}$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ is the check function, and $\lambda_{1,n}$ is the positive regularization parameter. We writhe the estimator in this step as $\tilde{\boldsymbol{\beta}}(\tau) = (\tilde{\beta}_1(\tau), \cdots, \tilde{\beta}_{p_n}(\tau))$.

Step 2 To reduce the bias of the ℓ_1- penalized estimator in step 1, and to proceed variable selection and estimation, we apply the adaptive LASSO penalty. Define $\tilde{S} = \{j : \tilde{\beta}_j(\tau) \neq 0\}$, and $\mathbf{x}_{i,\tilde{S}}$ and $\boldsymbol{\beta}(\tau)_{\tilde{S}}$ as the corresponding sub design vector and coefficient vector. We proceed to variable selection by minimizing

$$L_{2,n}(\boldsymbol{\beta}(\tau)) = Q_n(\boldsymbol{\beta}(\tau)_{\tilde{S}}) + \frac{\lambda_{2,n}}{n}\sum_{j\in\tilde{S}}\omega_j|\beta_j|, \tag{3}$$

where $Q_n(\boldsymbol{\beta}(\tau)_{\tilde{S}}) = n^{-1}\sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_{i,\tilde{S}}^T\boldsymbol{\beta}(\tau)_{\tilde{S}})$, and $\omega_j = |\tilde{\beta}_j|^{-2}$ is the adaptive weight.

In step 1, we borrow the methods in [10]. The dimension reduction in step 1 enjoys computation and theoretical readiness. In step 2, the choice of the adaptive LASSO penalty can significantly reduce the bias of the LASSO estimator and enjoys variable selection consistence.

The algorithms of Step 1 and Step 2 are the same as those of the interior-point method and dual method in penalized quantile regression as in [10].

3 Asymptotic properties

In this section, we discuss the asymptotic properties of the proposed two-step estimator as $n \to \infty$. Without loss of generality, we assume that $\boldsymbol{\beta}_{0,j}(\tau) \neq 0, j = 1, \cdots, s$, and $\boldsymbol{\beta}_{0,j}(\tau) \equiv 0, j = s+1, \cdots, p$. For convenience, we use $a \lesssim b$ to denote $a = O(b)$, and $a \lesssim_p b$ to denote $a = O_p(b)$. We express $\max(a,b)$ and $\min(a,b)$ by $a \vee b$ and $a \wedge b$. We use the notation $\|\cdot\|, \|\cdot\|_1, \|\cdot\|_0, \|\cdot\|_\infty$ to denote ℓ_2 norm, ℓ_1 norm, ℓ_0 norm and ℓ_∞ norm respectively, where ℓ_0 norm is the number of nonzero components in the vector. To give the asymptotic properties of the proposed estimator, we need the following assumptions.

C1. Data $(y_i, \mathbf{x}_i^T)^T, i = 1, 2, \cdots, n$ are i.i.d. copies of $(y, \mathbf{x}^T)^T$. The τth quantile of $e_i(\tau)$ conditional on $\mathbf{x}_i$ equals zero. The first component of $\mathbf{x}_i$ equals 1, and for $i = 1, \cdots, n$, $\| \mathbf{x}_i \|_\infty$ is bounded in probability.

C2. Assume that $\|\beta_0(\tau)\|_0 = s$, and $s\sqrt{\log(n \vee p)} = o(n^{3/20})$.

C3. Suppose $\mathbf{x}_i$ is absolutely continuous and y_i is absolutely continuous conditional on $\mathbf{x}_i$, for $i = 1, \cdots, n$. Let $f_{y_i|\mathbf{x}_i}(\cdot)$ denote the conditional density function of y_i given $\mathbf{x}_i$. There exist three positive constants, $\overline{f}, \underline{f}, \overline{f'}$, such that $f_{y_i|\mathbf{x}_i}(y|\mathbf{x}) \leqslant \overline{f}$, $|\frac{\partial}{\partial y} f_{y_i|\mathbf{x}_i}(y|\mathbf{x})| \leqslant \overline{f'}$ and $f_{y_i|\mathbf{x}_i}(\mathbf{x}^T\boldsymbol{\beta}_0(\tau)|\mathbf{x}) \geqslant \underline{f}$ uniformly in y and $\mathbf{x}$ over supports of y and $\mathbf{x}$, and uniformly in n.

C4. Define the k-sparse unit sphere in $\mathbb{R}^p$ as

$$\mathbb{B}_k^p = \{\alpha \in \mathbb{R}^p : \|\alpha\| = 1, \|\alpha\|_0 \leqslant k\}, 1 \leqslant k \leqslant n. \tag{4}$$

The eigenvalues of the design matrix $E(\mathbf{x}_i\mathbf{x}_i^T)$ are bounded by $0 < c_0 < \lambda_{min} \leqslant \lambda_{max} < c_1$ for some finite positive c_0, c_1. Assume that for $\alpha \in \mathbb{B}_k^p, k \leqslant n$, we have $c_2 \leqslant E(|\alpha^T\mathbf{x}_i|^2) \leqslant c_3$, $E(|\alpha^T\mathbf{x}_i|^3) \leqslant c_4$ for some finite positive constants c_2, c_3, c_4. Define $\varphi(k) = \sup_{\alpha\in\mathbb{B}_k^p} E[|\alpha^T\mathbf{x}_i|^2]$ and $\phi(k) = \sup_{\alpha\in\mathbb{B}_k^p} \frac{1}{n}\sum_{i=1}^n(\alpha^T\mathbf{x}_i)^2$. We assume that $\varphi(k) = O_p(\phi(k))$ for $1 \leqslant k \leqslant n$.

C5. Assume that the non-zero components of $\boldsymbol{\beta}_0(\tau)$ satisfy

$$\min_{1\leqslant j\leqslant s}|\beta_{0,j}(\tau)| > t_{n1}t_{n2}\sqrt{\frac{s\log(n\vee p)\phi(m_0+s)}{n}}\frac{\mu}{q^2},$$

for some diverging sequence of positive constants t_{n1}, t_{n2}, where μ, q satisfy $q = q(n) = \frac{\underline{f}\varrho(n)}{4}\{1 \wedge (\frac{3\underline{f}}{2\overline{f'}}\gamma(n))\} = O_p(1)$, and $\mu = \mu(n) = K(\varphi(n)\overline{f} + \sqrt{\varphi(n)}) = O_p(1)$. For the specific definitions of the constant K, $\varrho(n)$ and $\gamma(n)$, refer to the appendix.

The conditions C1-C5 ensure that Lemma 1-Lemma 4 in the appendix and Theorem 1 hold, which allow us to obtain the consistence and the converge rate of the estimators. The condition C1 imposes random sampling on the data. The condition C2 requires that the true model is sparse. The condition C3 imposes some smoothness on the conditional distribution of response variable, which is more relaxed than the Gaussian or sub-Gaussian error condition usually assumed in the ultra-high dimensional regression literature. The condition C4 requires the design matrix to be uniformly non-singular and the regressors' moments to be well-behaved and the condition C5 requires that the smallest signal should not decay too fast.

The following Theorem presents the properties of the first step estimator.

Theorem 1 Assume that the conditions C1-C5 hold. Define $m_0 = p_n \wedge \Big(\frac{n}{\log(n \vee p_n)} \frac{q^2}{\mu^2}\Big)$, and set $\lambda_{1,n} = t_{n1}\sqrt{n \log(n \vee p_n)\phi(m_0 + s)}\frac{\mu}{q}$, where t_{n1} is the diverging sequence of positive constants in C5. Assume that $\frac{\lambda_{1,n}\sqrt{s}}{qn} \to 0$. Then we have that

$$\|\tilde{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)\| \lesssim_p \frac{\lambda_{1,n}\sqrt{s}}{qn}, \tag{5}$$

$$\|\tilde{\boldsymbol{\beta}}(\tau)\|_0 \lesssim_p \Big(\frac{\mu}{q}\Big)^2 s, \tag{6}$$

and

$$P\{\tilde{S} \supset \{1, 2, \cdots, s\}\} \to 1. \tag{7}$$

Theorem 1 indicates that our proposed estimator $\tilde{\boldsymbol{\beta}}(\tau)$ is a consistent estimator of $\boldsymbol{\beta}_0(\tau)$, and after the first step of variable selection, we can reduce the dimension of the model to the same stochastic order as that of the true model. If the non-zero components of $\boldsymbol{\beta}_0(\tau)$ are well separated from zero, the true model is included with probability approaching one. Note that p_n affects the convergence rate only through the $\log(p_n)$ factor. Thus we allow the number of covariates to grow nearly exponentially in the sample size.

From Theorem 1, we can see that the support of the estimator $\tilde{\boldsymbol{\beta}}(\tau)$ may include some unnecessary components with the true coefficients equal to zero. The following Theorem 2 presents the asymptotic properties of the second step estimator, including the consistency for variable selection and estimation.

Theorem 2 Let $\hat{\boldsymbol{\beta}}(\tau)$ denote the minimizer of (3), $\tilde{S}$ be the selected model from step 1. Suppose the conditions in Theorem 1 hold. We set $\lambda_{2,n} = t_{n3}\sqrt{ns \log(n \vee p_n)\phi(s)}$, where t_{n3} is a diverging sequence of positive numbers. We assume that $\frac{\lambda_{2,n}\sqrt{s}}{qn} \to 0$, $\sqrt{\frac{s \log(n \vee p_n)\phi(s)}{n}}\frac{1}{q} \to 0$, and $\lambda_{2,n} n^{-\frac{1}{2}} (\log(p_n))^{-\frac{1}{2}} \to \infty$. Then with probability converging to one, we have

- $\hat{\beta}_j = 0, j \geqslant s + 1$;
- $\| \hat{\boldsymbol{\beta}}_1(\tau) - \boldsymbol{\beta}_{10}(\tau) \| \lesssim_p \sqrt{\frac{s \log(n \vee p_n)\phi(s)}{n}}\frac{1}{q}$,

where $\hat{\boldsymbol{\beta}}_1(\tau), \boldsymbol{\beta}_{10}(\tau)$ are the first s components of $\hat{\boldsymbol{\beta}}_1(\tau), \boldsymbol{\beta}_0(\tau)$ respectively.

Theorem 2 shows that by choosing a proper $\lambda_{2,n}$, the adaptive LASSO penalized quantile regression enjoys variable selection oracle properties, and the rate of convergence of the nonzero part is generally faster than the rate of the estimator in step 1. It can also be seen from Theorem 2 that in order to achieve variable selection consistence, we need a more heavier penalty in step 2. For example, we can choose $t_{n3} \leqslant n^{1/4}$ in $\lambda_{2,n}$.

4 Numerical Study

4.1 Tuning methods

To proceed to dimension reduction and variable selection, we need to select the tuning parameters $\lambda_{1,n}$ in the first step, and $\lambda_{2,n}$ in the second step. For $\lambda_{1,n}$, we follow the similar idea as is given in [10] and set $\lambda_{1,n}$ as the $1-\alpha_n$ quantile of Λ_n

$$\lambda_{1,n} = \inf\{c_n : P(\Lambda_n \leqslant c_n | \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) \geqslant 1-\alpha_n\},$$

where

$$\Lambda_n = n\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i(\tau - I\{u_i - \tau < 0\})\|_\infty,$$

$u_i, i = 1, \cdots, n$ are i.i.d. uniform $(0,1)$ random variables, independently distributed from the regressors, and α_n goes to zero at some rate, see [10].

In Step 2, we choose $\lambda_{2,n}$ by the minimizer of the extended Bayesian information criterion (EBIC, [13]), which is defined as

$$\mathrm{EBIC}(\lambda) = \log \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_{i,\tilde{S}}^T \beta(\tau)_{\tilde{S}}) I\{\hat{\pi}_i > 1-\tau+c_n\} + \frac{\log n}{2n} edf_\lambda + c\frac{\log p_n}{2n} edf_\lambda,$$

where edf_λ is defined as the number of interpolated $y_i's$, which gives a plausible estimation of the effective degree of freedom in quantile regression, and $0 \leqslant c \leqslant 1$ is a constant, and we use $c = 0.5$ in this paper.

4.2 Simulation examples

We study two cases to investigate the performance of the proposed penalized estimator. The sample size is $n = 300, 500$ and the dimension of covariates is $p_n = 1000$. We focus on three quantile levels $\tau = 0.25, \tau = 0.5$ and $\tau = 0.75$. For each case, the simulation is repeated 200 times.

Case 1 The data are generated from the following models

$$\begin{cases} y_i^* = \beta_0 + x_{i1} + x_{i2} + x_{i3} + x_{i4} + x_{i5} + e_i, \\ y_i = \max(y_i^*, 0) \end{cases}, \tag{8}$$

where $e_i \sim N(0,1)$, thus $\beta_0(\tau) = \beta_0 + \Phi^{-1}(\tau)$. Besides the 5 relevant covariates, we include rest 994 independent noise variables and $(x_{i2}, \cdots, x_{i1000})$ are generated from the multivariate normal distribution $N(\mathbf{0}, \mathbf{V})$, with $V_{ij} = 0.5^{|i-j|}$.

We compare the following estimators. The ″Step1″ and ″Step2″ are proposed estimators obtained in step 1 and step 2.

Table 1 summarizes the variable selection and parameter estimation results from all the methods. The average number of relevant (irrelevant) variables that are correctly (incorrectly) selected, denoted by ″TP″ (″FP″), and the percentage of times that the true model is correctly selected, denoted by ″OP″, are reported in Table 1. The median of absolute estimation error (MAE), defined by the median of $\sum_{j=1}^{p} |\hat{\beta}_j - \beta_{0,j}|$, is also reported in Table 1 to assess estimation accuracy.

As is expected, the ″Step1″ estimator can reduce the model from one with ultra-high dimension to a model that contains the true model as a valid sub model. This overfitting phenomenon in ″Step1″ has been greatly improved in ″Step2″, which performs further variable selection and has high oracle proportions in all scenarios considered. As

to parameter estimation, we can see from Table 1 that, the ″Step2″ has absolute estimation errors that are clearly smaller than those of ″Step1″.

Case 2 In this case, the covariates are generated in the same way as that in Case 1, but the random errors are generated from a heavier tailed distribution, the t-distribution with 3 degrees of freedom. Although the absolute estimate errors are a little higher than those under the normal errors, the variable select results are similar, which verifies that the proposed variable select method is robust to heavy tails.

5 Real data analysis

In this section, we use the proposed two stage variable selection method to analyze the Boston House Price data set, which is available online at http://lib.stat.cmu.edu/datasets/bosten-corrected.txt. There are 506 observations in this data set, including 15 predictor variables and one response variable: corrected median value of owner-occupied homes in $\$1000's$ (CMEDV). Predictors include the longitude (LON), latitude (LAT), proportion of area zoned with large lots (ZN), crime rate (CRIM), proportion of non-retail business acres per town (INDUS), Charles River as a dummy variable (1 if tract bounds river and 0 otherwise) (CHAS), nitric oxides concentration (NOX), average number of rooms per dwelling (RM), proportion of owner-occupied units built prior to 1940 (AGE), weighted distances to five Boston employment centres (DIS), index of accessibility to radial highways (RAD), property tax rate (TAX), pupil/teacher ratio by town (PTRAT), black population proportion town (B), and the lower status population proportion (LSTAT). One feature of this data set is that the response variable is the median price of a home in a given area, and the distributions of the response are left skewed, which can been seen from the histogram of the response. Therefore, quantile methods are particularly suited to the analysis of this dataset. There are several papers devoted to the analysis of this dataset in the literature. [14] employed the additive quantile technique to analyze the data. [15] used a functional coefficient quantile regression model to fit this dataset. Several authors have analyzed this dataset using varying coefficient models. See, for example, [16], [17] and so on.

To capture the underlying relationship between the predictors and the entire conditional distribution, we employ the proposed penalized quantile regression model on the $\tau = 0.25, 0.5, 0.75$ quantile for this dataset. For simplicity, we excluded the categorical variable RAD. We use 14 standardized predictor variables and 13 predictor's squares (aside from CHAS) to incorporate some potential nonlinear components. Besides these 27 original covariates and intercept, we include $p_n - 28$ artificial variables in the full model and consider the following model

$$y_i^* = \beta_0(\tau) + \sum_{j=1}^{p_n} x_{ij}\beta_j(\tau) + e_i, \tag{9}$$

where the artificial variables are independent from the original covariates and distributed from $N(0, \Sigma)$, with $\Sigma_{ij} = 0.5^{|i-j|}$, and the response data y_i is the standardized CMEDV. We consider three different model sizes, $p_n = 50, 200, 500$ and choose $\lambda_{1,n}$ as the $0.9-$th conditional quantile of the pivotal Λ_n given $\mathbf{x}_i$.

Then we conduct 50 random partitions. For each partition, we randomly split all the 506 observations into training and testing data sets of size 406 and 100 respectively. A 5-fold cross-validation is applied to the training data to select the tuning parameter $\lambda_{2,n}$, and we evaluate the performance over the test set for each partition.

The results in Table 3 suggest that the proposed two-stage procedure performs quite well in variable selection, since for each model size, even the large one, the artificial variables are not included in the final model.

As with every variable selection method, different repetitions may select different subsets of important predictors. In the following Tables 4–6, we report the frequency with that the important variables appear in the final model

of the 50 random partitions for $\tau = 0.25, 0.5, 0.75$ quantile respectively. The variables are ordered such that the frequency is decreasing.

From Tables 4,5,6, it is observed that some variables, such as the LSTAT, RM, PTRAT, RM^2 , B^2, have high frequencies across different quantiles, while some other variables such as the ZN^2, $INDUS^2$, NOX^2 do not. This implies that some variables are important across all quantiles, while some variables might be important only for certain quantiles. For further convince, we also computed the correlation coefficients between the response variable and the predictors, and we find that the first three highest correlated predictors are the LSTAT (-0.74), RM(0.70) and PTRAT (-0.51), which confirms our results.

References:

[1] KOENKER R. Quantile regression[M]. Cambridge: Cambridge University Press, 2005.

[2] TIBSHIRANI R J. Regression shrinkage and selection via the LASSO[J]. Journal of the Royal Statistical Society B, 1996, 58: 267-288.

[3] FAN J Q, LI R Z . Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96: 1348-1360.

[4] ZOU H. The adaptive LASSO and its oracle properties[J]. Journal of American Statistical Association, 2006, 101: 1418-1429.

[5] WU Y C, LIU Y. Variable selection in quantile regression[J]. Statistic Sinica, 2009, 19: 801-817.

[6] CANDES E, TAO T. The Dantzig selector: Statistical estimation when p is much larger than n[J]. The Annals of Statistics, 2007, 35: 2313-2351.

[7] ZHANG C, HUANG J. The sparsity and bias of the LASSO selection in high-dimensional linear regression[J]. The Annals of Statistics, 2008, 36: 1567-1594.

[8] MEINSHAUSEN N, YU B. LASSO-type recovery of sparse representations for high-dimensional data[J]. The Annals of Statistics,2009, 37: 246-270.

[9] FAN J Q, LV J . Sure independence screening for ultrahigh dimensional feature space[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology),2008, 70: 849-911.

[10] BELLONI A, CHERNOZHUKOV V . ℓ1-Penalized quantile regression in high dimensional sparse models[J]. Annals of Statistics, 2011, 39:82-130.

[11] WANG L, WU Y, LI R Z . Quantile regression for analyzing heterogeneity in ultra-high dimension[J]. Journal of the American Statistical Association, 2012, 107: 214-222.

[12] MEINSHAUSEN N, BUHLAMAN P. High dimensional graphs and variable selection with the LASSO.[J]. The Annals of Statistics, 2006, 34: 1436-1462.

[13] CHEN J, CHEN Z. Extended Bayesian information criteria for model selection with large model space[J]. Biometrika, 2008, 95: 759-771.

[14] YU K, LU Z. Local linear additive quantile regression[J]. Scandinavian Journal of Statistics, 2004, 31: 333-346.

[15] CAI Z, XU X . Nonparametric quantile estimations for dynamic smooth coefficient models[J]. Journal of the American Statistical Association, 2008, 103: 1595-1608.

[16] FAN J, HUANG T. Profile likelihood inferences on semiparametric varying-coefficient partially linear models[J]. Bernoulli, 2005, 11: 1031-1057.

[17] SENTURK D, MULLER H . Inference for covariate adjusted regression via varying coefficient models[J]. The Annals of Statistics, 2006, 34: 654-679.

Appendix: Sketch of proofs

First, we give some useful symbols and lemmas. Let $\|\tilde{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}_0(\tau)\|=r$, $\|\tilde{\boldsymbol{\beta}}(\tau)\|_0=m$, and

$$\mathbb{R}(p,r,m)=\{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p:\|\boldsymbol{\beta}(\tau)\|_0\leqslant m,\|\boldsymbol{\beta}(\tau)-\boldsymbol{\beta}_0(\tau)\|_0\leqslant r\}.$$

We define

$$\varrho(k)=\inf_{b\in\mathbb{B}_k^p}E[(\mathbf{x}^Tb)^2],\ \ \vartheta(k)=\inf_{b\in\mathbb{B}_k^p}\frac{E[|\mathbf{x}^Tb|^2]}{E[|\mathbf{x}^Tb|^3]}.$$

Lemma 1 gives an identification of the true sparse coefficient vector. Lemma 2 and lemma 3 show the sparseness characterizations of the solution in step 1. Lemma 4 presents the maximal inequalities applied to the empirical criterion function. These lemmas are similar to those given by Belloni and Chernozhukov (2011). The proof techniques also are similar, thus are omitted here.

Lemma 1 Suppose that the conditions C1-C5 hold. Let $Q_n^*(\boldsymbol{\beta}(\tau),\pi)=E(\rho_\tau(y-\mathbf{x}^T\boldsymbol{\beta}(\tau))$, and suppose that $\|\boldsymbol{\beta}(\tau)-\boldsymbol{\beta}_0(\tau)\|_0\leqslant n$. Then we have

$$Q_n^*(\boldsymbol{\beta}(\tau))-Q_n^*(\boldsymbol{\beta}_0(\tau))\geqslant q(\min(r^2,r)),$$

where $q=q(n)\lesssim 1$.

Lemma 2 Suppose that the conditions C1-C5 hold. For any $\lambda_{1,n}=\lambda>0$, we have

$$m\leqslant\min(n,p_n,\frac{n\phi(m)}{\lambda^2})$$

with probability one.

Lemma 3 Suppose that the conditions C1-C5 hold. We have that

$$\sqrt{m}\lesssim\frac{1}{\lambda}n\min(r\mu+C,\mu)+\frac{1}{\lambda}\sqrt{nm\log(p_n\vee n)\phi(m)},$$

where $\mu=\mu(n)=K(\varphi(n)\overline{f}+\sqrt{\varphi(n)})=O_p(1)$.

Lemma 4 Suppose that the conditions C1-C5 hold. We have that uniformly in $m\leqslant n$, and over region $\mathbb{R}(p,r,m)$,

$$|\,Q_n(\boldsymbol{\beta}(\tau))-Q_n^*(\boldsymbol{\beta}(\tau))-(Q_n(\boldsymbol{\beta}_0(\tau))-Q_n^*(\boldsymbol{\beta}_0(\tau)))\,|\lesssim_p \frac{r}{\sqrt{n}}\sqrt{(m+s)\log(p_n\vee n)\phi(m+s)}.$$

Proof of Lemma 4 is similar to the proof of Lemma 5 in Belloni and Chernozhukov (2011).

Proof of Theorem 1: Firstly, we show that $m\leqslant m_0=(\frac{n}{\log(n\vee p_n)}\frac{q^2}{\mu^2})$ hold with probability converging to one. By lemma 2, we have

$$m\leqslant\tilde{m}=\max\{m:m\leqslant n\wedge p_n\wedge\frac{n^2\phi(m)}{\lambda^2}\}.$$

Suppose that $\tilde{m}\geqslant m_0$. Then there exists a $k>1$ such that $\tilde{m}=m_0k$. By definition, $\tilde{m}$ satisfies

$$\tilde{m}\leqslant\frac{n\phi(\tilde{m})}{\lambda^2}.$$

Note that $\lambda \geqslant t_{n1}\sqrt{n\log(n\vee p_n)\phi(m_0)}\dfrac{\mu}{q}$ since $\phi(m_0) \leqslant \phi(m_0+s)$, by Lemma 11 of Belloni et al.(2011), we have

$$\begin{aligned}\tilde{m} \quad &\leqslant \quad \frac{n\phi(m_0k)}{t_{n1}^2 n\log(n\vee p_n)\phi(m_0)}\frac{q^2}{\mu^2} = \frac{1}{t_{n1}^2\log(n\vee p_n)}\frac{\phi(m_0k)}{\phi(m_0)}\frac{q^2}{\mu^2} < \\ &\frac{1}{t_{n1}^2\log(n\vee p_n)}2k\frac{q^2}{\mu^2} < \frac{n}{t_{n1}^2\log(n\vee p_n)}2k\frac{q^2}{\mu^2} = \frac{2}{t_{n1}^2}m_0k, \end{aligned} \tag{10}$$

provided $t_{n1} \geqslant \sqrt{2}$. Then we have $\tilde{m} = m_0k < m_0k$, which is a contradiction. Since $t_{n1} \to_p \infty$, we get that with probability converging to one, we have $m \leqslant m_0$.

Secondly, we show that

$$q(r^2 \wedge r) \lesssim_p \frac{\lambda}{n}\sqrt{s}r + r\sqrt{\frac{(m+s)\log(n\vee p_n)\phi(m+s)}{n}}.$$

By the condition C1, the support of $\boldsymbol{\beta}_0(\tau)$ has exactly s non-zero elements. Let S denote the support of $\boldsymbol{\beta}_0(\tau)$ and $\tilde{\boldsymbol{\beta}}_S(\tau)$ denote a vector whose s ($s \in S$) component agrees with s component of $\tilde{\boldsymbol{\beta}}(\tau)$ and the remaining components are equal to zero.

Since $\|\tilde{\boldsymbol{\beta}}_S(\tau)\|_1 \leqslant \|\tilde{\boldsymbol{\beta}}(\tau)\|_1$ and by definition of $\tilde{\boldsymbol{\beta}}(\tau)$, we have that

$$\begin{aligned}&Q_n(\tilde{\boldsymbol{\beta}}(\tau)) - Q_n(\boldsymbol{\beta}_0(\tau)) \leqslant \frac{\lambda}{n}(\|\boldsymbol{\beta}_0(\tau)\|_1 - \|\tilde{\boldsymbol{\beta}}(\tau)\|_1) \leqslant \\ &\frac{\lambda}{n}(\|\boldsymbol{\beta}_0(\tau)\|_1 - \|\tilde{\boldsymbol{\beta}}_S(\tau)\|_1) \leqslant \frac{\lambda}{n}(\|\boldsymbol{\beta}_0(\tau) - \tilde{\boldsymbol{\beta}}_S(\tau)\|_1) \leqslant \\ &\frac{\lambda}{n}\sqrt{|S|}\|\tilde{\boldsymbol{\beta}}_S(\tau) - \boldsymbol{\beta}_0(\tau)\| \leqslant \frac{\lambda}{n}\sqrt{s}r,\end{aligned}$$

where $|S|$ denotes the number of components in set S. Applying Lemma 4, we further get that

$$Q_n^*(\tilde{\boldsymbol{\beta}}(\tau)) - Q_n^*(\boldsymbol{\beta}_0(\tau)) \lesssim_p \frac{\lambda}{n}\sqrt{s}r + r\sqrt{\frac{(m+s)\log(n\vee p_n)\phi(m+s)}{n}}.$$

Invoking the result of Lemma 1, we obtain

$$q(r^2 \wedge r) \lesssim_p \frac{\lambda}{n}\sqrt{s}r + r\sqrt{\frac{(m+s)\log(n\vee p)\phi(m+s)}{n}}.$$

Thirdly, we show the consistency of $\tilde{\boldsymbol{\beta}}(\tau)$, namely, $r = o_p(1)$. The construction of λ, $t_{n1} \to_p \infty$, and the condition $\dfrac{\lambda\sqrt{s}}{qn} \to_p 0$ imply that

$$\begin{aligned}&(1) \quad \frac{\lambda\sqrt{s}}{n} = o_p(q); \\ &(2) \quad \sqrt{\frac{s\log(n\vee p)\phi(m_0+s)}{n}}\frac{\mu}{q} = o_p(q); \\ &(3) \quad \mu\frac{\sqrt{n\log(p_n\vee n)\phi(m_0+s)}}{\lambda} = o_p(q).\end{aligned} \tag{11}$$

By Lemma 3 and condition (3) and noting that $\mu \geqslant q$ by their definition, we have

$$\sqrt{m} \lesssim_p \frac{n}{\lambda}\mu + \sqrt{m}o_p(1),$$

which implies

$$\sqrt{m} \lesssim_p \frac{n\mu}{\lambda}. \tag{12}$$

Using (12) in (10) and noting $m \leqslant m_0$ with probability converging to one, we have that

$$\begin{aligned} q(r^2 \wedge r) &= I\{m > s\}q(r^2 \wedge r) + I\{m \leqslant s\}q(r^2 \wedge r) \lesssim_p \\ &\frac{\lambda}{n}\sqrt{s}r + r\frac{\sqrt{m\log(n \vee p_n)\phi(m_0+s)}\mu}{\lambda} + \\ &\frac{\lambda}{n}\sqrt{s}r + r\sqrt{\frac{s\log(n \vee p_n)\phi(m_0+s)}{n}} = ro_p(q) + ro_p(q) = ro_p(q). \end{aligned} \tag{13}$$

Dividing both sides of (13) by q and by r, we have $r = o_p(1)$.

Lastly, we derive the rate of convergence. Since $r = o_p(1)$, we improve the bound (12) to the following bound

$$\sqrt{m} \lesssim_p \frac{r\mu n}{\lambda}.$$

Plugging (13) into (10) and noting $r^2 = o_p(r)$, we have

$$qr^2 \lesssim_p r\frac{\lambda\sqrt{s}}{n} + r^2 o_p(q),$$

or equivalently,

$$r \lesssim_p \frac{\lambda\sqrt{s}}{nq}. \tag{14}$$

Inserting (14) into (13), we obtain

$$\sqrt{m} \lesssim_p \frac{\lambda\sqrt{s}}{nq}\frac{\mu n}{\lambda} = \frac{\mu}{q}\sqrt{s}.$$

Thus $m \lesssim_p \left(\frac{\mu}{q}\right)^2 s$. Finally, by the inequality $\|\hat{\beta}(\tau) - \beta(\tau)\|_\infty \leqslant \|\hat{\beta}(\tau) - \beta(\tau)\|$, with probability going to one, we have

$$\begin{aligned} \|\hat{\beta}(\tau) - \beta(\tau)\|_\infty &\leqslant \|\hat{\beta}(\tau) - \beta(\tau)\| \lesssim_p t_{n1}\sqrt{\frac{s\log(n \vee p_n)\phi(m_0+s)}{n}}\frac{\mu}{q^2} \leqslant \\ &t_{n2}t_{n1}\sqrt{\frac{s\log(n \vee p_n)\phi(m_0+s)}{n}}\frac{\mu}{q^2} < \min_{i \in support(\beta(\tau))}|\beta_i(\tau)|, \end{aligned} \tag{15}$$

If

$$support(\beta(\tau)) \supseteq support(\hat{\beta}(\tau)), \tag{16}$$

then

$$\|\hat{\beta}(\tau) - \beta(\tau)\|_\infty \geqslant \min_{i \in support(\beta(\tau))}|\beta_i(\tau)|, \tag{17}$$

Since (17) can occur only with probability approaching zero, we conclude that (16) occurs with probability converging to one.

Proof of Theorem 2 We first consider the minimizer of $L_{2,n}(\beta(\tau))$ in the subspace

$$\mathbb{A} = \{\beta(\tau) = (\beta_1^T(\tau), \mathbf{0}^T)^T \in R^{|\tilde{S}|} : \beta_1^T(\tau) \in R^s\}.$$

We use the symbol $w = (\omega_1, \cdots, \omega_{|\tilde{S}|})^T = (w_1^T, w_2^T)^T$, where $w_1 = (\omega_1, \cdots, \omega_s)^T$. For

$$\hat{\beta}(\tau) = argmin_{\beta(\tau) \in \mathbb{A}} L_{2,n}(\beta(\tau)),$$

we write $\| \hat{\beta}(\tau) - \beta_0(\tau) \| = \hat{r}$. Applying Lemma 4 and Lemma 1, and noting that $|\tilde{S}| \lesssim s$ implies that $\phi(|\tilde{S}| + s) \lesssim \phi(s)$ (by lemma 13 given by Belloni and Chernozhukov (2011)), we obtain that

$$\begin{aligned} q\min(\hat{r}^2, \hat{r}) \quad \lesssim \quad & Q_n^*(\hat{\beta}(\tau)) - Q_n^*(\beta_0(\tau)) = \\ & Q_n(\beta_0(\tau)) - Q_n^*(\beta_0(\tau)) - (Q_n(\hat{\beta}(\tau)) - Q_n^*(\hat{\beta}(\tau))) + \\ & L_{2,n}(\hat{\beta}(\tau)) - L_{2,n}(\beta_0(\tau)) + \frac{\lambda_{2,n}}{n} \| w_1 \circ (\hat{\beta}(\tau) - \beta_0(\tau)) \|_1 \lesssim_p \\ & \hat{r}\sqrt{\frac{s\log(n \vee p_n)\phi(s)}{n}} + \frac{\lambda_{2,n}}{n} \| w_1 \circ (\hat{\beta}(\tau) - \beta_0(\tau)) \|_1 \lesssim_p \\ & \hat{r}\sqrt{\frac{s\log(n \vee p_n)\phi(s)}{n}} + \hat{r}\frac{\lambda_{2,n}}{n} \| w_1 \|, \end{aligned} \tag{18}$$

where $\circ$ is the Hadamard product. By the assumptions of Theorem 2, (18) can be reduced to

$$\min(\hat{r}, 1) \lesssim_p \sqrt{\frac{s\log(n \vee p_n)\phi(s)}{n}}\frac{1}{q} + \frac{\lambda_{2,n}\sqrt{s}}{qn}.$$

Since $\frac{\lambda_{2,n}\sqrt{s}}{qn} \to 0$, we have $\hat{r} = O_p(\sqrt{\frac{s\log(n \vee p_n)\phi(s)}{n}}\frac{1}{q}) = o_p(1)$.

Next, we prove that $\hat{\beta}(\tau)$ is a global minimizer of $L_{2,n}(\beta(\tau))$ in the $R^{|\tilde{S}|}$space. Let $\mathbf{X}_{\tilde{S}} = (\mathbf{x}_{1,\tilde{S}}^T, \mathbf{x}_{2,\tilde{S}}^T, \cdots, \mathbf{x}_{n,\tilde{S}}^T)^T = (\mathbf{X}_1, \mathbf{X}_2)$, and $\mathbf{X}_1$ be the submatrix corresponding to non-zero coefficients. By the convexity of the $L_{2,n}(\beta(\tau))$, we only check that the following Karush-Kuhn-Tucker (KKT) condition holds for $\hat{\beta}(\tau)$,

$$\| w_2^{-1} \circ \mathbf{X}_2^T(\tau - I\{\mathbf{y} - \mathbf{X}_{\tilde{S}}\beta(\tau) < 0\}) \|_\infty \leqslant \lambda_{2,n}. \tag{19}$$

Similar to Lemma 3, we set

$$a_i^*(\tau) = (\tau - I\{y_i - \mathbf{x}_i^T\beta_0(\tau) < 0\}), \quad a_i(\tau) = (\tau - I\{y_i - \mathbf{x}_i^T\hat{\beta}(\tau) < 0\}).$$

Then,

$$\| \mathbf{X}_2^T(\tau - I\{\mathbf{y} - \mathbf{X}_{\tilde{S}}\beta(\tau) < 0\}) \|_\infty = \| \mathbf{X}_2^T a_i(\tau) \|_\infty \leqslant \max_{j>s} \cdot | \sum_{i=1}^n x_{ij}E(a_i(\tau) - a_i^*(\tau)) | +$$

$$\max_{j>s} \cdot | \sum_{i=1}^n x_{ij}[a_i(\tau) - E(a_i(\tau)) - (a_i^*(\tau) - E(a_i^*(\tau))] | + \| \mathbf{X}_2^T a^*(\tau) \|_\infty := A_1 + A_2 + A_3.$$

By lemma 10 and lemma 12 given by Belloni and Chernozhukov (2011), we can bound A_1, A_2 as follows

$$A_1 \leqslant \sup_{\beta \in \mathbb{R}(|\tilde{S}|, \hat{r}, s), \alpha \in \mathbb{B}_\beta^p} n|E(\alpha^T\mathbf{x}_i a_i(\tau)) - E(\alpha^T\mathbf{x}_i a_i^*(\tau))| \lesssim \sqrt{\phi(s)}, \tag{20}$$

$$A_2 \leqslant \sup_{\beta \in \mathbb{R}(|\hat{S}|, \hat{r}, s), \alpha \in \mathbb{B}_\beta^p} n^{\frac{3}{2}}|[\mathbb{G}_n(\alpha^T\mathbf{x}_i a_i(\tau)) - \mathbb{G}_n(\alpha^T\mathbf{x}_i a_i^*(\tau))]| \lesssim \sqrt{ns\log(n \vee p)\phi(s)}.$$

For A_3, since $E(a^*(\tau)) = 0$, we apply Hoeffding's inequality and have

$$P\{\| \mathbf{X}_2^T a^*(\tau) \|_\infty \geqslant \sqrt{Cn\log(p)}\} \leqslant 2\sum_{j>s} 2\exp\{-\frac{Cn\log(p)}{4\sum_{i=1}^n x_{ij}^2}\} = 4\exp\{\log(|\tilde{S}| - s) - \frac{C}{4}\log(p)\} \lesssim p_n^{-c},$$

where c is some positive constant depending only on C, and can be chosen arbitrarily large. So, if $n^{-\frac{1}{2}}\lambda_{2,n}\log(p)^{-\frac{1}{2}} \to \infty$, then with probability at least $1 - O(p^{-c})$,

$$\frac{\| \mathbf{X}_2^T a^*(\tau) \|_\infty}{\lambda_{2,n}} < \frac{\sqrt{Cn\log(p)}}{\lambda_{2,n}} \to 0,$$

that is

$$\| \mathbf{X}_2^T a^*(\tau) \|_\infty = o_p(\lambda_{2,n}). \tag{21}$$

Combining (20),(21),(22), we have that uniformly over region $\mathbb{R}(|\tilde{S}|, \hat{r}, s)$,

$$\| w_2^{-1} \circ \mathbf{X}_2^T(\tau - I\{\mathbf{y} - \mathbf{X}_{\tilde{S}}\beta(\tau) < 0\}) \|_\infty \leqslant \lambda_{2,n}. \tag{22}$$

Table 1 Simulation results in Case 1

	Methods	$\tau = 0.25$				$\tau = 0.5$				$\tau = 0.75$			
		TP	FP	OP	MAE	TP	FP	OP	MAE	TP	FP	OP	MAE
n=300	Step1	6	29.10	0%	1.92	6	33.45	0%	1.69	6	30.55	0%	1.63
	Step2	6	2.57	50%	0.69	6	1.88	66%	0.57	6	3.41	49%	0.66
n=500	Step1	6	28.58	0%	2.53	6	28.05	0%	2.06	6	26.73	0%	1.89
	Step2	6	2.98	77%	0.60	6	1.57	79%	0.46	6	3.55	80%	0.54

Table 2 Simulation results in Case 2

	Methods	$\tau = 0.25$				$\tau = 0.5$				$\tau = 0.75$			
		TP	FP	OP	MAE	TP	FP	OP	MAE	TP	FP	OP	MAE
n=300	Step1	6	30.85	0%	2.30	6	30.02	0%	1.97	6	31.38	0%	1.88
	Step2	6	2.48	68%	0.99	6	2.42	61%	0.68	6	3.38	45%	0.81
n=500	Step1	6	24.00	0%	2.78	6	28.80	0%	2.44	6	30.57	0%	2.30
	Step2	6	2.19	80%	0.75	6	2.81	75%	0.89	6	4.16	78%	1.03

Table 3 Variable selection results for the Boston Price Dataset

	Methods	$\#n$	$\#n.a$	PE
$\tau = 0.25$	Step1,p_n=50	9.10(0.1595)	0.00(0.0000)	0.3967(0.2407)
	Step1,p_n=200	9.22(0.1040)	0.00(0.0000)	0.5447(0.2526)
	Step1,p_n=500	9.32(0.1158)	0.00(0.0000)	0.4367(0.2550)
	Step2,p_n=50	8.24(0.1387)	0.00(0.0000)	0.3460(0.1040)
	Step2,p_n=200	8.28(0.1486)	0.00(0.0000)	0.4611(0.1141)
	Step2,p_n=500	8.44(0.1075)	0.00(0.0000)	0.2962(0.1268)
$\tau = 0.5$	Step1,p_n=50	10.08(0.1652)	0.00(0.0000)	0.3596(0.2145)
	Step1,p_n=200	10.38(0.1208)	0.00(0.0000)	0.3936(0.2006)
	Step1,p_n=500	10.78(0.1598)	0.00(0.0000)	0.3528(0.2149)
	Step2,p_n=50	8.30(0.1519)	0.00(0.0000)	0.3499(0.1442)
	Step2,p_n=200	8.32(0.1323)	0.00(0.0000)	0.3612(0.1428)
	Step2,p_n=500	8.56(0.1314)	0.00(0.0000)	0.3395(0.1622)
$\tau = 0.75$	Step1,p_n=50	10.18(0.2058)	0.00(0.0000)	0.3830(0.1925)
	Step1,p_n=200	10.26(0.1976)	0.00(0.0000)	0.3187(0.1852)
	Step1,p_n=500	10.44(0.2271)	0.00(0.0000)	0.3484(0.1728)
	Step2,p_n=50	7.58(0.1854)	0.00(0.0000)	0.3698(0.1456)
	Step2,p_n=200	7.74(0.1474)	0.00(0.0000)	0.2883(0.1443)
	Step2,p_n=500	7.79(0.1733)	0.00(0.0000)	0.2300(0.1320)

Table 4 Frequency (″Fre.″) and parameter estimation for the real data, where $p_n = 50, \tau = 0.25$

non-censored			15%-censored		
Varialbe	Fre.	estimation	Varialbe	Fre.	estimation
B^2	100%	-0.0105 (0.0085)	RM^2	100%	0.0667 (0.0480)
LSTAT	100%	-0.2010 (0.1457)	PTRAT	100%	-0.0425 (0.0342)
Inter.	100%	-0.1805 (0.1283)	RM	100%	0.1884 (0.1398)
RM	96%	0.1342 (0.1092)	Inter.	100%	-0.1319 (0.0975)
PTRAT	94%	-0.0344 (0.0323)	LSTAT	98%	-0.1250 (0.1022)
TAX	92%	-0.0540 (0.0478)	AGE	94%	-0.0343 (0.0325)
LON	88%	-0.0211 (0.0216)	B^2	86%	-0.0147 (0.0129)
$CRIM^2$	78%	-0.0037 (0.0034)	LON	60%	-0.0111 (0.0147)
RM^2	76%	0.0648 (0.0538)	$INDUS^2$	48%	-0.0096 (0.0129)
			LAT^2	34%	-0.0096 (0.0139)
			TAX	32%	-0.0304 (0.0423)

Table 5 Frequency (″Fre.″) and parameter estimation for the real data, where $p_n = 50, \tau = 0.5$

non-censored			15%-censored		
Varialbe	Fre.	estimation	Varialbe	Fre.	estimation
B^2	100%	-0.0140 (0.0104)	RM^2	100%	0.0786 (0.0562)
RM^2	100%	0.0762 (0.0544)	LSTAT	100%	-0.1225 (0.0964)
LSTAT	100%	-0.1936 (0.1379)	PTRAT	100%	-0.0721 (0.0524)
PTRAT	100%	-0.0636 (0.0465)	RM	100%	0.2033 (0.1462)
RM	100%	0.1696 (0.1209)	Inter.	98%	-0.0356 (0.0452)
Inter.	100%	-0.0829 (0.0595)	B^2	96%	-0.0144 (0.0115)
LON	80%	-0.0265 (0.0225)	AGE	86%	-0.0433 (0.0429)
$CRIM^2$	72%	-0.0038 (0.0044)	$INDUS^2$	74%	-0.0162 (0.0162)
TAX	60%	-0.0407 (0.0449)	LAT^2	50%	-0.0238 (0.0298)
			LON	36%	-0.0167 (0.0218)

Table 6 Frequency (″Fre.″) and parameter estimation for the real data, where $p_n = 50, \tau = 0.75$

non-censored			15%-censored		
Varialbe	Fre.	estimation	Varialbe	Fre.	estimation
RM^2	100%	0.1012 (0.0734)	RM^2	100%	0.1043 (0.0754)
LSTAT	100%	-0.1726 (0.1235)	LSTAT	100%	-0.1614 (0.1182)
PTRAT	100%	-0.0612 (0.0466)	PTRAT	100%	-0.0592 (0.0441)
RM	100%	0.2073 (0.1486)	RM	100%	0.2056 (0.1487)
Inter.	100%	0.0365 (0.0340)	Inter.	100%	0.0506 (0.0417)
B^2	94%	-0.0187 (0.0147)	LAT^2	88%	-0.0185 (0.0179)
LAT^2	54%	-0.0118 (0.0136)	B^2	84%	-0.0169 (0.0142)
$CRIM^2$	40%	-0.0055 (0.0074)	ZN^2	58%	0.0051 (0.0063)
ZN^2	36%	0.0042 (0.0058)	$INDUS^2$	48%	-0.0196 (0.0211)
NOX^2	28%	-0.0214 (0.0292)	$LSTAT^2$	46%	0.0234 (0.0301)

TP denotes the average number of relevant variables that are correctly selected. FP denotes the average number of irrelevant variables that are incorrectly selected. OP denotes the percentage of times that the true model is correctly selected. MAE denotes the median of the absolute estimation error defined by $\sum_{i=1}^{p}|\hat{\beta}_i-\beta_{0,i}|$

The columns show the number of nonzero coefficients selected by each method among original predictors($\#n$) and artificial variables ($\#n.a$), and the prediction errors (PE), which are defined by $\sum_{i=1}^{100}|\hat{y}_i-y_i|$.

分位数回归模型中的两步变量选择

樊亚莉

(上海理工大学 理学院，上海 200093)

摘 要：对于高维分位数回归模型提出了一种两步变量选择方法, 这里协变量的维数p_n 远远大于样本量n. 在第一步中，使用 ℓ_1 惩罚，并且证明第一步由 LASSO 惩罚所得到的惩罚估计量能够把模型从超高维降到同真实模型同阶的维数，并且所选模型能够覆盖真实模型. 第二步对第一步所得模型使用自适应的 LASSO 惩罚来剔除冗余变量. 在一些正则性条件下，证明了此方法具有变量选择的相合性. 还进行了数值模拟和实际数据分析，用来表明此方法在有限样本下的表现.

关键词：LASSO; 自适应 LASSO; 分位数回归；高维

(责任编辑：冯珍珍，郁 慧)