Quantile Regression for Analyzing Heterogeneity in Ultra-high Dimension

Lan Wang, Yichao Wu and Runze Li

Abstract

Ultra-high dimensional data often display heterogeneity due to either heteroscedastic variance or other forms of non-location-scale covariate effects. To accommodate heterogeneity, we advocate a more general interpretation of sparsity which assumes that only a small number of covariates influence the conditional distribution of the response variable given all candidate covariates; however, the sets of relevant covariates may differ when we consider different segments of the conditional distribution. In this framework, we investigate the methodology and theory of nonconvex penalized quantile regression in ultra-high dimension. The proposed approach has two distinctive features: (1) it enables us to explore the entire conditional distribution of the response variable given the ultra-high dimensional covariates and provides a more realistic picture of the sparsity pattern; (2) it requires substantially weaker conditions compared with alternative methods in the literature; thus, it greatly alleviates the difficulty of model checking in the ultra-high dimension. In theoretic development, it is challenging to deal with both the nonsmooth loss function and the nonconvex penalty function in ultra-high dimensional parameter space. We introduce a novel sufficient optimality condition which relies on a convex differencing representation of the penalized loss function and the subdifferential calculus. Exploring this optimality condition enables us to establish the oracle

[1]Lan Wang is Associate Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Email: wangx346@umn.edu. Yichao Wu is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695. Email: wu@stat.ncsu.edu. Runze Li is the corresponding author and Professor, Department of Statistics and the Methodology Center, the Pennsylvania State University, University Park, PA 16802-2111. Email: rli@stat.psu.edu. Wang's research is supported by a NSF grant DMS1007603. Wu's research is supported by a NSF grant DMS-0905561 and NIH/NCI grant R01 CA-149569. Li's research is supported by a NSF grant DMS 0348869 and a grant from NNSF of China, 11028103 and NIDA, NIH grants R21 DA024260 and P50 DA10075. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NCI, the NIDA or the NIH. The authors are indebted to the referees, the associate editor and the Co-editor for their valuable comments, which have significantly improved the paper.

property for sparse quantile regression in the ultra-high dimension under relaxed conditions. The proposed method greatly enhances existing tools for ultra-high dimensional data analysis. Monte Carlo simulations demonstrate the usefulness of the proposed procedure. The real data example we analyzed demonstrates that the new approach reveals substantially more information compared with alternative methods.

KEY WORDS: Penalized Quantile Regression, SCAD, Sparsity, Ultra-high dimensional data

1 Introduction

High-dimensional data are frequently collected in a large variety of research areas such as genomics, functional magnetic resonance imaging, tomography, economics and finance. Analysis of high-dimensional data poses many challenges for statisticians and calls for new statistical methodologies and theories (Donoho, 2000; Fan and Li, 2006). We consider the ultra-high dimensional regression setting in which the number of covariates p grows at an exponential rate of the sample size n.

When the primary goal is to identify the underlying model structure, a popular approach for analyzing ultra-high dimensional data is to use the regularized regression. For example, Candes and Tao (2007) proposed the Dantzig selector; Candes, Wakin and Boyd (2007) proposed weighted L_1-minimization to enhance the sparsity of the Dantzig selector; Huang, Ma and Zhang (2008) considered the adaptive lasso when a zero-consistent initial estimator is available; Kim, Choi and Oh (2008) demonstrated that the SCAD estimator still has the oracle property in ultra-high dimension regression when the random errors follow the normal distribution; Fan and Lv (2009) investigated non-concave penalized likelihood with ultra-high dimensionality; and Zhang (2010) proposed a minimax concave penalty (MCP) for penalized regression.

It is common to observe that real life ultra-high dimensional data display heterogeneity due to either heteroscedastic variance or other forms of non-location-scale covariate effects. This type of heterogeneity is often of scientific importance but tends to be overlooked by exiting procedures which mostly focus on the mean of the conditional distribution. Furthermore, despite significant recent developments in ultra-high dimensional regularized regression, the statistical theory of the existing methods generally requires conditions substantially stronger than those usually imposed in the classical $p < n$ framework. These conditions include homoscedastic random errors,

Gaussian or near Gaussian distributions, and often hard-to-check conditions on the design matrix, among others. These two main concerns motivate us to study nonconvex penalized quantile regression in ultra-high dimension.

Quantile regression (Koenker and Bassett, 1978) has become a popular alternative to least squares regression for modeling heterogeneous data. We refer to Koenker (2005) for a comprehensive introduction and to He (2009) for a general overview of many interesting recent developments. Welsh (1989), Bai and Wu (1994) and He and Shao (2000) established nice asymptotic theory for high-dimensional M-regression with possibly nonsmooth objective functions. Their results apply to quantile regression (without the sparseness assumption) but require that $p = o(n)$.

In this paper, we extend the methodology and theory of quantile regression to ultra-high dimension. To deal with the ultra-high dimensionality, we regularize quantile regression with a nonconvex penalty function, such as the SCAD penalty and the MCP. The choice of nonconvex penalty is motivated by the well-known fact that directly applying the L_1 penalty tends to include inactive variables and to introduce bias. We advocate a more general interpretation of sparsity which assumes that only a small number of covariates influence the conditional distribution of the response variable given all candidate covariates; however, the sets of relevant covariates may be different when we consider different segments of the conditional distribution. By considering different quantiles, this framework enables us to explore the entire conditional distribution of the response variable given the ultra-high dimensional covariates. In particular, it can provide a more realistic picture of the sparsity patterns, which may differ at different quantiles.

Regularized quantile regression with fixed p was recently studied by Li and Zhu (2008), Zou and Yuan (2008), Wu and Liu (2009) and Kai, Li and Zou (2011). Their

asymptotic techniques, however, are difficult to extend to the ultra-high dimension. For high dimensional p, Belloni and Chernozhukov (2011) recently derived a nice error bound for quantile regression with the L_1-penalty. They also showed that a post-L_1-quantile regression procedure can further reduce the bias. However, in general post-L_1-quantile regression does not possess the oracle property.

The main technical challenge of our work is to deal with both the nonsmooth loss function and the nonconvex penalty function in ultra-high dimension. Note that to characterize the solution to quantile regression with nonconvex penalty, the Karush-Kuhn-Tucker (KKT) local optimality condition is necessary but generally not sufficient. To establish the asymptotic theory, we novelly apply a sufficient optimality condition for the convex differencing algorithm; which relies on a convex differencing representation of the penalized quantile loss function (Section 2.2). Furthermore, we employ empirical process techniques to derive various error bounds associated with the nonsmooth objective function in high dimension. We prove that with probability approaching one, the oracle estimator, which estimates the zero coefficients as zero and estimates the nonzero coefficients as efficiently as if the true model is known in advance, is a local solution of the nonconvex penalized sparse quantile regression with either the SCAD penalty or the MCP penalty for ultra-high dimensionality.

The theory established in this paper for sparse quantile regression requires much weaker assumptions than those in the literature, which alleviates the difficulty of checking model adequacy in the ultra-high dimension settings. In particular, we do not impose restrictive distributional or moment conditions on the random errors and allow their distributions to depend on the covariates (Condition (C3) in the Appendix). Kim, Choi and Oh (2008) derived the oracle property of the high-dimensional SCAD penalized least squares regression under rather general conditions. They also discov-

ered that for the squared error loss, the upper bound of the dimension of covariates is strongly related to the highest existing moment of the error distribution. The higher the moment exists, the larger p is allowed for the oracle property; and for the normal random errors the covariate vector may be ultra-high dimensional. In fact, most of the theory in the literature for ultra-high dimensional penalized least squares regression requires either the Gaussian or Sub-Gaussian condition.

The rest of the paper is organized as follows. In Section 2, we propose sparse quantile regression with nonconvex penalty, and introduce a local optimality condition for nonsmooth nonconvex optimization. We further study the asymptotic theory for sparse quantile regression with ultra-high dimensional covariates. In Section 3, we conduct a Monte Carlo study and illustrate the proposed methodology by an empirical analysis of an eQTL microarray data set. All regularity conditions and technical proofs are relegated to the Appendix.

2 Nonconvex Penalized Quantile Regression

2.1 The Methodology

Let us begin with the notation and statistical setup. Suppose that we have a random sample $\{Y_i, x_{i1}, \ldots, x_{ip}\}$, $i = 1, \ldots, n$, from the following model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \stackrel{\Delta}{=} \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a $(p+1)$-dimensional vector of parameters, $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^T$ with $x_{i0} = 1$, and the random errors ϵ_i satisfy $P(\epsilon_i \leq 0 | \mathbf{x}_i) = \tau$ for some specified $0 < \tau < 1$. The case $\tau = 1/2$ corresponds to median regression.

The number of covariates $p = p_n$ is allowed to increase with the sample size n. It is possible that p_n is much larger than n.

The true parameter value $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01} \ldots, \beta_{0p_n})^T$ is assumed to be sparse; that is, the majority of its components are exactly zero. Let $A = \{1 \leq j \leq p_n : \beta_{0j} \neq 0\}$ be the index set of the nonzero coefficients. Let $|A| = q_n$ be the cardinality of the set A, which is allowed to increase with n. In the general framework of sparsity discussed in Section 1, both the set A and the number of nonzero coefficients q_n depend on the quantile τ. We omit such dependence in notation for simplicity. Without loss of generality, we assume that the last $p_n - q_n$ components of $\boldsymbol{\beta}_0$ are zero; that is, we can write $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^T, \mathbf{0}^T)^T$, where $\mathbf{0}$ denotes a $(p_n - q_n)$- dimensional vector of zeros. The oracle estimator is defined as $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T$, where $\widehat{\boldsymbol{\beta}}_1$ is the quantile regression estimator obtained when the model is fitted using only relevant covariates (i.e., those with index in A).

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ be the $n \times (p_n + 1)$ matrix of covariates, where $\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T$ are the rows of $\mathbf{X}$. We also write $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1, \ldots, \mathbf{X}_{p_n})$, where $\mathbf{1}, \mathbf{X}_1, \ldots, \mathbf{X}_p$ are the columns of $\mathbf{X}$ and $\mathbf{1}$ represents an n-vector of ones. Define $\mathbf{X}_A$ to be the submatrix of $\mathbf{X}$ that consists of its first $q_n + 1$ columns; similarly denoted by $\mathbf{X}_{A^c}$ the submatrix of $\mathbf{X}$ that consists of its last $p_n - q_n$ columns. For the rest of this paper, we often omit the subscript n for simplicity. In particular, we let p and q stand for p_n and q_n, respectively.

We consider the following penalized quantile regression model

$$Q(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2}$$

where $\rho_\tau(u) = u \{\tau - I(u < 0)\}$ is the quantile loss function (or check function), and

$p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ. The tuning parameter λ controls the model complexity and goes to zero at an appropriate rate. The penalty function $p_\lambda(t)$ is assumed to be nondecreasing and concave for $t \in [0, +\infty)$, with a continuous derivative $\dot{p}_\lambda(t)$ on $(0, +\infty)$. It is well known that penalized regression with the convex L_1 penalty tends to over-penalize large coefficients and to include spurious variables in the selected model. This may not be of much concern for predicting future observations, but is nonetheless undesirable when the purpose of the data analysis is to gain insights into the relationship between the response variable and the set of covariates.

In this paper, we consider two commonly used nonconvex penalties: the SCAD penalty and the MCP. The SCAD penalty function is defined by

$$\begin{aligned} p_\lambda(|\beta|) &= \lambda|\beta| I(0 \le |\beta| < \lambda) + \frac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a-1} I(\lambda \le |\beta| \le a\lambda) \\ &\quad + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda), \text{ for some } a > 2. \end{aligned}$$

The MCP function has the form

$$p_\lambda(|\beta|) = \lambda\Big(|\beta| - \frac{\beta^2}{2a\lambda}\Big) I(0 \le |\beta| < a\lambda) + \frac{a\lambda^2}{2} I(|\beta| \ge a\lambda), \quad \text{for some } a > 1.$$

2.2 Difference Convex Program and a Sufficient Local Optimality Condition

Kim, Choi and Oh (2008) investigated the theory of high-dimension SCAD penalized least squares regression by exploring the sufficient condition for local optimality. Their formulation is quite different from those in Fan and Li (2001), Fan and Peng (2004) and Fan and Lv (2009). For the SCAD-penalized least squares problem, the objective function is also nonsmooth and nonconvex. By writing $\beta_j = \beta_j^+ - \beta_j^-$, where β_j^+ and β_j^-

represent the positive part and the negative part of β_j respectively, the minimization problem can be equivalently expressed as a constrained smooth optimization problem. Therefore, the second-order sufficient condition (Proposition 3.3.2, Bertsekas, 2008) from constrained smooth optimization theory can be applied. Exploring local optimality condition directly leads to asymptotic theory for SCAD penalized least squares regression under more relaxed conditions.

However, in our case, the loss function itself is also nonsmooth in addition to the non-smoothness of the penalty function. As a result, the above local optimality condition for constrained smooth optimization is not applicable. In this paper, we novelly apply a new local optimality condition which can be applied to a much broader class of nonconvex nonsmoothing optimization problem. More specifically, we consider penalized loss functions belonging to the class

$$\mathbf{F} = \{f(\mathbf{x}) : f(\mathbf{x}) = g(\mathbf{x}) - h(\mathbf{x}), \quad g, h \text{ are both convex}\}.$$

This class of functions are very broad as it covers many other useful loss functions in addition to the quantile loss function, for example the least squares loss function, the Huber loss function for robust estimation and many loss functions used in the classification literature. Numerical algorithms based on the convex differencing representation and their convergence properties have been systematically studied by Tao and An (1997), An and Tao (2005), among others, in the filed of nonsmooth optimization. These algorithms have seen some recent applications in statistical learning, for example Liu, Shen and Doss (2005), Collobert et al. (2006), Kim, Choi and Oh (2008) and Wu and Liu (2009), among others.

Let $\text{dom}(g) = \{\mathbf{x} : g(\mathbf{x}) < \infty\}$ be the effective domain of g and let $\partial g(\mathbf{x}_0) =$

$\{\mathbf{t} : g(\mathbf{x}) \geq g(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^T\mathbf{t}, \ \forall \ \mathbf{x}\}$ be the *subdifferential* of a convex function $g(\mathbf{x})$ at a point $\mathbf{x}_0$. Although subgradient-based KKT condition has been commonly used for characterizing the necessary conditions for nonsmooth optimization; sufficient conditions for characterizing the set of local minima of nonsmooth nonconvex objective functions have not been explored in the statistical literature. The following lemma which was first presented and proved in Tao and An (1997), see also An and Tao (2005), provides a sufficient local optimization condition for the DC program based on the subdifferential calculus.

Lemma 2.1 *These exists a neighborhood U around the point $\mathbf{x}^*$ such that $\partial h(\mathbf{x}) \cap \partial g(\mathbf{x}^*) \neq \emptyset$, $\forall \ \mathbf{x} \in U \cap \mathrm{dom}(g)$. Then $\mathbf{x}^*$ is a local minimizer of $g(\mathbf{x}) - h(\mathbf{x})$.*

2.3 Asymptotic Properties

For notational convenience, we write $\mathbf{x}_i^T = (\mathbf{z}_i^T, \mathbf{w}_i^T)$, where $\mathbf{z}_i = (x_{i0}, x_{i1}, \ldots, x_{iq_n})^T$ and $\mathbf{w}_i = (x_{i(q_n+1)}, \ldots, x_{ip_n})^T$. We consider the case in which the covariates are from a fixed design. We impose the following regularity conditions to facilitate our technical proofs.

(C1) (Conditions on the design) There exists a positive constant M_1 such that $\frac{1}{n}\mathbf{X}_j^T\mathbf{X}_j \leq M_1$ for $j = 1, \ldots q$. Also $|x_{ij}| \leq M_1$ for all $1 \leq i \leq n$, $q+1 \leq j \leq p$.

(C2) (Conditions on the true underlying model) There exist positive constants $M_2 < M_3$ such that

$$M_2 \leq \lambda_{\min}\big(n^{-1}\mathbf{X}_A^T\mathbf{X}_A\big) \leq \lambda_{\max}\big(n^{-1}\mathbf{X}_A^T\mathbf{X}_A\big) \leq M_3,$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest eigenvalue and largest eigenvalue, re-

spectively. It is assumed that $\max_{1\leq i\leq n}||\mathbf{z}_i|| = O_p(\sqrt{q})$, $(\mathbf{z}_i, Y_i)$ are in general positions (Section 2.2, Koenker, 2005) and that there is at least one continuous covariate in the true model.

(C3) (Conditions on the random error) The conditional probability density function of ϵ_i, denoted by $f_i(\cdot|\mathbf{z}_i)$, is uniformly bounded away from 0 and ∞ in a neighborhood around 0 for all i.

(C4) (Condition on the true model dimension) The true model dimension q_n satisfies $q_n = O(n^{c_1})$ for some $0 \leq c_1 < 1/2$.

(C5) (Condition on the smallest signal) There exist positive constants c_2 and M_4 such that $2c_1 < c_2 \leq 1$ and $n^{(1-c_2)/2}\min_{1\leq j\leq q_n}|\beta_{0j}| \geq M_4$.

Conditions (C1), (C2), (C4) and (C5) are common in the literature on high-dimensional inference, for example, condition (C2) requires that the design matrix corresponding to the true model is well behaved, and condition (C4) requires that the smallest signal should not decay too fast. In particular, they are similar to those in Kim et al. (2008). Condition (C3), on the other hand, is more relaxed than the Gaussian or Subgaussian error condition usually assumed in the literature for ultra-high dimensional regression.

To formulate the problem in the framework of Section 2.2, we first note that the nonconvex penalized quantile objective function $\mathrm{Q}(\boldsymbol{\beta})$ in (2) can be written as the difference of two convex functions in $\boldsymbol{\beta}$:

$$Q(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) - h(\boldsymbol{\beta}),$$

where $g(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|$ and $h(\boldsymbol{\beta}) = \sum_{j=1}^{p}H_\lambda(\beta_j)$. The form of $H_\lambda(\beta_j)$ depends on the penalty function. For the SCAD penalty, we have

$$\begin{aligned} H_\lambda(\beta_j) &= \Big[(\beta_j^2 - 2\lambda|\beta_j| + \lambda^2)/(2(a-1))\Big]I\big(\lambda \le |\beta_j| \le a\lambda\big) \\ &\quad + \Big[\lambda|\beta_j| - (a+1)\lambda^2/2\Big]I\big(|\beta_j| > a\lambda\big); \end{aligned}$$

while for the MCP function, we have

$$H_\lambda(\beta_j) = \Big[\beta_j^2/(2a)\Big]I\big(0 \le |\beta_j| < a\lambda\big) + \Big[\lambda|\beta_j| - a\lambda^2/2\Big]I\big(|\beta_j| \ge a\lambda\big).$$

Next, we characterize the subdifferentials of $g(\boldsymbol{\beta})$ and $h(\boldsymbol{\beta})$, respectively. The subdifferential of $g(\boldsymbol{\beta})$ at $\boldsymbol{\beta}$ is defined as the following collection of vectors:

$$\begin{aligned} \partial g(\boldsymbol{\beta}) = \Big\{\boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_p) \in \mathcal{R}^{p+1} : \xi_j &= -\tau n^{-1}\sum_{i=1}^{n} x_{ij}I(Y_i - \mathbf{x}_i^T\boldsymbol{\beta} > 0) \\ &+ (1-\tau)n^{-1}\sum_{i=1}^{n} x_{ij}I(Y_i - \mathbf{x}_i^T\boldsymbol{\beta} < 0) - n^{-1}\sum_{i=1}^{n} x_{ij}v_i + \lambda l_j\Big\}, \end{aligned}$$

where $v_i = 0$ if $Y_i - \mathbf{x}_i^T\boldsymbol{\beta} \neq 0$ and $v_i \in [\tau - 1, \tau]$ otherwise; $l_0 = 0$; for $1 \le j \le p$ $l_j = \mathrm{sgn}(\beta_j)$ if $\beta_j \neq 0$ and $l_j \in [-1, 1]$ otherwise. In this definition, $\mathrm{sgn}(t) = I(t > 0) - I(t < 0)$. Furthermore, for both the SCAD penalty and the MCP penalty, $h(\boldsymbol{\beta})$ is differentiable everywhere. Thus the subdifferential of $h(\boldsymbol{\beta})$ at any point $\boldsymbol{\beta}$ is a singleton:

$$\partial h(\boldsymbol{\beta}) = \Big\{\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_p)^T \in \mathcal{R}^{p+1} : \mu_j = \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}\Big\}.$$

For both penalty functions, $\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = 0$ for $j = 0$. For $1 \leq j \leq p$,

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = \Big[(\beta_j - \lambda \text{sgn}(\beta_j))/(a-1)\Big] I\big(\lambda \leq |\beta_j| \leq a\lambda\big) + \lambda \text{sgn}(\beta_j) I\big(|\beta_j| > a\lambda\big),$$

for the SCAD penalty; while for the MCP penalty

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = \Big[\beta_j/a\Big] I\big(0 \leq |\beta_j| < a\lambda\big) + \lambda \text{sgn}(\beta_j) I\big(|\beta_j| \geq a\lambda\big).$$

The application of Lemma 2.1 utilizes the results from the following two lemmas. The set of the subgradient functions for the unpenalized quantile regression is defined as the collection of the vector $\mathbf{s}(\widehat{\boldsymbol{\beta}}) = (s_0(\widehat{\boldsymbol{\beta}}), s_1(\widehat{\boldsymbol{\beta}}), \ldots, s_p(\widehat{\boldsymbol{\beta}}))^T$, where

$$s_j(\widehat{\boldsymbol{\beta}}) = -\frac{\tau}{n} \sum_{i=1}^{n} x_{ij} I(Y_i - \mathbf{x}_i^T \boldsymbol{\beta} > 0) + \frac{1-\tau}{n} \sum_{i=1}^{n} x_{ij} I(Y_i - \mathbf{x}_i^T \boldsymbol{\beta} < 0) - \frac{1}{n} \sum_{i=1}^{n} x_{ij} v_i,$$

for $j = 0, 1, \ldots, p$, with $v_i = 0$ if $Y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} \neq 0$ and $v_i \in [\tau - 1, \tau]$ otherwise.

Lemmas 2.2 and 2.3 below characterize the properties of the oracle estimator and the subgradient functions corresponding to the active and inactive variables, respectively.

Lemma 2.2 *Suppose that conditions (C1)-(C5) in the Appendix hold and that* $\lambda = o(n^{-(1-c_2)/2})$. *For the oracle estimator* $\widehat{\boldsymbol{\beta}}$, *there exist* v_i^* *which satisfies* $v_i^* = 0$ *if* $Y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} \neq 0$ *and* $v_i^* \in [\tau - 1, \tau]$ *if* $Y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} = 0$, *such that for* $s_j(\widehat{\boldsymbol{\beta}})$ *with* $v_i = v_i^*$, *with probability approaching one, we have*

$$s_j(\widehat{\boldsymbol{\beta}}) = 0, \ j = 0, 1, \ldots, q, \quad \textit{and} \quad |\widehat{\beta}_j| \geq (a + 1/2)\lambda, \ j = 1, \ldots, q. \tag{3}$$

Lemma 2.3 *Suppose that conditions (C1)-(C5) in the Appendix hold and that* $qn^{-1/2} =$

$o(\lambda)$, $\log p = o(n\lambda^2)$ *and* $n\lambda^2 \to \infty$. *For the oracle estimator* $\widehat{\boldsymbol{\beta}}$ *and the* $s_j(\widehat{\boldsymbol{\beta}})$ *defined in Lemma 2.2, with probability approaching one, we have*

$$|s_j(\widehat{\boldsymbol{\beta}})| \leq \lambda, \ j = q+1, \ldots, p, \quad \textit{and} \quad |\widehat{\beta}_j| = 0, \quad j = q+1, \ldots, p. \tag{4}$$

Applying the above results, we will prove that with probability tending to one, for any $\boldsymbol{\beta}$ in a ball in $\mathcal{R}^{p+1}$ with the center $\widehat{\boldsymbol{\beta}}$ and radius $\lambda/2$, there exists a subgradient $\boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_p)^T \in \partial g(\widehat{\boldsymbol{\beta}})$ such that

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = \xi_j, \quad j = 0, 1, \ldots, p. \tag{5}$$

Then by Lemma 2.1, we can demonstrate that the oracle estimator $\widehat{\boldsymbol{\beta}}$ is itself a local minimizer. This is summarized in the following theorem.

Theorem 2.4 *Assume that conditions (C1)-(C5) in the Appendix hold. Let* $\mathcal{B}_n(\lambda)$ *be the set of local minima of the nonconvex penalized quantile objective function (2) with either the SCAD penalty or the MCP penalty and tuning parameter* λ. *The oracle estimator* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T$ *satisfies that*

$$P(\widehat{\boldsymbol{\beta}} \in \mathcal{B}_n(\lambda)) \to 1$$

as $n \to \infty$ *if* $\lambda = o(n^{-(1-c_2)/2})$, $n^{-1/2}q = o(\lambda)$ *and* $\log(p) = o(n\lambda^2)$.

REMARK. It can be shown that if we take $\lambda = n^{-1/2+\delta}$ for some $c_1 < \delta < \frac{c_2}{2}$, then these conditions are satisfied. We can also have $p = o(\exp(n^{\delta}))$. Therefore, the oracle property for sparse quantile regression holds in ultra-high dimension without the restrictive distributional or moment conditions on the random errors, which are commonly im-

posed for nonconvex penalized mean regression. This result greatly complements and enhances those in Fan and Li (2001) for fixed p, Fan and Peng (2004) for large p but $p < n$, Kim, Choi and Oh (2008) for $p > n$ and Fan and Lv (2009) for $p >> n$.

3 Simulation and Real Data Example

In this section, we investigate the performance of nonconvex penalized high-dimensional quantile regression with the SCAD penalty and the MCP (denoted by Q-SCAD and Q-MCP, respectively). We compare these two procedures with least-squares based high-dimensional procedures, including LASSO, adaptive LASSO, SCAD and MCP penalized least squares regression (denoted by LS-Lasso, LS-ALasso, LS-SCAD and LS-MCP, respectively). We also compare the proposed procedures with LASSO penalized and adaptive-LASSO penalized quantile regression (denoted by Q-Lasso and Q-ALasso, respectively). Our main interest is the performance of various procedures when $p > n$ and the ability of the nonconvex penalized quantile regression to identify signature variables that are overlooked by the least-squares based procedures.

While conducting our numerical experiments, we use the local linear approximation algorithm (LLA, Zou and Li, 2008) to implement the nonconvex penalized quantile regression. More explicitly, while minimizing $\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|)$, we initialize by setting $\tilde{\beta}_j^{(0)} = 0$ for $j = 1, 2, \cdots, p$. For each step $t \geq 1$, we update by solving

$$\min_{\boldsymbol{\beta}} \Big\{ \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + \sum_{j=1}^{p} w_j^{(t-1)}|\beta_j| \Big\}, \tag{6}$$

where $w_j^{(t-1)} = p'_\lambda(|\tilde{\beta}_j^{(t-1)}|) \geq 0$ denotes the weight and $p'_\lambda(\cdot)$ denotes the derivative of $p_\lambda(\cdot)$. Following the literature, when $\tilde{\beta}_j^{(t-1)} = 0$, we take $p'_\lambda(0)$ as $p'_\lambda(0+) = \lambda$. With the aid of slack variables ξ_i^+, ξ_i^-, and ζ_j, the convex optimization problem in (6) can

be equivalently rewritten as

$$\min_{\boldsymbol{\xi},\boldsymbol{\zeta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\tau \xi_i^+ + (1-\tau)\xi_i^-) + \sum_{j=1}^{p} w_j^{(t-1)} \zeta_j \right\} \tag{7}$$

$$\text{subject to} \quad \xi_i^+ - \xi_i^- = Y_i - \mathbf{x}_i^T \boldsymbol{\beta};\ i = 1, 2, \cdots, n,$$

$$\xi_i^+ \geq 0, \xi_i^- \geq 0;\ i = 1, 2, \cdots, n,$$

$$\zeta_j \geq \beta_j, \zeta_j \geq -\beta_j;\ j = 1, 2, \cdots, p.$$

Note that (7) is a linear programming problem and can be solved using many existing optimization software packages. We claim convergence of the LLA algorithm when the weights $w_j^{(t)}, j = 1, 2, \cdots, p$, stabilize, namely when $\sum_{j=1}^{p} (w_j^{(t-1)} - w_j^{(t)})^2$ is sufficiently small. Alternatively, one may obtain the solution by using the algorithm proposed in Li and Zhu (2008) for calculating the solution path of the L_1-penalized quantile regression.

3.1 Simulation Study

Predictors $X_1, X_2, \cdots, X_p$ are generated in two steps. We first generate $(\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_p)^T$ from the multivariate normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\sigma_{jk})_{p \times p}$ and $\sigma_{jk} = 0.5^{|j-k|}$. The next step is to set $X_1 = \Phi(\tilde{X}_1)$ and $X_j = \tilde{X}_j$ for $j = 2, 3, \cdots, p$. The scalar response is generated according to the heteroscedastic location-scale model

$$Y = X_6 + X_{12} + X_{15} + X_{20} + 0.7 X_1 \epsilon,$$

where $\epsilon \sim N(0, 1)$ is independent of the covariates. In this simulation experiment, X_1 plays an essential role in the conditional distribution of Y given the covariates; but does not directly influence the center (mean or median) of the conditional distribution.

We consider sample size $n = 300$ and covariate dimension $p = 400$ and 600. For quantile regression, we consider three different quantiles $\tau = 0.3$, 0.5 and 0.7. We generate an independent tuning data set of size $10n$ to select the regularization parameter by minimizing the estimated prediction error (based on either the squared error loss or the check function loss, depending on which loss function is used for estimation) calculated over the tuning data set; similarly as in Mazumder, Friedman and Hastie (2009). In the real data analysis in Section 3.2, we use cross-validation for tuning parameter selection.

For a given method, we denote the resulted estimate by $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T$. Based on simulation of 100 repetitions, we compare the performance of the aforementioned different methods in terms of the following criteria.

Size: the average number of non-zero regression coefficients $\widehat{\beta}_j \neq 0$ for $j = 1, 2, \cdots, p$;

P1: the proportion of simulation runs including all true important predictors, namely $\widehat{\beta}_j \neq 0$ for any $j \geq 1$ satisfying $\beta_j \neq 0$. For the LS-based procedures and conditional median regression, this means the percentage of times including X_5, X_{12}, X_{15} and X_{20}; for conditional quantile regression at $\tau = 0.3$ and $\tau = 0.7$, X_1 should also be included.

P2: the proportion of simulation runs X_1 is selected.

AE: the absolute estimation error defined by $\sum_{j=0}^{p} |\widehat{\beta}_j - \beta_j|$.

Tables 1 and 2 depict the simulation results for $p = 400$ and $p = 600$, respectively. In these two tables, the numbers in the parentheses in the columns labeled 'Size' and 'AE' are the corresponding sample standard deviations based on the 100 simulations. The simulation results confirm satisfactory performance of the nonconvex penalized

quantile regression when $p > n$ for selecting and estimating relevant covariates. In this example, the signature variable X_1 is often missed by least-squares based methods, but has high probability of being included when several different quantiles are examined together. This demonstrates that by considering several different quantiles, it is likely to gain a more complete picture of the underlying structure of the conditional distribution. From Tables 1 and 2, it can be seen that the penalized quantile median regression improves the corresponding penalized least squares methods in terms of AE due to the heteoscedastic error. Furthermore, it is observed that LASSO-penalized quantile regression tends to select a much larger model; on the other hand, the adaptive-Lasso penalized quantile regression tends to select a sparser model but with substantially higher estimation error for $\tau = 0.3$ and 0.7.

Table 1: Simulation results ($p = 400$)

Method	Size	P1	P2	AE
LS-Lasso	25.08 (0.60)	100%	6%	1.37(0.03)
Q-Lasso ($\tau = 0.5$)	24.43 (0.97)	100%	6%	0.95 (0.03)
Q-Lasso ($\tau = 0.3$)	29.83 (0.97)	99%	99%	1.67 (0.05)
Q-Lasso ($\tau = 0.7$)	29.65 (0.90)	98%	98%	1.58 (0.05)
LS-ALASSO	5.02 (0.08)	100%	0%	0.38 (0.02)
Q-Alasso ($\tau = 0.5$)	4.66 (0.09)	100%	1%	0.18 (0.01)
Q-Alasso ($\tau = 0.3$)	6.98 (0.20)	100%	92%	0.63 (0.02)
Q-Alasso ($\tau = 0.7$)	6.43 (0.15)	100%	98%	0.61 (0.02)
LS-SCAD	5.83 (0.20)	100%	0%	0.37 (0.01)
Q-SCAD ($\tau = 0.5$)	5.86 (0.24)	100%	0%	0.19 (0.01)
Q-SCAD ($\tau = 0.3$)	8.29 (0.34)	99%	99%	0.32 (0.02)
Q-SCAD ($\tau = 0.7$)	7.96 (0.30)	97%	97%	0.30 (0.02)
LS-MCP	5.43 (0.17)	100%	0%	0.37 (0.01)
Q-MCP ($\tau = 0.5$)	5.33 (0.18)	100%	1%	0.19 (0.01)
Q-MCP ($\tau = 0.3$)	6.76 (0.25)	99%	99%	0.31 (0.02)
Q-MCP ($\tau = 0.7$)	6.66 (0.20)	97%	97%	0.29 (0.02)

Table 2: Simulation results ($p = 600$)

Method	Size	P1	P2	AE
LS-Lasso	24.30 (0.61)	100%	7%	1.40 (0.03)
Q-Lasso ($\tau = 0.5$)	25.76 (0.94)	100%	10%	1.05 (0.03)
Q-Lasso ($\tau = 0.3$)	34.02 (1.27)	93%	93%	1.82 (0.06)
Q-Lasso ($\tau = 0.7$)	32.74 (1.22)	90%	90%	1.78 (0.05)
LS-ALASSO	4.68 (0.08)	100%	0%	0.37(0.02)
Q-Alasso ($\tau = 0.5$)	4.53 (0.09)	100%	0%	0.18 (0.01)
Q-Alasso ($\tau = 0.3$)	6.58 (0.21)	100%	86%	0.67 (0.02)
Q-Alasso ($\tau = 0.7$)	6.19 (0.16)	100%	86%	0.62 (0.01)
LS-SCAD	6.04 (0.25)	100%	0%	0.38 (0.02)
Q-SCAD ($\tau = 0.5$)	6.14 (0.36)	100%	7%	0.19 (0.01)
Q-SCAD ($\tau = 0.3$)	9.02 (0.45)	94%	94%	0.40 (0.03)
Q-SCAD ($\tau = 0.7$)	9.97 (0.54)	100%	100%	0.38 (0.03)
LS-MCP	5.56 (0.19)	100%	0%	0.38 (0.02)
Q-MCP ($\tau = 0.5$)	5.33 (0.23)	100%	3%	0.18 (0.01)
Q-MCP ($\tau = 0.3$)	6.98 (0.28)	94%	94%	0.38 (0.03)
Q-MCP ($\tau = 0.7$)	7.56 (0.32)	98%	98%	0.37 (0.03)

3.2 An Application

We now illustrate the proposed methods by an empirical analysis of a real data set. The data set came from a study that used expression quantitative trait locus (eQTL) mapping in laboratory rats to investigate gene regulation in the mammalian eye and to identify genetic variation relevant to human eye disease (Scheetz et al., 2006).

This microarray data set has expression values of 31042 probe sets on 120 twelve-week-old male offspring of rats. We carried out the following two preprocessing steps: remove each probe for which the maximum expression among the 120 rats is less than the 25th percentile of the entire expression values; and remove any probe for which the range of the expression among 120 rats is less than 2. After these two preprocessing steps, there are 18958 probes left. As in Huang, Ma, and Zhang (2008) and Kim, Choi, and Oh (2008), we study how expression of gene TRIM32 (a gene identified to be associated with human hereditary diseases of the retina), corresponding to probe

1389163_at, depends on expressions at other probes. As pointed out in Scheetz et al. (2006), "Any genetic element that can be shown to alter the expression of a specific gene or gene family known to be involved in a specific disease is itself an excellent candidate for involvement in the disease, either primarily or as a genetic modifier." We rank all other probes according to the absolute value of the correlation of their expression and the expression corresponding to 1389163_at and select the top 300 probes. Then we apply several methods on these 300 probes.

First, we analyze the complete data set of 120 rats. The penalized least squares procedures and the penalized quantile regression procedures studied in Section 3.1 were applied. We use five-fold cross validation to select the tuning parameter for each method. In the second column of Table 3, we report the number of nonzero coefficients (# nonzero) selected by each method.

Table 3: Analysis of microarray data set

	all data	random partition	
Method	# nonzero	ave # nonzero	prediction error
LS-Lasso	24	21.66(1.67)	1.57(0.03)
Q-Lasso ($\tau = 0.5$)	23	18.36(0.83)	1.51(0.03)
Q-Lasso ($\tau = 0.3$)	23	19.34(1.69)	1.54(0.04)
Q-Lasso ($\tau = 0.7$)	17	15.54(0.71)	1.29(0.02)
LS-ALASSO	16	15.22(10.72)	1.65(0.27)
Q-ALasso ($\tau = 0.5$)	13	11.28(0.65)	1.53(0.03)
Q-ALasso ($\tau = 0.3$)	19	12.52(1.38)	1.57(0.03)
Q-ALasso ($\tau = 0.7$)	10	9.16(0.48)	1.32(0.03)
LS-SCAD	10	11.32(1.16)	1.72(0.04)
Q-SCAD ($\tau = 0.5$)	23	18.32(0.82)	1.51(0.03)
Q-SCAD ($\tau = 0.3$)	23	17.66(1.52)	1.56(0.04)
Q-SCAD ($\tau = 0.7$)	19	15.72(0.72)	1.30(0.03)
LS-MCP	5	9.08(1.68)	1.82(0.04)
Q-MCP ($\tau = 0.5$)	23	17.64(0.82)	1.52(0.03)
Q-MCP ($\tau = 0.3$)	15	16.36(1.53)	1.57(0.04)
Q-MCP ($\tau = 0.7$)	16	13.92(0.72)	1.31(0.03)

There are two interesting findings. First, the sizes of the models selected by pe-

nalized least squares methods are different from that of models selected by penalized quantile regression. In particular, both LS-SCAD and LS-MCP, which focus on the mean of the conditional distribution, select sparser models compared to Q-SCAD and Q-MCP. A sensible interpretation is that a probe may display strong association with the target probe only at the upper tail or lower tail of the conditional distribution; it is also likely that a probe may display associations in opposite directions at the two tails. The least-squares based methods are likely to miss such heterogeneous signals. Second, a more detailed story is revealed when we compare the probes selected at different quantiles $\tau = 0.3, 0.5, 0.7$. The probes selected by Q-SCAD(0.3), Q-SCAD(0.5), and Q-SCAD(0.7) are reported in the first column of the left, center and right panels, respectively, of Table 4. Although Q-SCAD selects 23 probes at both $\tau = 0.5$ and $\tau = 0.3$, only 7 of the 23 overlap, and only 2 probes (1382835_at and 1393382_at) are selected at all three quantiles. We observe similar phenomenon with Q-MCP. This further demonstrates the heterogeneity in the data.

We then conduct 50 random partitions. For each partition, we randomly select 80 rats as the training data and the other 40 as the testing data. A five-fold cross-validation is applied to the training data to select the tuning parameters. We report the average number of nonzero regression coefficients (ave # nonzero), where numbers in the parentheses are the corresponding standard errors across 50 partitions, in the third column of Table 3. We evaluate the performance over the test set for each partition. For Q-SCAD and Q-MCP, we evaluate the loss using the check function at the corresponding τ. As the squared loss is not directly comparable with the check loss function, we use the check loss with $\tau = 0.5$ (i.e. L_1 loss) for the LS-based methods. The results are reported in the last column of Table 3, where the prediction error is defined as $\sum_{i=1}^{40} \rho_\tau(y_i - \hat{y}_i)$ and the numbers in the parentheses are the corresponding

standard errors across 50 partitions. We observe similar patterns as when the methods are applied to the whole data set. Furthermore, the penalized quantile regression procedures improves the corresponding penalized least squares in terms of prediction error. The performance of Q-Lasso, Q-ALasso, Q-SCAD and Q-MCP are similar in terms of prediction error, although the Q-Lasso tends to select less sparse models and the Q-ALasso tends to select sparser model, compared with Q-SCAD and Q-MCP.

As with every variable selection method, different repetitions may select different subsets of important predictors. In Table 4, we report in the left column the probes selected using the complete data set and in the right column the frequency these probes appear in the final model of these 50 random partitions for Q-SCAD(0.3), Q-SCAD(0.5), and Q-SCAD(0.7) in the left, middle and right panels, respectively. The probes are ordered such that the frequency is decreasing. From Table 4, we observe that some probes such as 1383996_at and 1382835_at have high frequencies across different τ's, while some other probes such as 1383901_at do not. This implies that some probes are important across all τ, while some probes might be important only for certain τ.

Wei and He (2006) proposed a simulation based graphical method to evaluate the overall lack-of-fit of the quantile regression process. We apply their graphical diagnosis method using the SCAD penalized quantile regression. More explicitly, we first generate a random $\tilde{\tau}$ from the uniform (0,1) distribution. We then fit the SCAD-penalized quantile regression model at the quantile $\tilde{\tau}$, where the regularization parameter is selected by five-fold cross-validation. Denote the penalized estimator by $\widehat{\boldsymbol{\beta}}(\tilde{\tau})$, and we generate a response $Y = \mathbf{x}^T\widehat{\boldsymbol{\beta}}(\tilde{\tau})$, where $\mathbf{x}$ is randomly sampled from the set of observed vector of covariates. We repeat this process 200 times and produce a sample of 200 simulated responses from the postulated linear model. The QQ plot

Table 4: Frequency table for the real data

Q-SCAD(0.3)		Q-SCAD(0.5)		Q-SCAD(0.7)	
Probe	Frequency	Probe	Frequency	Probe	Frequency
1383996_at	31	1383996_at	43	1379597_at	38
1389584_at	26	1382835_at	40	1383901_at	34
1393382_at	24	1390401_at	27	1382835_at	34
1397865_at	24	1383673_at	24	1383996_at	34
1370429_at	23	1393382_at	24	1393543_at	30
1382835_at	23	1395342_at	23	1393684_at	27
1380033_at	22	1389584_at	21	1379971_at	23
1383749_at	20	1393543_at	20	1382263_at	22
1378935_at	18	1390569_at	20	1393033_at	19
1383604_at	15	1374106_at	18	1385043_at	18
1379920_at	13	1383901_at	18	1393382_at	17
1383673_at	12	1393684_at	16	1371194_at	16
1383522_at	11	1390788_a_at	16	1383110_at	12
1384466_at	10	1394399_at	14	1395415_at	6
1374126_at	10	1383749_at	14	1383502_at	6
1382585_at	10	1395415_at	13	1383254_at	5
1394596_at	10	1385043_at	12	1387713_a_at	5
1383849_at	10	1374131_at	10	1374953_at	3
1380884_at	7	1394596_at	10	1382517_at	1
1369353_at	5	1385944_at	9		
1377944_at	5	1378935_at	9		
1370655_a_at	4	1371242_at	8		
1379567_at	1	1379004_at	8		

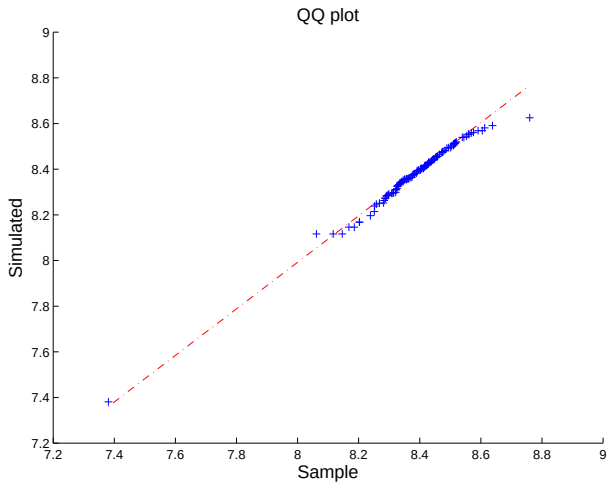

Figure 1: Lack-of-fit diagnosis QQ plot for the real data.

of the simulated sample vs the observed sample is given in Figure 1. The QQ plot is close to 45 degree line and thus indicates a reasonable fit.

4 Discussions

In this paper, we investigate nonconvex penalized quantile regression for analyzing ultra-high dimensional data under the assumption that at each quantile only a small subset of covariates are active but the active covariates at different quantiles may be

different. We establish the theory of the proposed procedures in ultra-high dimension under very relaxed conditions. In particular, the theory suggests that nonconvex penalized quantile regression with ultra-high dimensional covariates has the oracle property even when the random errors have heavy tails. In contrast, the existing theory for ultra-high dimensional nonconvex penalized least squares regression needs Gaussian or Sub-Gaussian condition for the random errors.

The theory was established by novelly applying a sufficient optimality condition based on a convex differencing representation of the penalized loss function. This approach can be applied to a large class of nonsmooth loss functions, for example the loss function corresponding to Huber's estimator and the many loss functions used for classification. As pointed out by a referee, the current theory only proves that the oracle estimator is a local minimum to the penalized quantile regression. How to identify the oracle estimator from potentially multiple minima is a challenging issue, which remains unsolved for nonconvex penalized least squares regression. This will be a good future research topic. The simulations suggest that the local minimum identified by our algorithm has fine performance.

Alternative methods for analyzing ultra-high dimensional data include two-stage approaches, in which a computationally efficient method screens all candidate covariates (for example Fan and Lv (2008), Wang (2009), Meinshausen and Yu (2009), among others) and reduces the ultra-high dimensionality to moderately high dimensionality in the first stage, and a standard shrinkage or thresholding method is applied to identify important covariates in the second stage.

References

[1] An, L.T.H. and Tao, P. D. (2005). The DC (Difference of Convex Functions) programming and DCA revisited with DC models of real world nonconvex opti-

mization problems. *Annals of Operations Research*, **133**, 23 - 46.

[2] Bai, Z. and Wu, Y. (1994). Limiting behavior of M-estimators of regression coefficients in high dimensional linear models, I. Scale-dependent case. *Journal of Multivariate Analysis*, **51**, 211-239.

[3] Belloni, A. and Chernozhukov, V. (2011). L1-Penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*. **39**, 82-130.

[4] Bertsekas, D. P. (2008), *Nonlinear programming*, third edition, Athena Scientific, Belmont, Massachusetts.

[5] Candes, E.J. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, **35**, 2313 - 2351.

[6] Cands, E. J., Wakin, M. and Boyd, S. (2007). Enhancing sparsity by reweighted L_1 minimization. *J. Fourier Anal. Appl.*, **14**, 877 - 905.

[7] Collobert, R., Sinz, F., Weston. J. and Bottou, L. (2006) Large scale transductive SVMs. *Journal of Machine Learning Research*, **7**, 1687 - 1712.

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

[9] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of Royal Statistical Society, Series B*, **70**, 849 - 911.

[10] Fan, J. and Lv, J. (2009). Non-concave penalized likelihood with NP-dimensionality. To appear in *IEEE Transactions on Information Theory*.

[11] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, **32**, 928 - 961.

[12] He, X. (2009) Modeling and inference by quantile regression. Technical report, Department of Statistics, University of Illinois at Urbana-Champaign.

[13] He, X. and Shao, Q. M.(2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis*, **73**, 120 - 135.

[14] He, X. and Zhu, L. X. (2003). A lack-of-fit test for quantile regression. *Journal of the American Statistical Association* , **98**, 1013-1022.

[15] Huang, J., Ma. S.G. and Zhang, C-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, **18**, 1603-1618.

[16] Kai, B., Li, R. and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics*, **39**, 305-332

[17] Kim, Y., Choi, H. and Oh, H-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, **103**, 1665 - 1673.

[18] Koenker, R. (2005), *Quantile regression*, Cambridge University Press.

[19] Koenker, R. and Bassett, G.W. (1978) Regression quantiles. *Econometrica*, **46**, 33 - 50.

[20] Li, Y. J. and Zhu. J. (2008) L_1-norm quantile regression. *Journal of Computational and Graphical Statistics*, **17**, 163-185.

[21] Liu, Y. F., Shen, X. and Doss, H. (2005). Multicategory psi-learning and support vector machine: computational tools. *Journal of Computational and Graphical Statistics*, **14**, 219-236.

[22] Mazumder, R, Friedman, J. and Hastie, T. (2009). SparseNet: Coordinate descent with nonconvex penalties. In press, *Journal of the American Statistical Association*.

[23] Scheetz, T.E., Kim, K.-Y. A., Swiderski, R.E., Philp, A.R., Braun, T.A., Knudtson, K.L., Dorrance, A.M., DiBona, G.F., Huang, J., Casavant, T.L., Sheffield, V.C. and Stone, E.M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, **103**, 14429-14434.

[24] Tao, P. D. and An, L.T.H. (1997). Convex analysis approach to D.C. programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, **22**, 289-355.

[25] Wang, H. (2009) Forward regression for ultra-highdimensional variable screening. *Journal of the American Statistical Association*, **104**, 1512 - 1524.

[26] Wei, Y and He, X. (2006). Conditional Growth Charts (with discussions). *Annals of Statistics*, **34**, 2069-2097 and 2126-2131.

[27] Welsh, A. H. (1989). On M-Processes and M-Estimation. *The Annals of Statistics*, **17**, 337 - 361.

[28] Wu, Y. C. and Liu, Y. F. (2009). Variable selection in quantile regression. *Statistica Sinica*, **19**, 801 - 817.

[29] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894 - 942.

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

[31] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *The Annals of Statistics*, **36**, 1509 - 1533.

[32] Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, **36**, 1108 - 1126.

Appendix: Technical Proofs

Throughout the proof, we use C to denote a generic positive constant, which does not depend on n and may vary from line to line.

Proof of Lemma 2.2 relies on the following Lemma 4.1.

Lemma 4.1 *Assume that conditions (C1)-(C5) are satisfied. The oracle estimator* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T$ *satisfies* $||\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}|| = O_p(\sqrt{q/n})$ *as* $n \to \infty$.

Proof. The result can be established by using the techniques of He and Shao (2000) on M-estimation. A more straightforward proof that directly explores the structure of quantile regression is given in the earlier version of this paper, which is available from the authors upon request.

Proof of Lemma 2.2. Note that the unpenalized quantile loss objective function is convex. By the convex optimization theory, $\mathbf{0} \in \partial \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{z}_i^T \widehat{\boldsymbol{\beta}}_1)$. Therefore there exists v_i^* such that $s_j(\widehat{\boldsymbol{\beta}}) = 0$ with $v_i = v_i^*$ for $j = 0, 1, \ldots, q$, and (3) holds. To prove (3), it suffices to show that

$$P\left(|\widehat{\beta}_j| \geq (a + 1/2)\lambda, \text{ for } j = 1, \ldots, q\right) \to 1 \tag{8}$$

as $n \to \infty$. Note that $\min_{1 \leq j \leq q} |\widehat{\beta}_j| \geq \min_{1 \leq j \leq q} |\beta_{0j}| - \max_{1 \leq j \leq q} |\widehat{\beta}_j - \beta_{0j}|$. Furthermore, $\min_{1 \leq j \leq q} |\beta_{0j}| \geq M_4 n^{-(1-c_2)/2}$ by condition (C5), and $\max_{1 \leq j \leq q} |\widehat{\beta}_j - \beta_{0j}| \leq ||\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}|| = O_p(\sqrt{q/n}) = O_p(n^{-(1-c_1)/2}) = o_p(n^{-(1-c_2)/2})$ by Lemma 4.1 and condition (C5). Thus (8) follows by the assumption $\lambda = o(n^{-(1-c_2)/2})$. □

The proof of Lemma 2.3 relies on technical results in Lemmas 4.2 and 4.3 below.

Lemma 4.2 *Assume that conditions (C1)-(C5) are satisfied and that* $\log p = o(n\lambda^2)$ *and* $n\lambda^2 \to \infty$. *We have*

$$P\left(\max_{q+1\le j\le p} n^{-1}\left|\sum_{i=1}^{n} x_{ij}[I(Y_i - \mathbf{z}_i^T\boldsymbol{\beta}_{01} \le 0) - \tau]\right| > \lambda/2\right) \to 0$$

as $n \to \infty$.

Proof. Since $I(Y_i - \mathbf{z}_i^T\boldsymbol{\beta}_{01} \le 0)$, $i = 1, \cdots, n.$ are i.i.d. Bernoulli random variables with mean τ, and x_{ij}, $q+1 \le j \le p$ are uniformly bounded, it holds

$$P\left(n^{-1}\left|\sum_{i=1}^{n} x_{ij}[I(Y_i - \mathbf{z}_i^T\boldsymbol{\beta}_{01} \le 0) - \tau]\right| > \lambda/2\right) \le \exp\left(-Cn\lambda^2\right).$$

by Hoeffding's inequality. We have

$$\begin{aligned} & P\left(\max_{q+1\le j\le p} n^{-1}\left|\sum_{i=1}^{n} x_{ij}[I(Y_i - \mathbf{z}_i^T\boldsymbol{\beta}_{01} \le 0) - \tau]\right| > \lambda/2\right) \\ \le\ & 2p\exp\left(-Cn\lambda^2\right) = 2\exp(\log p - Cn\lambda^2) \to 0, \end{aligned}$$

under the conditions of the lemma. □

Lemma 4.3 *Assume that conditions (C1)-(C5) are satisfied and that* $q\log(n) = o(n\lambda)$, $\log p = o(n\lambda^2)$ *and* $n\lambda \to \infty$. *Then* $\forall\ \Delta > 0$, *as* $n \to \infty$,

$$\begin{aligned} & P\Big(\max_{q+1\le j\le p} \sup_{||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}|| \le \Delta\sqrt{q/n}} \Big|\sum_{i=1}^{n} x_{ij}\big[I(Y_i - \mathbf{z}_i^T\boldsymbol{\beta}_1 \le 0) - I(Y_i - \mathbf{z}_i^T\boldsymbol{\beta}_{01} \le 0) \\ & - P(Y_i - \mathbf{z}_i^T\boldsymbol{\beta}_1 \le 0) + P(Y_i - \mathbf{z}_i^T\boldsymbol{\beta}_{01} \le 0)\big]\Big| > n\lambda\Big) \to 0. \end{aligned} \tag{9}$$

Proof. We generalize an approach by Welsh (1989). We cover the ball $\{\boldsymbol{\beta}_1 : ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}|| \le \Delta\sqrt{q/n}\}$ with a net of balls with radius $\Delta\sqrt{q/n^5}$. It can be shown that this

net can be constructed with cardinality $N \le d \cdot n^{4q}$ for some constant $d > 0$. Denote the N balls by $B(\mathbf{t}_1), \ldots, B(\mathbf{t}_N)$, where the ball $B(\mathbf{t}_k)$ is centered at $\mathbf{t}_k$, $k = 1, \ldots, N$.

To simplify the notation, let $\kappa_i(\boldsymbol{\beta}_1) = Y_i - \mathbf{z}_i^T \boldsymbol{\beta}_1$. Then

$$
\begin{aligned}
& P\Big(\sup_{||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}|| \le \Delta\sqrt{q/n}} \Big| \sum_{i=1}^{n} x_{ij} \big[I(\kappa_i(\boldsymbol{\beta}_1) \le 0) - I(\kappa_i(\boldsymbol{\beta}_{01}) \le 0) \\
& \qquad - P(\kappa_i(\boldsymbol{\beta}_1) \le 0) + P(\kappa_i(\boldsymbol{\beta}_{01}) \le 0) \big] \Big| > n\lambda \Big) \\
\le \; & \sum_{k=1}^{N} P\Big(\Big| \sum_{i=1}^{n} x_{ij} \big[I(\kappa_i(\mathbf{t}_k) \le 0) - I(\kappa_i(\boldsymbol{\beta}_{01}) \le 0) - P(\kappa_i(\mathbf{t}_k) \le 0) \\
& + P(\kappa_i(\boldsymbol{\beta}_{01}) \le 0) \big] \Big| > \frac{n\lambda}{2} \Big) \\
& + \sum_{k=1}^{N} P\Big(\sup_{||\widetilde{\boldsymbol{\beta}}_1 - \mathbf{t}_k|| \le \Delta\sqrt{q/n^5}} \Big| \sum_{i=1}^{n} x_{ij} \big[I(\kappa_i(\widetilde{\boldsymbol{\beta}}_1) \le 0) - I(\kappa_i(\mathbf{t}_k) \le 0) \\
& \qquad - P(\kappa_i(\widetilde{\boldsymbol{\beta}}_1) \le 0) + P(\kappa_i(\mathbf{t}_k) \le 0) \big] \Big| > \frac{n\lambda}{2} \Big) \\
\triangleq \; & J_{nj1} + J_{nj2}.
\end{aligned}
$$

To evaluate J_{nj1}, let $u_i = x_{ij}\big[I(\kappa_i(\mathbf{t}_k) \le 0) - I(\kappa_i(\boldsymbol{\beta}_{01}) \le 0) - P(\kappa_i(\mathbf{t}_k) \le 0) + P(\kappa_i(\boldsymbol{\beta}_{01}) \le 0) \big]$. Then the u_i are independent mean-zero random variables, and

$$
\begin{aligned}
\mathrm{Var}(u_i) \; = \; & x_{ij}^2 \big[F_i(\mathbf{z}_i^T(\mathbf{t}_k - \boldsymbol{\beta}_{01}) | \mathbf{z}_i)(1 - F_i(\mathbf{z}_i^T(\mathbf{t}_k - \boldsymbol{\beta}_{01}) | \mathbf{z}_i)) + F_i(0|\mathbf{z}_i)(1 - F_i(0|\mathbf{z}_i)) \\
& - 2F_i(\min(\mathbf{z}_i^T(\mathbf{t}_k - \boldsymbol{\beta}_{01}), 0) | \mathbf{z}_i) + 2F_i(\mathbf{z}_i^T(\mathbf{t}_k - \boldsymbol{\beta}_{01}) | \mathbf{z}_i) F_i(0|\mathbf{z}_i) \big] \\
\le \; & C \big| \mathbf{z}_i^T(\mathbf{t}_k - \boldsymbol{\beta}_{01}) \big|.
\end{aligned}
$$

Thus $\sum_{i=1}^{n} \mathrm{Var}(u_i) \le nC \max_i ||\mathbf{z}_i|| \cdot ||\mathbf{t}_k - \boldsymbol{\beta}_{01}|| = nO(\sqrt{q})O(\sqrt{q/n}) = O(\sqrt{n}q)$. Applying Bernstein's inequality,

$$
J_{nj1} \le N \cdot \exp\left(- \frac{n^2\lambda^2/4}{2\sqrt{n}q + Cn\lambda} \right) \le N \cdot \exp(-Cn\lambda) \le C \exp(4q \log n - Cn\lambda) \quad (10)
$$

To evaluate J_{nj2}, note that the function $I(x \le s)$ is increasing in s. Therefore,

$$\begin{aligned}
&\sup_{||\widetilde{\boldsymbol{\beta}}_1 - \mathbf{t}_k|| \le \Delta\sqrt{q/n^5}} \Big| \sum_{i=1}^{n} x_{ij} \big[I(\kappa_i(\widetilde{\boldsymbol{\beta}}_1) \le 0) - I(\kappa_i(\mathbf{t}_k) \le 0) - P(\kappa_i(\widetilde{\boldsymbol{\beta}}_1) \le 0) \\
&\qquad\qquad + P(\kappa_i(\mathbf{t}_k) \le 0) \big] \Big| \\
\le\ & \sum_{i=1}^{n} |x_{ij}| \big[I(\kappa_i(\mathbf{t}_k) \le \Delta\sqrt{q/n^5}||\mathbf{z}_i||) - I(\kappa_i(\mathbf{t}_k) \le 0) - P(\kappa_i(\mathbf{t}_k) \le -\Delta\sqrt{q/n^5}||\mathbf{z}_i||) \\
& + P(\kappa_i(\mathbf{t}_k) \le 0) \big] \\
=\ & \sum_{i=1}^{n} |x_{ij}| \big[I(\kappa_i(\mathbf{t}_k) \le \Delta\sqrt{q/n^5}||\mathbf{z}_i||) - I(\kappa_i(\mathbf{t}_k) \le 0) - P(\kappa_i(\mathbf{t}_k) \le \Delta\sqrt{q/n^5}||\mathbf{z}_i||) + \\
& P(\kappa_i(\mathbf{t}_k) \le 0) \big] + \sum_{i=1}^{n} |x_{ij}| \big[P(\kappa_i(\mathbf{t}_k) \le \Delta\sqrt{q/n^5}||\mathbf{z}_i||) - P(\kappa_i(\mathbf{t}_k) \le -\Delta\sqrt{q/n^5}||\mathbf{z}_i||) \big].
\end{aligned}$$

Note that

$$\begin{aligned}
& \sum_{i=1}^{n} |x_{ij}| \big[P(\kappa_i(\mathbf{t}_k) \le \Delta\sqrt{q/n^5}||\mathbf{z}_i||) - P(\kappa_i(\mathbf{t}_k) \le -\Delta\sqrt{q/n^5}||\mathbf{z}_i||) \big] \\
=\ & \sum_{i=1}^{n} |x_{ij}| \big[F_i(\Delta\sqrt{q/n^5}||\mathbf{z}_i|| + \mathbf{z}_i^T(\mathbf{t}_k - \boldsymbol{\beta}_{01})|\mathbf{z}_i) - F_i(-\Delta\sqrt{q/n^5}||\mathbf{z}_i|| + \mathbf{z}_i^T(\mathbf{t}_k - \boldsymbol{\beta}_{01})|\mathbf{z}_i) \big] \\
\le\ & C \sum_{i=1}^{n} |x_{ij}| \sqrt{q/n^5}||\mathbf{z}_i|| \le Cn\sqrt{q/n^5}\sqrt{q} = Cqn^{-3/2}.
\end{aligned}$$

Thus

$$\begin{aligned}
J_{nj2} \ =\ & \sum_{k=1}^{N} P\Big(\sum_{i=1}^{n} |x_{ij}| \big[I(\kappa_i(\mathbf{t}_k) \le \Delta\sqrt{q/n^5}||\mathbf{z}_i||) - I(\kappa_i(\mathbf{t}_k) \le 0) \\
& - P(\kappa_i(\mathbf{t}_k) \le \Delta\sqrt{q/n^5}||\mathbf{z}_i||) + P(\kappa_i(\mathbf{t}_k) \le 0) \big] \ge \frac{n\lambda}{2} \\
& - \sum_{i=1}^{n} |x_{ij}| \big[P(\kappa_i(\mathbf{t}_k) \le \Delta\sqrt{q/n^5}||\mathbf{z}_i||) - P(\kappa_i(\mathbf{t}_k) \le -\Delta\sqrt{q/n^5}||\mathbf{z}_i||) \big] \Big)
\end{aligned}$$

and

$$\begin{aligned} J_{nj2} &\leq \sum_{k=1}^{N} P\Big(\sum_{i=1}^{n} |x_{ij}| \big[I(\kappa_i(\mathbf{t}_k) \leq \Delta\sqrt{q/n^5}||\mathbf{z}_i||) - I(\kappa_i(\mathbf{t}_k) \leq 0) \\ &\quad - P(\kappa_i(\mathbf{t}_k) \leq \Delta\sqrt{q/n^5}||\mathbf{z}_i||) + P(\kappa_i(\mathbf{t}_k) \leq 0) \big] \geq \frac{n\lambda}{2} - Cqn^{-3/2} \Big) \\ &\leq \sum_{k=1}^{N} P\Big(\sum_{i=1}^{n} |x_{ij}| \big[I(\kappa_i(\mathbf{t}_k) \leq \Delta\sqrt{q/n^5}||\mathbf{z}_i||) - I(\kappa_i(\mathbf{t}_k) \leq 0) \\ &\quad - P(\kappa_i(\mathbf{t}_k) \leq \Delta\sqrt{q/n^5}||\mathbf{z}_i||) + P(\kappa_i(\mathbf{t}_k) \leq 0) \big] \geq \frac{n\lambda}{4} \Big), \end{aligned}$$

for all n sufficiently large since $qn^{-3/2} = o(1)$ by condition (C4). Let $v_i = |x_{ij}|\big[I(\kappa_i(\mathbf{t}_k) \leq \Delta\sqrt{q/n^5}||\mathbf{z}_i||) - I(\kappa_i(\mathbf{t}_k) \leq 0) - P(\kappa_i(\mathbf{t}_k) \leq \Delta\sqrt{q/n^5}||\mathbf{z}_i||) + P(\kappa_i(\mathbf{t}_k) \leq 0)\big]$. Then the v_i are independent zero-mean random variables, and

$$\begin{aligned} \mathrm{Var}(v_i) &\leq x_{ij}^2 E\left[I(\kappa_i(\mathbf{t}_k) \leq \Delta\sqrt{q/n^5}||\mathbf{z}_i||) - I(\kappa_i(\mathbf{t}_k) \leq 0) \right]^2 \\ &\leq C\sqrt{q/n^5}||\mathbf{z}_i|| = Cqn^{-5/2}. \end{aligned}$$

Thus by Bernstein's inequality, for some positive constants C_1, C_2 and C_3,

$$J_{nj2} \leq N \cdot 2\exp\left(-\frac{n^2\lambda^2/16}{C_1 qn^{-3/2} + C_2 n\lambda} \right) \leq 2N \cdot \exp(-Cn\lambda) \leq C\exp(4q\log n - Cn\lambda). \tag{11}$$

Finally, by (10) and (11), we have that the probability in (9) is bounded by

$$\sum_{j=q+1}^{p} (J_{nj1} + J_{nj2}) \leq C\exp(\log p + 4q\log n - Cn\lambda) = o(1),$$

under the assumptions of the lemma. This completes the proof. □

Proof of Lemma 2.3. By definition of the oracle estimator, $\widehat{\beta}_j = 0$, for $j =$

$q+1,\ldots,p$. We only need to show that

$$P(|s_j(\widehat{\boldsymbol{\beta}})| > \lambda, \quad \text{for some } j = q+1,\ldots,p) \to 0 \tag{12}$$

as $n \to \infty$. Let $\mathcal{D} = \{i : Y_i - \mathbf{z}_i^T \widehat{\boldsymbol{\beta}}_1 = 0\}$, then for $j = q+1,\ldots,p$

$$s_j(\widehat{\boldsymbol{\beta}}) = n^{-1}\sum_{i=1}^{n} x_{ij}[I(Y_i - \mathbf{z}_i^T\widehat{\boldsymbol{\beta}}_1 \le 0) - \tau] - n^{-1}\sum_{i\in\mathcal{D}} x_{ij}(v_i^* + (1-\tau)),$$

where $v_i^* \in [\tau - 1, \tau]$ with $i \in \mathcal{D}$ satisfies $s_j(\widehat{\boldsymbol{\beta}}) = 0$, for $j = 1,\ldots,q$, when $v_i = v_i^*$. By Condition (C2), with probability one there exists exactly $q+1$ elements in $\mathcal{D}$ (Section 2.2, Koenker, 2005). Then by condition (C1), with probability one

$$\max_{q+1\le j\le p} \left|n^{-1}\sum_{i\in\mathcal{D}} x_{ij}(v_i^* + (1-\tau))\right| = O(qn^{-1}) = o(\lambda),$$

under the assumptions of the lemma. Thus to prove (12), it suffices to show that

$$P\left(\left|n^{-1}\sum_{i=1}^{n} x_{ij}[I(Y_i - \mathbf{z}_i^T\widehat{\boldsymbol{\beta}}_1 \le 0) - \tau]\right| > \lambda, \text{ for some } j = q+1,\ldots,p\right) \to 0. \tag{13}$$

We have

$$\begin{aligned}
& P\left(\max_{q+1\le j\le p}\left|n^{-1}\sum_{i=1} x_{ij}[I(Y_i - \mathbf{z}_i^T\widehat{\boldsymbol{\beta}}_1 \le 0) - \tau]\right| > \lambda\right) \\
\le\ & P\left(\max_{q+1\le j\le p}\left|n^{-1}\sum_{i=1} x_{ij}[I(Y_i - \mathbf{z}_i^T\widehat{\boldsymbol{\beta}}_1 \le 0) - I(Y_i - \mathbf{z}_i^T\boldsymbol{\beta}_{01} \le 0)]\right| > \lambda/2\right) \\
& + P\left(\max_{q+1\le j\le p}\left|n^{-1}\sum_{i=1} x_{ij}[I(Y_i - \mathbf{z}_i^T\boldsymbol{\beta}_{01} \le 0) - \tau]\right| > \lambda/2\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq P\left(\max_{q+1\leq j\leq p}\left|n^{-1}\sum_{i=1}x_{ij}[I(Y_i-\mathbf{z}_i^T\widehat{\boldsymbol{\beta}}_1\leq 0)-I(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_{01}\leq 0)]\right|>\lambda/2\right)+o_p(1)\\
&\leq P\Big(\max_{q+1\leq j\leq p}\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\leq\Delta\sqrt{q/n}}\Big|n^{-1}\sum_{i=1}x_{ij}\big[I(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_1\leq 0)-I(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_{01}\leq 0)\\
&\quad -P(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_1\leq 0)+P(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_{01}\leq 0)\big]\Big|>\lambda/4\Big)\\
&\quad +P\Big(\max_{q+1\leq j\leq p}\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\leq\Delta\sqrt{q/n}}\Big|n^{-1}\sum_{i=1}x_{ij}\big[P(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_1\leq 0)\\
&\quad -P(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_{01}\leq 0)\big]\Big|>\lambda/4\Big)+o_p(1)\\
&\leq P\Big(\max_{q+1\leq j\leq p}\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\leq\Delta\sqrt{q/n}}\Big|n^{-1}\sum_{i=1}x_{ij}\big[P(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_1\leq 0)\\
&\quad -P(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_{01}\leq 0)\big]\Big|>\lambda/4\Big)+o_p(1),
\end{aligned}
$$

where the second inequality follows from Lemma 4.2, the their inequality follows from Lemma 4.1, and the last inequality follows from Lemma 4.3. Note that

$$
\begin{aligned}
&\max_{q+1\leq j\leq p}\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\leq\Delta\sqrt{q/n}}\Big|n^{-1}\sum_{i=1}x_{ij}\big[P(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_1\leq 0)-P(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_{01}\leq 0)\big]\Big|\\
=\;&\max_{q+1\leq j\leq p}\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\leq\Delta\sqrt{q/n}}n^{-1}\Big|\sum_{i=1}^{n}x_{ij}[F_i(\mathbf{z}_i^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})|\mathbf{z}_i)-F_i(0|\mathbf{z}_i)]\Big|\\
\leq\;&C\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\leq\Delta\sqrt{q/n}}n^{-1}\sum_{i=1}^{n}||\mathbf{z}_i||\cdot||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\\
=\;&O(\sqrt{q/n})O(\sqrt{q})=O(qn^{-1/2})=o(\lambda),
\end{aligned}
$$

where the inequality uses conditions (C1) and (C3). Thus

$$
\begin{aligned}
&P\Big(\max_{q+1\leq j\leq p}\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\leq\Delta\sqrt{q/n}}\Big|n^{-1}\sum_{i=1}x_{ij}\big[P(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_1\leq 0)-P(Y_i-\mathbf{z}_i^T\boldsymbol{\beta}_{01}\leq 0)\big]\Big|\\
&>\lambda/4\Big)=o(1).
\end{aligned}
$$

This proves (13). □

Proof of Theorem 2.4. We first check the condition in Lemma 2.1. From Lemma 2.2, there exist v_i^*, $i = 1, \ldots, n$, such that the subgradient function $s_j(\widehat{\boldsymbol{\beta}})$ defined with $v_i = v_i^*$ satisfies $P(s_j(\widehat{\boldsymbol{\beta}}) = 0,\ j = 0, 1, \ldots, q) \to 1$. Therefore, by the definition of the set $\partial g(\widehat{\boldsymbol{\beta}})$, we have $P(\mathbb{G} \subseteq \partial g(\widehat{\boldsymbol{\beta}})) \to 1$ where

$$\begin{aligned} \mathbb{G} \ = \ & \big\{\boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_p) : \xi_0 = 0;\ \xi_j = \lambda \mathrm{sgn}(\widehat{\beta}_j),\ j = 1, \ldots, q; \\ & \xi_j = s_j(\widehat{\boldsymbol{\beta}}) + \lambda l_j,\ j = q+1, \ldots, p.\big\} \end{aligned}$$

and l_j ranges over $[-1, 1]$, $j = q+1, \ldots, p$.

Consider any $\boldsymbol{\beta}$ in a ball in $\mathcal{R}^{p+1}$ with the center $\widehat{\boldsymbol{\beta}}$ and radius $\lambda/2$. To prove the theorem it is sufficient to show that there exists a vector $\boldsymbol{\xi}^* = (\xi_0^*, \xi_1^*, \ldots, \xi_p^*)^T$ in $\mathbb{G}$ such that

$$P\Big(\xi_j^* = \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}, \quad j = 0, 1, \ldots, p\Big) \to 1, \tag{14}$$

as $n \to \infty$.

By Lemma 2.3, $P(|s_j(\widehat{\boldsymbol{\beta}})| \le \lambda,\ j = q+1, \ldots, p) \to 1$; thus we can always find $l_j^* \in [-1, 1]$ such that $s_j(\widehat{\boldsymbol{\beta}}) + \lambda l_j^* = 0$, for $j = q+1, \ldots, p$. Let $\boldsymbol{\xi}^*$ be the vector in $\mathbb{G}$ with $l_j = l_j^*$, $j = q+1, \ldots, p$. We next verify that $\boldsymbol{\xi}^*$ satisfies (14).

(1) For $j = 0$, we have $\xi_0^* = 0$. Since $\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_0} = 0$ for both penalty functions, it is immediate that $\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_0} = \xi_0^*$.

(2) For $j = 1, \ldots, q$, we have $\xi_j^* = \lambda \mathrm{sgn}(\widehat{\beta}_j)$. We note that $\min_{1 \le j \le q} |\beta_j| \ge \min_{1 \le j \le q} |\widehat{\beta}_j| - \max_{1 \le j \le q} |\widehat{\beta}_j - \beta_j| \ge (a + 1/2)\lambda - \lambda/2 = a\lambda$ with probability approaching one by Lemma 2.2. Therefore, $P\big(\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = \lambda \mathrm{sgn}(\beta_j),\ j = 1, \ldots, q\big) \to 1$ as $n \to \infty$ for both the SCAD penalty and the MCP penalty. For n sufficiently large, $\widehat{\beta}_j$ and β_j have the same sign. Thus, $P\big(\xi_j^* = \frac{\partial h_\lambda(\boldsymbol{\beta})}{\partial \beta_j},\ j = 1, \ldots, q\big) \to 1$ as $n \to \infty$.

(3) For $j = q+1, \ldots, p$, we have $\xi_j^* = 0$ following the definition of $\boldsymbol{\xi}^*$. By Lemma 2.3, $P(|\beta_j| \leq |\widehat{\beta}_j| + |\widehat{\beta}_j - \beta_j| \leq \lambda,\ j = q+1, \ldots, p) \to 1$ as $n \to \infty$. Therefore $P\big(\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = 0,\ j = q+1, \ldots, p\big) \to 1$ for the SCAD penalty; and $P\big(\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = -\frac{\beta_j}{a},\ j = q+1, \ldots, p\big) \to 1$ for the MCP penalty. Note that for both penalty functions, we have $P\big(\big|\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}\big| \leq \lambda\big) \to 1$, for $j = q+1, \ldots, p$. By Lemma 2.3, with probability approaching one $|s_j(\widehat{\beta}_j)| \leq \lambda$, for $j = q+1, \ldots, p$. Thus, we can always find $l_j^* \in [-1, 1]$ such that $P\big(\xi_j^* = s_j(\widehat{\boldsymbol{\beta}}) + \lambda l_j^* = \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j},\ j = q+1, \ldots, p\big) \to 1$, as $n \to \infty$, for both penalty functions. This completes the proof. $\square$