Ultra-high dimensional variable selection for quantile regression

Xu Liu, Hongmei Jiang and Xingjie Shi

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China

Department of Statistics, Northwestern University, Evanston 60208

March 12, 2017

Abstract

Sparsity comes frequently in high or ultra-high dimensional data that arise nowadays in many research areas. In this paper, we study the problem of ultra-high dimensional variable section quantile regression. We develop a fast algorithm, coordinate descent minorization-maximization (CDMM), to compute effectively the solution path and to select the variables. To be specific, we first employ the minorization-maximization method to create a quadratic surrogate function for quantile regression. Then, the coordinate descent algorithm is used to solve the minimization of the penalized surrogate loss function. The proposed algorithm is significantly faster than the simplex algorithm which is usually used to solve problems with quatile regression. Under some regularity conditions, the convergence properties of the proposed algorithm are established. The algorithm can be easily extended to composite quantile regression. The simulation studies and one real data example demonstrate that the proposed method produces satisfactory results.

Keywords: Classification, LASSO, MCP, Penalized regression, Quantile regression, SCAD, SVM.

* Corresponding Author:

1 Introduction

High or ultra-high dimensional data arise nowadays in many research areas including genomics, signal processing, and biomedical imaging. It is common that the number of features p is much bigger than the sample size n. Recently, sparse modeling based on penalized likelihood or loss functions has been developed to handle high dimensionality. The widely used penalty functions include the L_1-norm as presented in the least absolute shrinkage and selection operator (LASSO, Tibshirani (1996)), the smoothly clipped absolute deviation (SCAD, Fan and Li (2001)), and the minimax concave penalty (MCP, Zhang (2010)).

Consider the linear regression model

$$\mathbf{y} = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is an n-dimensional response vector, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ is an $n \times p$ design matrix, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a p-dimensional vector of unknown parameters, and $\mathbf{e} \sim N(0, \sigma^2 \mathbf{I}_n)$ is the error term. When $p \gg n$, the ordinary least square method does not work. The penalized least square (PLS) is powerfully developed to minimize

$$PLS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \alpha - \mathbf{x}_i^T \boldsymbol{\beta})^2 + n \sum_{j=1}^{p} P(|\beta_j|; \lambda, \gamma),$$

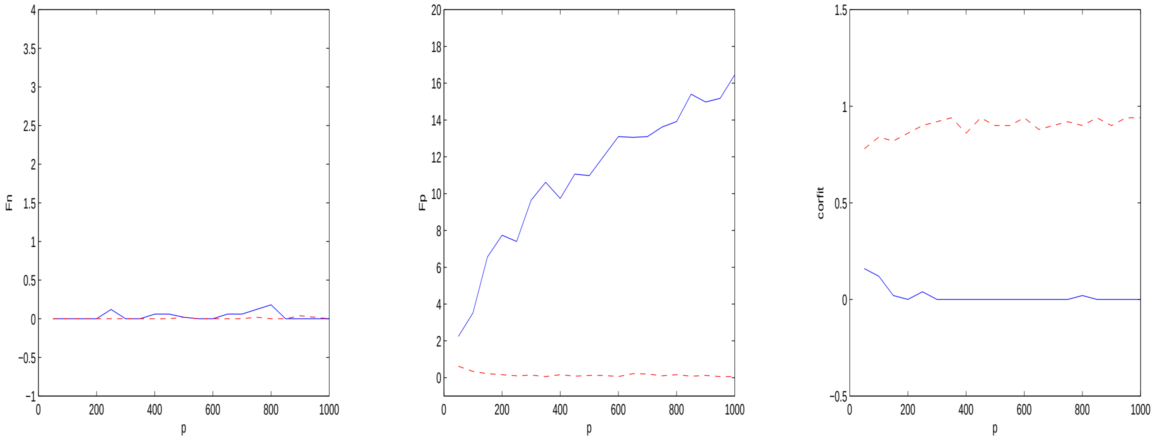
where $P(t; \lambda, \gamma)$ is a penalty function such as L_1-norm, MCP or SCAD.

When the error is normally distributed without large variance, PLS performs pretty well (Zhang 2010; Efron et al. 2004; Park and Hastie 2007; Zou and Li 2008). But if the error has a heavy-tailed distribution, for instance, normal distribution with large variance, t-distribution, or double exponential distribution, the PLS is not efficient. In the ordinary case with $p < n$, median or quantile regression (Koenker and Bassett 1978) has been used to deal with heavy-tailed distribution. For ultra-high dimensional data, the penalized median regression (PMR) or quantile regression (PQR) is preferred than PLS. The following example illustrates better performance of PMR comparing with PLS.

This example focuses on linear model (1.1) with $\mathbf{x} \sim N(0, \Sigma)$, where $\Sigma_{(i,j)} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$ and $\boldsymbol{\varepsilon} \sim t(3)$. We consider 20 different values of dimensionality with $p = 50, 100, \cdots, 1000$, and $\boldsymbol{\beta} = (2, 1, 0, 0, 3, \underline{0}_{p-5})^T$. For each value of p, 50 data sets are generated with sample size $n = 200$. In Table 1, we use three quantities to measure the variable

Table 1: Results for penalized least square (PLS) and penalized median regression (PMR).

	Fn		Fp		corfit	
p	PLS	PMR	PLS	PMR	PLS	PMR
200	0	0	7.74	.16	0	.86
400	.06	0	9.74	.16	0	.86
600	0	0	13.10	.06	0	.94
800	.18	0	13.92	.16	.02	.90
1000	0	0	16.46	.06	0	.94



(a) Fn

(b) Fp

(c) correct fit

Figure 1: *Results for penalized least square (PLS, solid line) and penalized median regression (PMR, dashed line) .*

selection performance of PLS and PMR with MCP penalty. False negative (Fn) is the average number of true non-zero coefficients incorrectly identified as zero, false positive (Fp) is the average number of true zero coefficients incorrectly identified as non-zero, and "corfit" is the proportion of simulation times where the underlying true model is selected. The tuning parameter is selected by 5-fold cross-validation (CV). Both PLS and PMR have low false negative rates which imply that the important variables are selected by both methods. However PLS selects more variables with true zero coefficient than PMR does. The chance to select exactly the underlying true model is almost zero using PLS, but is pretty high using PMR.

This example motivates us to consider quantile regression model (Koenker and Bassett

1978):

$$\rho_q(r) = qr - rI(r < 0) \tag{1.2}$$

where the indicator function $I(x \in A)$ equals to 1 if $x \in A$ and 0 otherwise. Especially, $\rho_{\frac{1}{2}}(r) = \frac{1}{2}|r|$ is the least absolute deviation with $q = 1/2$. Quantile regression (1.2) is a popular analysis tool in various areas such as economics (Koenker and Hollock 2001), time series (Cai and Xu 2009), and survival analysis (Koenker and Geling 2001), among others. More details can be referred to Koenker (2005) and He (2009) for general overview of many interesting developments.

In this paper, we study variable selection by minimizing the penalized piecewise linear loss function equipped with aforementioned penalty functions,

$$\arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} L(\boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{1.3}$$

where

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_q \left(y_i - \alpha - \mathbf{x}_i^T \boldsymbol{\beta} \right) + n \sum_{j=1}^{p} P(|\beta_j|; \lambda, \gamma).$$

There are two challenges to solve this problem. First of all, neither the piecewise linear loss function nor the penalty function is smooth at 0. We deal with the issue of non-smoothness of the piecewise linear loss function using the minorization-maximization algorithm (MM). The MM algorithm is widely used in various fields. The most important example of MM is the expectation-mamximization (EM) algorithm. Hunter and Lange (2000) studied carefully the MM algorithm for quantile regression. Hunter and Li (2005) considered variable selection using the MM algorithm to surrogate the non-smooth SCAD penalty function.

The second challenge is that this is a large p and small n problem. The coordinate descend algorithm (CD) can be used to solve the issues with non-smooth penalty function and large p problem. There exists a closed form for each of the three penalty functions, L_1-norm, SCAD and MCP, when only one predictor is considered for the penalized least square (Donoho and Johnstone 1994; Breheny and Huang 2011). Therefore we can estimate parameter $\boldsymbol{\beta}$ coordinately one by one. The convergence of coordinate algorithm for non-differentiable objective function has been fully discussed in Tseng (2001). Friedman et al. (2007) and Wu and Lange (2008) studied independently the L_1 penalized regression using the coordinate optimization. The coordinate descent algorithm is time-saving because it avoids computing the inverse of a big matrix for high dimensional data. Comparing with linear programming (LP) approach, the biggest merit of the CD algorithm is being computationally efficient.

Combining ideas of MM algorithm and CD algorithm, we propose a coordinate descend minorization-maximization algorithm (CDMM) for penalized piecewise linear loss function. The rest of the paper is organized as follows. We propose the CDMM algorithm for (1.2) in Section 2. In Section 3, the CDMM algorithm is extended to the composite quantile regression. How to select the tuning parameter and the perturbation ε is discussed in Section 4. In Section 5, We conduct several simulation studies and real data analysis to illustrate that our proposed procedure works well with small sample size. We conclude with discussion in Section 6. The technical proofs are given in the Appendix.

2 The CDMM algorithm for quantile regression

The existing method to solve penalized quantile regression is linear programming (LP). Wu and Liu (2009) and Wang et al. (2012) studied the penalized quantile regression using LP with difference of convex programming. Zou and Yuan (2008) used LP on minimizing penalized composite quantile regression. However the computational burden grows significantly as dimension p increases. The non-smoothness of piecewise linear loss function and ultra-high dimension are two great challenges. In current paper we propose a CDMM algorithm to deal with these two challenges, in which the MM algorithm is used to construct a surrogate function for the loss function and the CD algorithm is used to handle large p issue.

2.1 Penalties and soft-thresholding operator

The three popularly used penalty functions, the L_1-norm, the MCP, and the SCAD, belong to a family of quadratic spline form (Zhang 2010). For the L_1 penalty, $P(t;\lambda,\gamma) = \lambda t$, therefore it is biased (Fan and Li 2001). For the SCAD penalty,

$$P'(t;\lambda,\gamma) = \lambda I(t<\lambda) + \frac{(\gamma\lambda - t)_+}{\gamma - 1} I(t>\lambda),$$

for some $\gamma > 2$, where $P'(\cdot)$ denotes the derivative of function $P(\cdot)$. Often the regularization parameter $\gamma = 3.7$ is used (Fan and Li 2001). For the MCP penalty,

$$P'(t; \lambda, \gamma) = (\lambda - t/\gamma)\, I(t < \gamma\lambda),$$

where $\gamma > 1$. It can be seen that both the SCAD and the MCP are subject to unbiasedness and selection features. In addition, all three penalty functions produce continuous estimators so that instability in model prediction is avoided.

To simplify the explanation, we consider the case of linear regression with single standardized predictor. Let z denote the least squares estimate. The LASSO solution is the soft threshold of least squares estimate (Donoho and Johnstone 1994). That is, $\hat{\beta}^{lasso} = s(z, \lambda)$, where

$$s(z, \lambda) = \begin{cases} z - \lambda, & \text{if} \quad z > \lambda, \\ z + \lambda, & \text{if} \quad z < -\lambda, \\ 0 & \text{if} \quad |z| \le \lambda. \end{cases}$$

While the respective MCP solution and SCAD solution proposed by Breheny and Huang (2011) are

$$\hat{\beta}^{mcp} = f_{mcp}(z, \lambda, \gamma) = \begin{cases} \frac{s(z,\lambda)}{1-1/\gamma}, & \text{if} \quad |z| \le \gamma\lambda, \\ z & \text{if} \quad |z| \ge \gamma\lambda, \end{cases}$$

and

$$\hat{\beta}^{scad} = f_{scad}(z, \lambda, \gamma) = \begin{cases} s(z, \lambda), & \text{if} \quad |z| \le 2\lambda, \\ \frac{s(z,\gamma\lambda/(\gamma-1))}{1-1/(\gamma-1)}, & \text{if} \quad 2\lambda < |z| \le \gamma\lambda, \\ z & \text{if} \quad |z| > \gamma\lambda. \end{cases}$$

These estimates for singe predictor case can be easily generalized to multiple predictors by employing the coordinator descent algorithm (Friedman et al. 2007; Breheny and Huang 2011). In next section, we combine the MM algorithm and the CD algorithm to select the variables and to estimate the regression coefficients.

2.2 Surrogate function based on the MM algorithm

It is very important for the MM algorithm to construct a proper surrogate function. Suppose $\rho_q(\boldsymbol{\beta})$ is the objective function to be minimized. To facilitate computation, we usually construct the surrogate function using a twice-differentiable function, for example quadratic function. Let $\boldsymbol{\beta}^{(k)}$ be the parameter estimate of interest at iteration k. We construct a surrogate function $S_q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ satisfying

$$\begin{aligned} S_q(\boldsymbol{\beta}^{(k)}|\boldsymbol{\beta}^{(k)}) &= \rho_q(\boldsymbol{\beta}^{(k)}), \\ S_q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}) &\geq \rho_q(\boldsymbol{\beta}) \qquad \text{for all} \quad \boldsymbol{\beta}. \end{aligned} \tag{2.1}$$

The function $S_q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ majorizes the objective function $\rho_q(\boldsymbol{\beta})$ at $\boldsymbol{\beta}^{(k)}$, where $\boldsymbol{\beta}^{(k)}$ is treated as constant in the function $S_q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$. Then we find $\boldsymbol{\beta}^{(k+1)}$ by minimizing the surrogate function $S_q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$. In fact, the E-step of the EM algorithm is equivalent to the majorization step. MM algorithms also have been used for maximum likelihood estimation and missing data.

For quantile regression (1.2), we define a surrogate function, a quadratic function, to majorize $\rho_q(r)$ at current iterative estimator $r^{(k)}$,

$$S_q(r|r^{(k)}) = \frac{1}{4}\left[\frac{r^2}{|r^{(k)}|} + (4q-2)r + |r^{(k)}|\right]. \tag{2.2}$$

$S_q(r|r^{(k)})$ with $r = y - \mathbf{x}\boldsymbol{\beta}$ and $r^{(k)} = y - \mathbf{x}\boldsymbol{\beta}^{(k)}$ is the same as the surrogate function proposed by Hunter and Lange (2000). Figure 2 depicts the original loss function $\rho_q(r)$ and its surrogate function at some fixed point for quantile. Let $r^{(k+1)}$ be the next iterate to minimize $S_q(r|r^{(k)})$. Then we have

$$S_q(r^{(k+1)}|r^{(k)}) \leq S_q(r^{(k)}|r^{(k)}). \tag{2.3}$$

It is obvious from above inequality and the definition of surrogate function (2.1) that the descent property is satisfied, i.e.,

$$\rho_q(r^{(k+1)}) \leq \rho_q(r^{(k)}),$$

where the equality holds if equality holds in inequality (2.3). The convergence of MM algorithm is assured by this descent property which is the driving force behind an MM algorithm.

Because surrogate function $S_q(r|r^{(k)})$ in (2.2) is undefined when $r^{(k)} = 0$, we consider a perturbed surrogate function following the suggestions of Hunter and Lange (2000). We first define a perturbed version of piecewise linear loss function,

$$\rho_q^{\delta}(r) = \rho_q(r) - \frac{\delta}{2}\ln(\delta + |r|). \tag{2.4}$$

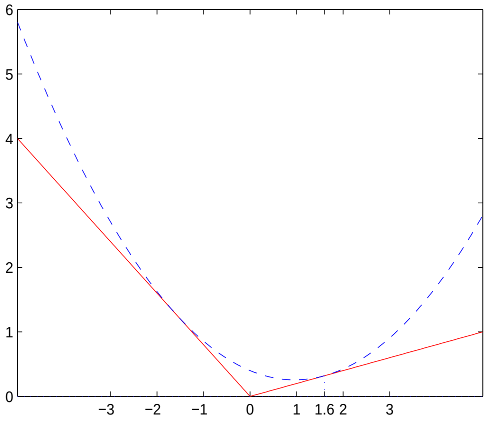


Figure 2: *The surrogate function $S_q(r|r^{(k)})$ (dashed line) with $r^{(k)} = 1.6$, and the original loss function $\rho_q(r)$ (solid line) for quantile loss function with $q = 0.2$.*

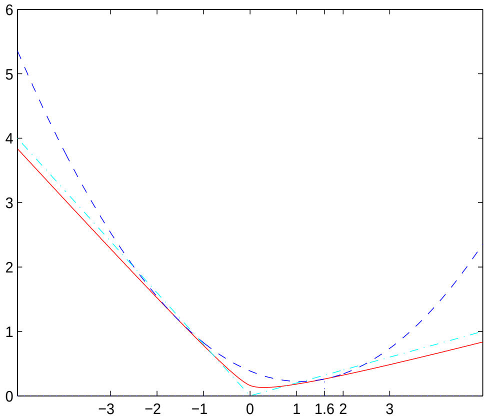


Figure 3: *The perturbed surrogate function $S_q^\delta(r|r^{(k)})$ (dashed line) with $r^{(k)} = 1.6$ and $\delta = 0.2$, the original loss function $\rho(r)$ (dotted line), and the perturbed loss function $\rho^\delta(r)$ (solid line) for quantile with $q = 0.2$.*

Based on $\rho_q^\delta(r)$, we construct a new perturbed surrogate function given the k-step estimator $r^{(k)}$,

$$S_q^\delta(r|r^{(k)}) = \frac{1}{4}\left[\frac{r^2}{\delta + |r^{(k)}|} + (4q-2)r + c^{(k)}\right], \tag{2.5}$$

where $c^{(k)} = |r^{(k)}|(|r^{(k)}|+2\delta)/(\delta+|r^{(k)}|)-2\delta\ln(\delta+|r^{(k)}|)$ is a constant such that $S_q^\delta(r^{(k)}|r^{(k)}) = \rho_q^\delta(r^{(k)})$.

Proposition 1 (a.) *The surrogate function $S_q(r|r^{(k)})$ defined by (2.2) majorizes the piecewise linear loss function $\rho_q(r)$ at $r = r^{(k)}$.*

(b.) *The surrogate function $S^{\delta}(r|r^{(k)})$ defined by (2.5) majorizes the perturbed piecewise linear loss function $\rho_q^{\delta}(r)$ defined by (2.4) at $r = r^{(k)}$.*

(c.) *As $\delta \downarrow 0$, $|S_q^{\delta}(r|r^{(k)}) - S_q(r|r^{(k)})| \to 0$ and $|\rho_q^{\delta}(r) - \rho_q(r)| \to 0$ uniformly on compact subsets of r.*

Let $r^{(k+1)}$ be the next iterate to minimize $S_q^{\delta}(r|r^{(k)})$. Due to (2.3), we have $S_q^{\delta}(r^{(k+1)}|r^{(k)}) \leq S_q^{\delta}(r^{(k)}|r^{(k)})$, which results in by Proposition 1 (a)

$$\rho_q^{\delta}(r^{(k+1)}) \leq \rho_q^{\delta}(r^{(k)}). \tag{2.6}$$

This descent property implies the convergence of a MM algorithm for perturbed loss function $\rho_q^{\delta}(r)$.

2.3 The CDMM algorithm

Instead of minimizing directly the objective function (1.3), we minimize the following penalized and perturbed function with perturbation δ,

$$\boxed{L_q^{\delta}(\alpha, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_q^{\delta}(y_i - \alpha - \mathbf{x}_i\boldsymbol{\beta}) + \sum_{j=1}^{p} P(|\beta_j|; \lambda, \gamma).} \tag{2.7}$$

Furthermore, due to the non-smoothness of the loss function, we minimize iteratively the following penalized and perturbed surrogate function given the k-iterate estimate $(\alpha^{(k)}, \boldsymbol{\beta}^{(k)})$,

$$Q_q^{\delta}(\alpha, \boldsymbol{\beta}|\alpha^{(k)}, \boldsymbol{\beta}^{(k)}) = \ell_q^{\delta}(\alpha, \boldsymbol{\beta}|\alpha^{(k)}, \boldsymbol{\beta}^{(k)}) + \sum_{j=1}^{p} P(|\beta_j|; \lambda, \gamma), \tag{2.8}$$

where

$$\ell_q^{\delta}(\alpha, \boldsymbol{\beta}|\alpha^{(k)}, \boldsymbol{\beta}^{(k)}) = \frac{1}{n}\sum_{i=1}^{n} S_q^{\delta}(r_i|r_i^{(k)}).$$

Since $\ell_q^{\delta}(\alpha, \boldsymbol{\beta}|\alpha^{(k)}, \boldsymbol{\beta}^{(k)})$ is quadratic in parameter $(\alpha, \boldsymbol{\beta})$, a Gauss-Newton approach is permitted with the gradient vector $\nabla \ell_q^{\delta}(\alpha, \boldsymbol{\beta}|\alpha^{(k)}, \boldsymbol{\beta}^{(k)})$ and Hessian matrix $\nabla^2 \ell_q^{\delta}(\alpha, \boldsymbol{\beta}|\alpha^{(k)}, \boldsymbol{\beta}^{(k)})$

with respect to $(\alpha, \boldsymbol{\beta})$. The first and second derivatives of the majorizer $\ell_q^\delta(\alpha, \boldsymbol{\beta}|\alpha^{(k)}, \boldsymbol{\beta}^{(k)})$ in (2.8) with respect to $\eta = \alpha + \mathbf{x}^T\boldsymbol{\beta}$ are given by

$$\begin{aligned}\nabla\ell_q^\delta(\eta|\eta^{(k)}) =& \frac{1}{2}\left(1 - 2q - \frac{r_1}{\delta + |r_1^{(k)}|}, \cdots, 1 - 2q - \frac{r_n}{\delta + |r_n^{(k)}|}\right),\\ \nabla^2\ell_q^\delta(\eta|\eta^{(k)}) =& \frac{1}{2}\text{diag}((\delta + |r_1^{(k)}|)^{-1}, \cdots, (\delta + |r_n^{(k)}|)^{-1}),\end{aligned}$$

where $\eta^{(k)} = \alpha^{(k)} + \mathbf{x}\boldsymbol{\beta}^{(k)}$. Let $\tilde{\mathbf{x}} = (\mathbf{1}, \mathbf{x})$, where $\mathbf{1}$ denotes a vector with all elements being 1. Thus, we have $\nabla\ell_q^\delta(\alpha, \boldsymbol{\beta}|\alpha^{(k)}, \boldsymbol{\beta}^{(k)}) = \tilde{\mathbf{x}}\nabla\ell_q^\delta(\eta|\eta^{(k)})$ and $\nabla^2\ell_q^\delta(\alpha, \boldsymbol{\beta}|\alpha^{(k)}, \boldsymbol{\beta}^{(k)}) = \tilde{\mathbf{x}}\nabla^2\ell_q^\delta(\eta|\eta^{(k)})\tilde{\mathbf{x}}^T$. Let $\mathbf{g} = \nabla\ell_q^\delta(\eta^{(k)}|\eta^{(k)})$ and $D = \nabla^2\ell_q^\delta(\eta^{(k)}|\eta^{(k)})$ denote the first and second derivative of majorizer $\ell_q^\delta(\alpha, \boldsymbol{\beta}|\alpha^{(k)}, \boldsymbol{\beta}^{(k)})$ with respect to η at $\eta^{(k)}$. One can obtain the iteratively reweighted least square,

$$Z_\lambda(\alpha, \boldsymbol{\beta}) = \frac{1}{2}(\tilde{\mathbf{y}}^{(k)} - \alpha - \mathbf{x}\boldsymbol{\beta})^T D(\tilde{\mathbf{y}}^{(k)} - \alpha - \mathbf{x}\boldsymbol{\beta}) + n\sum_{j=1}^{p} P(|\beta_j|; \lambda, \gamma), \tag{2.9}$$

where $\tilde{\mathbf{y}}^{(k)} = \eta^{(k)} - D^{-1}\mathbf{g}$.

Because intercept α is not penalized, from (2.9), we can update α by

$$\hat{\alpha}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{\mathbf{y}}_i^{(k)} - \mathbf{x}_i^T\boldsymbol{\beta}^{(k)}) = \alpha^{(k)} - \mathbf{g}^T\mathbf{1}/(\mathbf{1}^T D\mathbf{1}).$$

Let $\mathbf{y}^* = D^{1/2}\tilde{\mathbf{y}}^{(k)}$, $\mathbf{x}^* = D^{1/2}\mathbf{x}$, residual $\mathbf{r}^0 = \tilde{\mathbf{y}}^{(k)} - \alpha^{(k+1)} - \mathbf{x}\boldsymbol{\beta}^{(k)} = -D^{-1}\mathbf{g} + \alpha^{(k)} - \alpha^{(k+1)}$, and $\mathbf{r}^j = \mathbf{r}^{j-1} + (\beta_j^{(k)} - \beta_j^{(k+1)})\mathbf{x}_j$ for $j = 1, \cdots, p$. Denote leave-one out (the j^{th} coordinate removed) of $\mathbf{x}^*$ by $\mathbf{x}^*_{-j}$ and $\boldsymbol{\beta}^{(k)}$ by $\boldsymbol{\beta}^{(k)}_{-j}$. Let

$$z_j = \frac{1}{n\zeta_j}\mathbf{x}_j^{*T}(\mathbf{y}^* - \mathbf{x}^*_{-j}\boldsymbol{\beta}^{(k)}_{-j}) = \frac{1}{n\zeta_j}\mathbf{x}_j^T D\mathbf{r}^j + \beta_j^{(k)}, \tag{2.10}$$

where $\zeta_j = n^{-1}||\mathbf{x}_j^*||^2$. For the j^{th} coordinate, the estimator of β_j is

$$\hat{\beta}_j^{(k+1)} = \arg\min_{\beta_j} \frac{1}{2}(z_j - \beta_j)^2 + P(|\beta_j|; \lambda^*, \gamma^*),$$

where $\lambda^* = \lambda/\zeta_j$ and $\gamma^* = \gamma\zeta_j$. There is a closed form for single predictor as discussed in Section 2.1. The coordinate algorithm updates $\hat{\beta}_j^{(k)}$ by

$$\hat{\beta}_j^{(k+1)} = f.(z_j, \lambda^*, \gamma^*), \tag{2.11}$$

where $f.(z_j, \lambda^*, \gamma^*) = s(z_j, \lambda^*)$ for the LASSO, $f.(z_j, \lambda^*, \gamma^*) = f_{mcp}(z_j, \lambda^*, \gamma^*)$ for the MCP and $f.(z_j, \lambda^*, \gamma^*) = f_{scad}(z_j, \lambda^*, \gamma^*)$ for the SCAD.

Summarizing above discussion, we give the following CDMM algorithm for penalized piecewise linear loss regression.

CDMM Algorithm 1.

- Step 0. Input a grid of $\lambda = \{\lambda_1, \cdots, \lambda_N\}$ in decreasing order. For each λ_l, initialize $\hat{\alpha}^{(0)} = 0$ and $\hat{\boldsymbol{\beta}}^{(0)} = 0$. Repeat Step 1 and Step 2 until convergence.
- Step 1. Calculate D, $\mathbf{g}$. Update $\hat{\alpha}^{(k+1)} = \hat{\alpha}^{(k)} - n^{-1}\mathbf{1}^T D^{-1}\mathbf{g}$, and $\mathbf{r}_0 = \hat{\alpha}^{(k)} - \hat{\alpha}^{(k+1)} - D^{-1}\mathbf{g}$.
- Step 2. For $j = 1, 2, \cdots, p$,
  - 2.1. Calculate ζ_j, λ_j^*, γ_j^* and $z_j = \frac{1}{n\zeta_j}\mathbf{x}_j^T(D\mathbf{r}^{j-1}) + \beta_j^{(k)}$.
  - 2.2. Update $\hat{\beta}_j^{(k+1)} = f_\cdot(z_j, \lambda^*, \gamma^*)$.
  - 2.3. Update $\mathbf{r}_j = \mathbf{r}_{j-1} + (\hat{\beta}_j^{(k)} - \hat{\beta}_j^{(k+1)})\mathbf{x}_j$.
- Step 3. Output the path solution $\hat{\boldsymbol{\beta}}_{\lambda_1}, \cdots, \hat{\boldsymbol{\beta}}_{\lambda_N}$.

Remark 1. The CDMM algorithm does not depend on initial values of α and $\boldsymbol{\beta}$. Given each tuning parameter λ_l, it is very simple and fast to calculate the solution coordinately. Thus a solution path can be given. We will discuss how to select the tuning parameter and perturbation δ in section .

Proposition 2 *For any fixed δ, let $\{(\hat{\boldsymbol{\alpha}}_\delta^{(k)}, \hat{\boldsymbol{\beta}}_\delta^{(k)})\}$ denote the sequence of coefficients generated by CDMM algorithm. Then the limit point of the sequence is a stationary point of the objective function $L^\delta(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Furthermore,* (a) *for L_1 penalty,* (b) *for the MCP penalty with $\gamma > 1/c^*$ and* (c) *for the SCAD penalty with $\gamma > 1 + 1/c^*$, the limit point of the sequence is the minimum point of the objective function $L^\delta(\boldsymbol{\alpha}, \boldsymbol{\beta})$.*

3 Penalized composite quantile regression

The composite quantile regression was first introduced by Zou and Yuan (2008) and its variable selection was studied based on adaptive LASSO. Consider the M quantiles $0 < q_1 < q_2 < \cdots < q_M < 1$. Let $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_M)^T$, where α_m is the intercept for the

q_m^{th} quantile regression. In current paper, we consider the following penalized composite quantile regression

$$(\hat{\boldsymbol{\alpha}}^{cq}, \hat{\boldsymbol{\beta}}^{cq}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} L^{cq}(\boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{3.1}$$

$$\text{where} \quad L^{cq}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{n} \rho_{q_m}\left(y_i - \alpha_m - \mathbf{x}_i^T\boldsymbol{\beta}\right) + n\sum_{j=1}^{p} P(|\beta_j|; \lambda, \gamma),$$

and the penalty function $P(|\boldsymbol{\beta}|; \lambda, \gamma)$ may be the L_1 penalty, the MCP or the SCAD.

Corresponding to (3.1), the objective function of the penalized perturbed surrogate composite quantile regression at the $(k+1)$ step given $(\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)})$ is

$$L^{cq,\delta}(\boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}) = \frac{1}{M}\sum_{m=1}^{M} \ell^{\delta}_{q_m}(\alpha_m, \boldsymbol{\beta}|\alpha_m^{(k)}, \boldsymbol{\beta}^{(k)}) + \sum_{j=1}^{p} P(|\beta_j|; \lambda, \gamma). \tag{3.2}$$

To implement the coordinate descend algorithm in minimizing $L^{cq,\delta}(\boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)})$ by iteratively reweighted least squares, we define $\mathbf{g}_m = \nabla \ell^{\delta}_{q_m}(\eta|\eta_m^{(k)})$ and $D_m = \nabla^2 \ell^{\delta}_{q_m}(\eta|\eta_m^{(k)})$ for $m = 1, \cdots, M$ at $\eta = \eta_m^{(k)}$, where $\eta_m^{(k)} = \alpha_m^{(k)} + \mathbf{x}\boldsymbol{\beta}^{(k)}$. Furthermore, let $\zeta_j = (nM)^{-1}\sum_{j=1}^{M}(\mathbf{x}_j^T D_m \mathbf{x}_j)$, $\lambda_j^* = \lambda/\zeta_j$ and $\gamma_j^* = \gamma\zeta_j$. The following gives the details of the proposed algorithm.

CDMM Algorithm 2.

- Step 0. Input a grid of $\lambda = \{\lambda_1, \cdots, \lambda_N\}$ in decreasing order. For each λ_l, initialize $\hat{\alpha}^{(0)} = 0$ and $\hat{\boldsymbol{\beta}}^{(0)} = 0$. Repeat Step 1 and Step 2 until convergence.

- Step 1. Calculate D_m and $\mathbf{g}_m$. Update $\hat{\alpha}_m^{(k+1)} = \hat{\alpha}_m^{(k)} - n^{-1}\mathbf{1}^T D_m^{-1}\mathbf{g}_m$ and $\mathbf{r}_{0,m} = \hat{\alpha}_m^{(k)} - \hat{\alpha}_m^{(k+1)} - D_m^{-1}\mathbf{g}_m$ for $m = 1, \cdots, M$.

- Step 2. For $j = 1, 2, \cdots, p$,
  - 2.1. Calculate ζ_j, λ_j^*, γ_j^* and $z_j = \frac{1}{nM\zeta_j}\mathbf{x}_j^T\sum_{m=1}^{M}(D_m\mathbf{r}_m^{j-1}) + \beta_j^{(k)}$.
  - 2.2. Update $\hat{\beta}_j^{(k+1)} = f.(z_j, \lambda^*, \gamma^*)$.
  - 2.3. Update $\mathbf{r}_{j,m} = \mathbf{r}_{j-1,m} + (\hat{\beta}_j^{(k)} - \hat{\beta}_j^{(k+1)})\mathbf{x}_j$.

- Step 3. Output the path solution $\hat{\boldsymbol{\beta}}_{\lambda_1}, \cdots, \hat{\boldsymbol{\beta}}_{\lambda_N}$.

Corollary 1 *For any fixed δ, let $\{(\hat{\boldsymbol{\alpha}}_{\delta}^{(k)}, \hat{\boldsymbol{\beta}}_{\delta}^{(k)})\}$ denote the sequence of coefficients generated by coordinate descent algorithm. Then the limit point of the sequence is a stationary point of the objective function $L^{cq,\delta}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Furthermore,* (a) *for L_1 penalty,* (b) *for the MCP penalty with $\gamma > 1/c^*$ and* (c) *for the SCAD penalty with $\gamma > 1 + 1/c^*$, the limit point of the sequence is the minimum point of the objective function $L^{cq,\delta}(\boldsymbol{\alpha}, \boldsymbol{\beta})$.*

4 Selection of tuning parameter and perturbation

As mentioned by Kai, Li and Zou (2011), λ can be chosen by the BIC criterion as follows,

$$\mathrm{BIC}(\lambda) = \log\left(\frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{n}\rho_{q_m}(y_i - \hat{\boldsymbol{\alpha}}_m(\lambda) - \mathbf{x}_i\hat{\boldsymbol{\beta}}(\lambda))\right) + \log(n)\frac{df_\lambda}{n}, \tag{4.1}$$

where df_λ is the effective degrees of freedom for λ. Here $df_\lambda = \#\{j : \hat{\boldsymbol{\beta}}_j(\lambda) \neq 0\}$ is the number of the non-zero coefficients of $\hat{\boldsymbol{\beta}}(\lambda)$. Then we select $\hat{\lambda} = \arg\min_\lambda \mathrm{BIC}(\lambda)$.

??? How to choose perturbation parameter ???

5 Numerical studies

5.1 Simulation studies

In this Section, we consider the linear regression model (1.1) with differen error terms:

$$Y = \mathbf{x}\boldsymbol{\beta} + \Phi(x_1)\boldsymbol{\varepsilon}, \tag{5.1}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and error e follows one of the following six distributions.

a. Normal distribution, $\boldsymbol{\varepsilon} \sim N(0, 3)$.

b. Double exponential distribution with density $f(\boldsymbol{\varepsilon}) = \frac{1}{2}\exp(-|\boldsymbol{\varepsilon}|)$.

c. T-distribution with 3 degrees of freedom, $\boldsymbol{\varepsilon} \sim t(3)$.

d. Standard Cauchy distribution with density $f(\boldsymbol{\varepsilon}) = \frac{1}{\pi(1+\boldsymbol{\varepsilon}^2)}$.

e. The mixture of t-distribution and normal distribution $\boldsymbol{\varepsilon} \sim \frac{\sqrt{2}}{2}N(0, 1) + \frac{1}{2}t(4)$.

f. Logistic distribution with density $f(\varepsilon) = \frac{\exp(\varepsilon)}{(1+\exp(\varepsilon))^2}$.

EXAMPLE 1, Median regression

100 simulation data sets with sample size $n = 200$ are generated from the linear model (5.1). The predictors $\mathbf{x}$ are generated from a multivariate normal distribution $N(0, \Sigma)$ with $\Sigma_{(i,j)} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$. The true parameter $\boldsymbol{\beta}$ is a p-dimensional vector with $p = 50$ or 5000 whose first, second, fifth components are 2, 1, 3, respectively. Table 2 reports the estimation and selection results for the q^{th}-quantile regression with $q = 1/2$ (i.e., median regression) for the L_1, the MCP, and the SCAD penalty, respectively. In Table 2, the squared bias $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||^2$ is reported in column labeled "L_2". The average number of selected true non-zero coefficients is reported in column labeled with "C". False positive, the average number of coefficients erroneously set to be non-zero, is reported in column "Fp". We also report the proportion of simulation times where the underlying true model is selected in column "corfit". It can be seen that the three important variables are selected for all three penalty functions, however more variables with true zero coefficients are wrongly selected using L_1 penalty than using the other two penalty functions. The L_1 penalty yields large bias for the estimated coefficients and little chance to select exactly the true underlying model. While the MCP and the SCAD penalty perform similarly with small bias and high proportion of selecting the true model.

EXAMPLE 2, Composite quantile regression

100 simulation data sets with sample size $n = 200$ are generated from the linear model (5.1) for each of the six error distributions. The predictors $\mathbf{x}$ are generated from a multivariate normal distribution $N(0, \Sigma)$ with $\Sigma_{(i,j)} = 0.5$ for $i \neq j$ and $1 \leq i, j \leq p$. The dimensionality p is either 50 or 1000. The first four components of $\boldsymbol{\beta}$ are generated from $(4\log(n)/\sqrt{n} + |Z|)U$, and the last two components are equal to 2, where $Z \sim N(0, 1)$ is normally distributed, and U equals 1 with probability 0.5 and -1 with probability 0.5, and is independent of Z. In this Example the first four nonzero coefficients are $(\beta_1, \beta_2, \beta_3, \beta_4) = (-2.39, 1.62, 2.24, 2.86)$. The quantiles $q_m = m/20$ for $m = 1, \cdots, 19$ are used for composite quantile regression. Table 3 reports the results with the same notations as in Table 2. Similar conclusions can be drawn as for the median regression. Penalized composite qunatile regression has similar performance using the MCP and the SCAD, and outperforms that using the LASSO in terms of selection of important variables and estimation of regression

Table 2: The $q = 0.5^{\text{th}}$ quantile (median).

	LASSO				MCP				SCAD			
distribution and p	L_2	C	Fp	corfit	L_2	C	Fp	corfit	L_2	C	Fp	corfit
Normal Distribution												
$p = 500$	.45	3	.59	.56	.09	3	.40	.80	.08	3	0.28	.81
$p = 2000$	.46	3	1.02	.38	.13	2.99	.97	.61	.14	3	1.10	.60
T-Distribution												
$p = 500$	.08	3	.40	.65	.01	3	.01	.99	.01	3	0	1
$p = 2000$	.08	3	.44	.68	.01	3	.04	.97	.01	3	0	1
Logistic Distribution												
$p = 500$	.13	3	.33	.71	.03	3	.02	.98	.03	3	0	1
$p = 2000$	.14	3	.45	.68	.03	3	.02	.98	.03	3	0	1
Double Exponential												
$p = 500$	.07	3	.29	.76	.01	3	.02	.98	.01	3	.01	.99
$p = 2000$	.06	3	.33	.73	.01	3	.02	.98	.01	3	.01	.99
T-Normal-Mixed												
$p = 500$	.05	3	.31	.75	.01	3	.03	.97	.01	3	.01	.99
$p = 2000$	.06	3	.35	.71	.01	3	.02	.98	.01	3	.02	.98
Cauchy Distribution												
$p = 500$	2.95	2.47	.23	.62	2.31	2.48	.02	.78	2.31	2.48	.07	.75
$p = 2000$	3.15	2.42	.41	.56	2.59	2.42	.03	.76	2.58	2.42	.09	.72

coefficients.

5.2 Applications to gene expression data

We illustrate our proposed coordinate descent algorithm for composite quantile by a microarray data set that was reported in Scheetz et al. (2006). Similar to Wang et al. (2012), we split randomly the data into two sets for 100 times, training data including 80 sample size and test data including 40 sample size. The training data are used to select the important variables and estimate them. We use the test data to test the predict errors. For training data, ten-fold cross-validation (CV) method was employed to select the tuning parameter λ.

Table 4 summarizes the results the selected number and predict errors for the full data

Table 3: Composite quantile regression with the quantiles $q_m = m/20$ for $m = 1, \cdots, 19$.

	LASSO				MCP				SCAD			
distribution and p	L_2	C	Fp	corfit	L_2	C	Fp	corfit	L_2	C	Fp	corfit
Normal Distribution												
$p = 500$	2.64	5.93	8.23	0.01	0.14	6	0.10	0.91	0.19	6	0.27	0.83
$p = 2000$	4.63	5.76	10.38	0.00	0.20	6	0.39	0.86	0.32	6	0.81	0.72
T-Distribution												
$p = 500$	0.62	6.00	7.99	0.00	0.03	6	0.05	0.95	0.03	6	0.10	0.92
$p = 2000$	1.12	6.00	11.01	0.02	0.03	6	0.09	0.97	0.03	6	0.12	0.91
Logistic Distribution												
$p = 500$	0.29	6.00	6.47	0.01	0.04	6	0.03	0.98	0.04	6	0.03	0.98
$p = 2000$	0.39	6.00	11.09	0.01	0.04	6	0.19	0.89	0.05	6	0.47	0.80
Double Exponential												
$p = 50$	0.57	6.00	7.77	0.02	0.02	6	0.01	0.99	0.03	6	0.05	0.95
$p = 2000$	0.92	6.00	11.31	0.01	0.03	6	0.10	0.94	0.03	6	0.18	0.88
T-Normal-Mixed												
$p = 500$	0.37	6.00	7.36	0.03	0.03	6	0.02	0.98	0.02	6	0.03	0.97
$p = 2000$	0.55	6.00	12.45	0.00	0.02	6	0.19	0.90	0.03	6	0.39	0.87
Cauchy Distribution												
$p = 500$	5.36	5.43	4.96	0.02	1.35	5.75	0	0.95	1.35	5.74	0.09	0.94
$p = 2000$	7.59	5.16	6.42	0	1.75	5.65	0	0.92	1.75	5.65	0	0.92

Table 4: Composite quantile for microarray data.

	Full data		Split data	
Penalty	Number of Nonzeros	Predict errors	Number of Nonzeros	Predict errors
LASSO	26.0000	0.0282	30.5000	0.0366
MCP	22.0000	0.0248	5.9000	0.0412
SCAD	8.0000	0.0287	4.2000	0.0409

and split data. Table 5 list the selected genes for full data and split data. The frequencies of select genes for split data which appear in the genes selected for full data.

Table 5: Composite quantile for microarray data after screening by correlations between response and each covarate.

LASSO		MCP		SCAD	
Probe	Frequency	Probe	Frequency	Probe	Frequency
1370429_at	66	1370429_at	35	1370429_at	27
1370655_a_at	21	1374131_at	44	1378935_at	24
1371242_at	18	1377071_at	24	1379971_at	23
1374106_at	85	1377190_at	28	1382210_at	21
1374131_at	41	1377857_at	15	1382835_at	36
1377190_at	10	1378935_at	31	1383110_at	100
1378425_at	30	1379920_at	30	1383996_at	95
1378935_at	39	1379971_at	28	1389584_at	86
1379971_at	67	1380033_at	45		
1380033_at	46	1382835_at	49		
1382835_at	78	1383110_at	100		
1383110_at	95	1383502_at	12		
1383522_at	43	1383749_at	41		
1383673_at	85	1383996_at	95		
1383749_at	78	1384823_at	28		
1383996_at	99	1389584_at	94		
1384204_at	54	1390301_at	7		
1384466_at	38	1390401_at	31		
1389584_at	99	1393955_at	10		
1390401_at	47	1394107_at	9		
1390788_a_at	27	1395896_at	18		
1393382_at	58	1398255_at	25		
1393543_at	26				
1394399_at	49				
1395415_at	35				
1398255_at	14				

6 Discussion

In this paper, we studied the high dimensional variable selection for linear regression based on quantile regression. The usual penalized least square may not work when the error is high tail or when the variance is big. We developed the coordinate descent minorization-maximization algorithm to solve the minimizer of penalized quantile regression. The merit of this algorithm is much faster than existed methods, such as linear programming. The results of simulation and real data examples show our method performs pretty well.

Our method should have a good performance other piecewise linear loss function beyond the quantile regression, such as hinge loss function which is popularly used in classification. Comparing to linear programming, our method may be a good alternative for classification. This is our future work.

Appendix

Assumptions:

(1) The observations V_i are independent and identically distributed with probability density $f(V, \boldsymbol{\beta})$ with respect to some measure μ. $f(V, \boldsymbol{\beta})$ has a common support and the model is identifiable. Furthermore, $E_{\boldsymbol{\beta}}[\nabla \ell(\boldsymbol{\beta})] = 0$ and $E_{\boldsymbol{\beta}}[-\nabla^2 \ell(\boldsymbol{\beta})] = E_{\boldsymbol{\beta}}[\nabla \ell(\boldsymbol{\beta}) \nabla \ell(\boldsymbol{\beta})^T]$.

Proof of Proposition 1:

To prove (b), it suffices to verify that the conditions for surrogate function (2.1) are satisfied. The equality condition is obvious. For the second condition, it suffices to show that the difference $f(r) = S_q^\delta(r|r^{(k)}) - \rho^\delta(r)$ attains its minimum at $r = r^{(k)}$. It is easy to see that $f(r)$ is symmetric around 0, that is,

$$f(r) - f(-r) = 0,$$

because $S_q^\delta(r|r^{(k)}) - S_q^\delta(-r|r^{(k)}) = \rho^\delta(r) - \rho^\delta(-r) = (2q-1)r$. Thus, we only need to show

$f(r)$ attains its minimum when $r \geq 0$. Taking derivative of $f(r)$ with respect to r,

$$\begin{aligned} f'(r) =& \frac{1}{2}\left[\frac{r}{\delta + |r^{(k)}|} + 2q - 1\right] - \left(q - \frac{\delta}{2(\delta + r)}\right) \\ =& \frac{1}{2}\left(\frac{r}{\delta + |r^{(k)}|} + \frac{\delta}{\delta + r} - 1\right) \\ =& \frac{r(r - |r^{(k)}|)}{(\delta + |r^{(k)}|)(\delta + r)}. \end{aligned}$$

Obviously, $f'(|r^{(k)}|) = 0$, $f'(r) > 0$ if $r > |r^{(k)}|$, and $f'(r) < 0$ if $0 < r < |r^{(k)}|$, which implies $f(r)$ restricted $r \geq 0$ attains its minimum at $r = |r^{(k)}|$.

The proof of (a) can be completed by noting that $\delta \ln(\delta + |r|) \to 0$ as $\delta \downarrow 0$ for $|r| \leq C$ where $C > 0$ is a constant. Part (c) is a direct result of (a) and (b).

To facilitate the notation, rewrite $\beta_0 = \alpha$, $\tilde{\mathbf{x}} = (\mathbf{1}, \mathbf{x})$ and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}^T)^T$.

Lemma A.1 *Let $Q^{\delta}_{q,j}(\boldsymbol{\beta})$ denote the objective function, defined in (2.8), as a function of the single variable β_j with α and the remaining elements of $\boldsymbol{\beta}$ fixed. Then for the L_1 penalty,* (b) *for the MCP penalty with $\gamma > 1$ and* (c) *for the SCAD penalty with $\gamma > 2$, $Q^{\delta}_{q,j}(\boldsymbol{\beta})$ is a convex function of β_j for all j.*

Proof: The proof is similar to Breheny and Huang (2011). For all $\beta_j \in (-\infty, \infty)$,

$$\min\{d^2_- Q^{\delta}_{q,j}(\boldsymbol{\beta}), d^2_+ Q^{\delta}_{q,j}(\boldsymbol{\beta})\} \geq \frac{1}{n}\sum_{i=1}^{n} \frac{x_j^2}{2(\delta + |r_i|)} + \begin{cases} 0, & \text{for LASSO,} \\ 1 - \frac{1}{\gamma}, & \text{for MCP,} \\ 1 - \frac{1}{\gamma - 1}, & \text{for SCAD.} \end{cases}$$

where $d^2_- Q^{\delta}_{q,j}(\boldsymbol{\beta})$ and $d^2_+ Q^{\delta}_{q,j}(\boldsymbol{\beta})$ denote the second derivatives of $Q^{\delta}_{q,j}(\beta)$ in the direction $\beta < 0$ and $\beta > 0$, respectively, and $r_i = y_i - \alpha - x_i^T \boldsymbol{\beta}$ is the residual. The above inequality implies that $Q^{\delta}_{q,j}(\boldsymbol{\beta})$ is convex for all three penalties and that $Q^{\delta}_{q,j}(\boldsymbol{\beta})$ is strictly convex for the MCP penalty with $\gamma > 1$ and for the SCAD penalty with $\gamma > 2$.

Lemma A.2 *Let $L^{\delta}_{q,j}(\boldsymbol{\beta})$ denote the objective function, defined in (2.7), as a function of the single variable β_j with α and the remaining elements of $\boldsymbol{\beta}$ fixed. Then for the L_1 penalty,* (b) *for the MCP penalty with $\gamma > 1$ and* (c) *for the SCAD penalty with $\gamma > 2$, $L^{\delta}_{q,j}(\boldsymbol{\beta})$ is a convex function of β_j for all j.*

Proof: The proof is similar to Breheny and Huang (2011). For all $\beta_j \in (-\infty, \infty)$,

$$\min\{d_-^2 L_{q,j}^{\delta}(\boldsymbol{\beta}), d_+^2 L_{q,j}^{\delta}(\boldsymbol{\beta})\} \geq \frac{1}{n}\sum_{i=1}^{n} \frac{\delta x_j^2}{2(\delta + |r_i|)^2} + \begin{cases} 0, & \text{for LASSO,} \\ 1 - \frac{1}{\gamma}, & \text{for MCP,} \\ 1 - \frac{1}{\gamma - 1}, & \text{for SCAD.} \end{cases}$$

where $d_-^2 L_{q,j}^{\delta}(\boldsymbol{\beta})$ and $d_+^2 L_{q,j}^{\delta}(\boldsymbol{\beta})$ denote the second derivatives of $L_{q,j}^{\delta}(\beta)$ in the direction $\beta < 0$ and $\beta > 0$, respectively, and $r_i = y_i - \alpha - x_i^T \boldsymbol{\beta}$ is the residual. The above inequality implies that $L_{q,j}^{\delta}(\boldsymbol{\beta})$ is convex for all three penalties and that $L_{q,j}^{\delta}(\boldsymbol{\beta})$ is strictly convex for the MCP penalty with $\gamma > 1$ and for the SCAD penalty with $\gamma > 2$.

Proof of Proposition 2: If $||\boldsymbol{\beta}|| \to \infty$, the residual $|r| = |y - x^T \boldsymbol{\beta}| \to \infty$, which results in $L^{\delta}(\boldsymbol{\beta}) \to \infty$. The continuity of $L^{\delta}(\boldsymbol{\beta})$ implies the existence of the minimum of $L^{\delta}(\boldsymbol{\beta})$. Let $\Lambda = \{\boldsymbol{\beta} : L^{\delta}(\boldsymbol{\beta}) < L^{\delta}(\boldsymbol{\beta}_0)\}$, where $\boldsymbol{\beta}_0$ is the initial value of coordinate descent algorithm. It is easy to see that Λ is compact.

Let $\hat{\boldsymbol{\beta}}^{(k+1),j} = (\hat{\beta}_0^{(k+1)}, \cdots, \hat{\beta}_{j-1}^{(k+1)}, \hat{\beta}_j^{(k+1)}, \hat{\beta}_{j+1}^{(k)}, \cdots, \hat{\beta}_p^{(k)})^T$ for $j = 0, 1, \cdots, p$, and let $\hat{\boldsymbol{\beta}}^{(k+1),-1} = \hat{\boldsymbol{\beta}}^{(k)}$. Since, for $k \geq 1$,

$$Q^{\delta}(\hat{\boldsymbol{\beta}}^{(k+1),j} | \hat{\boldsymbol{\beta}}^{(k+1),j-1}) \leq Q^{\delta}(\hat{\boldsymbol{\beta}}^{(k+1),j-1} | \hat{\boldsymbol{\beta}}^{(k+1),j-1}) \quad \text{for any} \quad j = 0, 1, \cdots, p,$$

by Proposition 1, we have

$$\boxed{L^{\delta}(\hat{\boldsymbol{\beta}}^{(k+1),j}) \leq L^{\delta}(\hat{\boldsymbol{\beta}}^{(k+1),j-1}) \quad \text{for any} \quad j = 0, 1, \cdots, p.} \tag{A.1}$$

Furthermore, using inductive argument, we have

$$L^{\delta}(\hat{\boldsymbol{\beta}}^{(k+1),j}) \leq L^{\delta}(\hat{\boldsymbol{\beta}}^{(k),j}) \quad \text{for any} \quad j = 0, 1, \cdots, p, \tag{A.2}$$

and specially, $L^{\delta}(\hat{\boldsymbol{\beta}}^{(k+1)}) \leq L^{\delta}(\hat{\boldsymbol{\beta}}^{(k)})$. Since Λ is compact, $\{\hat{\boldsymbol{\beta}}^{(k)}\}_{k=1}^{\infty}$ have limit point, say $\boldsymbol{\beta}^*$. For each $j \in \{0, 1, ..., p\}$, $\{\hat{\boldsymbol{\beta}}^{(k),j}\}_{k=1}^{\infty}$ is bounded, we can assume that $\{\hat{\boldsymbol{\beta}}^{(k),j}\}_{k=1}^{\infty}$ converges to some $\boldsymbol{\beta}^{*,j} = (\beta_0^{*,j}, \beta_1^{*,j}, \cdots, \beta_p^{*,j})^T$ as k goes to infinity. By the continuity of $L^{\delta}(\boldsymbol{\beta})$, the facts that $L^{\delta}(\hat{\boldsymbol{\beta}}^{(k)})$ decreases monotonically as k increases and the minimum of $L^{\delta}(\boldsymbol{\beta})$ exists imply that

$$L^{\delta}(\boldsymbol{\beta}_0) \geq \lim_{k \to \infty} L^{\delta}(\hat{\boldsymbol{\beta}}^{(k)}) = L^{\delta}(\boldsymbol{\beta}^{*,0}) = \cdots = L^{\delta}(\boldsymbol{\beta}^{*,p}).$$

Let $\mathbf{e}_j$ is unit vector with the j^{th} element 1 and others 0. We claim that, for each $j = 0, 1, \cdots, p$, $\boldsymbol{\beta}^{*,j}$ is stationary point of $L^{\delta}(\boldsymbol{\beta})$ along the singe coordinate direction $\mathbf{e}_j$.

Denote by $\mathcal{M}_j(\boldsymbol{\beta})$ the map defined by the coordinate descent algorithm $\hat{\boldsymbol{\beta}}^{(k),j} \mapsto \hat{\boldsymbol{\beta}}^{(k+1),j}$ for $j = 0, 1, \cdots, p-1$. The continuity of map $\mathcal{M}_j(\boldsymbol{\beta})$ follows by the fact that z_j and $f_{\cdot}(z_j, \lambda, \gamma)$ are continuous. Let $\hat{\boldsymbol{\beta}}^{(k_n)}$ be a subsequence with limit $\boldsymbol{\beta}^*$. By (A.2), we have

$$L^\delta(\hat{\boldsymbol{\beta}}^{(k_n+1),j}) \leq L^\delta(\mathcal{M}_j(\hat{\boldsymbol{\beta}}^{(k_n),j})) \leq L^\delta(\hat{\boldsymbol{\beta}}^{(k_n),j}).$$

Invoking the continuity of map $\mathcal{M}_j(\boldsymbol{\beta})$ and objective function $L^\delta(\boldsymbol{\beta})$, taking limits in above inequalities gives

$$L^\delta(\boldsymbol{\beta}^{*,j}) = L^\delta(\lim_{n\to\infty} \mathcal{M}_j(\hat{\boldsymbol{\beta}}^{(k_n),j})) = L^\delta(\mathcal{M}_j(\boldsymbol{\beta}^{*,j})),$$

which implies that $\boldsymbol{\beta}^{*,j}$ is stationary point of $L^\delta(\boldsymbol{\beta})$ along the single coordinate direction $\mathbf{e}_j$. Thus, by Lemma A.2, we obtain $\boldsymbol{\beta}^{*,j}$ is minimum point of $L^\delta(\boldsymbol{\beta})$ along the single coordinate direction $\mathbf{e}_j$, which implies

$$\begin{aligned} L^\delta(\boldsymbol{\beta}^{*,j}) &\leq L^\delta(\boldsymbol{\beta}^{*,j} + d\mathbf{e}_j), \quad \forall d, j = 0, 1, \cdots, p, \\ \boldsymbol{\beta}_k^{*,j} &= \boldsymbol{\beta}_k^{*,j-1}, \forall k \neq j, \quad j = 1, \cdots, p, \qquad (7) \end{aligned}$$

Noting that $\rho_q(y - x^T\boldsymbol{\beta})$ is Gâteaux-differentiable about $\boldsymbol{\beta}$ and the convexity in each direct $\mathbf{e}_j$ established by Lemma A.2, the rest of the proof can be finished as the same as Theorem 4.1 in Tseng (2001).

References

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics.* **39**, 82-130.

Bradic, J., Fan, J., and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. R. Statist. Soc. B.* **73**, 325-349.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithm for nonconvex penalized regression, with application to biological feature selection. *Annals of Applied Statistics.* **5**, 232-253. [4, 6, 19, 20].

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics.* **24**, 2350-2383.

Boser, B. E., Guyon, I. and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proc. Fifth ACM Workshop on Computational Learning Theory (COLT)*, 144-152. ACM Press, New York.

Cai, Z. and Xu, X. (2009). Nonparametric Quantile Estimations for Dynamic Smooth Coefficient Models. *Journal of the American Statistical Association*, **103**, 1595-1608. [4].

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning.* **20**, 273-297.

Donoho, D. L., Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika.* **81**, 425-455. [4, 6].

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association.* **97**, 77-87.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics.* **32**, 407-499. [2].

Fan, J., Feng, Y. and Wu, Y. (2010). High-dimensional variable selection for Cox proportional hazards model. *Borrowing Strengh: Theory Powering Applications A Festschrift for Lawrence* . **6**, 70-86.

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. [2, 5, 6].

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics.* **30**, 74-99.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. *In proceedings of the Madrid International Congress of Mathematicians 2006.* **3**, 595-622.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc. B.* **70**, 894-911.

Fan, J. and Lv, J. (2011). Non-concave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57**, 5467-5484.

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*. **32**, 928-961.

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, **10**, 2013-2038.

FRIEDMAN, J., HASTIE, T., HOEFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.*, **2**, 302-332. [4, 6].

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1-22.

HE, X. (2009). *Modeling and inference by quantile regression.* Technical report, Departent of Statistics, University of Illinois at Urbana-Champain. [4].

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D. and LANDER, E. S. (1999). Molecular classifcation of cancer: Class discovery and class prediction by gene expression monitoring. *Science*. **286**, 531-537.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. New York: Springer.

HASTIE, T., ROSSET, S., TIBSHIRANI, R. and ZHU, J. (2004). The entire regularization path for the support vector machine. *J. Machine Learning Research*. **5**, 1391-1451.

HUNTER, D. R. and LANGE, K. (2000). Quantile regression MM algorithm. *J. Computational and Graphical Statistics*. **9**, 60-77. [4, 7].

HUNTER, D. R. and LI, R. (2005). Variable selection using MM algorithms. *Annals of Statistics*. **33**, 1617-1642. [4].

HUANG, J., MA. S. G. and ZHANG, C-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*. **18**, 1603-1618.

KAI, B., LI, R. and ZOU, H (2010). Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. *J. R. Statist. Soc. B*. **72**, 49-69.

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

Kato, K. (2010). Solving ℓ_1 regularization problems with piecewise linear losses. *J. Computational and Graphical Statistics*. **19**, 1024-1040.

Koenker, R. (2005). *Quantile regression.* Cambridge University Press. [4].

Koenker, R. and Bassett, G. S. (1978). Regression Quantiles. *Econometrica*, **46**, 33-50. [2, 3].

Koenker, R. and Hollock, G. S. (2001). Quantile regression: an introduction. *Journal of Economic Perspectives*, **15**, 43-56. [4].

Koenker, R. and Geling, O. (2001). Reappraising medfly longevity: A quantile regression survival analysis. *Journal of the American Statistical Association*. **96**, 458-468. [4].

Li, Y. and Zhu, J. (2008), L_1-Norm Quantile Regression. *J. Computational and Graphical Statistics*. **17**, 163?85.

Park, M. Y. and Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. R. Statist. Soc. B.* **69**, 659-677. [2].

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp1, 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. *Proc. Nat. Acad. Sci.* **103**, 14429-14434. [15].

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B.* **58**, 267-288. [2].

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statist. Med.* **16**, 385-395.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *J. R. Statist. Soc. B.* **73**, 273-282.

Tibshirani, R., Hoefling, H. and Tibshirani, R. (2011). Nearly isotonic regression. *Technometrics*, **53**, 54-61.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B*, **67**, 91-108.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109**, 475-494. [4, 21].

Vapnik, V. (1996). *The nature of statistical learning.* Springer-Verlag.

Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association.* **107**, 214-222. [5, 15].

Wei, Y. and He, X. (2006). Conditional growth charts (with discussions). *Annals of Statistics.* **34**, 2069-2097.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics.* **2**, 224-244. [4].

Wu, Y. C. and Liu, Y. F. (2009). Variable Selection in Quantile Regression. *Statistica Sinica.* **19**, 801-817. [5].

Wu, S., Zou, H. and Yuan, M. (2008). Structured variable selection in support vector machines. *Electronic Journal of Statistics.* **2**, 103-117.

Yuan, M. and Lin, Y. (2007) Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19?5.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics.* **38**, 894-942. [2, 5].

Zhang, C.-H. (2007). Penalized linear unbiased selection. *Technical Report 2007-003.* Dept. Statistics, Rutgers Univ.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in highdimensional regression. *Ann. Statist.*, **36**, 1567?594.

Zhao, P. and Yu, B. (2007). Stagewise lasso. *Journal of Machine Learning Research*, **8**, 2701-2726.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. statist. Assoc.* **101**, 1418-1429.

Zou, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika*. **95**, 241-247.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B.* **67**, 301-320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*. **36**, 1509-1533. [2].

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Annals of Statistics*. **36**, 1108-1126. [5, 11].

Zou, H. and Yuan, M. (2008). Regularized simultaneous model selection in multiple quantiles regression. *Computational Statistics and Data Analysis*. **52**, 5296-5304.