Statistica Sinica **19** (2009), 801-817

VARIABLE SELECTION IN QUANTILE REGRESSION

Yichao Wu and Yufeng Liu

North Carolina State University and University of North Carolina, Chapel Hill

Abstract: After its inception in Koenker and Bassett (1978), quantile regression has become an important and widely used technique to study the whole conditional distribution of a response variable and grown into an important tool of applied statistics over the last three decades. In this work, we focus on the variable selection aspect of penalized quantile regression. Under some mild conditions, we demonstrate the oracle properties of the SCAD and adaptive-LASSO penalized quantile regressions. For the SCAD penalty, despite its good asymptotic properties, the corresponding optimization problem is non-convex and, as a result, much harder to solve. In this work, we take advantage of the decomposition of the SCAD penalty function as the difference of two convex functions and propose to solve the corresponding optimization using the Difference Convex Algorithm (DCA).

Key words and phrases: DCA, LASSO, oracle, quantile regression, SCAD, variable selection.

1. Introduction

At the heart of statistics lies regression. Ordinary least squares regression (OLS) estimates the mean response as a function of the regressors or predictors. Least absolute deviation regression (LADR) estimates the conditional median function, which has been shown to be more robust to outliers. In the seminal paper of Koenker and Bassett (1978), they generalized the idea of LADR and introduced quantile regression (QR) to estimate the conditional quantile function of the response. As a result, QR provides much more information about the conditional distribution of a response variable. It includes LADR as a special case. After its introduction, QR has attracted tremendous interest in the literature. It has been applied in many different areas: economics (Hendricks and Koenker (1992) and Koenker and Hallock (2001)), survival analysis (Yang (1999) and Koenker and Geling (2001)), microarray study (Wang and He (2007)), growth chart (Wei et al. (2006) and Wei and He (2006)), and so on. Li et al. (2007) considered quantile regression in reproducing kernel Hilbert spaces, and proposed a very efficient algorithm to compute its entire solution path with respect to the tuning parameter.

Variable selection plays an important role in the model building process. In practice, it is common to have a large number of candidate predictor variables available, and they are included in the initial stage of modeling for the consideration of removing potential modeling bias (Fan and Li (2001)). However, it is undesirable to keep irrelevant predictors in the final model since this makes it difficult to interpret the resultant model and may decrease its predictive ability. In the regularization framework, many different types of penalties have been introduced to achieve variable selection. The L_1 penalty was used in the LASSO proposed by Tibshirani (1996) for variable selection. Fan and Li (2001) proposed a unified approach via nonconcave penalized least squares regression, which simultaneously performs variable selection and coefficient estimation. By choosing an appropriate nonconcave penalty function, this method keeps many merits of the best subset selection and of ridge regression: it produces sparse solution; it ensures the stability of model selection; it provides unbiased estimates for large coefficients. These are the three desirable properties of a good penalty (Fan and Li (2001)). An example of such nonconcave penalties is the smoothly clipped absolute deviation (SCAD) function first introduced in Fan (1997), and studied further by Fan and Li (2001) to show its oracle properties in the penalized likelihood setting. Later on, a series of papers Fan and Li (2002, 2004), Fan and Peng (2004) and Hunter and Li (2005) studied its further properties and produced new algorithms.

By using adaptive weights for penalizing different coefficients in the LASSO penalty, Zou (2006) introduced the adaptive LASSO and demonstrated its oracle properties. Similar results were also established in Yuan and Lin (2007) and Zhao and Yu (2006). Zhang and Lu (2007) studied the adaptive LASSO in proportional hazard models. Candes and Tao (2007) and Fan and Lv (2006) studied variable selection in the setting of dimensionality higher than the sample size.

Previously, Koenker (2004) applied the LASSO penalty to the mixed-effect quantile regression model for longitudinal data to encourage shrinkage in estimating the random effects. Li and Zhu (2005) developed the solution path of the L_1 penalized quantile regression. Wang, Li and Jiang (2007) considered LADR with the adaptive LASSO penalty. To our knowledge, there still lacks of study on variable selection in penalized quantile regression. In this work, we try to fill this void. Notice that the loss function used in quantile regression is not differentiable at the origin and, as a result, the general oracle properties for nonconcave penalized likelihood (Fan and Li (2001)) do not apply directly. Here, we extend the oracle properties of the SCAD and adaptive-LASSO penalties to the context of penalized quantile regression, including the LADR by Wang, Li and Jiang (2007) as a special case.

The SCAD penalty is nonconvex, and consequently it is hard to solve the corresponding optimization problem. Motivated by the fact that the SCAD penalty function can be decomposed as the difference of two convex functions, we propose to use the Difference Convex algorithm (DCA) (see An and Tao (1997)) to solve the corresponding non-convex optimization problem. DCA minimizes a non-convex objective function by solving a sequence of convex minimization problems. At each iteration, it approximates the second convex function by a linear function. As a result, the objective function at each step is convex and it is much easier to optimize than the original non-convex problem. In this sense, DCA turns out to be an instance of the MM algorithm since, at each step, DCA majorizes the nonconvex objective function and then performs minimization. One difference between DCA and Hunter and Li (2005)'s MM is that, at each iteration, DCA majorizes the nonconvex function using a linear approximation while Hunter and Li (2005)'s MM uses a quadratic approximation. We opt for DCA due to its clean formulation and simple implementation. In particular, for quantile regression, the resulting optimization at each iteration is a linear programming problem, thus more efficient. We recently learned that Zou and Li (2007) proposed a local linear approximation algorithm (LLA) to solve the SCAD optimization problem. Although both DCA and LLA perform iterative linear programming, unlike the LLA, DCA does not enforce symmetry in the approximation of the SCAD penalty.

The rest of the paper is organized as follows. Penalized quantile regressions with the SCAD and adaptive-LASSO penalties are introduced in Section 2. We present the asymptotic properties of the SCAD and adaptive-LASSO penalized quantile regressions in Section 3. Algorithms for handling their corresponding optimization problems are proposed in Section 4. Sections 5 and 6 present numerical results on simulations and on data, respectively. We conclude the paper with Section 7.

2. Penalized Linear Quantile Regression

Consider a sample $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ of size n from some unknown population, where $\boldsymbol{x}_i \in \mathbb{R}^d$. The conditional τth quantile function $f_\tau(\boldsymbol{x})$ is defined such that $P(Y \leq f_\tau(\boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}) = \tau$, for $0 < \tau < 1$. By tilting the absolute loss function, Koenker and Bassett (1978) introduced the check function which is defined by $\rho_\tau(r) = \tau r$ if $r > 0$, and $-(1-\tau)r$ otherwise. In this seminal paper, they demonstrated that the τth conditional quantile function can be estimated by solving the minimization problem

$$\min_{f_\tau \in \mathcal{F}} \sum_{i=1}^{n} \rho_\tau(y_i - f_\tau(\boldsymbol{x}_i)). \tag{2.1}$$

To avoid over-fitting and improve generalization ability, as in Koenker et al. (1994) and Koenker (2004), we consider the penalized version of (2.1) in the regularization framework

$$\min_{f_\tau \in \mathcal{F}} \sum_{i=1}^{n} \rho_\tau(y_i - f_\tau(\boldsymbol{x}_i)) + \lambda J(f_\tau), \tag{2.2}$$

where $\lambda \geq 0$ is the regularization parameter and $J(f_\tau)$ denotes the roughness penalty of the function $f_\tau(\cdot)$.

In this work, we focus on linear quantile regression by setting $f_\tau(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}_\tau$ where $\boldsymbol{\beta}_\tau = (\beta_{\tau,1}, \beta_{\tau,2}, \ldots, \beta_{\tau,d})^T$, namely, the conditional quantile function is a linear function of the regressor $\boldsymbol{x}$. This form can be easily generalized to handle nonlinear quantile regression via basis expansion. For functions of linear form, there are many different types of penalty functions available: the L_0 penalty (also known as the entropy penalty) used in best subsect selection (Breiman (1996)); the L_1 penalty (LASSO) (Tibshirani (1996)); the L_2 penalty used in ridge regression (Hoerl and Kennard (1988)); the combination of the L_0 and L_1 penalties (Liu and Wu (2007)); the L_q ($q \geq 0$) penalties in bridge regression (Frank and Friedman (1993)). Fan and Li (2001) argued that a good penalty should yield the following three properties in its estimator: unbiasedness, sparsity, and continuity. Unfortunately, none of the L_q penalty family satisfies these three properties simultaneously, but Fan and Li (2001) showed that the SCAD penalty in the penalized likelihood setting does. Another penalty falling into the latter category is the adaptive-LASSO penalty studied by Zou (2006).

2.1. SCAD

Fan and Li (2001) demonstrated the oracle properties for the SCAD in the variable selection aspect, and conjectured that the LASSO penalty does not possess the oracle properties. This conjecture was later confirmed by Zou (2006), who further proposed the adaptive LASSO and showed its oracle properties in penalized least squares regression.

The SCAD penalty is defined in terms of its first derivative and is symmetric around the origin. For $\theta > 0$, its first derivative is

$$p'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\}, \tag{2.3}$$

where $a > 2$ and $\lambda > 0$ are tuning parameters. Note that the SCAD penalty function is symmetric, non-convex on $[0, \infty)$, and singular at the origin. One instance of the SCAD penalty function is plotted in the right panel of Figure 4.1. We can see that, around the origin, it takes the same form as the LASSO penalty and this leads to its sparsity property. But, different from the LASSO penalty, the SCAD penalizes large coefficients equally while the LASSO penalty increases linearly as the magnitude of the coefficient increases. In this way, the SCAD results in unbiased penalized estimators for large coefficients. After putting the SCAD penalty in (2.2) with linear function $f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}_\tau$, the SCAD penalized quantile regression solves the minimization problem

$$\min_{\boldsymbol{\beta}_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_\tau) + \sum_{j=1}^{d} p_\lambda(\beta_{\tau,j}).$$

2.2. Adaptive-LASSO

The adaptive-LASSO can be viewed as a generalization of the LASSO penalty. Basically the idea is to penalize the coefficients of different covariates at a different level by using adaptive weights. In the case of least squares regression, Zou (2006) proposed to use as weights the reciprocal of the ordinary least squares estimates raised to some power. The straightforward generalization, for our case of quantile regression, is to use the non-penalized quantile regression estimates as weights. More explicitly, let

$$\tilde{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}_\tau}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_\tau). \tag{2.4}$$

It can be shown that $\tilde{\boldsymbol{\beta}}_\tau$ is a root-n consistent estimator of $\boldsymbol{\beta}_\tau$. Then the adaptive-LASSO penalized quantile regression minimizes

$$\sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_\tau) + \lambda \sum_{j=1}^{d} \tilde{w}_j \mid \beta_{\tau,j} \mid$$

with respect to $\boldsymbol{\beta}_\tau$, where the weights are set to be $\tilde{w}_j = 1/ \mid \tilde{\beta}_{\tau,j} \mid^\gamma$, $j = 1, \ldots, d$; for some appropriately chosen $\gamma > 0$.

3. Asymptotic Properties

In this section, we establish the oracle properties of the SCAD or adaptive-LASSO penalized quantile regression. We assume the data $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ consists of n observations from the linear model

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i = \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_1 + \boldsymbol{x}_{i2}^T\boldsymbol{\beta}_2 + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

with $P(\epsilon_i < 0) = \tau$ as in Condition (i). Here $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \boldsymbol{x}_{i2}^T)^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, $\boldsymbol{x}_{i1} \in \mathbb{R}^s$, $\boldsymbol{x}_{i2} \in \mathbb{R}^{d-s}$, and the true regression coefficients are $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$ with each component being nonzero, and $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20} = \mathbf{0}$ (as a result $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$). This means that the first s regressors are important while the remaining $p - s$ are noise variables.

For our theoretical results, we enforce the following technical conditions.

(i) Error assumption (cf Pollard (1991)): The regression errors $\{\epsilon_i\}$ are independent and identically distributed, with τth quantile zero and a continuous, positive density $f(\cdot)$ in a neighborhood of zero.

(ii) The design $\boldsymbol{x}_i, i = 1, \ldots, n$, is a deterministic sequence for which there exists a positive definite matrix Σ such that $\lim_{n\to\infty} (\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T)/n = \Sigma$. Denote the top-left s-by-s submatrix of Σ by Σ_{11} and the right-bottom $(d - s)$-by-$(d - s)$ submatrix of Σ by Σ_{22}.

3.1. SCAD penalty

The SCAD penalized quantile regression solves $\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$, where $Q(\boldsymbol{\beta}) = \sum_{i=1}^n \rho_\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) + n\sum_{j=1}^d p_{\lambda_n}(|\ \beta_j\ |)$. As in Fan and Li (2001), we establish the root-n consistency of our SCAD penalized estimator as in Theorem 1 when the tuning parameter $\lambda_n \to 0$ as $n \to \infty$.

Theorem 1(Consistency). *Consider a sample* $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ *from model* (3.1) *satisfying Conditions* (i) *and* (ii). *If* $\lambda_n \to 0$, *there exists a local minimizer* $\hat{\boldsymbol{\beta}}$ *such that* $\|\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\ \| = O_p(n^{-1/2})$.

Under some further conditions, the sparsity property $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ of the SCAD penalized estimator can be obtained.

Lemma 1(Sparsity). *Consider a sample* $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ *from model* (3.1) *satisfying Conditions* (i) *and* (ii). *If* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$ *as* $n \to \infty$, *then with probability tending to one, for any given* $\boldsymbol{\beta}_1$ *satisfying* $\|\ \boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\ \| = O_p(n^{-1/2})$ *and any constant* C,

$$Q((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T) = \min_{\|\boldsymbol{\beta}_2\| \le Cn^{-\frac{1}{2}}} Q((\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T).$$

Our next theorem addresses the asymptotic oracle property.

Theorem 2(Oracle). *Consider a sample* $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ *from model* (3.1) *satisfying Conditions* (i) *and* (ii). *If* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$ *as* $n \to \infty$, *then with probability tending to one, for the root-n consistent local minimizer* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ *in Theorem* 1, *one has*

(a) *Sparsity:* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;

(b) *Asymptotic normality:* $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \tau(1-\tau)\Sigma_{11}^{-1}/f(0)^2)$, *where* Σ_{11} *is defined in Condition* (ii).

Remark 1. Notice that the main difference between penalized quantile regression and the more general penalized likelihood, as considered in Fan and Li (2001), is that the check function in penalized quantile regression is non-differentiable at the origin. To handle the difficulty caused by this non-differentiability, we use the convexity lemma previously used by Pollard (1991).

3.2. Adaptive-LASSO

The adaptive-LASSO penalized quantile regression solves $\min_{\boldsymbol{\beta}} Q_1(\boldsymbol{\beta})$ where $Q_1(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) + n\lambda_n \sum_{j=1}^{d} \tilde{w}_j \mid \beta_j \mid$. Let $\hat{\boldsymbol{\beta}}^{(AL)}$ be its solution.

Theorem 3(Oracle). *Consider a sample* $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ *from model* (3.1) *satisfying Conditions* (i) *and* (ii). *If* $\sqrt{n}\lambda_n \to 0$ *and* $n^{(\gamma+1)/2}\lambda_n \to \infty$*, then we have*

1. *Sparsity:* $\hat{\boldsymbol{\beta}}_2^{(AL)} = \mathbf{0}$;
2. *Asymptotic normality:* $\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{(AL)} - \boldsymbol{\beta}_{10}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \tau(1-\tau)\Sigma_{11}^{-1}/f(0)^2)$.

3.3. Non i.i.d. random errors

The conclusions in Theorems 2 and 3 are based on the assumption of *i.i.d.* random errors. We can further extend the aforementioned oracle results to the case of non *i.i.d.* random errors. In the light of the work of Knight (1999), we make the following assumptions.

(N1) As $n \to \infty$, $\max_{1\le i\le n} \boldsymbol{x}_i^T\boldsymbol{x}_i/n \to 0$.

(N2) The random errors ϵ_i's are independent with $F_i(t) = P(\epsilon_i \le t)$ the distribution function of ϵ_i. We assume that each $F_i(\cdot)$ is locally linear near zero (with a positive slope) and $F_i(0) = \tau$.

Define $\psi_{ni}(t) = \int_0^t \sqrt{n}(F_i(s/\sqrt{n}) - F_i(0))ds$, which is a convex function for each n and i.

(N3) Assume that, for each $\boldsymbol{u}$, $(1/n)\sum_{i=1}^{n} \psi_{ni}(\boldsymbol{u}^T\boldsymbol{x}_i) \to \varsigma(\boldsymbol{u})$, where $\varsigma(\cdot)$ is a strictly convex function taking values in $[0, \infty)$.

Corollary 1. *Under Conditions* (ii) *and* (N1), *Theorems* 2 *and* 3 *hold provided the non i.i.d. random errors satisfy* (N2) *and* (N3).

Remark 2. The assumption (N2) covers a class of general models with non *i.i.d.* random errors, for example, it includes the common location-scale shift model (Koenker (2005)). Corollary 1 follows directly using the results of Knight (1999). Further details of all proofs are provided in the on-line supplement materials at `http://www.stat.sinica.edu.tw/statistica`.

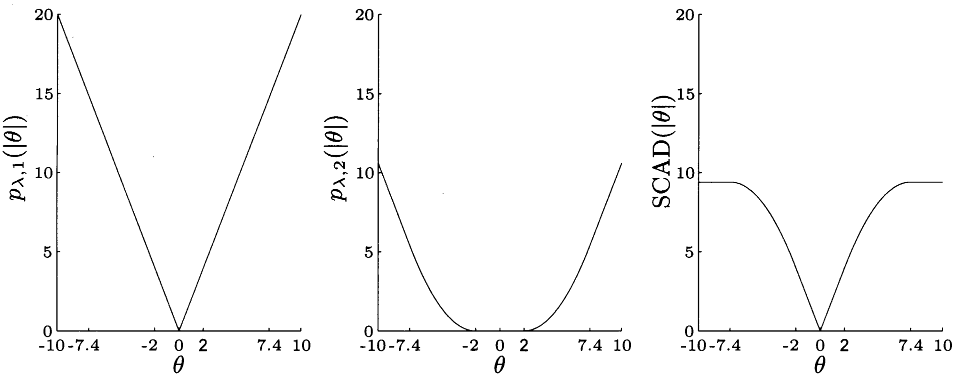

Figure 4.1. Decomposition of the SCAD penalty as $p_\lambda(\theta) = p_{\lambda,1}(\theta) - p_{\lambda,2}(\theta)$, with parameters $\lambda = 2$ and $a = 3.7$.

4. Algorithms

4.1. SCAD

Despite the excellent statistical properties of the SCAD penalized estimator, the corresponding optimization is a non-convex minimization problem and is much harder to solve than its LASSO penalized counterpart. In Fan and Li (2001), a unified least quadratic approximation (LQA) algorithm was proposed to solve the SCAD penalized likelihood optimization problem. Hunter and Li (2005) studied LQA under a more general MM-algorithm framework, where MM stands for minorize-maximize or majorize-minimize. A typical example of the MM algorithm is the well-known EM.

Notice that in (2.3), the first order derivative of the SCAD penalty function on $(0, \infty)$ is the sum of two components: the first is a constant and the second is a decreasing function on the range $(0, \infty)$. As a result, the SCAD penalty function can be decomposed as the difference of two convex functions. More explicitly, we have $p_\lambda(\theta) = p_{\lambda,1}(\theta) - p_{\lambda,2}(\theta)$ where both $p_{\lambda,1}(\cdot)$ and $p_{\lambda,2}(\cdot)$ are convex functions with derivatives, for $\theta > 0$, given by

$$\begin{cases} p'_{\lambda,1}(\theta) = \lambda \\ p'_{\lambda,2}(\theta) = \lambda(1 - \frac{(a\lambda-\theta)_+}{(a-1)\lambda})I(\theta > \lambda). \end{cases} \tag{4.1}$$

For the particular set of parameters $a = 3.7$ and $\lambda = 2$, this decomposition is graphically illustrated in Figure 4.1, where the left panel plots $p_{\lambda,1}(\theta)$, the central panel corresponds to $p_{\lambda,2}(\theta)$, and $p_\lambda(\theta) = p_{\lambda,1}(\theta) - p_{\lambda,2}(\theta)$ is given in the right panel. The above decomposition of the SCAD penalty allows us to use the well-studied DC algorithm. DCA was proposed by An and Tao (1997) to handle non-convex optimization; later on, it was applied in machine learning

Liu, Shen and Doss (2005b) and Wu and Liu (2007). DCA is a local algorithm and it decreases the objective value at each iteration. Due to its decomposition and approximation, DCA converges in a finite number of steps. More details on DCA can be found in Liu, Shen and Doss (2005b) and Liu, Shen and Wong (2005a).

Due to the above decomposition of the SCAD penalty, the objective function of the SCAD penalized quantile regression can be decomposed as $Q_{vex}(\boldsymbol{\beta}) + Q_{cav}(\boldsymbol{\beta})$, where $Q_{vex}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_{\lambda_n,1}(\mid \beta_j \mid)$ and $Q_{cav}(\boldsymbol{\beta}) = -n\sum_{j=1}^{d} p_{\lambda_n,2}(\mid \beta_j \mid)$.

Algorithm 1. Difference Convex Algorithm for minimizing $Q(\boldsymbol{\beta}) = Q_{vex}(\boldsymbol{\beta}) + Q_{cav}(\boldsymbol{\beta})$

1. Initialize $\boldsymbol{\beta}^{(0)}$.
2. Repeat $\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}(Q_{vex}(\boldsymbol{\beta}) + \left\langle Q'_{cav}(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(t)} \right\rangle)$ until convergence.

The difference convex algorithm solves the non-convex minimization problem via a sequence of convex subproblems (see Algorithm 1). Denote the solution at step t by $\boldsymbol{\beta}^{(t)} = (\beta_1^{(t)}, \ldots, \beta_p^{(t)})^T$. Then the derivative of the concave part at $\boldsymbol{\beta}^{(t)}$ is

$$Q'_{cav}(\boldsymbol{\beta}^{(t)}) = -n(p'_{\lambda_n,2}(\mid \beta_1^{(t)} \mid)\operatorname{sign}(\beta_1^{(t)}), p'_{\lambda_n,2}(\mid \beta_p^{(t)} \mid)\operatorname{sign}(\beta_p^{(t)}), \ldots,\\ p'_{\lambda_n,2}(\mid \beta_p^{(t)} \mid)\operatorname{sign}(\beta_p^{(t)}))^T,$$

where $p'_{\lambda_n,2}(\cdot)$ is defined in (4.1), and $\operatorname{sign}(\cdot)$ is the sign function. In the $(t+1)$th iteration, DCA approximates the second function by a linear function and solves the optimization problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_{\lambda_n,1}(\mid \beta_j \mid)\\ -n\sum_{j=1}^{d} p'_{\lambda_n,2}(\mid \beta_j^{(t)} \mid)\operatorname{sign}(\beta_j^{(t)})(\beta_j - \beta_j^{(t)}). \tag{4.2}$$

Here for the initialization in Step 1 of Algorithm 1, we use the solution of the non-penalized linear quantile regression $\tilde{\boldsymbol{\beta}}_\tau$ given by (2.4).

By introducing some slack variables, we can recast the above minimization problem (4.2) into the following linear programming problem.

$$\min \sum_{i=1}^{n}(\tau\xi_i + (1-\tau)\zeta_i) + n\lambda_n\sum_{j=1}^{d}\nu_j - n\sum_{j=1}^{d} p'_{\lambda_n,2}(\mid \beta_j^{(t)} \mid)\operatorname{sign}(\beta_j^{(t)})(\beta_j - \beta_j^{(t)})$$

$$\begin{aligned}\text{subject to } & \xi_i \geq 0,\ \zeta_i \geq 0,\ \xi_i - \zeta_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta},\ i = 1,\ldots,n \\ & \nu_j \geq \beta_j,\ \nu_j \geq -\beta_j,\ j = 1,\ldots,d.\end{aligned}$$

This can be easily solved by many optimization softwares. In contrast, at each iteration, the LQA (Fan and Li (2001) and Hunter and Li (2005)) needs to solve a quadratic programming problem and, as a result, it is less efficient. Our numerical studies in Section 5 confirm this.

4.2. Adaptive-LASSO

With the aid of slack variables, the adaptive-LASSO penalized quantile regression can also be casted into the linear programming problem

$$\begin{aligned}\min \quad & \sum_{i=1}^{n}(\tau\xi_i + (1-\tau)\zeta_i) + n\lambda_n \sum_{j=1}^{d} \tilde{w}_j\eta_j \\ \text{subject to} \quad & \xi_i \geq 0,\ \zeta_i \geq 0,\ \xi_i - \zeta_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta},\ i = 1,\ldots,n \\ & \eta_j \geq \beta_j,\ \eta_j \geq -\beta_j,\ j = 1,\ldots,d.\end{aligned} \tag{4.3}$$

Here the weights $\tilde{w}_j$'s are appropriately chosen, as discussed in Section 2.2. Note that the minimization problem (4.3) includes the LASSO penalized quantile regression as a special case, by setting $\tilde{w}_j = 1$ for $j = 1,\ldots,d$.

5. Monte Carlo Study

In this section, we first use one example to compare three different algorithms (LQA, MM, and DCA) for the SCAD penalized quantile regression, and thereby show the advantage of our new DC algorithm for the SCAD. Hence, we choose the DCA for the SCAD in the remaining numerical studies. In these examples, we study the finite-sample variable selection performance of different penalized quantile regressions. Here we want to point out that the intercept term is included in penalized quantile regression for all data analysis in this paper. For the SCAD penalty, we do not tune the parameter a. Following Fan and Li (2001)'s suggestion, we set $a = 3.7$ to reduce the computational burden. The number of zero coefficients is evaluated as follows: an estimate is treated as zero if its absolute value is smaller than 10^{-6}.

The data for Examples 5.1 and 5.2 were generated from the linear model

$$y = \boldsymbol{x}^T\boldsymbol{\beta} + \sigma\epsilon, \tag{5.1}$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. The components of $\boldsymbol{x}$ and ϵ were standard normal. The correlation between any two components x_i and x_j was set to $\rho^{|i-j|}$

Table 5.1. Simulation results for Example 5.1 with $n = 60$, $\tau = 0.5$.

			no. of zeros		
	Method	Test Error	Correct	Wrong	CPU-time in seconds
$\sigma = 1$	LQA	0.4189 (0.0158)	2.76 (1.40)	0.00 (0.00)	14.44 (20.63)
	MM	0.4189 (0.0161)	4.08 (1.20)	0.00 (0.00)	26.29 (155.87)
	DCA	0.4193 (0.0163)	4.40 (1.02)	0.00 (0.00)	0.26 (0.15)
$\sigma = 3$	LQA	1.2856 (0.0692)	2.30 (1.48)	0.06 (0.24)	9.56 (13.06)
	MM	1.2807 (0.0647)	3.68 (1.63)	0.14 (0.38)	12.39 (32.98)
	DCA	1.2822 (0.0642)	3.84 (1.57)	0.15 (0.39)	0.21 (0.11)

with $\rho = 0.5$. This model has been considered by many authors (Tibshirani (1996), Fan and Li (2001) and Zou (2006), to name a few).

Denote the sample size of training data sets by n. Throughout this section, an independent tuning data set and testing data set of size n and $100n$, respectively, were generated exactly in the same way as the training data set. The tuning parameter λ was selected via a grid search based on the tuning error in terms of the check loss function evaluated on the tuning data. Similarly defined testing errors on the testing data set are reported. More explicitly, a test error refers to the average check loss on the independent testing data set.

Example 5.1(Comparison of LQA, MM, and DCA for the SCAD). In this example, we generated data from model (5.1) with $n = 60$, and different algorithms for the SCAD penalized quantile regression were compared. Table 5.1 summarizes the results of 100 repetitions for two cases: $\sigma = 1$ and $\sigma = 3$. Average test errors, numbers of correct and wrong zero coefficients, and CPU times with standard deviations in their corresponding parentheses are reported. We found that, while giving very similar test errors, the three algorithms produced different numbers of zero coefficients. On average, DCA gave significantly more zeros. Remarkably, we notice that on average DCA took much less CPU-time than LQA and MM, as expected. The reason is that in each iteration, DCA solved a linear programming while LQA and MM required quadratic programming, as discussed at the end of Section 4.1. For the MM algorithm, we set Hunter and Li (2005)'s parameter τ to be 10^{-6} in their Equation (3.12).

Because of its superior performance, the DCA algorithm is used for the remaining data analysis to solve the SCAD penalized quantile regression.

Example 5.2(Comparison of finite-sample variable selection performance). We generated data from model (5.1) to compare the finite-sample variable selection performance of the L_1, the SCAD, and the adaptive-L_1 with the oracle. Simulation results of different settings are reported in Table 5.2. We can see that the

Table 5.2. Simulation results for Example 5.2

τ		$n = 100$, $\sigma = 1$			$n = 100$, $\sigma = 3$		
			no. of zeros			no. of zeros	
	Method	Test Error	Correct	Wrong	Test Error	Correct	Wrong
0.25	L_1	0.3378 (0.0111)	3.16 (1.47)	0.00 (0.00)	0.9976 (0.0347)	1.87 (1.49)	0.00 (0.00)
	SCAD	0.3296 (0.0091)	3.98 (1.66)	0.00 (0.00)	0.9968 (0.0364)	3.94 (1.60)	0.01 (0.10)
	adapt-L_1	0.3288 (0.0090)	4.21 (1.25)	0.00 (0.00)	0.9944 (0.0318)	3.00 (1.52)	0.00 (0.00)
	Oracle	0.3282 (0.0087)	5.00 (0.00)	0.00 (0.00)	0.9873 (0.0284)	5.00 (0.00)	0.00 (0.00)
0.5	L_1	0.4143 (0.0108)	2.92 (1.48)	0.00 (0.00)	1.2379 (0.0392)	2.00 (1.48)	0.00 (0.00)
	SCAD	0.4101 (0.0119)	4.00 (1.50)	0.00 (0.00)	1.2336 (0.0385)	4.01 (1.64)	0.02 (0.14)
	adapt-L_1	0.4081 (0.0101)	4.31 (1.00)	0.00 (0.00)	1.2339 (0.0385)	3.20 (1.48)	0.01 (0.10)
	Oracle	0.4072 (0.0099)	5.00 (0.00)	0.00 (0.00)	1.2248 (0.0348)	5.00 (0.00)	0.00 (0.00)
0.75	L_1	0.3364 (0.0093)	3.24 (1.42)	0.00 (0.00)	0.9893 (0.0324)	2.09 (1.36)	0.00 (0.00)
	SCAD	0.3286 (0.0100)	4.05 (1.42)	0.00 (0.00)	0.9866 (0.0336)	4.30 (1.35)	0.06 (0.24)
	adapt-L_1	0.3307 (0.0083)	4.51 (1.05)	0.00 (0.00)	0.9827 (0.0325)	3.73 (1.30)	0.01 (0.10)
	Oracle	0.3266 (0.0091)	5.00 (0.00)	0.00 (0.00)	0.9747 (0.0241)	5.00 (0.00)	0.00 (0.00)

reported test errors are very similar but, on average, the SCAD and the adaptive-L_1 gave more zero coefficients than the L_1. This confirms the superiority of the SCAD and the adaptive-L_1 as shown in our theoretical results.

Example 5.3(Dimensionality larger than the sample size). For the adaptive LASSO penalty, an initial consistent estimator is required to derive the adaptive weights. Due to the work of He and Shao (2000), the solution of the linear quantile regression at (2.4) is still consistent even if the dimension increases with the sample size, but at a speed slower than some root of the sample size. However, it is not clear how to find a consistent initial solution for deriving the adaptive weights in the case of dimension larger than sample size. For $p > n$, we performed the L_2 penalized quantile regression first and used this solution to derive the weights for the adaptive-L_1 penalty. In this example, we compared the performance of these different penalties in the case with more predictor variables than the sample size.

Our datasets in this example were generated from model (5.1), augmented with 102 more independent noise variables $x_9, x_{10}, \ldots, x_{110}$. Adding more independent noise variables makes the estimation harder. In order to make the estimation possible, the variance of random error ϵ was set at $\sigma^2 = 0.5^2$; each of these additional noise variable was $N(0, 0.5^2)$ and they were independent of each other. The results based on 100 repetitions with sample size 100 are reported in Table 5.3. It is evident from Table 5.3. that both the SCAD and adaptive-L_1 penalties improved over the L_1 penalty in terms of prediction accuracy as well as variable selection capability, even in the more difficult case of $p > n$. Such results also validate our proposed procedure of using the L_2 penalized solution to derive the adaptive weights for the adaptive-L_1 penalty.

Table 5.3. Simulation results for Example 5.3 with sample size $n = 100$. Here the supscript * in adapt-$L_1^\star$ indicates that the adaptive weights of the adaptive-L_1 penalty are based on the solution of the L_2 penalized quantile regression.

			no. of zeros	
τ	Method	Test Error	Correct	Wrong
0.25	L_1	0.1744 (0.0073)	113.65 (4.30)	0.00 (0.00)
	SCAD-DCA	0.1673 (0.0049)	116.72 (1.09)	0.00 (0.00)
	adapt-$L_1^\star$	0.1684 (0.0045)	115.47 (3.16)	0.00 (0.00)
	Oracle	0.1668 (0.0038)	117.00 (0.00)	0.00 (0.00)
0.5	L_1	0.2150 (0.0073)	112.37 (5.67)	0.00 (0.00)
	SCAD-DCA	0.2094 (0.0043)	116.38 (1.33)	0.00 (0.00)
	adapt-$L_1^\star$	0.2101 (0.0057)	114.54 (4.58)	0.00 (0.00)
	Oracle	0.2089 (0.0038)	117.00 (0.00)	0.00 (0.00)
0.75	L_1	0.1749 (0.0068)	112.47 (6.50)	0.00 (0.00)
	SCAD-DCA	0.1692 (0.0063)	116.59 (1.18)	0.00 (0.00)
	adapt-$L_1^\star$	0.1705 (0.0055)	115.02 (4.79)	0.00 (0.00)
	Oracle	0.1680 (0.0048)	117.00 (0.00)	0.00 (0.00)

Example 5.4(Non *i.i.d.* random errors). In this example, we considered the case of non *i.i.d.* random errors to check the robustness of our methods. Our data was generated from model 2 of Kocherginsky, He and Mu (2005), with

$$y = 1 + x_1 + x_2 + x_3 + (1 + x_3)\epsilon,$$

where x_1 and x_3 were generated from the standard normal distribution and the uniform distribution on $[0, 1]$, $x_2 = x_1 + x_3 + z$ with z being standard normal, and $\epsilon \sim N(0, 1)$. The variables x_1, x_3, z, and ϵ were mutually independent. To study the effect of variable selection, we included five more independent standard normal noise variables, $x_4, \ldots, x_8$, independent of each other.

The results based on 100 repetitions with sample size $n = 100$ are reported in Table 5.4, in the same format as in Example 5.2. Again we can see the improvement in test errors. Moreover, both the SCAD and adaptive-L_1 penalties can identify more correct zero coefficients than can L_1 . In this case, all three penalties tend to produce more wrong zero coefficients in the final model compared to Example 5.2. A possible reason is that x_2 is highly correlated with x_1 and x_3 since $x_2 = x_1 + x_3 + z$. Nevertheless, compared with the L_1 penalty, the SCAD and adaptive-L_1 penalties on average lead to fewer wrong zero coefficients.

6. Data

Harrison and Rubinfeld (1978) studied various methodological issues related to the use of housing data to estimate the demand for clean air. In particular,

Table 5.4. Simulation results for Example 5.4 with sample size $n = 100$.

			no. of zeros	
τ	Method	Test Error	Correct	Wrong
0.25	L_1	0.4944 (0.0110)	2.27 (1.72)	0.72 (0.45)
	SCAD-DCA	0.4919 (0.0136)	4.73 (0.62)	0.51 (0.50)
	adapt-L_1	0.4926 (0.0119)	3.64 (1.27)	0.65 (0.48)
	Oracle	0.4925 (0.0133)	5.00 (0.00)	0.00 (0.00)
0.5	L_1	0.6272 (0.0145)	1.37 (1.57)	0.36 (0.48)
	SCAD-DCA	0.6157 (0.0118)	4.52 (0.69)	0.19 (0.42)
	adapt-L_1	0.6196 (0.0107)	3.31 (1.32)	0.29 (0.46)
	Oracle	0.6157 (0.0117)	5.00 (0.00)	0.00 (0.00)
0.75	L_1	0.5081 (0.0146)	1.44 (1.59)	0.25 (0.44)
	SCAD-DCA	0.4942 (0.0140)	4.72 (0.55)	0.06 (0.24)
	adapt-L_1	0.5008 (0.0147)	3.59 (1.08)	0.16 (0.37)
	Oracle	0.4935 (0.0132)	5.00 (0.00)	0.00 (0.00)

the Boston House Price Dataset was used. This dataset is available online at `http://lib.stat.cmu.edu/datasets/boston_corrected.txt`, with some corrections and augmentation by the latitude and longitude of each observation; the result is called the Corrected Boston House Price Data. There are 506 observations, 15 non-constant predictor variables, and one response variable, corrected median value of owner-occupied homes (CMEDV). Predictors include longitude (LON), latitude (LAT), crime rate (CRIM), proportion of area zoned with large lots (ZN), proportion of non-retail business acres per town (INDUS), Charles River as a dummy variable (= 1 if tract bounds river; 0 otherwise) (CHAS), nitric oxides concentration (NOX), average number of rooms per dwelling (RM), proportion of owner-occupied units built prior to 1940 (AGE), weighted distances to five Boston employment centres (DIS), index of accessibility to radial highways (RAD), property tax rate (TAX), pupil-teacher ratio by town (PTRATIO), black population proportion town (B), and lower status population proportion (LSTAT). For simplicity, we excluded the categorical variable RAD. We also standardized the response variable CMEDV and the predictor variables aside from CHAS. Penalized quantile regression was applied with the standardized CMEDV as the response. We used 27 predictor variables in the penalized quantile regression, including the variable CHAS, the other 13 standardized predictor variables, and their squares.

In each repetition, we randomly split all the 506 observations into training, tuning and testing data sets of size 150, 150, and 206 respectively. The performance over 10 repetitions of the penalized quantile regression with different penalties and different quantiles is summarized in Table 6.5. The results indicate

Table 6.5. Results of the Corrected Boston House Price Data.

	$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
Method	Test Error	no. of zeros	Test Error	no. of zeros	Test Error	no. of zeros
L_1	0.1339 (0.0107)	11.10 (3.14)	0.1832 (0.0215)	9.30 (4.16)	0.1813 (0.0419)	7.10 (4.72)
SCAD	0.1367 (0.0164)	14.20 (2.78)	0.1862 (0.0257)	12.40 (4.40)	0.1920 (0.0799)	12.40 (3.86)
adapt-L_1	0.1346 (0.0130)	13.60 (3.20)	0.1840 (0.0216)	11.10 (5.67)	0.1776 (0.0403)	12.10 (3.98)

Note: In this table, the DCA is chosen for the SCAD.

that different penalties give similar test errors, but that SCAD and adaptive-L_1 use fewer variables than does L_1.

7. Discussion

In this work, we study penalized quantile regression with the SCAD and the adaptive-LASSO penalties. We show that they enjoy the oracle properties established by Fan and Li (2001) and Zou (2006), even though the check function is non-differentiable at the origin. To handle the non-convex optimization problem of the SCAD penalized quantile regression, we propose use of the Difference Convex algorithm. The new algorithm is very efficient, as confirmed by the simulation results in Example 5.1.

Notice that DCA is a very general algorithm. It can be easily extended to apply to a more general SCAD penalized likelihood setting, as long as the likelihood part is convex. For example, in SCAD penalized least squares regression, each iteration involves a quadratic programming problem. Similarly, DCA can be applied to the SCAD SVM (Zhang et al. (2006)).

Acknowledgement

We would like to thank Professor Jianqing Fan for his helpful comments and suggestions. This research is partially supported by NSF grant DMS-06-06577 and NIH grant R01-GM07261.

References

An, L. T. H. and Tao, P. D. (1997). Solving a class of linearly constrained indefinite quadratic problems by d.c. algorithms. *J. Global Optim.* **11**, 253-285.

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

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Amer. Statist.* **6**, 2313-2351.

Fan, J. (1997). Comments on "Wavelets in statistics: A review", by A. Antoniadis. *J. Amer. Statist. Assoc.* **6**, 131-138.

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

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

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99**, 710-723.

Fan, J. and Lv, J. (2006). Sure independence screening for ultra-high dimensional feature space. Submitted.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Amer. Statist.* **32**, 928-961.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-148.

Geyer, C. J. (1994). On the asymptotics of constrained m-estimation. *Amer. Statist.* **22**, 1993-2010.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *J. Environmental Economics and Management*, 81-102.

He, X. and Shao, Q.-M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73**, 120-135.

Hendricks, W. and Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *J. Amer. Statist. Assoc.* **87**, 58-68.

Hoerl, A. and Kennard, R. (1988). Ridge regression. In *Encyclopedia of Statistical Sciences* **8**, 129-136 Wiley, New York.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithm. *Amer. Statist.* **33**, .1617-1642.

Knight, K. (1999). Asymptotics for L_1-estimators of regression parameters under heteroscedasticity. *Canad. J. Statist.* **27**, 497-507.

Kocherginsky, M., He, X. and Mu, Y. (2005). Practical confidence intervals for regression quantiles. *J. Comput. Graph. Statist.* **14**, 41-55.

Koenker, R. (2004). Quantile regression for longitudinal data. *J. Multivariate Anal.* **91**, 74-89.

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press.

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

Koenker, R. and Geling, R. (2001). Reappraising medfly longevity: a quantile regression survival analysis. *J. Amer. Statist. Assoc.* **96**, 458-468.

Koenker, R. and Hallock, K. (2001). Quantile regression. *Journal of Economic Perspectives* **15**, 143-156.

Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81**, 673-680.

Li, Y., Liu, Y. and Zhu, J. (2007). Quantile regression in reproducing kernel Hilbert spaces. *J. Amer. Statist. Assoc.*, **102**, 255-268.

Li, Y. and Zhu, J. (2005). l_1-norm quantile regressions. *J. Comput. Graph. Statist.* To appear.

Liu, S., Shen, X. and Wong, W. (2005a). Computational development of ψ-learning. In *The SIAM* 2005 *International Data Mining Conf.*, 1-12.

Liu, Y., Shen, X. and Doss, H. (2005b). Multicategory ψ-learning and support vector machine: computational tools. *J. Comput. Graph. Statist.*, **14**, 219-236.

Liu, Y. and Wu, Y. (2007). Variable selection via a combination of the L_0 and L_1 penalties. *J. Comput. Graph. Statist.*, **16**, 782-798.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7**, 186-199.

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

Wang, H. and He, X. (2007). Detecting differential expressions in genechip microarray studies: A quantile approach. *J. Amer. Statist. Assoc.* **102**, 104-112.

Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the lad-lasso. *J. Business & Economic Statistics* **25**, 347-355.

Wei, Y. and He, X. (2006). Conditional growth charts (with discussions). *Ann. Statist.* **34**, 2069-2031.

Wei, Y., Pere, A., Koenker, R. and He, X. (2006). Quantile regression methods for reference growth curves. *Statist. Medicine* **25**, 1369-1382.

Wu, Y. and Liu, Y. (2007). Robust truncated-hinge-loss support vector machines. *J. Amer. Statist. Assoc.* **102**, 974-983.

Yang, S. (1999). Censored median regression using weighted empirical survival and hazard functions. *J. Amer. Statist. Assoc.* **94**, 137–145.

Yuan, M. and Lin, Y. (2007). On the nonnegative garrote estimator. *J. Roy. Statist. Soc. Ser. B* **69**, 143–161.

Zhang, H. H., Ahn, J., Lin, X. and Park, C. (2006). Gene selection using support vector machines with nonconvex penalty. *Bioinformatics* **22**, 88–95.

Zhang, H. H. and Lu, W. (2007). Adaptive-lasso for Cox's proportional hazard model. *Biometrika.* **94**, 691–703.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Machince Learning Research* **7**, 2541-2563.

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

Zou, H. and Li, R. (2007). One-step Sparse Estimates in Nonconcave Penalized Likelihood Models. *Ann. Statist.* To appear.

Department of Statistics, North Carolina State University, Raleigh NC 27695, U.S.A.

E-mail: wu@stat.ncsu.edu

Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

E-mail: yfliu@email.unc.edu

(Received July 2007; accepted October 2007)