Statistica Sinica **22** (2012), 000-000
doi:http://dx.doi.org/10.5705/ss.2010.203

ORACLE MODEL SELECTION FOR NONLINEAR MODELS BASED ON WEIGHTED COMPOSITE QUANTILE REGRESSION

Xuejun Jiang, Jiancheng Jiang and Xinyuan Song

Zhongnan University of Economics and Law,
University of North Carolina at Charlotte
and The Chinese University of Hong Kong

Abstract: In this paper we propose a weighted composite quantile regression (WCQR) estimation approach and study model selection for nonlinear models with a diverging number of parameters. The WCQR is augmented using a data-driven weighting scheme. With the error distribution unspecified, the proposed estimators share robustness from quantile regression and achieve nearly the same efficiency as the oracle maximum likelihood estimator for a variety of error distributions including the normal, mixed-normal, Student's t, Cauchy distributions, etc. Based on the proposed WCQR, we use the adaptive-LASSO and SCAD regularization to simultaneously estimate parameters and select models. Under regularity conditions, we establish asymptotic equivalency of the two model selection methods and show that they perform as well as if the correct submodels are known in advance. We also suggest an algorithm for fast implementation of the proposed methodology. Simulations are conducted to compare different estimators, and an example is used to illustrate their performance.

Key words and phrases: Adaptive WCQR, adaptive LASSO, high dimensionality, model selection, oracle property, SCAD.

1. Introduction

Various techniques have been developed for simultaneous variable selection and coefficient estimation, based on the penalized likelihood or least squares principles. Examples include the nonnegative garrote (Breiman (1995) and Yuan and Lin (2007)), the LASSO (Tibshirani (1996)), bridge regression (Fu (1998) and Knight and Fu (2000)), the SCAD (Fan and Li (2001)), the MC+ (Zhang (2010)), etc. These methods have advantages over traditional stepwise deletion and subset selection procedures in implementation and in the derivation of sampling properties, and have been extended by several authors to achieve robustness. For instance, for linear models, He and Shao (2000) considered M-estimator for general parametric models, Wang, Li, and Jiang (2007) considered the LASSO for least absolute regression (LAD-LASSO), and Zou and Yuan (2008a) studied the

LASSO for composite quantile regression (CQR-LASSO), among others. These endeavors have enriched the variable selection theory for different models by using different regularized estimation methods, with aim at oracle model selection procedures (see Fan and Li (2006) for a comprehensive overview) and robustness and efficiency of the estimation (Zou and Yuan (2008a)).

The CQR-LASSO in Zou and Yuan (2008a) is robust and performs nearly like a CQR-oracle model selector. The CQR they used is a sum of different quantile regression (QR) (Koenker and Bassett (1978)) at predetermined quantiles, which uses equal weights for different QR (see Section 2 for details). Intuitively, equal weights are not optimal in general, and hence a more efficient CQR should exist. In this article we suggest a "weighted CQR (WCQR)" estimation method and let the data decide the weights to improve efficiency, while keeping robustness from the QR. The WCQR method is applicable to various models, but here we focus on the nonlinear model

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where ε_i's are independent random errors with unknown distribution function $G(\cdot)$ and density $g(\cdot)$, and the function $f(\cdot, \boldsymbol{\beta})$ is known up to a p-dimensional vector of parameters $\boldsymbol{\beta}$. Model (1.1) contains many submodels of which linear models and generalized linear models with continuous responses are specific examples. The nonlinear model can also be used when the effects of some covariates are linear and the remaining are nonlinear. Note that the proposed WCQR is new even for linear models.

Model selection with a fixed number of parameters has been widely pursued in the last decades. However, to reduce possible modeling biases, many variables are introduced in practice. As noted in Huber (1973, 1988), Portnoy (1988) and Donoho (2000), the number of parameters p is often large and should be modeled as p_n, which tends to ∞. Fan and Peng (2004) and Lam and Fan (2008) advocated that, in most model selection problems, the number of parameters should be large and grow with the sample size. In a recent seminal paper, Fan and Lv (2010) also studied model selection for generalized linear models with the number of parameters much higher than the sample size. We allow p to depend on the sample size n. To stress dependence on the sample size, we denote the p_n-vector of parameters by $\boldsymbol{\beta}_n = (\beta_{n1}, \ldots, \beta_{np_n})'$ and rewrite (1.1) as:

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}_n) + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{1.2}$$

Without loss of generality, we partition the parameter vector as $\boldsymbol{\beta}_n = ({\boldsymbol{\beta}_{n1}}', {\boldsymbol{\beta}_{n2}}')'$ with $\boldsymbol{\beta}_{n1} \in \mathbf{R}^{s_n}$ and $\boldsymbol{\beta}_{2n} \in \mathbf{R}^{p_n - s_n}$, and assume the true regression coefficients are $\boldsymbol{\beta}_n^* = ({\boldsymbol{\beta}_{n1}^*}', \mathbf{0}')'$, where the s_n components in $\boldsymbol{\beta}_{n1}^*$ do not vanish.

We address the issue of variable/parameter selection using the penalized WCQR with the adaptive LASSO and SCAD penalties. Since the weights in the WCQR are allowed to be negative, the proposed WCQR is different from the common QR and the CQR (see also Section 2). When the weights are all equal and the model is linear with a fixed number of parameters, our method reduces to that of Zou and Yuan (2008a) if the LASSO penalty is employed. Since the proposed WCQR involves a vector of weights, we develop a data-driven weighting strategy that maximizes the efficiency of the WCQR estimators. The resulting estimation is adaptive in the sense that it performs asymptotically the same as if the theoretically optimal weights were used. The adaptive estimation is robust against outliers and heavy-tailed error distributions, such as the Cauchy distribution, and nearly as efficient as the oracle MLE for a variety of error distributions (see Theorem 4 and Table 1). This is a great advantage of the proposed estimation method, since the adaptive WCQR estimators does not require the form of error distribution and achieves nearly the Cramér-Rao lower bound.

The penalized WCQR estimators admit no close form and involve minimizing complicate nonlinear functions, so it is challenging to derive asymptotic properties and to implement the methodology. Theoretically, we establish asymptotic normality of the resulting estimators and show their optimality, no matter whether the error variance is finite or not. Practically, we develop an algorithm for fast implementation of the proposed methodology. This algorithm solves a succession of (penalized) linearized WCQR problems, each of whose dual problems is derived. We extend the "interior point algorithm" (Vanderbei, Meketon, and Freedman (1986) and Koenker and Park (1996)) to solve these dual problems. The resulting algorithm is easy to implement. Simulations endorse our discovery.

The rest of the article is organized as follows. In Section 2 we introduce the penalized WCQR for model (1.2). In section 3 we suggest a computation method for the proposed methodology. In Section 4 we conduct simulations and apply the proposed methods to analyse a dataset. Finally, in the Appendix we give proofs of the theorems.

2. Oracle Model Selection Based on Weighted Composite Quantile Regression

Our idea can be well motivated from the linear model,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \text{ for } i = 1, \ldots, n, \tag{2.1}$$

where $\{\varepsilon_i\}$ are i.i.d. noise with unknown distribution $G(\cdot)$ and density $g(\cdot)$.

By Koenker and Bassett (1978), the τ-th QR estimate of $\boldsymbol{\beta}$ can be obtained via minimizing

$$\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta} - b_\tau)$$

over $\boldsymbol{\beta}$ and b_τ, where $\rho_\tau(u) = u(\tau - I(u < 0))$ is the check function with derivative $\psi_\tau(u) = \tau - I(u < 0)$ for $u \neq 0$. Noticing that the regression coefficients are the same across different QR estimation methods, Zou and Yuan (2008a) proposed to estimate $\boldsymbol{\beta}$ by minimizing

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \mathbf{x}_i'\boldsymbol{\beta} - b_{\tau_k}), \tag{2.2}$$

over $\boldsymbol{\beta}$ and b_{τ_k} and to use the adaptive LASSO penalty (Zou (2006)) for (2.2) to select variables, where $\{\tau_k\}_{k=1}^K$ are predetermined over $(0, 1)$. This is the aforementioned CQR-LASSO.

Note that the CQR method uses the same weight for different QR models. Intuitively, it is more effective if different weights are used, which leads to minimizing

$$\sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \mathbf{x}_i'\boldsymbol{\beta} - b_{\tau_k}),$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)'$ is a vector of weights such that $\|\boldsymbol{\omega}\| = 1$ with $\|\cdot\|$ denoting the Euclidean norm. The weight ω_k controls the amount of contribution of the τ_k-th QR. The components in the weight vector $\boldsymbol{\omega}$ are allowed to be negative, since $\{\sum_{i=1}^{n} \rho_{\tau_k}(y_i - \mathbf{x}_i'\boldsymbol{\beta} - b_{\tau_k})\}_{k=1}^K$ may not be positively correlated. Thus, the WCQR is essentially different from the CQR. Applying the weighting scheme to (1.2), one can estimate $\boldsymbol{\beta}_n$ by minimizing

$$L_n(\boldsymbol{\beta}_n, \mathbf{b}; \boldsymbol{\omega}) \equiv \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - f(\mathbf{x}_i, \boldsymbol{\beta}_n) - b_{\tau_k}) \tag{2.3}$$

over $\boldsymbol{\beta}$ and $\mathbf{b} = (b_{\tau_1}, \ldots, b_{\tau_K})'$. Since this estimation method cannot directly be used to select variables/parameters, we resort to the penalized estimation by minimizing

$$L_n(\boldsymbol{\beta}_n, \mathbf{b}; \boldsymbol{\omega}) + n \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|) \tag{2.4}$$

over $(\boldsymbol{\beta}_n, \mathbf{b})$, where $p_{\lambda_n}(\cdot)$ is a penalty function and λ_n is a non-negative regularization parameter.

For convenience, the minimizer of $\boldsymbol{\beta}_n$ for (2.4) is referred to it as "the penalized WCQR estimator". For linear models, the CQR-LASSO method can be regarded as an example of the penalized WCQR estimation with $\omega_i = 1/\sqrt{K}$. In general, given K, one can use equally spaced quantiles at $\tau_k = k/(K+1)$ for $k = 1, 2, \ldots, K$. In practice, one can choose $K = 10$ to be efficient for most situations. See Table 1 for details.

There are various choices for the penalty function $p_{\lambda_n}(\cdot)$, as discussed in the beginning of the article. In the following we focus on only the SCAD and adaptive-LASSO penalties. The results can be extended to other penalty functions.

2.1. Model selection with SCAD penalty

The SCAD penalty $p_\lambda(\cdot)$ (Fan and Li (2001)) is defined in terms of its first order derivative and is symmetric about the origin. For $\theta > 0$,

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

where $a > 2$ and $\lambda > 0$ are tuning parameters. We obtain the SCAD penalized WCQR by solving

$$(\hat{b}_{\tau_1}, \ldots, \hat{b}_{\tau_K}, \hat{\boldsymbol{\beta}}_n) = \arg\min_{\mathbf{b}, \boldsymbol{\beta}_n} Q_n^{SC}(\boldsymbol{\beta}_n, \mathbf{b}), \tag{2.5}$$

where $Q_n^{SC}(\boldsymbol{\beta}_n, \mathbf{b}) = L_n(\boldsymbol{\beta}_n, \mathbf{b}; \boldsymbol{\omega}) + n\sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|)$. For convenience, the estimation is coined as WCQR-SCAD method.

We establish consistency and asymptotic normality of the SCAD penalized estimator. For clear exposition on the methodology, all regularity conditions are relegated to the Appendix.

Theorem 1. *[Consistency] Suppose the density $g(\cdot)$ satisfies Condition (C), The penalty function $p_{\lambda_n}(\cdot)$ satisfies Conditions $(A_2)-(A_4)$, and the regression function $f(\mathbf{x}_i, \boldsymbol{\beta}_n)$ satisfies Conditions $(B_1)-(B_2)$. If $p_n^3/n \to 0$ as $n \to \infty$, then there is a local minimizer $\hat{\boldsymbol{\beta}}_n$ in (2.5) such that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*\| = O_p(\sqrt{p_n/n})$.*

Let $n_p = np_n^{-1}$, $f_{ni}^* = f(\mathbf{x}_i, \boldsymbol{\beta}_{ni}^*)$, $\boldsymbol{\nabla} f_{ni}^* = [\partial f(\mathbf{x}_i, \boldsymbol{\beta}_n)/\partial \boldsymbol{\beta}_n] \,|_{\boldsymbol{\beta}_n = \boldsymbol{\beta}_n^*}$,

$$\begin{aligned}
\mathbf{c}_n &= \big\{p'_{\lambda_n}(|\beta_{n1}^*|)\mathrm{sgn}(\beta_{n1}^*), \ldots, p'_{\lambda_n}(|\beta_{ns_n}^*|)\mathrm{sgn}(\beta_{ns_n}^*)\big\}',\\
\boldsymbol{\Sigma}_{\lambda_n} &= \mathrm{diag}\big\{p''_{\lambda_n}(\boldsymbol{\beta}_{n1}^*), \ldots, p''_{\lambda_n}(\boldsymbol{\beta}_{ns_n}^*)\big\},\\
\sigma^2(\boldsymbol{\omega}) &= \{\sum_{k=1}^K \omega_k g(b_{\tau_k}^*)\}^{-2} \sum_{k,k'=1}^K \omega_k \omega_{k'} \min(\tau_k, \tau_{k'})(1 - \max(\tau_k, \tau_{k'})),
\end{aligned}$$

where $b_{\tau_k}^*$ is the τ_k-th quantile of ε. Put $\mathbf{g} = (g(b_{\tau_1}^*), \ldots, g(b_{\tau_K}^*))'$ and $\mathbf{G}_n = \mathrm{Var}(\boldsymbol{\nabla} f_{ni}^*)$. Let $\mathbf{G}_{n11}$ be the $s_n \times s_n$ sub-matrix of $\mathbf{G}_n$ corresponding to $\boldsymbol{\beta}_{n1}$, and let $\mathbf{e}_n$ be a $s_n \times 1$ unit vector.

Theorem 2 (Oracle property). *Suppose the conditions in Theorem 1, Condition* (A_1)*, and Condition* (B_3) *hold. If* $\lambda_n \to 0$, $\sqrt{n_p}\lambda_n \to \infty$, *and* $p_n^3/n \to 0$ *as* $n \to \infty$*, then, with probability tending to* 1*, the root-*n_p *consistent local minimizer* $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}'_{n1}, \hat{\boldsymbol{\beta}}'_{n2})'$ *in Theorem* 1 *satisfies*
(i) *Sparsity:* $\hat{\boldsymbol{\beta}}_{n2} = \boldsymbol{0}$; *and*
(ii) *Asymptotic normality:*"

$$\sqrt{n}\mathbf{e}'_n\mathbf{G}_{n11}^{-1/2}\big(\mathbf{G}_{n11}+\frac{\boldsymbol{\Sigma}_{\lambda_n}}{\boldsymbol{\omega}'\mathbf{g}}\big)[(\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}^*_{n1})+\big(\mathbf{G}_{n11}+\frac{\boldsymbol{\Sigma}_{\lambda_n}}{\boldsymbol{\omega}'\mathbf{g}}\big)^{-1}\frac{\mathbf{c}_n}{\boldsymbol{\omega}'\mathbf{g}}] \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0},\sigma^2(\boldsymbol{\omega})).$$

Fan and Peng (2004) established the oracle property of the penalized likelihood estimator under the assumption $p_n^5/n \to 0$. This condition has been relaxed to $p_n^3/n \to 0$ for the WCQR-SCAD method.

Remark 1. When n is finite and large enough, $\boldsymbol{\Sigma}_{\lambda_n} = \mathbf{0}$ and $\mathbf{c}_n = \mathbf{0}$. Hence, Theorem 2 (ii) becomes $\sqrt{n}\mathbf{e}'_n\mathbf{G}_{n11}^{1/2}(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}^*_{n1}) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega}))$, so $\hat{\boldsymbol{\beta}}_{n1}$ enjoys the same efficiency as the WCQR estimator of $\boldsymbol{\beta}_{n1}$ for the submodel with $\boldsymbol{\beta}_{n2} = 0$ known in advance.

As shown in Jennrich (1969) and Wu (1981), for a fixed number of parameters $p_n = p$, the asymptotic variance of the least squares estimator of $\boldsymbol{\beta}$ is $\sigma^2\mathbf{G}_n^{-1}$, where σ^2 is the variance of the error. The result can be extended to the case of a diverging number of parameters p_n. It follows that the asymptotic relative efficiency (ARE) of the WCQR-SCAD estimation with respect to the oracle least squares (OLS) estimation for the submodel with $\boldsymbol{\beta}_{n2} = 0$ known in advance is $\text{ARE}(\boldsymbol{\omega}, g) = \sigma^2\sigma^{-2}(\boldsymbol{\omega})$.

Since the asymptotic variance matrix depends on $\boldsymbol{\omega}$ only through $\sigma^2(\boldsymbol{\omega})$, the weights should be selected to minimize $\sigma^2(\boldsymbol{\omega})$. Let $\boldsymbol{\Omega}$ be a $K \times K$ matrix with the (k, k') element

$$\Omega_{kk'} = \min(\tau_k, \tau_{k'})(1 - \max(\tau_k, \tau_{k'})).$$

Then the optimal weight $\boldsymbol{\omega}_{opt}$, which minimizes $\sigma^2(\boldsymbol{\omega})$, is

$$\boldsymbol{\omega}_{opt} = (\mathbf{g}'\boldsymbol{\Omega}^{-2}\mathbf{g})^{-1/2}\boldsymbol{\Omega}^{-1}\mathbf{g},$$

and with this optimal weight, $\sigma^2(\boldsymbol{\omega}_{opt}) = (\mathbf{g}'\boldsymbol{\Omega}^{-1}\mathbf{g})^{-1}$. The optimal weight components can be very different, and some of them may even be negative, a fact seen in our simulations. The usual nonparametric density estimation methods, such as kernel smoothing based on estimated residuals $\hat{\varepsilon}_i$, can provide a consistent estimation $\hat{g}(\cdot)$ of $g(\cdot)$. Let the resulting estimate of $\mathbf{g}$ be $\hat{\mathbf{g}}$. Then $\hat{\boldsymbol{\omega}} = (\hat{\mathbf{g}}'\boldsymbol{\Omega}^{-2}\hat{\mathbf{g}})^{-1/2}\boldsymbol{\Omega}^{-1}\hat{\mathbf{g}}$ is a nonparametric estimator of $\boldsymbol{\omega}$. This leads to an adaptive estimator of $\boldsymbol{\beta}$ by minimizing

$$L_n(\boldsymbol{\beta}_n, \mathbf{b}; \hat{\boldsymbol{\omega}}) + n \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|) \tag{2.6}$$

over b_{τ_k} and $\boldsymbol{\beta}$. Let the resulting estimator of $\boldsymbol{\beta}$ be $\tilde{\boldsymbol{\beta}}_n$.

Theorem 3. *Under the conditions of Theorem* 2*, with probability tending to* 1*, there exists a root-n_p consistent local minimizer $\tilde{\boldsymbol{\beta}}_n = (\tilde{\boldsymbol{\beta}}'_{n1}, \tilde{\boldsymbol{\beta}}'_{n2})'$ satisfying*
(i) *Sparsity: $\tilde{\boldsymbol{\beta}}_{n2} = \boldsymbol{0}$; and*
(ii) *Asymptotic normality: $\sqrt{n}\mathbf{e}'_n \mathbf{G}_{n11}^{1/2}(\tilde{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}^*_{n1}) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, (\mathbf{g}'\boldsymbol{\Omega}^{-1}\mathbf{g})^{-1})$.*

Since $\sigma^2(\boldsymbol{\omega}_{opt}) = (\mathbf{g}'\boldsymbol{\Omega}^{-1}\mathbf{g})^{-1}$, $\tilde{\boldsymbol{\beta}}_{n1}$ has the same asymptotic variance matrix as $\hat{\boldsymbol{\beta}}_{n1}$, if $\boldsymbol{\omega}_{opt}$ were known. That is, the estimator $\tilde{\boldsymbol{\beta}}_n$ is adaptive. Therefore, $\hat{\boldsymbol{\omega}}$ is called the adaptive weight vector. By Theorem 3, the asymptotic relative efficiency (ARE) of adaptive WCQR estimation with respect to OLS estimation is $e(WCQR, OLS) = \sigma^2\mathbf{g}'\boldsymbol{\Omega}^{-1}\mathbf{g}$. It is easy to show that, for the oracle maximum likelihood (OML) estimator $\hat{\boldsymbol{\beta}}_{n1}^{OML}$ of $\boldsymbol{\beta}_{n1}$, $\sqrt{n}\mathbf{e}'_n\mathbf{G}_{n11}^{-1/2}[\hat{\boldsymbol{\beta}}_{n1}^{OML} - \boldsymbol{\beta}^*_{n1}]$ has asymptotic variance I_g^{-1}, where $I_g = \int [g'(t)]^2/g(t)\, dt$ is the Fisher information, and hence

$$e(WCQR, OML) = I_g^{-1}\mathbf{g}'\boldsymbol{\Omega}^{-1}\mathbf{g}.$$

The following theorem demonstrates that, for equally spaced $\{\tau_k\}_{k=1}^K$, the adaptive estimator $\tilde{\boldsymbol{\beta}}_n$ is nearly efficient as the OML estimators for various error distributions, a great advantage of the proposed methodology.

Theorem 4. *Suppose the derivative $g'(\cdot)$ of $g(\cdot)$ is uniformly continuous. Let $\tau_k = k/(K+1)$ for $k = 1, \ldots, K$. Then, for $K \to \infty$, the limiting ARE of the estimator $\tilde{\boldsymbol{\beta}}_2$ with respect to the OML estimator is*

$$\lim_{K\to\infty} e(WCQR, OML) = 1.$$

For each K, the AREs of the adaptive estimator $\tilde{\boldsymbol{\beta}}_n$ with respect to some common estimators can be calculated. To appreciate how much efficiency is gained in practice, we investigate the performance of common estimators. Table 1 reports AREs for linear models with various error distributions; it shows $\tilde{\boldsymbol{\beta}}_n$ is highly efficient for all distributions under consideration. For linear models, Leng (2009) demonstrated that his regularized rank regression estimator (R^2) was quite efficient and robust. Table 1 indicates that the proposed adaptive estimate dominates R^2 for all error distributions and is much more efficient than it when the error follows the Cauchy or chi-squared distribution. It also suggests that typically one could choose $K = 10$ in practice and such efficiency is largely

Table 1. The relative efficiency of estimators. LAD- Least absolute deviation.

	K	$e(WCQR, R^2)$	$e(WCQR, OML)$	$e(WCQR, OLS)$	$e(WCQR, LAD)$
	10	1.009	0.964	0.964	1.514
Normal	100	1.045	0.998	0.998	1.567
	1000	1.047	1.000	1.000	1.571
	10	1.003	0.961	1.378	1.380
Mixed	100	1.041	0.998	1.430	1.432
Normal	1000	1.044	1.000	1.434	1.436
	10	1.036	0.984	1.967	1.214
t_3	100	1.052	0.999	1.998	1.233
	1000	1.053	1.000	2.000	1.234
	10	1.387	0.585	1.755	2.913
$\chi^2(6)$	100	1.904	0.803	2.410	4.001
	1000	2.154	0.909	2.726	4.525
	10	1.601	0.973	inf	1.201
Cauchy	100	1.644	1.000	inf	1.233
	1000	1.645	1.000	inf	1.234

gained, as shown in simulations. Therefore, with $K = 10$ say, the computational burden associated with the penalized WCQR is not heavy.

2.2. Model selection with adaptive-LASSO

As a variable selection method, LASSO was proposed by Tibshirani (1996) using the L_1 penalty. Zou (2006) introduced the adaptive LASSO by penalizing different parameters with adaptive weights, which makes the LASSO an oracle method. In what follows we develop the adaptive LASSO theory for the WCQR estimation of model (1.2). Denote by $\tilde{\boldsymbol{\beta}}_n$ the solution to $\min_{\boldsymbol{\beta}_n, \mathbf{b}} L_n(\boldsymbol{\beta}_n, \mathbf{b}; \boldsymbol{\omega})$. Then using the same argument as for Theorem 1, $\tilde{\boldsymbol{\beta}}_n$ is $\sqrt{n_p}$-consistent. Thus, we can use $\tilde{\boldsymbol{\beta}}_n$ to construct the adaptive LASSO penalty. Let $\tilde{w}_{nj} = |\tilde{\beta}_{nj}|^{-\gamma}$ for some $\gamma > 0$, and take the adaptive LASSO penalized WCQR estimator to be

$$(\hat{b}_{\tau 1}, \ldots, \hat{b}_{\tau_K}, \hat{\boldsymbol{\beta}}_n^{AL}) = \arg\min_{\mathbf{b}, \boldsymbol{\beta}_n} Q_n^{AL}(\boldsymbol{\beta}_n, \mathbf{b}), \tag{2.7}$$

where $Q_n^{AL}(\boldsymbol{\beta}_n, \mathbf{b}) = L_n(\boldsymbol{\beta}_n, \mathbf{b}; \boldsymbol{\omega}) + nh_n \sum_{j=1}^{p_n} \tilde{w}_{nj} |\beta_{nj}|$, and h_n is a nonnegative regularization parameter. The estimation approach is referred to as the adaptive WCQR-LASSO, for convenience.

Theorem 5 (Consistency)**.** *Suppose the density $g(\cdot)$ satisfies Condition (C) and the regression function $f(\mathbf{x}_i, \boldsymbol{\beta}_n)$ satisfies Conditions (B_1)–(B_2). If $p_n^3/n \to 0$ and $\sqrt{n}h_n \to 0$ as $n \to \infty$, then there is a local minimizer $\hat{\boldsymbol{\beta}}_n^{AL}$ of $Q_n^{AL}(\boldsymbol{\beta}_n, \mathbf{b})$ such that $\|\hat{\boldsymbol{\beta}}_n^{AL} - \boldsymbol{\beta}_n^*\| = O_p(n_p^{-1/2})$.*

Let $\mathbf{d}_n = (\mathrm{sgn}(\beta^*_{n1})/|\tilde{\beta}_{n1}|^\gamma, \ldots, \mathrm{sgn}(\beta^*_{ns_n})/|\tilde{\beta}_{ns_n}|^\gamma)'$.

Theorem 6 (Oracle property)**.** *Suppose the conditions of Theorem* 5 *and condition* (B_4) *hold. If* $h_n n_p^{(\gamma+1)/2} \to \infty$*, then, with probability tending to* 1*, the* $\sqrt{n_p}$*-consistent local minimizer* $\hat{\boldsymbol{\beta}}_n^{AL} = (\{\hat{\boldsymbol{\beta}}_{n1}^{AL}\}', \{\hat{\boldsymbol{\beta}}_{n2}^{AL}\}')'$ *in Theorem 5 satisfies*

(i) *Sparsity:* $\hat{\boldsymbol{\beta}}_{n2}^{AL} = 0$; *and*

(ii) *Asymptotic normality:*

$$\sqrt{n}\mathbf{e}_n'\mathbf{G}_{n11}^{1/2}\Big[(\hat{\boldsymbol{\beta}}_{n1}^{AL} - \boldsymbol{\beta}^*_{n1}) + \frac{\mathbf{G}_{n11}^{-1} h_n \mathbf{d}_n}{\boldsymbol{\omega}'\mathbf{g}}\Big] \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})).$$

Note that $\mathbf{d}_n$ is not zero when n is finite and large enough, hence the bias term for the WCQR-LASSO in Theorem 6 cannot be ignored. By Condition (B_4), $\sqrt{n}h_n\mathbf{d}_n \to \mathbf{0}$, as $n \to \infty$. Therefore, Theorem 6(ii) becomes $\sqrt{n}\mathbf{e}_n'\mathbf{G}_{n11}^{1/2}(\hat{\boldsymbol{\beta}}_{n1}^{AL} - \boldsymbol{\beta}^*_{n1}) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega}))$. This combined with Remark 1 demonstrates that the adaptive WCQR-LASSO and WCQR-SCAD estimators enjoy the same oracle properties.

Remark 2. For model (2.1) with a fixed number of parameters, we have $\mathbf{G}_n \equiv \mathbf{G} = \mathrm{var}(\mathbf{x}_1)$. If all ω_k are equal, Theorem 6 reduces to the asymptotic normality of the adaptive lasso penalized CQR estimator in Zou and Yuan (2008a).

For the above model selection methods we require $p_n^3/n \to 0$. This condition is not the best available in the literature and is chosen partly for simplicity in proofs. He and Shao (2000) derived asymptotic normality of their M-estimator under $p^3(\log p)^2 = o(n)$ using a different argument (see Corollary 2.1 therein); this condition is weaker than ours. Recently, Belloni and Chernozhukov (2011) studied L_1-penalized quantile regression for high-dimensional sparse linear models and established nonasymptotic results and convergence rates of their estimators. We believe that our condition can be further relaxed to $p_n = O(\exp(n^\delta))$ for $0 < \delta < 1$ (NP-dimensionality; see Fan and Lv (2010) and Lv and Fan (2009)). However, establishing results for the WCQR under the current model with NP-dimensionality requires much more complicated techniques. We intend to study this in the future.

3. Numerical Implementation

We introduce a fast algorithm for computation. This algorithm solves a succession of penalized linearized WCQR problems, each of which is solved by extending the interior point algorithm (see Osborne and Watson (1971) and Koenker and Park (1996)). Matlab codes are available upon request for the proposed methods.

Minimization at (2.5) can be done using a similar method as for (2.7), so we first consider the minimization of (2.7). This is equivalent to

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - l_{ik}(\boldsymbol{\theta})) + nh_n \sum_{j=1}^{p_n} \tilde{w}_{nj}|\beta_{nj}|, \tag{3.1}$$

where $l_{ik}(\boldsymbol{\theta}) = f(\mathbf{x}_i, \boldsymbol{\beta}_n) + b_{\tau_k}$ with $\boldsymbol{\theta} = (\mathbf{b}', \boldsymbol{\beta}_n')'$. Following Osborne and Watson (1971), we solve (3.1) using the following algorithm.

(1) Given the current value, $\boldsymbol{\theta}^{(r)}$, of $\boldsymbol{\theta}$, calculate $\mathbf{t}$ to minimize

$$\sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}\{y_i - l_{ik}(\boldsymbol{\theta}^{(r)}) - \boldsymbol{\nabla} l_{ik}(\boldsymbol{\theta}^{(r)})\mathbf{t}\} + nh_n \sum_{j=1}^{p_n} \tilde{w}_{nj}|\beta_{nj}|, \tag{3.2}$$

where $\boldsymbol{\nabla} l_{ik}(\boldsymbol{\theta}^{(r)}) = \frac{\partial l_{ik}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(r)}}$ and β_{nj} is the $(K+j)$th component of $\boldsymbol{\theta}^{(r)} + \mathbf{t}$. Let the minimizer be $\mathbf{t} = \mathbf{t}^{(r)} = (t_1^{(r)}, \ldots, t_{K+p_n}^{(r)})'$.

(2) Calculate $\lambda \in [0, 1]$ to minimize

$$\sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}\{y_i - l_{ik}(\boldsymbol{\theta}^{(r)} + \lambda \mathbf{t}^{(r)})\} + nh_n \sum_{j=1}^{p_n} \tilde{w}_{nj}|\beta_{nj}^{(r)} + \lambda t_{K+j}^{(r)}|. \tag{3.3}$$

Let the minimizer be $\lambda = \lambda^{(r)}$.

(3) Put $\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} + \lambda^{(r)}\mathbf{t}^{(r)}$. Update the current value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}^{(r+1)}$, and repeat the above procedure until convergence.

Here the problem (3.3) can easily be solved by line search in the resulting direction $\mathbf{t} = \mathbf{t}^{(r)}$, but one has to solve a succession of penalized linearized WCQR problems in (3.2). Let $y_{ik}^* = y_i - l_{ik}(\boldsymbol{\theta}^{(r)})$ and $\mathbf{a}_{ik}' = \nabla l_{ik}(\boldsymbol{\theta}^{(r)})$. Then the problem (3.2) becomes

$$\min_{\mathbf{t}} \{\sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}(y_{ik}^* - \mathbf{a}_{ik}'\mathbf{t}) + nh_n \sum_{j=1}^{p_n} \tilde{w}_{nj}|\beta_{nj}|\}. \tag{3.4}$$

For $j = 1, \ldots, p_n$ and $k = 1, \ldots, K$, let $y_{(n+j)k}^* = 0$ and $\mathbf{a}_{(n+j)k} = nh_n\tilde{w}_{nj}\mathbf{e}_{K+j}$, where $\mathbf{e}_{K+j}$ is a $(K+p_n) \times 1$ vector with the $(K+j)$th entry being one and others being zeros. Then (3.4) ss the linear programming problem:

$$\min_{\mathbf{t}} \sum_{k=1}^{K} \omega_k \big\{\sum_{i=1}^{n} \rho_{\tau_k}(y_{ik}^* - \mathbf{a}_{ik}'\mathbf{t}) + \sum_{i=n+1}^{n+p_n} \omega_k|y_{ik}^* - \mathbf{a}_{ik}'\mathbf{t}|\big\}. \tag{3.5}$$

For $k = 1, \ldots, K$, let $\mathbf{y}_k^* = (y_{1k}^*, \ldots, y_{n+p_n,k}^*)'$, $\mathbf{u}_k = \text{vec}(\mathbf{y}_k^*, \mathbf{0}_{p_n \times 1})$, $\mathbf{u} = (\mathbf{u}_1', \ldots, \mathbf{u}_k')'$, $\mathbf{A}_k = (\mathbf{a}_{1k}, \ldots, \mathbf{a}_{n+p_n,k})'$, and $\mathbf{A} = (\mathbf{A}_1', \ldots, \mathbf{A}_K')'$. Then the dual problem of (3.5) is

$$\max_{\mathbf{d}} \{\mathbf{u}'\mathbf{d} \,|\, \mathbf{A}'\mathbf{d} = 0\}, \tag{3.6}$$

where $\mathbf{d} = \text{vec}(\mathbf{d}_1, \ldots, \mathbf{d}_K)$, $\mathbf{d}_k = \text{vec}(\mathbf{d}_k^{(1)}, \mathbf{d}_k^{(2)})$, $\mathbf{d}_k^{(1)} = (d_{1k}, \ldots, d_{nk})' \in [\omega_k(\tau_k - 1), \omega_k \tau_k]^n$, and $\mathbf{d}_k^{(2)} = (d_{n+1,k}, \ldots, d_{n+p_n,k})' \in [-\omega_k^2, \omega_k^2]^{p_n}$.

There are two methods, the simplex and the interior point, for solving (3.6). Here we opt for the latter due to its advantages (Bassett and Koenker (1992) and Koenker and Park (1996)): computational simplicity and natural extensions to nonlinear problems; unlike the simplex-based method, the interior point algorithm converges to the correct solution. Algorithmic details for the dual problem (3.6) proceed as follows.

1. For any initial feasible $\mathbf{d}$, e.g., $\mathbf{d} = \mathbf{0}$, following Vanderbei, Meketon, and Freedman (1986), take a $n \times n$ diagonal matrix $\mathbf{D}_k^{(1)}$ with (i, i) entry $\min\{\omega_k \tau_k - d_{ik}, d_{ik} - \omega_k(\tau_k - 1)\}$, and a $p_n \times p_n$ diagonal matrix $\mathbf{D}_k^{(2)}$ with (i, i) entry $\min\{\omega_k^2 - d_{ik}, d_{ik} + \omega_k^2\}$. Let $\mathbf{D}_k = \text{diag}(\mathbf{D}_k^{(1)}, \mathbf{D}_k^{(2)})$, $\mathbf{D} = \text{diag}(\mathbf{D}_1, \ldots, \mathbf{D}_K)$, $\mathbf{s} = \mathbf{D}^2(\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{D}^2\mathbf{A})^{-1}\mathbf{A}'\mathbf{D}^2)\mathbf{u}$, and $\mathbf{t} = (\mathbf{A}'\mathbf{D}^2\mathbf{A})^{-1}\mathbf{A}'\mathbf{D}^2\mathbf{u}$.
2. Set $\mathbf{d}^* = \mathbf{d} + (a_0/\gamma)\mathbf{s}$, where $\mathbf{s} = \text{vec}(\mathbf{s}_1, \ldots, \mathbf{s}_K)$, $\mathbf{s}_k = (s_{1k}, \ldots, s_{n+p_n,k})'$, $\gamma = \max(\gamma_1, \ldots, \gamma_K)$, $\gamma_k = \max(\gamma_k^{(1)}, \gamma_k^{(2)})$,
$$\gamma_k^{(1)} = \max_{1 \le i \le n} \left(\max \left\{ \frac{s_{ik}}{\omega_k \tau_k - d_{ik}}, -\frac{s_{ik}}{d_{ik} - \omega_k(\tau_k - 1)} \right\} \right),$$
$$\gamma_k^{(2)} = \max_{n+1 \le i \le n+p_n} \left(\max \left\{ \frac{s_{ik}}{\omega_k^2 - d_{ik}}, -\frac{s_{ik}}{d_{ik} + \omega_k^2} \right\} \right),$$
for $k = 1, \ldots, K$, and $a_0 \in (0, 1)$ is a constant chosen to insure feasibility. As suggested by Koenker and Park (1996), we take $a_0 = 0.97$.
3. Set $\mathbf{d} = \mathbf{d}^*$. Updating $\mathbf{D}$, $\mathbf{s}$ and $\mathbf{d}$ continues the iteration.

After solving (3.6) using this interior point algorithm, we arrive at the next loop that uses the current value $\boldsymbol{\theta} = \boldsymbol{\theta}^{(r+1)}$ for the primal problem in (3.5). This leads to the updated dual problem (3.6) with $y_{ik}^* = y_i - l_{ik}(\boldsymbol{\theta}^{(r+1)})$ and $\mathbf{a}_{ik}' = \nabla l_{ik}(\boldsymbol{\theta}^{(r+1)})$ for $i = 1, \ldots, n$. The current $\mathbf{d}$ should be adjusted to ensure that it is feasible for the new value of $\mathbf{A}$. Similar to Koenker and Park (1996), we project the current $\mathbf{d}$ onto the null space of the new $\mathbf{A}$, $\widehat{\mathbf{d}} = (\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')\mathbf{d}$, and then shrink it to insure that $\mathbf{d}_k^{(1)} \in [\omega_k(\tau_k - 1), \omega_k \tau_k]^n$ and $\mathbf{d}_k^{(2)} \in [-\omega_k^2, \omega_k^2]^{p_n}$. The adjusted $\mathbf{d}$ is $\mathbf{d} = \widehat{\mathbf{d}}/(m + \delta)$ for some $\delta > 0$, where $m = \max(m_1, m_2, \ldots, m_K)$, with $m_k = \max(m_k^{(1)}, m_k^{(2)})$, $m_k^{(1)} = \max_{1 \le i \le n}\{\max(\hat{d}_{ik}/\omega_k(\tau_k - 1), \hat{d}_{ik}/\omega_k \tau_k)\}$, and $m_k^{(2)} = \max_{n+1 \le i \le n+p_n}\{|\hat{d}_{ik}/\omega_k^2|\}$.

As noted by Koenker and Park (1996), the difficulty with the above method is twofold: one must solve a linearized problem (3.2) or equivalently (3.5) at each iteration; the resulting search directions may be inferior to directions determined by incomplete solutions to the sequence of linearized problems. As they suggest, when $f(\mathbf{x}_i, \boldsymbol{\beta}_n)$ is nonlinear there is no longer a compelling argument for fully solving (3.2), using only a few iterations to refine the dual vector is preferable. This reduces the computational burden.

Next, we consider (2.5). By Taylor's expansion for the SCAD penalty at an initial consistent estimate $\boldsymbol{\beta}_n^0$ (for example the common L_1-norm estimate), we have

$$p_{\lambda_n}(|\beta_{nj}|) \approx p'_{\lambda_n}(|\beta^0_{nj}|)|\beta_{nj}| + \{p_{\lambda_n}(|\beta^0_{nj}|) - p'_{\lambda_n}(|\beta^0_{nj}|)|\beta^0_{nj}|\},$$

where $p_{\lambda_n}(|\beta^0_{nj}|) - p'_{\lambda_n}(|\beta^0_{nj}|)|\beta^0_{nj}|$ is a constant. Therefore, (2.5) is reduced to

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - l_{ik}(\boldsymbol{\theta})) + n \sum_{j=1}^{p_n} p'_{\lambda_n}(|\beta^0_{nj}|)|\beta_{nj}|,$$

which can be solved using the same algorithm as for (3.1). Update the initial value for $\boldsymbol{\beta}_n$ and do iterations until convergence, where a few steps can lead to convergence since $\boldsymbol{\beta}_n^0$ is close to the true parameter.

4. Numerical Studies

4.1. Choice of the tuning parameters

For the penalized WCQR estimators, one has to select tuning parameters λ_n and h_n, respectively, for the SCAD and LASSO penalties. The two parameters can be chosen using the same method. We focus on the choice of λ_n.

There are several methods for selecting λ_n, including the generalized cross-validation (*GCV*) criterion (Wang, Li, and Tsai (2007)) and the Schwartz Information Criterion (*SIC*) (see Koenker, Ng, and Portnoy (1994) and Zou and Yuan (2008b)). Since the resulting estimators depend on λ_n, we denote the estimators by $(\hat{\boldsymbol{\beta}}_{\lambda_n}, \hat{\mathbf{b}}_{\lambda_n})$ to stress such dependence. Applying the *SIC* method, we propose to select λ_n by minimizing

$$SIC(\lambda_n) = \log\{\frac{1}{nK} L_n(\hat{\boldsymbol{\beta}}_{\lambda_n}, \hat{\mathbf{b}}_{\lambda_n})\} + \frac{\log(nK)}{2nK} df(\lambda_n)$$

over λ_n, where $df(\lambda_n)$ is the effective degrees of freedom of the fitted model that calibrates the complexity of model.

Following Koenker, Ng, and Portnoy (1994), for each given λ_n we take

$$\mathcal{E}_{\lambda_n} = \{(k, i) : y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\lambda_n}) - \hat{b}_{\lambda_n, \tau_k} = 0\}$$

and use the size $|\mathcal{E}_{\lambda_n}|$ of $\mathcal{E}_{\lambda_n}$ to estimate $df(\lambda_n)$. Nychka et al. (1995) and Yuan (2006) proposed to use Stein's (1981) SURE divergence formula $\sum_{i=1}^{n} \partial \hat{f}(\mathbf{x}_i)/\partial y_i$

to estimate df, where $\hat{f}(\mathbf{x}_i)$ is a fitted model. For the linear models $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, it is easy to see that $\sum_{i=1}^n \partial \hat{f}(\mathbf{x}_i)/\partial y_i$ is the dimension of $\boldsymbol{\beta}$ if the least squares estimation method is used. Li, Liu, and Zhu (2007) and Li and Zhu (2008) showed that, for quantile regression, $\sum_{i=1}^n \partial \hat{f}(\mathbf{x}_i)/\partial y_i = |\mathcal{E}_{\lambda_n}|$. Therefore, it is reasonable to use $|\mathcal{E}_{\lambda_n}|$ to estimate $df(\lambda_n)$. This leads to the tuning-parameter estimate

$$\hat{\lambda}_n = \arg\min_{\lambda_n} \Big\{ \log\big(\frac{1}{nK} L_n(\hat{\boldsymbol{\beta}}_{\lambda_n}, \hat{\mathbf{b}}_{\lambda_n}) \big) + \frac{\log(nK)}{2nK} |\mathcal{E}_{\lambda_n}| \Big\}.$$

4.2. Simulations

In this section we report on simulations to investigate finite sample performance of the WCQR estimation and the associated model selection. An exponential regression model was used:

$$y = 1 + b\exp(\mathbf{c}'\mathbf{x}) + \varepsilon,$$

where b and $\mathbf{c} = (c_1, c_2, c_3)'$ are parameters, ε is the error. The true values of parameters were set as $b = 1.5$, and $\mathbf{c} = (-0.6, -0.8, -0.7)'$.

When the penalized WCQR methods were considered, we allowed the lengths of $\mathbf{c}$ and the relevant $\mathbf{x}$ to increase with the sample size, setting $\mathbf{c} = (-0.6, -0.8, -0.7, 0, \ldots, 0)'$. Two penalties were employed: the adaptive LASSO penalty with $\gamma = 1$, defined by $nh_n \sum_{j=1}^{p_n} |\beta_j|/|\tilde{\beta}_j|$; the SCAD penalty, defined by $n \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|)/|\tilde{\beta}_j|$, where h_n and λ_n are tuning parameters and $\tilde{\beta}_j$'s are consistent estimators of β_j's. For simplicity, we used the LASSO with $\gamma = 1$ that closely relates to the nonnegative garotte (Breiman (1995)) as shown in in Zou (2006). Other values for γ are possible, since there is no optimal theoretical values for it. The tuning parameters were determined by *SIC* method, and the number of quantiles K was 10, as suggested in Section 2. Since the WCQR estimator involves a weighting scheme and the density of error is known in simulations, we took the optimal weight $\boldsymbol{\omega}_{opt}$ (see Section 2) for all simulations.

Following the suggestion of the AE, we compared the performance of the above penalized methods with the "naive" method that simply sets zero penalty for coefficients and hard-thresholds the resulting estimator. Specifically, we used the hard thresholding rule $\hat{\beta}_j(\lambda_n) = \hat{\beta}_j I(|\hat{\beta}_j| > \lambda_n)$, where $\hat{\beta}_j$ was the resulting estimate of β_j using the L_1 or CQR or $WCQR$ methods, and λ_n was the threshold parameter selected by SIC based on the naive estimator.

With $\boldsymbol{\beta}_n = (b, \mathbf{c}')'$ as the $p_n \times 1$ vector of parameters in the working model, we drew from the working model 400 samples of sizes 200 and 400 with $p_n = [n^{1/3}] + 3$. In each simulation, the first component of $\mathbf{x}$ was $U[-1, 1]$, and the remaining components of $\mathbf{x}$ were jointly normal distributed with the pairwise correlation coefficient 0.5 and standard normal as marginals. We considered four sets of errors: $N(0,1)$, $t(5)$, $0.1N(0,1)+0.9N(0,3^2)$ and $\chi^2(4)$. All of them were centralized and scaled so that the medians of the absolute errors were ones.

We compared five estimation methods: the penalized L_1, CQR, and WCQR estimation, naive estimation, and OML estimation. In each simulation the "root of mean squared errors (RMSE)" for different coefficient estimators were calculated, and their average over simulations is reported in Tables 2−5, where Σ denotes the sum of RMSE for all components in $\boldsymbol{\beta}$. Clearly, the OML estimator performed best, the penalized WCQR performed comparably to the oracle estimator, and the naive method was the worst. This is expected, since the hard-thresholding rule is discontinuous and creates unnecessary bias when the regression coefficients are large. The SCAD penalty function leaves large values of β_j not excessively penalized and makes the resulted solution continuous, and hence does not create excessive biases when β_j's are large (see Fan and Li (2001)). This exemplifies the theory about the penalized WCQR estimation: asymptotically the penalized WCQR estimation performed as well as if the correct submodel were known and had almost the same efficiency of OML estimation; the penalized WCQR performed much better than the penalized CQR and L_1 when the error was chi-squared, but the two methods were comparable when the errors were symmetric, such as normal, mixed normal and $t(5)$. In Table 6 we report the frequency that zero coefficients were set to zero correctly if their estimates were less than 10^{-8}; it shows that the frequency was higher for larger sample size. In this example, all non-zero coefficients were set to non-zero correctly.

As noted by the AE, it is not clear that SIC picks the best penalty level for model selection and the estimation of coefficients. We explored this issue in simulations and tested the limit of the algorithm by using larger p_n, $p_n = [n^{1/2}]$, $[n^{2/3}]$, and n. Our experience suggests that it works for $p_n = [n^{1/2}]$ but fails for the other two scenarios. The results here do not really support the conjecture in the last paragraph of Section 2, but the asymptotically weak correlation condition between the important variables and the unimportant variables did not hold in our simulations (see also Condition 2 of Fan and Lv (2010)). To save space we report only the results under normal error in Tables 7−8. Compared to other penalized estimators, the penalized WCQR still performed the best. However, the frequency of correctly identifying zero coefficients was not higher for larger sample size. Following the suggestion of a referee, we studied the effect of estimating optimal weights for the proposed estimators. Table 9 reports the results with chi-squared error. Compared with Table 5, it can be seen that, for large sample size, our estimators with estimated weights performed nearly as well as the estimators with optimal weights. This is expected from our theoretical results.

Table 2. RMSE(multiplied by 10^3) of penalized estimators under the normal error. SC- SCAD, LA - LASSO, NA - Naive.

	$n = 200$					$n = 400$				
Estimates	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ
SC-L_1	233	67	55	50	411	140	44	33	31	248
SC-CQR	194	57	47	42	351	120	38	27	27	213
SC-WCQR	192	57	46	41	342	119	38	27	26	210
LA-L_1	227	67	55	49	410	140	43	34	31	248
LA-CQR	191	57	47	41	350	120	38	27	27	214
LA-WCQR	188	57	46	41	341	119	39	27	26	213
L_1-NA	259	80	63	59	696	158	47	39	36	457
CQR-NA	211	69	54	50	576	130	41	32	30	377
WCQR-NA	210	68	53	50	569	128	41	31	29	372
WCQR-Oracle	191	57	46	41	336	120	39	27	26	213
OML	188	57	46	41	333	118	38	27	26	209

Table 3. RMSE (multiplied by 10^3) of penalized estimators under the normalized t(5) error. SC- SCAD, LA - LASSO, NA - Naive.

	$n = 200$					$n = 400$				
Estimates	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ
SC-L_1	223	69	56	47	407	146	47	34	31	258
SC-CQR	212	62	51	45	382	133	43	31	28	238
SC-WCQR	211	62	51	45	379	133	42	31	27	238
LA-L_1	217	71	56	47	407	147	46	34	31	258
LA-CQR	209	61	51	45	386	132	42	31	28	240
LA-WCQR	210	62	51	45	386	131	43	31	27	237
L_1-NA	259	80	63	59	696	158	47	39	36	457
CQR-NA	211	69	54	50	576	130	41	32	30	377
WCQR-NA	210	68	53	50	569	128	41	31	29	372
WCQR-oracle	212	62	52	46	372	131	41	31	28	232
OML	210	62	51	45	367	131	41	31	27	231

4.3. A data example

Patients in hospitals are at risk of infection. To study Efficacy of Nosocomial Infection Control (SENIC), the Hospital Infections Program was conducted by Robert W. Haley and his collaborators, Center for Infectious Diseases, Centers for Disease Control, Atlanta, Georgia 30333. This resulted in the SENIC dataset for the 1975-76 study period, consisting of a random sample of 113 hospitals selected from the original 338 hospitals surveyed (see Kutner et al. (2005)). For each hospital there are 11 variables.

- Infection risk (y): Average estimated probability of acquiring an infection in the hospital.

Table 4. RMSE (multiplied by 10^3) of penalized estimators under the mixed normal error. SC- SCAD, LA - LASSO, NA - Naive.

	$n=200$					$n=400$				
Estimates	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ
SC-L_1	256	73	58	54	456	150	45	35	32	265
SC-CQR	208	60	49	45	377	129	38	30	27	228
SC-WCQR	204	59	48	44	364	128	38	30	27	226
LA-L_1	250	71	58	54	451	147	44	35	32	263
LA-CQR	209	60	49	46	384	129	38	30	27	231
LA-WCQR	203	59	48	44	370	128	38	30	27	228
L_1-NA	285	82	69	66	740	165	48	39	36	466
CQR-NA	235	67	59	54	610	136	42	33	30	389
WCQR-NA	232	67	58	53	602	135	42	33	30	388
WCQR-oracle	205	59	49	44	357	130	38	30	28	227
OML	204	59	48	45	356	129	38	30	28	225

Table 5. RMSE (multiplied by 10^3) of penalized estimators under the normalized $\chi^2(4)$ error. SC- SCAD, LA - LASSO, NA - Naive.

	$n=200$					$n=400$				
Estimates	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ
SC-L_1	198	60	47	45	356	131	41	31	25	229
SC-CQR	156	49	38	36	289	98	32	23	20	179
SC-WCQR	121	40	30	29	219	79	27	19	17	141
LA-L_1	197	60	48	44	359	130	42	31	26	231
LA-CQR	155	48	38	36	296	98	32	23	20	183
LA-WCQR	121	39	31	29	224	78	26	19	17	140
L_1-NA	225	70	59	53	623	149	46	35	31	421
CQR-NA	175	55	46	43	492	112	36	28	24	328
WCQR-NA	139	45	39	35	401	90	28	23	21	267
WCQR-oracle	125	39	32	30	226	79	24	19	17	139
OML	99	34	28	25	185	60	20	15	14	108

- Length of stay (x_1): Average length of stay of all patients in the hospital (in days).
- Age (x_2): Average age of patients (in years).
- Routine culturing ratio (x_3): Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100.
- Routine chest X-ray ratio (x_4): Ratio of number of X-rays performed to numbers of patients without signs or symptoms of pneumonia, times 100.
- Number of beds (x_5): Average number of beds in the hospital during the study period.

Table 6. The frequency of correctly identifying zero coefficients.

Error\Method		$n=200$ Naive	LASSO	SCAD	$n=400$ Naive	LASSO	SCAD
$N(0,1)$	L_1	0.226	0.698	0.784	0.222	0.876	0.875
	CQR	0.266	0.725	0.768	0.267	0.937	0.930
	WCQR	0.265	0.845	0.857	0.275	0.933	0.935
$t(5)$	L_1	0.227	0.610	0.680	0.219	0.814	0.830
	CQR	0.250	0.878	0.889	0.257	0.841	0.853
	WCQR	0.246	0.836	0.894	0.259	0.854	0.885
mixed normal	L_1	0.227	0.619	0.699	0.225	0.791	0.835
	CQR	0.275	0.678	0.729	0.258	0.865	0.881
	WCQR	0.276	0.734	0.761	0.255	0.881	0.906
$\chi^2(4)$	L_1	0.244	0.692	0.746	0.225	0.738	0.800
	CQR	0.285	0.695	0.733	0.272	0.796	0.834
	WCQR	0.251	0.880	0.936	0.252	0.978	0.991

Table 7. RMSE (multiplied by 10^3) of penalized estimators when the error is normal and $p_n = [n^{1/2}]$. SC- SCAD, LA - LASSO, NA - Naive.

Estimates	$n=200$ $\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ	$n=400$ $\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ
SC-L_1	239	66	57	50	418	139	46	30	30	251
SC-CQR	190	55	46	43	363	115	38	25	25	226
SC-WCQR	184	53	44	43	343	116	38	25	25	217
LA-L_1	230	64	56	50	410	134	45	29	30	253
LA-CQR	191	56	46	43	379	116	38	26	26	250
LA-WCQR	184	53	45	43	350	116	38	25	25	227
L_1-NA	303	85	78	72	1073	175	56	43	40	855
CQR-NA	235	69	61	58	870	140	46	35	34	709
WCQR-NA	230	68	60	57	855	140	45	35	33	702
WCQR-Oracle	191	57	46	41	336	120	39	27	26	213
OML	188	57	46	41	333	118	38	27	26	209

Table 8. The frequency of correctly identifying zero coefficients when the error is normal and $p_n = [n^{1/2}]$.

Method\Penalty	$n=200$ Naive	LASSO	SCAD	$n=400$ Naive	LASSO	SCAD
L_1	0.125	0.810	0.862	0.100	0.736	0.800
CQR	0.159	0.655	0.747	0.134	0.650	0.758
WCQR	0.166	0.733	0.810	0.133	0.761	0.805

- Medical school affiliation (x_6): 1=Yes, 2=No.
- Region (x_7-x_9): Geographic region: 1=NE, 2=NC, 3=S, 4=W.
- Average daily census (x_{10}): Average number of patients in the hospital per

Table 9. RMSE (multiplied by 10^3) of penalized estimators with estimated weights under the normalized $\chi^2(4)$ error. SC- SCAD, LA - LASSO, NA - Naive.

	$n = 200$					$n = 400$				
Estimates	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ
SC-WCQR	133	41	34	30	242	81	26	19	17	143
LA-WCQR	133	40	34	31	246	80	26	19	17	142
WCQR-NA	146	47	39	37	422	92	29	23	21	273
WCQR-oracle	129	40	33	30	233	80	26	19	17	142

day during the study period.

- Number of nurses (x_{11}): Average number of full-time equivalent registered and licensed practical nurses during the study period (number full time plus one half the number part time).
- Available facilities and services (x_{12}): Percent of 35 potential facilities and services that are provided by the hospital.

We study whether the infection risk depends on the possible influential factors and target a good estimate for infection risk, after adjusting for contributions from confounding factors. Since the medical school affiliation and region are categorical, we introduced a dummy variable x_6 for the medical school affiliation and three dummy variables (x_7, x_8, x_9) for the region as covariates. Note that the response y (infection risk) is the average estimated probability of acquiring an infection in the hospital. It is sensible to use a logistic model with all of covariates,

$$y_i = \frac{\exp(\beta_0 + \sum_{i=1}^{12} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{12} \beta_i x_i)} + \varepsilon_i, \quad i = 1, \ldots, 113,$$

to model the relationship between the infection risk and all possible infection factors, where all of covariates are used to reduce possible modeling biases and the number of non-zero parameters is assumed to depend on the sample size.

We applied the L_2-penalized least squares estimation (LSE) and the penalized CQR and WCQR methods with adaptive LASSO and SCAD penalties to select the non-zero parameters or significant variables. We employed classic kernel smoothing over the residuals from CQR-estimation to estimate the density of error. The estimator takes the form $\hat{g}(x) = (1/nh) \sum_{i=1}^{n} K((\hat{\varepsilon}_i - x)/h)$, where K is a density kernel, h is the bandwidth controlling the amount of smoothing, and the $\hat{\varepsilon}_i$'s are residuals from the CQR method. Then we obtained the data-driven weight vector $\hat{\boldsymbol{\omega}}$. The SIC criterion (Section 5.1) was applied to choose the tuning parameters. The results of variable selection are presented in Table 10. From

Table 10. Estimates and standard errors (multiplied by 10^4).

Penalty	L_2	LASSO		SCAD	
Method	LSE	CQR	WCQR	CQR	WCQR
x_1	574 (335)	0 (-)	0 (-)	0 (-)	0 (-)
x_2	667 (105)	743 (113)	705 (102)	745 (114)	713 (103)
x_3	55 (40)	0 (-)	0 (-)	0 (-)	0 (-)
x_4	-31 (23)	-25 (36)	-32 (33)	-25 (37)	-23 (32)
x_5	-18 (12)	-12 (17)	-5 (16)	-10 (17)	-5 (16)
x_6	229 (1302)	0 (-)	0 (-)	0 (-)	0 (-)
x_7	66 (1512)	0 (-)	0 (-)	0 (-)	0 (-)
x_8	-100 (1359)	0 (-)	0 (-)	0 (-)	0 (-)
x_9	250 (1343)	0 (-)	0 (-)	0 (-)	0 (-)
x_{10}	15 (14)	23 (21)	12 (21)	21 (21)	12 (20)
x_{11}	9 (7)	0 (-)	0 (-)	0 (-)	0 (-)
x_{12}	-14 (46)	0 (-)	0 (-)	0 (-)	0 (-)

Table 10, we can see that penalized SCAD and penalized LASSO methods both selected four variables: age (x_2), routine chest X-ray ratio (x_4), number of beds (x_5), and average daily census (x_{10}), but the penalized LSE selected all variables (note that x_7-x_9 together represent the region). Similar to ridge regression for linear models, the LSE with L_2-penalty failed to shrink any coefficients directly to zero for the nonlinear model.

Since the estimated coefficients were negative for x_4 and x_5 and positive for x_2 and x_{10}, the above analysis indicates that, during the study period, infection risk (y) increases with the average age of patients (x_2) and the average number of patients in hospital per day (x_{10}), and decreases with the routine chest X-ray ratio (x_4) and average number of beds in hospital (x_5). This is expected, since elderly patients tend to have a weak resistance to infection, and a larger x_{10} results in a smaller value of x_5 and increases the chance of cross-infection among patients. In addition, routine chest X-ray may do harm to the body, and patients without signs or symptoms of pneumonia should receive it as little as possible.

To check the significance of the selected model, we considered the hypothesis testing problem:

$$H_0 : \beta_2 = \beta_4 = \beta_5 = \beta_{10} = 0 \text{ versus } H_1 : \text{at least one of them is non-zero.}$$

The LSE was used to estimate the parameters in the null and alternative models, with $SSE(H_0)$ and $SSE(H_1)$ the residual sum of squares under H_0 and H_1, respectively. Let

$$F = \frac{SSE(H_0) - SSE(H_1)}{df_0 - df_1} \Big/ \frac{SSE(H_1)}{df_1},$$

where $df_0 = n-1$ and $df_1 = n-5$ degrees of freedom for the null and alternative models, respectively. Then the approximate null distribution of F-statistic is $F(df_0 - df_1, df_1)$. The realized value of F was calculated as 124.541 with approximate p-value equal to zero. Hence, the selected model was significant.

Acknowledgements

The authors are thankful to an associate editor and the referees for their valuable comments and suggestions which have improved the article substantially. This research was supported in part by GRF grant 403109 from the Research Grant Council of the Hong Kong Special Administration Region. Additional partial support was also provided by NSFC grant 11101432 for Xunjun Jiang and NSF Grant DMS-09-06482 for Jiancheng Jiang. All correspondence should be addressed to Jiancheng Jiang, Department of Mathematics and Statistics, University of North Carolina at Charlotte, NC 28223, USA; E-mail: jjiang1@uncc.edu.

Appendix. Conditions and Proofs of Theorems

A.1. Regularity conditions

(i) *Regularity conditions on the penalty.* Let $a_n = \max_{1\le j\le p_n}\{p'_{\lambda_n}(|\beta^*_{nj}|), \beta^*_{nj} \neq 0\}$, and $b_n = \max_{1\le j\le p_n}\{p''_{\lambda_n}(|\beta^*_{nj}|), \beta^*_{nj} \neq 0\}$. The conditions on penalty functions are:

(A_1) $\liminf_{n\to+\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$;

(A_2) $a_n = O(n^{-1/2})$;

(A_3) $b_n \to 0$ as $n \to +\infty$;

(A_4) there are constants C and D such that $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \le D|\theta_1 - \theta_2|$, where $\theta_1, \theta_2 > C\lambda_n$.

Conditions (A_1)-(A_4) are also the regularity conditions on the penalty given in Fan and Peng (2004).

(ii) *Regularity conditions on the regression function.*

(B_1) There is a large enough open subset $\Omega_n \in \mathbf{R}^{p_n}$ that contains the true parameter point $\boldsymbol{\beta}^*_n$, such that for all $\mathbf{x}_i$ the second derivative matrix $\nabla^2 f(\mathbf{x}_i, \boldsymbol{\beta}_n)$ of $f(\mathbf{x}_i, \boldsymbol{\beta}_n)$ with respect to $\boldsymbol{\beta}_n$, satisfies

$$\|\nabla^2 f(\mathbf{x}_i, \boldsymbol{\beta}_{n1}) - \nabla^2 f(\mathbf{x}_i, \boldsymbol{\beta}_{n2})\| \le M(\mathbf{x}_i)\|\boldsymbol{\beta}_{n1} - \boldsymbol{\beta}_{n2}\|$$
$$\Big|\frac{\partial^2 f(\mathbf{x}_i, \boldsymbol{\beta}_n)}{\partial\beta_{nj}\partial\beta_{nk}}\Big| \le N_{jk}(\mathbf{x}_i)$$

for all $\boldsymbol{\beta}_n \in \Omega_n$, with $E[M^2(\mathbf{x}_i)] < \infty$, $E[N^2_{jk}(\mathbf{x}_i)] < C_1 < \infty$ for all j, k.

(B_2) $\mathrm{Var}(\nabla f_{ni}^*) = \mathbf{G}_n > \mathbf{0}$, $\mathrm{E}((\nabla f_{ni}^*)^{\otimes 2}) = \mathbf{\Gamma}_n$, and $0 < d_1 < \lambda_{\min}(\mathbf{\Gamma}_n) \leq \lambda_{\max}(\mathbf{\Gamma}_n) < d_2 < \infty$, for all n, where $\lambda_{\min}(\mathbf{\Gamma}_n)$ and $\lambda_{\max}(\mathbf{\Gamma}_n)$ are the smallest and largest eigenvalues of $\mathbf{\Gamma}_n$.

(B_3) $\beta_{n1}^*, \beta_{n2}^*, \ldots, \beta_{ns_n}^*$ satisfy $\min_{1\leq j\leq s_n} |\beta_{nj}^*|/\lambda_n \to \infty$ as $n \to \infty$.

(B_4) $\beta_{n1}^*, \beta_{n2}^*, \ldots, \beta_{ns_n}^*$ satisfy $\min_{1\leq j\leq s_n} |\beta_{nj}^*|/(\sqrt{n}h_n) \to \infty$ as $n \to \infty$.

Conditions $(B_1)-(B_2)$ are similar to the conditions (F)$-$(G) placed on the information matrix in Fan and Peng (2004). Condition (B_3) is the condition of Fan and Peng (2004) used to obtain the oracle property. Condition (B_4) is used to obtain the oracle property when using the adaptive LASSO penalty.

(iii) *Regularity conditions on the error distribution.*

(C) The errors ε_i have the distribution function $G(\cdot)$ with density $g(\cdot)$. The density g is positive and continuous at the τ_k-th quantiles $b_{\tau_k}^*$.

The condition (C) acts in accord with the condition placed on the error distribution for single quantile regression (Koenker (2005)).

A.2. Proofs of Theorems

Following the arguments for Theorem 2, we can show Theorem 3. Theorems 5 and 6 can be proved using the arguments for Theorems 1 and 2. Hence, we only discuss the proofs of Theorems 1, 2, and 4. The argument for likelihood estimation in Fan and Peng (2004) is based on Taylor's expansion on the loss function. Since the loss function $\rho(\cdot)$ is not differentiable here, we use some arguments from quantile regression.

To facilitate the proofs, we write $\eta_{n,k} = n^{-1/2}\omega_k \sum_{i=1}^{n}[I(\varepsilon_i < b_{\tau_k}^*) - \tau_k]$, $\boldsymbol{\eta}_n = (\eta_{n,1}, \ldots, \eta_{n,K})'$, $\mathbf{z}_n = n^{-1/2}\sum_{i=1}^{n} \boldsymbol{\nabla} f_{ni}^* \sum_{k=1}^{K} \omega_k[I(\varepsilon_i < b_{\tau_k}^*) - \tau_k]$, $\mathbf{b}^* = (b_{\tau_1}^*, \ldots, b_{\tau_K}^*)'$, and

$$S_n(\mathbf{u}, \mathbf{v}) = L_n(\boldsymbol{\beta}_n^* + n^{-1/2}\mathbf{u}, \mathbf{b}^* + n^{-1/2}\mathbf{v}) - L_n(\boldsymbol{\beta}_n^*, \mathbf{b}^*).$$

Proof of Theorem 1. Let $\alpha_n = \sqrt{p_n}(n^{-1/2} + a_n)$, $\mathbf{u}_n = \alpha_n^{-1}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^*)$, $\mathbf{v} = \alpha_n^{-1}(\mathbf{b} - \mathbf{b}^*)$, and $\mathcal{C}_n = \{(\mathbf{u}_n, \mathbf{v}) : \|(\mathbf{u}_n', \mathbf{v}')'\| = C\}$, where $\|\cdot\|$ denotes the L_2-norm. We show that, for any $\delta > 0$, there is a large constant C such that, for large n,

$$P\{\inf_{(\mathbf{u}_n,\mathbf{v})\in\mathcal{C}_n} Q_n^{SC}(\boldsymbol{\beta}_n^* + \alpha_n\mathbf{u}_n, \mathbf{b}^* + \alpha_n\mathbf{v}) > Q_n^{SC}(\boldsymbol{\beta}_n^*, \mathbf{b}^*)\} \geq 1 - \delta. \tag{A.1}$$

This implies that, with probability tending to one, there is a local minimum $\hat{\boldsymbol{\beta}}_n$ in the ball $\{(\boldsymbol{\beta}_n^*+\alpha_n\mathbf{u}_n, \mathbf{b}^*+\alpha_n\mathbf{v}) : \|(\mathbf{u}_n', \mathbf{v}')'\| \le C\}$ such that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*\| = O_p(\alpha_n)$.

Let $D_n^{SC}(\mathbf{u}_n, \mathbf{v}) = Q_n^{SC}(\boldsymbol{\beta}_n^* + \alpha_n\mathbf{u}_n, \mathbf{b}^* + \alpha_n\mathbf{v}) - Q_n^{SC}(\boldsymbol{\beta}_n^*, \mathbf{b}^*)$. Then

$$D_n^{SC}(\mathbf{u}_n, \mathbf{v}) = S_n(\mathbf{u}_n, \mathbf{v}) + P_{\lambda_n}(\mathbf{u}_n), \tag{A.2}$$

where $P_{\lambda_n}(\mathbf{u}_n) = n\sum_{j=1}^{p_n}[p_{\lambda_n}(|\beta_{nj}^* + \alpha_n u_{nj}|) - p_{\lambda_n}(|\beta_{nj}^*|)]$. By the Mean Value Theorem, there exists a $\tilde{\boldsymbol{\beta}}_n$ between $\boldsymbol{\beta}_n^*$ and $\boldsymbol{\beta}_n^* + \alpha_n\mathbf{u}_n$, such that

$$f(\mathbf{x}_i, \boldsymbol{\beta}_n^* + \alpha_n\mathbf{u}_n) = f_{ni}^* + \alpha_n \boldsymbol{\nabla} f(\mathbf{x}_i, \tilde{\boldsymbol{\beta}}_n)'\mathbf{u}_n.$$

Let $s_{ik} = \alpha_n v_k + \alpha_n \boldsymbol{\nabla} f(\mathbf{x}_i, \tilde{\boldsymbol{\beta}}_n)'\mathbf{u}_n$, $B_n^{(k)} = \sum\limits_{i=1}^{n}\int_0^{s_{ik}}[I(\varepsilon_i \le b_{\tau_k}^* + x) - I(\varepsilon_i \le b_{\tau_k}^*)]dx$, $\tilde{\mathbf{z}}_n = n^{-1/2}\sum\limits_{k=1}^{K}\omega_k\sum\limits_{i=1}^{n}\nabla f(\mathbf{x}_i, \tilde{\boldsymbol{\beta}}_n)[I(\varepsilon_i < b_{\tau_k}^*) - \tau_k]$, and $\delta_n(\mathbf{u}_n) = \sqrt{n}\alpha_n\mathbf{u}_n'(\tilde{\mathbf{z}}_n - \mathbf{z}_n)$. By (B_1) and direct computation of the mean and variance for each component, it is easy to show that $||\tilde{\mathbf{z}}_n - \mathbf{z}_n|| = o_p(1)$. Then, by the Cauchy-Schwartz inequality,

$$|\delta_n(\mathbf{u}_n)| = o_p(\sqrt{n}\alpha_n)\|\mathbf{u}_n\|. \tag{A.3}$$

By the identity (Knight (1998)),

$$|r-s| - |r| = -s(I(r>0) - I(r<0)) + 2\int_0^s [I(r \le x) - I(r \le 0)]dx,$$

we have

$$\rho_\tau(r-s) - \rho_\tau(r) = s[I(r<0) - \tau] + \int_0^s [I(r \le x) - I(r \le 0)]dx. \tag{A.4}$$

Then we can rewrite $S_n(\mathbf{u}_n, \mathbf{v})$ as

$$S_n(\mathbf{u}_n, \mathbf{v}) = \sqrt{n}\alpha_n(\boldsymbol{\eta}_n'\mathbf{v} + \mathbf{z}_n'\mathbf{u}_n) + \sum_{k=1}^{K}\omega_k B_n^{(k)} + \delta_n(\mathbf{u}_n). \tag{A.5}$$

Put $\boldsymbol{\mu}_n = E(\boldsymbol{\nabla} f_{n1}^*)$ and $\boldsymbol{\Gamma}_n = E[(\boldsymbol{\nabla} f_{n1}^*)^{\otimes 2}]$. Note that, by (B_2), $\|\boldsymbol{\Gamma}_n\| = O(1)$. It follows that $E(\mathbf{z}_n'\mathbf{u}_n) = 0$ and $E\{(\mathbf{z}_n'\mathbf{u}_n)^2\} = \mathbf{u}_n'E(\mathbf{z}_n\mathbf{z}_n')\mathbf{u}_n = \boldsymbol{\omega}'\mathbf{A}\boldsymbol{\omega}$ $\mathbf{u}_n'\boldsymbol{\Gamma}_n\mathbf{u}_n = O(\|\mathbf{u}_n\|^2)$. Hence, $\mathbf{z}_n'\mathbf{u}_n = O_p(\|\mathbf{u}_n\|)$. This, combined with (A.3) and (A.5), leads to

$$S_n(\mathbf{u}_n, \mathbf{v}) = \sum_{k=1}^{K}\omega_k B_n^{(k)} + o_p(n\alpha_n^2)\|\mathbf{u}_n\|. \tag{A.6}$$

By (B_1) and (C) and computation of the expectation and variance of $B_n^{(k)}$, we obtain

$$B_n^{(k)} = \frac{1}{2} g(b_{\tau_k}^*) n\alpha_n^2 (v_k^2 + \mathbf{u}_n' \mathbf{\Gamma}_n \mathbf{u}_n + 2v_k \boldsymbol{\mu}_n' \mathbf{u}_n)(1 + o_p(1)).$$

This, combined with (A.6), yields that

$$\begin{aligned} S_n(\mathbf{u}_n, \mathbf{v}) = & \frac{1}{2} n\alpha_n^2 \sum_{k=1}^{K} \omega_k g(b_{\tau_k}^*)(v_k^2 + \mathbf{u}_n' \mathbf{\Gamma}_n \mathbf{u}_n + 2v_k \boldsymbol{\mu}_n' \mathbf{u}_n)(1 + o_p(1)) \\ & + o_p(n\alpha_n^2) \|\mathbf{u}_n\|. \end{aligned} \tag{A.7}$$

Using $p_{\lambda_n}(0) = 0$ and $(A_2)-(A_4)$, we establish that

$$\begin{aligned} P_{\lambda_n}(\mathbf{u}_n) \geq & \sum_{j=1}^{s_n} \left[n\alpha_n p_{\lambda_n}'(|\beta_{nj}^*|) \mathrm{sgn}(\beta_{nj}^*) u_{nj} + \frac{1}{2} n\alpha_n^2 p_{\lambda_n}''(|\beta_{nj}^*|) u_{nj}^2 (1 + o(1)) \right] \\ \geq & -(n\alpha_n^2 \|\mathbf{u}_n\| + o_p(n\alpha_n^2)). \end{aligned} \tag{A.8}$$

It follows from (A.7)−(A.8) that $D_n^{SC}(\mathbf{u}_n, \mathbf{v})$ in (A.2) is dominated by the positive quadratic term $(1/2)n\alpha_n^2 \sum_{k=1}^{K} \omega_k g(b_\tau^*)(v_k^2 + \mathbf{u}_n' \mathbf{\Gamma}_n \mathbf{u}_n + 2v_k \boldsymbol{\mu}_n' \mathbf{u}_n)$, as long as $\|\mathbf{u}_n\|$ and $\|\mathbf{v}\|$ are large enough. Therefore, (A.1) holds and proof is complete.

Lemma A.1. Under $(A_1)-(A_4)$, $(B_1)-(B_3)$, and (C), if $\lambda_n \to 0$, $\sqrt{n_p}\lambda_n \to \infty$, and $p_n^3/n \to 0$ as $n \to \infty$, then with probability tending to 1, for any given $\boldsymbol{\beta}_{n1}$ satisfying $\|\boldsymbol{\beta}_{n1} - \boldsymbol{\beta}_{n1}^*\| = O_p(n_p^{-1/2})$, $\|\mathbf{b} - \mathbf{b}^*\| = O_p(n_p^{-1/2})$ and any constant C, we have

$$Q_n^{SC}((\boldsymbol{\beta}_{n1}', \mathbf{0}')', \mathbf{b}) = \min_{\|\boldsymbol{\beta}_{n2}\| \leq C n_p^{-1/2}} Q_n^{SC}((\boldsymbol{\beta}_{n1}', \boldsymbol{\beta}_{n2}')', \mathbf{b}).$$

Proof. Let $\alpha_n^{-1}(\boldsymbol{\beta}_{n1} - \boldsymbol{\beta}_{n1}^*) = \mathbf{u}_{n1}$, $\alpha_n^{-1}(\boldsymbol{\beta}_{n2} - \boldsymbol{\beta}_{n2}^*) = \mathbf{u}_{n2}$, and $\mathbf{u}_n = (\mathbf{u}_{n1}', \mathbf{u}_{n2}')'$. By the definition of $Q_n^{SC}(\boldsymbol{\beta}_n, \mathbf{b})$, we have

$$\begin{aligned} & Q_n^{SC}((\boldsymbol{\beta}_{n1}', \mathbf{0}')', \mathbf{b}) - Q_n^{SC}((\boldsymbol{\beta}_{n1}', \boldsymbol{\beta}_{n2}')', \mathbf{b}) \\ & \quad = S_n((\mathbf{u}_{n1}', \mathbf{0}'))', \mathbf{v}) - S_n((\mathbf{u}_{n1}', \mathbf{u}_{n2}')', \mathbf{v}) - n \sum_{j=s_n+1}^{p_n} p_{\lambda_n}(|\beta_{nj}|). \end{aligned}$$

From (A.7), we obtain that

$$\begin{aligned} S_n((\mathbf{u}_{n1}', \mathbf{u}_{n2}')', \mathbf{v}) = & \frac{1}{2} n\alpha_n^2 \sum_{k=1}^{K} \omega_k g(b_{\tau_k}^*)(v_k^2 + \mathbf{u}_n' \mathbf{\Gamma}_n \mathbf{u}_n + 2v_k \boldsymbol{\mu}_n' \mathbf{u}_n)(1 + o_p(1)) \\ & + o_p(n\alpha_n^2) \|\mathbf{u}_n\|. \end{aligned}$$

Since $\|\mathbf{u}_n\| = O_p(1)$ and $\mathbf{G}_n = \mathbf{\Gamma}_n - \boldsymbol{\mu}_n\boldsymbol{\mu}_n'$ is positive, by (B_2) we have $\mathbf{u}_n'\mathbf{\Gamma}_n\mathbf{u}_n \leq \|\mathbf{\Gamma}_n\|\|\mathbf{u}_n\|^2 = O_p(1)$ and $\|\boldsymbol{\mu}_n\|^2 = \mathrm{tr}(\boldsymbol{\mu}_n\boldsymbol{\mu}_n') \leq \mathrm{tr}(\mathbf{\Gamma}_n) = O_p(p_n)$. Hence, $\|\boldsymbol{\mu}_n\| = O_p(\sqrt{p_n})$. It follows that

$$S_n((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v}) = O_p(n\alpha_n^2\sqrt{p_n}) = O_p(p_n^{3/2}).$$

Similarly, $S_n((\mathbf{u}_{n1}', \mathbf{u}_{n2}')', \mathbf{v}) = O_p(p_n^{3/2})$. Using $p_{\lambda_n}(0) = 0$ and the Mean Value Theorem, we arrive at

$$\begin{aligned} n\sum_{j=s_n+1}^{p_n} p_{\lambda_n}(|\beta_{nj}|) &= n\sum_{j=s_n+1}^{p_n} p_{\lambda_n}'(|\beta_{nj}^{\dagger}|)|\beta_{nj}^{\dagger}| \\ &\geq p_n^2\sqrt{\frac{n}{p_n^3}}\sqrt{n_p}\lambda_n\Big(\liminf_{n\to+\infty}\liminf_{\theta\to 0+}\frac{p_{\lambda_n}'(\theta)}{\lambda_n}\Big)\sum_{j=s_n+1}^{p_n}|\beta_{nj}^{\dagger}|, \end{aligned}$$

where $0 < \beta_{nj}^{\dagger} < |\beta_j|$ $(j = s_n + 1, \ldots, p_n)$. Since $\sqrt{n_p}\lambda_n \to \infty$ and $p_n^3/n \to 0$, $p_n^2\sqrt{n/p_n^3}\sqrt{n_p}\lambda_n$ is of higher order than $O_p(p_n^{3/2})$. By (A_1), it follows that $Q_n^{SC}((\boldsymbol{\beta}_{n1}', \mathbf{0}')', \mathbf{b}) - Q_n^{SC}((\boldsymbol{\beta}_{n1}', \boldsymbol{\beta}_{n2}')', \mathbf{b})$ is dominated by the negative term $-n\sum_{j=s_n+1}^{p_n} p_{\lambda_n}(|\beta_{nj}|)$ for larger n. Hence, the lemma holds.

Proof of Theorem 2. (i) follows from Lemma A.1.
(ii) Let $\mathbf{u}_n = \alpha_n^{-1}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^*)$. Partition the vectors $\mathbf{u}_n = (\mathbf{u}_{n1}', \mathbf{u}_{n2}')'$ and $\boldsymbol{\nabla} f_{ni}^* = ((\boldsymbol{\nabla} f_{ni1}^*)', (\boldsymbol{\nabla} f_{ni2}^*)')'$ in the same way as $\boldsymbol{\beta}_n = (\boldsymbol{\beta}_{n1}', \boldsymbol{\beta}_{n2}')'$; partition $\mathbf{G}_n$ as a 2×2 block matrix $\mathbf{G}_n = (\mathbf{G}_{nij})$ (for $i, j = 1, 2$). By (A.2) and $P_{\lambda_n}(0) = 0$, we can write

$$D_n^{SC}((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v}) = S_n((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v}) + P_{\lambda_n}(\mathbf{u}_{n1}),$$

where $P_{\lambda_n}(\mathbf{u}_{n1}) = n\sum_{j=1}^{s_n}(p_{\lambda_n}(|\beta_{nj}^* + \alpha_n u_{nj}|) - p_{\lambda_n}(|\beta_{nj}^*|))$. By (A_4) and (B_3), and by taking Taylor's expansion for $P_{\lambda_n}(\mathbf{u}_{n1})$ at $\mathbf{u}_{n1} = 0$, we obtain that

$$P_{\lambda_n}(\mathbf{u}_{n1}) = n\alpha_n\mathbf{c}_n'\mathbf{u}_{n1} + \frac{1}{2}n\alpha_n^2\mathbf{u}_{n1}'\mathbf{\Sigma}_{\lambda_n}\mathbf{u}_{n1}(1 + o(1)).$$

Let $t_{ik}(\mathbf{u}_{n1}, \mathbf{u}_{n2}, v_k) = \alpha_n v_k + f(\mathbf{x}_i, \boldsymbol{\beta}_n^* + \alpha_n\mathbf{u}_n) - f(\mathbf{x}_i, \boldsymbol{\beta}_n^*)$. Then the minimizer $(\hat{\mathbf{u}}_{n1}, \hat{\mathbf{v}})$ of $D_n^{SC}((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v})$ satisfies the score equations

$$\begin{aligned} &n^{-1}\sum_{k=1}^{K}\omega_k\sum_{i=1}^{n}\psi_{\tau_k}(\varepsilon_i - b_{\tau_k}^* - t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k))\boldsymbol{\nabla} f_{ni1}^*(1 + o_p(1)) \\ &= \mathbf{c}_n + \alpha_n\mathbf{\Sigma}_{\lambda_n}\hat{\mathbf{u}}_{n1}(1 + o_p(1)), \end{aligned} \tag{A.9}$$

$$\omega_k\sum_{i=1}^{n}\psi_{\tau_k}(\varepsilon_i - b_{\tau_k}^* - t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k)) = 0. \tag{A.10}$$

Since $\psi_\tau(u) = \tau - I(u < 0)$, we can write

$$\begin{aligned} n^{-1}\sum_{k=1}^{K}\omega_k\sum_{i=1}^{n}\psi_{\tau_k}(\varepsilon_i - b^*_{\tau_k} - t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k))\boldsymbol{\nabla} f^*_{ni1} \\ = -n^{-1/2}\mathbf{z}_{n1} + \sum_{k=1}^{K}\omega_k(B^{(k)}_{n21} + B^{(k)}_{n22}), \end{aligned} \tag{A.11}$$

where $\mathbf{z}_{n1} = n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\nabla} f^*_{ni1}\sum_{k=1}^{K}\omega_k[I(\varepsilon_i < b^*_{\tau_k}) - \tau_k]$,

$$B^{(k)}_{n21} = n^{-1}\sum_{i=1}^{n}[G(b^*_{\tau_k}) - G(b^*_{\tau_k} + t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k))]\boldsymbol{\nabla} f^*_{ni1},$$

$$\begin{aligned} B^{(k)}_{n22} = n^{-1}\sum_{i=1}^{n}\big\{[I(\varepsilon_i < b^*_{\tau_k}) - I(\varepsilon_i < b^*_{\tau_k} + t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k))] \\ -[G(b^*_{\tau_k}) - G(b^*_{\tau_k} + t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k))]\big\}\boldsymbol{\nabla} f^*_{ni1}. \end{aligned}$$

Taking Taylor's explanation for $G(b^*_{\tau_k} + t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k))$ at $b^*_{\tau_k}$ gives

$$\begin{aligned} B^{(k)}_{n21} &= -n^{-1}\sum_{i=1}^{n} g(b^*_{\tau_k})t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k)\boldsymbol{\nabla} f^*_{ni1}(1 + o(1)) \\ &= -\alpha_n g(b^*_{\tau_k})(\boldsymbol{\Gamma}_{n11}\hat{\mathbf{u}}_{n1} + \boldsymbol{\mu}_{n1}\hat{v}_k)(1 + o_p(1)). \end{aligned} \tag{A.12}$$

By direct calculation of the mean and variance, we can show, as in Jiang, Zhao, and Hui (2001), that $B^{(k)}_{n22} = o_p(\alpha_n)$. This combined with (A.9), (A.11), and (A.12) leads to

$$-(n^{-1/2}\mathbf{z}_{n1} + \mathbf{c}_n) = \alpha_n\{\sum_{k=1}^{K}\omega_k g(b^*_{\tau_k})(\boldsymbol{\Gamma}_{n11}\hat{\mathbf{u}}_{n1} + \boldsymbol{\mu}_{n1}\hat{v}_k) + \boldsymbol{\Sigma}_{\lambda_n}\hat{\mathbf{u}}_{n1}\}(1 + o_p(1)). \tag{A.13}$$

Similarly, (A.10) can be simplified as

$$n^{-1/2}\eta_{n,k} + \alpha_n\omega_k g(b^*_{\tau_k})(\hat{v}_k + \boldsymbol{\mu}'_{n1}\hat{\mathbf{u}}_{n1}(1 + o_p(1))) = 0. \tag{A.14}$$

Solving (A.13) and (A.14), we obtain that

$$\alpha_n\big(\mathbf{G}_{n11} + \frac{\boldsymbol{\Sigma}_{\lambda_n}}{\boldsymbol{\omega}'\mathbf{g}}\big)\hat{\mathbf{u}}_{n1} + \frac{\mathbf{c}_n}{\boldsymbol{\omega}'\mathbf{g}} = -n^{-1/2}(\mathbf{z}_{n1} - \boldsymbol{\mu}_{n1}\sum_{k=1}^{K}\frac{\eta_{n,k})}{\boldsymbol{\omega}'\mathbf{g}} + o_p(n^{-1/2}),$$

where $\mathbf{e}_n'\mathbf{G}_{n11}^{-1/2}(\mathbf{z}_{n1} - \boldsymbol{\mu}_{n1}\sum_{k=1}^{K}\eta_{n,k})/\boldsymbol{\omega}'\mathbf{g} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega}))$. Note that $\hat{\mathbf{u}}_{n1} = \alpha_n^{-1}(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n1}^*)$. It follows that

$$\sqrt{n}\mathbf{e}_n'\mathbf{G}_{n11}^{-1/2}\big(\mathbf{G}_{n11} + \frac{\boldsymbol{\Sigma}_{\lambda_n}}{\boldsymbol{\omega}'\mathbf{g}}\big) \times [(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n1}^*) + \big(\mathbf{G}_{n11} + \frac{\boldsymbol{\Sigma}_{\lambda_n}}{\boldsymbol{\omega}'\mathbf{g}}\big)^{-1}\frac{\mathbf{c}_n}{\boldsymbol{\omega}'\mathbf{g}}] \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})).$$

Proof of Theorem 4. By the definition of $\boldsymbol{\Omega}$, we have $\boldsymbol{\Omega} = (K+1)^{-2}\mathbf{A}$, where the (i,j)th entry of $\mathbf{A}$ is $A_{ij} = \min(i,j)(K+1-\max(i,j))$. Let $\mathbf{C}$ be a $K \times K$ matrix with diagonal elements all $2/(K+1)^2$, superdiagonal and subdiagonal elements all $-1/(K+1)^2$, and others all zero. Then $\mathbf{AC} = (K+1)\mathbf{I}_{K\times K}$ and $\boldsymbol{\Omega}^{-1} = (K+1)\mathbf{C}$. Note that $\tau_i = \frac{i}{K+1}$ for $i = 1, \ldots, K$. Then

$$\begin{aligned}\mathbf{g}'\boldsymbol{\Omega}^{-1}\mathbf{g} &= 2(K+1)[\sum_{i=1}^{K} g^2(b_i) - \sum_{i=1}^{K} g(b_i)g(b_{i+1})]^2 \\ &= (K+1)\sum_{i=0}^{K}[g(b_i) - g(b_{i+1})]^2 = (K+1)\sum_{i=0}^{K}[g(G^{-1}(\tau_i)) - g(G^{-1}(\tau_{i+1}))]^2.\end{aligned}$$

Using Taylor's expansion, we obtain that

$$\begin{aligned}\mathbf{g}'\boldsymbol{\Omega}^{-1}\mathbf{g} &= (K+1)\sum_{i=0}^{K}\{(\tau_{i+1} - \tau_i)\frac{g'(G^{-1}(\tau_i))}{g(G^{-1}(\tau_i))} + o(\tau_{i+1} - \tau_i)\}^2. \\ &= \frac{1}{K+1}\sum_{i=0}^{K}\{\frac{g'(G^{-1}(\tau_i))}{g(G^{-1}(\tau_i))} + o(1)\}^2 \\ &= \int_0^1 \Big\{\frac{g'(G^{-1}(t)}{g(G^{-1}(t)}\Big\}^2\, dt + o(1) = \int_{-\infty}^{+\infty}\frac{(g'(t))^2}{g(t)}\, dt + o(1)\end{aligned}$$

as $K \to \infty$. Therefore, $\mathbf{g}'\boldsymbol{\Omega}^{-1}\mathbf{g} = I_g$, where $I_g = \int_{-\infty}^{+\infty}(g'(t))^2/g(t)\, dt$ is the Fisher information. It follows that $e(WCQR, OML) = I_g^{-1}\mathbf{g}'\boldsymbol{\Omega}\mathbf{g} \to 1$ as $K \to \infty$.

References

Bassett, G. W. and Koenker, R. W. (1992). A note on recent proposals for computing L_1 estimates. *Computational Statistics & Data Analysis* **14**, 207-211.

Belloni, A. and Chernozhukov, V. (2011). L_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39**, 82-130.

Breiman, L. (1995). Better subset regression using the non-negative garotte. *Technometrics* **37**, 373-384.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. Lecture on August 8, 2000, to the American Mathematical Society on Math Challenges of the 21st Century. Available at `http://www.inma.ucl.ac.be/~francois/these/papers/entry-Donoho-2000.html`.

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. (2006). Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. *Proceedings of the International Congress of Mathematicians* (Edited by M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera), Vol. III, 595-622. European Mathematical Society, Zurich.

Fan, J. and Lv, J. (2010). Properties of non-concave penalized likelihood with NP- dimensionality. Manuscript.

Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Ann. Statist.* **32**, 928-961.

Fu, W. J. (1998). Penalized regression: the bridge versus the LASSO. *J. Comput. Graph. Statist.* **7**, 397-416.

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

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. *Ann. Statist.* **1**, 799-821.

Huber, P. J. (1988). Robust regression: asymptotics, conjectures and monte carlo. *Ann. Statist.* **1**, 799-821.

Jennrich, R. (1969). Asymptotic properties of nonlinear least squares estimators. *Ann. Math. Statist.* **40**, 633-643.

Jiang, J., Zhao, Q. and Hui, Y. V. (2001). Robust modlling of ARCH models. *J. Forecasting* **20**, 111-133.

Knight, K. (1998). Limiting distributions for l_1 regression estimators under general conditions. *Ann. Statist.* **26**, 755-770.

Knight, K. and Fu, W. (2000). Asymptotics for LASSO-type estimators. *Ann. Statist.* **28**, 1356-1378.

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

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

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

Koenker, R. and Park, B. J. (1996). An interior point algorithm for nonlinear quantile regression. *J. Econometrics* **71**, 265-283.

Kutner, M. H. et, al. (2005). *Applied Linear Statistical Models*, 5th ed. McGraw-Hill/Irwin.

Lam, C. and Fan, J. (2008). Profile-Kernel likelihood inference with diverging number of parameters. *Ann. Statist.* **36**, 2232-2260.

Leng, C. (2009). Variable selection and coefficient estimation via regularized rank regression. *Statist. Sinica*, to appear.

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. (2008). The l_1 norm quantile regression. *J. Comput. Graph. Statist.* **17**, 163-185.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498-3528.

Nychka, D., Gray, G., Haaland, P., Martin, D. and O'Connell, M. (1995). A nonparametric regression approach to syringe grading for quality improvement. *J. Amer. Statist. Assoc.* **90**, 1171-1178.

Osborne, M. R. and Watson, G. A. (1971). On an algorithm for discrete nonlinear L_1 approximation. *Computer J.* **14**, 184-188.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**, 356-366.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135-1151.

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

Vanderbei, R. J., Meketon, M. S. and Freedman, B. A. (1986). A modification of Karmarkar's linear programming algorithm. *Algorithmica* **1**, 395-407.

Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. Bus. Econom. Statist.* **25**, 347-355.

Wang, H. Li, R. and Tsai, C-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

Wu, C. J. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.* **9**, 501-513.

Yuan, M. (2006). GACV for quantile smoothing splines. *Comput. Statist. Data Anal.* **5**, 813-829.

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

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894-942.

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

Zou, H. and Yuan, M. (2008a). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **3**, 1108-1126.

Zou, H. and Yuan, M. (2008b). Regularized simultaneous model selection in multiple quantiles regression. *Comput. Statist. Data Anal.* **52**, 5296-5304.

School of Statistics and Mathematics at Zhongnan University of Economics and Law, 430073, Wuhan, China.

E-mail: xjjiang@znufe.edu.cn

Department of Mathematics and Statistics, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223-0001, USA.

E-mail: jjiang1@uncc.edu

Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong.

E-mail: xysong@sta.cuhk.edu.hk

(Received September 2010; accepted September 2011)