Communications of the Korean Statistical Society
2009, Vol. 16, No. 1, 209–215

# Quantile Regression with Non-Convex Penalty on High-Dimensions

Hosik Choi[1,a], Yongdai Kim[b], Sang-Tae Han[c], Hyuncheol Kang[c]

[a]Dept. of Informational Statistics and Institute of Basic Science, Hoseo Univ.,
[b]Dept. of Statistics, Seoul National Univ., [c]Dept. of Informational Statistics, Hoseo Univ.

**Abstract**

In regression problem, the SCAD estimator proposed by Fan and Li (2001), has many desirable property such as continuity, sparsity and unbiasedness. In this paper, we extend SCAD penalized regression framework to quantile regression and hence, we propose new SCAD penalized quantile estimator on high-dimensions and also present an efficient algorithm. From the simulation and real data set, the proposed estimator performs better than quantile regression estimator with $L_1$ norm.

Keywords: Quantile regression, SCAD penalty, high-dimensions.

## 1. Introduction

Suppose that $\{(\mathbf{x}_i, y_i)_{i=1}^n : \mathbf{x} \in \mathbb{R}^p,\ y_i \in \mathbb{R}\}$ is a set of independent observation from a distribution. Let $F(y|\mathbf{x})$ be the conditional distribution function of $y$ given $\mathbf{x}$. Then for any $\tau \in (0, 1)$, the $\tau^{th}$ quantile of $y$ given $\mathbf{x}$ is defined as follows:

$$f(\mathbf{x}; \tau) = F^{-1}(\tau|\mathbf{x}) = \inf\{y : F(y|\mathbf{x}) \geq \tau\}. \tag{1.1}$$

Then the objective of quantile regression is to estimate the $\tau^{th}$ conditional quantile defined by (1.1). Although there are many several approaches, the semiparametric quantile regression using the check function, $\rho_\tau(z) = \tau z I_{[0,\infty)}(z) - (1-\tau) z I_{(-\infty,0)}(z)$ introduced by Konecker and Bassett (1978) has been successful in many applications due to its high prediction accuracy and flexibility.

Under linear model, $f(\mathbf{x}; \tau) = \beta_0(\tau) + \mathbf{x}'\beta(\tau)$, the quantile regression model adopting the $L_1$ penalty on coefficients is introduced by Li and Zhu (2008). They developed the efficient computation algorithm which is based on the LARs algorithm (Efron *et al.*, 2004) minimizing the following regularized cost functional

$$\sum_{i=1}^{n} \rho_\tau(y_i - f(\mathbf{x}_i; \tau)) + \lambda \sum_{j=1}^{p} |\beta_j(\tau)|, \tag{1.2}$$

where some positive $\lambda$ is regularization parameter which controls the balance between the fidelity and complexity of $f$. The $L_1$-norm penalty shrinks to fitted coefficients toward zero and benefits from the reduction of the variance of fitted coefficients. For simplicity of notation, given $\tau$, we use $f(\mathbf{x})$, $\beta_0$

This research was supported by the Academic Research fund of Hoseo University in 2008 (20080093)

[1] Corresponding author: Full Time Lecturer, Department of Informational Statistics and Institute of Basic Science, Hoseo University, Asan Campus, San 165, Sechul-ri, Baebang-myun, Asan, Chungnam, 336-795, Korea.
E-mail: choi.hosik@gmail.com.

and $\beta$ instead of $f(\mathbf{x};\tau)$, $\beta_0(\tau)$ and $\beta(\tau)$ respectively and also, denote the minimizer of (1.2) by L1QR ($L_1$-norm penalized quantile regression).

Though the method using the $L_1$ penalty give a sparse solutions, the estimates can be biased for large coefficients since larger penalties are imposed on larger coefficients. Related with this issue, in linear regression models, Fan and Li (2001) explained that there are three desirable properties of penalized estimators - unbiasedness, sparsity and continuity. They showed that the smoothly clipped absolute deviation (SCAD) penalty satisfies all the three properties. However, since SCAD penalty function is nonconvex, more special algorithm is needed.

Recently, Kim *et al.* (2008) (referred to as the HSCAD paper henceforth) propose the high-dimensional SCAD regression estimator and develop a new stage-wise building algorithm. Motivated by the work of HSCAD, in this paper, we are to develop an efficient algorithm for SCAD penalized quantile regression estimator on high-dimensions.

When we find the appropriate $\lambda$ on high-dimensions, cross-validation demands much computational burden. If we know solution paths for the whole $\lambda$, we can find the optimal solution without much efforts. The algorithms which provides such works are called the entire solution path algorithm or path-following algorithm. The entire solution path according to the regularization parameter $\lambda$ enables us to achieve the set of solutions for all values of the regularization parameter and hence, we can reduce a computational burden which can be consumed in selecting an appropriate regularization parameters. These type algorithms depend on the "kinks" in the path. Note that the entire solution path of (1.2) can be constructed by linear interpolating $\beta^1, \beta^2, \ldots, \beta^L$ where $L$ denotes the number of steps that the algorithm takes. However, in quantile regression using SCAD penalty, it is difficult to find the entire solution path because of the nonconvexity of SCAD penalty. In spite of the defect, we can find the approximated entire solution path of the method using SCAD penalty.

The rest of the paper is organized as follows. In the next Section, the problem of the quantile regression with SCAD penalty is explicitly defined and an efficient path following optimization procedure is addressed. Numerical results including a detailed simulations and real data are presented in Section 3. Finally, we end with a concluding remark in Section 4.

## 2. Method

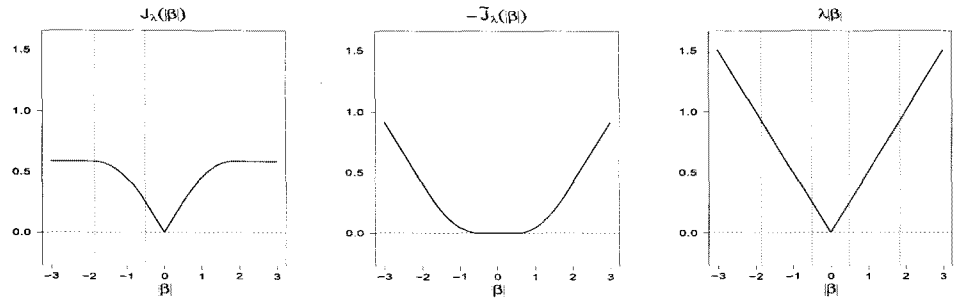
### 2.1. The proposed method: HscadQR

For the check function loss $\rho_\tau(\cdot)$, quantile regression estimator with the SCAD penalty is defined by the solution minimizing

$$\sum_{i=1}^{n} \rho_\tau(y_i - f(\mathbf{x}_i)) + \sum_{j=1}^{p} J_\lambda(|\beta_j|) \tag{2.1}$$

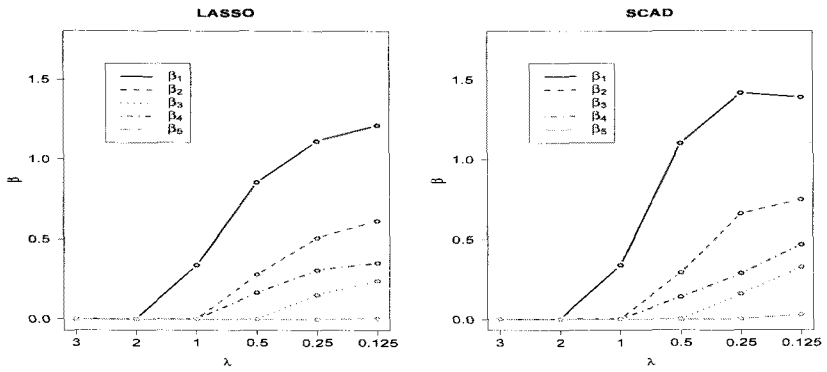
where $J_\lambda(\beta)$ is the SCAD penalty given as

$$J_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & \text{if } |\beta| < \lambda, \\ -\dfrac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)}, & \text{if } \lambda \le |\beta| \le a\lambda, \\ \dfrac{(a+1)\lambda^2}{2}, & \text{if } |\beta| > a\lambda. \end{cases}$$




(a) $J_\lambda(|\beta|)$ (b) $-\tilde{J}_\lambda(|\beta|)$ (c) $\lambda|\beta|$

Figure 1: *SCAD penalty when* $a = 3.7, \lambda = 0.5$




(a) The solution path of L1QR (b) The solution path of HscadQR

Figure 2: *When* $\tau = 0.5$, *the solution paths of the L1QR and HscadQR estimates of the first 5 quantile regression coefficients from the simulated model (3.1) with* $r = 0$

We can rewrite (2.1) as

$$\sum_{i=1}^{n} \rho_\tau(y_i - f(\mathbf{x}_i)) + \sum_{j=1}^{p} \tilde{J}_\lambda(|\beta_j|) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.2}$$

where $\tilde{J}_\lambda(\beta) = J_\lambda(\beta) - \lambda\beta$. Denote the SCAD penalized quantile regression estimator by *HscadQR*. The decomposition about $J_\lambda(\beta)$ is founded by Kim *et al.* (2008). The functional forms of $\tilde{J}_\lambda(|\beta|)$ as well as $J_\lambda(|\beta|)$ are depicted in Figure 1. In (2.2), the only difference comparative to (1.2) is the second term $\sum_{j=1}^{p} \tilde{J}_\lambda(|\beta_j|)$. From Figure 1, we can see that the term $\tilde{J}_\lambda(|\beta_j|)$ plays an important role of making the minimizer with SCAD penalty have a sparser solution than that with $L_1$ penalty.

In fact, the term $\tilde{J}_\lambda(|\beta_j|)$ encourages larger coefficients more to be included in the solution. Such phenomena are identified from results of simple simulated example. Figure 2 draws the solution paths of the L1QR and HscadQR estimates of the first 5 regression coefficients from the simulated model (3.1) with $r = 0$ for various values of $\lambda$s. The coefficient values of the SCAD estimates are larger than those of the L1QR estimates, which implies that the HscadQR estimates are less biased than the L1QR estimates.

## 2.2. Computation

In this section, we present an efficient optimization algorithm for the SCAD penalized quantile regression estimator, which is a hybrid algorithm of the convex concave procedure (CCCP; Yuille and Rangarajan, 2003) and $L_1$ quantile regression algorithm (Li and Zhu, 2008). The CCCP is an optimization method for non-convex problems when a given non-convex objective function can be decomposed by the sum of convex and concave functions. The idea of the CCCP is to update the solution by minimizing the convex function which is the tight upper bound of the objective function at the current solution. To explain more, let $Q_{\text{vex}}(\beta)$ be the convex part and $Q_{\text{cave}}(\beta)$ be the concave part of $Q(\beta)$. That is, $Q(\beta) = Q_{\text{vex}}(\beta) + Q_{\text{cave}}(\beta)$. Let $\nabla Q_{\text{cave}}(\beta)$ be the gradient of $Q_{\text{cave}}(\beta)$ with respect to $\beta$. Denote $<,>$ by the inner product. For a given current solution $\beta^{(l)}$, the tight upper convex bound $Q^{(l)}(\beta)$ is defined by $Q^{(l)}(\beta) = Q_{\text{vex}}(\beta) + < \nabla Q_{\text{cave}}(\beta^{(l)}), \beta >$ . Then, the CCCP updates the solution as the minimizer of $Q^{(l)}(\beta)$. One important property of the CCCP is that the cost functional $Q(\beta)$ decreases monotonically (See Theorem 2 in Yuille and Rangarajan, 2003), and hence the solutions of the CCCP always converge to a local minimum.

Now, to compute the SCAD penalized quantile regression estimator, let $Q(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - f(\mathbf{x}_i)) + \sum_{j=1}^{p} \tilde{J}_\lambda(|\beta_j|) + \lambda \sum_{j=1}^{p} |\beta_j|$. Then, $Q(\beta)$ can be decomposed by the sum of $Q_{\text{vex}}(\beta)$ and $Q_{\text{cave}}(\beta)$ where $Q_{\text{vex}}(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - f(\mathbf{x}_i)) + \lambda \sum_{j=1}^{p} |\beta_j|$ and $Q_{\text{cave}}(\beta) = \sum_{j=1}^{p} \tilde{J}_\lambda(|\beta_j|)$. Therefore, the CCCP's $l^{th}$ subproblem is equivalent to following problem

$$Q^{(l)}(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - (\beta_0 + \mathbf{x}_i'\beta)) + \sum_{j=1}^{p} \nabla \tilde{J}_\lambda(|\beta_j^{(l)}|)'\beta_j$$

for given the previous solution $\beta^{(l)}$. To sum up, we note the algorithm about the SCAD penalized quantile regression estimator.

---

**Algorithm for HscadQR**

Step 1 Initialization: $l = 1$ and let $\beta^{(l)} = 0$.

Step 2 Do until convergence:

(a) $\arg\min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} Q^{(l)}(\beta)$ given $\beta^{(l)}$ using the algorithm of Li and Zhu (2008).

(b) $\beta^c \leftarrow \beta$ and $l \leftarrow l + 1$.

---

Note that, 1) if we use zero vector as the initial solution, the solution of Step 2 equals to the solution of L1QR. In this paper, we use the L1QR's solution as the initial solution to save the computational complexity. 2) HscadQR algorithm using the hybrid of CCCP and path following algorithm can gives the chance to alter zero coefficients of L1QR to nonzero.

## 2.3. Tuning parameter selection

For any regularization methods, an important issue is to find a good choice of the regularization parameter such that the corresponding model is optimal according to some criterions. Two commonly used criteria for quantile regression are the Schwarz information criterion (SIC, Schwarz, 1978; Koenker and Portnoy, 1994) and the generalized approximate cross-validation criterion. Especially,

Yuan (2006) proposed to use generalized approximate cross-validation criterion (GACV) to select the tuning parameter defined by

$$\text{GACV}(\lambda) = \frac{\sum_{i=1}^{n} \rho_\tau \left(y_i - \hat{f}(\mathbf{x}_i)\right)}{n - \text{df}}, \tag{2.3}$$

where df is the degree of freedom. In this paper, we select the optimal $\lambda$ among the solutions from entire solution paths from GACV.

## 3. Numerical Studies

In this section, we investigate the finite sample performance of the HscadQR estimator via simulation experiments as well as real data analysis. In particular, we compare the HscadQR estimator with the L1QR estimator in terms of selectivity and prediction accuracy.

### 3.1. Simulation: Prediction accuracy and selectivity

We consider the sparse model. The simulation model is

$$y = \sum_{k=1}^{p} \beta_k x_k + \sigma\epsilon, \tag{3.1}$$

where $\mathbf{x} = (x_1, \ldots, x_p)'$ is a multivariate Gaussian random vector with mean 0 and covariances of $x_k$ and $x_l$ being $r^{|k-l|}$ for some $r \in [0, 1)$. The $\epsilon$ is a Gaussian random variable with mean 0 and variance 1. We fix the sample size 50 and the number of covariates 500 such that $p > n$. All the $\beta$s except the some $\beta$s are set to zero. First scenario considers the moderate sparse setting, $\beta = (3, -1.5, 0, 0, 2, 0, 0.5, -0.5, 0, 0.2, 0.2, 0, -0.1, -0.1, -0.1, \mathbf{0}_{485})$. Second scenario considers the very sparse setting with only large signal coefficients $\beta = (3, -1.5, 0, 0, 2, \mathbf{0}_{495})$. We investigate the performances with $\sigma = (1, \sqrt{3})$ and $r = (0.0, 0.5)$, respectively.

The results about prediction accuracy are the averages based on 20 repetitions of the simulation. The regularization parameter $\lambda$ is selected by the GACV. The mean absolute deviation(MAD) are measured on independent test samples of size 1,000 defined by

$$\text{Test MAD} = \frac{1}{1{,}000} \sum_{i=1}^{1{,}000} \left| f^\tau(\mathbf{x}_i) - \hat{f}^\tau(\mathbf{x}_i) \right|$$

for the test sample, where the fitted quantile function is $\hat{f}^\tau(\mathbf{x})$ and the true quantile function is $f^\tau(\mathbf{x})$.

In Table 1, "Test MAD" denotes the mean of MAD. "inczeros" is average number of incorrect 0 coefficients, "cnzeros" is average number of correct nonzeros coefficients in the models chosen in 20 repetitions.

Based on replicated experiments according to specific conditions from the results of first scenario as shown in Table 1, we observe HscadQR estimator tends to be better in the MAD than the L1QR estimator when the correlation between predictive variables is large or $\tau$ is large, but those are similar when the correlation is small. Also, HscadQR has more power in selecting nonzero coefficients in the correlated setting($r = 0.5$). Additionally, similar conclusion is observed in results of the low quantile $\tau = 0.1$ case.

Table 1: Simulation results of first scenario when $\tau = 0.5$ and 0.9 the prediction accuracy of L1QR and HscadQR: mean absolute deviation (standard errors)

| $\sigma^2$ | $\tau$ | $r$ | Method | Test MAD | inczeros | cnzeros |
|---|---|---|---|---|---|---|
| 1 | 0.5 | 0.0 | L1QR | 1.058 (.051) | 31.45 (1.672) | 7.20 (.247) |
| | | | HscadQR | 1.129 (.065) | 32.00 (1.930) | 6.95 (.294) |
| | | 0.5 | L1QR | 1.140 (.058) | 31.40 (1.024) | 6.95 (.285) |
| | | | HscadQR | 1.268 (.066) | 34.95 (1.157) | 6.90 (.228) |
| | 0.9 | 0.0 | L1QR | 1.161 (.044) | 31.50 (1.009) | 6.95 (.303) |
| | | | HscadQR | 1.151 (.071) | 31.05 (1.050) | 7.25 (.260) |
| | | 0.5 | L1QR | 1.263 (.058) | 28.40 (1.079) | 6.35 (.335) |
| | | | HscadQR | 1.135 (.047) | 27.85 (1.286) | 6.55 (.344) |
| 3 | 0.5 | 0.0 | L1QR | 1.518 (.081) | 23.45 (1.904) | 5.80 (.304) |
| | | | HscadQR | 1.519 (.101) | 22.30 (2.096) | 5.80 (.287) |
| | | 0.5 | L1QR | 1.590 (.091) | 21.85 (1.600) | 5.40 (.311) |
| | | | HscadQR | 1.562 (.097) | 21.40 (2.177) | 5.50 (.356) |
| | 0.9 | 0.0 | L1QR | 1.803 (.093) | 22.40 (1.134) | 5.25 (.307) |
| | | | HscadQR | 1.688 (.075) | 22.65 (1.094) | 5.55 (.235) |
| | | 0.5 | L1QR | 1.973 (.079) | 21.80 (0.999) | 5.40 (.285) |
| | | | HscadQR | 1.776 (.068) | 22.50 (1.077) | 5.85 (.344) |

Table 2: Simulation results of second scenario when $\tau = 0.5$ and 0.9 the prediction accuracy of L1QR and HscadQR: mean absolute deviation (standard errors)

| $\sigma^2$ | $\tau$ | $r$ | Method | Test MAD | inczeros | cnzeros |
|---|---|---|---|---|---|---|
| 1 | 0.5 | 0.0 | L1QR | 1.119 (.153) | 18.90 (1.114) | 3.00 (.000) |
| | | | HscadQR | 1.053 (.089) | 26.95 (1.895) | 3.00 (.000) |
| | | 0.5 | L1QR | 2.218 (.273) | 25.75 (1.258) | 2.65 (.109) |
| | | | HscadQR | 1.717 (.262) | 30.50 (1.786) | 2.80 (.092) |
| | 0.9 | 0.0 | L1QR | 2.869 (1.201) | 16.60 (1.001) | 3.00 (.000) |
| | | | HscadQR | 1.222 (.151) | 28.05 (1.745) | 3.00 (.000) |
| | | 0.5 | L1QR | 3.058 (.389) | 22.55 (1.097) | 2.20 (.117) |
| | | | HscadQR | 2.778 (.268) | 21.45 (1.069) | 2.35 (.131) |
| 3 | 0.5 | 0.0 | L1QR | 1.733 (.189) | 13.40 (1.379) | 2.90 (.069) |
| | | | HscadQR | 1.744 (.179) | 19.45 (1.468) | 2.85 (.082) |
| | | 0.5 | L1QR | 2.766 (.318) | 24.20 (1.130) | 2.10 (.124) |
| | | | HscadQR | 2.370 (.177) | 26.70 (1.298) | 2.20 (.117) |
| | 0.9 | 0.0 | L1QR | 4.577 (1.087) | 15.50 (0.783) | 2.55 (.153) |
| | | | HscadQR | 2.390 (.290) | 22.90 (1.796) | 3.00 (.000) |
| | | 0.5 | L1QR | 4.472 (.652) | 20.35 (0.862) | 1.70 (.105) |
| | | | HscadQR | 3.454 (.523) | 30.30 (1.223) | 2.10 (.145) |

In Table 2, the results of second scenario says that except case of $\tau = 0.5$ and $r = 0$, HscadQR performs better than L1QR. Such difference between L1QR and HscadQR is more intensified comparative to the results of the first scenario.

## 3.2. Microarray data

In this section, we apply the HscadQR to a quantile regression problem of gene microarrays in high dimensional settings. We employ the data set used in Scheetz *et al.* (2006), which consists of gene expression levels of 18,975 genes obtained from 120 rats. The main objective of the analysis is to find genes that are correlated with gene TRIM32 known to cause Bardet-Biedl syndrome. We first select 3000 genes with the largest variance in expression level, and then choose the top 500 genes that have the largest absolute correlation with gene TRIM32 among the selected 3000 genes.

Each data set is divided into two parts, training and test data sets, by randomly selecting 2/3

Table 3: Results of Microarray Data

| $\tau$ | Method | Test MAD | Nonzeros |
|---|---|---|---|
| 0.5 | L1QR | 1.096 (0.058) | 76.66 (1.361) |
| | HscadQR | 0.844 (0.019) | 75.64 (1.403) |
| 0.9 | L1QR | 1.087 (0.042) | 76.60 (0.928) |
| | HscadQR | 0.873 (0.023) | 75.94 (0.876) |

observations and 1/3 observations, respectively. The optimal values of the regularization parameters are chosen by GACV. Results of 20 replicated experiments are summarized in Table 3 according to $\tau$ quantiles.

As shown in Table 3, the HscadQR performs best in terms of Test MAD. Also, the number of nonzero coefficients of the HscadQR is similar with those of the L1QR.

## 4. Concluding Remark

In this paper, we have developed an efficient optimization algorithm for SCAD penalized quantile regression, specially which is appropriate on high dimensions. We show good performance from numerical results in sparse model. Also, its superiority is more confirmed in very sparse setting than moderate sparse setting.

## References

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

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

Kim, Y., Choi, H. and Oh, H. (2008). Smoothly clipped absolute deviation on high-dimensions, *Journal of the American Statistical Association,* To appear.

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

Konecker, R. and Portnoy, S. (1994). Quantile smoothing splines, *Biometrika,* **81**, 673–680.

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

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

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, 461–464.

Yuan, M. (2006). GACV for quantile smoothing splines, *Computational Statistics and Data Analysis,* **50**, 813–829.

Yuille, A. and Rangarajan, A. (2003). The concave-convex procedure, *Neural Computation,* **15**, 915–936.

*Received November 2008; Accepted January 2009*