L_1-Norm Quantile Regression

Youjuan LI and Ji ZHU

Classical regression methods have focused mainly on estimating conditional mean functions. In recent years, however, quantile regression has emerged as a comprehensive approach to the statistical analysis of response models. In this article we consider the L_1-norm (LASSO) regularized quantile regression (L_1-norm QR), which uses the sum of the absolute values of the coefficients as the penalty. The L_1-norm penalty has the advantage of simultaneously controlling the variance of the fitted coefficients and performing automatic variable selection. We propose an efficient algorithm that computes the entire solution path of the L_1-norm QR. Furthermore, we derive an estimate for the effective dimension of the L_1-norm QR model, which allows convenient selection of the regularization parameter.

Key Words: Effective dimension; LASSO; Linear programming; L_1-norm penalty; Variable selection.

1. INTRODUCTION

Classical regression methods have focused mainly on estimating conditional mean functions. In recent years, however, quantile regression has emerged as a comprehensive approach to the statistical analysis of response models, and it has been widely used in many real applications, such as reference charts in medicine (Cole and Green 1992; Heagerty and Pepe 1999), survival analysis (Yang 1999; Koenker and Geling 2001) and economics (Hendricks and Koenker 1992; Koenker and Hallock 2001).

Suppose we have a set of training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ are the predictors, and $y_i \in \mathbb{R}$ is the response. We consider the following regularized model fitting for finding the $100\tau\%$ quantile function:

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \boldsymbol{\beta}^\top \boldsymbol{x}_i) + \lambda \|\boldsymbol{\beta}\|_1, \tag{1.1}$$

Youjuan Li is a Ph.D. Student, and Ji Zhu is Assistant Professor, Department of Statistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: *jizhu@umich.edu.*).

Journal of Computational and Graphical Statistics, Volume 17, Number 1, Pages 1–23
DOI: 10.1198/106186008X289155

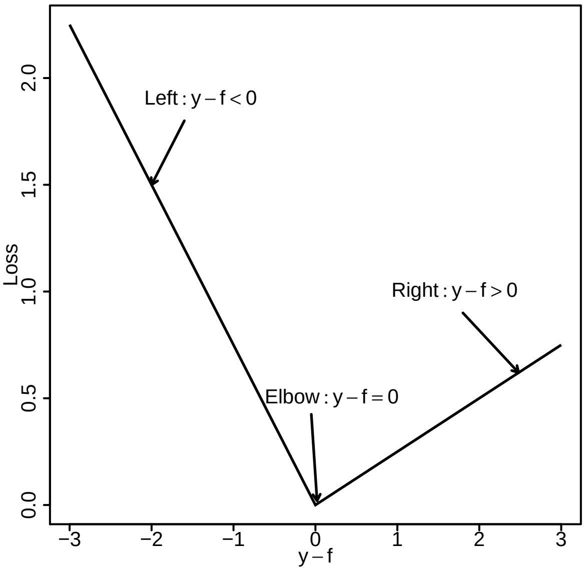

Figure 1. The check function with $\tau = 0.25$. We divide data points into three sets based on their associated residuals $y_i - f(\boldsymbol{x}_i)$. The three sets are left, elbow, and right.

which we will refer to as the L_1*-norm quantile regression* (L_1-norm QR). The loss $\rho_\tau(\cdot)$ is called the check function of Koenker and Bassett (1978):

$$\rho_\tau(y - f(\boldsymbol{x})) = \begin{cases} \tau \cdot (y - f(\boldsymbol{x})) & \text{if } y - f(\boldsymbol{x}) > 0, \\ -(1-\tau)\cdot(y - f(\boldsymbol{x})) & \text{otherwise,} \end{cases} \tag{1.2}$$

where $f(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}$. Here $\tau \in (0,1)$ indicates the quantile of interest. The check function is an analogue of the squared error loss in the context of least squares regression. One can verify that given $\boldsymbol{X} = \boldsymbol{x}$, the population minimizer to the check function is the $100\tau\%$ conditional quantile, that is,

$$100\tau\% \text{ quantile of } (Y|\boldsymbol{X} = \boldsymbol{x}) = \arg\min_f \mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}[\rho_\tau(Y - f(\boldsymbol{X}))].$$

Figure 1 shows the check function with $\tau = 0.25$.

The models considered are of the form $f(\boldsymbol{x}_i) = \beta_0 + \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_i$, and we penalize the model's complexity using the L_1-norm of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$. $\lambda > 0$ is a regularization parameter that balances the quantile loss and the penalty.

Canonical examples using the explicit L_1-norm penalty include the basis pursuit model (Chen, Donoho, and Saunders 1998) and the LASSO model (Tibshirani 1996) for least squares regression:

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^n (y_i - \beta_0 - \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_i)^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

The L_1-norm penalty not only shrinks the fitted coefficients toward zero but also causes some of the fitted coefficients to be *exactly* zero when making λ sufficiently large. The latter property is not shared by other types of penalties such as the L_2-norm penalty (Hoerl and Kennard 1970). Thus, in situations where there are a lot of irrelevant noise variables, the L_1-norm penalty may prove superior to the L_2-norm penalty from a prediction error perspective. From an inference/interpretation perspective, the L_1-norm penalty allows smooth variable selection and offers more compact models than the L_2-norm penalty.

Note that (1.1) has an L_1 *loss* + L_1 *penalty* structure. Some previous work also considered this format. For example, in the case of $p = 1$, Koenker, Ng, and Portnoy (1994) propose using $\lambda \int_0^1 |f''(x)|dx$ as the penalty, that is,

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \rho_\tau(y_i - f(x_i)) + \lambda \int_0^1 |f''(x)|dx,$$

where $\mathcal{F}$ is a certain model space. With an appropriately chosen $\mathcal{F}$, Koenker, Ng, and Portnoy (1994) showed that the solution is a linear spline with knots at $x_i, i = 1, \ldots, n$, which leads essentially also to an L_1 loss + L_1 penalty problem.

As in every regularized model fitting, choice of the regularization parameter is critical. In practice, people usually prespecify a finite set of values for the regularization parameter, then use either a validation dataset or a certain model-selection criterion to pick the regularization parameter. Two commonly used criteria in the quantile regression literature are the Schwarz information criterion (Schwarz 1978; Koenker, Ng, and Portnoy 1994) (SIC) and the generalized approximate cross-validation criterion (Yuan 2006) (GACV):

$$\mathrm{SIC}(\lambda) = \ln\left(\frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - f(\boldsymbol{x}_i))\right) + \frac{\ln n}{2n} df, \tag{1.3}$$

$$\mathrm{GACV}(\lambda) = \frac{\sum_{i=1}^{n} \rho_\tau(y_i - f(\boldsymbol{x}_i))}{n - df}, \tag{1.4}$$

where *df* is a measure of the effective dimensionality of the fitted model. Koenker, Ng, and Portnoy (1994) heuristically argued that in the case of one-dimensional quantile smoothing splines, the number of interpolated y_i's is a plausible measure for the effective dimension of the fitted model. In the case of GACV, Yuan (2006) used a smooth approximation of the check function to estimate *df*.

For the rest of the article, we rewrite (1.1) as an equivalent constrained optimization problem (the reason will become clear in Section 2):

$$\min_{\beta_0, \boldsymbol{\beta}} \quad \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_i), \tag{1.5}$$

$$\text{subject to} \quad |\beta_1| + \cdots + |\beta_p| \le s, \tag{1.6}$$

where s is the regularization parameter, playing the same role as λ. This constrained optimization problem is equivalent to (1.1), in the sense that for every given positive value of s, there exists a positive value of λ, such that the solutions to the two problems are identical.

In this article, we make two main contributions:

- We show that $\boldsymbol{\beta}(s)$, the fitted coefficients by solving (1.5)–(1.6) for a given s, is *piecewise linear* as a function of s, and we derive an efficient algorithm that computes the *exact entire solution path* $\{\boldsymbol{\beta}(s), 0 \le s \le \infty\}$. We make a note that the algorithm is fundamentally *different* from the LARS/LASSO algorithm in Efron et al.

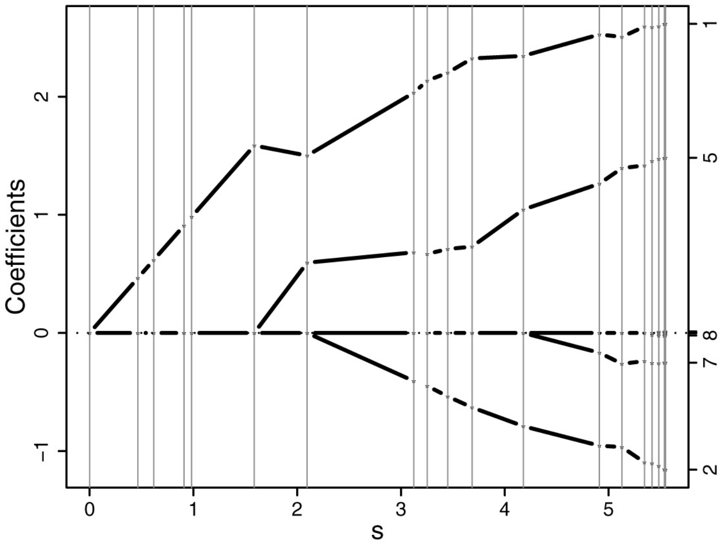

Figure 2. The solution path $\boldsymbol{\beta}(s)$ as a function of s. Any segment between two adjacent vertical lines is linear, hence the whole solution path is piecewise linear. The indices of predictor variables are labeled on the right side axis. Predictors 1, 2, and 5 are relevant variables, and the corresponding true coefficients are 3, -1.5, and 2, respectively. As we can see, over a range of s, only these three predictors have nonzero fitted coefficients.

(2004) and the kernel quantile regression (KQR) algorithm we have developed earlier in Li, Liu, and Zhu (2007), because we are now dealing with a nondifferentiable loss function *and* a nondifferentiable penalty.

- We prove that the number of interpolated y_i's is an estimate of the effective dimension of the fitted model, which allows convenient selection of the regularization parameter s, and also justifies the conjecture of Koenker, Ng, and Portnoy (1994).

Before delving into the technical details, we illustrate the concept of sparsity and piecewise linearity of the solution path with a simple example: $y = \beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{x} + \varepsilon$, where $\beta_0 = 0$, $\boldsymbol{\beta}^{\mathsf{T}} = (3, -1.5, 0, 0, 2, 0, 0, 0)$, $\boldsymbol{x}$ is distributed as Normal$(0, \boldsymbol{I}_{8\times 8})$, and ϵ is distributed as the standard double exponential. We generate 30 observations and fit the L_1-norm QR model with $\tau = 0.5$. Figure 2 shows the solution path $\boldsymbol{\beta}(s)$ as a function of s. As we can see, any segment between two adjacent vertical lines is linear, hence the whole solution path is piecewise linear. Another important feature of this solution path is, over a certain range of values of s, only the relevant predictor variables, that is, x_1, x_2, and x_5, have nonzero fitted coefficients.

The rest of the article is organized as follows: In Section 2, we derive an efficient algorithm for computing the entire solution path. In Section 3, we propose the number of interpolated y_i's as an estimate for the effective dimension of the fitted L_1-norm QR model. In Section 4, we present numerical results on simulation datasets and a real-world microarray dataset. We conclude the article in Section 5.

2. ALGORITHM

In this section, we derive an efficient algorithm that computes the exact solution path $\{\boldsymbol{\beta}(s), 0 \leq s \leq \infty\}$. We assume the $(\boldsymbol{x}_i, y_i)$ are in general positions and the solution $\boldsymbol{\beta}(s)$ is unique (except for the initial solution which is described in Section 2.2).

2.1 PROBLEM SETUP

Criterion (1.5)–(1.6) can be rewritten in an equivalent way:

$$\begin{aligned} \min_{\beta_0, \boldsymbol{\beta}} \quad & \tau \sum_{i=1}^{n} \xi_i + (1-\tau) \sum_{i=1}^{n} \zeta_i, \\ \text{subject to} \quad & \sum_{j=1}^{p} |\beta_j| \leq s, \\ & -\zeta_i \leq y_i - f(\boldsymbol{x}_i) \leq \xi_i, \\ & \zeta_i, \xi_i \geq 0, \quad i = 1, \ldots, n, \end{aligned}$$

where $f(\boldsymbol{x}_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$. The above setting gives the Lagrangian primal function

$$\begin{aligned} L_p : \tau \sum_{i=1}^{n} \xi_i + (1-\tau) \sum_{i=1}^{n} \zeta_i + \lambda^* (\sum_{j=1}^{p} |\beta_j| - s) + \sum_{i=1}^{n} \alpha_i (y_i - f(\boldsymbol{x}_i) - \xi_i) \quad (2.1) \\ - \sum_{i=1}^{n} \gamma_i (y_i - f(\boldsymbol{x}_i) + \zeta_i) - \sum_{i=1}^{n} \kappa_i \xi_i - \sum_{i=1}^{n} \eta_i \zeta_i, \end{aligned}$$

where $\lambda^*, \alpha_i, \gamma_i, \kappa_i$, and η_i are non-negative Lagrangian multipliers. Setting the derivatives of L_p to zero, we arrive at

$$\frac{\partial}{\partial \boldsymbol{\beta}} : \lambda^* \cdot \operatorname{sign}(\beta_j) - \sum_{i=1}^{n} (\alpha_i - \gamma_i) x_{ij} = 0, \quad \forall j \text{ with } \beta_j \neq 0, \tag{2.2}$$

$$\frac{\partial}{\partial \beta_0} : \sum_{i=1}^{n} (\alpha_i - \gamma_i) = 0, \tag{2.3}$$

$$\frac{\partial}{\partial \xi_i} : \tau = \alpha_i + \kappa_i, \tag{2.4}$$

$$\frac{\partial}{\partial \zeta_i} : 1 - \tau = \gamma_i + \eta_i, \tag{2.5}$$

and the Karush–Kuhn–Tucker (KKT) conditions are

$$\alpha_i (y_i - f(\boldsymbol{x}_i) - \xi_i) = 0, \tag{2.6}$$

$$\gamma_i (y_i - f(\boldsymbol{x}_i) + \zeta_i) = 0, \tag{2.7}$$

$$\kappa_i \xi_i = 0, \tag{2.8}$$

$$\eta_i \zeta_i = 0. \tag{2.9}$$

Since the Lagrange multipliers must be non-negative, we conclude from (2.4) and (2.5) that both $0 \le \alpha_i \le \tau$ and $0 \le \gamma_i \le 1 - \tau$. Furthermore, when $y_i - f(\boldsymbol{x}_i) > 0$ (hence $\xi_i > 0$), we have $\alpha_i = \tau$ and $\gamma_i = 0$; when $y_i - f(\boldsymbol{x}_i) < 0$ (hence $\zeta_i > 0$), we have $\alpha_i = 0$, and $\gamma_i = 1 - \tau$. These lead to the following relationships:

$$\begin{array}{lllll} y_i - f(\boldsymbol{x}_i) > 0 \quad \Rightarrow & \alpha_i = \tau, & \xi_i > 0, & \gamma_i = 0, & \zeta_i = 0; \\ y_i - f(\boldsymbol{x}_i) < 0 \quad \Rightarrow & \alpha_i = 0, & \xi_i = 0, & \gamma_i = 1 - \tau, & \zeta_i > 0; \\ y_i - f(\boldsymbol{x}_i) = 0 \quad \Rightarrow & \alpha_i \in [0, \tau], & \xi_i = 0, & \gamma_i \in [0, 1 - \tau], & \zeta_i = 0. \end{array}$$

Notice that (2.2) and (2.3) depend on α_i and γ_i only in their difference. Let $\theta_i = \alpha_i - \gamma_i$, hence using these relationships, we can define the following four sets that will be used later when we calculate the solution path of the L_1-norm QR:

- $\mathcal{E} = \{i : y_i - f(\boldsymbol{x}_i) = 0, -(1 - \tau) \le \theta_i \le \tau\}$ (elbow)
- $\mathcal{L} = \{i : y_i - f(\boldsymbol{x}_i) < 0, \theta_i = -(1 - \tau)\}$ (left of the elbow)
- $\mathcal{R} = \{i : y_i - f(\boldsymbol{x}_i) > 0, \theta_i = \tau\}$ (right of the elbow)
- $\mathcal{V} = \{j : \beta_j \neq 0\}$ (active set)

Since our goal is to compute the solution path $\boldsymbol{\beta}(s)$, we are interested in how the KKT conditions change when the regularization parameter s increases. When s increases, we define an *event* to be

- either a data point hits the elbow, that is, a residual $y_i - f(\boldsymbol{x}_i)$ changes from nonzero to zero, or
- a coefficient β_j changes from nonzero to zero, that is, a variable leaves the active set, $\mathcal{V}$.

These two changes correspond to the nonsmooth points of $\sum_i \rho_\tau(y_i - f(\boldsymbol{x}_i))$ and $||\boldsymbol{\beta}||_1$, respectively. Note that it is also possible for a residual to change from zero to nonzero, or a coefficient to change from zero to nonzero, and we handle these two cases towards the end of Section 2.3. Given the above definition of the events, we can see:

- As s increases, the sets $\mathcal{V}$, $\mathcal{L}$, $\mathcal{R}$, and $\mathcal{E}$ will not change (or equivalently, the KKT conditions will not change), unless an event happens. When the KKT conditions do not change, from (2.2)–(2.3), there are $|\mathcal{E}| + 1$ unknowns, that is, λ^* and $\theta_i = \alpha_i - \gamma_i$ with $i \in \mathcal{E}$, and $|\mathcal{V}| + 1$ equations. For the solution to be unique, we have the number of observations in the elbow equal to the number of variables in the active set, that is, $|\mathcal{E}| = |\mathcal{V}|$.
- As s increases, points in $\mathcal{E}$ stay in the elbow, unless an event happens. Therefore, nonzero β_j's satisfy:

$$y_i - (\beta_0 + \sum_{j \in \mathcal{V}} \beta_j x_{ij}) = 0 \quad \text{for} \quad i \in \mathcal{E}.$$

Since $|\mathcal{V}| = |\mathcal{E}|$, there is one free unknown in this set of equations, which allows $\boldsymbol{\beta}$ to change linearly when s increases, unless an event happens.

The basic idea of our algorithm is as follows: We start with $s = 0$ and increase it, keeping track of the location of all data points relative to the elbow and also of the magnitude of the fitted coefficients along the way. As s increases, for a point to pass through $\mathcal{E}$, the corresponding θ_i must change from τ to $-(1-\tau)$ or vice versa, hence by continuity, points in $\mathcal{E}$ must linger in the elbow. Since all points in the elbow have $y_i - f(\boldsymbol{x}_i) = 0$, we can establish a path for $\boldsymbol{\beta}$. The elbow set will stay stable until either some other point comes to the elbow or one nonzero fitted coefficient has dropped to zero.

2.2 INITIALIZATION

Initially, at $s = 0$, we can see from (1.6) that $f(\boldsymbol{x}) = \beta_0$. We can determine the value of β_0 via a simple one-dimensional optimization. For expositional simplicity, we focus on the case that all the values of y_i are distinct and ordered $y_1 < y_2 < \cdots < y_n$. This is the usual case for quantitative data and can always be realized by adding a small jitter to the y values. We distinguish between two cases: the initial β_0 is unique, and the initial β_0 is nonunique.

2.2.1 Case 1: The Initial β_0 is Unique

This happens when $n\tau$ is a noninteger, for example, when $\tau = 0.5$, and the number of data points n is odd. In this case, it is easy to show that β_0 must be equal to one of the observed y_i's and $\beta_0 = y_{\lfloor n\tau \rfloor + 1}$; we denote it as y_{i^*}. All data points are therefore initially divided into the three sets:

- $\mathcal{E} = \{i^* : \text{point } (\boldsymbol{x}_{i^*}, y_{i^*})\}$,
- $\mathcal{L} = \{i : y_i < y_{i^*}\}$, and
- $\mathcal{R} = \{i : y_i > y_{i^*}\}$.

From (2.3), we have

$$\theta_{i^*} = (1-\tau)n_{\mathcal{L}} - \tau n_{\mathcal{R}},$$

where $n_{\mathcal{L}} = |\mathcal{L}|$ and $n_{\mathcal{R}} = |\mathcal{R}|$. To find the initial $\mathcal{V}$, we compute $\max_j |\sum_i \theta_i x_{ij}|$, and according to (2.2) it corresponds to the largest feasible λ^*. Thus we get:

$$\mathcal{V} = \{j^\star : \arg\max_j |\sum_i \theta_i x_{ij}|\}.$$

Therefore, for small enough s, we have

$$f(\boldsymbol{x}) = \beta_0 + s \cdot \text{sign}(\sum_i \theta_i x_{ij^\star}) x_{j^\star},$$

where $\beta_0 = y_{i^*} - s \cdot \text{sign}(\sum_i \theta_i x_{ij^\star}) x_{i^* j^\star}$ since $(\boldsymbol{x}_{i^*}, y_{i^*})$ stays in the elbow.

2.2.2 Case 2: The Initial β_0 is Nonunique

This happens when $n\tau$ is an integer, for example, when $\tau = 0.5$ and the number of data points n is even. In this case, it is easy to show that β_0 can take any value between two adjacent y_i's and $\beta_0 \in (y_{n\tau}, y_{n\tau+1})$; we denote them as (y_{i^*}, y_{i^*+1}).

Although β_0 is not unique, all the θ_i's are fully determined, that is,

- $\theta_i = -(1-\tau),\ y_i \leq y_{i^*}$, and
- $\theta_i = \tau,\ y_i \geq y_{i^*+1}$.

Hence again, we can divide all data points into the three sets:

- $\mathcal{E} = \emptyset$,
- $\mathcal{L} = \{i : y_i \leq y_{i^*}\}$, and
- $\mathcal{R} = \{i : y_i \geq y_{i^*+1}\}$.

Similar to case 1, the initial $\mathcal{V}$ can be found by

$$\mathcal{V} = \left\{ j^\star : \arg\max_j |\sum_i \theta_i x_{ij}| \right\},$$

and for sufficiently small s, we have

$$f(\boldsymbol{x}) = \beta_0 + s \cdot \text{sign}\left(\sum_i \theta_i x_{ij^\star} \right) x_{j^\star}.$$

When s increases, Equation (2.3) imposes a constraint on all the θ_i's. Since to pass through $\mathcal{E}$, a θ_i must change from τ to $-(1-\tau)$ or vice versa, by continuity, the sets $\mathcal{L}$ and $\mathcal{R}$ will stay stable. Therefore

$$y_i - \beta_0 - s \cdot \text{sign}\left(\sum_i \theta_i x_{ij^\star} \right) x_{ij^\star} < 0, \quad i \in \mathcal{L},$$

$$y_i - \beta_0 - s \cdot \text{sign}\left(\sum_i \theta_i x_{ij^\star} \right) x_{ij^\star} > 0, \quad i \in \mathcal{R}.$$

These inequalities imply that the solution for β_0 is not unique, and β_0 can be any value in the interval

$$\left(\max_{i \in \mathcal{L}} (y_i - s \cdot \text{sign}(\theta_i x_{ij^\star}) x_{ij^\star}), \quad \min_{i \in \mathcal{R}} (y_i - s \cdot \text{sign}(\sum_i \theta_i x_{ij^\star}) x_{ij^\star}) \right).$$

When s increases, the length of this interval will shrink toward zero, which corresponds to two data points (from different sets) hitting the elbow simultaneously.

2.3 THE REGULARIZATION PATH

The algorithm focuses on the set of points $\mathcal{E}$ and the set of nonzero coefficients $\mathcal{V}$. Until an event (as defined in Section 2.1) has occurred, all sets will remain the same. Points in $\mathcal{E}$ have $f(\boldsymbol{x}_i) = y_i$. Relying on this fact, we can calculate how $\boldsymbol{\beta}$ changes.

We use the subscript ℓ to index the sets above immediately after the ℓth event has occurred, and let β_0^ℓ, $\boldsymbol{\beta}^\ell$, and s^ℓ be the parameter values immediately after the ℓth event. Also let f^ℓ be the function at this point. For $s^\ell < s < s^{\ell+1}$, we can write

$$\begin{aligned} f(\boldsymbol{x}) &= f(\boldsymbol{x}) - f^\ell(\boldsymbol{x}) + f^\ell(\boldsymbol{x}) \\ &= (\beta_0 - \beta_0^\ell) + \sum_{j \in \mathcal{V}^\ell} (\beta_j - \beta_j^\ell) x_j + f^\ell(\boldsymbol{x}). \end{aligned}$$

Suppose $n_{\mathcal{E}}^\ell = |\mathcal{E}^\ell|$ and $n_{\mathcal{V}}^\ell = |\mathcal{V}^\ell|$, so for the points staying at the elbow, we have that

$$y_i = (\beta_0 - \beta_0^\ell) + \sum_{j \in \mathcal{V}^\ell} (\beta_j - \beta_j^\ell) x_{ij} + y_i, \quad \forall i \in \mathcal{E}^\ell.$$

Also we have that

$$\sum_{j \in \mathcal{V}^\ell} (\beta_j - \beta_j^\ell) \cdot \text{sign}(\beta_j^\ell) = s - s^\ell.$$

To simplify, let $\nu_0 = (\beta_0 - \beta_0^\ell)/(s - s^\ell)$ and $\nu_j = (\beta_j - \beta_j^\ell)/(s - s^\ell)$. Then

$$\begin{aligned} \nu_0 + \sum_{j \in \mathcal{V}^\ell} \nu_j x_{ij} = 0, \quad \forall i \in \mathcal{E}^\ell \\ \sum_{j \in \mathcal{V}^\ell} \nu_j \cdot \text{sign}(\beta_j^\ell) = 1. \end{aligned}$$

Recall $n_{\mathcal{E}}^\ell = n_{\mathcal{V}}^\ell$, thus this gives us $n_{\mathcal{E}}^\ell + 1$ linear equations we can use to solve for each of the $n_{\mathcal{V}}^\ell + 1$ unknown variables ν_0 and ν_j.

Now, define $\boldsymbol{X}^\ell$ to be a $n_{\mathcal{E}}^\ell \times n_{\mathcal{V}}^\ell$ matrix with the entries equal to x_{ij} where $i \in \mathcal{E}^\ell$, $j \in \mathcal{V}^\ell$, and let $\boldsymbol{\nu}$ denote the vector with the components equal to ν_j, $j \in \mathcal{V}^\ell$. Using these we have the following two equations

$$\nu_0 \mathbf{1} + \boldsymbol{X}^\ell \boldsymbol{\nu} = \mathbf{0}, \tag{2.10}$$

$$\boldsymbol{\nu}^\mathsf{T} \text{sign}(\boldsymbol{\beta}_{\mathcal{V}}^\ell) = 1. \tag{2.11}$$

Simplifying further, if we let

$$\boldsymbol{X}_0^\ell = \begin{pmatrix} \mathbf{1} & \boldsymbol{X}^\ell \\ 0 & \text{sign}^\mathsf{T}(\boldsymbol{\beta}_{\mathcal{V}}^\ell) \end{pmatrix}, \quad \boldsymbol{\nu}_0 = \begin{pmatrix} \nu_0 \\ \boldsymbol{\nu} \end{pmatrix}, \quad \text{and } \mathbf{1}_0 = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix},$$

Equations (2.10) and (2.11) can be combined to be

$$\boldsymbol{X}_0^\ell \boldsymbol{\nu}_0 = \mathbf{1}_0.$$

Since we assume $(\boldsymbol{x}_i, y_i)$ are in general positions, $\boldsymbol{X}_0^\ell$ is usually of full rank, then we have

$$\beta_0 = \beta_0^\ell + (s - s^\ell)\nu_0, \tag{2.12}$$

$$\beta_j = \beta_j^\ell + (s - s^\ell)\nu_j, \quad \forall j \in \mathcal{V}^\ell. \tag{2.13}$$

Thus for $s^\ell < s < s^{\ell+1}$, the β_j and β_0 proceed linearly in s. Also

$$f(\boldsymbol{x}) = (s - s^\ell)\left(\nu_0 + \sum_{j \in \mathcal{V}^\ell} \nu_j x_j\right) + f^\ell(\boldsymbol{x}). \tag{2.14}$$

Given s^ℓ, Equations (2.13) and (2.14) allow us to compute $s^{\ell+1}$, the s at which the next event will occur. This will be the smallest s larger than s^ℓ, such that either $f(\boldsymbol{x}_i)$ for $i \notin \mathcal{E}^\ell$ reaches y_i, or one of the coefficients β_j for $j \in \mathcal{V}^\ell$ reaches zero.

To update sets $\mathcal{V}$ and $\mathcal{E}$ when the ℓth event occurs, by the definition of an event in Section 2.1, there will be $|\mathcal{V}|$ variables with nonzero coefficients and $|\mathcal{V}| + 1$ points in the elbow. Therefore, to maintain the KKT conditions, we need to either add a variable not in $\mathcal{V}$ into $\mathcal{V}$, or remove a point in $\mathcal{E}$ from $\mathcal{E}$. The choice is such that the resulting loss in (1.5) decreases with the fastest rate. We compute the rate of the change in the loss function as the following:

$$\begin{aligned}\frac{\Delta\text{loss}}{\Delta s} &= \frac{\sum_i \rho_\tau(y_i - f(\boldsymbol{x}_i)) - \sum_i \rho_\tau(y_i - f^\ell(\boldsymbol{x}_i))}{s - s^\ell} \\ &= (1-\tau)\sum_{i \in \mathcal{L}}\left(\nu_0 + \sum_{j \in \mathcal{V}} \nu_j x_{ij}\right) - \tau \sum_{i \in \mathcal{R}}\left(\nu_0 + \sum_{j \in \mathcal{V}} \nu_j x_{ij}\right).\end{aligned}$$

We choose the update that corresponds to the smallest (negative) $\Delta\text{loss}/\Delta s$, and terminate the algorithm when all $\Delta\text{loss}/\Delta s$ are non-negative.

In fact, we can show that $\Delta\text{loss}/\Delta s$ is related to the parameter λ^* in (2.2).

Theorem 1. *The rate that the loss decreases along the solution path is the same as the value of the parameter λ^*, that is,*

$$\frac{\Delta\text{loss}}{\Delta s} = -\lambda^*.$$

The details of the proof are in the Appendix.

2.4 COMPUTATIONAL COST

The major computational cost for updating the solutions at any step ℓ involves two things: solving the system of $n_{\mathcal{E}}^\ell + 1$ linear equations, and finding the correct update for $\mathcal{V}$ and $\mathcal{E}$. The former takes $O(n_{\mathcal{E}}^{\ell 2})$ calculations by using inverse updating and downdating since the sets usually differ by only one element between consecutive events, and the latter requires $O(pn_{\mathcal{E}}^{\ell 2})$ computations.

According to our experience, the total number of steps taken by the algorithm is on average $O(\min(n, p))$. This can be heuristically understood in the following way: if $n < p$, the data points can be perfectly interpolated by a linear model, then it takes $O(n)$ steps for every data point to reach the elbow; if $n > p$, then it takes $O(p)$ steps to include all variables into the fitted model. Since the maximum value of $n_{\mathcal{E}}^{\ell}$ is $\min(n, p)$, it suggests the worst computational cost is $O(p \min(n, p)^3)$.

3. EFFECTIVE DIMENSION

It is well known that an appropriate value of the regularization parameter is crucial for the performance of the fitted model in any regularized model fitting. One advantage of computing the entire solution path is to facilitate the selection of the regularization parameter. In practice, one can first use the efficient algorithm in Section 2 to compute the entire solution path, then select the value of s that minimizes a certain model selection criterion. This avoids a more computationally intensive cross-validation approach.

Two commonly used criteria for quantile regression are the SIC (1.3) and the GACV (1.4), where both depend on the quantity *df* which should be an informative measure of the complexity of a fitted model. Nychka et al. (1995) and Yuan (2006) proposed to use the SURE divergence formula (Stein 1981)

$$\sum_{i=1}^{n} \frac{\partial \hat{f}(\boldsymbol{x}_i)}{\partial y_i} \tag{3.1}$$

to estimate *df*, where $\hat{f}(\boldsymbol{x})$ is a fitted model. To compute (3.1), they approximated the check function with a differentiable function $\rho_{\tau,\delta}(\cdot)$, which differs from $\rho_\tau(\cdot)$ within an interval $(-\delta, \delta)$:

$$\rho_{\tau,\delta}(r) = \begin{cases} \tau r & r \geq \delta \\ \tau r^2/\delta & 0 \leq r < \delta \\ (1-\tau) r^2/\delta & -\delta \leq r < 0 \\ -(1-\tau) r & r < -\delta \end{cases}$$

where δ is a small positive number.

Notice that (3.1) measures the sum of the sensitivity of each fitted value with respect to the corresponding observed value. This quantity first appeared under the framework of Stein's unbiased risk estimation (SURE) theory (Stein 1981). Given $\boldsymbol{x}$, assuming y is generated according to a homoscedastic model:

$$y \sim (\mu(\boldsymbol{x}), \sigma^2),$$

where μ is the true mean and σ^2 is the common variance, then the degrees of freedom of a fitted model $\hat{f}(\boldsymbol{x})$ can be defined as

$$\mathrm{df}(\hat{f}) = \sum_{i=1}^{n} \mathrm{cov}(\hat{f}(\boldsymbol{x}_i), y_i)/\sigma^2.$$

Stein showed that under mild conditions, $\sum_{i=1}^{n} \partial \hat{f}(\boldsymbol{x}_i)/\partial y_i$ is an unbiased estimate of $\mathrm{df}(\hat{f})$. Later on, Efron (1986) proposed the concept *expected optimism* based on (3.1), and

Ye (1998) developed Monte Carlo methods to estimate (3.1) for general modeling procedures. Meyer and Woodroofe (2000) discussed (3.1) in shape-restricted regression and also argued that it provides a measure of the effective dimension. For detailed discussion and complete references, we refer the readers to Efron (2004).

It turns out that in the case of L_1-norm QR, for every fixed s and almost all $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathsf{T}}$, $\sum_{i=1}^{n} \partial \hat{f}(\boldsymbol{x}_i)/\partial y_i$ has an extremely simple formula:

$$\sum_{i=1}^{n} \frac{\partial \hat{f}(\boldsymbol{x}_i)}{\partial y_i} = |\mathcal{E}|. \tag{3.2}$$

That is, the number of interpolated data points is a convenient estimate for the effective dimension of $\hat{f}(\boldsymbol{x})$, and this agrees with the heuristic conjecture of Koenker, Ng, and Portnoy (1994). We outline the proof of (3.2) in this section, and leave all the details to the Appendix.

As we have seen in Section 2, for a fixed response vector $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathsf{T}}$, there is a sequence of s's, $0 = s_0 < s_1 < s_2 < \cdots < s_L = \infty$, such that in the interior of any interval $(s_\ell, s_{\ell+1})$, the sets $\mathcal{R}$, $\mathcal{L}$, $\mathcal{E}$, and $\mathcal{V}$ are constant with respect to s. These sets only change at each s_ℓ. We thus define these s_ℓ's as *event points*.

Lemma 1. *For any fixed $s > 0$, the set of $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathsf{T}}$ such that s is an event point is a finite collection of hyperplanes in $\mathbb{R}^n$.*

Denote this set as $\mathcal{N}_s$. Then for any $\boldsymbol{y} \in \mathbb{R}^n \backslash \mathcal{N}_s$, s is not an event point. Notice $\mathcal{N}_s$ is a null set, and $\mathbb{R}^n \backslash \mathcal{N}_s$ is of full measure.

Lemma 2. *For any fixed $s > 0$, $\hat{\boldsymbol{\beta}}(\boldsymbol{y})$ is a continuous function of $\boldsymbol{y}$, where $\hat{\boldsymbol{\beta}}(\boldsymbol{y})$ is the fitted coefficient vector when the response vector is $\boldsymbol{y}$.*

Lemma 3. *For any fixed $s > 0$ and any $\boldsymbol{y} \in \mathbb{R}^n \backslash \mathcal{N}_s$, the sets $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{E}$ are locally constant with respect to $\boldsymbol{y}$.*

Theorem 2. *For any fixed $s > 0$ and any $\boldsymbol{y} \in \mathbb{R}^n \backslash \mathcal{N}_s$, we have the divergence formula*

$$\sum_{i=1}^{n} \frac{\partial \hat{f}(\boldsymbol{x}_i)}{\partial y_i} = |\mathcal{E}|.$$

4. NUMERICAL RESULTS

In this section, we use both simulation data and a real-world data to demonstrate our algorithm and the selection of s via the SIC criterion and the GACV criterion with *df* estimated by $|\mathcal{E}|$ (Section 3). We also compare the performance of the L_1-norm QR with that of the L_2-norm QR, that is, the L_1-norm in (1.1) is replaced by the L_2-norm.

4.1 SIMULATION DATA

We consider two scenarios:

1. $p < n$ case

We mimicked the simulation found in Tibshirani (1996). Data were generated using the mechanism:

$$y = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x} + \sigma\varepsilon,$$

where $\beta_0 = 0$, $\boldsymbol{\beta} \in \mathbb{R}^8$, and $\boldsymbol{x} \sim \text{Normal}(0, \boldsymbol{\Sigma}_{8\times 8})$. The pairwise correlation between x_j and x_k was $\rho^{|j-k|}$, with $\rho = 0$ for the independent case or $\rho = 0.5$ for the correlated case.

Three different error distributions were used: standard normal (N), double exponential (DE), and a mixture distribution (Mix):

$$0.1 \cdot N(0, 5^2) + 0.9 \cdot N(0, 1).$$

We also considered three different settings of $\boldsymbol{\beta}$ as follows:

a. Dense: $\beta_j = 0.85$, $j = 1, \dots, 8$, which corresponds to a dense scenario. We chose $\sigma = 1.75$, $\sqrt{1.5}$, and $\sqrt{0.9}$, respectively, for the three error distributions when $\rho = 0$, and $\sigma = 3$, 2, and 1.6, respectively, for the three error distributions when $\rho = 0.5$. The resulting signal-to-noise (S/N) ratios are all about 1.8. The S/N ratio is defined as $\text{var}(\boldsymbol{\beta}^\top \boldsymbol{X})/\text{var}(\sigma\varepsilon)$.

b. Sparse: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$, which corresponds to a moderately sparse case. We chose $\sigma = \sqrt{3}$, $\sqrt{1.5}$, and $\sqrt{0.9}$, respectively, for the three error distributions when $\rho = 0$, and $\sigma = 2$, $\sqrt{2}$, and 1, respectively, for the three error distributions when $\rho = 0.5$. The resulting S/N ratios are all about 5.

c. Very Sparse: $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)^\top$, which mimics a very sparse situation. We chose $\sigma = 2$, $\sqrt{2}$, and 1, respectively, for the three error distributions when $\rho = 0$ and 0.5. The resulting S/N ratios are all about 6.5.

We generated 100 training observations from each $\boldsymbol{\beta}$ setting, associated with each of the three error distributions and each of the two dependence relationships among covariates, along with 10, 000 validation observations and 10,000 test observations. We considered three different values of τ: 10%, 30%, and 50%. Since the error distributions are all symmetric, these τ's are also representative of the upper quantiles 70% and 90%. We then found the s's that minimized the SIC criterion and the GACV criterion, respectively. The validation set was used to select the *gold standard* s, which minimized the prediction error, that is, $\sum_{i=1}^{10{,}000} \rho_\tau(y_i - \hat{f}^\tau(\boldsymbol{x}_i))$. Using these s's we calculated the mean absolute deviations on the test dataset in order to evaluate different models' "goodness of fit." Suppose the fitted quantile function is $\hat{f}^\tau(\boldsymbol{x})$ and the true quantile function is $f^\tau(\boldsymbol{x})$, the mean absolute deviation is defined as

$$\text{Mean Absolute Deviation} = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \left| f^\tau(\boldsymbol{x}_i) - \hat{f}^\tau(\boldsymbol{x}_i) \right|.$$

We repeated the procedure 100 times. We computed the mean absolute deviations and recorded the effective dimensions of the selected models, that is, $|\mathcal{E}|$.

Table 1. Mean absolute deviations over 100 repetitions under the dense and very sparse scenarios. The numbers in parentheses are the corresponding standard deviations. "L_1-norm" and "L_2-norm" are for the L_1-norm QR and the L_2-norm QR, respectively. We report results on two model-selection methods, the SIC and the GACV, via our solution path algorithm and our formula for the effective dimension of the fitted model. The "Gold" (gold standard) serves as a benchmark. For the error distribution, "N" is normal, "DE" is double exponential, and "Mix" is a mixture distribution. In all settings, $n = 100$, $p = 8$, $\tau = 0.5$, $\rho = 0.5$.

	L_1-norm			L_2-norm		
	SIC	GACV	Gold	SIC	GACV	Gold
			Dense			
N	1.108 (0.169)	1.098 (0.168)	1.089 (0.170)	0.426 (0.077)	0.414 (0.079)	0.367 (0.069)
DE	1.139 (0.149)	1.126 (0.148)	1.120 (0.147)	0.339 (0.077)	0.322 (0.073)	0.274 (0.059)
Mix	1.151 (0.175)	1.138 (0.170)	1.132 (0.169)	0.264 (0.046)	0.252 (0.041)	0.221 (0.043)
			Very sparse			
N	0.340 (0.095)	0.366 (0.098)	0.270 (0.089)	0.543 (0.099)	0.524 (0.088)	0.457 (0.086)
DE	0.226 (0.085)	0.248 (0.081)	0.176 (0.072)	0.392 (0.096)	0.370 (0.081)	0.331 (0.078)
Mix	0.202 (0.056)	0.215 (0.057)	0.162 (0.050)	0.326 (0.052)	0.310 (0.047)	0.275 (0.046)

Table 1 shows the mean absolute deviation results for $\tau = 0.5$ and $\rho = 0.5$, under the dense and very sparse scenarios. Since the results for $\tau = 0.1$ and $\tau = 0.3$ are similar to those of $\tau = 0.5$, for lack of space, we omit them here. As we can see, in terms of the mean absolute deviation, both the SIC and the GACV perform closely to the gold standard. In the dense scenario, the L_2-norm QR performs better than the L_1-norm QR; while in the very sparse scenario, the L_1-norm QR performs better than the L_2-norm QR.

Table 2 shows the effective dimensions of the selected models and Table 3 shows the results on how frequently each variable was selected by the L_1-norm QR when ϵ has a double exponential distribution. As we can see, in terms of model selection, the SIC tends to select a simpler model than the GACV and the gold standard. In the L_1-norm QR model, both the SIC and the GACV perform reasonably well in selecting relevant variables, however, the SIC tends to be more effective in removing irrelevant variables than the GACV and the gold standard, especially in the very sparse case. It is interesting to observe that in the very sparse scenario, the gold standard did not identify the true model (Table 2). The true model contains only one relevant variable, while the gold standard on average selected a little more than three variables. In fact, as Leng, Lin, and Wahba (2006) pointed out, when the prediction accuracy is used as the criterion to choose the regularization parameter, the LASSO-type procedure is not consistent in selecting variables, that is, the probability that the LASSO-type procedure correctly identifies the set of relevant variables does not approach to 1 as the sample size goes to infinity. Our simulation results agree with Leng, Lin, and Wahba (2006)'s statement.

Table 2. Estimated effective dimensions, that is, $|\mathcal{E}|$, over 100 repetitions under the dense and very sparse scenarios. The numbers in the parentheses are the corresponding standard deviations. Descriptions of the columns and rows are the same as the caption in Table 1.

Dense	L_1-norm			L_2-norm		
	SIC	GACV	Gold	SIC	GACV	Gold
			Dense			
N	7.1 (0.4)	7.3 (0.4)	8.0 (0.1)	4.5 (0.9)	5.7 (0.9)	7.0 (0.9)
DE	7.1 (0.5)	7.4 (0.4)	8.0 (0.0)	4.7 (0.8)	5.4 (0.9)	7.6 (0.8)
Mix	7.1 (0.6)	7.5 (0.4)	8.0 (0.0)	5.2 (0.8)	5.9 (0.7)	7.5 (0.8)
			Very sparse			
N	1.8 (0.8)	3.6 (1.2)	3.4 (1.1)	6.3 (0.7)	6.8 (0.7)	8.2 (0.5)
DE	1.6 (0.6)	3.0 (1.2)	3.4 (1.2)	6.4 (0.7)	6.9 (0.5)	8.3 (0.5)
Mix	1.6 (0.5)	3.1 (1.1)	3.5 (1.2)	6.4 (0.6)	7.0 (0.5)	8.3 (0.4)

Table 3. Number of times each predictor variable was selected (out of 100 repetitions) by the L_1-norm QR: $n = 100$, $p = 8$, $\tau = 0.5$, $\rho = 0.5$, and $\epsilon \sim$ double exponential.

	X1	X2	X3	X4	X5	X6	X7	X8
			Dense: $\boldsymbol{\beta} = (0.85, \ldots, 0.85)$					
SIC	74	88	94	86	90	82	80	88
GACV	84	88	92	88	90	91	80	95
Gold	98	98	97	96	96	95	96	98
			Sparse: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$					
SIC	96	92	22	26	94	11	9	31
GACV	95	96	30	32	94	40	36	35
Gold	98	96	40	52	95	45	39	51
			Very Sparse: $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)$					
SIC	95	7	9	22	6	8	13	20
GACV	96	30	40	40	27	39	25	28
Gold	98	25	39	37	35	20	24	26

4.1.1 $p > n$ case

We mimicked the simulation found in Friedman et al. (2004). Data were generated using the mechanism

$$y = \beta_0 + \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{x} + \sigma \epsilon,$$

where $\beta_0 = 0$, $\boldsymbol{\beta} \in \mathbb{R}^{300}$, and $\boldsymbol{x} \sim \text{Normal}(0, \boldsymbol{\Sigma}_{300\times 300})$. The pairwise correlation between x_j and x_k was $\rho^{|j-k|}$, with $\rho = 0$ for the independent case or $\rho = 0.5$ for the correlated case.

Three different error distributions were used: standard normal (N), double exponential (DE), and a mixture distribution (Mix):

$$0.1 \cdot N(0, 5^2) + 0.9 \cdot N(0, 1).$$

Again, we considered the dense, sparse, and very sparse scenarios as the following:

a. Dense: all 300 coefficients were generated from the standard normal distribution.

b. Sparse: 30 nonzero coefficients were generated from the standard normal distribution.

c. Very sparse: only three coefficients are nonzero.

In each case, the coefficients are scaled such that the signal var($\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}$) is 1, and the noise var($\sigma \varepsilon$) ranges in 0.1, 0.3, and 0.5.

We generated 50 training observations from each function, associated with each of the three error distributions and each of the two dependence relationships among covariates, along with 10,000 validation observations and 10,000 test observations. We considered three different values of τ: 10%, 30%, and 50%. Since the SIC and the GACV criteria break down in the $p > n$ case, we only chose the *golden standard* models to compare the L_1-norm QR and the L_2-norm QR. The validation set was used to select the gold standard s. Using these s's we calculated the mean absolute deviations with the test data.

We repeated the procedure 100 times, and computed the average mean absolute deviations and the corresponding standard deviations. The results for $\tau = 50\%$ and $\rho = 0.5$ are reported in Table 4.

As we can see, in the $p > n$ case, the two types of regularized quantile regression models perform quite differently. In the dense scenario when all 300 coefficients are nonzero, neither the L_2-norm QR nor the L_1-norm QR performs very well, since there were too few data (only 50 observations) available to estimate the 300 coefficients. However, in the very sparse scenario when only three coefficients are nonzero, the L_1-norm QR performs significantly better than the L_2-norm QR.

4.2 REAL DATA

In this section, we apply the L_1-norm QR to a microarray-based study of cardiomyopathy in transgenic mice. The data are provided by Professor Mark Segal (Segal, Kam, and Bruce 2003). The study applied inducible gene expression techniques to control the

Table 4. Mean absolute deviations over 100 repetitions under the dense and very sparse scenarios. The numbers in the parentheses are the corresponding standard deviations. In all settings, $n = 50$, $p = 300$, $\tau = 0.5$, and $\rho = 0.5$. Descriptions of the columns and rows are the same as the caption in Table 1.

	0.1		0.3		0.5	
N/S Ratio	L1	L2	L1	L2	L1	L2
			Dense			
N	0.678 (0.007)	0.629 (0.010)	0.647 (0.012)	0.515 (0.023)	0.656 (0.013)	0.540 (0.023)
DE	0.677 (0.008)	0.628 (0.010)	0.646 (0.012)	0.514 (0.024)	0.653 (0.014)	0.535 (0.022)
Mix	0.677 (0.006)	0.629 (0.011)	0.641 (0.014)	0.503 (0.022)	0.646 (0.016)	0.516 (0.024)
			Very sparse			
N	0.143 (0.030)	0.556 (0.019)	0.246 (0.053)	0.585 (0.022)	0.306 (0.065)	0.606 (0.020)
DE	0.113 (0.030)	0.557 (0.020)	0.193 (0.051)	0.582 (0.021)	0.236 (0.060)	0.599 (0.019)
Mix	0.087 (0.022)	0.555 (0.020)	0.146 (0.036)	0.575 (0.022)	0.199 (0.041)	0.589 (0.019)

expression of a G protein-coupled receptor, designated Ro1, which is a transgene modified from human kappa-opioid receptor. Thirty mice were divided into four experimental groups (Redfern et al. 2000): The two-week group of six transgenic mice expressed Ro1 for two weeks, which is approximately the amount of time required to reach maximal expression of Ro1; these mice did not show symptoms of disease. The eight-week group of nine transgenic mice expressed Ro1 for eight weeks and exhibited cardiomyopathy symptoms. The recovery group of seven transgenic mice expressed Ro1 for eight weeks before expression was turned off for four weeks. The control group of eight mice was treated exactly the same as the eight-week group except that they did not have the Ro1 transgene. In Figure 3, the measures of Ro1 expression for these groups are denoted as 2, 8, R, and C, respectively. The experiment reported a Ro1 model of cardiomyopathy: When Ro1 was overexpressed in the heart of an adult mouse, the mouse developed a lethal cardiomyopathy. There is further evidence that cardiomyopathy is due to overexpression of Ro1 since it does not occur at more moderate expression levels of Ro1 (Redfern et al. 2000).

Identifying genes involved in the progression of cardiomyopathy, that is, gene expression changes associated with the Ro1 expression changes, may provide new diagnostic markers for cardiomyopathy. In our analysis, the response of interest was Ro1 expression and the predictors were all the 6,319 gene expressions measured using microarray technology. We fitted 50%, 75%, and 90% L_1-norm QR functions and used five-fold cross-validation to select the regularization parameter. The selected genes are listed in Table 5. As we can see, there are quite a few overlaps between the genes selected by the L_1-norm QR and those selected by Segal, Kam, and Bruce (2003) (which is essentially LASSO). We also notice that when there are a group of highly correlated genes, the L_1-norm QR tends to select only one or a few genes from the group (Zou and Hastie 2005). For example, gene AA044561 has high pairwise correlations with gene AA061310 ($\rho = 0.80$), gene W75373 ($\rho = 0.79$), and gene AA111168 ($\rho = 0.90$). Consequently, gene AA044561 was the only

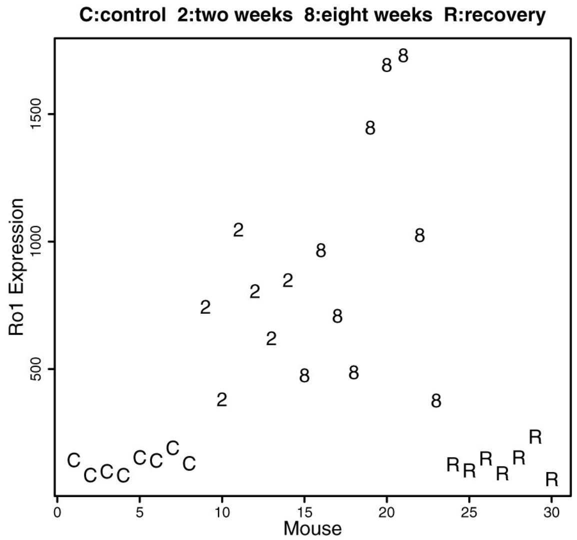

Figure 3. Ro1 expression for the 30 mice.

gene from the group selected by the $\tau = 75\%$ L_1-norm QR model, and it was excluded when other genes from the group were selected by other models. We recognize that these data analysis results need to be validated by further biological experiments.

5. SUMMARY

Our work can be considered as an extension of the LASSO model (Tibshirani 1996), where we use the check loss (for estimating the quantile function), and the LASSO uses the squared error loss (for estimating the mean function).

Table 5. Genes selected by three L_1-norm quantile regression models. The first column contains the gene IDs. Other columns indicate genes selected by different models.

GeneBank	$\tau = 50\%$	$\tau = 75\%$	$\tau = 90\%$	Segal, Kam, and Bruce (2003)
D31717	√	√	√	√
U73744	√	√	√	√
U25708	√	√		√
AA061310	√			√
M30127	√	√	√	
L38971	√			
Z32675	√			
W75373	√			
AA044561		√		
AA111168			√	
M18194				√

Our work is also connected with Koenker, Ng, and Portnoy (1994), where $p = 1$, and $\lambda \int_0^1 |f''(x)|dx$ was used as the penalty. With an appropriately chosen model space, Koenker, Ng, and Portnoy (1994) showed that the solution is a linear spline with knots at the points $x_i, i = 1, \ldots, n$, which leads essentially also to an L_1 loss $+$ L_1 penalty problem.

We have gone beyond the spline model, and considered general L_1 regularized quantile regression. In particular, we have proposed an efficient algorithm that computes the entire regularization path for the L_1-norm QR. Our path algorithm was inspired by the LARS/LASSO algorithm (Efron et al. 2004). Since we are dealing with a nondifferentiable loss function, our algorithm is fundamentally different from the LARS/LASSO algorithm. We have also proposed an estimate for the effective dimension of the fitted model that can be used to select the regularization parameter. This estimate seems to work sufficiently well on the simulation data (when $n > p$).

A. APPENDIX: PROOFS

A.1 PROOF OF THEOREM 1

From (2.2) we have

$$\lambda^* \cdot \text{sign}(\beta_j) = \tau \sum_{i\in\mathcal{R}} x_{ij} - (1-\tau)\sum_{i\in\mathcal{L}} x_{ij} + \sum_{i\in\mathcal{E}} \theta_i x_{ij}, \quad j \in \mathcal{V}.$$

Multiply each equation with ν_j and add them up, we have

$$\begin{aligned}
\lambda^* \cdot \sum_{j\in\mathcal{V}} \nu_j \cdot \text{sign}(\beta_j) &= \tau \sum_{i\in\mathcal{R}}\sum_{j\in\mathcal{V}} \nu_j x_{ij} - (1-\tau)\sum_{i\in\mathcal{L}}\sum_{j\in\mathcal{V}} \nu_j x_{ij} + \sum_{i\in\mathcal{E}}\theta_i \sum_{j\in\mathcal{V}} \nu_j x_{ij} \\
&= \tau \sum_{i\in\mathcal{R}} \left(\sum_{j\in\mathcal{V}} \nu_j x_{ij} + \nu_0\right) - (1-\tau)\sum_{i\in\mathcal{L}} \left(\sum_{j\in\mathcal{V}} \nu_j x_{ij} + \nu_0\right) \\
&\quad + \sum_{i\in\mathcal{E}}\theta_i \sum_{j\in\mathcal{V}} \left(\nu_j x_{ij} + \nu_0\right) - \tau \sum_{i\in\mathcal{R}} \nu_0 + (1-\tau)\sum_{i\in\mathcal{L}} \nu_0 - \sum_{i\in\mathcal{E}} \theta_i \nu_0 \\
&= \tau \sum_{i\in\mathcal{R}} \left(\sum_{j\in\mathcal{V}} \nu_j x_{ij} + \nu_0\right) - (1-\tau)\sum_{i\in\mathcal{L}} \left(\sum_{j\in\mathcal{V}} \nu_j x_{ij} + \nu_0\right) \\
&= -\frac{\Delta\text{loss}}{\Delta s},
\end{aligned}$$

where we used the facts that

$$\sum_{j\in\mathcal{V}} \nu_j x_{ij} + \nu_0 = 0, \quad \forall i \in \mathcal{E},$$

and

$$\begin{aligned}
\tau \sum_{i\in\mathcal{R}} \nu_0 - (1-\tau)\sum_{i\in\mathcal{L}} \nu_0 + \sum_{i\in\mathcal{E}} \theta_i \nu_0 &= \nu_0 \sum_{i=1}^{n} \theta_i \\
&= 0.
\end{aligned}$$

Also notice that

$$\sum_{j\in\mathcal{V}} \nu_j \cdot \operatorname{sign}(\beta_j) = 1.$$

Hence we conclude

$$\frac{\Delta\text{loss}}{\Delta s} = -\lambda^*.$$

□

A.2 PROOF OF LEMMA 1

For any fixed $s > 0$, suppose $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{L}$ are given, then we have

$$\beta_0 + \sum_{j\in\mathcal{V}} \beta_j x_{kj} = y_k, \quad \forall\, k \in \mathcal{E}$$
$$\sum_{j\in\mathcal{V}} |\beta_j| = s.$$

These can be re-expressed as

$$\begin{pmatrix} 0 & \operatorname{sign}(\boldsymbol{\beta}_{\mathcal{V}})^{\mathsf{T}} \\ \mathbf{1} & \boldsymbol{X} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_{\mathcal{V}} \end{pmatrix} = \begin{pmatrix} s \\ \boldsymbol{y}_{\mathcal{E}} \end{pmatrix},$$

where $\boldsymbol{X}$ is a $n_{\mathcal{E}} \times n_{\mathcal{V}}$ square matrix (since $n_{\mathcal{E}} = n_{\mathcal{V}}$) with entries $x_{kj}, k \in \mathcal{E}, j \in \mathcal{V}$. $\boldsymbol{\beta}_{\mathcal{V}}$ is a vector of length $n_{\mathcal{V}}$, with elements equal to β_j, $j \in \mathcal{V}$, and $\boldsymbol{y}_{\mathcal{E}}$ is a vector of length $n_{\mathcal{E}}$, with elements equal to y_k, $k \in \mathcal{E}$.

Then β_0 and $\boldsymbol{\beta}_{\mathcal{V}}$ can be expressed as

$$\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_{\mathcal{V}} \end{pmatrix} = \boldsymbol{H} \begin{pmatrix} s \\ \boldsymbol{y}_{\mathcal{E}} \end{pmatrix},$$

where

$$\boldsymbol{H} = \begin{pmatrix} 0 & \operatorname{sign}(\boldsymbol{\beta}_{\mathcal{V}})^{\mathsf{T}} \\ \mathbf{1} & \boldsymbol{X} \end{pmatrix}^{-1}.$$

Notice that β_0 and $\boldsymbol{\beta}_{\mathcal{V}}$ are linear in $\boldsymbol{y}_{\mathcal{E}}$.

Now corresponding to the two events listed in Section 2.1, if s is an event point, one of the following conditions has to be satisfied:

- $\exists j \in \mathcal{V}$ s.t. $\beta_j = 0$
- $\exists i \in \mathcal{R} \cup \mathcal{L}$ s.t. $y_i = \beta_0 + \sum_{j\in\mathcal{V}} \beta_j x_{ij}$.

For any fixed $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{L}$, each of the above conditions defines a hyperplane of $\boldsymbol{y}$ in $\mathbb{R}^n$. Taking into account all possible combinations of $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{L}$, the set of $\boldsymbol{y}$ such that s is an event point is a collection of finite number of hyperplanes. □

A.3 PROOF OF LEMMA 2

Let $g(\boldsymbol{\beta}, \boldsymbol{y})$ denote the function $\sum_{i=1}^n \rho_\tau(y_i - f(\boldsymbol{x}_i)) + \lambda\|\boldsymbol{\beta}\|_1$. We note that we consider $g(\cdot, \cdot)$ as a function of $\boldsymbol{\beta}$ and $\boldsymbol{y}$. Let $\boldsymbol{\beta}(\boldsymbol{y}_0)$ be the unique minimizer of $g(\boldsymbol{\beta}, \boldsymbol{y}_0)$, that is, when the response vector $\boldsymbol{y}$ is fixed at $\boldsymbol{y}_0$, and similarly $\boldsymbol{\beta}(\boldsymbol{y}_m)$ be the unique minimizer of $g(\boldsymbol{\beta}, \boldsymbol{y}_m)$, i.e., when the response vector $\boldsymbol{y}$ is fixed at $\boldsymbol{y}_m$.

For any fixed $\boldsymbol{y}_0 \in \mathbb{R}^n$, we wish to show that if a sequence $\boldsymbol{y}_m$ converges to $\boldsymbol{y}_0$, then $\boldsymbol{\beta}(\boldsymbol{y}_m)$ converges to $\boldsymbol{\beta}(\boldsymbol{y}_0)$.

Since $\boldsymbol{\beta}(\boldsymbol{y}_m)$ are bounded, it is equivalent to show that for every converging subsequence, say $\boldsymbol{\beta}(\boldsymbol{y}_{m_k})$, the subsequence converges to $\boldsymbol{\beta}(\boldsymbol{y}_0)$. Suppose $\boldsymbol{\beta}(\boldsymbol{y}_{m_k})$ converges to $\boldsymbol{\beta}_\infty$, we will show $\boldsymbol{\beta}_\infty = \boldsymbol{\beta}(\boldsymbol{y}_0)$.

Let

$$\Delta g(\boldsymbol{\beta}(\boldsymbol{y}), \boldsymbol{y}, \boldsymbol{y}') = g(\boldsymbol{\beta}(\boldsymbol{y}), \boldsymbol{y}) - g(\boldsymbol{\beta}(\boldsymbol{y}), \boldsymbol{y}'),$$

where $g(\boldsymbol{\beta}(\boldsymbol{y}), \boldsymbol{y}')$ is the value of $g(\cdot, \cdot)$ by plugging in $\boldsymbol{\beta}(\boldsymbol{y})$ for the regression coefficients and $\boldsymbol{y}'$ for the response vector. Then we have

$$\begin{aligned} g(\boldsymbol{\beta}(\boldsymbol{y}_0), \boldsymbol{y}_0) &= g(\boldsymbol{\beta}(\boldsymbol{y}_0), \boldsymbol{y}_{m_k}) + \Delta g(\boldsymbol{\beta}(\boldsymbol{y}_0), \boldsymbol{y}_0, \boldsymbol{y}_{m_k}) \\ &\geq g(\boldsymbol{\beta}(\boldsymbol{y}_{m_k}), \boldsymbol{y}_{m_k}) + \Delta g(\boldsymbol{\beta}(\boldsymbol{y}_0), \boldsymbol{y}_0, \boldsymbol{y}_{m_k}) \\ &= g(\boldsymbol{\beta}(\boldsymbol{y}_{m_k}), \boldsymbol{y}_0) + \Delta g(\boldsymbol{\beta}(\boldsymbol{y}_{m_k}), \boldsymbol{y}_{m_k}, \boldsymbol{y}_0) + \Delta g(\boldsymbol{\beta}(\boldsymbol{y}_0), \boldsymbol{y}_0, \boldsymbol{y}_{m_k}). \end{aligned} \tag{A.1}$$

Using the fact that $|a| - |b| \leq |a - b|$ and $\boldsymbol{y}_{m_k} \to \boldsymbol{y}_0$, it is easy to show that for large enough m_k, we have

$$\begin{aligned} &|\Delta g(\boldsymbol{\beta}(\boldsymbol{y}_{m_k}), \boldsymbol{y}_{m_k}, \boldsymbol{y}_0) + \Delta g(\boldsymbol{\beta}(\boldsymbol{y}_0), \boldsymbol{y}_0, \boldsymbol{y}_{m_k})| \\ &\quad = |g(\boldsymbol{\beta}(\boldsymbol{y}_{m_k}), \boldsymbol{y}_{m_k}) - g(\boldsymbol{\beta}(\boldsymbol{y}_{m_k}), \boldsymbol{y}_0) + g(\boldsymbol{\beta}(\boldsymbol{y}_0), \boldsymbol{y}_0) - g(\boldsymbol{\beta}(\boldsymbol{y}_0), \boldsymbol{y}_{m_k})| \\ &\quad \leq |g(\boldsymbol{\beta}(\boldsymbol{y}_{m_k}), \boldsymbol{y}_{m_k}) - g(\boldsymbol{\beta}(\boldsymbol{y}_{m_k}), \boldsymbol{y}_0)| + |g(\boldsymbol{\beta}(\boldsymbol{y}_0), \boldsymbol{y}_0) - g(\boldsymbol{\beta}(\boldsymbol{y}_0), \boldsymbol{y}_{m_k})| \\ &\quad \leq c_1\|\boldsymbol{y}_0 - \boldsymbol{y}_{m_k}\|_1 + c_2\|\boldsymbol{y}_0 - \boldsymbol{y}_{m_k}\|_1 \\ &\quad \leq c\|\boldsymbol{y}_0 - \boldsymbol{y}_{m_k}\|_1, \end{aligned} \tag{A.2}$$

where $c_1 > 0$, $c_2 > 0$, and $c > 0$ are constants. Furthermore, using $\boldsymbol{y}_{m_k} \to \boldsymbol{y}_0$ and $\boldsymbol{\beta}(\boldsymbol{y}_{m_k}) \to \boldsymbol{\beta}_\infty$, we reduce (A.1) to

$$g(\boldsymbol{\beta}(\boldsymbol{y}_0), \boldsymbol{y}_0) \geq g(\boldsymbol{\beta}_\infty, \boldsymbol{y}_0).$$

Since $\boldsymbol{\beta}(\boldsymbol{y}_0)$ is the unique minimizer of $g(\boldsymbol{\beta}, \boldsymbol{y}_0)$, we have $\boldsymbol{\beta}_\infty = \boldsymbol{\beta}(\boldsymbol{y}_0)$.

Similarly, one can prove that for any fixed $s > 0$, $\boldsymbol{\theta}(\boldsymbol{y})$ is also a continuous function of $\boldsymbol{y}$. □

A.4 PROOF OF LEMMA 3

For any fixed $s > 0$ and any fixed $\boldsymbol{y}_0 \in \mathbb{R}^n \backslash \mathcal{N}_s$, since $\mathbb{R}^n \backslash \mathcal{N}_s$ is an open set, we can always find a small enough $\epsilon > 0$, such that $\text{Ball}(\boldsymbol{y}_0, \epsilon) \subset \mathbb{R}^n \backslash \mathcal{N}_s$. So s is not an event point for any $\boldsymbol{y} \in \text{Ball}(\boldsymbol{y}_0, \epsilon)$.

We claim that if ϵ is small enough, the sets $\mathcal{V}$, $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{E}$ stay the same for all $\boldsymbol{y} \in \text{Ball}(\boldsymbol{y}_0, \epsilon)$.

Consider $\boldsymbol{y}$ and $\boldsymbol{y}_0$. Let $\mathcal{V}_{\boldsymbol{y}}, \mathcal{R}_{\boldsymbol{y}}, \mathcal{L}_{\boldsymbol{y}}, \mathcal{E}_{\boldsymbol{y}}, \mathcal{V}_0, \mathcal{R}_0, \mathcal{L}_0, \mathcal{E}_0$ denote the corresponding sets, and $\boldsymbol{\theta}^{\boldsymbol{y}}, \boldsymbol{\beta}^{\boldsymbol{y}}, f^{\boldsymbol{y}}, \boldsymbol{\theta}^0, \boldsymbol{\beta}^0, f^0$ denote the corresponding fits.

For any $i \in \mathcal{E}_0$, we have $-(1-\tau) < \theta_i^0 < \tau$. Therefore, by continuity, we also have $-(1-\tau) < \theta_i^{\boldsymbol{y}} < \tau,\ i \in \mathcal{E}_0$ for $\boldsymbol{y}$ close enough to $\boldsymbol{y}_0$; or equivalently, $\mathcal{E}_0 \subseteq \mathcal{E}_{\boldsymbol{y}}, \forall \boldsymbol{y} \in \text{Ball}(\boldsymbol{y}_0, \epsilon)$ for small enough ϵ.

Similarly, for any $i \in \mathcal{R}_0$, since $y_i^0 - f^0(\boldsymbol{x}_i) > 0$, again by continuity, we have $y_i - f^{\boldsymbol{y}}(\boldsymbol{x}_i) > 0$ for $\boldsymbol{y}$ close enough to $\boldsymbol{y}_0$; or equivalently, $\mathcal{R}_0 \subseteq \mathcal{R}_{\boldsymbol{y}}, \forall \boldsymbol{y} \in \text{Ball}(\boldsymbol{y}_0, \epsilon)$ for small enough ϵ. The same applies to $\mathcal{L}_0$ and $\mathcal{L}_{\boldsymbol{y}}$ as well.

Overall, we then have $\mathcal{E}_0 = \mathcal{E}_{\boldsymbol{y}}$, $\mathcal{R}_0 = \mathcal{R}_{\boldsymbol{y}}$ and $\mathcal{L}_0 = \mathcal{L}_{\boldsymbol{y}}$ for all $\boldsymbol{y} \in \text{Ball}(\boldsymbol{y}_0, \epsilon)$ when ϵ is small enough.

Regarding $\mathcal{V}_0$, by definition, $\beta_j^0 \neq 0$ for any $j \in \mathcal{V}_0$. By continuity, we have $\beta_j^{\boldsymbol{y}} \neq 0,\ j \in \mathcal{V}_0$ for $\boldsymbol{y}$ close enough to $\boldsymbol{y}_0$. Therefore, $\mathcal{V}_0 \subseteq \mathcal{V}_{\boldsymbol{y}}, \forall \boldsymbol{y} \in \text{Ball}(\boldsymbol{y}_0, \epsilon)$ for small enough ϵ. On the other hand, we have

$$\lambda^* \cdot \text{sign}(\beta_j) = \sum_{i=1}^{n} \theta_i x_{ij}, \quad \forall j \in \mathcal{V}_0,$$

and

$$\lambda^* \cdot \text{sign}(\beta_j) > \sum_{i=1}^{n} \theta_i x_{ij}, \quad \forall j \notin \mathcal{V}_0.$$

Using continuity again, we have $\mathcal{V}_{\boldsymbol{y}} \subseteq \mathcal{V}_0$. Therefore, we have $\mathcal{V}_0 = \mathcal{V}_{\boldsymbol{y}}, \forall \boldsymbol{y} \in \text{Ball}(\boldsymbol{y}_0, \epsilon)$ for small enough ϵ. □

A.5 PROOF OF THEOREM 2

Using Lemma 3, we know that there exists $\epsilon > 0$, such that for all $\boldsymbol{y} \in \text{Ball}(\boldsymbol{y}, \epsilon)$, the sets $\mathcal{V}$, $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{E}$ stay the same. This implies that for points in $\mathcal{E}$, we have

$$\frac{\partial f(\boldsymbol{x}_i)}{\partial y_i} = 1, \quad i \in \mathcal{E}.$$

Furthermore, since $\boldsymbol{\beta}_{\mathcal{V}}$ is determined by $\boldsymbol{y}_{\mathcal{E}}$, hence for points in $\mathcal{R}$ and $\mathcal{L}$, we have

$$\frac{\partial f(\boldsymbol{x}_i)}{\partial y_i} = 0, \quad i \in \mathcal{R} \cup \mathcal{L}.$$

Overall, we have

$$\sum_{i=1}^{n} \frac{\partial f(\boldsymbol{x}_i)}{\partial y_i} = |\mathcal{E}|.$$

□

ACKNOWLEDGMENTS

We would like to thank two reviewers, an Associate Editor, and Editor L. Tierney for their thoughtful and useful comments. Li and Zhu are partially supported by grant DMS-0505432 and grant DMS-0705532 from the National Science Foundation.

[Received May 2006. Revised July 2007.]

REFERENCES

Chen, S., Donoho, D., and Saunders, M. (1998), "Atomic Decomposition by Basis Pursuit," *SIAM Journal of Scientific Computing*, 20, 33–61.

Cole, T., and Green, P. (1992), "Smoothing Reference Centile Curves: The LMS Method and Penalized Likelihood," *Statistics in Medicine*, 11, 1305–1319.

Efron, B. (1986), "How Biased is the Apparent Error Rate of a Prediction Rule?," *Journal of the American Statistical Association*, 81, 461–470.

——— (2004), "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation," *Journal of the American Statistical Association*, 99, 619–632.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics*, 32, 407–451.

Friedman, J., Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004), Discussion of "Consistency in Boosting" by W. Jiang, G. Lugosi, N. Vayatis, and T. Zhang, *The Annals of Statistics*, 32, 102–107.

Heagerty, P., and Pepe, M. (1999), "Semiparametric Estimation of Regression Quantiles with Application to Standardizing Weight for Height and Age in U.S. Children," *Journal of the Royal Statistical Society*, Series C, 48, 533–551.

Hendricks, W., and Koenker, R. (1992), "Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity," *Journal of the American Statistical Association*, 87, 58–68.

Hoerl, A., and Kennard, R. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R., and Geling, R. (2001), "Reappraising Medfly Longevity: A Quantile Regression Survival Analysis," *Journal of the American Statistical Association*, 96, 458–468.

Koenker, R., and Hallock, K. (2001), "Quantile Regression," *Journal of Economic Perspectives*, 15, 143–156.

Koenker, R., Ng, P., and Portnoy, S. (1994), "Quantile Smoothing Splines," *Biometrika*, 81, 673–680.

Leng, C., Lin, Y., and Wahba, G. (2006), "A Note on the Lasso and Related Procedures in Model Selection," *Statistica Sinica*, 16, 1273–1284.

Li, Y., Liu, Y., and Zhu, J. (2007), "Quantile Regression in Reproducing Kernel Hilbert Spaces," *Journal of the American Statistical Association*, 102, 255–268.

Meyer, M., and Woodroofe, M. (2000), "On the Degrees of Freedom in Shape-Restricted Regression," *The Annals of Statistics*, 28, 1083–1104.

Nychka, D., Gray, G., Haaland, P., Martin, D., and O'Connell, M. (1995), "A Nonparametric Regression Approach to Syringe Grading for Quality Imporovement," *Journal of the American Statistical Association*, 90, 1171–1178.

Redfern, C., Degtyarev, M., Kwa, A., Salomonis, N., Cotte, N., Nanevicz, T., Fidelman, N., Desai, K., Vranizan, K., Lee, E., Coward, P., Shah, N., Warrington, J., Fishman, G., Bernstein, D., Baker, A., and Conklin, B. (2000), "Conditional Expression of a gi-coupled Receptor Causes Ventricular Conduction Delay and a Lethal Cardiomyopathy," *PNAS*, 97, 4826–4831.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Segal, M., Kam, D., and Bruce, C. (2003), "Regression Approaches for Microarray Data Analysis," *Journal of Computational Biology*, 10, 961–980.

Stein, C. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 9, 1135–1151.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288.

Yang, S. (1999), "Censored Median Regression using Weighted Empirical Survival and Hazard Functions," *Journal of the American Statistical Association*, 94, 137–145.

Ye, J. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection," *Journal of the American Statistical Association*, 93, 120–131.

Yuan, M. (2006), "GACV for Quantile Smoothing Splines," *Computational Statistics and Data Analysis*, 5, 813–829.

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320.