Bayesian Analysis (2010) **5**, Number 3, pp. 533–556

Bayesian Regularized Quantile Regression

Qing Li*, Ruibin Xi† and Nan Lin‡

Abstract. Regularization, e.g. lasso, has been shown to be effective in quantile regression in improving the prediction accuracy (Li and Zhu 2008; Wu and Liu 2009). This paper studies regularization in quantile regressions from a Bayesian perspective. By proposing a hierarchical model framework, we give a generic treatment to a set of regularization approaches, including lasso, group lasso and elastic net penalties. Gibbs samplers are derived for all cases. This is the first work to discuss regularized quantile regression with the group lasso penalty and the elastic net penalty. Both simulated and real data examples show that Bayesian regularized quantile regression methods often outperform quantile regression without regularization and their non-Bayesian counterparts with regularization.

Keywords: Quantile regression; Regularization; Gibbs sampler; Bayesian analysis; Lasso; Elastic net; Group lasso

1 Introduction

Quantile regression (Koenker and Bassett 1978) has gained increasing popularity as it provides richer information than the classic mean regression. Suppose that we have a sample $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$. Then the linear quantile regression model for the θth quantile $(0 < \theta < 1)$ is $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i, i = 1, \ldots, n$, where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T \in \mathbb{R}^p$ and u_i's are independent with their θth quantiles equal to 0. It can be shown that the coefficients $\boldsymbol{\beta}$ can be estimated consistently by the solution to the following minimization problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\theta \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right), \tag{1}$$

where $\rho_\theta(\cdot)$ is the check loss function

$$\rho_\theta(t) = \begin{cases} \theta t, & \text{if } t \geq 0, \\ -(1-\theta)t, & \text{if } t < 0. \end{cases}$$

Although the asymptotic theory for quantile regression has been well developed (Koenker and Bassett 1978; Koenker 2005), a Bayesian approach enables exact inference even when the sample size is small. Yu and Moyeed (2001) proposed a Bayesian formulation of quantile regression using the skewed Laplace distribution for the errors

*Department of Mathematics, Washington University in St. Louis, St. Louis, MO, mailto:qli@math.wustl.edu

†Center for Biomedical Informatics, Harvard Medical School, Cambridge, MA, mailto:Ruibin_Xi@hms.harvard.edu

‡Department of Mathematics, Washington University in St. Louis, St. Louis, MO, mailto:nlin@math.wustl.edu

 DOI:10.1214/10-BA521

and sampling $\boldsymbol{\beta}$ from its posterior distribution using a random walk Metropolis-Hastings algorithm. Similar formulations were also employed by Tsionas (2003) and Kozumi and Kobayashi (2009), both of which developed Gibbs samplers to estimate their models. Reed and Yu (2009) considered a similar scale-mixture expression of the skewed Laplace distribution as in Kozumi and Kobayashi (2009) and derived efficient Gibbs samplers. Reed et al. (2009) studied the stochastic search variable selection (SSVS) algorithm based on the scale-mixture expression in Reed and Yu (2009). Kottas and Krnjajić (2009) extended this idea to Dirichlet process scale mixture of skewed Laplace distributions and another scale mixture of uniform distributions. The special case of median regression was discussed by Walker and Mallick (1999), Kottas and Gelfand (2001) and Hanson and Johnson (2002). They modeled the error distribution as mixture distributions based on either the Dirichlet process or the Pólya tree. Hjort and Walker (2009) introduced the quantile pyramids and discussed briefly on its application to Bayesian quantile regression. Geraci and Bottai (2007) and Reich et al. (2009) considered Bayesian quantile regression models for clustered data.

One crucial problem in building a quantile regression model is the selection of predictors. The prediction accuracy can often be improved by choosing an appropriate subset of predictors. Also, in practice, it is often desired to identify a smaller subset of predictors from a large set of predictors to obtain better interpretation. There has been active research on sparse representation of linear regression. Tibshirani (1996) introduced the least absolute shrinkage and selection operator (lasso) technique which can simultaneously perform variable selection and parameter estimation. The lasso estimate is an L_1-regularized least squares estimate, i.e., the lasso estimate is the solution to $\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_1$, for some $\lambda \geq 0$ and $\|\boldsymbol{\beta}\|_1 = \sum_{k=1}^{p}|\beta_k|$. Several constraints and corresponding improvement of the lasso are as follows. When categorical predictors are present in the regression model, the lasso is not satisfactory since it only selects individual dummy variables instead of the whole predictor. To solve this problem, Yuan and Lin (2006) introduced the group lasso by generalizing the lasso penalty. Another extension of the lasso is the elastic net (Zou and Hastie 2005), which has an improved performance for correlated predictors and can select more than n variables in the case of $p > n$. Other related approaches include the smoothly clipped absolute deviation (SCAD) model (Fan and Li 2001) and the fused lasso (Tibshirani et al. 2005).

The first use of regularization in quantile regression is made by Koenker (2004), which put the lasso penalty on the random effects in a mixed-effect quantile regression model to shrink the random effects towards zero. Wang et al. (2007) considered the least absolute deviance (LAD) estimate with adaptive lasso penalty (LAD-lasso) and proved its oracle property. Recently, Li and Zhu (2008) considered quantile regression with the lasso penalty and developed its piecewise linear solution path. Wu and Liu (2009) demonstrated the oracle properties of the SCAD and adaptive lasso regularized quantile regression.

Our goal is to develop a Bayesian framework for regularization in linear quantile regression. For linear regression, Bae and Mallick (2004) and Park and Casella (2008) treated the lasso from a Bayesian perspective, and proposed hierarchical models which can be solved efficiently through Gibbs samplers. These works shed light on incorpo-

rating the regularization methods into the Bayesian quantile regression framework. In this paper, we consider different penalties including lasso, group lasso and elastic net penalties. Gibbs samplers are developed for these three types of regularized quantile regression. As demonstrated later by simulation studies, these Bayesian regularized quantile regression methods provide more accurate estimates and better prediction accuracy than their non-Bayesian peers.

The rest of the paper is organized as follows. In Section 2, we introduce Bayesian regularized quantile regression with lasso, group lasso and elastic net penalties, and derive the corresponding Gibbs samplers. Simulation studies are then presented in Sections 3 followed by a real data example in Section 4. Discussions and conclusions are put in Section 5. An appendix contains technical proofs and derivations.

2 Bayesian formulation of the regularized quantile regression

Assume that $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + u_i$, $i = 1, \ldots, n$, with u_i being i.i.d. random variables from the skewed Laplace distribution with density

$$f(u \mid \tau) = \theta(1-\theta)\tau \exp[-\tau\rho_\theta(u)]. \tag{2}$$

Then the joint distribution of $\mathbf{y} = (y_1, \cdots, y_n)$ given $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ is

$$f(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \tau) = \theta^n(1-\theta)^n\tau^n \exp\left\{-\tau\sum_{i=1}^{n}\rho_\theta(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\right\}. \tag{3}$$

Hence, maximizing the likelihood (3) is equivalent to minimizing (1). Recently, Kozumi and Kobayashi (2009) proved that the skewed Laplace distribution (2) can be viewed as a mixture of an exponential and a scaled normal distribution. More specifically, we have the following lemma.

Lemma 1. *Suppose that v is a standard exponential random variable and z is a standard normal random variable. For $\theta \in (0,1)$, denote*

$$\xi_1 = \frac{1-2\theta}{\theta(1-\theta)} \quad \textit{and} \quad \xi_2 = \sqrt{\frac{2}{\theta(1-\theta)}}.$$

It follows that the random variable $u = \xi_1 v + \xi_2\sqrt{v}z$ follows the skewed distribution (2) with $\tau = 1$.

From Lemma 1, the response y_i can be equivalently written as $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \tau^{-1}\xi_1 v_i + \tau^{-1}\xi_2\sqrt{v_i}z_i$, where v_i and z_i follow the standard exponential distribution, Exp(1), and the standard normal distribution, $N(0,1)$, respectively. Let $\tilde{v}_i = \tau^{-1}v_i$, then it follows the exponential distribution Exp (τ^{-1}), i.e. the density function of $\tilde{v}_i$ is

$f(\tilde{v}_i \mid \tau) = \tau \exp(-\tau \tilde{v}_i)$. Denote $\tilde{\mathbf{v}} = (\tilde{v}_1, \cdots, \tilde{v}_n)$ and $\mathbf{z} = (z_1, \cdots, z_n)$. Then, we have the hierarchical model

$$\begin{aligned} y_i &= \mathbf{x}_i^T \boldsymbol{\beta} + \xi_1 \tilde{v}_i + \tau^{-1/2} \xi_2 \sqrt{\tilde{v}_i} z_i, \\ \tilde{\mathbf{v}} \mid \tau &\sim \prod_{i=1}^{n} \tau \exp(-\tau \tilde{v}_i), \\ \mathbf{z} &\sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z_i^2\right). \end{aligned} \tag{4}$$

The lasso, group lasso and elastic net estimates are all regularized least squares estimates and the differences among them are only at their penalty terms. Specifically, they are all solutions to the following form of minimization problem $\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda_1 h_1(\boldsymbol{\beta}) + \lambda_2 h_2(\boldsymbol{\beta})$, for some $\lambda_1,\ \lambda_2 \geq 0$ and penalty functions $h_1(\cdot)$ and $h_2(\cdot)$. The lasso corresponds to $\lambda_1 = \lambda$, $\lambda_2 = 0$, $h_1(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ and $h_2(\boldsymbol{\beta}) = 0$. Suppose that the predictors are classified into G groups and $\boldsymbol{\beta}_g$ is the coefficient vector of the gth group. Denote $\|\boldsymbol{\beta}_g\|_{\mathbf{K}_g} = \left(\boldsymbol{\beta}_g^T \mathbf{K}_g \boldsymbol{\beta}_g\right)^{1/2}$ for positive definite matrices $\mathbf{K}_g$ $(g = 1, \cdots, G)$. Then, the group lasso corresponds to $\lambda_1 = \lambda$, $\lambda_2 = 0$, $h_1(\boldsymbol{\beta}) = \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_{\mathbf{K}_g}$ and $h_2(\boldsymbol{\beta}) = 0$. The elastic net corresponds to $\lambda_1, \lambda_2 > 0$, $h_1(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ and $h_2(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2 = \sum_{k=1}^{p} \beta_k^2$. Similarly, we form the minimization problem for regularized quantile regression as

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\theta(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda_1 h_1(\boldsymbol{\beta}) + \lambda_2 h_2(\boldsymbol{\beta}). \tag{5}$$

That is, we replace the squared error loss by the the check loss, while keeping the corresponding penalty terms unchanged. Starting from (4) we can show that, by introducing suitable priors on $\boldsymbol{\beta}$, the solution to (5) is equivalent to the maximum a posteriori (MAP) estimate in the Bayesian formulation. In the following three subsections, we will discuss (5) with lasso, elastic net and group lasso penalties separately. For each penalty, we present the Bayesian hierarchical model and derive its corresponding Gibbs sampler.

2.1 Quantile regression with the lasso penalty

We first consider quantile regression with the lasso penalty. The lasso regularized quantile regression (Li and Zhu 2008) is given by

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\theta(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda \sum_{k=1}^{p} |\beta_k|. \tag{6}$$

If we put a Laplace prior $\pi(\boldsymbol{\beta} \mid \tau, \lambda) = (\tau\lambda/2)^p \exp\{-\tau\lambda \sum_{k=1}^{p} |\beta_k|\}$ and assume that the residuals u_i come from the skewed Laplace distribution (2), then the posterior distribution of $\boldsymbol{\beta}$ is

$$f(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \tau, \lambda) \propto \exp\left\{-\tau \sum_{i=1}^{n} \rho_\theta(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) - \tau\lambda \sum_{k=1}^{p} |\beta_k|\right\}. \tag{7}$$

So minimizing (6) is equivalent to maximizing the likelihood (7). For any $a \geq 0$, we have the following equality (Andrews and Mallows 1974),

$$\frac{a}{2}e^{-a|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left(-\frac{z^2}{2s}\right) \frac{a^2}{2} \exp\left(\frac{-a^2}{2}s\right) ds. \tag{8}$$

Let $\eta = \tau\lambda$. Then, the Laplace prior on $\boldsymbol{\beta}$ can be written as

$$\begin{aligned} \pi(\boldsymbol{\beta} \mid \tau, \lambda) &= \prod_{k=1}^{p} \frac{\eta}{2} \exp\{-\eta|\beta_k|\} \\ &= \prod_{k=1}^{p} \int_0^\infty \frac{1}{\sqrt{2\pi s_k}} \exp\left(-\frac{\beta_k^2}{2s_k}\right) \frac{\eta^2}{2} \exp\left(\frac{-\eta^2}{2}s_k\right) ds_k. \end{aligned}$$

Denote $\mathbf{s} = (s_1, \cdots, s_p)$. We further put Gamma priors on the parameter τ and η^2 and have the following Bayesian hierarchical model.

$$\begin{aligned} y_i &= \mathbf{x}_i^T\boldsymbol{\beta} + \xi_1\tilde{v}_i + \xi_2\tau^{-1/2}\sqrt{\tilde{v}_i}z_i, \\ \tilde{\mathbf{v}} \mid \tau &\sim \prod_{i=1}^{n} \tau \exp\left(-\tau\tilde{v}_i\right), \\ \mathbf{z} &\sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z_i^2\right), \\ \boldsymbol{\beta},\ \mathbf{s} \mid \eta^2 &\sim \prod_{k=1}^{p} \frac{1}{\sqrt{2\pi s_k}} \exp\left(-\frac{\beta_k^2}{2s_k}\right) \prod_{k=1}^{p} \frac{\eta^2}{2} \exp\left(-\frac{\eta^2}{2}s_k\right), \\ \tau,\ \eta^2 &\sim \tau^{a-1}\exp(-b\tau)\left(\eta^2\right)^{c-1}\exp(-d\eta^2). \end{aligned} \tag{9}$$

If $a = b = c = d = 0$, the priors on τ and η^2 become noninformative priors.

As for the Gibbs sampler, the full conditional distribution of β_k is a normal distribution and those of τ and η^2 are Gamma distributions. And the full conditional distribution of $\tilde{v}_i$ and s_k are generalized inverse Gaussian distributions (Jørgensen 1982). The details of the Gibbs sampler and full conditional distributions are given in Appendix A.

2.2 Quantile regression with the elastic net penalty

We consider quantile regression with the elastic net penalty, which solves the following

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\theta(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + \lambda_1 \sum_{k=1}^{p} |\beta_k| + \lambda_2 \sum_{k=1}^{p} \beta_k^2. \tag{10}$$

Let $\eta_1 = \tau\lambda_1$ and $\eta_2 = \tau\lambda_2$ and put the prior of β_k as

$$\pi(\beta_k \mid \eta_1, \eta_2) = C(\eta_1, \eta_2)\frac{\eta_1}{2}\exp\{-\eta_1|\beta_k| - \eta_2\beta_k^2\}, \tag{11}$$

where $C(\eta_1, \eta_2)$ is a normalizing constant depending on η_1 and η_2. The posterior of $\boldsymbol{\beta}$ becomes

$$f(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \tau, \lambda_1, \lambda_2) \propto \exp\left\{-\tau \sum_{i=1}^{n} \rho_\theta(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) - \tau\lambda_1 \sum_{k=1}^{p} |\beta_k| - \tau\lambda_2 \sum_{k=1}^{p} \beta_k^2\right\}. \quad (12)$$

Maximizing the posterior distribution (12) is thus equivalent to minimizing (10). Calculation of the constant $C(\eta_1, \eta_2)$ is in Appendix B. Let $\tilde{\eta}_1 = \eta_1^2/(4\eta_2)$. Putting Gamma priors on τ, $\tilde{\eta}_1$ and η_2, we then have the following hierarchical model.

$$\begin{aligned}
y_i &= \mathbf{x}_i^T \boldsymbol{\beta} + \xi_1 \tilde{v}_i + \xi_2 \tau^{-1/2} \sqrt{\tilde{v}_i} z_i, \\
\tilde{\mathbf{v}} \mid \tau &\sim \prod_{i=1}^{n} \tau \exp(-\tau \tilde{v}_i), \\
\mathbf{z} &\sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z_i^2\right), \\
\beta_k \mid t_k, \eta_2 &\overset{ind.}{\sim} \frac{1}{\sqrt{2\pi(t_k - 1)/(2\eta_2 t_k)}} \exp\left\{-\frac{1}{2}\left(\frac{t_k - 1}{2\eta_2 t_k}\right)^{-1} \beta_k^2\right\}, \quad (13) \\
t_k | \tilde{\eta}_1 &\overset{ind.}{\sim} \Gamma^{-1}(1/2,\ \tilde{\eta}_1) t_k^{-1/2} \tilde{\eta}_1^{1/2} \exp\{-\tilde{\eta}_1 t_k\} I(t_k > 1), \\
\tau,\ \tilde{\eta}_1,\ \eta_2 &\sim \tau^{a-1} \exp(-b\tau)\ \tilde{\eta}_1^{c_1 - 1} \exp(-d_1 \tilde{\eta}_1)\ \eta_2^{c_2 - 1} \exp(-d_2 \eta_2),
\end{aligned}$$

where $a, b, c_1, c_2, d_1, d_2 \geq 0$, $I(\cdot)$ is the indicator function and $\Gamma(\cdot, \cdot)$ is the upper incomplete gamma function.

Appendix B gives the details of the full conditionals for the Gibbs sampler. The full conditional distributions are all common distributions except the full conditional of $\tilde{\eta}_1$. The full conditional distribution of $\tilde{\eta}_1$ is

$$f(\tilde{\eta}_1 \mid \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{t}, \tau, \eta_2) \propto \Gamma^{-p}(1/2,\ \tilde{\eta}_1) \tilde{\eta}_1^{p/2 + c_1 - 1} \exp\left\{-\tilde{\eta}_1 \left[d_1 + \sum_{k=1}^{p} t_k\right]\right\}.$$

Since it is difficult to directly sample from $f(\tilde{\eta}_1 |\ \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{t}, \tau, \eta_2)$, we use a Metropolis-Hastings within Gibbs algorithm. The proposal distribution for the Metropolis-Hastings step is $q(\tilde{\eta}_1 |\ \mathbf{t}) \propto \tilde{\eta}_1^{p + c_1 - 1} \exp\{-\tilde{\eta}_1 [d_1 + \sum_{k=1}^{p} (t_k - 1)]\}$. Notice that

$$\lim_{\tilde{\eta}_1 \to \infty} \frac{\tilde{\eta}_1^{1/2} \exp(\tilde{\eta}_1)}{\Gamma^{-1}(1/2, \tilde{\eta}_1)} = 1$$

and hence $\lim_{\tilde{\eta}_1 \to \infty} f(\tilde{\eta}_1 |\ \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{t}, \tau, \eta_2) q^{-1}(\tilde{\eta}_1 |\ \mathbf{t})$ exists and equals to some positive constant. So the tail behaviors of $q(\tilde{\eta}_1 |\ \mathbf{t})$ and $f(\tilde{\eta}_1 |\ \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{t}, \tau, \eta_2)$ are similar. At each iteration of the Gibbs sampler, we sample from $f(\tilde{\eta}_1 \mid \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{t}, \tau, \eta_2)$ using a one-step Metropolis-Hastings sampling.

3 Simulation studies

In this section, we carry out Monte Carlo simulations to study the performance of Bayesian regularized quantile regression with comparison to some non-Bayesian approaches. The methods in the comparison include:

- Bayesian regularized quantile regressions with the lasso penalty (BQR.L), the elastic net penalty (BQR.EN) and the group lasso penalty (BQR.GL).
- Regularized mean regression methods including the lasso and the elastic net (EN).
- The standard quantile regression (QR).

The data in the simulation studies are generated by

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i, \ i = 1, \ldots, n, \tag{17}$$

where u_i's have the θth quantile equal to 0. We first consider models with homogeneous errors and then those with heterogenous errors.

3.1 Independent and identically distributed random errors

For the i.i.d. random errors, we consider five simulation studies. The first three are similar to those in Li and Zhu (2008). The fourth simulation study corresponds to the case where elastic net regularization is more proper and the fifth corresponds to the case where group lasso regularization is recommended.

- Simulation 1: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$, which corresponds to the sparse case.
- Simulation 2: $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$, which corresponds to dense case.
- Simulation 3: $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)$, which corresponds to the very sparse case.
- Simulation 4: $\boldsymbol{\beta} = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{10}, \underbrace{3, \ldots, 3}_{15})$, which corresponds to the case with more predictors than the sample size.
- Simulation 5: $\boldsymbol{\beta} = ((-1.2, 1.8, 0), (0, 0, 0), (0.5, 1, 0), (0, 0, 0), (1, 1, 0))$, which corresponds to the case with group structures in the predictors.

Table 1: MMADs for Simulation 1. In the parentheses are standard deviations of the MMADs obtained by 500 bootstrap resampling. The bold numbers correspond to the smallest MMAD in each category.

θ	Method	Error Distribution			
		normal	normal mixture	Laplace	Laplace mixture
$\theta = 0.5$	BQR.L	**0.939** (0.048)	1.461 (0.089)	**1.090** (0.103)	**1.615** (0.149)
	BQR.EN	1.051 (0.044)	1.521 (0.075)	1.174 (0.076)	1.642 (0.095)
	lasso	1.089 (0.072)	1.510 (0.098)	1.384 (0.142)	2.026 (0.167)
	EN	1.169 (0.066)	1.628 (0.102)	1.649 (0.098)	1.921 (0.133)
	QR	1.262 (0.048)	1.812 (0.073)	1.636 (0.091)	1.965 (0.085)
	QR-L	1.102 (0.038)	**1.458** (0.090)	1.228 (0.081)	1.719 (0.125)
$\theta = 0.3$	BQR.L	**0.967** (0.059)	**1.618** (0.073)	**1.117** (0.102)	**1.952** (0.140)
	BQR.EN	0.973 (0.066)	1.623 (0.102)	1.180 (0.091)	2.006 (0.076)
	lasso	1.135 (0.071)	1.762 (0.107)	1.436 (0.082)	2.144 (0.131)
	EN	1.280 (0.063)	1.876 (0.091)	1.436 (0.105)	2.025 (0.135)
	QR	1.213 (0.070)	1.712 (0.069)	1.406 (0.059)	2.157 (0.075)
	QR-L	1.035 (0.071)	1.745 (0.109)	1.223 (0.104)	2.002 (0.115)
$\theta = 0.1$	BQR.L	1.079 (0.060)	2.057 (0.143)	**1.097** (0.103)	2.886 (0.214)
	BQR.EN	**1.078** (0.049)	2.099 (0.152)	1.168 (0.109)	2.959 (0.209)
	lasso	1.076 (0.052)	2.652 (0.244)	1.361 (0.091)	3.136 (0.187)
	EN	1.185 (0.074)	2.591 (0.212)	1.445 (0.103)	3.049 (0.261)
	QR	1.318 (0.036)	**1.873** (0.073)	1.453 (0.088)	**1.913** (0.149)
	QR-L	1.460 (0.101)	2.310 (0.150)	2.035 (0.114)	3.308 (0.215)

In the first three simulation studies, the rows of $\mathbf{X}$ in (17) are generated independently from $N(\mathbf{0}, \mathbf{\Sigma})$, where the (i,j)th element of $\mathbf{\Sigma}$ is $0.5^{|i-j|}$. In Simulation 4, we first generate Z_1 and Z_2 independently from $N(0,1)$. Then let $x_j = Z_1 + \epsilon_j$, $j = 1, \ldots, 10$, $x_j \sim N(0,1)$, $j = 11, \ldots, 20$, $x_j = Z_2 + \epsilon_j$, $j = 21, \ldots, 30$, where $\epsilon_j \sim N(0, 0.01)$, $j = 1, \ldots, 10, 21, \ldots, 30$. In Simulation 5, we first simulate a latent variable, $\mathbf{Z} = (Z_1, \ldots, Z_5)^T$, from $N(\mathbf{0}, \mathbf{\Sigma})$, where the (i,j)th element of $\mathbf{\Sigma}$ is $0.5^{|i-j|}$. Then each Z_j is trichotomized as 0, 1 or 2, depending on whether it is smaller than $\Phi^{-1}(1/3)$, between $\Phi^{-1}(1/3)$ and $\Phi^{-1}(2/3)$, or larger than $\Phi^{-1}(2/3)$. Here $\Phi(\cdot)$ is the cumulative distribution function for standard normal distribution. The rows of $\mathbf{X}$ are given by $(I(Z_1 = 0), I(Z_1 = 1), I(Z_1 = 2), \ldots, I(Z_5 = 0), I(Z_5 = 1), I(Z_5 = 2))$.

Within each simulation study, we consider four different choices for the distribution of u_i's.

- The first choice is a normal distribution $N(\mu, \sigma^2)$, with μ chosen so that the θth quantile is 0. σ^2 is set as 9.
- The second choice is a mixture of two normal distributions, $0.1N(\mu, \sigma_1^2) + 0.9N(\mu, \sigma_2^2)$, with μ chosen so that the θth quantile is 0. σ_1^2 is set as 1 and σ_2^2 is set as 5.

Table 2: MMADs for Simulation 4. In the parentheses are standard deviations of the MMADs obtained by 500 bootstrap resampling. The bold numbers correspond to the smallest MMAD in each category.

θ	Method	Error Distribution			
		normal	normal mixture	Laplace	Laplace mixture
$\theta = 0.5$	BQR.L	3.736 (0.116)	5.613 (0.202)	4.907 (0.209)	6.941 (0.128)
	BQR.EN	3.399 (0.143)	**5.114** (0.177)	4.587 (0.122)	**6.007** (0.134)
	lasso	4.707 (0.268)	6.512 (0.388)	6.187 (0.365)	7.861 (0.196)
	EN	**3.172** (0.193)	5.329 (0.224)	**4.326** (0.231)	6.490 (0.171)
	QR	4.812 (0.300)	7.059 (0.229)	6.017 (0.263)	10.263 (0.552)
	QR-L	7.480 (0.289)	7.766 (0.278)	7.899 (0.219)	8.581 (0.201)
$\theta = 0.3$	BQR.L	3.840 (0.122)	6.141 (0.162)	5.135 (0.228)	7.012 (0.220)
	BQR.EN	3.660 (0.189)	5.869 (0.172)	**4.636** (0.239)	**6.288** (0.145)
	lasso	4.822 (0.399)	6.644 (0.227)	6.242 (0.157)	8.184 (0.274)
	EN	**3.490** (0.155)	**5.645** (0.151)	4.841 (0.277)	6.585 (0.338)
	QR	4.613 (0.155)	9.088 (0.401)	6.556 (0.411)	10.476 (0.388)
	QR-L	6.641 (0.148)	7.726 (0.168)	7.180 (0.104)	8.366 (0.178)
$\theta = 0.1$	BQR.L	3.920 (0.194)	7.861 (0.204)	5.474 (0.238)	9.217 (0.309)
	BQR.EN	3.773 (0.147)	**7.233** (0.197)	**4.987** (0.198)	**8.374** (0.183)
	lasso	4.742 (0.301)	8.654 (0.225)	6.479 (0.310)	9.821 (0.265)
	EN	**3.577** (0.132)	8.046 (0.334)	4.836 (0.262)	9.581 (0.436)
	QR	4.793 (0.288)	13.553 (0.704)	6.669 (0.319)	16.808 (0.921)
	QR-L	6.890 (0.136)	9.865 (0.331)	7.637 (0.221)	10.760 (0.206)

- The third choice is a Laplace distribution Laplace(μ, b), with μ chosen so that the θth quantile is 0. b is set to 3 so that the variance is $2b^2 = 18$.
- The fourth choice is a mixture of two Laplace distributions, 0.1Laplace(μ, b_1) + 0.9Laplace(μ, b_2), with μ chosen so that the θth quantile is 0. b_1 is set as 1 and b_2 is set as $\sqrt{5}$.

Note that we intentionally choose error distributions that are different from the skewed Laplace distribution to see how the Bayesian regularized quantile regression depends on this error assumption, as there has been criticisms that assigning a specific error distribution is departing from the semiparametric nature of quantile regression since quantile regression treats the error distribution nonparametrically. Our simulation results show that, in terms of parameter estimation accuracy and quantile estimation accuracy, the Bayesian regularized quantile regression methods still perform well even when this error distribution assumption is violated.

For each simulation study and each choice of the error distribution, we run 50 simulations. In each simulation, we generate a training set with 20 observations, a validation set with 20 observations, and a testing set with 200 observations. The validation set is used to choose the penalty parameters in lasso (λ), EN (λ_1 and λ_2) and QR-L (λ). After

Table 3: MMADs for Simulation 5. In the parentheses are standard deviations of the MMADs obtained by 500 bootstrap resampling. The bold numbers correspond to the smallest MMAD in each category.

θ	Method	Error Distribution			
		normal	normal mixture	Laplace	Laplace mixture
$\theta = 0.5$	BQR.L	1.353 (0.058)	1.857 (0.144)	**1.397** (0.109)	2.181 (0.085)
	BQR.EN	1.228 (0.098)	1.827 (0.150)	1.563 (0.078)	2.263 (0.081)
	lasso	1.269 (0.056)	2.197 (0.135)	1.895 (0.120)	2.530 (0.083)
	EN	1.251 (0.058)	1.909 (0.124)	1.651 (0.107)	2.272 (0.134)
	QR-L	1.498 (0.089)	**1.676** (0.080)	1.627 (0.138)	1.836 (0.120)
	BQR.GL	**1.222** (0.087)	1.800 (0.070)	1.466 (0.077)	**1.834** (0.080)
$\theta = 0.3$	BQR.L	1.275 (0.066)	2.202 (0.058)	1.416 (0.101)	2.330 (0.091)
	BQR.EN	1.295 (0.066)	2.240 (0.075)	1.653 (0.129)	2.434 (0.038)
	lasso	1.308 (0.083)	2.326 (0.159)	1.783 (0.119)	2.791 (0.141)
	EN	1.346 (0.061)	2.122 (0.219)	1.658 (0.064)	2.530 (0.036)
	QR-L	1.360 (0.051)	2.257 (0.056)	1.621 (0.076)	2.492 (0.100)
	BQR.GL	**1.232** (0.048)	**2.013** (0.110)	**1.361** (0.136)	**2.182** (0.075)
$\theta = 0.1$	BQR.L	1.259 (0.047)	2.424 (0.018)	1.450 (0.060)	2.457 (0.027)
	BQR.EN	1.315 (0.071)	2.491 (0.016)	1.524 (0.066)	2.510 (0.011)
	lasso	1.292 (0.055)	2.693 (0.190)	1.856 (0.119)	2.530 (0.148)
	EN	1.240 (0.058)	2.530 (0.101)	1.779 (0.086)	2.530 (0.022)
	QR-L	1.746 (0.068)	6.025 (0.611)	2.343 (0.109)	6.972 (0.742)
	BQR.GL	**1.225** (0.058)	**2.350** (0.045)	**1.424** (0.070)	**2.366** (0.049)

the penalty parameters are selected, we combine the training set and the validation set together to estimate $\boldsymbol{\beta}$. Since QR is not a regularization method, it does not need the validation set to choose any penalty parameters, so we directly combine the training and validation sets together to estimate $\boldsymbol{\beta}$. Similarly, BQR.L, BQR.EN and BQR.GL do not need the validation set since they estimate the penalty parameter automatically, so we also combine the training and validation sets together for estimation. The testing set is used to evaluate the performance for these methods.

Priors for the Bayesian methods are taken to be almost noninformative. In BQR.L and BQR.GL, the parameters a, b, c, d in the Gamma priors for τ and η^2 are all set to be 0.1. Similarly, in BQR.EN, the parameters a, b, c_1, d_1, c_2, d_2 in the Gamma priors for $\tau, \tilde{\eta}_1$ and η_2 are also chosen to be 0.1.

In the first four simulation studies, there is no group structure in the predictors, so BQR.GL reduces to BQR.L. In Simulation 5, we choose the positive definite matrices $\mathbf{K}_g = d_g \mathbf{I}_{d_g}$, where d_g's are dimensions of $\boldsymbol{\beta}_g$'s, as suggested by Yuan and Lin (2006).

We considered five different values of the given quantile θ: 10%, 30%, 50%, 70% and 90%. But since all four distributions of u_i's are symmetric, only 10%, 30% and 50% need to be reported. Due to the space limit, we only present the results for Simulation

Table 4: The parameter estimations for the first three simulation studies. The error distribution is chosen to be normal and the quantile θ is chosen to be 0.1. Within each simulation study, the median of 50 estimates of $\boldsymbol{\beta}$ is reported.

Simulation Study	Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
Simulation 1	$\boldsymbol{\beta}_{\text{true}}$	3.000	1.500	0.000	0.000	2.000	0.000	0.000	0.000
	BQR.L	2.917	1.436	0.053	0.038	1.623	-0.001	0.045	0.019
	BQR.EN	2.852	1.522	0.091	0.099	1.638	-0.055	0.126	0.018
	lasso	2.605	1.276	0.000	0.000	1.292	0.000	0.000	0.000
	EN	2.606	1.569	0.198	0.021	1.249	0.000	0.000	0.000
	QR	3.007	1.642	-0.030	0.123	2.000	-0.109	0.218	-0.007
	QR-L	2.501	1.453	0.000	0.000	1.548	0.000	0.000	0.000
Simulation 2	$\boldsymbol{\beta}_{\text{true}}$	0.850	0.850	0.850	0.850	0.850	0.850	0.850	0.850
	BQR.L	0.655	0.885	0.616	0.734	0.630	0.548	0.807	0.585
	BQR.EN	0.714	0.879	0.759	0.797	0.726	0.556	0.926	0.729
	lasso	0.534	0.756	0.758	0.839	0.630	0.592	0.728	0.527
	EN	0.562	0.905	0.960	1.102	0.964	0.845	0.864	0.499
	QR	0.884	1.000	0.808	0.962	0.864	0.729	1.057	0.831
	QR-L	0.403	0.721	0.746	0.853	0.772	0.620	0.626	0.574
Simulation 3	$\boldsymbol{\beta}_{\text{true}}$	5.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	BQR.L	4.802	0.0851	0.034	0.032	0.008	-0.081	0.021	0.031
	BQR.EN	4.689	0.1634	0.037	0.008	0.040	-0.140	0.093	-0.016
	lasso	4.018	0.0000	0.000	0.000	0.000	0.000	0.000	0.000
	EN	3.956	0.1568	0.000	0.000	0.000	0.000	0.000	0.000
	QR	4.981	0.1615	-0.030	0.123	0.025	-0.109	0.218	-0.007
	QR-L	4.377	0.0017	0.000	0.000	0.000	0.000	0.000	0.000

1, 4 and 5 (Table 1, 2 and 3). In these tables, the MMAD stands for the median of mean absolute deviations, i.e. median$(1/200\sum_{i=1}^{200}|\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{\text{true}}|)$, where the median is taken over the 50 simulations. In the parentheses are the standard deviations of the MMAD obtained by 500 bootstrap resampling of the 50 mean absolute deviations (MAD).

Simulation 1, 2 and 3 show that, in terms of the MMAD, the two Bayesian regularized quantile regression methods perform better than the other four methods in general, especially for the last three non-normal error distributions. In Simulation 2, whose results are not shown here, BQR.EN performs the best — BQR.EN has the smallest MMAD in 9 out of 12 simulation setups. In Simulation 3, with results not shown either, BQR.L performs the best — BQR.L has the smallest MMAD in 9 out of 12 simulation setups. Although none of these four error distributions are assumed in the Bayesian regularized quantile regression methods, performance of BQR.L and BQR.EN shows that the Bayesian methods are quite robust to the error distribution assumption. Secondly, as the Bayesian counterpart for QR-L,BQR.L generally behaves better in terms of the MMAD. Thirdly, while QR-L performs well for $\theta = 0.5$ and $\theta = 0.3$, its performance drops noticeably when $\theta = 0.1$. This phenomenon also shows up in

Simulation 4 and 5 and in the real data example in Section 4.

Simulation 4 is the case where we have more predictors than the sample size. Usually the elastic net penalty is recommended in such situations, as can be seen from Table 2. Simulation 5 corresponds to the case with group structures among the predictors, and group lasso penalty tends to behave better. The results are summarized in Table 3, where we can see that BQR.GL gives the best MMADs most of the times. Another comment on Table 3 is that, due to the categorical nature of the predictors, the design matrix is singular. As a result, the standard QR fails in this situation, while the other six methods still work. This is another example showing the advantage of regularization based methods.

Instead of looking at the MMADs, we may also look at the estimation of $\boldsymbol{\beta}$'s directly. Since the results would be too many to put in a table, we only choose the case where $\theta = 0.1$ and normally distributed errors in the first three simulation studies for illustrations. These are summarized in Table 4. From Table 4 we can see that, QR tends to give less biased parameter estimates for $\boldsymbol{\beta}$, but this not necessarily guarantees good quantile prediction, as implied by the MMADs in all previous tables. Similar patterns are also observed for the cases not shown here.

3.2 Heterogeneous random errors

Now we consider the case where u_i's are not i.i.d. random errors. The data are generated according to Example 5.4 in Wu and Liu (2009) from the model

$$y_i = 1 + x_{1i} + x_{2i} + x_{3i} + (1 + x_{3i})\varepsilon_i,$$

where x_{1i} is generated independently from $N(0, 1)$, x_{3i} is generated independently from the uniform distribution on $[0, 1]$, $x_{2i} = x_{1i} + x_{3i} + z_i$, where z_i follows $N(0, 1)$. ε_i's are independent normally distributed with variance 1 and mean as the negative of the θth standard normal quantile. There are also five more independent noise random variables, distributed as $N(0, 1)$, $x_4, \ldots, x_8$. The results are summarized in Table 5. Here we use the same prior specification as those in Section 3.1, setting $a = b = c = d = 0.1$ for BQR.L and $a = b = c_1 = d_1 = c_2 = d_2 = 0.1$ for BQR.EN. Table 5 contains another performance criterion, the test error, which refers to the average check loss on the independent testing data set (Wu and Liu 2009).

From Table 5 we can see that BQR.L, BQR.EN, QR and QR-L behave significantly better than lasso and EN, which demonstrates the broader applicability of quantile regression based methods. Secondly, within the set of quantile regression based methods, all four methods considered here behave similarly in terms of the test error. However, if we consider MMAD, BQR.L and BQR.EN outperform others uniformly. Thirdly, BQR.L performs better than its non-Bayesian counterpart, QR-L, uniformly for all quantiles considered.

Table 5: MMADs and test errors for the simulation with heterogeneous random errors. The bold numbers correspond to the smallest MMAD or test error in each category.

θ	Method	MMAD(SD)	Test Error(SD)
0.25	BQR.L	**0.2995** (0.012)	**0.4942** (0.0030)
	BQR.EN	0.3056 (0.009)	0.4945 (0.0031)
	lasso	1.1015 (0.047)	0.6291 (0.0113)
	EN	1.2389 (0.080)	0.6281 (0.0081)
	QR	0.3180 (0.006)	0.4967 (0.0018)
	QR-L	0.3211 (0.011)	0.4950 (0.0028)
0.5	BQR.L	0.2764 (0.010)	0.6128 (0.0038)
	BQR.EN	**0.2650** (0.009)	0.6128 (0.0030)
	lasso	1.2976 (0.064)	0.8256 (0.0178)
	EN	1.4991 (0.007)	0.8822 (0.0060)
	QR	0.2984 (0.005)	0.6151 (0.0043)
	QR-L	0.2902 (0.005)	**0.6122** (0.0031)
0.75	BQR.L	0.2945 (0.008)	**0.4914** (0.0017)
	BQR.EN	**0.2809** (0.008)	0.4931 (0.0020)
	lasso	1.5853 (0.036)	0.8073 (0.0223)
	EN	1.5448 (0.007)	0.7836 (0.0058)
	QR	0.3165 (0.006)	0.4967 (0.0019)
	QR-L	0.3186 (0.008)	0.4929 (0.0029)

4 A real data example

In this section, we compare the performance of the six methods in Section 3, BQR.L, BQR.EN, lasso, EN, QR, QR-L, on the Boston Housing data. This data set was first analyzed by Harrison and Rubinfeld (1978) in a study on the influence of "clean air" on house prices. The version of the data set in this paper is a corrected version of the original data, corrected for a few minor errors and augmented with the latitude and longitude of the observations, and is available in the "spdep" package in `R` (R Development Core Team 2005). It has 506 rows and 20 columns. The response variable is the log-transformed corrected median value of owner-occupied housing in USD 1000 (LCMEDV). Predictors that we considered include point longitudes in decimal degrees (LON), point latitudes in decimal degrees (LAT), per capita crime (CRIM), proportions of residential land zoned for lots over 25000 square feet per town (ZN), proportions of non-retail business acres per town (INDUS), a factor indicating whether tract borders Charles River (CHAS), nitric oxides concentration (parts per 10 million) per town (NOX), average numbers of rooms per dwelling (RM), proportions of owner-occupied units built prior to 1940 (AGE), weighted distances to five Boston employment centers (DIS), index of accessibility to radial highways per town (RAD), full-value property-tax rate per USD 10,000 per town (TAX), pupil-teacher ratios per town (PTRATIO), transformed African American population proportion (B) and percentage values of lower status population (LSTAT).

Similar as in Section 3, we consider three choices of θ, 0.1, 0.3 and 0.5. Also, we run 10-fold cross-validation to evaluate the performance of the four methods under the check-loss criterion (1). The prior specifications are the same as those in Section 3, $a = b = c = d = 0.1$ for BQR.L and $a = b = c_1 = d_1 = c_2 = d_2 = 0.1$ for BQR.EN. The results are summarized in Table 6 and show that the Bayesian quantile regression methods perform uniformly better than the lasso and EN for all selected quantiles, and similarly to QR and QR-L. Again, as mentioned in the discussion of Tables 1, 2 and 3, the performance of QR-L drops when θ changes from 0.5 and 0.3 to 0.1.

Also, instead of comparing the check loss, we can look at the point and interval estimations for the parameters. These are summarized in Figure 1, where we choose the case with $\theta = 0.1$ as an illustration. 95% credible intervals are plotted for BQR.L and BQR.EN estimators in these plots. As there are too many estimators in a plot, we add a slight horizontal shift to the estimators given by BQR.L and BQR.EN to make it more readable. Note that the Bayesian methods can easily provide the interval estimations while the frequentist methods usually do not have simply implemented interval estimators. From this figure we can see that the four quantile regression based methods tend to behave similarly while lasso and EN also behave similarly. Also, among the four quantile regression based methods, QR-L tends to behave more differently from the other three methods, as there are many QR-L estimators lying outside the 95% credible intervals of BQR.L and BQR.EN. This difference is less apparent for $\theta = 0.5$ and $\theta = 0.3$, the figures for which are not shown here.

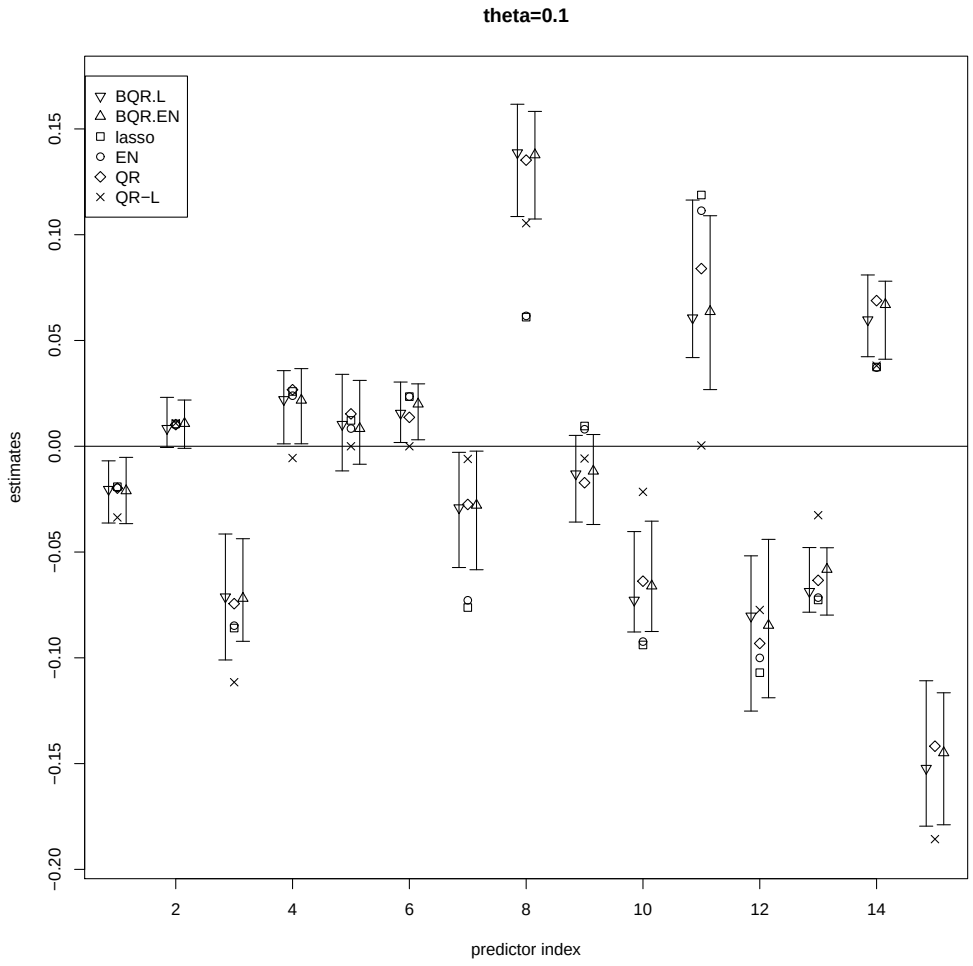

Figure 1: The estimates of the predictor effects for the Boston Housing data using different methods. The given quantile is $\theta = 0.1$. The 95% credible intervals given by BQR.L and BQR.EN are also plotted.

Table 6: 10-fold cross-validation results for the Boston Housing data. The bold numbers correspond to the smallest MMAD in each category.

Method	Test error		
	$\theta = 0.1$	$\theta = 0.3$	$\theta = 0.5$
BQR.L	0.06626	0.06639	**0.06597**
BQR.EN	**0.06613**	0.06680	0.06600
lasso	0.06806	0.06806	0.06806
EN	0.06803	0.06807	0.06807
QR	0.06623	**0.06623**	0.06623
QR-L	0.07398	0.06745	0.06755

5 Conclusion and Discussion

In this paper, we propose the Bayesian regularized quantile regression and treat generically three different types of penalties, lasso, elastic net and group lasso. Bayesian hierarchical models are developed for each regularized quantile regression problem and Gibbs samplers are derived. Simulation studies and real data examples show that Bayesian regularized quantile regression methods generally perform better compared with current existing non-Bayesian regularized quantile regression methods. In particular, as counterparts to each other, BQR.L behaves better than QR-L.

One of the most valued advantages of quantile regression is its model robustness in the sense that it makes no distributional assumption to the error term other than its quantile. However, in the parametric Bayesian quantile regression framework, a common practice is to assume the error to have the skewed Laplace distribution (2). While this may cause some concern on losing the nonparametric nature of quantile regression, our results show that the Bayesian methods are quite insensitive to this assumption and behave well for data generated from other distributions.

One issue we found in the numerical studies is that, compared to the Bayesian methods, the performance of QR-L often deteriorates for extreme quantiles like $\theta = 0.1$, and further study is needed to explain this phenomenon. Another future direction is to develop Bayesian regularization for multiple quantiles. Zou and Yuan (2008a,b) considered regularized quantile regression models for a finite number of quantiles. Our Bayesian formulation shall extend naturally to this context by imposing suitable functional priors on the coefficients.

Appendix

A. The Gibbs sampler for the lasso regularized quantile regression

Let $\boldsymbol{\beta}_{-k}$ be the parameter vector $\boldsymbol{\beta}$ excluding the component β_k, $\mathbf{s}_{-k}$ be the variable $\mathbf{s}$ excluding the component s_k and $\tilde{\mathbf{v}}_{-i}$ be the variable $\tilde{\mathbf{v}}$ excluding the component $\tilde{v}_i$. Denote $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$. Then, the conditional distribution $f(\mathbf{y}|\mathbf{X}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}, \tau, \eta^2)$ in the lasso regularized quantile regression is

$$\begin{aligned} f(\mathbf{y}|\mathbf{X}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}, \tau, \eta^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\tau^{-1}\xi_2^2\tilde{v}_i}} \exp\left\{ -\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2}{2\tau^{-1}\xi_2^2\tilde{v}_i} \right\} \\ &= \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2}{\tau^{-1}\xi_2^2\tilde{v}_i} \right\} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\tau^{-1}\xi_2^2\tilde{v}_i}}. \end{aligned}$$

The full conditional distribution $f(\tilde{v}_i \mid \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}_{-i}, \boldsymbol{\beta}, \mathbf{s}, \tau, \eta^2)$ is

$$\begin{aligned} f(\tilde{v}_i \mid \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}_{-i}, \boldsymbol{\beta}, \mathbf{s}, \tau, \eta^2) &\propto f(\mathbf{y}|\mathbf{X}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}, \tau, \eta)\pi(\tilde{v}_i|\tau) \\ &\propto \frac{1}{\sqrt{\tilde{v}_i}} \exp\left\{ -\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2}{2\tau^{-1}\xi_2^2\tilde{v}_i} \right\} \exp(-\tau\tilde{v}_i) \\ &\propto \frac{1}{\sqrt{\tilde{v}_i}} \exp\left\{ -\frac{1}{2}\left[\left(\frac{\tau\xi_1^2}{\xi_2^2} + 2\tau \right)\tilde{v}_i + \frac{\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{\xi_2^2}\tilde{v}_i^{-1} \right] \right\}. \end{aligned}$$

Thus, the full conditional distribution of $\tilde{v}_i$ is a generalized inverse Gaussian distribution. The full conditional distribution $f(s_k \mid \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}_{-k}, \tau, \eta^2)$ of s_k is

$$\begin{aligned} f(s_k \mid \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}_{-k}, \tau, \eta^2) &\propto \pi(\beta_k \mid s_k)\pi(s_k \mid \eta^2) \\ &\propto \frac{1}{\sqrt{2\pi s_k}} \exp\left(-\frac{\beta_k^2}{2s_k} \right) \exp\left(-\frac{\eta^2}{2}s_k \right) \\ &\propto \frac{1}{\sqrt{s_k}} \exp\left\{ -\frac{1}{2}\left[\eta^2 s_k + \beta_k^2 s_k^{-1} \right] \right\}. \end{aligned}$$

The full conditional $f(s_k \mid \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}_{-k}, \tau, \eta^2)$ is then again a generalized inverse Gaussian distribution. Now we consider the full conditional distribution of β_k which is given by

$$\begin{aligned} f(\beta_k \mid \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}_{-k}, \mathbf{s}, \tau, \eta^2) &\propto f(\mathbf{y}|\mathbf{X}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}, \tau, \eta)\pi(\beta_k \mid s_k) \\ &\propto \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2}{\tau^{-1}\xi_2^2\tilde{v}_i} \right\} \exp\left(-\frac{\beta_k^2}{2s_k} \right) \\ &\propto \exp\left\{ -\frac{1}{2}\left[\left(\sum_{i=1}^{n} \frac{\tau x_{ik}^2}{\xi_2^2\tilde{v}_i} + \frac{1}{s_k} \right)\beta_k^2 - 2\sum_{i=1}^{n} \frac{\tau\tilde{y}_{ik}x_{ik}}{\xi_2^2\tilde{v}_i}\beta_k \right] \right\}, \end{aligned}$$

where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$ and $\tilde{y}_{ik} = y_i - \xi_1\tilde{v}_i - \sum_{j=1, j\neq k}^{p} x_{ij}\beta_j$. Let $\tilde{\sigma}_k^{-2} = \tau\xi_2^{-2}\sum_{i=1}^{n} x_{ik}^2\tilde{v}_i^{-1} + s_k^{-1}$ and $\tilde{\mu}_k = \tilde{\sigma}_k^2\tau\xi_2^{-2}\sum_{i=1}^{n} \tilde{y}_{ik}x_{ik}\tilde{v}_i^{-1}$. Then the full conditional

distribution is just the normal distribution $N(\tilde{\mu}_k, \tilde{\sigma}_k^2)$. The full conditional distribution of τ is

$$\begin{aligned} & f(\tau \mid \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}, \eta^2) \\ \propto\ & f(\mathbf{y} \mid \mathbf{X}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}, \tau, \eta)\pi(\tilde{v} \mid \tau)\pi(\tau) \\ \propto\ & \tau^{n/2} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2}{\tau^{-1}\xi_2^2\tilde{v}_i} \right\} \tau^n \exp\left(-\tau \sum_{i=1}^{n} \tilde{v}_i \right) \tau^{a-1} \exp(-b\tau) \\ \propto\ & \tau^{a+3n/2-1} \exp\left\{ -\tau \left[\sum_{i=1}^{n} \left(\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2}{2\xi_2^2\tilde{v}_i} + \tilde{v}_i \right) + b \right] \right\}. \end{aligned}$$

That is, the full conditional distribution of τ is a Gamma distribution.

B. Calculation of $C(\eta_1, \eta_2)$ and the Gibbs sampler for quantile regression with the elastic net penalty

B.1 Calculation of $C(\eta_1, \eta_2)$ in (11)

Before putting priors on η_1 and η_2, let us first calculate the quantity $C(\eta_1, \eta_2)$. By (8), we have

$$\begin{aligned} \pi(\beta_k|\ \eta_1, \eta_2) &= C(\eta_1, \eta_2)\frac{\eta_1}{2}\exp\{-\eta_1|\beta_k| - \eta_2\beta_k^2\} \\ &= C(\eta_1, \eta_2)\int_0^\infty \frac{1}{\sqrt{2\pi s_k}} \exp\left\{ -\frac{1 + 2\eta_2 s_k}{2s_k}\beta_k^2 \right\} \frac{\eta_1^2}{2} \exp\left(\frac{-\eta_1^2}{2} s_k \right) ds_k. \end{aligned}$$

Let $t_k = 1 + 2\eta_2 s_k$. Then we have

$$\begin{aligned} \pi(\beta_k|\ \eta_1, \eta_2) &= C(\eta_1, \eta_2)\int_1^\infty \frac{t_k^{-1/2}}{\sqrt{2\pi(t_k - 1)/(2\eta_2 t_k)}} \exp\left\{ -\frac{1}{2}\left(\frac{t_k - 1}{2\eta_2 t_k} \right)^{-1} \beta_k^2 \right\} \times \\ &\quad \frac{\eta_1^2}{4\eta_2} \exp\left\{ \frac{-\eta_1^2}{4\eta_2}(t_k - 1) \right\} dt_k. \end{aligned}$$

Since $\int_{-\infty}^{\infty} \pi(\beta_k|\ \eta_1, \eta_2) d\beta_k = 1$, by Fubini's theorem, we have

$$\begin{aligned}\int_{-\infty}^{\infty} \pi(\beta_k|\ \eta_1, \eta_2) d\beta_k &= C(\eta_1, \eta_2) \int_1^{\infty} t_k^{-1/2} \frac{\eta_1^2}{4\eta_2} \exp\left\{\frac{-\eta_1^2}{4\eta_2}(t_k - 1)\right\} dt_k \\ &= C(\eta_1, \eta_2)\left(\frac{\eta_1^2}{4\eta_2}\right)^{1/2} \exp\left\{\frac{\eta_1^2}{4\eta_2}\right\} \int_{4^{-1}\eta_1^2\eta_2^{-1}}^{\infty} t^{-1/2} \exp(-t) dt \\ &= 1.\end{aligned}$$

Notice that $\int_{4^{-1}\eta_1^2\eta_2^{-1}}^{\infty} t^{-1/2} \exp(-t) dt = \Gamma(1/2,\ 4^{-1}\eta_1^2\eta_2^{-1})$ is the upper incomplete gamma function. Therefore, we have

$$C(\eta_1, \eta_2) = \Gamma^{-1}\left(1/2,\ \frac{\eta_1^2}{4\eta_2}\right)\left(\frac{\eta_1^2}{4\eta_2}\right)^{-1/2} \exp\left\{-\frac{\eta_1^2}{4\eta_2}\right\}.$$

B.2 The Gibbs sampler for quantile regression with the elastic net penalty

The full conditional distributions of $\tilde{v}_i$ and τ are again the same as in the lasso regularized quantile regression case. Let

$$\tilde{y}_{ik} = y_i - \xi_1 \tilde{v}_i - \sum_{j=1, j\neq k}^{p} x_{ij}\beta_j,$$

$$\tilde{\sigma}_k^{-2} = \tau\xi_2^{-2} \sum_{i=1}^{n} x_{ik}^2 \tilde{v}_i^{-1} + 2\eta_2 t_k (t_k - 1)^{-1},$$

$$\tilde{\mu}_k = \tilde{\sigma}_k^2 \tau \xi_2^{-2} \sum_{i=1}^{n} \tilde{y}_{ik} x_{ik} \tilde{v}_i^{-1}.$$

Then, similar to the lasso regularized quantile regression case, we can show that the full conditional $f(\beta_k \mid \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}_{-k}, \mathbf{t}, \tau, \tilde{\eta}_1, \eta_2)$ of β_k is the normal distribution $N(\tilde{\mu}_k, \tilde{\sigma}_k^2)$.

The full conditional distribution of $t_k - 1$ is

$$\begin{aligned} & f(t_k - 1 \mid \mathbf{y}, \mathbf{X}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{t}_{-k}, \tau, \tilde{\eta}_1, \eta_2) \\ \propto\ & \frac{1}{\sqrt{t_k - 1}} \exp\left\{ -\frac{1}{2}\left(\frac{t_k - 1}{2\eta_2 t_k}\right)^{-1} \beta_k^2 \right\} \exp\left\{ -\tilde{\eta}_1 t_k \right\} I(t_k > 1) \\ \propto\ & \frac{1}{\sqrt{t_k - 1}} \exp\left\{ -\frac{1}{2}\left[2\tilde{\eta}_1(t_k - 1) + \frac{2\eta_2\beta_k^2}{t_k - 1}\right]\right\} I(t_k - 1 > 0). \end{aligned}$$

Therefore, the full conditional distribution of $t_k - 1$ is a generalized Gaussian distribution. The full conditional distribution of $\tilde{\eta}_1$ is

$$\begin{aligned} f(\tilde{\eta}_1 \mid \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{t}, \tau, \eta_2) &\propto \tilde{\eta}_1^{c_1 - 1} \exp(-d_1\tilde{\eta}_1) \prod_{k=1}^{p} \Gamma^{-1}(1/2,\ \tilde{\eta}_1) \tilde{\eta}_1^{1/2} \exp\left\{ -\tilde{\eta}_1 t_k \right\} \\ &\propto \Gamma^{-p}(1/2,\ \tilde{\eta}_1) \tilde{\eta}_1^{p/2 + c_1 - 1} \exp\left\{ -\tilde{\eta}_1 \left[d_1 + \sum_{k=1}^{p} t_k \right] \right\}. \end{aligned}$$

The full conditional distribution of η_2 is

$$\begin{aligned} f(\eta_2 \mid \mathbf{X}, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{t}, \tau, \tilde{\eta}_1) &\propto \eta_2^{c_2-1} \exp(-d_2\eta_2) \prod_{k=1}^{p} \eta_2^{1/2} \exp\left\{ -\frac{1}{2}\left(\frac{t_k - 1}{2\eta_2 t_k}\right)^{-1} \beta_k^2 \right\} \\ &\propto \eta_2^{p/2+c_2-1} \exp\left\{ -\eta_2 \left(d_2 + \sum_{k=1}^{p} t_k (t_k - 1)^{-1} \beta_k^2 \right) \right\}, \end{aligned}$$

which is a Gamma distribution.

References

Andrews, D. F. and Mallows, C. L. (1974). "Scale Mixtures of Normal Distributions." *Journal of the Royal Statistical Society, Ser. B*, 36: 99–102. 537

Bae, K. and Mallick, B. (2004). "Gene Selection Using a Two-Level Hierarchical Bayesian Model." *Bioinformatics*, 20: 3423–3430. 534

Bakin, S. (1999). "Adaptive Regression and Model Selection in Data Mining Problems." PhD thesis, Australian National University, Canberra. 539

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. 534

Geraci, M. and Bottai, M. (2007). "Quantile Regression for Longitudinal Data Using the Asymmetric Laplace Distribution." *Biostatistics*, 8: 140–154. 534

Hanson, T. and Johnson, W. (2002). "Modeling Regression Error with a Mixture of Pólya trees." *Journal of the American Statistical Association*, 97: 1020–1033. 534

Harrison, D. and Rubinfeld, D. L. (1978). "Hedonic Prices and the Demand for Clean Air." *Journal of Environmental Economics and Management*, 5: 81–102. 547

Hjort, N. and Walker, S. (2009). "Quantile Pyramids for Bayesian Nonparametrics." *Annals of Statistics*, 37: 105–131. 534

Jørgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution, Lecture Notes in Statistics*, volume 9. Springer-Verlag. 537

Koenker, R. (2004). "Quantile Regression for Longitudinal Data." *Journal of Multivariate Analysis*, 91: 74–89. 534

— (2005). *Quantile Regression*. New York: Cambridge University Press. 533

Koenker, R. and Bassett, G. W. (1978). "Regression Quantiles." *Econometrica*, 46: 33–50. 533

Kottas, A. and Gelfand, A. (2001). "Bayesian Semiparametric Median Regression Modeling." *Journal of the American Statistical Association*, 96: 1458–1468. 534

Kottas, A. and Krnjajić, M. (2009). "Bayesian Semiparametric Modeling in Quantile Regression." *Scandinavian Journal of Statistics*, 36: 297–319. 534

Kozumi, H. and Kobayashi, G. (2009). "Gibbs Sampling Methods for Bayesian Quantile Regression." Technical report, Graduate School of Business Administration, Kobe University.
URL http://www.b.kobe-u.ac.jp/paper/2009_02.pdf 534, 535

Li, Y. and Zhu, J. (2008). "L_1-Norm Quantile Regression." *Journal of Computational and Graphical Statistics*, 17: 163–185. 533, 534, 536, 540

Meier, L., van de Geer, S., and Bühlmann, P. (2008). "The Group Lasso for Logistic Regression." *Journal of the Royal Statistical Society, Ser. B*, 70(1): 53–71. 539

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103: 681–686. 534

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org 547

Reed, C., Dunson, D., and Yu, K. (2009). "Bayesian variable selection in quantile regression." Technical report, Department of Mathematical Sciences, Brunel University. 534

Reed, C. and Yu, K. (2009). "An efficient Gibbs sampler for Bayesian quantile regression." Technical report, Department of Mathematical Sciences, Brunel University. 534

Reich, B., Bondell, H., and Wang, H. (2009). "Flexible Bayesian Quantile Regression for Independent and Clustered Data." *Biostatistics*. Doi:10.1093/biostatistics/kxp049. 534

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Ser. B*, 58: 267–288. 534

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). "Sparsity and Smoothness via the Fused Lasso." *Journal of the Royal Statistical Society: Ser. B*, 67: 91–108. 534

Tsionas, E. (2003). "Bayesian Quantile Inference." *Journal of Statistical Computation and Simulation*, 73: 659–674. 534

Walker, S. and Mallick, B. (1999). "A Bayesian Semiparametric Accelerated Failure Time Model." *Biometrics*, 55: 477–483. 534

Wang, H., Li, G., and Jiang, G. (2007). "Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso." *Journal of Business & Economic Statistics*, 25: 347–355. 534

Wu, Y. and Liu, Y. (2009). "Variable Selection in Quantile Regression." *Statistica Sinica*, 19: 801–817. 533, 534, 545

Yu, K. and Moyeed, R. (2001). "Bayesian Quantile Regression." *Statistics & Probability Letters*, 54: 437–447. 533

Yuan, M. and Lin, Y. (2006). "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society, Ser. B*, 68: 49–67. 534, 539, 543

Zou, H. and Hastie, T. (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society, Ser. B*, 67: 301–320. 534

Zou, H. and Yuan, M. (2008a). "Composite Quantile Regression and The Oracle Model Selection Theory." *Annals of Statistics*, 36: 1108–1126. 549

— (2008b). "Regularized Simultaneous Model Selection in Multiple Quantiles Regression." *Computational Statistics & Data Analysis*, 52: 5296–5304. 549