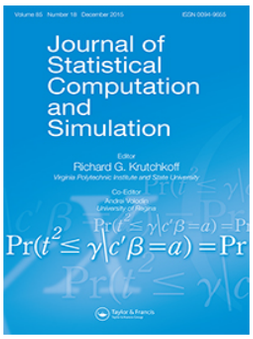

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Bayesian composite quantile regression

Hanwen Huang & Zhongxue Chen

To cite this article: Hanwen Huang & Zhongxue Chen (2015) Bayesian composite quantile regression, Journal of Statistical Computation and Simulation, 85:18, 3744-3754, DOI: 10.1080/00949655.2015.1014372

To link to this article: https://doi.org/10.1080/00949655.2015.1014372

Journal of Statistical Computation and Simulation, 2015
Vol. 85, No. 18, 3744–3754, http://dx.doi.org/10.1080/00949655.2015.1014372

Taylor & Francis
Taylor & Francis Group

Bayesian composite quantile regression

Hanwen Huang[a*] and Zhongxue Chen[b]

[a]*Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA 30605, USA;*
[b]*Department of Epidemiology and Biostatistics, Indiana University, Bloomington, IN 47405, USA*

(*Received 21 November 2014; accepted 29 January 2015*)

One advantage of quantile regression, relative to the ordinary least-square (OLS) regression, is that the quantile regression estimates are more robust against outliers and non-normal errors in the response measurements. However, the relative efficiency of the quantile regression estimator with respect to the OLS estimator can be arbitrarily small. To overcome this problem, composite quantile regression methods have been proposed in the literature which are resistant to heavy-tailed errors or outliers in the response and at the same time are more efficient than the traditional single quantile-based quantile regression method. This paper studies the composite quantile regression from a Bayesian perspective. The advantage of the Bayesian hierarchical framework is that the weight of each component in the composite model can be treated as open parameter and automatically estimated through Markov chain Monte Carlo sampling procedure. Moreover, the lasso regularization can be naturally incorporated into the model to perform variable selection. The performance of the proposed method over the single quantile-based method was demonstrated via extensive simulations and real data analysis.

Keywords: Laplace prior; mixture model; quantile regression; variable selection

1. Introduction

It is well known that the ordinary least-square (OLS) estimator is not robust to problem with response variable subject to heavy-tailed errors or outliers. In practice, quantile regression [1] has been widely used to overcome this problem. While the OLS regression focuses on the mean outcome given predictor variables, quantile regression considers the impact of the covariates on the entire distribution of the response variable and thus provides a richer characterization of the data. Historically, quantile regression has focused on estimation of a single quantile (e.g. the median) given predictor variables. However, when the error has no heavy tail or does not suffer from outliers, the relative efficiency of the quantile regression based on a single quantile can be arbitrarily small in contrast to the OLS regression. Moreover, for a given distribution, the quantile regression at one quantile may deliver more efficient estimators than the quantile regression at another quantile. But in practice, the distribution is unknown and thus we do not know whether the quantile we choose is the most appropriate one or not.

In the last five years, there has been significant advancement in the simultaneous estimation of multiple quantiles. Zou and Yuan [2] proposed an equally weighted composite quantile regression (WCQR) method to simultaneously consider multiple quantile regression models. Zhao and

*Corresponding author. Email: huanghw@uga.edu

Xiao [3] argued that equal weight in general is not an efficient way of using distributional information from quantile regressions. To further improve the efficiency, they proposed a WCQR model to allow different components to have different weights. The asymptotic properties of the WCQR estimator have been derived in [3] and the optimal weights were determined by minimizing the corresponding asymptotic variance. However, their method requires estimation of the probability density function of the error process which is quite challenging for situations with relatively small sample size. Bradic et al. [4] developed a composite quasi-likelihood method which allows the linear combination of different convex loss functions and treats WCQR regression as a specific example.

This paper studies the WCQR model from a Bayesian perspective. In Bayesian hierarchical framework, the weight of each component can be naturally treated as open parameters and estimated through Markov chain Monte Carlo (MCMC) sampling. Bayesian formulation of single quantile-based quantile regression has been extensively studied in the literature. Examples include [5–10] among many others. In addition to the computational convenience, Bayesian approach also has the advantage of enabling exact inference even when the sample size is small.

On the other hand, an important topic in linear regression analysis is variable selection. Variable selection is particularly important when the underlying model has sparse representation. The least absolute shrinkage and selection operator (lasso) is a popular technique for simultaneous estimation and variable selection.[11] The use of lasso penalty in quantile regression has been considered in [12–14] among many others. Li et al. [15] developed a Bayesian framework for regularization in linear quantile regression. Alhamzawi et al. [16] proposed adaptive lasso quantile regression from a Bayesian perspective.

This paper is the first attempt to use Bayesian technique to study the composite quantile regression model. We propose a Bayesian framework to combine weighted composite quantile and lasso regularization together to perform estimation and variable selection simultaneously. The rest of the article is organized as follows. In Section 2, we introduce the Bayesian composite quantile regression (BCQR) method with lasso penalty. An efficient MCMC sampling scheme is derived in Section 3. Section 4 presents some numerical results including applications to both simulated and real data. Discussion and conclusions are put in Section 5.

2. Bayesian composite quantile regression

Consider the following linear model:

$$y = q_0 + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\mathbf{x} \in R^p$ is the p-dimensional covariate, $\boldsymbol{\beta} \in R^p$ is the vector of unknown parameters, $q_0 \in R$ is the intercept, $y \in R$ is the response, and ϵ is the noise. The conditional θth quantile of $y \mid \mathbf{x}$ is

$$q_0 + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + q_\theta = b_\theta + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta},$$

where q_θ is the θth quantile of ϵ and uniquely defined for any $0 < \theta < 1$. Suppose we have samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, then the linear θth quantile regression model estimates $\boldsymbol{\beta}$ and b_θ by solving

$$(\hat{b}_\theta, \hat{\boldsymbol{\beta}}) = \mathrm{argmin}_{b_\theta, \boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\theta(y_i - b_\theta - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \tag{2}$$

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

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

To combine the information over quantiles, Zou and Yuan [2] proposed the composite quantile regression to simultaneously consider multiple quantile regression models. Denote $0 < \theta_1 \cdots < \theta_K < 1$. Composite quantile regression solve the following problem:

$$(\hat{b}_{\theta_1}, \ldots, \hat{b}_{\theta_K}, \hat{\boldsymbol{\beta}}) = \text{argmin}_{b_{\theta_1},\ldots,b_{\theta_K},\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K} \rho_{\theta_k}(y_i - b_{\theta_k} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \right\}. \tag{3}$$

The objective function of Equation (3) is a mixture of the objective functions from different quantile regression models with each component having equal weight. Zhao and Xiao [3] generalize Equation (3) to the WCQR which considers the following optimization problem:

$$\text{argmin}_{b_{\theta_1},\ldots,b_{\theta_K},\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K} w_k \rho_{\theta_k}(y_i - b_{\theta_k} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \right\}, \tag{4}$$

where $0 \le w_k \le 1$ is the weight for the kth component and $\sum_{k=1}^{K} w_k = 1$.

It can be easily shown that the minimization of the loss function (2) is exactly equivalent to the maximization of a likelihood function formed by combining independently distributed skewed Laplace densities

$$l = \prod_{i=1}^{n} \exp\{-\rho_\theta(y_i - b_\theta - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}.$$

Therefore, in Bayesian framework for θth quantile regression, it is naturally to assume that the noise ϵ in Equation (1) follows a skewed Laplace distribution with density

$$p_\theta(u \mid \tau) = \theta(1-\theta)\tau \exp[-\tau\rho_\theta(u)], \tag{5}$$

where $\tau > 0$ is the scale parameter. Denote $\mathbf{b} = (b_{\theta_1}, \ldots, b_{\theta_K})$. Then the joint distribution of $\mathbf{y} = (y_1, \ldots, y_n)$ given $X = (\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}}$ for a composite model is

$$p(\mathbf{y} \mid X, \mathbf{b}, \boldsymbol{\beta}, \tau) \sim \prod_{i=1}^{n} \left(\sum_{k=1}^{K} w_k p_{\theta_k}(y_i \mid \mathbf{x}_i, b_{\theta_k}, \boldsymbol{\beta}, \tau) \right), \tag{6}$$

where

$$p_{\theta_k}(y_i \mid \mathbf{x}_i, b_{\theta_k}, \boldsymbol{\beta}, \tau) = \theta_k(1-\theta_k)\tau \exp\{-\tau\rho_{\theta_k}(y_i - b_{\theta_k} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\} \quad \text{for } k = 1, \ldots, K.$$

It is difficult to solve Equation (6) directly because of the mixture of K components. We first introduce a cluster assignment matrix C whose i,kth element C_{ik} is equal to 1 if the ith subject belongs to the kth cluster, otherwise $C_{ik} = 0$. We treat C_{ik} as missing value and start from the complete likelihood which has the form

$$\prod_{i=1}^{n} \prod_{k=1}^{K} [w_k p_{\theta_k}(y_i \mid \mathbf{x}_i, b_{\theta_k}, \boldsymbol{\beta}, \tau)]^{C_{ik}}. \tag{7}$$

We further consider introducing regularization terms to the model such that variable selection can be performed. In Bayesian framework, the lasso penalty is equivalent to putting a Laplace

prior for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, i.e.

$$\pi(\boldsymbol{\beta} \mid \tau, \lambda) = \left(\frac{\tau\lambda}{2}\right)^p \exp\left\{-\tau\lambda \sum_{j=1}^{p} |\beta_j|\right\}.$$

Denote $\mathbf{w} = (w_1, \ldots, w_K)$ and put a Dirichlet prior on $\mathbf{w}$

$$\pi(\mathbf{w}) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_K),$$

where $\alpha_k > 0$ for $k = 1, \ldots, K$ are hyperparameters. Then the posterior distribution for the regularized WCQR method is

$$\prod_{i=1}^{n}\prod_{k=1}^{K}[w_k p_{\theta_k}(y_i \mid \mathbf{x}_i, b_{\theta_k}, \boldsymbol{\beta}, \tau)]^{C_{ik}} \pi(\boldsymbol{\beta} \mid \tau, \lambda)\pi(\mathbf{w}). \tag{8}$$

3. MCMC sampling

In this section, we will derive the MCMC sampling procedure for Equation (8). Note that it is difficult to get closed-form full conditional distributions for all parameters directly from Equation (8) due to the presence of the Laplace distributions in both likelihood term $p_{\theta_k}(y_i \mid \mathbf{x}_i, b_{\theta_k}, \boldsymbol{\beta}, \tau)$ and prior term $\pi(\boldsymbol{\beta} \mid \tau, \lambda)$. We need to introduce some auxiliary variables to the model such that efficient Gibbs sampling schemes can be used.

Kozumi and Kobayashi [6] proved that the skewed Laplace distribution (5) can be reformulated as a mixture of an exponential and a scaled normal distribution. More specifically, suppose that v is a standard exponential random variable and z is a standard normal random variable. Denote $\xi_{1k} = (1 - 2\theta_k)/\theta_k(1 - \theta_k)$ and $\xi_{2k} = \sqrt{2/\theta_k(1 - \theta_k)}$. Then the random variable $u = \xi_{1k}v + \xi_{2k}\sqrt{v}z$ follows the skewed Laplace distribution (5) with $\theta = \theta_k$ and $\tau = 1$. If the ith subject belongs to the kth cluster, then the response y_i can be equivalently written as $y_i = \mathbf{x}_i^{\mathrm{T}}\beta + \tau^{-1}\xi_{1k}v_i + \tau^{-1}\xi_{2k}\sqrt{v_i}z_i$, where v_i and z_i follow the standard exponential distribution and the standard normal distribution, respectively. Let $\tilde{v}_i = \tau^{-1}v_i$, then the density function of $\tilde{v}_i$ is $p(\tilde{v}_i \mid \tau) = \tau \exp(-\tau\tilde{v}_i)$.

According to Andrews and Mallows,[17] for any $a \geq 0$, the Laplace density function can be expressed as

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

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

$$\pi(\boldsymbol{\beta} \mid \tau, \lambda) = \prod_{j=1}^{p} \int_0^\infty \frac{1}{\sqrt{2\pi s_j}} \exp\left(-\frac{\beta_j^2}{2s_j}\right) \frac{\eta^2}{2} \exp\left(-\frac{\eta^2}{2}s_j\right) \mathrm{d}s_j.$$

After introducing the auxiliary variables $\tilde{v}_i$, z_i and s_j and further putting Gamma priors on the parameters τ and η^2, we have the following hierarchical model:

$$y_i = \prod_{k=1}^{K}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \xi_{1k}\tilde{v}_i + \tau^{-1/2}\xi_{2k}\sqrt{\tilde{v}_i}z_i)^{C_{ik}} \quad \text{for } i = 1, \ldots, n,$$

$$p(\tilde{v}_i \mid \tau) = \tau \exp(-\tau\tilde{v}_i) \quad \text{for } i = 1, \ldots, n,$$

$$p(z_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) \quad \text{for } i = 1, \ldots, n,$$

$$p(\beta_j, s_j \mid \eta^2) = \frac{1}{\sqrt{2\pi s_j}} \exp\left(-\frac{\beta_j^2}{2s_j}\right) \frac{\eta^2}{2} \exp\left(-\frac{\eta^2}{2} s_j\right) \quad \text{for } j = 1, \ldots, p,$$

$$p(\tau, \eta^2) = \tau^{a-1} \exp(-b\tau)(\eta^2)^{c-1} \exp(-d\eta^2),$$

where a, b, c, d are hyperparameters.

In the above hierarchical model, the parameters that need to be estimated include β_j, b_{θ_k}, τ, η^2, $\tilde{v}_i$, s_j, w_k and C_{ik} for $i = 1, \ldots, n$, $j = 1, \ldots, p$ and $k = 1, \ldots, K$. Denote $\tilde{\boldsymbol{\nu}} = (\tilde{v}_1, \ldots, \tilde{v}_n)$ and $\mathbf{s} = (s_1, \ldots, s_p)$. The above hierarchical model yields a normal conditional distribution for the observed response variable $\mathbf{y}$

$$p(\mathbf{y} \mid X, \tilde{v}, \boldsymbol{\beta}, \mathbf{b}, s, \tau, \eta^2, C, \mathbf{w}) = \exp\left\{-\frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{C_{ik}(y_i - b_{\theta_k} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \xi_{1k}\tilde{v}_i)^2}{\tau^{-1}\xi_{2k}^2 \tilde{v}_i}\right\} \prod_{i=1}^{n} \left(\frac{1}{2\pi\tau^{-1}\xi_{2k}^2\tilde{v}_i}\right)^{C_{ik}}.$$

The posterior distribution of all parameters is given by

$$p(\boldsymbol{\beta}, \mathbf{b}, \tau, \eta^2, \tilde{\mathbf{v}}, \mathbf{s}, \mathbf{w}, C \mid \mathbf{y}, X) \propto p(\mathbf{y} \mid X, \tilde{v}, \boldsymbol{\beta}, \mathbf{b}, s, \tau, \eta^2, C, \mathbf{w}) \prod_{i=1}^{n} p(\tilde{v}_i \mid \tau) \prod_{j=1}^{p} p(\beta_j, s_j \mid \eta^2) p(\tau, \eta^2) \pi(\mathbf{w}). \tag{9}$$

The expression (9) yields a tractable and efficient Gibbs sampler. The full conditional distributions of b_{θ_k} and β_k are normal distributions and those for τ and η^2 are gamma distributions. And the full conditional distributions of $\tilde{v}_i$ and s_j are generalized inverse Gaussian distributions. The full conditional distribution of w_k is Dirichlet distribution and those of C_{ik} is multinomial distribution. The details of the full conditional distribution are given in appendix.

4. Numerical studies

4.1. *Simulation*

In this section, we carry out Monte Carlo simulations to study the performance of Bayesian regularized composite quantile regression (BCQR) with comparison to the Bayesian regularized quantile regression (BQR) approach based on a single quantile. It has been shown in [15] that the Bayesian regularized quantile regression has advantage over its non-Bayesian counterpart.

We consider the linear model for testing variable selection methods that was used by Fan and Li [18] where the data are generated from

$$y_i = x_i^{\mathrm{T}}\boldsymbol{\beta} + u_i, \quad i = 1, \ldots, n. \tag{10}$$

We sample $\mathbf{x}$ from $N(0, \Sigma)$ with $\Sigma_{ij} = \rho^{|i-j|}$. The error u_i follows six different distributions: a normal distribution $N(0, 1)$, a mixture of normal distribution (MN) $0.5N(-2, 1) + 0.5N(2, 1)$, a Laplace distribution Laplace(3,1) with location 3 and scale 1, a mixture of Laplace distribution (ML) $0.5\text{Laplace}(-3, 1) + 0.5\text{Laplace}(3, 1)$, a gamma distribution $\Gamma(10, 1)$ and a beta distribution $B(0.5, 0.5)$.

The data have sample size $n = 100$. We simulate two settings with $p = 8$ and $p = 20$. The first setting corresponds to the dense case where $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$. The second setting corresponds to the sparse case where $(\beta_1, \beta_2, \beta_5) = (3, 1.5, 2)$ and the other 17 coefficients are equal to 0. For each setting and each error distribution, we repeat simulation 100 times. The criterion we used to compare different methods is the model error which is defined as

$$\mathrm{ME}(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \Sigma (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

For each simulated datum, the sampled parameter values from the first 3000 sweeps of the chain (burn-in period) were discarded from the analysis. Then we performed an additional 10,000 MCMC sweeps. After the burn in, the final sample of observations was selected every 20 sweeps to reduce serial correlation, resulting in 500 samples from the posterior. The estimated parameters are taken as the posterior mean from the 500 samples. We choose hyperparameters $a = b = c = d = 0.1$ and $\alpha_1 = \cdots = \alpha_K = 0.1$. Similar to the simulation setting in [3], we choose $K = 9$ and use the equally spaced quantiles: $\theta_k = k/(K + 1)$ for $k = 1, \ldots, K$. It has been shown in [2] that the relative efficient of composite quantile regression over single quantile regression converges to a limit as K approaches infinity. We have tried several other values of K from 5 to 20 and found that the numerical results are not sensitive to this choice (results are not shown for saving space).

Table 1 shows the model errors for two settings with random error generated from six distributions. For distributions with single mode such as normal, Laplace and gamma distributions, results from the two methods are very close. But for distributions with more than one modes such as normal mixture, Laplace mixture and beta distributions, the model errors from the BCQR method is much smaller than from the BQR method.

Table 2 presents the estimated $\boldsymbol{\beta}$ values from both the BQR and BCQR methods for Setting 1 with normally and mixture normally distributed random errors. Similar to Table 1, the results from the two methods are quite close to each other for normally distributed errors. But for the mixture normally distributed errors, estimations from our proposed BCQR method are much more accurate than from the BQR method. The results for the data generated from the other four choices of the error distributions are pretty similar and not shown here to save space. Table 2 indicates that the BCQR method gives smaller bias for the parameter estimation than the BQR method for distributions with more than one modes. The BQR method tends to underestimate the parameters in that situation.

For the sparse situation Setting 2, we can further test the variable selection property. Denote TP the number of correctly classified non-zero coefficients, i.e. the true positive number, and FP the number of incorrectly classified zero coefficients, i.e. the false positive number. In Bayesian framework, coefficient is classified as non-zero if the 95% highest posterior density interval of

Table 1. Summary table of model error over 100 replications for simulated data.

		$N(0,1)$	MN	Laplace	ML	Gamma	Beta
Setting 1	BQR	0.18	1.90	0.23	4.80	1.72	0.15
		(0.07)	(0.62)	(0.10)	(1.56)	(0.69)	(0.05)
	BCQR	0.18	0.28	0.25	0.29	1.69	0.01
		(0.07)	(0.12)	(0.12)	(0.13)	(0.70)	(0.01)
Setting 2	BQR	0.38	2.85	0.53	7.48	3.18	0.26
		(0.11)	(0.73)	(0.17)	(1.86)	(0.96)	(0.06)
	BCQR	0.46	0.99	0.70	0.84	3.13	0.08
		(0.13)	(0.38)	(0.20)	(0.29)	(1.04)	(0.03)

Note: Mean and STD (in parentheses) are reported.

Table 2. Summary table of $\boldsymbol{\beta}$ values over 100 replications for simulated data.

		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
True		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
$N(0,1)$	BQR	0.84 (0.01)	0.86 (0.01)	0.85 (0.01)	0.87 (0.02)	0.85 (0.01)	0.84 (0.01)	0.85 (0.01)	0.83 (0.01)
	BCQR	0.84 (0.01)	0.87 (0.01)	0.85 (0.01)	0.87 (0.01)	0.85 (0.01)	0.84 (0.01)	0.86 (0.01)	0.83 (0.01)
MN	BQR	0.77 (0.04)	0.83 (0.05)	0.81 (0.05)	0.77 (0.04)	0.85 (0.05)	0.89 (0.05)	0.79 (0.05)	0.78 (0.04)
	BCQR	0.85 (0.01)	0.85 (0.02)	0.86 (0.01)	0.84 (0.02)	0.85 (0.02)	0.87 (0.02)	0.85 (0.02)	0.83 (0.01)

Note: Mean and STD (in parentheses) are reported.

Table 3. Summary table of variable selection based on 100 replications of simulated data.

		$N(0,1)$	MN	Laplace	ML	Gamma	Beta
TP	BQR	3	2.67	3	1.52	2.63	3
	BCQR	3	2.98	3	3	2.67	3
FP	BQR	0.08	0.08	0.02	0.07	0.03	0.15
	BCQR	0.09	0.10	0.14	0.11	0.04	0.03

Note: Mean is reported.

this parameter does not cover zero. Otherwise, it is classified as zero. Table 3 shows the results of the TP and FP based on two methods for all six distributions. The number of true non-zero coefficients are 3 for this setting. It is shown that the two methods have similar performance for single mode distributions: normal, Laplace and gamma. But for multiple mode distributions like mixture normal, mixture laplace and beta, the BCQR method is more powerful than the BQR method in the sense that it can correctly identify the true coefficients more often than the BQR method and at the same time it can control the falsely discovered coefficients as well as the BQR method.

4.2. *Real data*

In this section, we apply the proposed method to Boston Housing Data which was first analysed by Harrison and Rubinfeld [19] and also used by Li et al. [15] for testing the Bayesian regularized quantile regression method. The data were downloaded from the 'spdep' package in R.[20] The data include 506 samples. We consider the relationship between the variable of the log-transformed corrected median value of owner-occupied housing (in 1000 USD) and 15 other explanatory variables. The detailed description of each variable is given in [19]. Similar to Li et al.,[15] we run 10-fold cross-validation to evaluate the performance of different methods. We first apply each method to the training dataset to estimate the corresponding parameters. The prediction accuracies of these methods are measured by the mean absolute prediction error (MAPE) and the corresponding standard deviation (STD), evaluated on the testing data. The MAPE of the BCQR method is 0.067 with STD 0.011. The results for the BQR method depend on the choice of θ. The first part of Table 4 summarizes the MAPE of BQR for five different choices of θ. The error of BCQR is slightly larger than the smallest error among the five choices of θ ($\theta = 0.5$) and much smaller than the other four choices. This indicates that the single quantile-based method

Table 4. First part: 10-fold cross-validation MAPE results for the Boston Housing Data based on the BQR method for five different choices of θ.

θ	0.1	0.3	0.5	0.7	0.9
MAPE	0.104 (0.008)	0.073 (0.010)	0.065 (0.009)	0.0076 (0.0015)	0.124 (0.007)
DIC	− 0.45	− 0.73	− 0.71	− 0.49	0.08

Note: Second part: DIC after fitting the BQR models to the whole dataset for five different choices of θ.

can achieve good performance if the quantile is chosen appropriately. But its performance can be worse if θ is not chosen appropriately. In contrast, our proposed composite method can always perform reasonably well because it combines regressions from multiple quantiles.

The deviance information criterion (DIC) is a hierarchical modelling generalization of the Akaike information criterion and Bayesian information criterion. It is particularly useful in Bayesian model selection problems where the posterior distributions of the models have been obtained by MCMC simulation.[21] The larger the DIC, the worse the fit. The DIC for fitting the BCQR model to the whole dataset is -0.68 and the results for the BQR model is listed in the second part of Table 4. The result from BQCR is slightly worse than the two best fits from BQR ($\theta = 0.3$ and 0.5), but much better than the fits from the other choices of θ. This is consistent with the conclusion based on the analysis of MAPE.

Figure 1 summarizes the point and interval estimations for the 15 regression parameters included in the analysis based on the two methods. A shift in horizontal direction is given to separate the two methods. The estimations from the two methods are very similar for most of the variables except variables 3 (per capita crime) and 9 (proportions of owner-occupied units built prior to 1940). The estimation of variable 3 from our method is much smaller than the estimation from the BQR method. The variable 9 is not significant in the BQR method but tends to be significant in our method. An interesting finding from Figure 1 is that for most of the variables, the 95% credible interval estimated from our method is smaller than the estimation from the BQR

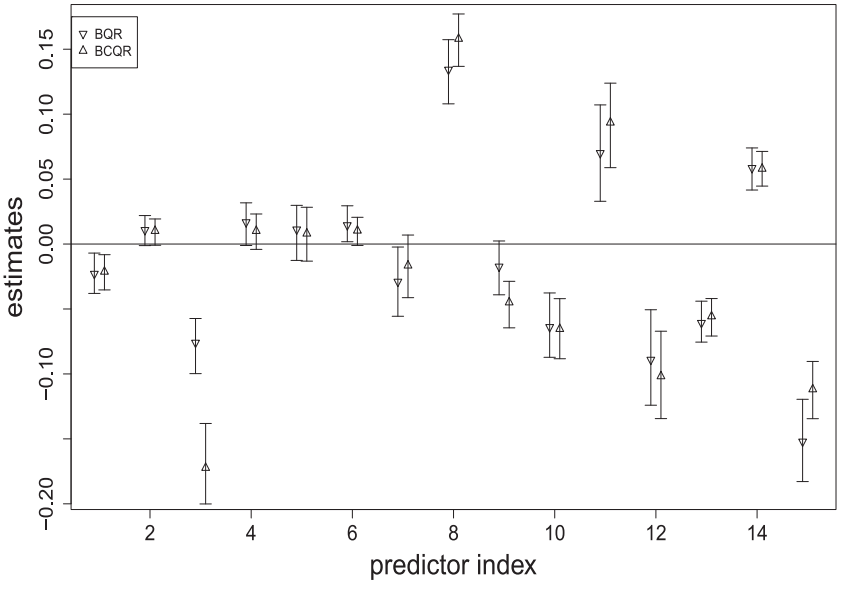

Figure 1. The estimates of the predictor effects for the Boston Housing Data using different methods. The 95% credible intervals given by BQR and BCQR are also plotted.

method. This indicates that the composite method can provide more precise prediction than the single quantile-based quantile regression method.

5. Discussion

In this paper, an efficient Bayesian method is proposed to combine multiple quantile regressions together with lasso regularization to conduct inference and variable selection simultaneously. Numerical studies show that our method is adaptive to unknown error distributions and outperforms the traditional single quantile-based quantile regression.

It is well known that the lasso regularization is not satisfactory for categorical predictors in the regression model since it only selects individual dummy variables instead of the whole predictor. To improve this, Yuan and Lin [22] introduced the grouped lasso by generalizing the lasso penalty. Other extensions of lasso include elastic net,[23] smoothly clipped absolute deviation [18] and the fussed lasso.[24] Li et al. [15] gave a general treatment to a set of regularizations in the Bayesian quantile regression (BQR) model including lasso, grouped lasso and elastic net penalties. In future research, we will also study the performance of different regularizations in out BCQR model.

In our application, we assume that the number of components K is fixed. But this restriction can be easily relaxed and K can be treated as a tuning parameter. We can further generalize the discretized θ_k in Equation (8) to continuous θ so that an infinite number of components can be included in the model. In that situation, we need to consider the joint distribution

$$p(y \mid X, \mathbf{b}, \boldsymbol{\beta}, \tau) \sim \prod_{i=1}^{n} \left(\int_0^1 w(\theta) p_\theta(y_i \mid \mathbf{x}_i, b_\theta, \boldsymbol{\beta}, \tau) \, \mathrm{d}\theta \right),$$

where the weight function $w(\theta)$ is a density function over $(0, 1)$. Computationally, this is a quite challenging problem for the frequentist method. In Bayesian framework, it can be solved by imposing a Dirichlet process prior on the density function $w(\theta)$.

Another future direction is to study a more general framework based on the weighted linear combination of different types of convex loss functions from Bayesian perspective. Bradic et al. [4] developed a composite quasi-likelihood approach which allows the combination of different regressions, e.g. the combination of quantile regression and least-square regression. Our Bayesian framework can be extended naturally to this context as well.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Koenker R, Bassett GW. Regression quantiles. Econometrica. 1978;46:33–50.
[2] Zou H, Yuan M. Composite quantile regression and the oracle model selection theory. Ann Stat. 2008;36:1108–1126.
[3] Zhao Z, Xiao Z. Efficient regressions via optimally combining quantile information. Econ Theory. 2014;30:1272–1314.
[4] Bradic J, Fan J, Wang W. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. J R Stat Soc: Ser B (Stat Methodol). 2011;73:325–349.
[5] Kottas A, Gelfand AE. Bayesian semiparametric median regression modeling. J Am Stat Assoc. 2001;96:1458–1468.
[6] Kozumi A, Kobayashi G. Gibbs sampling methods for Bayesian quantile regression. J Stat Comput Simul. 2011;81:1565–1578.

[7] Tsionas EG. Bayesian quantile inference. J Stat Comput Simul. 2003;73:659–674.
[8] Walker S, Mallick BK. A Bayesian semiparametric accelerated failure time model. Biometrics. 1999;55:477–483.
[9] Yu K, Chen CWS, Reed C, Dunson DB. Bayesian variable selection in quantile regression. Stat Interface. 2013;6:261–274.
[10] Yu K, Moyeed RA. Bayesian quantile regression. Stat Probab Lett. 2001;54:437–447.
[11] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc: Ser B. 1996;58(1):267–288.
[12] Koenker R. Quantile regression for longitudinal data. J Multivariate Anal. 2004;91:74–89.
[13] Li Y, Zhu J. l_1-Norm quantile regression. J Comput Graph Stat. 2004;17:163–185.
[14] Wang H, Li G, Jiang G. Robust regression shrinkage and consistent variable selection through the lad-lasso. J Bus Econ Stat. 2007;25:347–355.
[15] Li Q, Lin N, Xi R. Bayesian regularized quantile regression. Bayesian Anal. 2010;5:429–618.
[16] Alhamzawi R, Yu K, Benoit DF. Bayesian adaptive lasso quantile regression. Stat Model. 2013;13:223–252.
[17] Andrews DF, Mallows CL. Scale mixtures of normal distributions. J R Stat Soc: Ser B. 1974;36:99–102.
[18] Fan J, Li R. Variable selection via nonconcave penalized likelihood and it oracle properties. J Am Stat Assoc. 2001;96:1348–1360.
[19] Harrison D, Rubinfeld DL. Hedonic housing prices and the demand for clean air. J Environ Econ Manage. 1978;5:81–102.
[20] R Development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2013. Available from: http://www.R-project.org/
[21] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc: Ser B (Stat Methodol). 2002;64(4):583–639.
[22] Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika. 2007;94(1):19–35.
[23] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc: Ser B (Stat Methodol). 2005;67:301–320.
[24] Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. J R Stat Soc: Ser B (Stat Methodol). 2005;67:91–108.

Appendix

Let $\boldsymbol{\beta}_{-j}$ be the parameter vector $\boldsymbol{\beta}$ excluding the component $\boldsymbol{\beta}_j$, $\mathbf{s}_{-j}$ be the variable $\mathbf{s}$ excluding the component s_j, $\mathbf{b}_{-\theta_k}$ be the variable $\mathbf{b}$ excluding the component b_{θ_k} and $\tilde{\mathbf{v}}_{-i}$ be the variable $\tilde{\mathbf{v}}$ excluding the component $\tilde{v}_i$. The full conditional distribution of η^2 is a Gamma distribution

$$\begin{aligned} p(\eta^2 \mid X, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}, \tau, C, \mathbf{w}) &\propto \pi(\mathbf{s} \mid \eta^2)\pi(\eta^2) \\ &\propto (\eta^2)^{p+c-1} \exp\left\{-\left(\frac{1}{2}\sum_{j=1}^{p} s_j + d\right)\eta^2\right\} \\ &\propto \text{Gamma}\left(p + c, \frac{1}{2}\sum_{j=1}^{p} s_j + d\right). \end{aligned}$$

The full conditional distribution of τ is a Gamma distribution

$$\begin{aligned} &p(\tau \mid X, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}, \eta^2, C, \mathbf{w}) \\ &\quad\propto p(\mathbf{y} \mid X, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}, \tau, \eta^2, C, \mathbf{w})\pi(\tilde{\mathbf{v}} \mid \tau)\pi(\tau) \\ &\quad\propto \tau^{3n/2+a-1} \exp\left\{-\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{C_{ik}(y_i - b_{\theta_k} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \xi_{1k}\tilde{v}_i)^2}{\xi_{2k}^2\tilde{v}_i} + \sum_{i=1}^{n}\tilde{v}_i + b\right)\tau\right\} \\ &\quad\propto \text{Gamma}\left(\frac{3n}{2} + a, \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{C_{ik}(y_i - b_{\theta_k} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \xi_{1k}\tilde{v}_i)^2}{\xi_{2k}^2\tilde{v}_i} + \sum_{i=1}^{n}\tilde{v}_i + b\right). \end{aligned}$$

The full conditional distribution of $\tilde{v}_i$ is an inverse Gaussian distribution

$$\begin{aligned} &p(\tilde{v}_i \mid X, \mathbf{y}, \tilde{\mathbf{v}}_{-i}, \boldsymbol{\beta}, s, \tau, \eta^2, C, \mathbf{w}) \\ &\quad\propto p(\mathbf{y} \mid X, \tilde{\mathbf{v}}, \beta, s, \tau, \eta^2, C, \mathbf{w})\pi(\tilde{v}_i \mid \tau) \\ &\quad\propto \tilde{v}_i^{-1/2} \exp\left\{-\frac{1}{2}\sum_{k=1}^{K} C_{ik}\left[\left(\frac{\xi_{1k}^2}{\xi_{2k}^2} + 2\right)\tilde{v}_i + \frac{(y_i - b_{\theta_k} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2}{\xi_{2k}^2\tilde{v}_i}\right]\tau\right\} \end{aligned}$$

$$\propto \text{inverse Gaussian}\left(\lambda = \left(\sum_{k=1}^{K} \frac{C_{ik}\xi_{1k}^2}{\xi_{2k}^2} + 2\right)\tau, \mu = \sqrt{\sum_{k=1}^{K} \frac{C_{ik}(\xi_{1k}^2 + 2\xi_{2k}^2)}{(y_i - b_{\theta_k} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2}}\right).$$

The full conditional distribution of s_j is an inverse Gaussian distribution

$$\begin{aligned} p(s_j \mid X, \mathbf{y}, \tilde{\mathbf{v}}, \boldsymbol{\beta}, s_{-j}, \tau, \eta^2, C, \mathbf{w}) &\propto \pi(s_j \mid \eta^2)\pi(\beta_j \mid \eta^2) \\ &\propto s_j^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{\beta_j^2}{s_j} + \eta^2 s_j\right)\right\} \\ &\propto \text{inverse Gaussian}(\lambda = \eta^2, \mu = (\eta^2/\beta_j^2)^{1/2}). \end{aligned}$$

The full conditional distribution of β_j is a Gaussian distribution

$$\begin{aligned} &p(\beta_j \mid \mathbf{y}, X, \tilde{\mathbf{v}}, \boldsymbol{\beta}_{-j}, \mathbf{s}, \tau, \eta^2, C, \mathbf{w}) \\ &\quad\propto p(\mathbf{y} \mid X, \tilde{\mathbf{v}}, \boldsymbol{\beta}, \mathbf{s}, \tau, \eta^2, C, \mathbf{w})\pi(\beta_j \mid s_j) \\ &\quad\propto \exp\left\{-\frac{1}{2}\left[\left(\sum_{i=1}^{n} \frac{x_{ij}^2}{\tilde{\sigma}_i^2} + \frac{1}{s_j}\right)\beta_j^2 - 2\sum_{i=1}^{n} \frac{x_{ij}\tilde{y}_i}{\tilde{\sigma}_i^2}\beta_j\right]\right\} \\ &\quad\propto N\left(\frac{\sum_{i=1}^{n} x_{ij}\tilde{y}_i/\tilde{\sigma}_i^2}{\sum_{i=1}^{n} x_{ij}^2/\tilde{\sigma}_i^2 + 1/s_j}, \frac{1}{\sum_{i=1}^{n} x_{ij}^2/\tilde{\sigma}_i^2 + 1/s_j}\right), \end{aligned}$$

where

$$\begin{aligned} \tilde{\sigma}_i^2 &= \sum_{k=1}^{K} C_{ik}\tau^{-1}\xi_{2k}^2\tilde{v}_i, \\ \tilde{y}_i &= \sum_{k=1}^{K} C_{ik}(y_i - b_{\theta_k} - \mathbf{x}_{i,-j}^{\mathrm{T}}\boldsymbol{\beta}_{-j} - \xi_{1k}\tilde{v}_i). \end{aligned}$$

The full conditional distribution of b_{θ_k} is a Gaussian distribution

$$\begin{aligned} &p(b_{\theta_k} \mid \mathbf{y}, X, \tilde{\mathbf{v}}, \mathbf{b}_{-\theta_k}, \boldsymbol{\beta}, \mathbf{s}, \tau, \eta^2, C, \mathbf{w}) \\ &\quad\propto p(\mathbf{y} \mid X, \tilde{\mathbf{v}}, \mathbf{b}, \boldsymbol{\beta}, \mathbf{s}, \tau, \eta^2, C, \mathbf{w}) \\ &\quad\propto \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n} \frac{C_{ik}\tau}{\xi_{2k}^2\tilde{v}_i} b_{\theta_k}^2 - 2\sum_{i=1}^{n} \frac{C_{ik}\bar{y}_{ik}\tau}{\xi_{2k}^2\tilde{v}_i} b_{\theta_k}\right)\right\} \\ &\quad\propto N\left(\frac{\sum_{i=1}^{n}(C_{ik}\bar{y}_{ik}/\xi_{2k}^2\tilde{v}_i)}{\sum_{i=1}^{n}(C_{ik}/\xi_{2k}^2\tilde{v}_i)}, \frac{1}{\sum_{i=1}^{n}(C_{ik}\tau/\xi_{2k}^2\tilde{v}_i)}\right), \end{aligned}$$

where $\bar{y}_{ik} = y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \xi_{1k}\tilde{v}_i$. The full conditional distribution of $\mathbf{w}$ is a Dirichlet distribution

$$\begin{aligned} p(\mathbf{w} \mid \mathbf{y}, X, \tilde{\mathbf{v}}, \mathbf{b}, \boldsymbol{\beta}, \tau, \eta^2, C) &\propto \prod_{k=1}^{K} w_k^{n_k + \alpha_k} \\ &\propto \text{Dirichlet}(n_1 + \alpha_1, \ldots, n_K + \alpha_K), \end{aligned}$$

where n_k is the total number of objects in the kth cluster, i.e. $n_k = \sum_{i=1}^{n} C_{ik}$. The full conditional distribution of $\mathbf{C}_i = (C_{i1}, \ldots, C_{iK})$ is a multinomial distribution

$$\begin{aligned} &p(\mathbf{C}_i \mid \mathbf{y}, X, \tilde{\mathbf{v}}, \mathbf{b}, \beta, \tau, \eta^2, C_{-i}, \mathbf{w}) \\ &\quad\propto \prod_{k=1}^{K}\left\{\frac{w_k}{\xi_{2k}} \exp\left[-\frac{(y_i - b_{\theta_k} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \xi_{1k}\tilde{v}_i)^2\tau}{2\xi_{2k}^2\tilde{v}_i}\right]\right\}^{C_{ik}} \\ &\quad\propto \text{Multinomial}(1, \hat{p}_1, \ldots, \hat{p}_K), \end{aligned}$$

where

$$\hat{p}_k = \frac{(w_k/\xi_{2k}) \exp[-(y_i - b_{\theta_k} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \xi_{1k}\tilde{v}_i)^2\tau/2\xi_{2k}^2\tilde{v}_i]}{\sum_{k=1}^{K}(w_k/\xi_{2k}) \exp[-(y_i - b_{\theta_k} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \xi_{1k}\tilde{v}_i)^2\tau/2\xi_{2k}^2\tilde{v}_i]}.$$