Bernoulli **21**(2), 2015, 832–850
DOI: 10.3150/13-BEJ589

Bayesian quantile regression with approximate likelihood

YANG FENG[1], YUGUO CHEN[2] and XUMING HE[3]

[1]*Ads Metrics, Google Inc., Pittsburgh, PA 15206, USA. E-mail: yfeng@google.com*
[2]*Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA. E-mail: yuguo@illinois.edu*
[3]*Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA. E-mail: xmhe@umich.edu*

Quantile regression is often used when a comprehensive relationship between a response variable and one or more explanatory variables is desired. The traditional frequentists' approach to quantile regression has been well developed around asymptotic theories and efficient algorithms. However, not much work has been published under the Bayesian framework. One challenging problem for Bayesian quantile regression is that the full likelihood has no parametric forms. In this paper, we propose a Bayesian quantile regression method, the linearly interpolated density (LID) method, which uses a linear interpolation of the quantiles to approximate the likelihood. Unlike most of the existing methods that aim at tackling one quantile at a time, our proposed method estimates the joint posterior distribution of multiple quantiles, leading to higher global efficiency for all quantiles of interest. Markov chain Monte Carlo algorithms are developed to carry out the proposed method. We provide convergence results that justify both the algorithmic convergence and statistical approximations to an integrated-likelihood-based posterior. From the simulation results, we verify that LID has a clear advantage over other existing methods in estimating quantities that relate to two or more quantiles.

Keywords: Bayesian inference; linear interpolation; Markov chain Monte Carlo; quantile regression

1. Introduction

Quantile regression, as a supplement to the mean regression, is often used when a comprehensive relationship between the response variable y and the explanatory variables x is desired. Consider the following linear model:

$$y_i = x_i^T \beta + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{1.1}$$

where y_i is the response variable, x_i is a $p \times 1$ vector consisting of p explanatory variables, β is a $p \times 1$ vector of coefficients for the explanatory variables, and ε_i is the error term. The quantile regression analysis models the τth conditional quantile of y given x as:

$$Q_{y_i}(\tau | x_i) = x_i^T \beta(\tau), \qquad i = 1, 2, \ldots, n, \tag{1.2}$$

1350-7265 © 2015 ISI/BS

which is equivalent to (1.1) with $Q_{\varepsilon_i}(\tau|x_i) = 0$. The τ-specific coefficient vector $\beta(\tau)$ can be estimated by minimizing the loss function:

$$\min_{\beta(\tau)} \sum_{i=1}^{n} \rho_\tau\big(y_i - x_i^T \beta(\tau)\big), \tag{1.3}$$

where $\rho_\tau(u) = u\tau$ if $u \geq 0$, and $\rho_\tau(u) = u(\tau - 1)$ if $u < 0$; see Koenker [6].

To make inference on the quantile regression, one could use the asymptotic normal distribution of the estimates or use the bootstrap method. Aside from the regular bootstrap such as the residual bootstrap and the (x, y) bootstrap, one could also use Parzen, Wei and Ying [10]'s method or the Markov chain marginal bootstrap method (He and Hu [5]).

In contrast to the rich literature on quantile regression with the frequentist view, not much work has been done under the Bayesian framework. The most challenging problem for Bayesian quantile regression is that the likelihood is usually not available unless the conditional distribution for the error is assumed.

Yu and Moyeed [17] proposed an idea of employing a likelihood function based on the asymmetric Laplace distribution. In their work, Yu and Moyeed assumed that the error term follows an independent asymmetric Laplace distribution

$$f_\tau(u) = \tau(1 - \tau)\mathrm{e}^{-\rho_\tau(u)}, \qquad u \in R, \tag{1.4}$$

where $\rho_\tau(u)$ is the loss function of quantile regression. The asymmetric Laplace distribution is very closely related to quantile regression since the mode of $f_\tau(u)$ is the solution to (1.3). Reich, Bondell and Wang [11] developed a Bayesian approach for quantile regression assuming that the error term follows an infinite mixture of Gaussian densities and their prior for the residual density is stochastically centered on the asymmetric Laplace distribution. Kottas and Gelfand [7] implemented a Bayesian median regression by introducing two families of distributions with median zero and the Dirichlet process prior. Dunson and Taylor [4] used a substitution likelihood proposed by Lavine [9] to make inferences based on the posterior distribution. One property of Dunson and Taylor's method is that it allows regression on multiple quantiles simultaneously. Tokdar and Kadane [15] proposed a semiparametric Bayesian approach for simultaneous analysis of quantile regression models based on the observation that when there is only a univariate covariate, the monotonicity constraint can be satisfied by interpolating two monotone curves, and the Bayesian inference can be carried out by specifying a prior on the two monotone curves. Taddy and Kottas [14] developed a fully nonparametric model-based quantile regression based on Dirichlet process mixing. Kottas and Krnjajić [8] extended this idea to the case where the error distribution changes nonparametrically with the covariates. Recently, Yang and He [16] proposed a Bayesian empirical likelihood method which targets on estimating multiple quantiles simultaneously, and justified the validity of the posterior based inference.

In this paper, we propose a Bayesian method, which aims at estimating the joint posterior distribution of multiple quantiles and achieving "global" efficiency for quantiles of interest. We consider a Bayesian approach to estimating multiple quantiles as follows. Let $\tau_1, \ldots, \tau_m$ be m quantiles in model (1.2) and $B_m = (\beta(\tau_1), \ldots, \beta(\tau_m))$. Let $X = (x_1, \ldots, x_n)'$ and $Y = (y_1, \ldots, y_n)$ be the observations of size n. For each pair of observation (x_i, y_i), the likelihood

$L(B_m|x_i, y_i) = p(y_i|x_i, B_m)$ is not available. However if we include f_i, the probability density function (pdf) of the conditional distribution $y|x_i$, as the nuisance parameter, then the likelihood $L(B_m, f_i|x_i, y_i) = p(y_i|x_i, B_m, f_i) = f_i(y_i)$. This is to treat Bayesian quantile regression as a semi-parametric problem: the parameter of interest is finite dimensional and the nuisance parameter is nonparametric. To eliminate the nuisance parameter, we use the integrated likelihood methods recommended by Berger, Liseo and Wolpert [1]. More specifically, let θ_{f_i} be all the quantiles of f_i, and $\theta_{m,i} = x_i B_m$ be the m quantiles of interest. We can define $p(y_i|x_i, B_m)$ as

$$p(y_i|x_i, B_m) = \int_{f_i \in \mathscr{F}_{\theta_{m,i}}} p(y_i|\theta_{f_i})\, \mathrm{d}\Pi_{\theta_{m,i}}(f_i), \tag{1.5}$$

where $\mathscr{F}_{\theta_{m,i}}$ denotes the subset of well-behaved pdfs (will be defined precisely in Section 3.2) with those m quantiles equal to $\theta_{m,i}$, $\Pi_{\theta_{m,i}}(\cdot)$ denotes the prior on $f_i|\theta_{m,i} \in \mathscr{F}_{\theta_{m,i}}$ (will be specified in Section 3.2), and $p(y_i|\theta_{f_i}) = f_i(y_i)$ because $f_i(y|x_i)$ is determined by the conditional quantile functions. Here, $p(y_i|x_i, B_m)$ can be viewed as an integral of a function or an expectation with the densities as the random variable. The posterior distribution of $B_m|X, Y$ can be written as

$$p(B_m|X, Y) \propto \pi_m(B_m|X) L(Y|X, B_m), \tag{1.6}$$

where $\pi_m(B_m|X)$ is the prior on B_m and $L(Y|X, B_m) = \prod_{i=1}^n p(y_i|x_i, B_m)$.

One practical difficulty with the above approach is that the integration step to remove the nuisance parameter is computationally infeasible except for the case of $m = 1$ (Doss [3]). To circumvent this issue, we consider a different approximation to the likelihood. Note that $x_i B_m$ gives the m quantiles of the conditional distribution $y|x_i$ based on model (1.2). These m quantiles can be used to construct an approximate conditional distribution $y|x_i$ through linear interpolation. With this approximate likelihood, an approximate posterior distribution becomes available. We show that the total variation distance between the approximate posterior distribution and $p(B_m|X, Y)$ (the posterior based on the integrated likelihood) goes to 0 as $\tau_1, \ldots, \tau_m$ becomes dense in $(0, 1)$ as $m \to \infty$. A Markov chain Monte Carlo (MCMC) algorithm can then be developed to sample from the approximate posterior distribution. The recent work of Reich, Fuentes and Dunson [12] used large-sample approximations to the likelihood to do Bayesian quantile regression. Their approach also aims to achieve global efficiency over multiple quantiles, and can adapt to account for spatial correlation. In contrast, our work uses approximations at a fixed sample size n and provides a Bayesian interpretation of the posterior quantities.

The rest of the paper is organized as follows. Section 2 introduces the proposed method. Section 3 provides the convergence property of the algorithm as well as the approximate posterior distribution. Section 4 compares the proposed method with some existing methods through simulation studies and applies the proposed method to real data. Section 5 provides concluding remarks.

2. Methodology

In this section, we describe the linearly interpolated density to be used in approximating the likelihood, and then give the layout of our MCMC algorithm for posterior inference. We list

again the basic setting introduced in Section 1. Let $X = (x_1, \ldots, x_n)'$ and $Y = (y_1, \ldots, y_n)$ be the observations. Let $\tau_1, \ldots, \tau_m$ be m quantiles in model (1.2) and $B_m = (\beta(\tau_1), \ldots, \beta(\tau_m))$. We are interested in the posterior distribution $B_m|X, Y$.

2.1. Linearly interpolated density

The likelihood is generally not assumed under the quantile regression model, but $x_i B_m$ gives the m quantiles of the conditional distribution $y|x_i$. With the linearly interpolated density based on the m quantiles, we can approximate the true likelihood from a sequence of specified quantile functions.

Here is how the linear interpolation idea works in a simple setting. Suppose $Z \sim F(z)$, where $F(z)$ is the cumulative distribution function (cdf) of Z. Let $f(z)$ be the pdf of Z. Let $\tau_z = F(z)$, and τ_1, τ_2 be two constants such that $0 \le \tau_1 < \tau_z < \tau_2 \le 1$. Then $F^{-1}(\tau_1) < z < F^{-1}(\tau_2)$ if $f(z)$ is continuous and non-zero on the support of Z. We can approximate $f(z)$ by

$$\frac{\tau_2 - \tau_1}{F^{-1}(\tau_2) - F^{-1}(\tau_1)}, \tag{2.1}$$

because

$$\frac{\tau_2 - \tau_1}{F^{-1}(\tau_2) - F^{-1}(\tau_1)} = \frac{\tau_2 - \tau_1}{\frac{\mathrm{d}}{\mathrm{d}\tau} F^{-1}(\tau^*)(\tau_2 - \tau_1)} = f\big(z^*\big), \tag{2.2}$$

where $\tau_1 < \tau^* < \tau_2$ and $z^* = F^{-1}(\tau^*) \in (F^{-1}(\tau_1), F^{-1}(\tau_2))$.

Now we extend the interpolation idea to model (1.2). Given $B_m = (\beta(\tau_1), \beta(\tau_2), \ldots, \beta(\tau_m))$, we could calculate the linearly interpolated density $\hat{f}_i(y_i|x_i, B_m)$, $i = 1, 2, \ldots, n$, by

$$\begin{aligned} \hat{f}_i(y_i|x_i, B_m) = {} & \left[\sum_{j=1}^{m-1} I_{\{y_i \in (x_i\beta(\tau_j), x_i\beta(\tau_{j+1}))\}} \frac{\tau_{j+1} - \tau_j}{x_i\beta(\tau_{j+1}) - x_i\beta(\tau_j)}\right] \\ & + I_{\{y_i \in (-\infty, x_i\beta(\tau_1))\}} \tau_1 f_1(y_i) + I_{\{y_i \in (x_i\beta(\tau_m), \infty)\}} (1 - \tau_m) f_2(y_i), \end{aligned} \tag{2.3}$$

where f_1 is distributed as the left half of $N(x_i\beta(\tau_1), \sigma^2)$, f_2 is distributed as the right half of $N(x_i\beta(\tau_m), \sigma^2)$, and σ^2 is some pre-specified parameter.

Let $\hat{p}_m(Y|X, B_m) = \prod_{i=1}^n \hat{f}_i(y_i|x_i, B_m)$ denote the approximate likelihood. One possible prior $\pi_m(B_m|X)$ on B_m is a truncated normal $N(\mu, \Sigma)$ satisfying

$$x_i\beta(\tau_1) < x_i\beta(\tau_2) < \cdots < x_i\beta(\tau_m), \qquad i = 1, 2, \ldots, n. \tag{2.4}$$

Since we include the intercept in model (1.2), the first element of x_i is 1, and at least the parallel quantile regression lines satisfy (2.4). The corresponding posterior is

$$\hat{p}_m(B_m|X, Y) = \frac{\pi_m(B_m|X)\hat{p}_m(Y|X, B_m)}{\hat{p}_m(Y|X)}, \tag{2.5}$$

where $\hat{p}_m(Y|X) = \int \pi_m(B_m|X)\hat{p}_m(Y|X, B_m)\,\mathrm{d}B_m$. In the next section, we give a MCMC algorithm to sample B_m from this posterior. We show later that the total variation distance between this posterior distribution and the target posterior $p(B_m|X, Y)$ goes to 0 as m goes to infinity.

2.2. Algorithm of the linearly interpolated density (LID) method

We incorporate the linearly interpolated density into the following modified Metropolis–Hastings algorithm to draw samples from $\hat{p}_m(B_m|X, Y)$.

1. Choose an initial value B_m^0 for B_m. One good choice is to use the parallel quantile estimates, that is, all the slopes for the quantiles are the same and the intercepts are different. We could use the *quantreg* (a function in R) estimates of the slopes for the median as the initial slopes, and use the *quantreg* estimates of the intercepts for each quantile as the initial intercepts. In case a lower quantile has a larger intercept than an upper quantile, we could order the intercepts such that the intercepts increase with respect to τ. If there are ties, we could add an increasing sequence with respect to τ to the intercepts to distinguish them. Another possible choice for the initial value is to use Bondell, Reich and Wang [2]'s estimate which guarantees the non-crossing of the quantiles.
2. Approximate the densities. With the initial values of the parameters, we can calculate the linearly interpolated density $\hat{f}_i^0(y_i|x_i, B_m^0)$, $i = 1, 2, \ldots, n$, by plugging B_m^0 into equation (2.3). Let $L^0 = \prod_{i=1}^n \hat{f}_i^0(y_i|x_i, B_m^0)$.
3. Propose a move. Suppose we are at the kth iteration. Randomly pick a number τ_j from $\tau_1, \tau_2, \ldots, \tau_m$ and then randomly pick a component $\beta_l^{k-1}(\tau_j)$ of $\beta^{k-1}(\tau_j)$ to update. To make sure that the proposed point $\beta_l^*(\tau_j)$ satisfies constraint (2.4), we can calculate a lower bound $l_{j,l}$ and an upper bound $u_{j,l}$ for $\beta_l^*(\tau_j)$ and generate a value for $\beta_l^*(\tau_j)$ from Uniform$(l_{j,l}, u_{j,l})$. In case $l_{j,l} = -\infty$ or $u_{j,l} = \infty$, we will use a truncated normal as the proposal distribution. The details on how to find the bounds are in Appendix A.1. Denote $\beta^*(\tau_j)$ as the updated $\beta^{k-1}(\tau_j)$ by replacing its lth component $\beta_l^{k-1}(\tau_j)$ by the proposed value $\beta_l^*(\tau_j)$.
4. Set $B_m^* = (\beta^{k-1}(\tau_1), \ldots, \beta^{k-1}(\tau_{j-1}), \beta^*(\tau_j), \beta^{k-1}(\tau_{j+1}), \ldots, \beta^{k-1}(\tau_m))$. We can calculate the linearly interpolated density $\hat{f}_i^*(y_i|x_i, B_m^*)$, $i = 1, 2, \ldots, n$, by plugging B_m^* into equation (2.3). Let $L^* = \prod_{i=1}^n \hat{f}_i^*(y_i|x_i, B_m^*)$.
5. Calculate the acceptance probability

$$r = \min\left(1, \frac{\pi_m(B_m^*|X)L^*q(B_m^* \to B_m^{k-1})}{\pi_m(B_m^{k-1}|X)L^{k-1}q(B_m^{k-1} \to B_m^*)}\right), \tag{2.6}$$

   where $q(B_m^{k-1} \to B_m^*)$ denotes the transition probability from B_m^{k-1} to B_m^*. Notice that these two transition probabilities cancel out if we choose symmetric proposals. Let $B_m^k = B_m^*$ with probability r, and $B_m^k = B_m^{k-1}$ with probability $1 - r$. If $B_m^k = B_m^*$, then $L^k = L^*$; otherwise $L^k = L^{k-1}$.
6. Repeat steps 3–5 until the desired number of iterations is reached.

3. Theoretical properties

In this section, we give the stationary distribution of the Markov chain in Section 2.2 for fixed m, and study the limiting behavior of the stationary distribution as $m \to \infty$.

3.1. Stationary distribution

Since we replace the true probability density function by the linearly interpolated density in the Metropolis–Hastings algorithm in Section 2.2, it is not obvious what the stationary distribution of the Markov chain is. The following theorem, whose proof is in Appendix A.2, says that the Markov chain converges to $\hat{p}_m(B_m|X, Y)$ defined in (2.5).

Theorem 3.1. *The stationary distribution of the Markov chain constructed in Section 2.2 is* $\hat{p}_m(B_m|X, Y)$.

This theorem implies that we can use the algorithm in Section 2.2 to draw samples from $\hat{p}_m(B_m|X, Y)$.

3.2. Limiting distribution

In this section, we show that as $m \to \infty$, the total variation distance between the stationary distribution $\hat{p}_m(B_m|X, Y)$ and the target distribution $p(B_m|X, Y)$ (defined in (1.6)) goes to 0. The proof requires the following assumption about f_i, the probability density function of the conditional distribution $y|x_i$. All the results are stated for a given sample size n.

Assumption 3.1. *Let $q_{f,\tau}$ be the τth quantile of f, and M_1, M_2 and c be constants. The densities of $y|x_i$ are in the set $\mathscr{F} = \{f | \int f\,\mathrm{d}x = 1, 0 \le f \le M_1, |f'| < M_2$, and $f(x) < c/\sqrt{m}$ for $x < q_{f,1/m}$ and for $x > q_{f,(m-1)/m}, m = 2, 3, \ldots\}$.*

The assumption implies that $\mathscr{F}$ is a set of bounded probability density functions with bounded first derivatives and controlled tails. The restrictions on the tails are not hard to satisfy. The Cauchy distribution, for example, is in the set. For the Cauchy distribution, the $\frac{1}{m}$th quantile is $q_{1/m} = \tan(\pi(\frac{1}{m} - \frac{1}{2})) = -\operatorname{ctan}(\frac{\pi}{m})$, so $f(q_{1/m}) = \frac{1}{\pi}\frac{1}{1+\operatorname{ctan}^2(\pi/m)} = \frac{1}{\pi}\sin^2(\frac{\pi}{m}) = \mathrm{O}(\frac{1}{m^2}) < \frac{c}{\sqrt{m}}$ for some c. The set $\mathscr{F}_{\theta_{m,i}}$ appeared in (1.5) denotes the subset of $\mathscr{F}$ that contains all the pdfs with those m quantiles equal to $\theta_{m,i} = x_i B_m$.

We now specify the prior on $f_i(\cdot|x_i) \in \mathscr{F}$, denoted by $\Pi(f_i)$, and the prior on $f_i|\theta_{m,i} \in \mathscr{F}_{\theta_{m,i}}$, denoted by $\Pi_{\theta_{m,i}}(f_i)$. We know from (1.2) that the τth quantile of $f_i(\cdot|x_i)$, the conditional distribution of y given $x = x_i$, is $x_i^T\beta(\tau)$. Let us consider $\beta(\tau)$ as a function of τ, where $0 \le \tau \le 1$. Because $x_i^T\beta(\tau)$, $0 \le \tau \le 1$, determines all the quantiles of $f_i(\cdot|x_i)$ based on (1.2), and therefore determines $f_i(\cdot|x_i)$ (Koenker [6]), the prior on $f_i(\cdot|x_i)$ can be induced from the prior on $\beta(\tau)$. To satisfy Assumption 3.1, we use a Gaussian process prior on $\beta''(\tau)$ so that $\beta(\tau)$ has the second derivative, and then f_i's have the first derivative. The prior $\Pi(f_i)$ on $f_i(\cdot|x_i)$ is induced from the

prior on $\beta(\tau)$. The prior $\Pi_{\theta_{m,i}}(f_i)$ on $f_i|\theta_{m,i}$ is induced by $\Pi(f_i)$. The prior on B_m can be obtained from the prior on $\beta(\tau)$, because B_m is a vector of m points on $\beta(\tau)$. With the specification of these priors, $p(y_i|x_i, B_m)$ and $p(B_m|X, Y)$ given in (1.5) and (1.6) are well-defined.

To study the limiting distribution as $m \to \infty$, we assume the sequence of quantile levels satisfies the following condition:

$$\Delta\tau = \max_{0\le j\le m}(\tau_{j+1} - \tau_j) = \mathrm{O}\left(\frac{1}{m}\right), \tag{3.1}$$

where $\tau_0 = 0$ and $\tau_{m+1} = 1$. This condition is not difficult to satisfy. For example, we can start from $m_0 = M_0$ quantile levels: $\tau = \frac{1}{M_0+1}, \frac{2}{M_0+1}, \ldots, \frac{M_0}{M_0+1}$, which include the quantiles of interest. We add new τ's one by one so that the new τ divides one of the previous intervals in halves, that is, $\tau = \frac{1}{2(M_0+1)}, \frac{3}{2(M_0+1)}, \ldots, \frac{2M_0+1}{2(M_0+1)}, \frac{1}{4(M_0+1)}, \frac{3}{4(M_0+1)}, \ldots, \frac{4M_0+3}{4(M_0+1)}$ and so on. For this sequence of quantiles, we have $\Delta\tau = \max_{0\le j\le m}(\tau_{j+1} - \tau_j) \le \frac{2}{m} = \mathrm{O}(\frac{1}{m})$.

To prove the convergence of distributions, we use the total variation norm, $\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \sup_A |\mu_1(A) - \mu_2(A)|$ for two probability measures μ_1 and μ_2, where A denotes any measurable set. It is more convenient to use the following equivalent definition (Robert and Casella [13], page 253): $\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \frac{1}{2}\sup_{|h|\le 1} |\int h(x)\mu_1(\mathrm{d}x) - \int h(x)\mu_2(\mathrm{d}x)|$. The following theorem gives the limiting distribution of the stationary distribution as $m \to \infty$.

Theorem 3.2. $\|\hat{p}_m(B_m|X, Y) - p(B_m|X, Y)\|_{\mathrm{TV}} \to 0$ *as* $m \to \infty$, *assuming* $\tau_{j+1} - \tau_j = \mathrm{O}(\frac{1}{m})$.

The proof is in Appendix A.3. As a consequence of Theorem 3.2, we have the following corollary.

Corollary 3.1. *Let* η *be the quantiles of interest, which is contained in* B_m. *We have* $\|\hat{p}_m(\eta|X, Y) - p(\eta|X, Y)\|_{\mathrm{TV}} \to 0$ *as* $m \to \infty$, *assuming* $\tau_{j+1} - \tau_j = \mathrm{O}(\frac{1}{m})$.

The above corollary says that by the linearly interpolated density approximation the posterior distribution of the quantiles of interest converges to the target distribution. The theorem requires that we need to increase m in the algorithm. Although m is fixed in applications, the convergence result lends support to $\hat{p}_m(B_m|X, Y)$ as an approximation.

4. Comparison of LID with other methods

In this section, we compare the proposed method with some existing methods through three simulation studies. In the quantile regression model (1.2), if the conditional densities $f_i(y|x_i)$ are different for different observation i, one could apply weighted quantile regression to improve the efficiency of estimates (Koenker [6], page 160). In this case, the loss function would be:

$$\min_{\beta(\tau)} \sum_{i=1}^{n} w_i \rho_\tau\big(y_i - x_i^T \beta(\tau)\big), \tag{4.1}$$

where w_i denotes the weight for the ith observation. The optimal weight is the conditional density $f_i(y|x_i)$ at the τth quantile. Because the density is not available generally, one could approximate the density by a nonparametric density estimate. One simple way is to use

$$\hat{w}_i = \frac{2\Delta\tau}{x_i^T(\beta^{rq}(\tau+\Delta\tau) - \beta^{rq}(\tau-\Delta\tau))}, \qquad i = 1, 2, \ldots, n, \tag{4.2}$$

where β^{rq} denotes the unweighted quantile regression estimate. When the weight is negative due to crossing of quantile estimates, we just set the weight to be 0. This occurs with probability tending to 0 as n increases. To make inference, one could use the asymptotic normal distribution of the estimates or use the bootstrap method.

4.1. Example 1

The data were generated from the following model

$$y_i = a + bx_i + (1 + x_i)\varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{4.3}$$

where ε_i's are independent and identically distributed (i.i.d.) as $N(0, 1)$. We chose $n = 100$, $a = 5$ and $b = 1$. The covariate x_i was generated from lognormal(0, 1). The corresponding quantiles of interest are

$$Q_{y_i}(\tau|x_i) = a(\tau) + b(\tau)x_i, \qquad i = 1, 2, \ldots, n, \tau = \frac{1}{m+1}, \ldots, \frac{m}{m+1}. \tag{4.4}$$

Here we report the results on the 0.25, 0.5 and 0.75 quantiles and the difference between the 0.75 and 0.5 quantiles by comparing the mean squared error (MSE) for the slope estimates from five different methods: the proposed linearly interpolated density method (LID), the regular regression of quantiles (RQ), the weighted RQ with estimated weights (EWRQ) (Koenker [6]), the pseudo-Bayesian method of Yu and Moyeed [17], and the approximate Bayesian method of Reich, Fuentes and Dunson [12]. We generated 100 data sets for computing the MSE.

For LID and Yu and Moyeed's method, we used the normal prior $N(0, 100)$ for each parameter $a(\tau)$ and $b(\tau)$. For LID, we chose $m = 49$, equally spaced quantiles between 0 and 1 (which include the quantiles of interest: 0.25, 0.5 and 0.75), and the length of the Markov chain is 1 000 000 (half of the samples were used as burn-in). We ran such a long chain because we updated 98 parameters one at a time, which means we updated each parameter about 10 000 times on average. Every thousandth sample in the chain is taken for the posterior inference. For Yu and Moyeed's method, a Markov chain with length 5 000 (half of the samples were used as burn-in) seems enough for the inference, partially because Yu and Moyeed's method is dealing with one quantile at a time and has only two parameters. For Reich et al.'s method, we simply used their code and set the length of the chain to be 2 000 (half of the samples were used as burn-in). Notice that for LID and Reich et al.'s method, only one run is needed to provide all results in the table, and other methods have to run for each τ.

From the results in Table 1, we can see that LID did better than RQ and Yu and Moyeed's method. Comparing with weighted RQ and Reich et al.'s method, LID gave better estimates

Table 1. $n \times$ MSE and its standard error (in parentheses) for Example 1

Methods	$b(0.25)$	$b(0.5)$	$b(0.75)$	$b(0.75) - b(0.5)$
RQ	23 (4)	19 (2)	19 (3)	15 (2)
EWRQ	16 (2)	13 (2)	15 (3)	11 (2)
LID	22 (4)	15 (2)	13 (1)	3 (0.6)
Yu and Moyeed	21 (4)	17 (2)	16 (3)	10 (1)
Reich et al.	16 (2)	15 (2)	23 (3)	11 (1)

for upper quantiles but poorer estimates for lower quantiles. For estimating the differences of quantiles, LID is clearly the best among all the methods.

4.2. Example 2

The data were generated from the following model

$$y_i = a + bx_{1,i} + cx_{2,i} + (1 + x_{1,i} + x_{2,i})\varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{4.5}$$

where ε_i's are i.i.d. from $N(0, 1)$. In the simulations, we chose $n = 100$, $a = 5$, $b = 1$, and $c = 1$. The covariates $x_{1,i}$ was generated from lognormal$(0, 1)$ and $x_{2,i}$ was generated from Bernoulli(0.5). The corresponding quantiles of interest are

$$Q_{y_i}(\tau|x_i) = a(\tau) + b(\tau)x_{1,i} + c(\tau)x_{2,i}, \qquad i = 1, 2, \ldots, n, \tau = \frac{1}{m+1}, \ldots, \frac{m}{m+1}. \tag{4.6}$$

We compared the five methods with the same performance criterion as Example 1. We generated 400 data sets for computing the MSE. The results are in Table 2. We see that for the quantile estimates, LID (with $m = 15$) and EWRQ perform similarly, and LID outperforms RQ and Yu and Moyeed's method. For estimating the difference between quantiles, LID outperforms RQ, EWRQ, and Yu and Moyeed's method. Comparing with Reich et al.'s method, LID gave better estimates for parameter b but poorer estimates for parameter c.

From the two simulation studies, we can see that most of the time the proposed LID method works as well as the weighted RQ, and outperforms RQ and Yu and Moyeed's method, for

Table 2. $n \times$ MSE and its standard error (in parenthesis) for Example 2

Methods	$b(0.5)$	$b(0.75)$	$b(0.75) - b(0.5)$	$c(0.5)$	$c(0.75)$	$c(0.75) - c(0.5)$
RQ	22 (3)	25 (3)	20 (3)	47 (9)	52 (7)	42 (6)
EWRQ	15 (2)	19 (3)	16 (2)	46 (8)	49 (8)	40 (6)
LID	17 (2)	18 (2)	2.9 (0.4)	36 (5)	42 (6)	18 (2)
Yu and Moyeed	20 (2)	21 (3)	13 (2)	42 (7)	45 (6)	28 (4)
Reich et al.	20 (3)	29 (5)	11 (1)	4.2 (0.6)	8.6 (1.1)	3.1 (0.3)

estimating quantiles. LID performs better than Reich et al.'s method in some cases and is outperformed by Reich et al.'s method in others. LID has a significant advantage over other methods in estimating the difference of quantiles. When several quantiles are of interest, including their differences, there is a clear efficiency gain in using LID.

4.3. Empirical studies

In this section, we look at the June 1997 Detailed Natality Data published by the National Center for Health Statistics. Following the analysis in Koenker ([6], page 20), we use 65 536 cases of recorded singleton births. We consider the following quantile model for the birth weight data:

$$Q_{y_i}(\tau|x_i) = a(\tau) + b(\tau)x_{i,1} + c(\tau)x_{i,2} + d(\tau)x_{i,3} + e(\tau)x_{i,4}, \qquad i = 1, 2, \ldots, n, \tag{4.7}$$

where $x_{i,1}$ is the indicator function that indicates whether the mother went to prenatal care for at least two times, $x_{i,2}$ is the indicator function that indicates whether the mother smoked or not, $x_{i,3}$ is mother's weight gain in pounds during pregnancy, and $x_{i,4}$ is the square of mother's weight gain. The mother's weight gain enters the model as a quadratic following the discussion in Koenker ([6], page 23). To make the results more comparable, we consider a slight modification of model (4.7):

$$Q_{y_i}(\tau|x_i) = a(\tau) + b(\tau)x_{i,1} + c(\tau)x_{i,2} + d^*(\tau)x^*_{i,3} + e^*(\tau)x^*_{i,4}, \qquad i = 1, 2, \ldots, n, \tag{4.8}$$

where $x^*_{i,3}$ denotes the standardized mother's weight gain during pregnancy and $x^*_{i,4}$ denotes the standardized square of mother's weight gain. We compared the results from RQ and LID (with $m = 39$) for the full data set. Here we focus on the 0.1, 0.25, and 0.5 quantiles. The results are in Table 3. From the results, we can see that the estimates from both methods are very close. The standard error from LID seems to be smaller than that from RQ.

To see how good the estimates are, we compared the estimated conditional quantile with the local quantile estimated nonparametrically. We considered two subsets of the full data. For the first subset of the data, we selected $x_{i,1} = 1$, $x_{i,2} = 1$, and $24.5 < x_{i,3} < 25.5$, within which range there are 96 observations. For the second subset of the data, we selected $x_{i,1} = 1$, $x_{i,2} = 0$, and $44.5 < x_{i,3} < 45.5$, within which range there are 1318 observations. Then we calculated the quantile of y_i in each subset of the data as the local quantile, and compared it with the predicted quantiles from RQ and LID. The results are presented in Table 4. From the results, we can see that all the estimated quantiles are very close to the local quantile estimates.

Table 3. Estimates of the parameters and their standard errors (in parentheses) for the birth weight data

Methods	$b(0.1)$	$c(0.1)$	$d^*(0.1)$	$e^*(0.1)$	$b(0.25)$	$c(0.25)$	$d^*(0.25)$	$e^*(0.25)$	$b(0.5)$	$c(0.5)$	$d^*(0.5)$	$e^*(0.5)$
RQ	−0.030	−0.22	0.37	−0.21	−0.049	−0.22	0.19	−0.075	−0.061	−0.22	0.127	−0.020
	(0.009)	(0.01)	(0.02)	(0.02)	(0.008)	(0.008)	(0.011)	(0.012)	(0.006)	(0.007)	(0.008)	(0.008)
LID	−0.045	−0.22	0.36	−0.22	−0.052	−0.23	0.20	−0.081	−0.061	−0.23	0.131	−0.026
	(0.007)	(0.003)	(0.002)	(0.003)	(0.001)	(0.002)	(0.008)	(0.007)	(0.003)	(0.003)	(0.002)	(0.002)

Table 4. Estimates of the local quantile

	$x_{i,1}=1$, $x_{i,2}=1$, and $x_{i,3}=25$			$x_{i,1}=1$, $x_{i,2}=0$, and $x_{i,3}=45$		
Quantile	Local quantile	RQ	LID	Local quantile	RQ	LID
0.1	2.54	2.44	2.43	2.89	2.90	2.88
0.25	2.81	2.76	2.75	3.18	3.17	3.17
0.5	3.02	3.07	3.07	3.54	3.47	3.46

Another way to check the model fitness is to build the model by leaving out a portion of the data, and then evaluate the model performance on the out-of-bag portion of the data. Here we compared the out-of-bag quantile coverage (the percentage of the testing data that fall below the τth quantile line) by randomly selecting 10% of the data as the out-of-bag testing data and using the rest as the training data. The results based on a random splitting are summarized in Table 5. We can see that both RQ and LID have coverages similar to the nominal values.

From this example we can see that the model parameter estimates, including the quantiles, from both RQ and LID are very similar, but LID estimates are associated with lower standard errors, which corroborates our findings in simulation studies.

5. Conclusion

In this paper we proposed a Bayesian method for quantile regression which estimates multiple quantiles simultaneously. We proved the convergence of the proposed algorithm, i.e., the stationary distribution of the Markov chain constructed by LID would converge to the target distribution as the number of quantiles m goes to infinity. In the simulation studies, we found that choosing $m = 15$ already gave satisfactory results. In the comparison of the proposed LID method with other methods, LID provides comparable results for quantile estimation, and gives much better estimates of the difference of the quantiles than other methods (RQ, weighted RQ, and Yu and Moyeed's method).

The LID method is computationally intensive, and it requires longer time than other methods to obtain the results. Therefore, it is of interest to optimize LID to reduce the computational cost.

The LID method uses m quantiles to construct an approximation to the likelihood through linear interpolation. For large m, it would be useful to impose regularization to make inference more efficient. We may assume that $\beta(\tau)$ can be characterized by a few parameters, so we have a

Table 5. Out-of-bag quantile coverage

Methods	$\tau=0.1$	$\tau=0.25$	$\tau=0.5$	$\tau=0.75$	$\tau=0.9$
RQ	0.100	0.251	0.504	0.749	0.895
LID	0.093	0.249	0.506	0.748	0.909

low-dimensional parameter space no matter what m is, and the computation of LID would simplify. On the other hand, this approach involves additional assumption or approximation which would require additional work for its theoretical justification.

Appendix: Technical details

A.1. Find the bounds for the proposal distribution

This is for step 3 of the algorithm in Section 2.2. For each observation (y_i, x_i), $i = 1, 2, \ldots, n$, we can calculate a lower bound $l_{j,l,i}$ and an upper bound $u_{j,l,i}$. Then $l_{j,l} = \max_i(l_{j,l,i})$ is taken as the maximum of all these lower bounds and $u_{j,l} = \min_i(u_{j,l,i})$ is taken as the minimum of all these upper bounds. The formula to calculate $l_{j,l,i}$ and $u_{j,l,i}$ is given as follows.

If $1 < j < m$ and $x_{i,l} > 0$, where $x_{i,l}$ denotes the lth element of x_i, then

$$l_{j,l,i} = \frac{x_i^T \beta^{k-1}(\tau_{j-1}) - \sum_{t \neq l} x_{i,t} \beta_t^{k-1}(\tau_j)}{x_{i,l}} \quad \text{and}$$

$$u_{j,l,i} = \frac{x_i^T \beta^{k-1}(\tau_{j+1}) - \sum_{t \neq l} x_{i,t} \beta_t^{k-1}(\tau_j)}{x_{i,l}}.$$

If $1 < j < m$ and $x_{i,l} < 0$, then

$$l_{j,l,i} = \frac{x_i^T \beta^{k-1}(\tau_{j+1}) - \sum_{t \neq l} x_{i,t} \beta_t^{k-1}(\tau_j)}{x_{i,l}} \quad \text{and}$$

$$u_{j,l,i} = \frac{x_i^T \beta^{k-1}(\tau_{j-1}) - \sum_{t \neq l} x_{i,t} \beta_t^{k-1}(\tau_j)}{x_{i,l}}.$$

If $j = 1$ and $x_{i,l} > 0$, then

$$l_{j,l,i} = -\infty \quad \text{and} \quad u_{j,l,i} = \frac{x_i^T \beta^{k-1}(\tau_{j+1}) - \sum_{t \neq l} x_{i,t} \beta_t^{k-1}(\tau_j)}{x_{i,l}}.$$

If $j = 1$ and $x_{i,l} < 0$, then

$$l_{j,l,i} = \frac{x_i^T \beta^{k-1}(\tau_{j+1}) - \sum_{t \neq l} x_{i,t} \beta_t^{k-1}(\tau_j)}{x_{i,l}} \quad \text{and} \quad u_{j,l,i} = \infty.$$

If $j = m$ and $x_{i,l} > 0$, then

$$l_{j,l,i} = \frac{x_i^T \beta^{k-1}(\tau_{j-1}) - \sum_{t \neq l} x_{i,t} \beta_t^{k-1}(\tau_j)}{x_{i,l}} \quad \text{and} \quad u_{j,l,i} = \infty.$$

If $j = m$ and $x_{i,l} < 0$, then

$$l_{j,l,i} = -\infty \quad \text{and} \quad u_{j,l,i} = \frac{x_i^T \beta^{k-1}(\tau_{j-1}) - \sum_{t \neq l} x_{i,t} \beta_t^{k-1}(\tau_j)}{x_{i,l}}.$$

If $x_{i,l} = 0$, then

$$l_{j,l,i} = -\infty \quad \text{and} \quad u_{j,l,i} = \infty.$$

A.2. Proof of Theorem 3.1

We will verify the detailed balance condition to show that the stationary distribution is $\hat{p}_m(B_m|X, Y)$. Denote the probability of moving from B_m to B'_m by $K(B_m \to B'_m)$ and the proposal distribution by $q(B_m \to B'_m)$. We have

$$\begin{aligned}
&\hat{p}_m(B_m|X, Y) K\big(B_m \to B'_m\big) \\
&= \hat{p}_m(B_m|X, Y) q\big(B_m \to B'_m\big) \min\left(1, \frac{\pi_m(B'_m|X)\hat{p}_m(Y|X, B'_m) q(B'_m \to B_m)}{\pi_m(B_m|X)\hat{p}_m(Y|X, B_m) q(B_m \to B'_m)}\right) \\
&= \frac{\pi_m(B_m|X)\hat{p}_m(Y|X, B_m)}{\hat{p}_m(Y|X)} q\big(B_m \to B'_m\big) \min\left(1, \frac{\pi_m(B'_m|X)\hat{p}_m(Y|X, B'_m) q(B'_m \to B_m)}{\pi_m(B_m|X)\hat{p}_m(Y|X, B_m) q(B_m \to B'_m)}\right) \\
&= \frac{\pi_m(B'_m|X)\hat{p}_m(Y|X, B'_m)}{\hat{p}_m(Y|X)} q\big(B'_m \to B_m\big) \min\left(\frac{\pi_m(B_m|X)\hat{p}_m(Y|X, B_m) q(B_m \to B'_m)}{\pi_m(B'_m|X)\hat{p}_m(Y|X, B'_m) q(B'_m \to B_m)}, 1\right) \\
&= \hat{p}_m\big(B'_m|X, Y\big) K\big(B'_m \to B_m\big).
\end{aligned}$$

So the detailed balance condition is satisfied.

A.3. Proof of Theorem 3.2

To prove Theorem 3.2, we need three lemmas.

Lemma A.1. *Let* $\hat{p}_m(y_i|\theta_{m,i}) = \hat{f}_i(y_i|x_i, B_m)$ *given in* (2.3). *Assume* $\tau_{j+1} - \tau_j = \mathrm{O}(\frac{1}{m})$. *Then*

(a) $|\hat{p}_m(y_i|\theta_{m,i}) - p(y_i|\theta_{f_i})| = \mathrm{O}(\frac{1}{\sqrt{m}})$ *uniformly in the support of* y *as well as uniformly in* i.

(b) $|\hat{p}_m(Y|X, B_m) - p(Y|X, B_m)| = \mathrm{O}(\frac{1}{\sqrt{m}})$ *uniformly in the support of* Y.

Proof. (a) We will prove this proposition in two different cases.

Case 1: If y_i is between two quantiles we are using, in which case we can find two consecutive quantiles q_{i,τ_j} and $q_{i,\tau_{j+1}}$ such that $y_i \in [q_{i,\tau_j}, q_{i,\tau_{j+1}})$, where $1 \leq j \leq m-1$, then by the

mechanism of linear interpolation, we have the following equation

$$\begin{aligned}\hat{p}_m(y_i|\theta_{m,i}) &= \frac{\tau_{j+1}-\tau_j}{q_{i,\tau_{j+1}}-q_{i,\tau_j}}\\ &= \frac{\tau_{j+1}-\tau_j}{F_i^{-1}(\tau_{j+1})-F_i^{-1}(\tau_j)}\\ &= \frac{\tau_{j+1}-\tau_j}{(F_i^{-1})'(\tau^*)(\tau_{j+1}-\tau_j)}\\ &= \frac{\tau_{j+1}-\tau_j}{(1/f_i(y_i^*))(\tau_{j+1}-\tau_j)}\\ &= f_i\big(y_i^*\big),\end{aligned}$$

where $\tau^* \in [\tau_j, \tau_{j+1})$, $y_i^* \in [q_{i,\tau_j}, q_{i,\tau_{j+1}})$, F_i denotes the cdf of $y_i|\theta_f$, $F_i(y_i^*) = \tau^*$, and f_i denotes the pdf of $y_i|\theta_f$.

Now we want to show that

$$\big|f_i\big(y_i^*\big) - f_i(y_i)\big| \le \sup_{y\in[q_{i,\tau_j},q_{i,\tau_{j+1}})} f_i(y) - \inf_{y\in[q_{i,\tau_j},q_{i,\tau_{j+1}})} f_i(y) \le M_2\delta, \tag{A.1}$$

where $\delta = \sqrt{2(\tau_{j+1}-\tau_j)/M_2}$ and M_2 is given in Assumption 3.1. If $q_{i,\tau_{j+1}} - q_{i,\tau_j} \le \delta$, then $|f_i(y_i^*) - f_i(y_i)| = |f_i'(y^\dagger)(y_i^* - y_i)| \le M_2\delta$, where $y^\dagger \in [q_{i,\tau_j}, q_{i,\tau_{j+1}})$. Now let us consider the case that $q_{i,\tau_{j+1}} - q_{i,\tau_j} > \delta$. We will show that

$$\int_{q_{i,\tau_j}}^{q_{i,\tau_{j+1}}} f_i(y)\,\mathrm{d}y > \tau_{j+1} - \tau_j, \tag{A.2}$$

if

$$\sup_{y\in[q_{i,\tau_j},q_{i,\tau_{j+1}})} f_i(y) - \inf_{y\in[q_{i,\tau_j},q_{i,\tau_{j+1}})} f_i(y) > M_2\delta. \tag{A.3}$$

Letting $y_{\inf} = \arg\inf_{y\in[q_{i,\tau_j},q_{i,\tau_{j+1}})} f_i(y)$, $y_{\sup} = \arg\sup_{y\in[q_{i,\tau_j},q_{i,\tau_{j+1}})} f_i(y)$, without loss of generality, we can assume that $y_{\inf} < y_{\sup}$. It is obvious that $y_{\sup} - y_{\inf} > \delta$, because if $y_{\sup} - y_{\inf} \le \delta$, then

$$\begin{aligned}\sup_{y\in[q_{i,\tau_j},q_{i,\tau_{j+1}})} f_i(y) - \inf_{y\in[q_{i,\tau_j},q_{i,\tau_{j+1}})} f_i(y) &= f_i(y_{\sup}) - f_i(y_{\inf})\\ &= \big|f_i'\big(y^\dagger\big)\big|(y_{\sup} - y_{\inf}) \le M_2\delta.\end{aligned} \tag{A.4}$$

We can find a line with slope M_2 that goes through $(y_{\sup}, f_i(y_{\sup}))$. This line would be below the curve $f_i(y)$ in $[y_{\inf}, y_{\sup})$, since $f_i(y) - f_i(y_{\sup}) = f_i'(y^{\dagger\dagger})(y - y_{\sup}) \ge M_2(y - y_{\sup})$ for $y < y_{\sup}$, which leads to $f_i(y) \ge f_i(y_{\sup}) + M_2(y - y_{\sup})$.

Now we can check the area S formed by the line, $y = y_{\inf}$, $y = y_{\sup}$, and $f_i(y) = 0$. Figure 1 shows two possible cases. The shaded region is S.

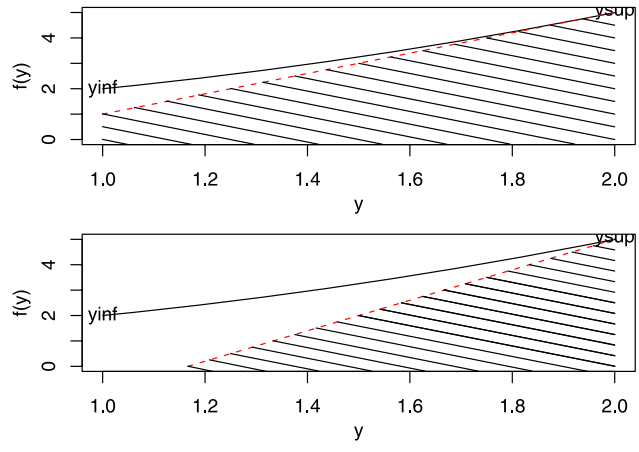

Figure 1. Illustration of the two possible cases of the area S: trapezoid and triangle. The solid curve stands for $f(y)$. The dotted line stands for the line with slope M_2. The shaded area is S.

If $f_i(y_{\sup}) - M_2(y_{\sup} - y_{\inf}) \geq 0$, the area is equal to

$$\begin{aligned} \frac{[2f_i(y_{\sup}) - M_2(y_{\sup} - y_{\inf})](y_{\sup} - y_{\inf})}{2} &\geq \frac{f_i(y_{\sup})(y_{\sup} - y_{\inf})}{2} \\ &> \frac{M_2\delta^2}{2} = \tau_{j+1} - \tau_j. \end{aligned} \tag{A.5}$$

If $f_i(y_{\sup}) - M_2(y_{\sup} - y_{\inf}) < 0$, the area is equal to

$$\frac{f_i(y_{\sup})^2}{2M_2} > \frac{(M_2\delta)^2}{2M_2} = \tau_{j+1} - \tau_j. \tag{A.6}$$

Therefore, in both cases, we have

$$\int_{q_{i,\tau_j}}^{q_{i,\tau_{j+1}}} f_i(y)\,\mathrm{d}y \geq \int_{y_{\inf}}^{y_{\sup}} f_i(y)\,\mathrm{d}y \geq S > \tau_{t+1} - \tau_j, \tag{A.7}$$

which contradicts with the fact that $\int_{q_{i,\tau_j}}^{q_{i,\tau_{j+1}}} f_i(y)\,\mathrm{d}y = \tau_{j+1} - \tau_j$. Hence

$$\begin{aligned} \left|f_i\left(y_i^*\right) - f_i(y_i)\right| &\leq \sup_{y\in[q_{i,\tau_j}, q_{i,\tau_{j+1}})} f_i(y) - \inf_{y\in[q_{i,\tau_j}, q_{i,\tau_{j+1}})} f_i(y) \\ &\leq M_2\delta = \sqrt{2M_2(\tau_{j+1} - \tau_j)} \\ &= \mathrm{O}\left(\frac{1}{\sqrt{m}}\right), \end{aligned}$$

given that $\tau_{j+1} - \tau_j = \mathrm{O}(\frac{1}{m})$.

Now let us consider the second case.

Case 2: If y_i is a point in the tail, which means $y_i \le q_{i,\tau_1}$ or $y_i > q_{i,\tau_m}$, then we have $p(y_i|\theta_{f_i}) = f_i(y_i) < \frac{c}{\sqrt{m}}$ from Assumption 3.1. For the tail part, we can use a truncated normal for the interpolation so that $\hat{p}_m(y_i|\theta_{m,i}) < \frac{c}{\sqrt{m}}$. Therefore, we have $|\hat{p}_m(y_i|\theta_{m,i}) - p(y_i|\theta_{f_i})| < \frac{2c}{\sqrt{m}} = \mathrm{O}(\frac{1}{\sqrt{m}})$.

Thus for both Cases 1 and 2, we showed $|\hat{p}_m(y_i|\theta_{m,i}) - p(y_i|\theta_{f_i})| = \mathrm{O}(\frac{1}{\sqrt{m}})$.

(b) Let us first show $|\hat{p}_m(y_i|\theta_{m,i}) - p(y_i|x_i, B_m)| = \mathrm{O}(\frac{1}{\sqrt{m}})$.

$$
\begin{aligned}
&\big|\hat{p}_m(y_i|\theta_{m,i}) - p(y_i|x_i, B_m)\big| \\
&= \left| \int_{f_i \in \mathscr{F}_{\theta_{m,i}}} \hat{p}_m(y_i|\theta_{m,i})\, \mathrm{d}\Pi_{\theta_{m,i}}(f_i) - \int_{f_i \in \mathscr{F}_{\theta_{m,i}}} p(y_i|\theta_{f_i})\, \mathrm{d}\Pi_{\theta_{m,i}}(f_i) \right| \\
&\le \int_{f_i \in \mathscr{F}_{\theta_{m,i}}} \big|\hat{p}_m(y_i|\theta_{m,i}) - p(y_i|\theta_{f_i})\big|\, \mathrm{d}\Pi_{\theta_{m,i}}(f_i) \\
&= \mathrm{O}\left(\frac{1}{\sqrt{m}}\right).
\end{aligned}
$$

Because $\hat{p}_m(Y|X, B_m) = \prod_{i=1}^n \hat{p}_m(y_i|x_i, B_m)$ and $p(Y|X, B_m) = \prod_{i=1}^n p(y_i|x_i, B_m)$, we can show $|\hat{p}_m(Y|X, B_m) - p(Y|X, B_m)| = \mathrm{O}(\frac{1}{\sqrt{m}})$ simply by induction. We will show the case with $n = 2$ here.

$$
\begin{aligned}
&\big|\hat{p}_m(Y|X, B_m) - p(Y|X, B_m)\big| \\
&\quad = \big|\hat{p}_m(y_1|X, B_m)\hat{p}_m(y_2|X, B_m) - p(y_1|X, B_m)p(y_2|X, B_m)\big| \\
&\quad = \big|\hat{p}_m(y_1|X, B_m)\hat{p}_m(y_2|X, B_m) - \hat{p}_m(y_1|X, B_m)p(y_2|X, B_m) \\
&\qquad + \hat{p}_m(y_1|X, B_m)p(y_2|X, B_m) - p(y_1|X, B_m)p(y_2|X, B_m)\big| \\
&\quad \le \big|\hat{p}_m(y_1|X, B_m)\big[\hat{p}_m(y_2|X, B_m) - p(y_2|X, B_m)\big]\big| \\
&\qquad + \big|\big[\hat{p}_m(y_1|X, B_m) - p(y_1|X, B_m)\big]p(y_2|X, B_m)\big| \\
&\quad = M_1\mathrm{O}\left(\frac{1}{\sqrt{m}}\right) + M_1\mathrm{O}\left(\frac{1}{\sqrt{m}}\right) \\
&\quad = \mathrm{O}\left(\frac{1}{\sqrt{m}}\right),
\end{aligned}
$$

where M_1 is given in Assumption 3.1. The proof can be easily generalized to the case with $n > 2$. □

Lemma A.2.

(a) $E_{\pi_m}(|\hat{p}_m(Y|X, B_m) - p(Y|X, B_m)|) = \mathrm{O}(\frac{1}{\sqrt{m}})$.

(b) $E_{\pi_m}(|\hat{p}_m(Y|X, B_m) - \hat{p}_{m-1}(Y|X, B_{m-1})|) = \mathrm{O}(\frac{1}{\sqrt{m}})$.

Proof. Part (a) of Lemma A.2 follows immediately from Lemma A.1(b). Part (b) of Lemma A.2 can be obtained by applying Lemma A.2(a) twice. □

Lemma A.3. $|\hat{p}_m(Y|X) - p(Y|X)| = \mathrm{O}(\frac{1}{\sqrt{m}})$.

Proof.

$$\begin{aligned}
&\big|\hat{p}_m(Y|X) - p(Y|X)\big| \\
&\quad = \left|\int \pi_m(B_m|X)\big[\hat{p}_m(Y|X, B_m) - p(Y|X, B_m)\big]\,\mathrm{d}B_m\right| \\
&\quad \le \int \pi_m(B_m|X)|\hat{p}_m(Y|X, B_m) - p(Y|X, B_m)|\,\mathrm{d}B_m \\
&\quad = E_{\pi_m}\big(\big|\hat{p}_m(Y|X, B_m) - p(Y|X, B_m)\big|\big) \\
&\quad = \mathrm{O}\left(\frac{1}{\sqrt{m}}\right).
\end{aligned}$$

□

Now we are ready to prove Theorem 3.2. We have

$$\begin{aligned}
&\big\|\hat{p}_m(B_m|X, Y) - p(B_m|X, Y)\big\|_{\mathrm{TV}} \\
&\quad = \frac{1}{2}\sup_{|h|\le 1}\left|\int h(B_m)\left(\frac{\pi_m(B_m|X)\hat{p}_m(Y|X, B_m)}{\hat{p}_m(Y|X)} - \frac{\pi_m(B_m|X)p(Y|X, B_m)}{p(Y|X)}\right)\mathrm{d}B_m\right| \\
&\quad \le \frac{1}{2}\int \pi_m(B_m|X)\left|\frac{\hat{p}_m(Y|X, B_m)}{\hat{p}_m(Y|X)} - \frac{p(Y|X, B_m)}{p(Y|X)}\right|\mathrm{d}B_m \\
&\quad = \frac{1}{2}\int \pi_m(B_m|X)\left|\frac{\hat{p}_m(Y|X, B_m)p(Y|X) - \hat{p}_m(Y|X)p(Y|X, B_m)}{\hat{p}_m(Y|X)p(Y|X)}\right|\mathrm{d}B_m \\
&\quad = \frac{1}{2}\int \pi_m(B_m|X) \\
&\qquad\quad \times \left|\frac{[\hat{p}_m(Y|X, B_m) - p(Y|X, B_m)]p(Y|X) + p(Y|X, B_m)[p(Y|X) - \hat{p}_m(Y|X)]}{\hat{p}_m(Y|X)p(Y|X)}\right|\mathrm{d}B_m \\
&\quad \le \frac{1}{2}\int \pi_m(B_m|X) \\
&\qquad\quad \times \frac{|\hat{p}_m(Y|X, B_m) - p(Y|X, B_m)|p(Y|X) + p(Y|X, B_m)|p(Y|X) - \hat{p}_m(Y|X)|}{\hat{p}_m(Y|X)p(Y|X)}\,\mathrm{d}B_m \\
&\quad = \frac{1}{2}\left[\frac{E_{\pi_m}(|\hat{p}_m(Y|X, B_m) - p(Y|X, B_m)|)}{\hat{p}_m(Y|X)} + \frac{|\hat{p}_m(Y|X) - p(Y|X)|}{\hat{p}_m(Y|X)}\right].
\end{aligned}$$

We already know from Lemma A.3 that $\hat{p}_m(Y|X) \to p(Y|X)$ as $m \to \infty$, so for any $e^* \in (0, p(Y|X))$, there exists an m^* such that $|\hat{p}_m(Y|X) - p(Y|X)| < e^*$ for $m \ge m^*$. We can see

that

$$LB = \min\big(\hat{p}_{m_0}(Y|X), \hat{p}_{m_0+1}(Y|X), \ldots, \hat{p}_{m^*-1}(Y|X), p(Y|X) - e^*\big)$$

is a lower bound for $\hat{p}_m(Y|X)$, where m_0 is the minimum number of quantiles we use. Therefore, $\|\hat{p}_m(B_m|X,Y) - p(B_m|X,Y)\|_{\mathrm{TV}} \le \frac{1}{2LB}[E_{\pi_m}(|\hat{p}_m(Y|X,B_m) - p(Y|X,B_m)|) + |\hat{p}_m(Y|X) - p(Y|X)|] = \mathrm{O}(\frac{1}{\sqrt{m}}) \to 0$ as $m \to \infty$ (because of Lemmas A.2 and A.3).

Acknowledgements

The research of Yuguo Chen was supported in part by NSF Grant DMS-1106796. The research of Xuming He was supported in part by NSF Grants DMS-1237234 and DMS-1307566, and National Natural Science Foundation of China Grant 11129101.

References

[1] Berger, J.O., Liseo, B. and Wolpert, R.L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14** 1–28. MR1702200

[2] Bondell, H.D., Reich, B.J. and Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika* **97** 825–838. MR2746154

[3] Doss, H. (1985). Bayesian nonparametric estimation of the median. I. Computation of the estimates. *Ann. Statist.* **13** 1432–1444. MR0811501

[4] Dunson, D.B. and Taylor, J.A. (2005). Approximate Bayesian inference for quantiles. *J. Nonparametr. Stat.* **17** 385–400. MR2129840

[5] He, X. and Hu, F. (2002). Markov chain marginal bootstrap. *J. Amer. Statist. Assoc.* **97** 783–795. MR1941409

[6] Koenker, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge: Cambridge Univ. Press. MR2268657

[7] Kottas, A. and Gelfand, A.E. (2001). Bayesian semiparametric mdeian regression modeling. *J. Amer. Statist. Assoc.* **96** 1458–1468. MR1946590

[8] Kottas, A. and Krnjajić, M. (2009). Bayesian semiparametric modelling in quantile regression. *Scand. J. Stat.* **36** 297–319. MR2528986

[9] Lavine, M. (1995). On an approximate likelihood for quantiles. *Biometrika* **82** 220–222. MR1332852

[10] Parzen, M.I., Wei, L.J. and Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81** 341–350. MR1294895

[11] Reich, B.J., Bondell, H.D. and Wang, H.J. (2010). Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics* **11** 337–352.

[12] Reich, B.J., Fuentes, M. and Dunson, D.B. (2011). Bayesian spatial quantile regression. *J. Amer. Statist. Assoc.* **106** 6–20. MR2816698

[13] Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. *Springer Texts in Statistics*. New York: Springer. MR2080278

[14] Taddy, M.A. and Kottas, A. (2010). A Bayesian nonparametric approach to inference for quantile regression. *J. Bus. Econom. Statist.* **28** 357–369. MR2723605

[15] Tokdar, S.T. and Kadane, J.B. (2012). Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Anal.* **7** 51–72. MR2896712

[16] Yang, Y. and He, X. (2012). Bayesian empirical likelihood for quantile regression. *Ann. Statist.* **40** 1102–1131. MR2985945

[17] Yu, K. and Moyeed, R.A. (2001). Bayesian quantile regression. *Statist. Probab. Lett.* **54** 437–447. MR1861390

Received July 2012 and revised November 2012