BAYESIAN QUANTILE LINEAR REGRESSION

BY

YANG FENG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Yuguo Chen, Chair
Professor Xuming He
Professor Feng Liang
Professor Stephen L. Portnoy

Abstract

Quantile regression, as a supplement to the mean regression, is often used when a comprehensive relationship between the response variable and the explanatory variables is desired. The traditional frequentists' approach to quantile regression was well developed with asymptotic theories and efficient algorithms. However 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 a certain distribution for the error is assumed. In this dissertation, we propose two Bayesian quantile regression methods: the data generating process based method (DG) and the linearly interpolated density based method (LID). Markov chain Monte Carlo algorithms are developed to implement the proposed methods. We provide the convergence property of the algorithms and numerically verify the theoretical results. We compare the proposed methods with some existing methods through simulation studies, and apply our method to the birth weight data.

Unlike most of the existing methods which aim at tackling one quantile at a time, our proposed methods aim at estimating the joint posterior distribution of multiple quantiles and achieving global efficiency for all quantiles of interest and functions of those quantiles. From the simulation results, we found that LID could produce more efficient estimates than some existing methods. In particular, for estimating the difference of quantiles, LID has a big advantage over other existing methods.

Keywords: Bayesian inference; Markov chain Monte Carlo (MCMC); Quantile regression; Linearly interpolated density (LID);

To My family

Acknowledgments

I am heartily thankful to my supervisor, Prof. Yuguo Chen, whose encouragement, supervision and support during my years of study greatly helped me to develop an understanding of the subject. I also would like to thank Prof. Xuming He, who made lots of very insightful advice for my work. Besides, I would like to thank Prof. Feng Liang and Prof. Stephen Portnoy for their suggestions during my preliminary exam and defense, Prof. Ying Wei for the birth weight data, and Yunwen Yang for the temperature data.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the project.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Regression, a term defined by Galton (1885) [5] to describe a biological phenomenon, is one of the most widely used statistical tools. Most of the time it refers to the linear mean regression, which takes the form of

$$y_i = x_i\beta + \epsilon_i, \quad i = 1, 2, ..., n,$$

where y_i is the response variable, x_i is a $1 \times p$ 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 which is usually assumed to have mean zero.

The earliest regression, which used the method of least squares proposed by Legendre and Gauss, however, was used for determining the orbits of the bodies around the sun. After about a century, Yule and Pearson studied the theoretical property of the regression by assuming that the joint distribution of y_i and x_i is normal. Later, Fisher found that only the conditional distribution of $y_i|x_i$ needs to be normal. This is one of the most commonly used assumptions in regression analysis. It turns out that the mean regression could solve many problems under such simple assumptions.

However, for some data, the assumptions of the linear mean regression do not hold or the objective of interests is no longer the mean. For example, people would like to study why some infants are born with relatively low birth weights, that is, the lower quantiles of infant birth weights are of main interests. Moreover, sometimes the error term does not even come from a distribution with finite mean, e.g., the Cauchy distribution. In such cases, modeling other quantities, such as quantiles, might be more appealing. For quantile regression, there are no specific assumptions about the error term. In Sections 1.1, 1.2.1, and 1.2.2, we follow Chapters 1 and 3 of Koenker (2005) [10] to introduce quntile regression.

1.1 Introduction of quantile regression

As early as 1755, Boscovich published his work about calculating the ellipticity of the earth. His model is

$$y_i = a + bsin^2\lambda_i, \quad i = 1, 2, 3, 4, 5,$$

where y is the arc-length of 1° of latitude and λ is the latitude, and the ellipticity is computed as $3a/b$. In his work, he estimated a and b by minimizing the sum of absolute residuals under the constraint that the sum of residuals equals to 0, that is,

$$\begin{aligned} &\min_{a,b} \sum_{i=1}^{5} |y_i - a - bsin^2\lambda_i|, \\ \text{subject to:} \quad &\sum_{i=1}^{5} y_i - a - bsin^2\lambda_i = 0. \end{aligned}$$

Edgeworth (1888) [3] revised Boscovich's idea by throwing away the constraint on the sum of residuals, and thereby developed the median regression.

In order to model quantiles other than the median, an asymmetric version of absolute errors is used as the objective function. For the τth quantile, the asymmetric function is defined as:

$$\rho_\tau(u) = u(\tau - I_{\{u<0\}}), \tag{1.1}$$

where $I_{\{u<0\}}$ is an indicator function taking value 1 if $u < 0$, and 0 otherwise. An example of the ρ function with $\tau = 0.25$ is given in Figure 1.1. One motivation to use this function as the loss function is that for a random variable $X \sim F$,

$$E\rho_\tau(X - \hat{x}) = (\tau - 1)\int_{-\infty}^{\hat{x}} (x - \hat{x})dF(x) + \tau\int_{\hat{x}}^{\infty} (x - \hat{x})dF(x).$$

In order to minimize this loss function, we can take the derivative with respect to $\hat{x}$ and set it to 0:

$$\begin{aligned} (\tau - 1)\int_{-\infty}^{\hat{x}} dF(x) + \tau\int_{\hat{x}}^{\infty} dF(x) &= 0 \\ \tau - F(\hat{x}) &= 0. \end{aligned}$$

One solution is the τ-th quantile, which is defined as $\hat{x} = F^{-1}(\tau)$, where

$$F^{-1}(\tau) = inf\{x : F(x) \geq \tau\}.$$

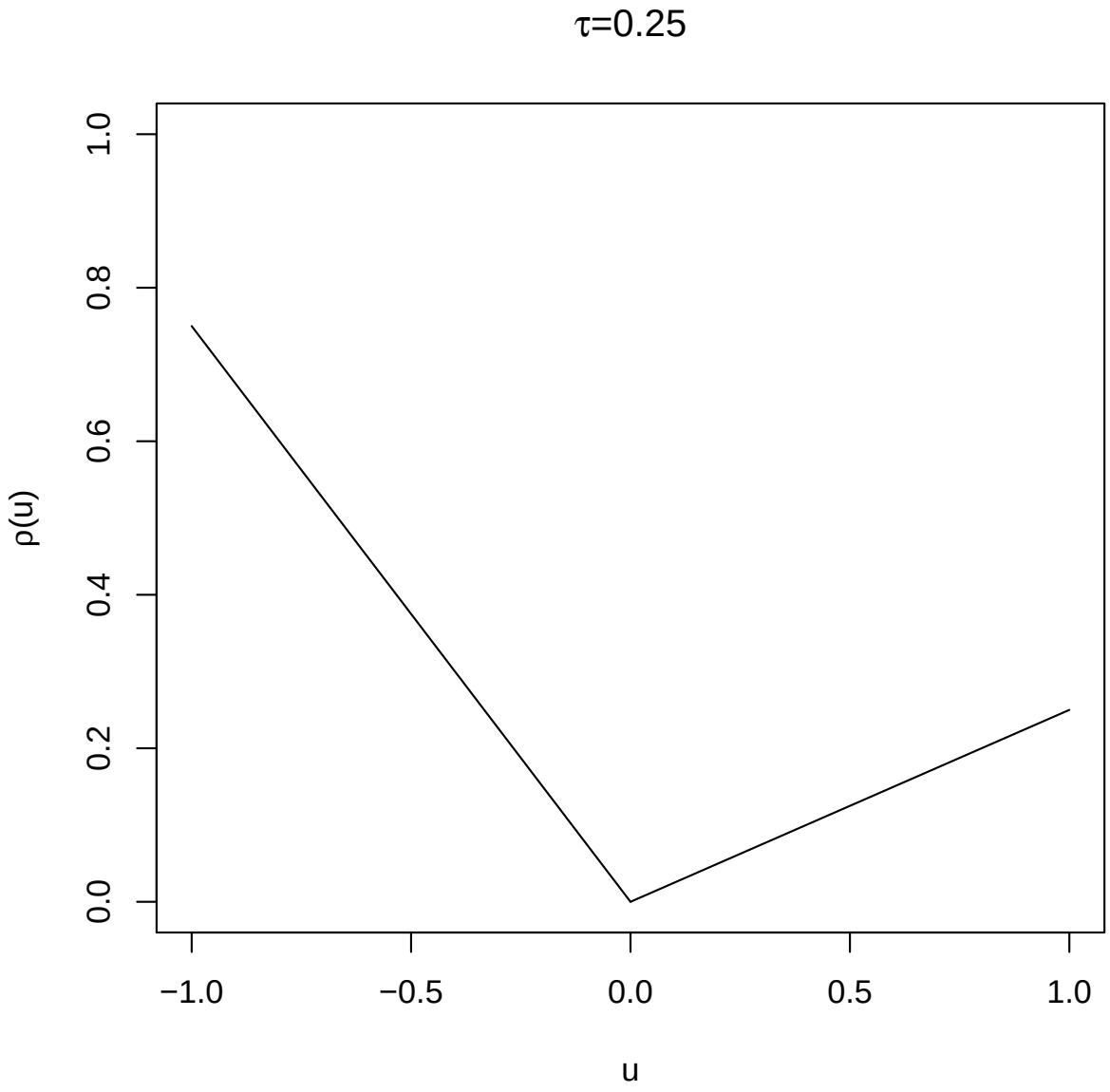

Figure 1.1: Example of the ρ function when $\tau = 0.25$.

For the linear regression case, we can specify the τ-th conditional quantile function as

$$Q_\tau(y_i|x_i) = x_i\beta(\tau), \quad i = 1, 2, ..., n.$$

The parameter $\beta(\tau)$ can be estimated by $\hat{\beta}(\tau)$ which solves the objective function

$$\min_{\beta(\tau)} \sum_{i=1}^{n} \rho_\tau(y_i - x_i\beta(\tau)).$$

Let $Y = (y_1, y_2, ..., y_n)'$, $X = (x_1', x_2', ..., x_n')'$, $1_n = (\underbrace{1, 1, ..., 1}_{n})'$. The above quantile regression problem is equivalent to a linear programming problem,

$$\min_{(\beta(\tau),u,v)} \{\tau 1_n' u + (1-\tau)1_n' v | X\beta(\tau) + u - v = Y\}. \tag{1.2}$$

where u and v correspond to the positive and negative parts of the residual vector $Y - X\beta$. One can obtain $\hat{\beta}(\tau)$ through solving the linear programming problem.

1.2 Inference for quantile regression

1.2.1 Inference based on asymptotic distributions

Consider data from the linear model

$$y_i = x_i\beta + \epsilon_i, \quad i = 1, 2, ..., n, \tag{1.3}$$

where ϵ_i's are independent and identically distributed (iid) from a distribution F with density f. Koenker and Bassett (1978) [11] showed that the joint asymptotic distribution of the m quantile regression estimators $\hat{\zeta}_n = (\hat{\beta}_n(\tau_1)', ..., \hat{\beta}_n(\tau_m)')'$ is

$$\sqrt{n}(\hat{\zeta}_n - \zeta) \xrightarrow{D} N(0, \Omega \otimes Q_0^{-1}), \quad \text{as} \quad n \to \infty, \tag{1.4}$$

where $\zeta = (\beta(\tau_1)', ..., \beta(\tau_m)')'$, the positive definite matrix $Q_0 = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} x_i' x_i$, the symbol $\otimes$ denotes Kronecker product and Ω is an $m \times m$ matrix with elements

$$\omega_{ij} = \frac{max(\tau_i, \tau_j) - \tau_i \tau_j}{f(F^{-1}(\tau_i)) f(F^{-1}(\tau_j))}. \quad i = 1, 2, ..., m, j = 1, 2, ..., m.$$

If ϵ_i's are not iid, then the asymptotic distribution of $\sqrt{n}(\hat{\beta}(\tau_j) - \beta(\tau_j))$ takes the following form.

$$\sqrt{n}(\hat{\beta}(\tau_j) - \beta(\tau_j)) \xrightarrow{D} N(0, \tau_j(1-\tau_j) H_n^{-1}(\tau_j) J_n(\tau_j) H_n^{-1}(\tau_j)), \quad j = 1, 2..., m, \quad \text{as} \quad n \to \infty, \tag{1.5}$$

where

$$J_n = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} x_i' x_i,$$

and

$$H_n(\tau_j) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} x_i' x_i f_i(\xi_i(\tau_j)).$$

Here $\xi_i(\tau) = Q_\tau(y_i|x_i)$ and f_i denotes the conditional density of $y_i|x_i$. The asymptotic covariance matrix for $\hat{\beta}(\tau_i)$ and $\hat{\beta}(\tau_j)$ is as follows.

$$Acov(\sqrt{n}(\hat{\beta}(\tau_i) - \beta(\tau_i), \sqrt{n}(\hat{\beta}(\tau_j) - \beta(\tau_j)) = [max(\tau_i, \tau_j) - \tau_i \tau_j] H_n(\tau_i)^{-1} J_n H_n(\tau_j)^{-1}, \tag{1.6}$$

where $i = 1, 2, ..., m$ and $j = 1, 2, ..., m$. Therefore, to test hypotheses or construct confidence intervals, one has to estimate $s(\tau_j) = [f(F^{-1}(\tau_j))]^{-1}$ for iid errors and $f_i(\xi_i(\tau))$ for non-iid errors. Because $F(F^{-1}(\tau_j)) = \tau_j$, if we take the derivative with respect to τ on both sides, then

$$f(F^{-1}(\tau_j)) \frac{d}{d\tau_j} F^{-1}(\tau_j) = 1, \quad j = 1, 2, ...,$$

which leads to $s(\tau_j) = \frac{d}{d\tau_j} F^{-1}(\tau_j)$. Siddiqui (1960) [17] suggests that one could approximate $s(\tau_j)$ by

$$\hat{s}_n(\tau_j) = \frac{\hat{F}_n^{-1}(\tau_j + h_n) - \hat{F}_n^{-1}(\tau_j - h_n)}{2h_n}, \quad j = 1, 2, ...,$$

where $\hat{F}_n^{-1}$ is an estimate of F^{-1} and h_n is the bandwidth which converges to 0 as $n \to \infty$. Similarly, for non-iid errors, Hendreicks and Koenker (1991) [7] suggests that one could estimate $f_i(\xi_i(\tau))$ by

$$\hat{f}_i(\xi_i(\tau)) = \frac{2h_n}{x_i(\hat{\beta}(\tau + h_n) - \hat{\beta}(\tau - h_n))}.$$

However, there is no guarantee that

$$d_i = x_i(\hat{\beta}(\tau + h_n) - \hat{\beta}(\tau - h_n)) > 0.$$

One could replace $\hat{f}_i$ by $max\{0, \frac{2h_n}{d_i - \delta}\}$, where $\delta > 0$ is a small quantity to avoid the denominator being 0.

1.2.2 Inference based on bootstrap

When the variance of the estimate is difficult to calculate, the bootstrap method is usually one way to circumvent the difficulty. For example, if the data come from Model (1.3), one can implement the residual bootstrap to calculate confidence intervals for $\beta(\tau)$. Let

$$\hat{\epsilon}_i = y_i - x_i\hat{\beta}(\tau), \quad i = 1, 2, ..., n. \tag{1.7}$$

The bootstrap samples $\epsilon_1^*, \epsilon_2^*, ..., \epsilon_n^*$ are drawn from $\hat{\epsilon}_1, \hat{\epsilon}_2, ..., \hat{\epsilon}_n$ with replacement. Letting

$$y_i^* = x_i\hat{\beta}(\tau) + \epsilon_i^*, \quad i = 1, 2, ..., n, \tag{1.8}$$

we can obtain the bootstrap estimate of $\beta(\tau)$, denoted as $\beta^*(\tau)$, by

$$\beta^*(\tau) = \underset{\beta}{argmin} \sum_{i=1}^{n} \rho_\tau(y_i^* - x_i\beta).$$

DeAngelis et al. (1993) [1] showed that the distribution of $\sqrt{n}(\beta^*(\tau) - \hat{\beta}(\tau))$ conditional on D, where $D = \{(x_i, y_i), i = 1, 2, ..., n\}$, converges to the limiting distribution of $\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau))$. There are two ways to calculate confidence intervals based on B bootstrap samples $\beta_1^*(\tau), \beta_2^*(\tau), ..., \beta_B^*(\tau)$. The first way is to estimate the covariance matrix of $\hat{\beta}(\tau)$ by

$$\frac{1}{B}\sum_{i=1}^{B}(\beta_i^*(\tau) - \hat{\beta}(\tau))(\beta_i^*(\tau) - \hat{\beta}(\tau))',$$

and then calculate confidence intervals based on the asymptotical normal distribution of $\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau))$. The other method, discussed by Efron and Tibshirani (1993) [4], is based on quantiles of

$$\beta^*_{1,j}(\tau), \beta^*_{2,j}(\tau), ..., \beta^*_{B,j}(\tau), \quad j = 1, 2, ..., p,$$

where $\beta^*_{i,j}(\tau)$ denotes the j-th component of the i-th bootstrap estimate $\beta^*_i(\tau)$. A 95% confidence interval for $\beta_j(\tau)$, where $\beta_j(\tau)$ is the j-th component of $\beta(\tau)$, could be estimated by $(\beta^*_{0.025,j}(\tau), \beta^*_{0.975,j}(\tau))$, where $\beta^*_{0.025,j}(\tau)$ and $\beta^*_{0.975,j}(\tau)$ denote the 2.5% and 97.5% quantiles of $\beta^*_{1,j}(\tau), \beta^*_{2,j}(\tau), ..., \beta^*_{B,j}(\tau)$ correspondingly.

When the errors are not iid, we have to switch to the (x, y)-paired bootstrap (Efron, 1982). In the (x, y)-paired bootstrap, we will draw samples (x^*_i, y^*_i), $i = 1, 2, ..., n$, from $\{(x_i, y_i), i = 1, 2, ..., n\}$ with replacement and equal weights. The bootstrap estimate $\beta^*(\tau)$ is computed by

$$\beta^*(\tau) = \underset{\beta}{argmin} \sum_{i=1}^{n} \rho_\tau(y^*_i - x^*_i\beta).$$

One could use the same methods as that for the residual bootstrap to calculate confidence intervals after one has B bootstrap estimates.

1.2.3 Inference based on MCMB

He and Hu (2002) [6] proposed a Markov chain marginal bootstrap (MCMB) method based on bootstrapping estimating equations.

Let $Y_1, Y_2, ..., Y_n$ be n independent random variables and let $\theta = (\theta_1, \theta_2, ..., \theta_p)$ be the p-dimensional parameter that relates to the distribution of Y_i. Suppose that $g_i(Y_i, \theta)$ is a p-dimensional function with $E_\theta(g_i(Y_i, \theta)) = 0$, $i = 1, 2, ..., n$. Then

$$S(Y, \theta) = \sum_{i=1}^{n} g_i(Y_i, \theta) = 0$$

is called an *unbiased estimating equation*. Assume that $g_i(Y_i, \theta) = a_i z_i$, $i = 1, 2, ..., n$, where a_i's are constant and z_i's are random. The MCMB algorithm, quoted from He and Hu (2002) [6], is as follows.

1. Initialize $\hat{\theta}^{(0)} = \hat{\theta}$ and $k = 1$.
2. Draw a bootstrap sample $\{z^{*(k)}_{1j}, ..., z^{*(k)}_{nj}\}$ from $\{z_1, ..., z_n\}$ for each $j = 1, ..., p$.

3. In the sequence of $j = 1, 2, ..., p$, solve for $\hat{\theta}_j^{(k)}$ from

$$S_j(Y, \hat{\theta}_1^{(k)}, ..., \hat{\theta}_{j-1}^{(k)}, \hat{\theta}_j^{(k)}, \hat{\theta}_{j+1}^{(k-1)}, ..., \hat{\theta}_p^{(k-1)}) = S_j^{*(k)},$$

where S_j is the j-th component of $S(Y, \theta)$, and $S_j^{*(k)}$ is the j-th component of $\sum_{i=1}^n a_i z_{ij}^{*(k)}$.

4. Increase k by 1 and go to step 2, or stop if k has reached a prespecified level.

MCMB is very computationally efficient since it only solves a one-dimensional equation every time. The computational cost of MCMB is in the order of $O(np)$ instead of $O(np^{5/2})$ for traditional bootstrap methods. Kocherginsky and He (2007) [8] modified the MCMB method by applying two transformations in order to decrease the potential high autocorrelation of the MCMB sequence and broaden the applicability of MCMB.

For the quantile regression case, the corresponding estimating equation for the τ-th quantile is

$$S(D, \beta(\tau)) = \sum_{i=1}^{n} x_i \psi_\tau(y_i - x_i\beta(\tau)),$$

where $\psi_\tau(y_i - x_i\beta(\tau)) = \tau - I_{\{y_i \le x_i\beta(\tau)\}}$. At the resampling step in the MCMB algorithm, one can set $a_i = 1$ and $z_i = x_i\psi_\tau(y_i - x_i\beta(\tau)) - \bar{z}$, where $\bar{z} = \frac{1}{n}\sum_{i=1}^n x_i\psi_\tau(y_i - x_i\beta(\tau))$, as proposed by Kocherginsky, He and Mu (2005) [9]. Alternatively, one can make use of the pivotal property of $S(D, \beta(\tau))$, which is observed by Parzen, Wei and Ying (1994) [15]. They suggest that $I_{\{y_i \le x_i\beta(\tau)\}}$ can be generated from a Bernoulli distribution with success probability τ, so the distribution of $S(D, \beta(\tau))$ is independent of $\beta(\tau)$, which makes $S(D, \beta(\tau))$ pivotal. Therefore, Step 2 in the algorithm could be modified as follows.

2.a) Sample $I_{\{y_i \le x_i\beta(\tau)\}}^{*ik}$ from a Bernoulli distribution with success probability equal to τ.

2.b) set $z_{ij}^{*k} = x_{ij}\psi_\tau^{*ik}(y_i - x_i\beta(\tau)) = x_{ij}(\tau - I_{\{y_i \le x_i\beta(\tau)\}}^{*ik})$.

1.3 Bayesian regression of quantiles

Similar to the linear mean regression, it is of interest to study regression of quantiles under the Bayesian framework. A good property of the Bayesian method is that once we have the posterior distribution or samples from the posterior distribution, it is relative easy to construct credible intervals for the parameters. Generally, one can use the posterior quantiles or posterior sample

quantiles to construct credible intervals for the parameters. However, the most challenging problem for Bayesian quantile regression is that the likelihood is usually not available unless a certain distribution for the error is assumed. Due to this difficulty, not much work has been done under the Bayesian framework.

The following are some work in the literature that I am aware of. Yu and Moyeed (2001) [19] 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

$$f_\tau(u) = \tau(1-\tau)e^{-\rho_\tau(u)}, \quad u \in \mathbb{R},$$

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 the quantile regression objective function. Kottas and Gelfand (2001) [12] implemented a Bayesian median regression. They introduced two distribution families with median zero, and they also employed the Dirichlet process prior. Dunson and Taylor (2005) [2] tried to use a substitution likelihood proposed by Lavine (1995) [13], to make inferences based on the posterior distribution. Here is the description of the substitution likelihood. If $y_1, y_2, ..., y_n$ are iid from a distribution F, then for m quantiles $\theta = (\theta_{\tau_1}, \theta_{\tau_2}, ..., \theta_{\tau_m})$ of F with $\tau_1 < \tau_2 < ... < \tau_m$, the substitution likelihood is

$$s(\theta) = \begin{pmatrix} n \\ u_1(\theta) \cdots u_{m+1}(\theta) \end{pmatrix} \prod_{i=1}^{m+1} \Delta\tau_i^{u_i(\theta)},$$

where $u_i(\theta) = \sum_{i=1}^{n} I_{y_i \in (\theta_{\tau_{i-1}}, \theta_{\tau_i}]}$ and $(\Delta\tau_1, \Delta\tau_2, ..., \Delta\tau_{m+1}) = (\tau_1, \tau_2 - \tau_1, ..., 1 - \tau_m)$. One property of Dunson and Taylor's method is that it allows regression on multiple quantiles simultaneously. Taddy and Kottas (2009) [18] developed a fully nonparametric model-based quantile regression based on Dirichlet process mixing.

1.4 Our contribution

In this dissertation, we introduce a Bayesian method based on linearly interpolated density (LID) or a data-generating process (DG), which avoids calculating densities directly. The proposed methods can estimate multiple quantiles simultaneously. In particular, we found that LID has a big advantage for estimating the difference of quantiles.

The rest of the dissertation is organized as follow. Chapter 2 describes the algorithms of the two methods. Chapter 3 shows theoretical results of the proposed methods. Chapter 4 gives numerical results based on simulated and real data. Chapter 5 gives more numerical studies. Chapter 6 gives the conclusion and discusses possible directions for future work.

Chapter 2

Algorithms of the Two Proposed Methods

Suppose that we have a linear model

$$y_i = x_i\beta + \epsilon_i, \quad i = 1, 2, ..., n, \tag{2.1}$$

where x_i is a $1 \times p$ 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 corresponding quantile model for the τ_j-th quantile is

$$Q_{\tau_j}(y_i|x_i) = x_i\beta(\tau_j), \quad i = 1, 2, ..., n, \quad j = 1, 2, ..., m. \tag{2.2}$$

Because our methods are under the Bayesian framework, we need to put priors on the parameters and make posterior analysis. Let $B_m = (\beta(\tau_1), \beta(\tau_2), ..., \beta(\tau_m))$ and denote $\pi(B_m|X)$ as the prior for $B_m|X$. We are interested in the posterior distribution of $B_m|X, Y$, where $X = (x_1', ..., x_n')'$ and $Y = (y_1, ..., y_n)$.

As introduced in the first chapter, it is not an easy task to find the posterior distribution since the likelihood is not available unless the error distribution is specified. Here we will introduce two ways to deal with this problem. The first method approximates the density based on linear interpolation. The second method employs a data-generating process to avoid the likelihood.

2.1 Interpretations of $B_m|X, Y$

Before we introduce the algorithms, it is important to know how to interpret $B_m|X, Y$. Let us consider an ideal case first, where B_∞ is infinite dimensional and covers all the quantiles. In this case, $Y|X, B_\infty$ is equivalent to $Y|F_1, F_2, ..., F_n$, where F_i is the cumulative distribution function (cdf) of $y_i|x_i$. Assuming that all these conditional distributions have probability density functions (pdf), we can calculate the likelihood function $L(Y|X, B_\infty)$ through $\prod_{i=1}^n f_i(y_i)$, where f_i is the pdf of $y_i|x_i$. Denoting $\pi(B_\infty)$ as the prior for B_∞, we can define the posterior distribution of $B_\infty|X, Y$

as

$$p(B_\infty|Y,X) = \frac{\pi(B_\infty)L(Y|X,B_\infty)}{p(Y|X)} \propto \pi(B_\infty)L(Y|X,B_\infty),$$

where $p(Y|X)$ denotes the marginal density of $Y|X$.

Similarly, we can define the posterior distribution of $B_m|X,Y$ as

$$p(B_m|X,Y) \propto \pi(B_m)L(Y|X,B_m).$$

Now the interpretation of $L(Y|X,B_m)$ is not straightforward. Our definition of $L(Y|X,B_m)$ is that $L(Y|X,B_m) = \prod_{i=1}^n \bar{f}_i(y_i)$, where $\bar{f}_i(y_i)$ is defined as an average over all the possible conditional distributions of $y_i|x_i$ with the same m quantiles $(Q_{\tau_1}(y_i|x_i), Q_{\tau_2}(y_i|x_i), ..., Q_{\tau_m}(y_i|x_i)) = x_iB_m$. Denote $\pi(f_i|x_iB_m)$ as the prior on all the possible conditional distributions of $y_i|x_i$ with the same m quantiles $(Q_{\tau_1}(y_i|x_i), Q_{\tau_2}(y_i|x_i), ..., Q_{\tau_m}(y_i|x_i))$, then

$$\bar{f}_i(y_i) = \int f_i(y_i|x_i)\pi(f_i|x_iB_m)df_i.$$

We will revisit this concept in Chapter 3, where we will discuss more about the priors and the interpretation.

2.2 The linearly interpolated density method

We illustrate the basic of the method through the following example. Suppose $Z \sim F(z)$, where $F(z)$ is the cdf for Z. Let $f(z)$ be the pdf of Z and z be an observed sample. Let $\tau_z = F(z)$ and τ_1, τ_2 be two constants such that $0 \leq \tau_1 < \tau_z < \tau_2 \leq 1$. Then $F^{-1}(\tau_1) < z < F^{-1}(\tau_2)$ if we assume $f(z)$ is continuous and non-zero. We can approximate $f(z)$ by

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

since

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

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

$$\tau^* = F(F^{-1}(\tau^*)) \Rightarrow 1 = f(F^{-1}(\tau^*))\frac{d}{d\tau}F^{-1}(\tau^*) = f(z^*)\frac{d}{d\tau}F^{-1}(\tau^*) \Rightarrow \frac{d}{d\tau}F^{-1}(\tau^*) = \frac{1}{f(z^*)}.$$

Moreover, if we shrink (τ_1, τ_2) towards τ_z, then $(F^{-1}(\tau_1), F^{-1}(\tau_2))$ will shrink towards z, given $f(z)$ is continuous and nonzero. Because $z^* \in (F^{-1}(\tau_1), F^{-1}(\tau_2))$, we have

$$\lim_{\tau_1 \to \tau_z, \tau_2 \to \tau_z} f(z^*) = f(z).$$

We will discuss more about the convergence property of the linear interpolated density in Chapter 3.

Algorithm of the linearly interpolated density method

We want to run an MCMC algorithm on B_m and obtain samples from the posterior distribution. Here we introduce the algorithm step by step.

1. Pick m quantiles, say, the τ_1-th, the τ_2-th,..., and the τ_m-th quantiles, which should include the quantiles of interest. One possible choice is to make them equally spaced, that is, $\tau_i = \frac{i}{m}$.

2. Put a prior $\pi(B_m)$ on B_m. One possible prior is a truncated normal $N(\mu, \Sigma)$ satisfying

$$x_i\beta(\tau_1) < x_i\beta(\tau_2) < ... < x_i\beta(\tau_m), \quad i = 1, 2, ..., n. \tag{2.3}$$

3. Choose an initial value B_m^0 for B_m. A good choice is the quantile regression estimate, which could be calculated by "*quantreg*" package in R. Since quantile regression estimates does not guarantee that

$$x_i\hat{\beta}^{rq}(\tau_1) < x_i\hat{\beta}^{rq}(\tau_2) < ... < x_i\hat{\beta}^{rq}(\tau_m), \quad i = 1, 2, ..., n,$$

we need to make some adjustments to the quantile regression estimates if necessary. If all the covariates are positive, one possible modification is to use the order statistic of

$$(\hat{\beta}_k^{rq}(\tau_1), \hat{\beta}_k^{rq}(\tau_2), ..., \hat{\beta}_k^{rq}(\tau_m)), \quad k = 1, 2, ..., p,$$

denoted as

$$(\hat{\beta}_{k,(1)}^{rq}(\tau_1), \hat{\beta}_{k,(2)}^{rq}(\tau_2), ..., \hat{\beta}_{k,(m)}^{rq}(\tau_m)), \quad k = 1, 2, ..., p,$$

where $\hat{\beta}_k^{rq}(\tau_j)$ denotes the k-th element of $\hat{\beta}_{rq}(\tau_j)$. Therefore, the k-th row of B_m^0 is

$$(\hat{\beta}_{k,(1)}^{rq}(\tau_1), \hat{\beta}_{k,(2)}^{rq}(\tau_2), ..., \hat{\beta}_{k,(m)}^{rq}(\tau_m)), \quad k = 1, 2, ..., p.$$

If some covariates are not positive, we can shift them to a positive region.

4. Approximate the densities. With the initial values of the parameters, we can calculate the linear interpolated densities by

$$\begin{aligned}\hat{f}_i^0(y_i|x_i) &= [\sum_{j=1}^{m-1} I_{\{y_i\in(x_i\beta^0(\tau_j),x_i\beta^0(\tau_{j+1}))\}}\frac{\tau_{j+1}-\tau_j}{x_i\beta^0(\tau_{j+1})-x_i\beta^0(\tau_j)}] + I_{\{y_i\in(-\infty,x_i\beta^0(\tau_1))\}}\tau_1 f_1(y_i)\\ &\quad + I_{\{y_i\in(x_i\beta^0(\tau_m),\infty)\}}(1-\tau_m)f_2(y_i), \quad i=1,2,...,n,\end{aligned}$$

where f_1 and f_2 are two densities for the tail, which could be chosen as truncated normal densities. Let $L^0=\prod_{i=1}^n \hat{f}_i^0$.

5. Propose a move. Suppose we are at the k-th iteration. Randomly pick a number τ_j from $\tau_1,\tau_2,...,\tau_m$ and then randomly pick a component $\beta_l^k(\tau_j)$ of $\beta^k(\tau_j)$ to update. To make sure that the proposed point $\beta_l^*(\tau_j)$ satisfying constraint (2.3), we can calculate a lower bound lb and an upper bound ub for $\beta_l^*(\tau_j)$ and generate a value for $\beta_l^*(\tau_j)$ from Uniform(lb,ub), and we will use a truncated normal as the proposal distribution in case $lb=-\infty$ or $ub=\infty$. For each observation (y_i,x_i), $i=1,2,...,n$ we can calculate a lower bound lb_i and an upper bound ub_i, $i=1,2,...,n$, and then $lb=\underset{i}{max}(lb_i)$ is taken as the maximum of all these lower bounds and $ub=\underset{i}{min}(ub_i)$ is taken as the minimum of all these upper bounds. The formula to calculate ub_i and lb_i is given as follows. If $1<j<m$ and $x_{i,l}>0$, where $x_{i,l}$ denote the l-th element of x_i, then

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

and

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

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

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

and

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

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

$$lb_i=-\infty,$$

and

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

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

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

and

$$ub_i = \infty.$$

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

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

and

$$ub_i = \infty.$$

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

$$lb_i = -\infty,$$

and

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

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

$$lb_i = 0,$$

and

$$ub_i = 0.$$

6. Once a move is proposed, let us set $B_m^* = (\beta^{k-1}(\tau_1), ..., \beta^{k-1}(\tau_{j-1}), \beta^*(\tau_j), \beta^{k-1}(\tau_{j+1}), ..., \beta^{k-1}(\tau_m))$. We can calculate the linear interpolated density $\hat{f}_i^*(y_i|x_i)$, $i = 1, 2, .., n$, through

$$\begin{aligned}\hat{f}_i^*(y_i|x_i) &= [\sum_{t\neq j, t\neq j-1, t<m} I_{\{y_i\in(x_i\beta^{k-1}(\tau_t), x_i\beta^{k-1}(\tau_{t+1}))\}} \frac{\tau_{t+1}-\tau_t}{x_i\beta^{k-1}(\tau_{t+1}) - x_i\beta^{k-1}(\tau_t)}] \\ &+ I_{\{y_i\in(x_i\beta^{k-1}(\tau_{j-1}), x_i\beta^*(\tau_j))\}} \frac{\tau_j-\tau_{j-1}}{x_i\beta^*(\tau_j) - x_i\beta^{k-1}(\tau_{j-1})} \\ &+ I_{\{y_i\in(x_i\beta^*(\tau_j), x_i\beta^{k-1}(\tau_{j+1}))\}} \frac{\tau_{j+1}-\tau_j}{x_i\beta^{k-1}(\tau_{j+1}) - x_i\beta^*(\tau_j)} \\ &+ I_{\{y_i\in(-\infty, x_i\beta^{k-1}(\tau_1))\}}\tau_1 f_1(y_i) + I_{\{y_i\in(x_i\beta^{k-1}(\tau_m), \infty)\}}(1-\tau_m) f_2(y_i), \quad i = 1, 2, ..., n.\end{aligned}$$

Let $L^* = \prod_{i=1}^n \hat{f}_i^*$.

7. Calculate the Metropolis-Hastings ratio

$$r = \frac{\pi(B_m^*)L^* p(B_m^{k-1} \to B_m^*)}{\pi(B_m^{k-1})L^{k-1} p(B_m^* \to B_m^{k-1})},$$

where $p(B_m^{k-1} \to B_m^*)$ denotes the transition probability from B_m^{k-1} to B_m^* and $p(B_m^* \to B_m^{k-1})$ denotes the transition probability from B_m^* to B_m^{k-1}. Notice that these two transition probabilities can cancel out if we choose symmetric proposals for the tails. If $r \geq 1$ then $B_m^k = B_m^*$; otherwise, 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 $L^k = L^{k-1}$.

8. Repeat Steps 5 - 7 until the desired number of iterations is reached.

2.3 The data generating method

Marjoram et. al. (2003) [14] proposed a Markov chain Monte Carlo (MCMC) without likelihoods method to deal with the case that the likelihood is not available or very hard to compute while generating data from the model is relative easy. Let us review their algorithm before we introduce the alternative method for the Bayesian quantile regression problem.

2.3.1 MCMC without likelihoods

Suppose data D come from a discrete distribution $f(D|\theta)$. If we want to draw samples from $p(\theta|D) \propto \pi(\theta)f(D|\theta)$, where $\pi(\theta)$ is the prior of θ, we can take the following steps.

1. Generate θ from $\pi(\theta)$.

2. Generate D' from $f(D|\theta)$.

3. Accept θ if $D = D'$.

If D follows a continuous distribution, we can change Step 3 to "accept θ if $\rho(D, D') < \epsilon$", where $\rho(D, D')$ is a measure of distance between D and D' and ϵ is a small quantity to control the accuracy. If S is the sufficient statistic for θ, then the acceptance rate may be improved by comparing the sufficient statistics. Combining this accept-reject algorithm with the Metropolis-Hastings algorithm, we can get the MCMC without likelihoods algorithm as follows.

1. Suppose the chain is at θ. We can propose a move θ' from a proposal distribution $t(\theta, \theta')$.

2. Generate D' from $f(D|\theta)$.

3. Calculate S' from D'.

4. If $S' = S$, then go to next step, and stay at θ otherwise.

5. Calculate the Metropolis-Hastings ratio

$$r = \frac{\pi(\theta')t(\theta' \to \theta)}{\pi(\theta)t(\theta \to \theta')},$$

and accept θ' with probability $min(r, 1)$.

Again, if D is continuous, we will introduce a distance measure and a tolerance quantity.

2.3.2 Generating data based on quantiles

If we can generate data based on quantiles, then we can use the MCMC without likelihoods method. Let us start from a simple case, where $Z_1, Z_2, ..., Z_n$ are iid with cdf $F(z)$. We know that if $F(z)$ is invertible, we can use the inverse cdf method to generate n samples from $F(z)$ through following steps.

1. Generate $u_1, u_2, ..., u_n$ from Uniform$(0, 1)$.

2. Calculate $z_i' = F^{-1}(u_i)$, $i = 1, 2, ..., n$.

If we only know m quantiles instead of the cdf F, say, $q_{\tau_1}, q_{\tau_2}, ..., q_{\tau_m}$, where $0 < \tau_1 < \tau_2 < ... < \tau_m < 1$, then we can use linear interpolations to generate samples as follows.

1. Generate $u_1, u_2, ..., u_n$ from Uniform$(0, 1)$.

2. If $\tau_j < u_i < \tau_{j+1}$, then

$$z_i' = q_{\tau_j} + \frac{q_{\tau_{j+1}} - q_{\tau_j}}{\tau_{j+1} - \tau_j}(u_i - \tau_j). \tag{2.4}$$

If $u_i < \tau_1$ or $u_i > \tau_m$, then z_i' can be generated from a truncated normal where z_i' need to be smaller than q_{τ_1} or greater than q_{τ_m}.

Now let us consider the linear regression case, where

$$y_i = x_i\beta + \epsilon_i, \quad i = 1, 2, ..., n.$$

For each point (x_i, y_i), we can calculate m quantiles $Q_{\tau_1}(y_i|x_i), Q_{\tau_2}(y_i|x_i), ..., Q_{\tau_m}(y_i|x_i)$ if we know B_m. To generate samples in this case, we need to modify the interpolation step to the following.

If $\tau_j < u_i < \tau_{j+1}$, then

$$y_i' = Q_{\tau_j}(y_i|x_i) + \frac{Q_{\tau_{j+1}}(y_i|x_i) - Q_{\tau_j}(y_i|x_i)}{\tau_{j+1} - \tau_j}(u_i - \tau_j). \tag{2.5}$$

If $u_i < \tau_1$ or $u_i > \tau_m$, then y_i' can be generated from a truncated normal where y_i' need to be smaller than $Q_{\tau_1}(y_i|x_i)$ or greater than $Q_{\tau_m}(y_i|x_i)$.

2.3.3 The algorithm of the data generating method

For the univariate case, where no covariates are involved, one can use order statistics as the sufficient statistic and use the Euclidean distance as the measure if the data are continuous. Suppose that the observed data is $Z = (z_1, z_2, ..., z_n)$ and continuous. In this case, the algorithm is as follows.

1. Pick m quantiles, say, the τ_1-th, the τ_2-th,..., and the τ_m-th quantiles, which should include the quantiles of interest.

2. Put a prior $\pi(q_{\tau_0:\tau_m})$ on $q_{\tau_0:\tau_m} = (q_{\tau_1}, q_{\tau_2}, ..., q_{\tau_m})$.

3. Choose an initial value $q^0_{\tau_0:\tau_m}$ for $q_{\tau_0:\tau_m}$. One can use the sample quantiles as the initial point.

4. Propose a move at the k-th iteration. One possible proposal can be chosen as follows. Randomly choose a number τ_j from $\tau_1, \tau_2, ..., \tau_m$. If $1 < j < m$, then $q^*_{\tau_j} \sim \text{Uniform}(q_{\tau_{j-1}}, q_{\tau_{j+1}})$. If $j = 1$ or $j = m$, then $q^*_{\tau_j}$ can be generated from a truncated normal, which should guarantee that $q^*_{\tau_j} < q_{\tau_1}$ or $q^*_{\tau_j} > q_{\tau_m}$. Let $q^*_{\tau_1:\tau_m} = (q_{\tau_1}, ..., q_{\tau_{j-1}}, q^*_{\tau_j}, q_{\tau_{j+1}}, ..., q_{\tau_m})$.

5. Generate $u_1, u_2, ..., u_n$ from Uniform$(0, 1)$. Use the interpolation scheme (2.4) proposed in Section 2.3.2 and plug in $q^*_{\tau_1:\tau_m}$. Denote the generated data as $Z' = (z_1', ..., z_n')$.

6. Calculate the order statistic $S' = (z'_{(1)}, ..., z'_{(n)})$.

7. Calculate the Euclidean distance between S and S'

$$\rho(S, S') = \sqrt{\sum_{i=1}^{n}(z'_{(i)} - z_{(i)})^2}.$$

Go to the next step if $\rho(S, S') < \epsilon$, where ϵ is a pre-specified tolerance quantity. Otherwise $q^k_{\tau_1:\tau_m} = q^{k-1}_{\tau_1:\tau_m}$.

8. Calculate the Metropolis-Hastings ratio

$$r = \frac{\pi(q^*_{\tau_1:\tau_m})t(q^{k-1}_{\tau_1:\tau_m} \to q^*_{\tau_1:\tau_m})}{\pi(q^{k-1}_{\tau_1:\tau_m})t(q^*_{\tau_1:\tau_m} \to q^{k-1}_{\tau_1:\tau_m})}.$$

Let $q^k_{\tau_1:\tau_m} = q^*_{\tau_1:\tau_m}$ with probability $min(r, 1)$.

9. Repeat Steps 4-8 until the desired number of iterations is reached.

For the linear model with one covariate case,

$$y_i = a + x_i\beta + \epsilon_i, \quad i = 1, 2, ..., n,$$

we would like to introduce summary statistics d_1, d_2 defined as

$$\begin{cases} d_{1,\tau_1} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(I_{(y_i \le q_{\tau_1}(y_i|x_i))} - \tau_1)/\sqrt{\tau_1(1-\tau_1)} \\ \quad ... \\ d_{1,\tau_m} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(I_{(y_i \le q_{\tau_m}(y_i|x_i))} - \tau_m)/\sqrt{\tau_m(1-\tau_m)}, \end{cases} \tag{2.6}$$

$$\begin{cases} d_{2,\tau_1} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(I_{(y_i \le q_{\tau_1}(y_i|x_i))} - \tau_1)x^*_i/\sqrt{\tau_1(1-\tau_1)} \\ \quad ... \\ d_{2,\tau_m} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(I_{(y_i \le q_{\tau_m}(y_i|x_i))} - \tau_m)x^*_i/\sqrt{\tau_m(1-\tau_m)}, \end{cases} \tag{2.7}$$

where $x^*_i = \frac{x_i}{max(|x_i|)}$ and $q_{\tau_j}(y_i|x_i)) = \hat{a}(\tau_j) + x_i\hat{\beta}(\tau_j)$, where $\hat{a}(\tau_j)$ and $\hat{\beta}(\tau_j)$ are the "quantreg" (a package in R) estimates based on the originally observed data (X, Y).

Let $D' = \{(x_i, y_i'), i = 1, 2, ..., n\}$ be the generated data. We can calculate d_1', d_2' by

$$\begin{cases} d'_{1,\tau_1} = \dfrac{1}{n}\sum_{i=1}^{n}(I_{(y_i' \le q_{\tau_1}(y_i|x_i))} - \tau_1)/\sqrt{\tau_1(1-\tau_1)} \\ \quad ... \\ d'_{1,\tau_m} = \dfrac{1}{n}\sum_{i=1}^{n}(I_{(y_i' \le q_{\tau_m}(y_i|x_i))} - \tau_m)/\sqrt{\tau_m(1-\tau_m)}, \end{cases} \tag{2.8}$$

$$\begin{cases} d'_{2,\tau_1} = \dfrac{1}{n}\sum_{i=1}^{n}(I_{(y_i' \le q_{\tau_1}(y_i|x_i))} - \tau_1)x_i^*/\sqrt{\tau_1(1-\tau_1)} \\ \quad ... \\ d'_{2,\tau_m} = \dfrac{1}{n}\sum_{i=1}^{n}(I_{(y_i' \le q_{\tau_m}(y_i|x_i))} - \tau_m)x_i^*/\sqrt{\tau_m(1-\tau_m)}. \end{cases} \tag{2.9}$$

We calculate the Euclidean distance between d_1, d_2 and d_1', d_2' to decide whether we will reject the move or not.

The algorithm is as follows.

1. Pick m quantiles, say, the τ_1-th, the τ_2-th,..., and the τ_m-th quantiles, which should include the quantiles of interests.

2. Put a prior $\pi(B_m)$ on $B_m = (a(\tau_1), a(\tau_2), ..., a(\tau_m), \beta(\tau_1), \beta(\tau_2), ..., \beta(\tau_m))$. One can use a truncated normal prior same as the one for the linear interpolated densities method.

3. Choose an initial value B_m^0. One can choose the initial value discussed at Step 3 of the linearly interpolated density method.

4. Calculate d_1 and d_2 through equations (2.6) and (2.7).

5. Propose a move. One can use the same strategy as Step 4 of the linearly interpolated density method. Denote the new point as B_m^*.

6. Generate data. Generate $u_1, u_2, ..., u_n$ from Uniform$(0, 1)$. Use the interpolation scheme (2.5) discussed in the Section 2.3.2 and plug in B_m^* to generate D'.

7. Calculate d_1' and d_2' through equations (2.8) and (2.9). Go to the next step if $\rho(d_1, d_1') < \epsilon_1$ and $\rho(d_2, d_2') < \epsilon_2$, where $\rho(\cdot, \cdot)$ is the Euclidean distance and ϵ_1 and ϵ_2 are two pre-specified tolerance quantities. Otherwise $B_m^k = B_m^{k-1}$.

8. Calculate the Metropolis-Hastings ratio

$$r = \frac{\pi(B^*)t(B^* \to B^{k-1})}{\pi(B_m^{k-1})t(B_m^{k-1} \to B_m^*)},$$

and accept B^* with probability $min(r, 1)$.

9. Repeat Steps 5 - 8 until the desired number of iterations is reached.

For the multi-covariates case, this method could be easily generalized, but more distances may be needed to compare instead of only d_1 and d_2.

Chapter 3

Theoretical Results of the Proposed Methods

In the previous chapter, we introduced two methods to solve the quantile regression problem. Both methods used the Markov chain Monte Carlo (MCMC) method, so it is important to know the stationary distribution of the Markov chain. The limiting property of the stationary distribution is also of interest. In this chapter, we will show the limiting property of the stationary distribution as $m \to \infty$, where m is the number of quantiles we use.

3.1 Stationary distribution of the linearly interpolated density method

Let us consider the linear model

$$y_i = x_i\beta + \epsilon_i, \quad i = 1, 2, ..., n, \tag{3.1}$$

where x_i is a $1 \times p$ 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 corresponding quantile model for the τ_j-th quantile is

$$Q_{\tau_j}(y_i|x_i) = x_i\beta(\tau_j), \quad i = 1, 2, ..., n, \quad j = 1, 2, ..., m. \tag{3.2}$$

Let $\pi_m(B_m)$ be the prior for $B_m = (\beta(\tau_1), \beta(\tau_2), ..., \beta(\tau_m))$ and $\hat{P}_m(Y|X, B_m) = \prod_{i=1}^{n} \hat{f}_{i,m}(y_i)$ denote the linear interpolated likelihood, where $\hat{f}_{i,m}(y_i)$ denotes the linear interpolated density for the i-th observation and can be calculated by

$$\begin{aligned}\hat{f}_{i,m}(y_i|x_i) &= [\sum_{j=1}^{m-1} I_{\{y_i \in (x_i\beta(\tau_j), x_i\beta(\tau_{j+1}))\}} \frac{\tau_{j+1} - \tau_j}{x_i\beta(\tau_{j+1}) - x_i\beta(\tau_j)}] + I_{\{y_i \in (-\infty, x_i\beta(\tau_1))\}}\tau_1 f_1(y_i) \\ &\quad + I_{\{y_i \in (x_i\beta(\tau_m), \infty)\}}(1 - \tau_m) f_2(y_i), \quad i = 1, 2, ..., n,\end{aligned}$$

where f_1 and f_2 are two densities for the tail. Therefore, the posterior distribution of B_m based on the linear interpolated likelihood is $\hat{P}_m(B_m|X,Y) = \pi_m(B_m)\hat{P}_m(Y|X,B_m)/\hat{P}_m(Y|X)$, where $\hat{P}_m(Y|X) = \int \pi_m(B_m)\hat{P}_m(Y|X,B_m)dB_m$.

Proposition 3.1.1. *$\hat{P}_m(B_m|X,Y)$ is the stationary distribution of the Markov chain constructed through the linearly interpolated density method.*

Proof: We will verify the detailed balance condition to show the stationary distribution. Denote the probability 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

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

□

3.2 Limiting distribution of the stationary distribution for the linearly interpolated density method

In this section, we will first show that the stationary distribution will converge to some distribution in terms of the total variation norm as $m \to \infty$, and then we will show that after we construct a Markov chain with increasing m the distribution at each step of the Markov chain will converge to some distribution as $m \to \infty$.

To study the limiting distribution as $m \to \infty$, we need to define a way to increase the number of quantiles. Let us start from $m_0 = M_0 - 1$ quantiles: the $\frac{1}{M_0}$-th, $\frac{2}{M_0}$-th, ..., $\frac{M_0-1}{M_0}$-th quantiles. We can add new quantiles one by one in the following way: the $\frac{1}{2M_0}$-th, $\frac{3}{2M_0}$-th,..., $\frac{2M_0-1}{2M_0}$-th, $\frac{1}{4M_0}$-th,

$\frac{3}{4M_0}$-th,..., $\frac{4M_0-1}{4M_0}$-th quantiles and so on. Through this definition, we can see that

$$\Delta\tau = \max_{0\le j\le m}(\tau_{j+1}-\tau_j) \le \frac{2}{m} = O(\frac{1}{m}), \tag{3.3}$$

where $\tau_0 = 0$ and $\tau_{m+1} = 1$. We need the following assumption to show the limiting distribution.

Assumption 3.2.1. *Let $\mathscr{F} = \{f | \int f dx = 1, 0 \le f \le M_1, |f'| < M_2$, and $f(x) < \frac{c}{\sqrt{m}}$ for $x < q_{\frac{1}{m}}$ and $x > q_{\frac{m-1}{m}}\}$, where m is any positive integer, the quantities $q_{\frac{1}{m}}$ and $q_{\frac{m-1}{m}}$ are the $\frac{1}{m}$-th and $\frac{m-1}{m}$-th quantile of f and M_1, M_2 and c are constants. We need to assume that all the densities of $y_i|x_i$ we considered are in this set.*

From the assumption, we can see that $\mathscr{F}$ is a set of probability density functions with bounded value, bounded first derivative and not too heavy tails. We can show that the Cauchy distribution, which has fairly heavy tails, is in the set. For the Cauchy distribution, the $\frac{1}{m}$-th quantile is $q_{\frac{1}{m}} = tan(\pi(\frac{1}{m}-\frac{1}{2})) = -ctan(\frac{\pi}{m})$, so $f(q_{\frac{1}{m}}) = \frac{1}{\pi}\frac{1}{1+ctan^2(\frac{\pi}{m})} = \frac{1}{\pi}sin^2(\frac{\pi}{m}) = O(\frac{1}{m^2}) < \frac{c}{\sqrt{m}}$ for some c.

Let us consider $\beta(\tau)$ as a function of τ, where $0 \le \tau \le 1$. One possible prior for $\beta(\tau)$ is the Gaussian process prior. The prior of $f_i(y_i|x_i)$ can be induced from the prior of $\beta(\tau)$ because $x_i\beta(\tau)$ can give all the quantiles of $f_i(y_i|x_i)$, which will determine $f_i(y_i|x_i)$. We can also obtain the priors on B_m from the prior of $\beta(\tau)$ because B_m is a vector of m point on $\beta(\tau)$. Denote the prior on $f_i(y_i|x_i)$ by $\pi(f_i)$ and the prior on B_m by $\pi_m(B_m)$.

Definition 3.2.1. *Let θ_{f_i} be all the quantiles of f_i and $\theta_m = x_iB_m$ to be the m quantiles we are using.*

Proposition 3.2.1. *Let $\hat{P}_m(y_i|\theta_m)$ denote the linear interpolated density of y_i given that the m quantiles are θ_m. Let $P(y_i|\theta_{f_i}) = f_i(y_i)$ denote the true density given that the pdf of $y_i|x_i$ is f_i. Then $\hat{P}_m(y_i|\theta_m) \to P(y_i|\theta_{f_i})$ as $m \to \infty$.*

Proof:

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,τ_j+1} such that $y_i \in [q_{i,\tau_j}, q_{i,\tau_{j+1}})$, where $1 \le j \le m-1$, then by the mechanism of linear interpolation, we have the following equation.

$$
\begin{aligned}
& \hat{P}_m(y_i|\theta_m) \\
= & \frac{\tau_{j+1}-\tau_j}{q_{i,\tau_{j+1}}-q_{i,\tau_j}} \\
= & \frac{\tau_{j+1}-\tau_j}{F_i^{-1}(\tau_{j+1})-F_i^{-1}(\tau_j)} \\
= & \frac{\tau_{j+1}-\tau_j}{(F_i^{-1})'(\tau^*)(\tau_{j+1}-\tau_j)} \quad \text{(By the mean value theorem)} \\
= & \frac{\tau_{j+1}-\tau_j}{\frac{1}{f_i(y_i^*)}(\tau_{j+1}-\tau_j)} \\
= & f_i(y_i^*)
\end{aligned}
$$

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

$$
|f_i(y_i^*) - f_i(y_i)| \leq \sup_{y \in [q_{i,\tau_j}, q_{i,\tau_{j+1}})} f_i(y) - \inf_{y \in [q_{i,\tau_j}, q_{i,\tau_{j+1}})} f_i(y) \leq M_2\delta, \tag{3.4}
$$

where $\delta = \sqrt{\frac{2(\tau_{j+1}-\tau_j)}{M_2}}$. If $q_{i,\tau_{j+1}} - q_{i,\tau_j} \leq \delta$, then $|f_i(y_i^*) - f_i(y_i)| = |f_i'(y^\dagger)(y_i^* - y_i)| \leq M_2\delta$, 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

$$
\int_{q_{i,\tau_j}}^{q_{i,\tau_{j+1}}} f_i(y)dy > \tau_{j+1} - \tau_j,
$$

if

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

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 $y_{sup} - y_{inf} \leq \delta$, then

$$
\sup_{y \in (q_{i,\tau_j}, q_{i,\tau_{j+1}})} f_i(y) - \inf_{y \in (q_{i,\tau_j}, q_{i,\tau_{j+1}})} f_i(y) = f_i(y_{sup}) - f_i(y_{inf}) = |f_i'(y^\dagger)|(y_{sup} - y_{inf}) \leq M_2\delta.
$$

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 $f_i(y) - f_i(y_{sup}) = f'(y^{\dagger\dagger})(y - y_{sup}) \geq M_2(y - y_{sup})$ for $y < y_{sup}$,

which leads to $f_i(y) \geq f_i(y_{sup}) + M_2(y - y_{sup})$.

Now we can check the area S formed by the line, $y = y_{inf}$, $y = y_{sup}$, and $f_i(y) = 0$. Figure 3.1 shows two possible cases. The shaded region is S.

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

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

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

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

Therefore, in both cases, we show

$$\int_{q_{i,\tau_j}}^{q_{i,\tau_{j+1}}} f_i(y)dy \geq \int_{y_{inf}}^{y_{sup}} f_i(y)dy \geq S > \tau_{t+1} - \tau_j,$$

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

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

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

Now let us consider the second case.

Case2: If y_i is a point in the tail, which means that $y_i \leq q_{i,\tau_1}$ or $y_i > q_{i,\tau_m}$, then we can see $P(y_i|\theta_{f_i}) = f_i(y_i) < \frac{c}{\sqrt{m}}$ from the Assumption 3.2.1. For the tail part, we can use a truncated normal to do the interpolation so that $\hat{P}_m(y_i|\theta_m) < \frac{c}{\sqrt{m}}$. Therefore, we find $|\hat{P}_m(y_i|\theta_m) - P(y_i|\theta_{f_i})| < \frac{2c}{\sqrt{m}} = O(\frac{1}{\sqrt{m}})$.

In both cases, we showed $|\hat{P}_m(y_i|\theta_m) - P(y_i|\theta_{f_i})| = O(\frac{1}{\sqrt{m}})$. □

Definition 3.2.2. *Let $P(y_i|\theta_m) = \int_{f_i \in \mathscr{F}_{\theta_m}} P(y_i|\theta_{f_i})\pi(f_i|\theta_m)df_i$, where $\mathscr{F}_{\theta_m}$ denotes the subset of $\mathscr{F}$ that contains all the pdfs with those m quantiles equal to θ_m and $\pi(f_i|\theta_m)$ denotes the prior of $f_i|\theta_m$ which is induced by $\pi(f_i)$.*

Under this definition of $P(y_i|\theta_m)$, we can show the following proposition.

Proposition 3.2.2. $|\hat{P}_m(Y|X, B_m) - P(Y|X, B_m)| = O(\frac{1}{\sqrt{m}})$.

Proof: Let us show $|\hat{P}_m(y_i|\theta_m) - P(y_i|\theta_m)| = O(\frac{1}{\sqrt{m}})$ first.

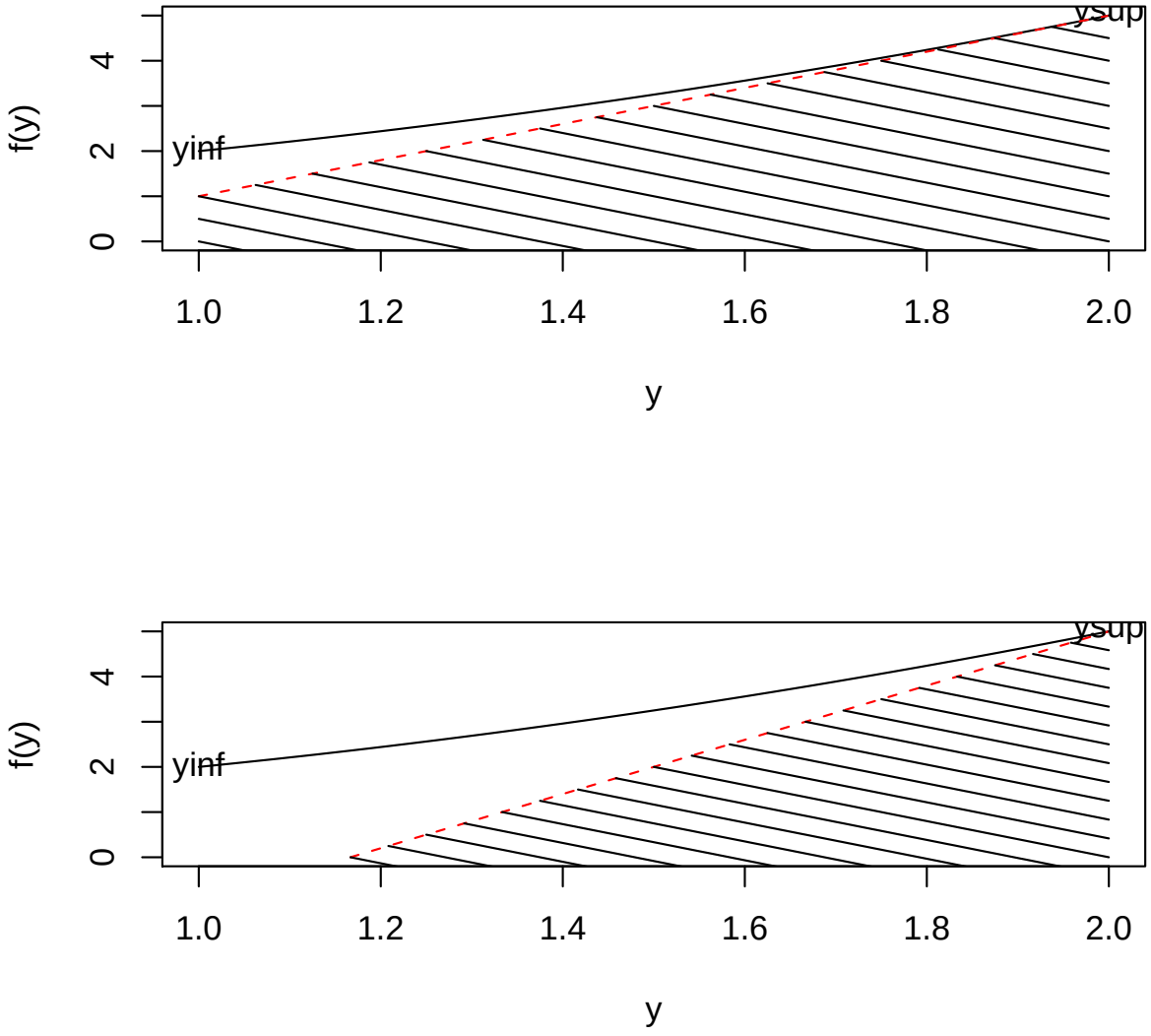

Figure 3.1: Example of the 2 possible cases of the area: trapezia or triangle. The solid curve stands for f(y). The dotted line stands for the line we constructed. And the shaded area is S.

$$
\begin{aligned}
& |\hat{P}_m(y_i|\theta_m) - P(y_i|\theta_m)| \\
= \; & |\int_{f\in\mathscr{F}_{\theta_m}} \hat{P}_m(y_i|\theta_m)\pi(f|\theta_m)df - \int_{f\in\mathscr{F}_{\theta_m}} P(y_i|\theta_f)\pi(f|\theta_m)df| \\
\leq \; & \int_{f\in\mathscr{F}_{\theta_m}} \pi(f|\theta_m)|\hat{P}_m(y_i|\theta_m) - P(y_i|\theta_f)|df \\
= \; & O(\frac{1}{\sqrt{m}}).
\end{aligned}
$$

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

$$
\begin{aligned}
& |\hat{P}_m(Y|X,B_m) - P(Y|X,B_m)| \\
= \; & |\hat{P}_m(y_1|X,B_m)\hat{P}_m(y_2|X,B_m) - P(y_1|X,B_m)P(y_2|X,B_m)| \\
= \; & |\hat{P}_m(y_1|X,B_m)\hat{P}_m(y_2|X,B_m) - \hat{P}_m(y_1|X,B_m)P(y_2|X,B_m) \\
& +\hat{P}_m(y_1|X,B_m)P(y_2|X,B_m) - P(y_1|X,B_m)P(y_2|X,B_m)| \\
\leq \; & |\hat{P}_m(y_1|X,B_m)(\hat{P}_m(y_2|X,B_m) - P(y_2|X,B_m))| + |(\hat{P}_m(y_1|X,B_m) - P(y_1|X,B_m))P(y_2|X,B_m)| \\
= \; & M_1O(\frac{1}{\sqrt{m}}) + M_1O(\frac{1}{\sqrt{m}}) \\
= \; & O(\frac{1}{\sqrt{m}})
\end{aligned}
$$

For the case that $n > 2$, the proof can be easily generalized. □

Proposition 3.2.3. $E_{\pi_m}(|\hat{P}_m(Y|X,B_m) - P(Y|X,B_m)|) = O(\frac{1}{\sqrt{m}})$

The proof for this proposition is simply using the conclusion of Proposition 3.2.2. □

Proposition 3.2.4. $E_{\pi_m}(|\hat{P}_m(Y|X,B_m) - \hat{P}_{m-1}(Y|B_{m-1},X)|) = O(\frac{1}{\sqrt{m}})$

The proof for this proposition is simply using the conclusion of Proposition 3.2.3 twice. □

In order to prove the convergence of one distribution to another, we need to introduce a norm to measure the discrepancy. Here we will use the total variation norm, which will be denoted by $\|\cdot\|_{TV}$

Definition 3.2.3. *If μ_1 and μ_2 are probability measures, $||\mu_1 - \mu_2||_{TV} = \sup_A |\mu_1(A) - \mu_2(A)|$, where A denotes any measurable set.*

The following proposition, which appears as a homework problem of Robert and Casella (2004) [16] p. 253, gives an equivalent definition.

Proposition 3.2.5. $||\mu_1 - \mu_2||_{TV} = \frac{1}{2}\sup_{|h|\leq 1} |\int h(x)\mu_1(dx) - \int h(x)\mu_2(dx)|$.

Proof: Assuming that M is a measurable set such that $\sup_A |\mu_1(A) - \mu_2(A)| = \mu_1(M) - \mu_2(M)$, we can see that $M^- = \{a|a \in M, \mu_1'(a) - \mu_2'(a) < 0\}$ has measure 0 on both μ_1 and μ_2. Otherwise, $\mu_1(M^-) - \mu_2(M^-) = \int_{M^-} \mu_1'(x) - \mu_2'(x)dx < 0$, and in this case if we define $M^+ = \{a|a \in M, \mu_1'(a) - \mu_2'(a) \geq 0\}$, then

$$\begin{aligned} & \mu_1(M^+) - \mu_2(M^+) \\ = \; & \mu_1(M) - \mu_1(M^-) - (\mu_2(M) - \mu_2(M^-)) \\ = \; & \mu_1(M) - \mu_2(M) - (\mu_1(M^-) - \mu_2(M^-)) \\ > \; & \mu_1(M) - \mu_2(M), \end{aligned}$$

which contradicts with the definition of M. Without loss of generality, let us assume that $\mu_1'(x) - \mu_2'(x) \geq 0$ for all $x \in M$, so $\mu_1'(x) - \mu_2'(x) < 0$ for all $x \in \bar{M}$, where $\bar{M}$ is the complementary set of M. Define $h_0(x) = 1$ if $x \in M$, $h_0(x) = -1$ if $x \in \bar{M}$. Then,

$$\begin{aligned} & \int h_0(x)\mu_1(dx) - \int h_0(x)\mu_2(dx) \\ = \; & \mu_1(M) - \mu_2(M) + \mu_2(\bar{M}) - \mu_1(\bar{M}) \\ = \; & \mu_1(M) - \mu_2(M) + (1 - \mu_2(M)) - (1 - \mu_1(M)) \\ = \; & 2(\mu_1(M) - \mu_2(M)). \end{aligned}$$

Hence,

$$\begin{aligned} & ||\mu_1 - \mu_2||_{TV} \\ = \; & \frac{1}{2}(\int h_0(x)\mu_1(dx) - \int h_0(x)\mu_2(dx)) \\ \leq \; & \frac{1}{2}\sup_{|h|\leq 1} |\int h(x)\mu_1(dx) - \int h(x)\mu_2(dx)|. \end{aligned}$$

We can also show the other direction as follows.

$$\begin{aligned}
& \frac{1}{2}\sup_{|h|\leq 1}|\int h(x)\mu_1(dx)-\int h(x)\mu_2(dx)| \\
\leq\ & \frac{1}{2}\int|\mu_1'(x)-\mu_2'(x)|dx \\
=\ & \frac{1}{2}\int_M(\mu_1'(x)-\mu_2'(x))dx+\int_{\bar{M}}(\mu_2'(x)-\mu_1'(x))dx \\
=\ & \frac{1}{2}(\mu_1(M)-\mu_2(M)+\mu_2(\bar{M})-\mu_1(\bar{M})) \\
=\ & \mu_1(M)-\mu_2(M)=||\mu_1-\mu_2||_{TV}
\end{aligned}$$

Therefore, $||\mu_1-\mu_2||_{TV}=\frac{1}{2}\sup_{|h|\leq 1}|\int h(x)\mu_1(dx)-\int h(x)\mu_2(dx)|$. □

Now we would like to prove that $\hat{P}_m(B_m|X,Y)\to P(B_m|X,Y)$ as $m\to\infty$. We need to show the following proposition first.

Proposition 3.2.6. $|\hat{P}_m(Y|X)-P(Y|X)|=O(\frac{1}{\sqrt{m}})$

Proof:

$$\begin{aligned}
& |\hat{P}_m(Y|X)-P(Y|X)| \\
=\ & |\int\pi_m(B_m)(\hat{P}_m(Y|X,B_m)-P(Y|X,B_m))dB_m| \\
\leq\ & \int\pi_m(B_m)|(\hat{P}_m(Y|X,B_m)-P(Y|X,B_m))|dB_m \\
=\ & E_{\pi_m}(|\hat{P}_m(Y|X,B_m)-P(Y|X,B_m)|)=O(\frac{1}{\sqrt{m}}) \quad \text{(By Proposition 3.2.3)}
\end{aligned}$$

□

Now we can prove the following theorem which gives the limiting distribution of the stationary distribution as $m\to\infty$.

Theorem 3.2.1. $\|\hat{P}_m(B_m|X,Y)-P(B_m|X,Y)\|_{TV}\to 0$ *as* $m\to\infty$.

Proof:

$$
\begin{aligned}
& ||\hat{P}_m(B_m|X,Y) - P(B_m|X,Y)||_{TV} \\
= \; & \frac{1}{2} \sup_{|h|\leq 1} | \int h(B_m)(\frac{\pi_m(B_m)\hat{P}_m(Y|X,B_m)}{\hat{P}_m(Y|X)} - \frac{\pi_m(B_m)P(Y|X,B_m)}{P(Y|X)})dB_m| \\
\leq \; & \frac{1}{2} \int \pi_m(B_m)|\frac{\hat{P}_m(Y|X,B_m)}{\hat{P}_m(Y|X)} - \frac{P(Y|X,B_m)}{P(Y|X)}|dB_m \\
= \; & \frac{1}{2} \int \pi_m(B_m)|\frac{\hat{P}_m(Y|X,B_m)P(Y|X) - \hat{P}_m(Y|X)P(Y|X,B_m)}{\hat{P}_m(Y|X)P(Y|X)}|dB_m \\
= \; & \frac{1}{2} \int \pi_m(B_m)|\frac{(\hat{P}_m(Y|X,B_m) - P(Y|X,B_m))P(Y|X) + P(Y|X,B_m)(P(Y|X) - \hat{P}_m(Y|X))}{\hat{P}_m(Y|X)P(Y|X)}|dB_m \\
\leq \; & \frac{1}{2} \int \pi_m(B_m)\frac{|(\hat{P}_m(Y|X,B_m) - P(Y|X,B_m)|P(Y|X) + P(Y|X,B_m)|P(Y|X) - \hat{P}_m(Y|X)|}{\hat{P}_m(Y|X)P(Y|X)}dB_m \\
= \; & \frac{1}{2}(\frac{E_{\pi_m}(|\hat{P}_m(Y|X,B_m) - P(Y|X,B_m)|)}{\hat{P}_m(Y|X)} + \frac{|\hat{P}_m(Y|X) - P(Y|X)|}{\hat{P}_m(Y|X)})
\end{aligned}
$$

As we already know that $\hat{P}_m(Y|X) \to P(Y|X)$ as $m \to \infty$ by Proposition 3.2.6, we can choose any $e^* > 0$ such that $e^* < P(Y|X)$. Given this e^*, there exists an m^* such that $|\hat{P}_m(Y|X) - P(Y|X)| < e^*$ for $m > m^*$. We can see that $LB = min(\hat{P}_{m_0}(Y|X), \hat{P}_{m_0+1}(Y|X), ..., \hat{P}_{m^*-1}(Y|X), P(Y|X) - e^*)$ is a lower bound for $\hat{P}_m(Y|X)$, where m_0 is the minimum number of quantiles we use.
Therefore, $||\hat{P}_m(B_m|X,Y) - P(B_m|X,Y)||_{TV} < \frac{1}{2LB}(E_{\pi_m}(|\hat{P}_m(Y|X,B_m) - P(Y|X,B_m)|) + |\hat{P}_m(Y|X) - P(Y|X)|) = O(\frac{1}{\sqrt{m}}) \to 0$ as $m \to \infty$. □

Definition 3.2.4. *Let η denote the parameters of quantiles on which we want to make inference. They should be included in B_m. Let $\pi(\eta)$ denote the prior distribution of η, which is induced by $\pi_m(B_m)$.*

Definition 3.2.5. *Let $f_{t,m}(B_m)$ denotes the density of the t-th step of the chain that uses m quantiles and $g_{t,m}(\eta)$ denotes the marginal density of η by integrating out other variables of B_m from $f_{t,m}(B_m)$.*

Proposition 3.2.7. *Suppose that $f_1(B_m)$ and $f_2(B_m)$ are two pdfs of B_m and $g_1(\eta)$ and $g_2(\eta)$ are the marginal pdfs by integrating out other variables of B_m from $f_1(B_m)$ and $f_2(B_m)$ respectively. If $||f_1 - f_2||_{TV} < \epsilon$, then $||g_1 - g_2||_{TV} \leq ||f_1 - f_2||_{TV} < \epsilon$.*

Proof: Denote $\bar{\eta} = B_m \setminus \eta$, then we will have:

$$
\begin{aligned}
& ||g_1(\eta) - g_2(\eta)||_{TV} \\
= \; & \frac{1}{2} \sup_{|h|\leq 1} |\int h(\eta)(g_1(\eta) - g_2(\eta))d\eta| \\
= \; & \frac{1}{2} \sup_{|h|\leq 1} |\int h(\eta) \int (f_1(B_m) - f_2(B_m))d\bar{\eta}d\eta| \\
= \; & \frac{1}{2} \sup_{|h^*|\leq 1} |\int h^*(B_m)(f_1(B_m) - f_2(B_m))dB_m| \quad (\text{where } h^*(B_m) = h(\eta)) \\
\leq \; & \frac{1}{2} \sup_{|h|\leq 1} |\int h(B_m)(f_1(B_m) - f_2(B_m))dB_m| \\
= \; & ||f_1(\eta) - f_2(\eta)||_{TV} \\
< \; & \epsilon
\end{aligned}
$$

□

Definition 3.2.6. *Denote $\hat{P}_m(\eta|X,Y)$ and $P_m(\eta|X,Y)$ as the distributions of $\eta|X,Y$ by integrating out other variables of B_m from $\hat{P}_m(B_m|X,Y)$ and $P_m(B_m|X,Y)$, respectively.*

From the result of Proposition 3.2.7, we can obtain the following corollary.

Corollary 3.2.1. *If $||\hat{P}_m(B_m|X,Y) - P(B_m|X,Y)||_{TV} \leq \epsilon$, then $||\hat{P}_m(\eta|X,Y) - P(\eta|X,Y)||_{TV} \leq \epsilon$.*

Proposition 3.2.8. *If a sequence $\{a_n\}$ converges to 0, which is to say that $|a_n| \to 0$ as $n \to \infty$, then we can find a strictly decreasing sequence $\{\epsilon_n\}$ such that $\epsilon_n \to 0$ as $n \to \infty$ and $|a_n| < \epsilon_n$.*

Proof: Because $|a_n| \to 0$ as $n \to \infty$, we can find a strictly decreasing sequence $\{\delta_m\}$ such that $\delta_m \to 0$ as $m \to \infty$ and $\delta_1 > |a_1|$. For any m, there exists some N_m, where N_m is increasing with respect to m and $N_1 = 1$, such that $|a_n| < \delta_m$ for $n \geq N_m$.
we can choose $\epsilon_{N_m} = \delta_{m-1}$, where $\delta_0 = \delta_1 + 1$. Let $\epsilon_k = \epsilon_{N_m} - \frac{k-N_m}{N_{m+1}-N_m}(\epsilon_{N_m} - \epsilon_{N_{m+1}})$ if $N_m < k < N_{m+1}$. We want to check that this sequence $\{\epsilon_k\}$ satisfies all the conditions. First, it is obvious that this sequence is strictly decreasing because ϵ_{N_m} is strictly decreasing with respect to m. Second, we need to check that $|a_n| < \epsilon_n$. We can see that $\epsilon_{N_m} = \delta_{m-1} > |a_n|$ for $n \geq N_{m-1}$ and $N_m > N_{m-1}$, so $\epsilon_{N_m} > |a_{N_m}|$. If $N_m < k < N_{m+1}$, then $\epsilon_k > \epsilon_{N_{m+1}} = \delta_m > |a_n|$ for $n > N_m$, which implies $\epsilon_k > |a_k|$. Therefore, for any $k \in \mathbb{N}$, we have $|a_k| < \epsilon_k$. □.
Now let us construct a chain with increasing number of quantiles as follows. First, choose a strictly

decreasing sequence $\{e_m\}_{m=m_0}^{\infty}$ such that $||\hat{P}_m(B_m|X,Y) - P(B_m|X,Y)||_{TV} \leq e_m$. Second, assume that the chain is Harris positive and aperiodic. Start the chain with m_0 quantiles, which should include η. After generating T_{m_0} samples such that $||f_{T_{m_0},m_0}(B_{m_0}) - \hat{P}_{m_0}(B_{m_0}|X,Y)||_{TV} < e_{m_0}$, we can add one more quantile, using the strategy discussed at the beginning of the section. Let us denote the prior of this new quantile conditionally on other quantiles to be $\pi(B_{new}|B_m)$. Then after generating T_{m_0+1} samples such that $||f_{T_{m_0+1},m_0+1}(B_{m_0+1}) - \hat{P}_{m_0+1}(B_{m_0+1}|X,Y)||_{TV} < e_{m_0+1}$, we can add another quantile, and so on.

Theorem 3.2.2. *$\|g_{t,m}(\eta) - P(\eta|X,Y)\|_{TV} \to 0$ as $m \to \infty$.*

Proof: By Propositions 3.2.4 and 3.2.8, we can find a decreasing sequence δ_m such that $E_{\pi_m}(|\hat{P}_m(Y|X,B_m) - \hat{P}_{m-1}(Y|X,B_{m-1})|) \leq \delta_m$ and $\delta_m \to 0$ as $m \to 0$.

We will divide the proof into two parts.

Part1: we will show that $||g_{1,m}(\eta) - P(\eta|X,Y)||_{TV} \to 0$ as $m \to \infty$.

Suppose that $m > m_0$, then $f_{1,m}(B_m) = \int \pi(B'_{new}|B'_{m-1}) f_{T_{m-1},m-1}(B'_{m-1}) K_m(B'_m, B_m) dB'_m$, where $B'_m = (B'_{m-1}, B'_{new})$ and K_m is the transition kernel for the m-th step.

Let's check the following equation first.

$$
\begin{aligned}
& \|f_{1,m}(B_m) - \hat{P}_m(B_m|X,Y)\|_{TV} \\
= \; & \frac{1}{2} \sup_{|h|\leq 1} \Big| \int h(B_m) \Big(\int \pi(B'_{new}|B'_{m-1}) f_{T_{m-1},m-1}(B'_{m-1}) K_m(B'_m, B_m) dB'_m \\
& - \hat{P}_m(B_m|X,Y) \Big) dB_m \Big| \\
= \; & \frac{1}{2} \sup_{|h|\leq 1} \Big| \int h(B_m) \Big(\int \pi(B'_{new}|B'_{m-1}) f_{T_{m-1},m-1}(B'_{m-1}) K_m(B'_m, B_m) dB'_m \\
& - \int \hat{P}_m(B'_m|X,Y) K_m(B'_m, B_m) dB'_m \Big) dB_m \Big| \quad \text{(Property of the stationary distribution)} \\
= \; & \frac{1}{2} \sup_{|h|\leq 1} \Big| \int [\pi(B'_{new}|B'_{m-1}) f_{T_{m-1},m-1}(B'_{m-1}) - \hat{P}_m(B'_m|X,Y)] \int h(B_m) K_m(B'_m, B_m) dB_m dB'_m \Big|.
\end{aligned}
$$

Let $h^*(B'_m) = \int h(B_m) K_m(B'_m, B_m) dB_m$. It is not difficult to see that $h^*(B'_m) \leq 1$, so we can

rewrite the above equation as follows.

$$
\begin{aligned}
& ||f_{1,m}(B_m) - \hat{P}_m(B_m)||_{TV} \\
= \; & \frac{1}{2} \sup_{|h|\leq 1} | \int h^*(B'_m)(\pi(B'_{new}|B'_{m-1}) f_{T_{m-1},m-1}(B'_{m-1}) - \hat{P}_m(B'_m|X,Y)) dB'_m| \\
\leq \; & \frac{1}{2} \sup_{|h|\leq 1} | \int h(B'_m)(\pi(B'_{new}|B'_{m-1}) f_{T_{m-1},m-1}(B'_{m-1}) - \hat{P}_m(B'_m|X,Y)) dB'_m| \\
= \; & ||\pi(B_{new}|B_{m-1}) f_{T_{m-1},m-1}(B_{m-1}) - \hat{P}_m(B_m|X,Y)||_{TV} \\
\leq \; & ||\pi(B_{new}|B_{m-1}) f_{T_{m-1},m-1}(B_{m-1}) - \pi(B_{new}|B_{m-1}) \hat{P}_{m-1}(B_{m-1}|X,Y)||_{TV} \\
& + ||\pi(B_{new}|B_{m-1}) \hat{P}_{m-1}(B_{m-1}|X,Y) - \hat{P}_m(B_m|X,Y)||_{TV}.
\end{aligned}
$$

Now we want to show that

a) $||\pi(B_{new}|B_{m-1}) f_{T_{m-1},m-1}(B_{m-1}) - \pi(B_{new}|B_{m-1}) \hat{P}_{m-1}(B_{m-1}|X,Y)||_{TV} \leq e_{m-1}$.

b) $||\pi(B_{new}|B_{m-1}) \hat{P}_{m-1}(B_{m-1}|X,Y) - \hat{P}_m(B_m|X,Y)||_{TV} \leq C\delta_m$, where C is some constant.

Let us show a) first,

$$
\begin{aligned}
& ||\pi(B_{new}|B_{m-1}) f_{T_{m-1},m-1}(B_{m-1}) - \pi(B_{new}|B_{m-1}) \hat{P}_{m-1}(B_{m-1}|X,Y)||_{TV} \\
= \; & \frac{1}{2} \sup_{|h|\leq 1} | \int h(B_m)(\pi(B_{new}|B_{m-1}) f_{T_{m-1},m-1}(B_{m-1}) - \pi(B_{new}|B_{m-1}) \hat{P}_{m-1}(B_{m-1}|X,Y)) dB_m| \\
= \; & \frac{1}{2} \sup_{|h|\leq 1} | \int h(B_m) \pi(B_{new}|B_{m-1})(f_{T_{m-1},m-1}(B_{m-1}) - \hat{P}_{m-1}(B_{m-1}|X,Y)) dB_m| \\
= \; & \frac{1}{2} \sup_{|h|\leq 1} | \int (f_{T_{m-1},m-1}(B_{m-1}) - \hat{P}_{m-1}(B_{m-1}|X,Y))(\int h(B_m) \pi(B_{new}|B_{m-1}) dB_{new}) dB_{m-1}|.
\end{aligned}
$$

Let $h^*(B_{m-1}) = \int h(B_m)\pi(B_{new}|B_{m-1})dB_{new}$, It is easy to see that $h^*(B_{m-1}) \leq 1$, so we can rewrite the above equation as follows.

$$
\begin{aligned}
& ||\pi(B_{new}|B_{m-1}) f_{T_{m-1},m-1}(B_{m-1}) - \pi(B_{new}|B_{m-1}) \hat{P}_{m-1}(B_{m-1}|X,Y)||_{TV} \\
= \; & \frac{1}{2} \sup_{|h|\leq 1} | \int h^*(B_{m-1})(f_{T_{m-1},m-1}(B_{m-1}) - \hat{P}_{m-1}(B_{m-1}|X,Y)) dB_{m-1} \\
\leq \; & \frac{1}{2} \sup_{|h|\leq 1} | \int h(B_{m-1})(f_{T_{m-1},m-1}(B_{m-1}) - \hat{P}_{m-1}(B_{m-1}|X,Y)) dB_{m-1} \\
\leq \; & ||f_{T_{m-1},m-1}(B_{m-1}) - \hat{P}_{m-1}(B_{m-1}|X,Y)||_{TV} \\
\leq \; & e_{m-1}.
\end{aligned}
$$

Now let us prove b),

$$
\begin{aligned}
& ||\pi(B_{new}|B_{m-1})\hat{P}_{m-1}(B_{m-1}|X,Y) - \hat{P}_m(B_m|X,Y)||_{TV} \\
= \; & \frac{1}{2}\sup_{|h|\le 1}|\int h(B_m)(\pi(B_{new}|B_{m-1})\hat{P}_{m-1}(B_{m-1}|X,Y) - \hat{P}_m(B_m|X,Y))dB_m| \\
= \; & \frac{1}{2}\sup_{|h|\le 1}|\int h(B_m)(\pi(B_{new}|B_{m-1})\frac{\pi_{m-1}(B_{m-1})\hat{P}_{m-1}(Y|X,B_{m-1})}{\hat{P}_{m-1}(Y|X)} - \frac{\pi_m(B_m)\hat{P}_m(Y|X,B_m)}{\hat{P}_m(Y|X)})dB_m| \\
\le \; & \frac{1}{2}\int \pi_m(B_m)|\frac{\hat{P}_{m-1}(Y|X,B_{m-1})}{\hat{P}_{m-1}(Y|X)} - \frac{\hat{P}_m(Y|X,B_m)}{\hat{P}_m(Y|X)}|dB_m \\
= \; & \frac{1}{2}\int \pi_m(B_m)|\frac{\hat{P}_{m-1}(Y|X,B_{m-1})\hat{P}_m(Y|X) - \hat{P}_m(Y|X,B_m)\hat{P}_{m-1}(Y|X)}{\hat{P}_{m-1}(Y|X)\hat{P}_m(Y|X)}|dB_m \\
= \; & \frac{1}{2}\int \pi_m(B_m)|\frac{\hat{P}_{m-1}(Y|X,B_{m-1})\hat{P}_m(Y|X) - \hat{P}_{m-1}(Y|X,B_{m-1})\hat{P}_{m-1}(Y|X)}{\hat{P}_{m-1}(Y|X)\hat{P}_m(Y|X)} \\
& + \frac{\hat{P}_{m-1}(Y|X,B_{m-1})\hat{P}_{m-1}(Y|X) - \hat{P}_m(Y|X,B_m)\hat{P}_{m-1}(Y|X)}{\hat{P}_{m-1}(Y|X)\hat{P}_m(Y|X)}|dB_m \\
\le \; & \frac{1}{2}\int \pi_m(B_m)(\frac{\hat{P}_{m-1}(Y|X,B_{m-1})|\hat{P}_m(Y|X) - \hat{P}_{m-1}(Y|X)|}{\hat{P}_{m-1}(Y|X)\hat{P}_m(Y|X)} + \frac{|\hat{P}_m(Y|X,B_m) - \hat{P}_{m-1}(Y|X,B_{m-1})|}{\hat{P}_m(Y|X)})dB_m \\
= \; & \frac{1}{2}(\int \pi(B_{new}|B_{m-1})\hat{P}_{m-1}(B_{m-1}|X,Y)\frac{|\hat{P}_m(Y|X) - \hat{P}_{m-1}(Y|X)|}{\hat{P}_m(Y|X)}dB_m \\
& + \frac{E_{\pi_m}(|\hat{P}_m(Y|X,B_m) - \hat{P}_{m-1}(Y|X,B_{m-1})|}{\hat{P}_m(Y|X)}) \\
= \; & \frac{1}{2}(\frac{|\hat{P}_m(Y|X) - \hat{P}_{m-1}(Y|X)|}{\hat{P}_m(Y|X)} + \frac{E_{\pi_m}(|\hat{P}_m(Y|X,B_m) - \hat{P}_{m-1}(Y|X,B_{m-1})|)}{\hat{P}_m(Y|X)}) \\
\le \; & \frac{|\int \pi_{m-1}(B_{m-1})\hat{P}_{m-1}(Y|X,B_{m-1})dB_{m-1} - \int \pi_m(B_m)\hat{P}_m(Y|X,B_m)dB_m|}{2\hat{P}_m(Y|X)} + \frac{\delta_m}{2\hat{P}_m(Y|X)} \\
& \text{(The second term is by Proposition 3.2.4)} \\
= \; & \frac{|\int \pi_m(B_m)\hat{P}_{m-1}(Y|X,B_{m-1})dB_m - \int \pi_m(B_m)\hat{P}_m(Y|X,B_m)dB_m|}{2\hat{P}_m(Y|X)} + \frac{\delta_m}{2\hat{P}_m(Y|X)} \\
\le \; & \frac{\int \pi_m(B_m)|\hat{P}_{m-1}(Y|X,B_{m-1}) - \hat{P}_m(Y|X,B_m)|dB_m}{2\hat{P}_m(Y|X)} + \frac{\delta_m}{2\hat{P}_m(Y|X)} \\
= \; & \frac{E_{\pi_m}(|\hat{P}_m(Y|X,B_m) - \hat{P}_{m-1}(Y|X,B_{m-1})|)}{2\hat{P}_m(Y|X)} + \frac{\delta_m}{2\hat{P}_m(Y|X)} \\
\le \; & \frac{\delta_m}{2\hat{P}_m(Y|X)} + \frac{\delta_m}{2\hat{P}_m(Y|X)} \quad \text{(The first term is by Proposition 3.2.4)} \\
= \; & \frac{\delta_m}{\hat{P}_m(Y|X)} \\
\le \; & \frac{\delta_m}{LB} \quad (LB \text{ is defined in the proof of Theorem 3.2.1}) \\
= \; & C\delta_m
\end{aligned}
$$

Hence, $||f_{1,m}(B_m) - \hat{P}_m(B_m|X,Y)||_{TV} \le e_{m-1} + C\delta_m$.

Now, by the convexity of the norm, we can show the following.

$$\begin{aligned} & ||g_{1,m}(\eta) - P(\eta|X,Y)||_{TV} \\ = \; & ||g_{1,m}(\eta) - \hat{P}_m(\eta|X,Y) + \hat{P}_m(\eta|X,Y) - P(\eta|X,Y)||_{TV} \\ \leq \; & ||g_{1,m}(\eta) - \hat{P}_m(\eta|X,Y)||_{TV} + ||\hat{P}_m(\eta|X,Y) - P(\eta|X,Y)||_{TV} \\ \leq \; & c_{m-1} + C\delta_m + e_m, \quad \text{(by Theorem 3.2.1 and Corollary 3.2.1)} \end{aligned}$$

Since $e_m \to 0$ and $\delta_m \to 0$ as $m \to \infty$, we have $||g_{1,m}(\eta) - P(\eta|X,Y)||_{TV} \to 0$ as $m \to \infty$.

Part2: We need to show that for any point on the chain with $t > 1$ and $m^* \geq m$, we have $||g_{t,m^*}(\eta) - P(\eta|X,Y)||_{TV} \leq e_{m-1} + C\delta_m + e_m$.

By Proposition 6.52 in Robert and Casella (2004) [16], we have, $||f_{t,m}(B_m) - \hat{P}_m(B_m|X,Y)||_{TV} \leq ||f_{1,m}(B_m) - \hat{P}_m(B_m|X,Y)||_{TV}$. Using corollalry 3.2.1, we obtain $||g_{t,m}(\eta) - \hat{P}_m(\eta|X,Y)||_{TV} \leq ||f_{1,m}(B_m) - \hat{P}_m(B_m|X,Y)||_{TV}$.

Still by the convexity of norm, we have the following.

$$\begin{aligned} & ||g_{t,m}(\eta) - P(\eta|X,Y)||_{TV} \\ \leq \; & ||g_{t,m}(\eta) - \hat{P}_m(\eta|X,Y) + \hat{P}_m(\eta|X,Y) - P(\eta|X,Y)||_{TV} \\ \leq \; & ||g_{t,m}(\eta) - \hat{P}_m(\eta|X,Y)||_{TV} + ||\hat{P}_m(\eta|X,Y) - P(\eta|X,Y)||_{TV} \\ \leq \; & ||f_{1,m}(B_m) - \hat{P}_m(B_m|X,Y)||_{TV} + ||\hat{P}_m(\eta|X,Y) - P(\eta|X,Y)||_{TV} \\ \leq \; & e_{m-1} + C\delta_m + e_m \end{aligned}$$

By the same argument, for $m^* > m$, we can obtain,

$||g_{t,m^*}(\eta) - P(\eta|X,Y)||_{TV} \leq e_{m^*-1} + C\delta_{m^*} + e_{m^*}$.

By the monotonic property of e_m and δ_m, we have, $e_{m^*-1} + C\delta_{m^*} + e_{m^*} < e_{m-1} + C\delta_m + e_m$.

Therefore, combining these two parts, we show $||g_{t,m}(\eta) - P(\eta|X,Y)||_{TV} \leq e_{m-1} + C\delta_m + e_m \to 0$ as $m \to \infty$. □

3.3 Stationary distribution of the data-generating method

First, let us consider the method that accepts data Y' in a neighborhood of Y. According to the algorithm, we will reject any proposed point B'_m if Y' is not in the neighborhood of Y, $\mathcal{N}(Y,\epsilon)$, where

$\mathscr{N}(Y,\epsilon) = \{Y' | \rho(Y',Y) = \sqrt{\sum_{i=1}^{n}(y_i' - y_i)^2} < \epsilon\}$. Let $\hat{P}_m(\mathscr{N}(Y,\epsilon)|X, B_m)$ denote the probability that the generated data Y' is in $\mathscr{N}(Y,\epsilon)$. The posterior distribution of $B_m|X, \mathscr{N}(Y,\epsilon)$ is

$$\hat{P}_m(B_m|X, \mathscr{N}(Y,\epsilon)) = \frac{\pi_m(B_m)\hat{P}_m(\mathscr{N}(Y,\epsilon)|X, B_m)}{\hat{P}_m(\mathscr{N}(Y,\epsilon)|X)}$$

Proposition 3.3.1. *$\hat{P}_m(B_m|X, \mathscr{N}(Y,\epsilon))$ is the stationary distribution of the Markov chain constructed through the data generating method.*

Proof: We will verify the detailed balance condition to show the stationary distribution. Denote the probability from B_m to B_m' by $K(B_m \to B_m')$ and the proposal distribution by $q(B_m \to B_m')$. Assume $\frac{\pi_m(B_m')q(B_m' \to B_m)}{\pi_m(B_m)q(B_m \to B_m')} \leq 1$. We have

$$\begin{aligned}
& \hat{P}_m(B_m|X, \mathscr{N}(Y,\epsilon))K(B_m \to B_m') \\
= \; & \hat{P}_m(B_m|X, \mathscr{N}(Y,\epsilon))q(B_m \to B_m')\hat{P}_m(\mathscr{N}(Y,\epsilon)|X, B_m')\frac{\pi_m(B_m')q(B_m' \to B_m)}{\pi_m(B_m)q(Q_m \to B_m')} \\
= \; & \frac{\pi_m(B_m)\hat{P}_m(\mathscr{N}(Y,\epsilon)|X, B_m)}{\hat{P}_m(\mathscr{N}(Y,\epsilon)|X)}q(B_m \to B_m')\hat{P}_m(\mathscr{N}(Y,\epsilon)|X, B_m')\frac{\pi_m(B_m')q(B_m' \to B_m)}{\pi_m(B_m)q(B_m \to B_m')} \\
= \; & \frac{\pi_m(B_m')\hat{P}_m(\mathscr{N}(Y,\epsilon)|X, B_m')}{\hat{P}_m(\mathscr{N}(Y,\epsilon)|X)}q(B_m' \to B_m)\hat{P}_m(\mathscr{N}(Y,\epsilon)|X, B_m) \\
= \; & \hat{P}_m(B_m'|X, \mathscr{N}(Y,\epsilon))K(B_m' \to B_m).
\end{aligned}$$

The proof is analogous when $\frac{\pi_m(B_m')q(B_m' \to B_m)}{\pi_m(B_m)q(B_m \to B_m')} \geq 1$. □

If we consider the linear model with one covariate,

$$y_i = a + x_i\beta + \epsilon_i, \quad i = 1, 2, ..., n,$$

then we will consider the following neighborhood,

$$\mathscr{N}_1(d_1, d_2, e_1, e_2) = \{D' = (X, Y') | \rho(d_1, d_1') < e_1 \& \rho(d_2, d_2') < e_2\}.$$

If we assume $\pi(B_m)$ is the prior for $B_m = (a(\tau_1), a(\tau_2), ..., a(\tau_m), \beta(\tau_1), \beta(\tau_2), ..., \beta(\tau_m))$ and denote the probability that the generated data is in $\mathscr{N}_1(d_1, d_2, e_1, e_2)$ as $\hat{P}_m(\mathscr{N}_1(d_1, d_2, e_1, e_2)|B_m)$, then the posterior distribution of $B_m|\mathscr{N}_1(d_1, d_2, e_1, e_2)$ is

$$\hat{P}_m(B_m|\mathscr{N}_1(d_1, d_2, e_1, e_2)) = \frac{\pi_m(B_m)\hat{P}_m(\mathscr{N}_1(d_1, d_2, e_1, e_2)|B_m)}{\hat{P}_m(\mathscr{N}_1(d_1, d_2, e_1, e_2))},$$

where

$$\hat{P}_m(\mathscr{N}_1(d_1, d_2, e_1, e_2)) = \int \pi_m(B_m)\hat{P}_m(\mathscr{N}_1(d_1, d_2, e_1, e_2)|B_m)dB_m.$$

Proposition 3.3.2. *$\hat{P}_m(B_m|\mathscr{N}_1(d_1, d_2, e_1, e_2))$ is the stationary distribution of the Markov chain constructed through the data generating method.*

The proof is similar to the previous one. □

Using similar arguments as we presented in Section 3.2, we can show the following theorems.

Theorem 3.3.1. *$||\hat{P}_m(B_m|X, \mathscr{N}(Y, \epsilon)) - P(B_m|X, Y)||_{TV} \to 0$ as $m \to \infty$ and $\epsilon \to 0$.*

Theorem 3.3.2. *$||g_{t,m}(\eta) - P(B_m|X, Y)||_{TV} \to 0$ as $m \to \infty$ and $\epsilon \to 0$.*

Theorem 3.3.3. *$||\hat{P}_m(B_m|\mathscr{N}_1(d_1, d_2, e_1, e_2)) - P(B_m|X, Y)||_{TV} \to 0$ as $m \to \infty$, $e_1 \to 0$ and $e_2 \to 0$.*

Theorem 3.3.4. *$||g_{t,m}(\eta) - P(B_m|X, Y)||_{TV} \to 0$ as $m \to \infty$, $e_1 \to 0$ and $e_2 \to 0$.*

Chapter 4

Simulation Studies and a Real Data Example

In this chapter, we will check the performance of the algorithms proposed in Chapter 2. We will also compare our methods with some other methods including Regression of Quantiles (RQ) and Markov Chain Marginal bootstrap (MCMB). For all the simulation and real data studies in this chapter, we will focus on the inferences on the first quartile, the median, and the third quartile.

4.1 Performance of proposed methods

In this section, we will check the performance of the two proposed algorithms: the linearly interpolated density algorithm and the data-generating algorithm. We will compare the posterior estimates of the parameters with the true value and the RQ estimates. For all the simulations in this chapter, we always center the covariates before running our proposed algorithms and the RQ and MCMB algorithms. When calculating the estimates, we transform the parameters back to original ones.

4.1.1 Performance of the linearly interpolated density method (LID)

Consider the following two models:

$$y_i = a + bx_i + \epsilon_i, \quad i = 1, 2, ..., n, \tag{4.1}$$

and

$$y_i = a + bx_i + \epsilon_i x_i, \quad i = 1, 2, ..., n, \tag{4.2}$$

where ϵ_i's are iid from $N(0, 1)$, $i = 1, 2, ..., n$. The quantile model associated with these models is

$$Q_\tau(y_i|x_i) = a(\tau) + b(\tau)x_i, \quad i = 1, 2, ..., n. \tag{4.3}$$

It is not difficult to see that $a^{true}(\tau) = a + \Phi^{-1}(\tau)$ and $b^{true}(\tau) = b$ for Model (4.1), where $\Phi^{-1}(\tau)$ denotes the τ-th quantile of the standard normal distribution. For Model (4.2), we have $a^{true}(\tau) = a$ and $b^{true}(\tau) = b + \Phi^{-1}(\tau)$.

In the simulations, we set $a = 5$, $b = 5$, and generated $n = 200$ obervations from Models (4.1) and (4.2). The covariate x_i was generated from Uniform$(1, 5)$. To get the posterior distribution of $(a(\tau_j), b(\tau_j))|X, Y$, $j = 1, 2, ..., m$, we used $m = 7, 11$, and 15 quantiles and put truncated normal priors on $a(\tau_1), a(\tau_2), ..., a(\tau_m)$ and $b(\tau_1), b(\tau_2), ..., b(\tau_m)$. The truncated normal priors are $N(0, \Sigma_a)$ and $N(0, \Sigma_b)$ with the order constraint that $a(\tau_1)+b(\tau_1)x_i < a(\tau_2)+b(\tau_2)x_i < ... < a(\tau_m)+b(\tau_m)x_i$, $i = 1, 2, ..., n$. The covariance matrices $\Sigma_a = \Sigma_b = diag(1/100, ..., 1/100)$ are both $m \times m$ diagonal matrices. We let $\tau_j = j/(m + 1)$, $j = 1, 2, ..., m$.

We also provided the posterior estimates of the parameters based on the true densities (TD), where we used the same normal prior as that for the LID method but ignored the order constraint to simplify the computation. For Model (4.1), we used Gibbs sampler to draw samples from $P_{true}((a(\tau_j), b(\tau_j))|X, Y)$, which denotes the posterior distribution of the parameters based on the true densities, through the following conditional distributions:

$$a(\tau_j)|X, Y, b(\tau_j) \sim N(\frac{\sum_{i=1}^n y_i + \Phi^{-1}(\tau_j) - b(\tau_j)x_i}{n + 1/100}, \frac{1}{n + 1/100}), \tag{4.4}$$

and

$$b(\tau_j)|X, Y, a(\tau_j) \sim N(\frac{\sum_{i=1}^n (y_i - a(\tau_j) + \Phi^{-1}(\tau_j))x_i}{1/100 + \sum_{i=1}^n x_i^2}, \frac{1}{1/100 + \sum_{i=1}^n x_i^2}). \tag{4.5}$$

For Model (4.2), we used the following conditional distributions:

$$a(\tau_j)|X, Y, b(\tau_j) \sim N(\frac{\sum_{i=1}^n \frac{y_i - (b(\tau_j) - \Phi^{-1}(\tau_j))x_i}{x_i^2}}{1/100 + \sum_{i=1}^n \frac{1}{x_i^2}}, \frac{1}{1/100 + \sum_{i=1}^n \frac{1}{x_i^2}}), \tag{4.6}$$

and

$$b(\tau_j)|X, Y, a(\tau_j) \sim N(\frac{\sum_{i=1}^n \frac{y_i + \Phi^{-1}(\tau_j)x_i - a(\tau_j)}{x_i}}{n + 1/100}, \frac{1}{n + 1/100}). \tag{4.7}$$

For the model with iid errors, the RQ standard errors are calculated by the "iid" method of the "quantreg" package in R, and for models with non-iid errors, the RQ standard errors are calculated by the "nid" method of the "quantreg" package. The LID estimates are based on 200,000 samples (we ran the Markov chain for 400,000 steps and used the first half as burn-in). The TD estimates are based on 10,000 samples.

From the results in Table 4.1, we can see that the RQ estimates and the standard errors are very

Table 4.1: Comparison of the LID method with the RQ method for model (4.1)

Methods	$a(0.25)$	$b(0.25)$	$a(0.5)$	$b(0.5)$	$a(0.75)$	$b(0.75)$
RQ	4.47 (0.23)	4.95 (0.07)	4.94 (0.22)	5.03 (0.07)	5.74 (0.19)	4.93 (0.06)
LID $m = 7$	4.15 (0.15)	5.03 (0.05)	4.53 (0.20)	5.14 (0.06)	5.57 (0.18)	5.01 (0.06)
LID $m = 11$	4.33 (0.17)	4.99 (0.05)	4.80 (0.25)	5.06 (0.07)	5.71 (0.23)	4.97 (0.07)
LID $m = 15$	4.29 (0.20)	4.98 (0.07)	4.80 (0.29)	5.05 (0.09)	5.72 (0.23)	4.96 (0.08)
TD	4.43 (0.18)	4.97 (0.06)	5.10 (0.18)	4.97 (0.06)	5.78 (0.18)	4.97 (0.06)
True value	4.33	5	5	5	5.67	5

Note: For the LID and TD estimates, the values in each cell are the posterior mean and standard deviation. For the RQ estimates, the values in each cell are the estimate and standard error.

Table 4.2: Comparison of the LID method with the RQ method for model (4.2)

Methods	$a(0.25)$	$b(0.25)$	$a(0.5)$	$b(0.5)$	$a(0.75)$	$b(0.75)$
RQ	5.48 (0.37)	4.15 (0.21)	4.74 (0.43)	5.11 (0.20)	5.30 (0.49)	5.41 (0.23)
LID $m = 7$	5.02 (0.37)	4.28 (0.19)	4.89 (0.53)	4.99 (0.25)	5.32 (0.37)	5.41 (0.18)
LID $m = 11$	5.29 (0.31)	4.22 (0.11)	5.28 (0.40)	4.90 (0.20)	5.76 (0.45)	5.26 (0.18)
LID $m = 15$	5.27 (0.28)	4.20 (0.11)	5.50 (0.33)	4.75 (0.15)	5.96 (0.41)	5.16 (0.14)
TD	5.28 (0.34)	4.21 (0.16)	5.28 (0.34)	4.89 (0.16)	5.28 (0.34)	5.56 (0.16)
True value	5	4.33	5	5	5	5.67

Note: For the LID and TD estimates, the values in each cell are the posterior mean and standard deviation. For the RQ estimates, the values in each cell are the estimate and standard error.

close to the estimates and standard deviations based on the true densities, though the estimates for the median are a little different. The posterior mean of the LID method is closer to the TD posterior mean when m is increased from 7 to 11, and the posterior means are similar for $m = 11$ and $m = 15$. From the results in Table 4.2, we can see that the RQ estimates are still close to the estimates based on the true densities, but the standard errors are a little larger than the TD standard deviation. The posterior means of the LID method for are closer to the TD posterior means when m increases from 7 to 11. It is a little strange that the posterior standard deviations in Table reft1 increase when m increases and in Table reft2 the posterior means of the LID method with $m = 11$ are closer to the TD posterior means than the posterior means of the LID method with $m = 15$. This may be due to the following reasons. First, when m is large the Markov chain may need a longer time to converge. Second, the prior distribution of the TD method is not the same as the LID method, which may result in a different posterior distribution from the limiting distribution of the LID method. Notice that the standard deviations of the LID posterior distribution when $m = 11$ and 15 are smaller than the standard errors of the RQ estimate for Model (4.2). This may suggest that the LID estimates sometimes are more efficient than the RQ estimates.

4.1.2 Performance of the data-generating method (DG)

In Chapter 2, we introduced two scenarios for the data-generating method under two different cases, the univariate case and the regression case. Here we will check the performances of the method for both cases.

First, let us consider the univariate case with the following model,

$$z_i \sim N(\theta, \sigma^2), \quad i = 1, 2, ..., n. \tag{4.8}$$

The corresponding quantile model is

$$Q_\tau(z_i) = a(\tau). \tag{4.9}$$

In the simulation, we set $\theta = 10$, $\sigma = 4$, and generated $n = 200$ observations from Model (4.8). To get the posterior distribution of $a(\tau_j)|Z$, $j = 1, 2, ..., m$, we used $m = 7$, 11, and 15 quantiles and put truncated normal priors on $a(\tau_1), a(\tau_2), ..., a(\tau_m)$. The truncated normal distribution is $N(0, \Sigma_a)$ with the order constraint that $a(\tau_1) < a(\tau_2) < ... < a(\tau_m)$. The covariance matrix $\Sigma_a = diag(1/100, ..., 1/100)$ is $m \times m$ and diagonal. Let $\tau_j = j/(m+1)$, $j = 1, 2, ..., m$. For this algorithm, we need to specify a tolerance quantity ϵ which defines the neighborhood of the observed data. In this example, we set $\epsilon = 0.8$ and 0.6. The LID estimates are based on 50,000 samples (we ran the Markov chain for 100,000 steps and used the first half as burn-in). The TD posterior mean and standard deviation are calculated directly from the the following distribution:

$$a(\tau)|Z \sim N(\frac{\sum_{i=1}^{n} \frac{z_i + \Phi^{-1}(\tau)}{\sigma^2}}{n/\sigma^2 + 1/100}, \frac{1}{n/\sigma^2 + 1/100}). \tag{4.10}$$

From the results in Table 4.3, we can see that the DG estimates give a bigger standard deviation than that of RQ estimates and the estimates based on the true density. We can also see that the standard deviations of the DG method are smaller when we increase number of quantiles or decrease the tolerance quantity. This is consistent with the theoretical results.

Now, let us consider the regression model 4.1 and 4.2, and this time we apply the data-generating method. As introduced in Chapter 2, in this case we need to calculate the d_1 and d_2 distances and set corresponding tolerance quantities ϵ_1 and ϵ_2 for them. In the simulations, we used the same settings as the ones in Section 4.1.1 and set $\epsilon_1 = 0.2\sqrt{m}$ and $\epsilon_2 = 0.1\sqrt{m}$ or $\epsilon_1 = 0.1\sqrt{m}$ and $\epsilon_2 = 0.05\sqrt{m}$.

Table 4.3: Comparison of the DG method with the RQ method for Model (4.8)

Methods	$a(0.25)$	$a(0.5)$	$a(0.75)$
RQ	7.45 (0.33)	10.28 (0.33)	12.88 (0.37)
DG m=7 $\epsilon = 0.8$	7.20 (0.83)	10.09 (0.64)	12.96 (0.95)
DG m=7 $\epsilon = 0.6$	7.23 (0.71)	10.08 (0.57)	12.84 (0.83)
DG m=11 $\epsilon = 0.8$	7.23 (0.80)	10.08 (0.61)	13.02 (0.86)
DG m=11 $\epsilon = 0.6$	7.20 (0.62)	10.09 (0.48)	12.91 (0.75)
DG m=15 $\epsilon = 0.8$	7.26 (0.72)	10.12 (0.63)	13.08 (0.80)
DG m=15 $\epsilon = 0.6$	7.18 (0.59)	10.10 (0.48)	12.88 (0.72)
DG m=19 $\epsilon = 0.8$	7.17 (0.65)	10.16 (0.58)	13.22 (0.78)
DG m=19 $\epsilon = 0.6$	7.13 (0.67)	10.12 (0.52)	13.00 (0.72)
DG m=23 $\epsilon = 0.8$	7.08 (0.65)	10.04 (0.66)	13.18 (0.77)
DG m=23 $\epsilon = 0.6$	7.04 (0.59)	10.13 (0.49)	13.02 (0.68)
TD	7.50 (0.28)	10.20 (0.28)	12.90 (0.28)
True value	7.30	10	12.70

Table 4.4: Comparison of the DG method with the RQ method for Model (4.1)

Methods	$a(0.25)$	$b(0.25)$	$a(0.5)$	$b(0.5)$	$a(0.75)$	$b(0.75)$
RQ	4.47 (0.23)	4.95 (0.07)	4.94 (0.22)	5.03 (0.07)	5.74 (0.19)	4.93 (0.06)
DG $m = 7*$	3.99 (1.15)	4.98 (0.33)	5.01 (0.44)	4.98 (0.14)	5.82 (0.83)	4.99 (0.25)
DG $m = 7**$	4.44 (0.31)	4.97 (0.10)	4.98 (0.29)	5.00 (0.09)	5.59 (0.35)	5.00 (0.11)
DG $m = 11*$	4.04 (0.89)	4.90 (0.22)	4.87 (0.45)	4.98 (0.12)	5.46 (0.60)	5.05 (0.17)
DG $m = 11**$	4.39 (0.30)	4.96 (0.10)	4.95 (0.26)	5.00 (0.08)	5.59 (0.36)	5.00 (0.13)
DG $m = 15*$	4.69 (0.45)	4.81 (0.16)	5.41 (0.43)	4.85 (0.14)	6.09 (0.53)	4.84 (0.16)
DG $m = 15**$	4.45 (0.20)	4.94 (0.06)	4.83 (0.11)	5.03 (0.04)	5.59 (0.30)	5.00 (0.08)
TD	4.43 (0.18)	4.97 (0.06)	5.10 (0.18)	4.97 (0.06)	5.78 (0.18)	4.97 (0.06)
True value	4.33	5	5	5	5.67	5

NOTE: Here $*$ denotes that $\epsilon_1 = 0.2\sqrt{m}$ and $\epsilon_2 = 0.1\sqrt{m}$, and $**$ denote $\epsilon_1 = 0.1\sqrt{m}$ and $\epsilon_2 = 0.05\sqrt{m}$.

Table 4.5: Comparison of the DG method with the RQ method for Model (4.2)

Methods	$a(0.25)$	$b(0.25)$	$a(0.5)$	$b(0.5)$	$a(0.75)$	$b(0.75)$
RQ	5.48 (0.37)	4.15 (0.21)	4.74 (0.43)	5.11 (0.20)	5.30 (0.49)	5.41 (0.23)
DG $m = 7*$	3.82 (1.86)	4.50 (0.59)	5.18 (0.83)	4.85 (0.37)	6.19 (1.38)	5.18 (0.46)
DG $m = 7**$	4.82 (1.08)	4.33 (0.40)	5.10 (0.52)	4.89 (0.25)	5.37 (0.77)	5.38 (0.31)
DG $m = 11*$	3.59 (1.81)	4.55 (0.50)	5.19 (0.84)	4.79 (0.32)	6.32 (1.48)	5.16 (0.47)
DG $m = 11**$	4.96 (0.50)	4.37 (0.30)	5.06 (0.37)	4.96 (0.16)	5.72 (0.71)	5.31 (0.23)
DG $m = 15*$	4.06 (1.29)	4.16 (0.47)	5.56 (0.47)	4.57 (0.22)	6.95 (0.93)	4.89 (0.29)
DG $m = 15**$	4.91 (0.62)	4.37 (0.26)	5.17 (0.53)	4.90 (0.22)	5.92 (0.55)	5.19 (0.18)
TD	5.28 (0.34)	4.21 (0.16)	5.28 (0.34)	4.89 (0.16)	5.28 (0.34)	5.56 (0.16)
True value	5	4.33	5	5	5	5.67

NOTE: Here $*$ denotes that $\epsilon_1 = 0.2\sqrt{m}$ and $\epsilon_2 = 0.1\sqrt{m}$, and $**$ denote $\epsilon_1 = 0.1\sqrt{m}$ and $\epsilon_2 = 0.05\sqrt{m}$.

From the results in Tables 4.4 and 4.5, we can see that the DG standard deviations tend to be smaller with larger m and smaller ϵ_1 and ϵ_2. Also, the DG posterior means tend to be closer to the TD posterior means as ϵ_1 and ϵ_2 decrease. These are all consistent with the theoretical results.

4.2 Comparison of several methods under models with multiple covariates

In this section we will compare the performance of the LID, RQ, and MCMB estimates. Let us consider the following model,

$$y_i = a + bx_{1,i} + cx_{2,i} + (1 + x_{1,i} + x_{2,i})\epsilon_i, \quad i = 1, 2, ..., n, \tag{4.11}$$

where $\epsilon_i \sim N(0, 1)$. The corresponding quantile model is

$$Q_\tau(y_i|x_{1,i}, x_{2,i}) = a(\tau) + b(\tau)x_{1,i} + c(\tau)x_{2,i}. \tag{4.12}$$

It is not difficult to see that the true values of $a(\tau)$, $b(\tau)$, and $c(\tau)$ are $a + \Phi^{-1}(\tau)$, $b + \Phi^{-1}(\tau)$, and $c + \Phi^{-1}(\tau)$.

In the simulations, we set $a = 5$, $b = 1$, $c = 1$ and generated $n = 200$ obervations. The covariates $x_{1,i}$ and $x_{2,i}$ were generated from lognormal$(0, 1)$ and $N(0, 1)$, respectively. For the LID algorithm, we used $m = 11$ quantiles and put the truncated normal prior on the parameters similar as those introduced in Section 4.1.1 with the constraint changed to $a(\tau_k) + b(\tau_k)x_{1,i} + c(\tau_k)x_{2,i} < a(\tau_l) + b(\tau_l)x_{1,i} + c(\tau_l)x_{2,i}$, where $1 \le k < l \le m$ and $i = 1, 2, ..., n$. For the TD method, we used the same normal prior as for the LID method but ignored the order constraint for the prior setting to simplify the computation. For Model (4.11) The TD samples were drawn through the following conditional distributions:

$$a(\tau_j)|X, Y, b(\tau_j), c(\tau_j) \sim N\left(\frac{\sum_{i=1}^n \frac{y_i + \Phi^{-1}(\tau_j)(1+x_{1,i}+x_{2,i}) - b(\tau_j)x_{1,i} - c(\tau_j)x_{2,i}}{(1+x_{1,i}+x_{2,i})^2}}{1/100 + \sum_{i=1}^n \frac{1}{(1+x_{1,i}+x_{2,i})^2}}, \frac{1}{1/100 + \sum_{i=1}^n \frac{1}{(1+x_{1,i}+x_{2,i})^2}}\right), \tag{4.13}$$

$$b(\tau_j)|X, Y, a(\tau_j), c(\tau_j) \sim N\left(\frac{\sum_{i=1}^n \frac{x_{1,i}(y_i + \Phi^{-1}(\tau_j)(1+x_{1,i}+x_{2,i}) - c(\tau_j)x_{2,i})}{(1+x_{1,i}+x_{2,i})^2}}{1/100 + \sum_{i=1}^n \frac{x_{1,i}^2}{(1+x_{1,i}+x_{2,i})^2}}, \frac{1}{1/100 + \sum_{i=1}^n \frac{x_{1,i}^2}{(1+x_{1,i}+x_{2,i})^2}}\right), \tag{4.14}$$

Table 4.6: Simulation results for Model (4.11)

Methods	$a(0.25)$	$b(0.25)$	$c(0.25)$	$a(0.5)$	$b(0.5)$	$c(0.5)$	$a(0.75)$	$b(0.75)$	$c(0.75)$
RQ	4.35 (0.41)	0.16 (0.56)	0.38 (0.15)	5.50 (0.27)	1.48 (0.34)	1.38 (0.09)	5.50 (0.27)	1.48 (0.34)	1.38 (0.09)
MCMB	4.22 (0.24)	0.34 (0.40)	0.40 (0.10)	4.93 (0.22)	0.91 (0.29)	0.90 (0.12)	5.61 (0.22)	1.48 (0.26)	1.38 (0.11)
LID	4.81 (0.30)	-0.21 (0.27)	0.33 (0.47)	5.53 (0.27)	0.51 (0.15)	0.87 (0.54)	6.37 (0.35)	1.33 (0.20)	1.41 (0.58)
True density	4.62 (0.25)	-0.01 (0.21)	0.45 (0.35)	5.29 (0.25)	0.66 (0.21)	1.12 (0.35)	5.96 (0.25)	1.33 (0.21)	1.79 (0.35)
True value	4.33	0.33	0.33	5	1	1	5.57	1.67	1.67

and

$$c(\tau_j)|X,Y,a(\tau_j),b(\tau_j) \sim N(\frac{\sum_{i=1}^{n} \frac{x_{2,i}(y_i+\Phi^{-1}(\tau_j)(1+x_{1,i}+x_{2,i})-b(\tau_j)x_{1,i})}{(1+x_{1,i}+x_{2,i})^2}}{1/100+\sum_{i=1}^{n} \frac{x_{2,i}^2}{(1+x_{1,i}+x_{2,i})^2}}, \frac{1}{1/100+\sum_{i=1}^{n} \frac{x_{2,i}^2}{(1+x_{1,i}+x_{2,i})^2}}), \tag{4.15}$$

For the LID method, the estimates are based on 500,000 samples, which are the second half of the 1,000,000 samples generated. For the MCMB method, we used 200 bootstrap samples and set the length of the MCMB sequence equal to 100. For the TD method, the estimates are based on 10,000 samples.

From the results in Table 4.6, we can see that the RQ estimates and the MCMB estimates are similar, but the MCMB estimates tend to give smaller standard errors. Compared with the estimates based on the true densities, the RQ and MCMB estimates underestimate $a(\tau)$ and overestimate $b(\tau)$, whereas the LID estimates performs in the opposite direction. All these three algorithm underestimate $c(\tau)$. The overall view is that the RQ and MCMB estimates are closer to the true value while the LID estimates are closer to the estimates based on the true densities.

Let us consider the following model with iid errors.

$$y_i = a + bx_{1,i} + cx_{2,i} + \epsilon_i, \quad i = 1, 2, ..., n, \tag{4.16}$$

The corresponding quantile model is the same as (4.12). We can see that the true values for $a(\tau)$, $b(\tau)$ and $c(\tau)$ are $a + \Phi^{-1}(\tau)$, b and c. For Model (4.16), the TD samples were drawn through the

Table 4.7: Simulation results for Model (4.16)

Methods	$a(0.25)$	$b(0.25)$	$c(0.25)$	$a(0.5)$	$b(0.5)$	$c(0.5)$	$a(0.75)$	$b(0.75)$	$c(0.75)$
RQ	4.36	0.91	1.09	5.13	0.95	0.86	5.67	1.01	0.89
	(0.15)	(0.037)	(0.20)	(0.11)	(0.027)	(0.14)	(0.17)	(0.040)	(0.21)
MCMB	4.46	0.90	1.03	5.11	0.94	0.92	5.79	0.96	0.95
	(0.21)	(0.090)	(0.22)	(0.12)	(0.041)	(0.16)	(0.17)	(0.082)	(0.25)
LID	4.52	0.82	0.94	5.17	0.93	0.87	5.71	1.06	1.00
	(0.11)	(0.079)	(0.19)	(0.072)	(0.035)	(0.13)	(0.13)	(0.079)	(0.22)
True density	4.36	0.98	1.01	5.03	0.98	1.01	5.70	0.98	1.01
	(0.11)	(0.027)	(0.14)	(0.11)	(0.027)	(0.14)	(0.11)	(0.027)	(0.14)
True value	4.33	1	1	5	1	1	5.67	1	1

following conditional distributions:

$$a(\tau_j)|X,Y,b(\tau_j),c(\tau_j) \sim N(\frac{\sum_{i=1}^n y_i + \Phi^{-1}(\tau_j)(1+x_{1,i}+x_{2,i}) - b(\tau_j)x_{1,i} - c(\tau_j)x_{2,i}}{1/100+n}, \frac{1}{1/100+n}), \tag{4.17}$$

$$b(\tau_j)|X,Y,a(\tau_j),c(\tau_j) \sim N(\frac{\sum_{i=1}^n x_{1,i}(y_i + \Phi^{-1}(\tau_j)(1+x_{1,i}+x_{2,i}) - c(\tau_j)x_{2,i})}{1/100+\sum_{i=1}^n x_{1,i}^2}, \frac{1}{1/100+\sum_{i=1}^n x_{1,i}^2}), \tag{4.18}$$

and

$$c(\tau_j)|X,Y,a(\tau_j),b(\tau_j) \sim N(\frac{\sum_{i=1}^n x_{2,i}(y_i + \Phi^{-1}(\tau_j)(1+x_{1,i}+x_{2,i}) - b(\tau_j)x_{1,i})}{1/100+\sum_{i=1}^n x_{2,i}^2}, \frac{1}{1/100+\sum_{i=1}^n x_{2,i}^2}), \tag{4.19}$$

In the simulations, we used the same settings as that for Model (4.11) except that we generated $x_{i,1}$'s from $lognormal(0,1)$ and $x_{i,2}$'s from $Bernoulli(0.5)$.

From the results in Tables 4.7, we can see that all these three methods performs similarly, especially for the median. All these three methods have smaller standard errors for the median and larger standard errors for the first and third quartiles. However, there are still some minor differences. Unlike the results for Model (4.11), the MCMB algorithm tends to give larger standard errors than the RQ estimates, while the LID standard deviations are in-between except the standard deviations for $a(\tau)$'s, which are always smaller than the other two.

4.3 Real data study

In this section, our study is based on the June 1997 Detailed Natality Data, which is published by the National Center for Health Statistics. It is also analyzed in Koenker (2005) [10]. The following background information is quoted from Pg. 20 of Koenker (2005) [10].

Table 4.8: Results for the birth weight data with $\tau = 0.25$

Methods	*Intercept*	*mom.age*	*smoke*	*m.wtgain*
RQ	2.35 (0.042)*	0.013* (0.0043)	-0.11 (0.084)	0.011 (0.0019)*
MCMB	2.42 (0.034)*	0.012* (0.0055)	-0.15 (0.065)*	0.010 (0.0020)*
LID	2.26 (0.034)*	0.016* (0.0035)	-0.16 (0.069)*	0.011 (0.0016)*

NOTE: The symbol $*$ denotes statistical significance.

> "[T]he sample is restricted to singleton births, with mothers recorded as either black or white, between the age of 18 and 45, resident in the United States. Observations with missing data for any of the variables described in the following were also dropped from the analysis. This process yielded a sample of 198,377 babies. Education of the mother is divided into four categories: less than high school, high school, some college, and college graduate." "The prenatal medical care of the mother is also divided into four categories: those with no prenatal visit, those whose first prenatal visit was the first trimester of the pregnancy, those with the first visit in the second trimester, and those with the first visit in the last trimester."

With the infant birth weight being the response variable, we are interested in the following explanatory variables: *mom.age*, *smoke*, and *m.wtgain*, where the variable *mom.age* denotes the age of the mother, the variable *smoke* is a dummy variable indicates whether the mother smokes during pregnancy, and the variable *m.wtgain* denotes mother's weight gain during pregnancy. The quantile model is

$$Q_\tau(y_i|x_i) = a(\tau) + b(\tau)x_{i,1} + c(\tau)x_{i,2} + d(\tau)x_{i,3}, \quad i = 1, 2, ..., n, \tag{4.20}$$

where y_i denotes the infant birth weight for the i-th observation, the value $x_{i,1}$ is the i-th observation of *mom.age*, the value $x_{i,2}$ is the i-th observation of *smoke*, and the value $x_{i,3}$ is the i-th observation of *m.wtgain*. Because the original data set is quite large and our algorithm is quite computationally intensive, we will analyze a portion of the original data set, which are the first 1000 observations. We centered the covariates before we implement the algorithms. For the two continuous variables, *mom.age* and *m.wtgain*, we subtract the mean from them, while for the dummy variable *smoke* we subtract 0.5 from it. The reason to center the covariates is that this helps reducing the standard error of the intercept. For the LID method, the estimates are based on 5,000,000 samples, which are the second half of the 10,000,000 samples generated. For the MCMB algorithm, we used 200

Table 4.9: Results for the birth weight data with $\tau = 0.5$

Methods	*Intercept*	*mom.age*	*smoke*	*m.wtgain*
RQ	2.79 (0.017)*	0.010 (0.0027)*	-0.17 (0.033)*	0.010 (0.0012)*
MCMB	2.78 (0.019)*	0.010 (0.0036)*	-0.18 (0.037)*	0.010 (0.0016)*
LID	2.61 (0.011)*	0.015 (0.0018)*	-0.15 (0.027)*	0.012 (0.0009)*

NOTE: The symbol $*$ denotes statistical significance.

Table 4.10: Results for the birth weight data with $\tau = 0.75$

Methods	*Intercept*	*mom.age*	*smoke*	*m.wtgain*
RQ	3.15 (0.035)*	0.0050 (0.0035)	-0.11 (0.068)	0.0123 (0.0017)*
MCMB	3.14 (0.037)*	0.0056 (0.0039)	-0.088 (0.068)	0.0124 (0.0017)*
LID	3.05 (0.030)*	0.011 (0.0035)*	-0.078 (0.064)	0.013 (0.0015)*

Note: The symbol $*$ denotes statistical significance.

bootstrap samples and set the length of the MCMB sequence equal to 100.

From the results in Tables 4.8, 4.9, and 4.10, we notice the following things. First, these three methods agree on the effects of the covariates. Mother's age and the weight gain during pregnancy have positive effects, while smoking during pregnancy have negative effects. Second, we used a simple Z-test ($|estimate/se|$ compared with 2) to decide the significance and find that these methods agree on the significance of almost all the parameters. For the relative low birth weight and the normal birth weight, all parameters seem significant, except for the RQ estimate of *smoke* for the relative low birth weight. For the relative high birth weight, only *m.wtgain* seems to be significant, although the LID estimate suggests that *mom.age* is also significant. Third, the standard deviations by the LID algorithm are almost always the smallest, which again suggest that the LID estimates sometimes are more efficient.

4.4 Some conclusions

From the results based on the simulation data and the real data. We find the followings. First, both the LID and DG algorithms can give comparable results with those from RQ or MCMB. Second, the numerical results show that the DG algorithm will have better performance with large m and small tolerance quantities, which is consistent with the theoretical results. Third, the numerical results suggest that with larger m the LID algorithm may need a longer time to converge. Last, the numerical results also suggest that the LID algorithm sometimes produces more efficient estimates

than the RQ estimates or the MCMB estimates.

Chapter 5

More Numerical Explorations for LID

In this chapter, we compare our proposed method with other methods for several different models. We compare the mean squared error, level, and power based on simulation studies. We also compare the estimation accuracy based on a real data example.

5.1 Comparison of mean squared errors

In this section, we compare the mean squared errors (MSEs) among different methods for several different models.

5.1.1 The MSE for single quantiles

Let us consider the following non-i.i.d-error model:

$$y_i = a + bx_i + (1 + x_i)\epsilon_i, \quad i = 1, 2, ..., n, \tag{5.1}$$

where ϵ_i's are i.i.d. from N(0,1)
In the simulation, we choose $a = 5$ and $b = 1$. The covariate x_i is generated from lognormal(0,1). We compared the MSEs of different methods based on 400 data sets generated from Model (5.1). We also considered the following parametric model for the MLE calculation and the Bayesian method that uses the true underlying density.

$$y_i = a + bx_i + (\gamma_1 + \gamma_2 x_i)\epsilon_i, \quad i = 1, 2, ..., n. \tag{5.2}$$

We used the following abbreviation for different methods. RQ denotes the estimates by the "quantreg" package in R. EWRQ denotes the weighted RQ with estimated weights [10], and OWRQ denotes the

weighted RQ with optimal weights [10]. LID* denotes the method using modified likelihood estimates, i.e., using $\frac{\tau_{i+1}-\tau_{i-1}}{q_{i+1}-q_{i-1}}$ as the estimates of the densities instead of $\frac{\tau_{i+1}-\tau_i}{q_{i+1}-q_i}$. LID* nc denotes the modified method LID* applying to the original data, which is not centered. TD denotes the Bayesian method using the true densities based on Model 5.1. TD (5.2) denotes the Bayesian method using the true densities based on Model (5.2). MLE denotes the maximum likelihood estimates based on Model (5.1), and MLE (5.2) denotes the maximum likelihood estimates based on Model (5.2). TQ (5.2) denotes the Bayesian method using the linear interpolated densities with normal quantiles based on Model (5.2).

First consider the case with data size $n = 100$ for each of the 400 data sets generated from Model (5.1). We used $m = 15$ quantiles for LID based methods. For all the Bayesian methods, we constructed a Markov chain with length 1,000,000, used the first half as the burn-in period, and took every 1,000-th samples.

The results are given in Tables 5.1 and 5.2. We can see that LID behaved similarly to two weighted RQ methods, and all of them are better than RQ. The Bayesian methods based on the true densities performed very similar as MLE, which is not surprising, because we used a very flat prior $\mathrm{N}(0, 100)$ for each parameter. TQ (4.2) seems to be the limiting case of LID, that is, the best performance that LID can achieve as $m \to \infty$.

Table 5.1: Comparison of the MSEs of the median from different methods ($n = 100$ and $m = 15$).

Methods	MSE of $a(0.5)$	SE of MSE	MSE of $b(0.5)$	SE of MSE
RQ	0.18	0.013	0.18	0.013
EWRQ	0.11	0.008	0.11	0.007
OWRQ	0.11	0.008	0.10	0.007
LID	0.11	0.008	0.11	0.008
LID*	0.11	0.008	0.10	0.007
LID* nc	0.10	0.008	0.10	0.006
TD	0.07	0.005	0.06	0.005
MLE	0.07	0.005	0.07	0.005
TD (5.2)	0.07	0.005	0.07	0.005
MLE (5.2)	0.07	0.005	0.07	0.005
TQ (5.2)	0.09	0.006	0.08	0.006

Table 5.2: Comparison of the MSEs of the third quartile from different methods ($n = 100$ and $m = 15$).

Methods	MSE of $a(0.75)$	SE of MSE	MSE of $b(0.75)$	SE of MSE
RQ	0.22	0.015	0.21	0.013
EWRQ	0.15	0.010	0.14	0.009
OWRQ	0.14	0.010	0.13	0.009
LID	0.16	0.012	0.12	0.009
LID*	0.16	0.012	0.12	0.009
LID* nc	0.14	0.011	0.14	0.010
TD	0.07	0.005	0.06	0.005
MLE	0.07	0.005	0.07	0.005
TD (5.2)	0.09	0.006	0.08	0.006
MLE (5.2)	0.09	0.006	0.08	0.006
TQ (5.2)	0.12	0.008	0.12	0.011

We did more simulations with different values of n and m. We changed the size for each data set from $n = 100$ to 200. We checked the performance of LID for $m = 15, 19$, and 23. The results are in Tables 5.3 and 5.4. From the results, we can see that with $n = 200$ and different values of m, the MSE of LID and it variations behaved similarly. It seems $m = 15$ is enough to give a reasonable approximation to the limiting distribution, and increasing m does not help much. In this example, the MSE of LID seems to be a little worse than the MSE of weighted RQ, but still a little better than that of RQ.

Table 5.3: Comparison of the MSEs of the median from different methods ($n = 200$).

Methods	MSE of $a(0.5)$	SE of MSE	MSE of $b(0.5)$	SE of MSE
RQ	0.09	0.006	0.10	0.007
EWRQ	0.06	0.004	0.06	0.004
OWRQ	0.06	0.004	0.06	0.004
LID 15	0.07	0.005	0.07	0.005
LID 19	0.07	0.005	0.08	0.005
LID 23	0.08	0.005	0.08	0.006
LID* 15	0.07	0.005	0.07	0.005
LID* 19	0.07	0.005	0.07	0.005
LID* 23	0.07	0.005	0.07	0.006
TD (5.2)	0.03	0.002	0.04	0.003
MLE (5.2)	0.03	0.002	0.04	0.003
TQ (5.2)	0.04	0.003	0.05	0.004

Table 5.4: Comparison of the MSEs of the third quartile from different methods ($n = 200$).

Methods	MSE of $a(0.75)$	SE of MSE	MSE of $b(0.75)$	SE of MSE
RQ	0.11	0.008	0.10	0.007
EWRQ	0.07	0.005	0.07	0.004
OWRQ	0.07	0.005	0.07	0.004
LID 15	0.10	0.007	0.08	0.006
LID 19	0.10	0.008	0.08	0.006
LID 23	0.10	0.008	0.08	0.006
LID* 15	0.11	0.010	0.09	0.010
LID* 19	0.10	0.007	0.08	0.006
LID* 23	0.10	0.008	0.08	0.006
TD (5.2)	0.04	0.003	0.04	0.003
MLE (5.2)	0.04	0.003	0.04	0.003
TQ (5.2)	0.06	0.004	0.06	0.006

Next, we consider the following model:

$$y_i = a + bx_i + (1 + x_i)\epsilon_i, \quad i = 1, 2, ..., n, \tag{5.3}$$

where ϵ_i's are i.i.d. from F. The distribution F has a piecewise linear CDF between the $1/m$-th and the $(m-1)/m$-th quantile, where the $1/m, 2/m, ..., (m-1)/m$-th quantiles are the same as those of $N(0, 1)$. Between the i/m-th and $(i+1)/m$-th quantiles, $i = 1, 2, ..., m-2$, the CDF is linear. The left tail of F between $-\infty$ and the $1/m$-th quantile is proportional to a truncated normal, the left half of $N(\Phi^{-1}(1/m), 2^2)$, and the right tail between the $(m-1)/m$-th quantile and ∞ is proportional to the right half of $N(\Phi^{-1}((m-1)/m), 2^2)$. The only difference between Model (5.1) and Model (5.3) is the error term. All other settings are the same.

Correspondingly, we have the following parametric model for the Bayesian method that uses the true density:

$$y_i = a + bx_i + (\gamma_1 + \gamma_2 x_i)\epsilon_i, \quad i = 1, 2, ..., n, \tag{5.4}$$

where ϵ_i's are i.i.d. from F.

We use TQ (5.3) to denote the Bayesian method based on Model (5.3), assuming the underlying distribution of ϵ_i is unknown. For TQ (5.3) there are $m+1$ parameters: a, b, and $m-2$ quantiles of ϵ_i. TD (5.4) denotes the Bayesian method using the true densities based on Model (5.4). For this example, we added Yu and Moyeed's (2005) method, denoted by YM, in the comparison.

Table 5.5: Comparison of the MSEs of the median from different methods ($n = 100$ and $m = 15$).

Methods	MSE of $a(0.5)$	SE of MSE	MSE of $b(0.5)$	SE of MSE
RQ	0.19	0.015	0.19	0.014
EWRQ	0.12	0.008	0.11	0.008
OWRQ	0.12	0.008	0.11	0.007
LID	0.12	0.008	0.12	0.008
LID*	0.10	0.007	0.10	0.007
YM	0.16	0.013	0.17	0.013
TQ (5.3)	0.15	0.011	0.14	0.009
TD (5.4)	0.07	0.005	0.06	0.005

Table 5.6: Comparison of the MSEs of the third quartile from different methods ($n = 100$ and $m = 15$).

Methods	MSE of $a(0.75)$	SE of MSE	MSE of $b(0.75)$	SE of MSE
RQ	0.23	0.015	0.21	0.014
EWRQ	0.19	0.036	0.17	0.020
OWRQ	0.15	0.010	0.14	0.010
LID	0.17	0.013	0.13	0.010
LID*	0.14	0.011	0.11	0.009
YM	0.20	0.014	0.18	0.012
TQ (5.3)	0.15	0.011	0.14	0.011
TD (5.4)	0.08	0.006	0.08	0.006

From Tables 5.5 and 5.6, we can see that LID and LID* work well. Their performance is similar to that of weighted RQ for $\tau = 0.5$. For $\tau = 0.75$, LID and LID* are better than EWRQ, and LID* seems to be the best among all the methods except TD (5.4), which should be the optimal result that the Bayesian method could achieve. Other than these, we can see that the performance of Yu and Moyeed's method is only slightly better than RQ, which is not surprising because there are some similarities between these two methods.

We also increased the size of each data set from 100 to 200, and the results are in Tables 5.7 and 5.8. We can see that in this case EWRQ, OWRQ, LID, LID* and TQ (5.3) perform very similarly.

Table 5.7: Comparison of the MSEs of the median from different methods ($n = 200$ and $m = 15$).

Methods	MSE of $a(0.5)$	SE of MSE	MSE of $b(0.5)$	SE of MSE
RQ	0.09	0.007	0.10	0.007
EWRQ	0.05	0.004	0.06	0.004
OWRQ	0.05	0.004	0.06	0.004
LID	0.06	0.004	0.05	0.004
LID*	0.06	0.004	0.06	0.005
YM	0.08	0.006	0.09	0.006
TQ (5.3)	0.06	0.005	0.06	0.004
TD (5.4)	0.03	0.002	0.03	0.002

Table 5.8: Comparison of the MSEs of the third quartile from different methods ($n = 200$ and $m = 15$).

Methods	MSE of $a(0.75)$	SE of MSE	MSE of $b(0.75)$	SE of MSE
RQ	0.11	0.008	0.11	0.008
EWRQ	0.06	0.004	0.07	0.005
OWRQ	0.06	0.004	0.06	0.004
LID	0.06	0.004	0.06	0.005
LID*	0.07	0.006	0.08	0.008
YM	0.10	0.007	0.10	0.007
TQ (5.3)	0.06	0.004	0.06	0.004
TD (5.4)	0.03	0.002	0.04	0.003

5.1.2 The MSE for difference of quantiles

Because our proposed method estimates many quantiles simultaneously and RQ only tackles one quantile at a time, it is possible that our method may produce better estimates for some functions of multiple quantiles. Here we consider the estimation of the difference of quantiles on three examples. In the first example, We used $m = 15$ quantiles and each data set contains 100 or 200 observations generated from Model (5.3). We compared the MSE of the difference of the parameters of the 0.75 quantile and the 0.5 quantile for the following five methods: RQ, EWRQ, OWRQ, LID and YM. For all the Bayesian methods, we constructed a Markov chain with length 1,000,000, used the first half as the burn-in period, and took every 1,000-th samples. The results are in Tables 5.9 and 5.10.

Table 5.9: The MSE and its standard error of the difference between the median and the third quartile with $n = 100$ for Model (5.3)

Methods	MSE of $a(0.75) - a(0.5)$	SE of MSE	MSE of $b(0.75) - b(0.5)$	SE of MSE
RQ	0.16	0.011	0.15	0.010
EWRQ	0.12	0.008	0.11	0.007
OWRQ	0.11	0.007	0.09	0.006
LID	0.07	0.005	0.03	0.002
YM	0.10	0.007	0.10	0.007

Table 5.10: The MSE and its standard error of the difference between the median and the third quartile with $n = 200$ for Model (5.3)

Methods	MSE of $a(0.75) - a(0.5)$	SE of MSE	MSE of $b(0.75) - b(0.5)$	SE of MSE
RQ	0.09	0.006	0.09	0.007
EWRQ	0.05	0.004	0.06	0.004
OWRQ	0.05	0.003	0.05	0.004
LID	0.03	0.002	0.02	0.002
YM	0.07	0.004	0.07	0.005

We can see that the MSE of LID is the smallest among all the methods. When $n = 100$, the MSE of LID is about half of that of EWRQ and OWRQ for $a(0.75) - a(0.5)$, and the MSE of LID is about one fourth of that of EWRQ and OWRQ for $b(0.75) - b(0.5)$. When $n = 200$, the MSE of LID for both $a(0.75) - a(0.5)$ and $b(0.75) - b(0.5)$ are about half of that of EWRQ and OWRQ.
In the second example, We used $m = 15$ quantiles and each data set contains 100 or 200 observations generated from Model (5.1). We compared the MSE of the difference of the parameters of the 0.75 quantile and the 0.5 quantile for the following five methods: RQ, EWRQ, OWRQ, LID and YM. For all the Bayesian methods, we constructed a Markov chain with length 1,000,000, used the first half as the burn-in period, and took every 1,000-th samples. The results are in Tables 5.11 and 5.12.

Table 5.11: The MSE and its standard error of the difference between the median and the third quartile with $n = 100$ for Model (5.1)

Methods	MSE of $a(0.75) - a(0.5)$	SE of MSE	MSE of $b(0.75) - b(0.5)$	SE of MSE
RQ	0.17	0.013	0.16	0.012
EWRQ	0.11	0.008	0.12	0.009
OWRQ	0.11	0.008	0.11	0.007
LID	0.06	0.004	0.03	0.003
YM	0.11	0.008	0.11	0.008)

Table 5.12: The MSE and its standard error of the difference between the median and the third quartile with $n = 200$ for Model (5.1)

Methods	MSE of $a(0.75) - a(0.5)$	SE of MSE	MSE of $b(0.75) - b(0.5)$	SE of MSE
RQ	0.085	0.006	0.08	0.005
EWRQ	0.051	0.004	0.05	0.004
OWRQ	0.048	0.004	0.05	0.003
LID	0.038	0.004	0.03	0.002
YM	0.063	0.005	0.06	0.004

From Tables 5.11 and 5.12, we can see that LID outperforms other methods for estimating the difference of quantiles for this model.

In the third example, the data are generated from the following model:

$$y_i = a + bx_{1,i} + cx_{2,i} + (1 + x_{1,i} + x_{2,i})\epsilon_i, \quad i = 1, 2, ..., n, \tag{5.5}$$

where ϵ_i's are i.i.d. from N(0,1). The corresponding quantile model is

$$Q_\tau(y_i|x_i) = a(\tau) + b(\tau)x_{1,i} + c(\tau)x_{2,i}, \quad i = 1, 2, ..., n, \quad \tau = \frac{1}{m+1}, ..., \frac{m}{m+1}. \tag{5.6}$$

In the simulations, we chose $a = 5$, $b = 1$, and $c = 1$. The covariate $x_{1,i}$ was generated from lognormal(0,1) and $x_{2,i}$ was generated from Bernoulli(0.5). We set $m = 15$ and $n = 100$ or 200. We compared the MSE of the difference of the parameters of the 0.75 quantile and the 0.5 quantile for the following five methods: RQ, EWRQ, OWRQ, LID and YM. For all the Bayesian methods, we constructed a Markov chain with length 1,000,000, used the first half as the burn-in period, and took every 1,000-th samples. The results are in Tables 5.13 and 5.14.

Table 5.13: The MSE and its standard error (in parenthesis) of the difference between the median and the third quartile with $n = 100$ for Model (5.5)

Methods	MSE of $a(0.75) - a(0.5)$	MSE of $b(0.75) - b(0.5)$	MSE of $c(0.75) - c(0.5)$
RQ	0.28 (0.021)	0.20 (0.014)	0.42 (0.031)
EWRQ	0.19 (0.014)	0.16 (0.011)	0.40 (0.028)
OWRQ	0.19 (0.014)	0.14 (0.009)	0.39 (0.028)
LID	0.28 (0.012)	0.03 (0.002)	0.18 (0.012)
YM	0.18 (0.014)	0.13 (0.010)	0.28 (0.023)

Table 5.14: The MSE and its standard error (in parenthesis) of the difference between the median and the third quartile with $n = 200$ for Model (5.5)

Methods	MSE of $a(0.75) - a(0.5)$	MSE of $b(0.75) - b(0.5)$	MSE of $c(0.75) - c(0.5)$
RQ	0.13 (0.008)	0.10 (0.006)	0.22 (0.016)
EWRQ	0.09 (0.006)	0.07 (0.005)	0.20 (0.013)
OWRQ	0.09 (0.006)	0.07 (0.005)	0.20 (0.013)
LID	0.07 (0.005)	0.03 (0.002)	0.12 (0.008)
YM	0.09 (0.006)	0.07 (0.005)	0.17 (0.012)

From Tables 5.13 and 5.14, we can see that for estimating the difference between quantiles, LID outperforms all other methods except for $a(0.75) - a(0.5)$ with $n = 100$.
Therefore, when the main interest is the difference of the parameters for different quantiles, LID showed a big advantage over all the other methods.

5.2 Level and Power studies

In this section, we study the level and power for our proposed method in hypotheses testing. We are interested in knowing whether our method can achieve the claimed level and whether our method could be more powerful than RQ. First we consider the following model:

$$y_i = a + bx_{1,i} + cx_{2,i} + (1 + 0.2x_{1,i} + x_{2,i})\epsilon_i, \quad i = 1, 2, ..., n, \tag{5.7}$$

where ϵ_i's are i.i.d. from N(0,1). Model (5.7) has non-i.i.d. errors.

In the simulations, we chose $a = 5$, $b = 1$, and $c = 1$. The covariate $x_{1,i}$ was generated from lognormal(0,1) and $x_{2,i}$ was generated from Bernoulli(0.5). We set $m = 15$ and $n = 1000$. We simulated 100 data sets from the model and applied LID and RQ to each data set. We know the true value of the differences between the parameters for different quantiles, so we can subtract this value from the parameter and test whether this parameter is 0. To determine whether the parameter is significant or not, we checked whether the 95% confidence/credible interval contains 0. We recorded the number of times that the parameter is significant under the claimed level 0.05. The results are in Tables 5.15 to 5.17.

Table 5.15: The number of times of significance of the difference between the median and the first quartile for Model (5.7)

Methods	$b(0.5) - b(0.25) - 0.1348980$	$c(0.5) - c(0.25) - 0.6744898$
RQ	3	4
LID	2	7

Table 5.16: The number of times of significance of the difference between the median and the 0.125 quantile for Model (5.7)

Methods	$b(0.5) - b(0.125) - 0.2300698$	$c(0.5) - c(0.125) - 1.150349$
RQ	2	5
LID	2	15

Table 5.17: The number of times of significance of the difference between the first quartile and the 0.125 quantile for Model (5.7)

Methods	$b(0.25) - b(0.125) - 0.09517192$	$c(0.25) - c(0.125) - 0.4758596$
RQ	4	4
LID	1	6

We can see that RQ gives roughly the correct level. LID works reasonably well for this example,

with the level a little low for b and a little high for c.

We also compared the power of the test for Model (5.7). Because we know that the true value of $b(0.5) - b(0.25)$ and $c(0.5) - c(0.25)$ are not 0, we tested whether these parameters are significant or not for each data set and we recorded the number of times that the parameter is significant under the claimed level 0.05. The results are in Tables 5.18 to 5.20. We can see that the power of LID is better than that of RQ for the differences of $b(\tau)$. For the differences of $c(\tau)$, RQ seems to be slightly better than LID.

Table 5.18: The number of times of significance of the difference between the median and the first quartile for Model (5.7)

Methods	$b(0.5) - b(0.25)$	$c(0.5) - c(0.25)$
RQ	60	100
LID	94	96

Table 5.19: The number of times of significance of the difference between the median and the 0.125 quantile for Model (5.7)

Methods	$b(0.5) - b(0.125)$	$c(0.5) - c(0.125)$
RQ	86	100
LID	99	100

Table 5.20: The number of times of significance of the difference between the first quartile and the 0.125 quantile for Model (5.7)

Methods	$b(0.25) - b(0.125)$	$c(0.25) - c(0.125)$
RQ	29	89
LID	43	78

We also tested the following model:

$$y_i = a + bx_{1,i} + cx_{2,i} + (1 + 0.2x_{1,i} + 0.5x_{2,i})\epsilon_i, \quad i = 1, 2, ..., n, \tag{5.8}$$

where the covariate $x_{1,i}$ was generated from lognormal(0,1) and $x_{2,i}$ was generated from Gamma(1,1/2).

For this model, we can see that the differences of $c(\tau)$ is only half of that for Model (5.7), and we should be able to see better whether RQ is truly better at detecting the differences of $c(\tau)$. The results are in Tables 5.21 to 5.23. We can see that LID has a better power for the difference of both $b(\tau)$ and $c(\tau)$ in this case.

Table 5.21: The number of times of significance of the difference between the median and the first quartile for Model (5.8)

Methods	$b(0.5) - b(0.25)$	$c(0.5) - c(0.25)$
RQ	66	58
LID	99	72

Table 5.22: The number of times of significance of the difference between the median and the 0.125 quantile for Model (5.8)

Methods	$b(0.5) - b(0.125)$	$c(0.5) - c(0.125)$
RQ	85	76
LID	99	89

Table 5.23: The number of times of significance of the difference between the first quartile and the 0.125 quantile for Model (5.8)

Methods	$b(0.25) - b(0.125)$	$c(0.25) - c(0.125)$
RQ	38	30
LID	53	33

5.3 Bootstrap testing

In this section we used the bootstrap idea to study the level and power of hypotheses testing. We bootstrapped the data and used LID and RQ to give the estimates for each bootstrapped data set, and then used the standard deviation of the estimates as the standard error.

Consider the following model:

$$y_i = a + bx_{1,i} + \epsilon_i, \quad i = 1, 2, ..., n, \tag{5.9}$$

where ϵ_i's are i.i.d. from N(0,1). In the simulations, we chose $a = 5$, and $b = 1$. The covariates $x_{1,i}$ was generated from Normal(0,1). We set $m = 15$ and $n = 200$. The corresponding quantile model is

$$Q_\tau(y_i|x_i) = a(\tau) + b(\tau)x_{1,i}, \quad i = 1, 2, ..., n, \quad \tau = \frac{1}{m+1}, ..., \frac{m}{m+1}. \tag{5.10}$$

We simulated 100 data sets from this model and compared the number of times of significance of LID and RQ. Here we focus on the following parameters: $a(0.5) - a(0.25)$ and $b(0.5) - b(0.25)$. We can see that the true value of $a(0.5) - a(0.25)$ is not 0 and the true value of $b(0.5) - b(0.25)$ should be 0. For each data set, we used 40 bootstrap samples to give the standard error. Based on this standard error, we constructed the 95% confidence/credible interval and checked whether 0 is in the interval. We recorded the number of times that the parameter is significant under the claimed level 0.05.

For this model, we treat the number of times that $a(0.5) - a(0.25)$ is significant as a measurement of the power and the number of times that $b(0.5) - b(0.25)$ is significant as a measurement of the level. The results are in Table 5.24. The expected number of times of significance for $b(0.5) - b(0.25)$ is 5. LID gives the right level and the level for RQ is a little high.

Table 5.24: The number of times of significance of the difference between the median and the first quartile

Methods	$a(0.5) - a(0.25)$	$b(0.5) - b(0.25)$
RQ	100	12
LID	100	6

We did more simulations to confirm the findings. We simulated 500 data sets from Model (5.9). The results are in Table 5.25. We can see that for the 500 data sets, the estimated level of $b(0.5) - b(0.25)$ is very close to 25 for LID, but it is a little high for RQ.

Table 5.25: The number of times of significance of the difference between the median and the first quartile for 500 data sets

Methods	$a(0.5) - a(0.25)$	$b(0.5) - b(0.25)$
RQ	500	47
LID	496	22

Because in the simulation RQ is not giving the right level, we used the asymptotic standard error instead of the bootstrap variance to determine the significance of the parameters for RQ. The results are in Table 5.26. We can see that RQ now also gives roughly the right level. In Figures 5.5 and 5.6, we have the plots of the bootstrap variance versus the asymptotic variance. We can clearly see that the bootstrap variance is usually larger.

Table 5.26: The number of times of significance of the difference between the median and the first quartile for 500 data sets (corrected for RQ)

Methods	$a(0.5) - a(0.25)$	$b(0.5) - b(0.25)$
RQ	500	15
LID	496	22

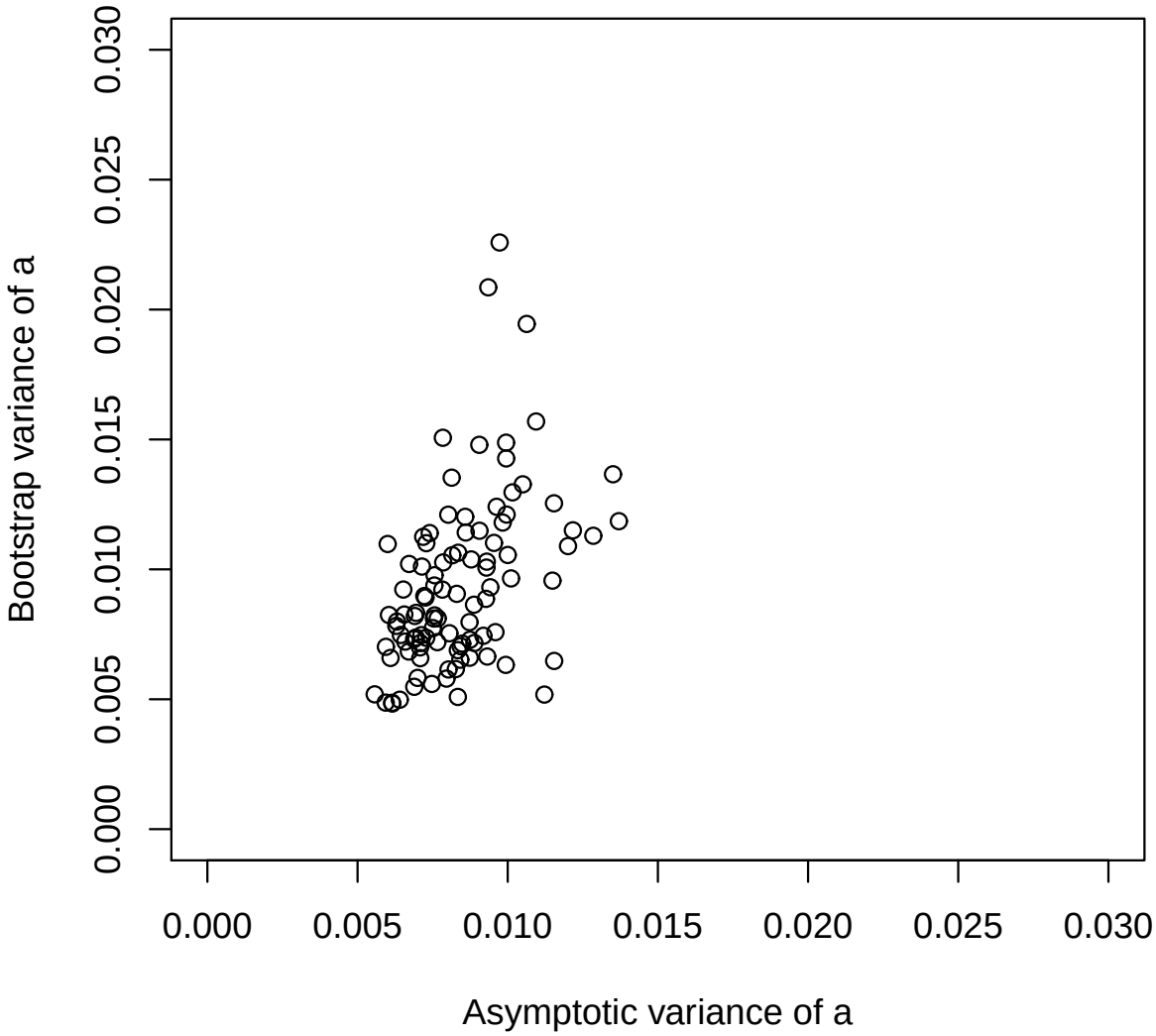

Figure 5.1: The plot of the bootstrap variance versus the asymptotic variance for $a(0.5) - a(0.25)$

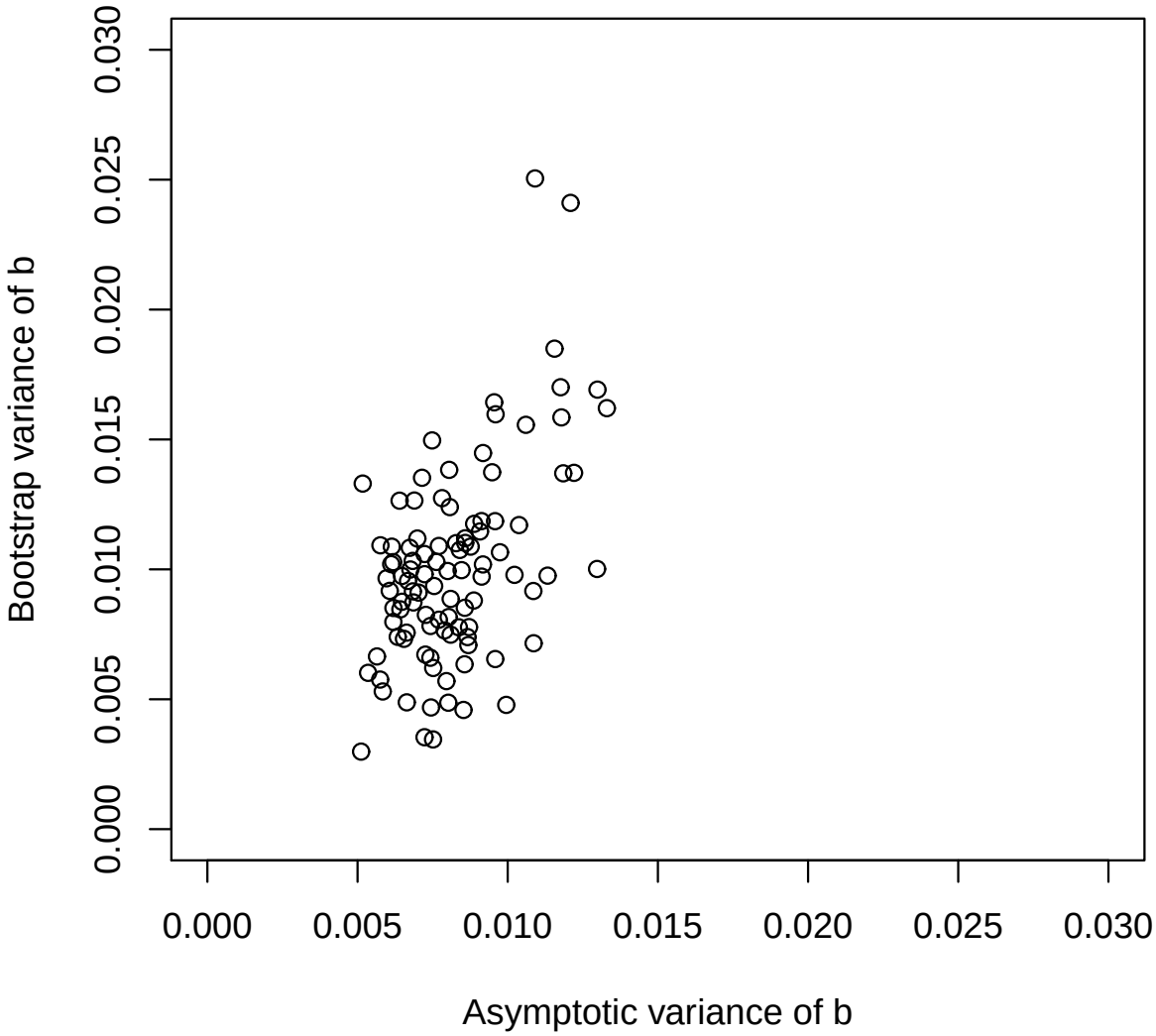

Figure 5.2: The plot of the bootstrap variance versus the asymptotic variance for $b(0.5) - b(0.25)$

5.4 Birth weight data

In this section, we revisit the birth weight data. We consider the following quantile model for the birth weight data:

$$Q_\tau(y_i|x_i) = a(\tau) + b(\tau)x_{i,1} + c(\tau)x_{i,2} + d(\tau)x_{i,3}, \quad i = 1, 2, ..., n, \tag{5.11}$$

where $x_{i,1}$ is the indicator function that indicates whether the mother went to prenatal care for more than or equal to two times, $x_{i,2}$ is the indicator function that indicates whether the mother smoked or not, and $x_{i,3}$ is mother's weight gain during pregnancy. We compared the results from RQ and LID for the full data set. Here we focus on the 0.25, 0.5 and 0.75 quantiles and the difference between the 0.25 and 0.5 quantiles. The results are in Tables 5.27 to 5.29.

Table 5.27: Estimates of the parameters and their standard errors (in parentheses) for the birth weight data with $\tau = 0.25$.

Methods	$a(0.25)$	$b(0.25)$	$c(0.25)$	$d(0.25)$
RQ	2.94 (0.0045)	−0.048 (0.0069)	−0.22 (0.0081)	0.0091 (0.00020)
LID	2.94 (0.0032)	−0.036 (0.0058)	−0.21 (0.0020)	0.0085 (0.00003)

Table 5.28: Estimates of the parameters and their standard errors (in parentheses) for the birth weight data with $\tau = 0.5$.

Methods	$a(0.5)$	$b(0.5)$	$c(0.5)$	$d(0.5)$
RQ	3.26 (0.0040)	−0.064 (0.0063)	−0.23 (0.0070)	0.0084 (0.00018)
LID	3.27 (0.0038)	−0.057 (0.0048)	−0.23 (0.0046)	0.0084 (0.00013)

Table 5.29: Estimates of the parameters and their standard errors (in parentheses) for the birth weight data with $\tau = 0.75$.

Methods	$a(0.75)$	$b(0.75)$	$c(0.75)$	$d(0.75)$
RQ	3.59 (0.0044)	−0.058 (0.0071)	−0.22 (0.0076)	0.0078 (0.00019)
LID	3.61 (0.0023)	−0.062 (0.0061)	−0.26 (0.0041)	0.0083 (0.00024)

From the results, we can see that the estimates from both methods are close for most parameters with a few exceptions, such as $d(0.25)$ and $c(0.75)$. The standard error from LID seems to be smaller than that from RQ. For $d(0.25)$, the standard error is extremely small, so we checked the histogram and the trace plot, which are in Figures 5.3 and 5.4. The trace plot of the chain looks fine and the Markov chain does not get stuck in a local mode.

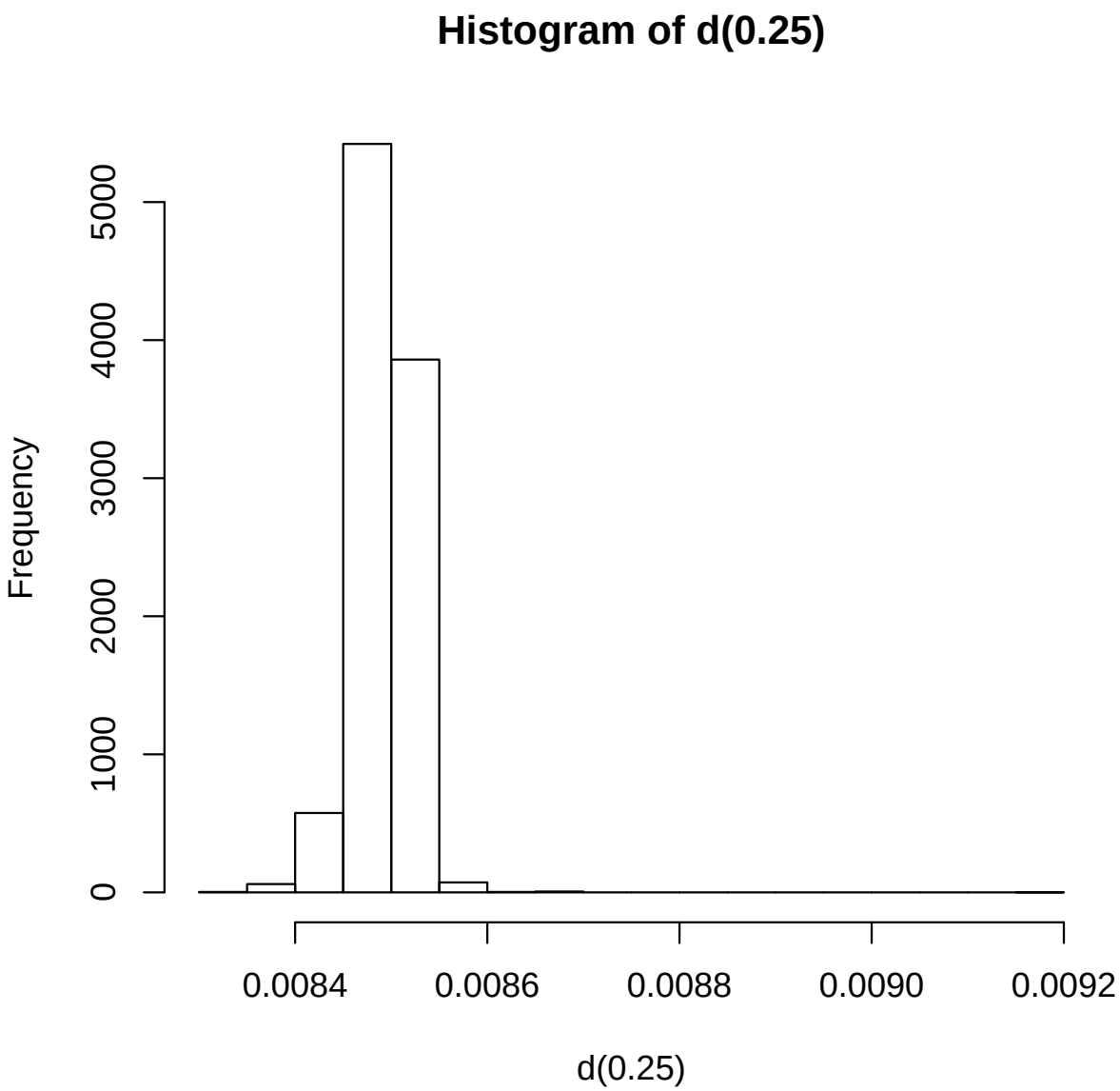

Figure 5.3: The histogram of $d(0.25)$

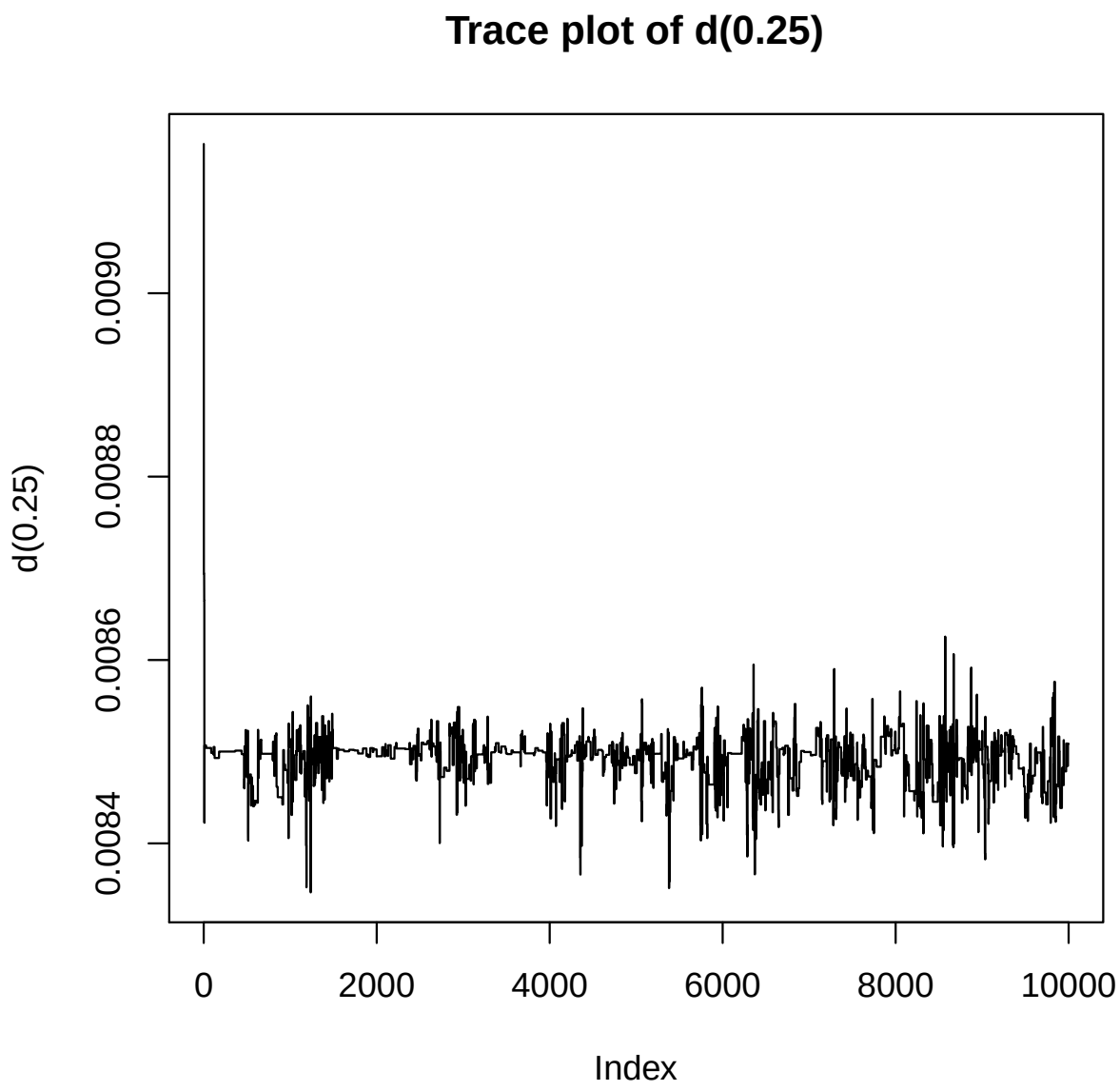

Figure 5.4: The trace plot of $d(0.25)$

To see how well 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 $44.5 < x_{i,5} < 45.5$, 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 Tables 5.30 and 5.31. From the results, we can see that all the estimated quantiles are very close to the local quantile estimates.

Table 5.30: Estimates of the local quantile at $x_{i,1} = 1$, $x_{i,2} = 1$, and $x_{i,3} = 25$.

Quantile	Local quantile	RQ estimate	LID estimate
0.25	2.81	2.76	2.77
0.5	3.02	3.07	3.08
0.75	3.41	3.40	3.40

Table 5.31: Estimates of the local quantile at $x_{i,1} = 1$, $x_{i,2} = 0$, and $x_{i,3} = 45$.

Quantile	Local quantile	RQ estimate	LID estimate
0.25	3.18	3.21	3.19
0.5	3.54	3.53	3.53
0.75	3.86	3.84	3.88

Then, we compared the performance of both methods for randomly selected subsets to check the variability of the methods. We randomly sampled 50 data sets from the full data set, with 1000 observations in each data set. For each data set, we sampled from the full data set without replacement. Then, based on each data set, we computed the estimates for the parameters, and compared them with the estimates from the full data set, which we treated as the "truth". In this way, we can calculate the MSE for all the parameters. The results are in Tables 5.32 to 5.35. From the results, we can see that for single quantile estimation, RQ has slightly smaller MSEs than LID. In Table 5.35, we looked at the MSE of the difference of the quantiles. The results show that LID has smaller MSEs except for the intercept. In particular, for $d(0.5) - d(0.25)$, which is the parameter for mother's weight gain, LID has a much smaller MSE than that of RQ.

Table 5.32: MSE of the parameters and their standard errors (in parentheses) for the birth weight data with $\tau = 0.25$.

Methods	$a(0.25)$	$b(0.25)$	$c(0.25)$	$d(0.25)$
RQ	0.10 (0.003)	0.0031 (0.0005)	0.0047 (0.0009)	3.30×10^{-6} (6×10^{-7})
LID	0.14 (0.004)	0.0048 (0.0009)	0.0066 (0.0014)	6.25×10^{-6} (1.2×10^{-6})

Table 5.33: MSE of the parameters and their standard errors (in parentheses) for the birth weight data with $\tau = 0.5$.

Methods	$a(0.5)$	$b(0.5)$	$c(0.5)$	$d(0.5)$
RQ	0.10 (0.003)	0.0031 (0.0006)	0.0026 (0.0006)	3.55×10^{-6} (6×10^{-7})
LID	0.11 (0.003)	0.0042 (0.0008)	0.0049 (0.0011)	4.91×10^{-6} (9×10^{-7})

Table 5.34: MSE of the parameters and their standard errors (in parentheses) for the birth weight data with $\tau = 0.75$.

Methods	$a(0.75)$	$b(0.75)$	$c(0.75)$	$d(0.75)$
RQ	0.0014 (0.0003)	0.0037 (0.0008)	0.0037 (0.0007)	2.84×10^{-6} (5×10^{-7})
LID	0.0031 (0.0009)	0.0045 (0.0010)	0.0084 (0.0021)	4.56×10^{-6} (6×10^{-7})

Table 5.35: MSE of the difference between the 0.5 and the 0.25 quantile and their standard errors (in parentheses) for the birth weight data.

Methods	$a(0.5) - a(0.25)$	$b(0.5) - b(0.25)$	$c(0.5) - c(0.25)$	$d(0.5) - d(0.25)$
RQ	0.40 (0.006)	0.0036 (0.0006)	0.0030 (0.0006)	5.16×10^{-6} (1.0×10^{-6})
LID	0.49 (0.007)	0.0029 (0.0006)	0.0028 (0.0006)	1.23×10^{-6} (2×10^{-7})

To provide an explanation of the different performance between the estimates of single quantiles and the difference of quantiles, we looked at the correlation between $d(0.5)$ and $d(0.25)$ estimated from both methods. Figures 5.5 and 5.6 are the plots of $d(0.5)$ versus $d(0.25)$ from both methods. From the two plots, we can see that the correlation between $d(0.5)$ and $d(0.25)$ is much stronger for

LID than that for RQ. The correlation between $d(0.5)$ and $d(0.25)$ for LID is 0.89 and the correlation for RQ is 0.57. The reason that the correlation between $d(0.5)$ and $d(0.25)$ is larger for LID is because LID assumes more about the global likelihood than individual RQ. This strong correlation decreased the variability of the estimate of $d(0.5) - d(0.25)$ for LID.

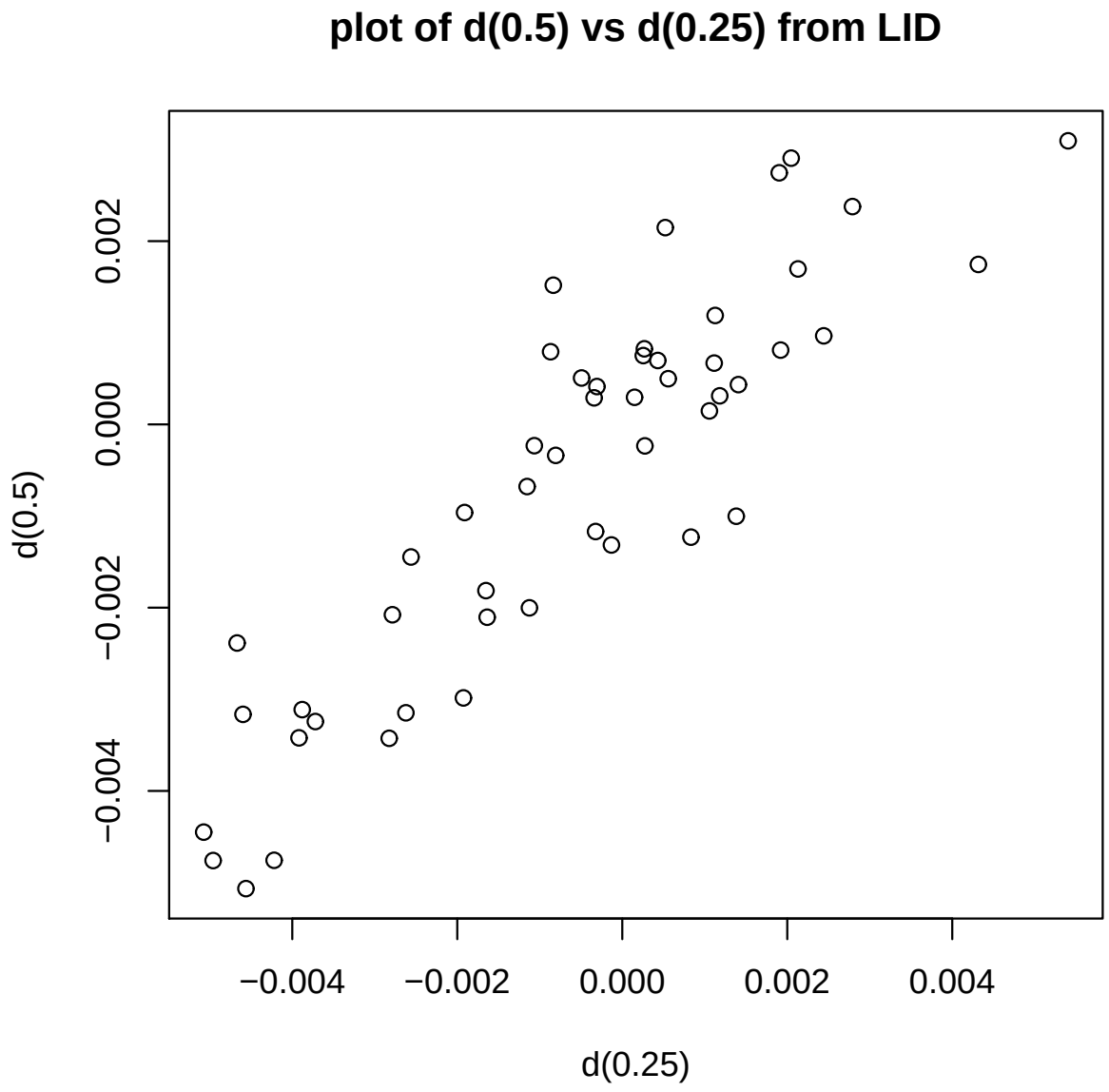

Figure 5.5: The plot of d(0.5) versus d(0.25) from LID over the 50 data sets. The correlation is about 0.89.

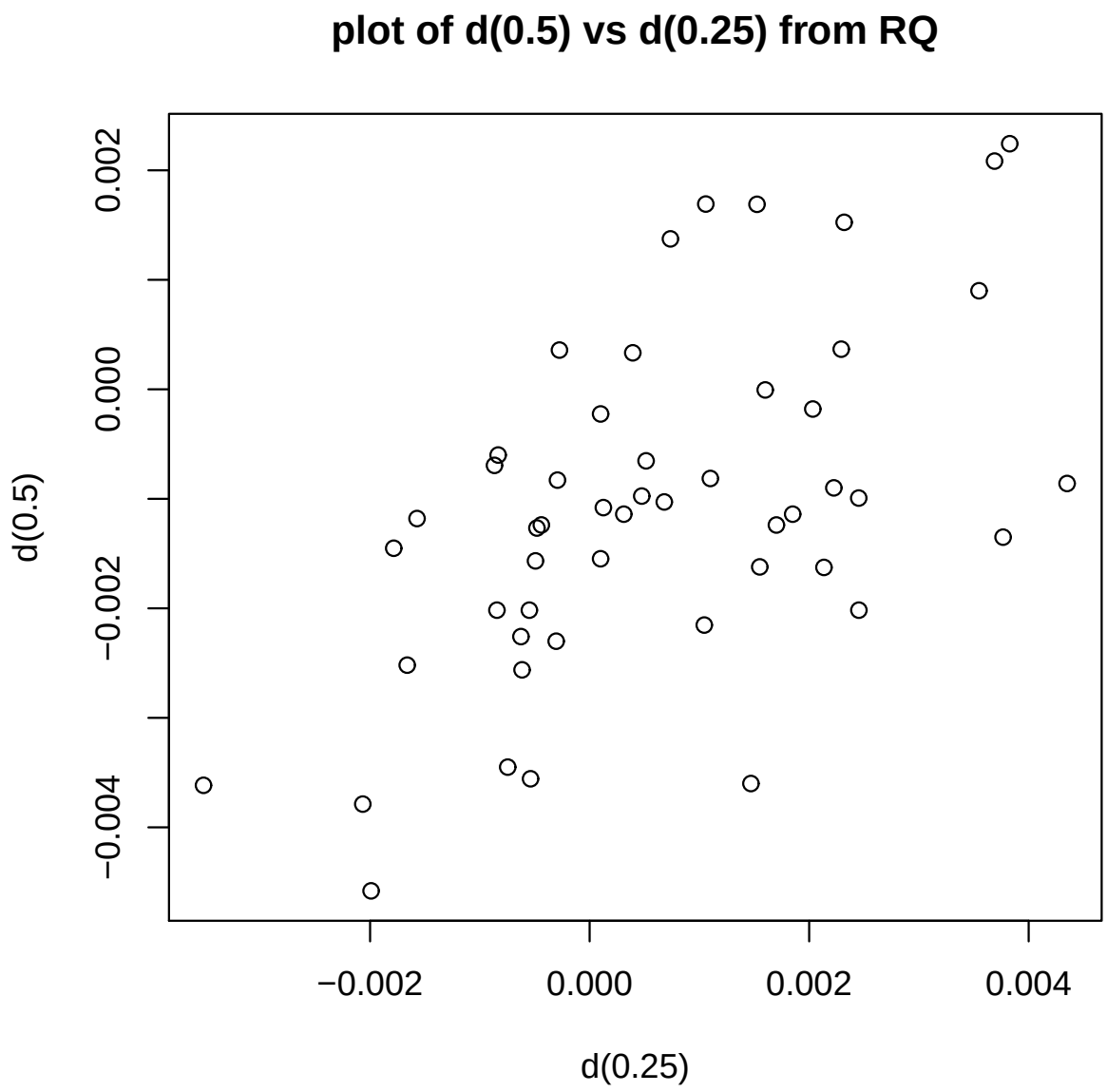

Figure 5.6: The plot of d(0.5) versus d(0.25) from RQ over the 50 data sets. The correlation is about 0.57

5.5 Conclusions

From the simulation results, we can see that for estimating single quantiles, LID performs similarly as weighted RQ. For differences of quantiles, LID performs better than other methods. For a non-i.i.d. error model, LID has a reasonable level and good power. In bootstrap testing, LID gives the

correct level.

For the birth weight data, both LID and RQ give good estimates of the quantiles. RQ has slightly smaller MSEs for estimating single quantiles and LID has smaller MSEs for estimating the difference of quantiles. The large correlation between the parameters estimated by LID may explain why LID performs better for estimating the difference of quantiles.

Chapter 6

Conclusions and Future Work

In this thesis, we introduced two Bayesian methods, DG and LID, for the quantile regression problem. We proved the convergence of these two methods under some mild conditions and numerically verified the theoretical results. From the simulation results, we found that LID could produce more efficient estimates than some existing methods. In particular, for estimating the difference of quantiles, LID has a big advantage over other existing methods. Besides, we tried two ways to do hypotheses testing based on LID estimates: one is to use the posterior distribution, and the other is to use the bootstrap idea. We found that for a non-i.i.d. error model, LID is more powerful than RQ with the first testing method. With the bootstrap testing, LID can provide the right level.

The followings are some possible future directions. First, we would like to generalize our methods for censored data. One challenging issue is how to interpolate the densities for the censored parts. Second, we would like to generalize our algorithms to some non-linear models. As long as it is possible to find some proposal distribution satisfying the order constraint, our algorithms should be able to be generalized in this direction. Third, we only implemented linear interpolation up to now, so it is of our interest to see whether other interpolations could enhance the algorithm. For example, we can try some smooth interpolations so that the interpolated densities will be continuous or even differentiable. Then, the assumptions in Chapter 3 will be easier to check. Fourth, LID is a computationally intensive algorithm. If we can find some way to reduce the computational complexity, it will make the method more widely applicable in practice.

References

[1] Angelis, D., Hall, P., and Young, G. A., "Analytical and bootstrap approximations to estimator distributions in l^1 regression," *Journal of the American Statistical Association*, vol. 88, pp. 1310–1316, 1993.

[2] Dunson, D. and Taylor, J., "Approximate bayesian inference for the quantiles," *Journal of Nonparametric Statistics*, vol. 17:3, pp. 385–400, 2005.

[3] Edgeworth, F. Y., "On a new method of reducing observations relating to several quantities," *Philosophical Magazine*, vol. 25, pp. 184–191, 1888.

[4] Efron, B. and Tibshirani, R., *An Introduction to the Bootstrap*. New York: Chapman & Hall, 2004.

[5] Galton, F., "Regression towards mediocrity in hereditary stature," *Journal of the Anthropological Institute*, vol. 15, pp. 246–263, 1885.

[6] He, X. and Hu, F., "Markov chain marginal bootstrap," *Journal of the American Statistical Association*, vol. 97(459), pp. 783–795, 2002.

[7] Hendricks, W. and Koenker, R., "Hierarchical spline models for conditional quantiles and the demand for electricity," *Journal of the American Statistical Association*, vol. 87, pp. 58–68, 1991.

[8] Kocherginsky, M. and He, X., "Extensions of the markov chain marginal bootstrap," *Statistics and Probability Letters*, vol. 77, pp. 1258–1268, 2007.

[9] Kocherginsky, M., He, X., and Mu, Y., "Practical confidence intervals for regression quantiles," *Journal of Computational an Graphical Statistics*, vol. 14, pp. 41–45, 2005.

[10] Koenker, R., *Regression Quantiles*. New York: Cambridge University Press, 2005.

[11] Koenker, R. and Bassett, G., "Regression quantiles," *Econometrica*, vol. 46, pp. 33–50, 1978.

[12] Kotts, A. and Gelfand, A., "Bayesian semiparametric median regression modeling," *Journal of the American Statistical Association*, vol. 96, pp. 1458–1468, 2001.

[13] Lavine, M., "On an approximate likelihood for quantiles," *Biometrica*, vol. 82, pp. 220–222, 1995.

[14] Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S., "Markov chain monte carlo without likelihoods," *Proceedings of the National Academy of Sciences*, vol. 26, pp. 15324–15328, 2003.

[15] Parzen, M. I., Wei, L., and Ying, Z., "A resampling method based on pivotal estimating functions," *Biometrika*, vol. 81, pp. 341–350, 1994.

[16] Robert, C. and Casella, G., *Monte Carlo Statistical Methods*. New York: Springer, 2004.

[17] Siddiqui, M., "Distribution of quantiles from a bivariate population," *Journal of Research of the national Bureau of Standaeds*, vol. 64, pp. 145–150, 1960.

[18] Taddy, M. and Kottas, A., "A bayesian nonparametric approach to inference for quantile regression," *Journal of Business and Economic Statistics*, vol. ahead of print, 2009.

[19] Yu, K. and Moyeed, R. A., "Bayesian quantile regression," *Statistics & Probability Letters*, vol. 54, pp. 437–447, 2001.

vita

Yang Feng was born in Yizheng, Jiangsu, China on March 12, 1985, the son of Tianlang Feng and Cuihua Yang. After completing high school at Yizheng High School, Jiangsu, China in 2002, he attended Nanjing University in Jiangsu, China from 2002-2006. He graduated with Bachelor of Science degree in 2006. From 2006-2011, he attended the University of Illinois at Urbana-Champaign, Illinois. He graduated with a Ph.D degree in 2011.