Bayesian Quantile Regression using a Mixture of Pólya Tree Prior

Minzhao Liu, Mike Daniels

June 10, 2013

1 Introduction

Quantile regression is an attractive way of studying the relationship between response and covariates when one (or several) quantiles are of interest as compared to mean regression. The dependence between upper or lower quantiles of the response variable and the covariates are expected to vary differentially relative to that of the average. This is often of interest in econometrics, educational studies, biomedical studies, and environment studies (Yu and Moyeed, 2001; Buchinsky, 1994, 1998; He et al., 1998; Koenker and Machado, 1999; Wei et al., 2006; Yu et al., 2003).

A comprehensive review of quantile regression was presented in (Koenker, 2005). Furthermore, mean regression provides less information about the relationship of the average with linear combination of covariates; quantile regression can offer a more complete description of the conditional distribution of the response.

The traditional frequentist approach was proposed by (Koenker and Bassett, 1978) for a single quantile (τ) with estimators derived by minimizing a loss function. The popularity of this approach is due to its computational efficiency by linear programming, well-developed asymptotic properties, and straightforward extensions to simultaneous quantile regression and random effect models. However, asymptotic inference may not be accurate for small sample sizes.

Bayesian approaches offer exact inference. Motivated by the loss (check) function, (Yu and Moyeed, 2001) proposed an asymmetric Laplace distribution for the error term, such that maximizing the posterior distribution is equivalent to minimizing the check function. Other than parametric Bayesian approaches, some semiparametric methods have been proposed for median regression. (Walker and Mallick, 1999) used a diffuse finite Polya Tree prior for the error term. (Kottas and Gelfand, 2001) modeled the error by two families of median zero distribution using a mixture Dirichlet process priors, which is very useful for unimodal error distributions. (Hanson and Johnson, 2002) adopted mixture of Pólya Tree prior on error term to make inference in regression model. They illustrated the implementation on AFT model for the median survival time, which showed robustness of Pólya in terms of multimodality and skewness. Other recent approaches include quantile pyramid priors, mixture of Dirichlet process priors of multivariate normal distributions and infinite mixture of Gaussian densities which put quantile constraints on the residuals (Hjort and Petrone, 2007; Hjort and Walker, 2009; Kottas and Krnjajić, 2009; Reich et al., 2010).

Like the asymmetric Laplace distribution, all of the above methods are single semiparametric quantile regression methods, which have some limitations. The densities have their (restrictive) mode at the quantile of interest, which is not appropriate when extreme quantiles are being investigated. Other criticisms include crossed quantile lines, monotonicity constraints, difficulty in making inference for quantile regression parameter for an interval of τs. Joint inference is poor in borrowing information through single quantile regressions. It is not coherent to pool from every individual quantile regression. Meanwhile, the sampling distribution of response for τ_1 might not be the same as that under quantile τ_2. **might be more specific about why the above issues are problems** *Minzhao: show mike tokdar2011*

In order to solve those problems, simultaneous linear quantile regression have been proposed by (Tokdar and Kadane, 2011). Another popular approach is to assign a nonparametric model for the error term to avoid the monotonicity problem ((Scaccia and Green, 2003), (Geweke and Keane, 2007), (Taddy and Kottas, 2010)).

We use Pólya Tree (PT) priors in our approach. PT priors were introduced decades ago ((Freedman, 1963), (Fabius, 1964), (Ferguson, 1974)) and Lavine extended them to Pólya Tree models ((Lavine, 1992, 1994)). The major advantage of Pólya Tree over Dirichlet process is that it can be absolutely continuous with probability 1 and it can be easily tractable **isn't the DP tractable too??** *Minzhao: DP is tractable too, here I only mean PT is better than DP in terms of continuous pdf.*. In a regression context, (Walker and Mallick, 1997, 1999) assigned a finite Pólya Tree prior to the random effects in a generalized linear mixed model. (Berger and Guglielmi, 2001) used a mixture of Pólya Tree comparing data distribution coming from parametric distribution or mixture of Pólya Tree. **this is not clear??***Minzhao: used a Pólya tree process to test the fit of data to a parametric model by embedding the parametric model in a nonparametric alternative and computing the Bayes factor of the parametric model to the nonparametric alternative.* (Hanson and Johnson, 2002) as mentioned earlier modeled the error term as a mixture of Pólya tree prior.

Multivariate regression is also possible with Pólya Tree. (Paddock, 1999, 2002) studied multivariate Pólya Tree in a k-dimensional hypercube. (Hanson, 2006) constructed a general framework for multivariate random variable with a Pólya Tree distribution. (Jara et al., 2009) extended the multivariate mixture of Pólya Tree prior from Hanson with directional orthogonal matrix. He also demonstrated how to fit a generalized mixed effect model by modeling a multivariate random effects with multivariate mixture of Pólya Tree priors.

In this article, we present a Bayesian approach by adopting a mixture of Pólya Tree prior for the regression error term, and we account for the change of quantile regression parameter via heterogeneity of the error term. As a result, several quantile regression can be fit simultaneously and there is a closed form for posterior quantile regression parameter. Exact inference can be made through MCMC, and our method avoids the problem of crossing quantile lines that occurs in the traditional frequentist quantile regressions.

The rest of the paper is organized as follows. In Section 2, we introduce the heterogeneity model and derive a closed form for marginalized posterior quantile regression parameter with mixture of Pólya tree prior. We conduct some simulation studies in section 3 and use a real data example to illustrate our approach in Section 4. Finally, conclusions are presented in section 5.

2 Model, Priors, and Computations

2.1 Heterogeneity Model

Let Y be a random variable with CDF F. The τth quantile of Y is defined as

$$Q_Y(\tau) = \inf_y \{y : F(y) \geq \tau\}.$$

If covariates $\boldsymbol{x_1, \ldots, x_n}$ are of interest, then the quantile regression parameter satisfies this condition:

$$Q_Y(\tau) = \boldsymbol{X'\beta}(\tau),$$

where $\boldsymbol{X}$ is the matrix of covariates including an intercept (ie. $\boldsymbol{x_1} = \mathbf{1}$). If F is continuous, then $F(\boldsymbol{X'\beta}(\tau)) = \tau$, i.e., $\mathrm{p}(Y \leq \boldsymbol{X'\beta}(\tau)) = \tau$.

Now, consider a location shift model,

$$y_i = \boldsymbol{x}_i\beta + \epsilon_i,$$

where $\epsilon_i \overset{\text{i.i.d}}{\sim} F_\epsilon$. Then, the τth quantile regression parameter can be expressed as

$$\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + F_\epsilon^{-1}(\tau)\boldsymbol{e}_1, \tag{1}$$

where $\boldsymbol{e}_1 = [1, 0, \ldots, 0]^T$, **put the next little bit in an appendix** and $F_\epsilon^{-1}(\tau)$ is the τth quantile for error ϵ. **up to here**

As we can see from equation (1), if the model is homogeneous , i.e., i.i.d case, then for different quantiles τ, the corresponding quantile regression parameters only vary in the first component, the intercept. The rest of the quantile regression parameters stay the same. Therefore, quantile lines for different quantiles are parallel to each other. **i would put figure 1 in the dissertation, but not the paper**

Now, consider the heterogeneous linear regression model from (He et al., 1998)

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + (\boldsymbol{x}_i'\boldsymbol{\gamma})\epsilon_i, \tag{2}$$

need an equation # for this one where $\boldsymbol{x_i'\gamma}$ is positive for all i. Under this model, the τth quantile regression parameter is

$$\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + F_\epsilon^{-1}(\tau)\boldsymbol{\gamma}, \tag{3}$$

put this little 'proof' in the appendix too figure for dissertation

Quantile lines are no longer parallel under the heterogeneous linear model which adds considerably more flexibility in the model.

this paragraph should probably be merged with material in the introduction *Minzhao: merge to introduction* Traditional single quantile regression make different assumptions on the error term. The frequentist approach of (Koenker and Bassett, 1978) does not assign distributions for the residual, and uses linear programming technique to minimize the check function $\sum_{i=1}^n \rho_\tau(y_i - \boldsymbol{x_i'\beta})$, where $\rho_\tau(\epsilon) = \epsilon(\tau - \mathrm{I}(\epsilon < 0))$. Some Bayesian approaches specify the error distribution as an asymmetric Laplace distribution ((Yu and Moyeed, 2001)), or Dirichlet process prior ((Kottas and Gelfand, 2001), (Kottas and Krnjajić, 2009), (Taddy and Kottas,

2010)) or Pólya tree prior ((Walker and Mallick, 1999), (Hanson and Johnson, 2002)). (Reich et al., 2010) uses an infinite mixture of Gaussian densities on the residual. However, all these densities keep their restrictive mode at the quantile of interest, i.e., in equation (3), $\boldsymbol{\beta}(\tau) \equiv \beta$. Other limitations exist as well such as crossing quantile lines, monotonicity constraints, non-coherent joint quantile regression inference from each single quantile regression. The sampling distribution of response in τ_1th quantile regression is not even the same as that in τ_2th quantile regression.

We use a mixture of Pólya tree prior for the error term in (#) and derive a closed form for posterior quantile regression parameter in (3). Since Pólya tree are a very flexible way to model the unknown distribution, our approach makes fewer assumptions. Exact inference can be made through MCMC and functional of posterior samples. The next subsection briefly reviews the Pólya tree priors and their relevant properties.

2.2 Pólya Tree

(Lavine, 1992, 1994) and (Mauldin et al., 1992) developed theory for Pólya tress priors as a generalization of the Dirichlet process ((Ferguson, 1974)). Denote $E = \{0, 1\}$ and E^m as the m-fold product of E, $E^0 = \emptyset$, $E^* = \cup_0^\infty E^m$ and Ω be a separable measurable space , $\pi_0 = \Omega$, $\Pi = \{\pi_m : m = 0, 1, \ldots\}$ be a separating binary tree of partitions of Ω. In addition, define $B_\emptyset = \Omega$ and $\forall \epsilon = \epsilon_1 \cdots \epsilon_m \in E^*$, $B_{\epsilon 0}$ and $B_{\epsilon 1}$ are the two partition of B_ϵ.

Definition 2.2.1. *A random probability measure G on $(\Omega, \mathcal{F})$ is said to have a Pólya tree distribution, or a Pólya tree prior with parameter $(\Pi, \mathcal{A})$, written as $G|\Pi, \mathcal{A} \sim \mathrm{PT}(\Pi, \mathcal{A})$, if there exist nonnegative numbers $\mathcal{A} = \{\alpha_\epsilon, \epsilon \in E^*\}$ and random vectors $\mathcal{Y} = \{Y_\epsilon : \epsilon \in E^*\}$ such that the following hold:*

1. *all the random variables in $\mathcal{Y}$ are independent;*
2. *$Y_\epsilon = (Y_{\epsilon 0}, Y_{\epsilon 1}) \sim \mathrm{Dirichlet}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1}), \forall \epsilon \in E^*$;*
3. *$\forall m = 1, 2, \ldots,$ and $\forall \epsilon \in E^*$, $G(B_{\epsilon_1, \ldots, \epsilon_m}) = \prod_{j=1}^m Y_{\epsilon_1 \cdots \epsilon_j}$.*

2.2.1 Pólya Tree Parameters

There are two parameters in the Pólya tree distribution $(\Pi, \mathcal{A})$. *Minzhao: If a Pólya tree is centered around a pre-specified distribution G_0, which is called the baseline measure,* the $\mathcal{A}$ family determines how much G can deviate from G_0. ~~*, the baseline measure*~~ **need to define G_0 as the base measure before this,** $\Pi(G_0)$. (Ferguson, 1974) pointed out $\alpha_{\epsilon=1}$ yields a G that is continuous singular with probability 1, and $\alpha_{\epsilon_1, \ldots, \epsilon_m} = m^2$ yields G that is absolutely continuous with probability 1. (Walker and Mallick, 1999) and (Paddock, 1999) considered $\alpha_{\epsilon_1, \ldots, \epsilon_m} = cm^2$, where $c > 0$. (Berger and Guglielmi, 2001) considered $\alpha_{\epsilon_1, \ldots, \epsilon_m} = c\rho(m)$. In general, any $\rho(m)$ such that $\sum_{m=1}^\infty \rho(m)^{-1} < \infty$ guarantees G to be absolutely continuous. In our case, we adopt $\alpha_{\epsilon_1, \ldots, \epsilon_m} = cm^2$.

As to the partition parameter Π, the canonical way of constructing a Pólya tree distribution G centering on G_0, a continuous CDF is to choose $B_0 = G_0^{-1}([0, 1/2])$, $B_1 = G_0^{-1}((1/2, 1])$, such that $G(B_0) = G(B_1) = 1/2$. Furthermore, for all $\epsilon \in E^*$, choose $B_{\epsilon 0}$ and $B_{\epsilon 1}$ to satisfy $G(B_{\epsilon 0}|B_\epsilon) = G(B_{\epsilon 1}|B_\epsilon) = 1/2$, then any choice of $\mathcal{A}$ makes G coincide with G_0. A simple

example is to choose $B_{\epsilon 0}$ and $B_{\epsilon 1}$ in level m by setting them as $G_0^{-1}\left((k/2^m,(k+1)/2^m]\right)$, for $k=0,\ldots,2^m-1$.

2.2.2 Some properties of Pólya Tree

Suppose $G\sim \mathrm{PT}(\Pi,\mathcal{A})$ is a random probability measure and $\epsilon_1,\epsilon_2,\ldots$ are a random sample from G.

Definition 2.2.2 (Expectation of Pólya Tree)**.** *$F=E(G)$ as a probability measure is defined by $F(B)=E(G(B)),\forall B\in\mathcal{B}$. By the definition of Pólya tree, for any $\epsilon\in E^*$,*

$$F(B_\epsilon)=E(G(B_\epsilon))=\prod_{j=1}^{m}\frac{\alpha_{\epsilon_1,\ldots,\epsilon_j}}{\alpha_{\epsilon_1,\ldots,\epsilon_{j-1},0}+\alpha_{\epsilon_1,\ldots,\epsilon_{j-1},1}}.$$

Remark 2.2.3. *If G is constructed based on baseline measure G_0 and we set $\alpha_{\epsilon_1,\ldots,\epsilon_m}=cm^2$, $\epsilon_{\epsilon_0}=\alpha_{\epsilon_1}$, then $\forall B\in\mathcal{B},F(B)=G_0(B)$; thus, $F=G_0$, if there is no data.*

Definition 2.2.4 (Density Function)**.** *Suppose $F=E(G),G|\Pi,\mathcal{A}\sim\mathrm{PT}(\Pi,\mathcal{A})$, where G_0 is the baseline measure. Then, using the canonical construction, $F=G_0$ (as shown above), the density function is*

$$f(x)=\left[\prod_{j=1}^{m}\frac{\alpha_{\epsilon_1,\ldots,\epsilon_j}(x)}{\alpha_{\epsilon_1,\ldots,\epsilon_{j-1},0}(x)+\alpha_{\epsilon_1,\ldots,\epsilon_{j-1},1}(x)}\right]2^m g_0(x), \tag{4}$$

where g_0 is the pdf of G_0.

Remark 2.2.5. *When using the canonical construction with no data, $\alpha_{\epsilon_0}=\alpha_{\epsilon_1}$, equation (4) simplifies to*

$$f(x)=g_0(x).$$

Remark 2.2.6 (Conjugacy)**. should use y's instead of x's here??** *If $y_1,\ldots,y_n|G\sim G,G|\Pi,\mathcal{A}\sim\mathrm{PT}(\Pi,\mathcal{A})$, then $G|y_1,\ldots,y_n,\Pi,\mathcal{A}\sim\mathrm{PT}(\Pi,\mathcal{A}^*)$, where in $\mathcal{A}^*,\forall\epsilon\in E^*$,*

$$\alpha_\epsilon^*=\alpha_\epsilon+n_\epsilon(y_1,\ldots,y_n),$$

where $n_\epsilon(y_1,\ldots,y_n)$ indicates the count how many samples of $y_1,\ldots,y_n$ drop in B_ϵ.

2.2.3 Mixture of Pólya Tree

The behavior of a single Pólya tree highly depends on how the partition is separated. A random probability measure G_θ is said to be a mixture of Pólya tree if there exists a random variable θ with distribution h_θ , and Pólya tree parameters $(\Pi_\theta,\mathcal{A}_\theta)$ such that $G_\theta|\theta=\theta\sim\mathrm{PT}(\Pi^\theta,\mathcal{A}^\theta)$.

Example 2.2.7. *Suppose $G_0=\mathrm{N}(\mu,\sigma^2)$ is the baseline measure. For $\epsilon\in E^*,\alpha_{\epsilon_m}=cm^2$, $\boldsymbol{\theta}=(\mu,\sigma^2,c)$ is the mixing index and the distribution on $\Theta=(\mu,\sigma^2,c)$ is the mixing distribution.*

With the mixture of Pólya tree, the influence of the partition is lessened. Thus, inference will not be affected greatly by a single Pólya tree distribution.

2.2.4 Predictive Error Density, Cumulative Density Function and Quantiles

Suppose $G_\theta = \mathrm{N}(0, \sigma^2)$ is the baseline measure, $g_0(x) = \phi(x; 0, \sigma^2)$ is the density function. Π^θ is defined as

$$B^\theta_{\epsilon_1,\ldots,\epsilon_m} = \left(G_\theta^{-1}\left(\frac{k}{2^m}\right), G_\theta^{-1}\left(\frac{k+1}{2^m}\right)\right),$$

where k is the index of partition $\epsilon_1, \ldots, \epsilon_m$ in level m. $\mathcal{A}^c$ is defined as

$$\alpha_{\epsilon_1,\ldots,\epsilon_m} = cm^2.$$

Therefore, the error model is

$$\begin{aligned} x_1, \ldots, x_n | G_\theta &\overset{\text{i.i.d}}{\sim} G, \\ G|\Pi^\theta, \mathcal{A}^c &\sim \mathrm{PT}(\Pi^\theta, \mathcal{A}^c). \end{aligned}$$

The predictive density function of $X|x_1, \ldots, x_n, \theta$, marginalizing out G , is

$$f^\theta_x(x|x_1, \ldots, x_n) = \lim_{m\to\infty} \left[\prod_{j=2}^{m} \frac{cj^2 + n_{\epsilon_1\cdots\epsilon_j(x)}(x_1, \ldots, x_n)}{2cj^2 + n_{\epsilon_1\cdots\epsilon_{j-1}(x)}(x_1, \ldots, x_n)}\right] 2^{m-1} g_0(x), \tag{5}$$

where $n_{\epsilon_1\cdots\epsilon_j(x)}(x_1, \ldots, x_n)$ denotes the number of observations $x_1, \ldots, x_n$ dropping in the slot $\epsilon_1 \cdots \epsilon_j$ where x stays in the level j. Notice that, if we restrict the first level weight as $\alpha_0 = \alpha_1 = 1$, then we only need to update levels other than the first level.

Remark 2.2.8 (The predictive density for Finite Pólya Tree)**.** *In practice, a finite M level Pólya Tree is usually adopted to approximate the full Pólya tree, in which, only up to M levels are updated. The corresponding predictive density becomes*

$$f^{\theta,M}_x(x|x_1, \ldots, x_n) = \left[\prod_{j=2}^{M} \frac{cj^2 + n_{\epsilon_1\cdots\epsilon_j(x)}(x_1, \ldots, x_n)}{2cj^2 + n_{\epsilon_1\cdots\epsilon_{j-1}(x)}(x_1, \ldots, x_n)}\right] 2^{M-1} g_0(x). \tag{6}$$

The rule of thumb for choosing M is to set $M = \log_2 n$, where n is the sample size.

(Hanson and Johnson, 2002) showed the approximation to (4) given in (6) is exact for M large enough. We now derive the predictive cdf and the predictive quantile(s).

Theorem 2.2.9. *Based on the predictive density function (6) of a finite Pólya tree distribution, the predictive cumulative density function is*

$$F^{\theta,M}_X(x|x_1, \ldots, x_n) = \sum_{i=1}^{N-1} P_i + P_N\left(G_\theta(x)2^M - (N-1)\right), \tag{7}$$

where

$$\begin{aligned} P_i &= \frac{1}{2}\left\{\prod_{j=2}^{M} \frac{cj^2 + n_{j,\lceil i2^{j-M}\rceil}(x_1, \ldots, x_n)}{2cj^2 + n_{j-1,\lceil i2^{j-1-M}\rceil}(x_1, \ldots, x_n)}\right\} \text{ and} \\ N &= \left[2^M G_\theta(x) + 1\right], \end{aligned}$$

in which $n_{j,\lceil i2^{j-M}\rceil}(x_1, \ldots, x_n)$ denotes the number of observations $x_1, \ldots, x_n$ in the $\lceil i2^{j-M}\rceil$ slot at level j, $\lceil\cdot\rceil$ is the ceiling function, and $[\cdot]$ is the floor function.

Proof.

$$
\begin{aligned}
F_X^{\theta,M}(x|x_1,\ldots,x_n) &= \int_{-\infty}^{x} f_x^{\theta,M}(x|x_1,\ldots,x_n)dx \\
&= \int_{-\infty}^{x} \left[\prod_{j=2}^{M} \frac{cj^2 + n_{\epsilon_1\cdots\epsilon_j(x)}(x_1,\ldots,x_n)}{2cj^2 + n_{\epsilon_1\cdots\epsilon_{j-1}(x)}(x_1,\ldots,x_n)}\right] 2^{M-1} g_\theta(x)dx \\
&= \sum_{i=1}^{N-1} \left[\prod_{j=2}^{M} \frac{cj^2 + n_{j,\lceil i2^{j-M}\rceil}(x_1,\ldots,x_n)}{2cj^2 + n_{j-1,\lceil i2^{j-1-M}\rceil}(x_1,\ldots,x_n)} 2^{M-1} \int_{\epsilon_{M,i}} g_\theta(x)dx\right] \\
&+ \int_{G_\theta^{-1}((N-1)/2^M)}^{x} \left[\prod_{j=2}^{M} \frac{cj^2 + n_{j,\lceil N2^{j-M}\rceil}(x_1,\ldots,x_n)}{2cj^2 + n_{j-1,\lceil N2^{j-1-M}\rceil}(x_1,\ldots,x_n)}\right] 2^{M-1} g_\theta(x)dx \\
&= \sum_{i=1}^{N-1} P_i + P_N 2^M \left(G_\theta(x) - G_\theta(G_\theta^{-1}\left(\frac{N-1}{2^M}\right)\right) \\
&= \sum_{i=1}^{N-1} P_i + P_N \left(G_\theta(x)2^M - (N-1)\right),
\end{aligned}
$$

where $\epsilon_{M,i}$ is the ith partition in level M. □

Theorem 2.2.10. *The posterior predictive quantile of finite Pólya tree distribution is*

$$
Q_{X|x_1,\ldots,x_n}^{\theta,M}(\tau) = G_\theta^{-1}\left[\frac{\tau - \sum_{i=1}^{N} P_i + NP_N}{2^M P_N}\right], \tag{8}
$$

where N satisfies $\sum_{i=1}^{N-1} P_i < \tau \leq \sum_{i=1}^{N} P_i$.

Proof. From equation (7),

$$
\begin{aligned}
\tau = F_X^{\theta,M}(x|x_1,\ldots,x_n) &= \sum_{i=1}^{N-1} P_i + P_N \left(G_\theta(x)2^M - (N-1)\right) \\
\Rightarrow G_\theta(x) &= \frac{\tau - \sum_{i=1}^{N} P_i + NP_N}{2^M P_N} \\
x &= G_\theta^{-1}\left[\frac{\tau - \sum_{i=1}^{N} P_i + NP_N}{2^M P_N}\right].
\end{aligned}
$$

□

Now explicitly state result (corollary) for closed form for (2) Now the explicit form for quantile regression coefficients in equation (3) becomes:

$$
\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + \boldsymbol{\gamma} G_\theta^{-1}\left[\frac{\tau - \sum_{i=1}^{N} P_i + NP_N}{2^M P_N}\right], \tag{9}
$$

where P_i and N are the notations in equation (7) and (8).

2.3 Fully Bayesian Quantile Regression Specification with Mixture of Pólya Tree Priors

The full Bayesian specification of our quantile regression is given as follows,

$$\begin{aligned} y_i &= \boldsymbol{x_i'}\boldsymbol{\beta} + (\boldsymbol{x_i'}\boldsymbol{\gamma})\epsilon_i, i = 1, \ldots, n \\ \epsilon_i | G_\theta &\overset{\text{i.i.d}}{\sim} G_\theta \\ G_\theta | \Pi^\theta, \mathcal{A}^\theta &\sim \text{PT}(\Pi^\theta, \mathcal{A}^\theta) \\ \boldsymbol{\theta} = (\sigma^2, c) &\sim \pi_\theta(\theta) \\ \boldsymbol{\beta} &\sim \pi_\beta(\beta) \\ \boldsymbol{\gamma} &\sim \pi_\gamma(\gamma). \end{aligned}$$

In order to not confound the location parameter, ϵ_i or G is set to have median 0 by fixing $\alpha_0 = \alpha_1 = 1$. For the similar reasons, the first component of $\boldsymbol{\gamma}$ is fixed at 1.

The posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, c)$ is given as

$$\begin{aligned} \text{p}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, c | \boldsymbol{Y}) &\propto L(\boldsymbol{Y} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, c) \pi_\beta(\beta) \pi_\gamma(\gamma) \pi_{\sigma^2}(\sigma^2) \pi_c(c) \\ &= \frac{1}{\prod_{i=1}^n (\boldsymbol{x_i'}\boldsymbol{\gamma})} \text{p}\left(\epsilon_1, \ldots, \epsilon_n | \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, c\right) \pi_\beta(\beta) \pi_\gamma(\gamma) \pi_{\sigma^2}(\sigma^2) \pi_c(c) \\ &= \frac{1}{\prod_{i=1}^n (\boldsymbol{x_i'}\boldsymbol{\gamma})} \text{p}\left(\epsilon_n | \epsilon_1, \ldots, \epsilon_{n-1}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, c\right) \cdots \text{p}\left(\epsilon_2 | \epsilon_1, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, c\right) \text{p}\left(\epsilon_1 | \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, c\right) \\ &\quad \pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \pi_\gamma(\boldsymbol{\gamma}) \pi_{\sigma^2}(\sigma^2) \pi_c(c). \end{aligned}$$

Need to discuss choices of priors here

Usually priors for parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ could be diffused p-dimensional normal distribution . Diffused gamma distribution could be chosen as priors for σ^2 and c. For shrinkage model, spike priors could be adopted to shrink the parameter estimates to pre-specified values. In addition, spike priors can also help variable selection in Bayesian model and shrink heterogeneity parameters to zero to find homogeneous model. Moreover, spike and slab priors can help to accommodate zero-inflated situation and research hypothesis in variable selection.

3 Simulations

We conduct a simulation study to compare our approach with other existing methods, specifically, the 'rq' function in the 'quantreg' package in R (the standard frequentist quantile regression method) and flexible Bayesian quantile regression approach by Reich ('BQR'). We compare quantile regression approaches for both homogeneous and heterogeneous models.

3.1 Design

We generated data from the following 3 models,

M1: $y_i = 1 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_{1i}$,

M2: $y_i = 1 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_{2i}$,

M3: $y_i = 1 + x_{i1}\beta_1 + x_{i2}\beta_2 + (1 - 0.5x_{i1} + 0.5x_{i2})\epsilon_{2i}$,

where $x_{i1}, x_{i2} \overset{\text{iid}}{\sim} \text{N}(0,1), \epsilon_{1i} \sim \text{N}(0,1), \epsilon_{2i} \overset{\text{iid}}{\sim} 0.5 \times \text{N}(-2,1) + 0.5 \times \text{N}(2,1)$. All covariates and error terms are mutually independent. All coefficients are set to be 1. For each model, we generate 100 data sets with the sample size $n = 100$. $\tau = 0.5, 0.9$ are the quantiles of interest.

Each simulated data set is analyzed using the three methods. For the proposed Bayesian linear quantile regression with Pólya Tree prior (PT), we adopt the following prior specifications: $\mu_\beta = (0,0,0)^T, \Sigma_\beta = \text{diag}(\sqrt{1000}, \sqrt{1000}, \sqrt{1000}), \mu_\gamma = (0,0,0)^T, \Sigma_\gamma = \text{diag}(\sqrt{1000}, \sqrt{1000}, \sqrt{1000})$, and $\tau = (0.01, 0.01)$. A partial Pólya tree with $M = 6$ levels was adopted in the model. For Monte Carlo Markov chain parameter, 220,000 iterations of a single Markov chain were used, during which, 10,000 samples were saved through every 20 steps after a burn-in period of 20,000 samples. **can comment on time??** *Minzhao: 330 seconds for one simulation under Linux 2.6.32 kernel, x86_64 machine, 8g memory, 4 core 2.66GHz cpu.* Acceptance rates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ candidates during the adaptive Metropolis-Hastings algorithm were set to approach 25%. We also use the method proposed by Reich (BQR), which conducts a single τ quantile regression for linear model and assigns an infinite mixture of Gaussian densities for the error term and the standard frequentist quantile regression approach , 'rq' function in the 'quantreg' package in R (RQ).

Methods are evaluated based on mean squared error:

$$\text{MSE} = \frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j(\tau) - \beta_j(\tau))^2,$$

where p is the number of covariates except the intercept (so here, $p = 2$). $\beta_j(\tau)$ is the $j-$component of the true quantile regression parameters. $\hat{\beta}_j(\tau)$ is the jth component of estimated quantile regression parameters. (we use the posterior median for the Bayesian approaches).

Table 1: Mean squared error (reported as 100*average) and standard error (reported as 100*standard error) for each quantile regression method. **identify abbreviations in table in the table caption here** *Minzhao: The three columns ('rq', 'BQR', 'PT') stand for frequentist method (rq) function from 'quantreg' R package, flexible Bayesian method by Reich , and our Bayesian approach using Pólya tree separately.*

Model	quantile	rq	BQR	PT
M1	0.5	1.39(0.13)	0.96(0.10)	0.96(0.09)
M2		17.2(1.48)	4.09(0.5)	1.89(0.26)
M3		95.4(6.69)	16.5(1.90)	6.29(0.86)
M1	0.9	2.35(0.26)	1.96(0.25)	1.79(0.17)
M2		5.73(1.03)	4.32(0.54)	3.83(0.49)
M3		25.1(2.66)	12.4(1.44)	14.06(1.39)

3.2 Results

The simulation results are shown in Table 1. The proposed Bayesian quantile regression method with Pólya tree prior (PT) does very well (in terms of MSE) relative to Reich's method (BQR) and traditional frequentist approach (rq). The differences becomes quite large in the non-unimodal case (M2) and heterogeneous model (M3).

In Model 2 and Model 3 with $\tau = 0.5$, where the error is distributed as a bimodal distribution (mixture of normal distributions), the rq method performs poorly in terms of MSE since the mode of the error is no longer the quantile of interest. In contrast, our method (PT) is not impacted by lack of unimodality and heterogeneity and provides more information for the relationship between responses and covariates. In Model 3 with $\tau = 0.9$, Reich's method (BQR) outperforms our approach (PT), since in his model, the error is assigned an infinite mixture of normal distribution with mode at $\tau = 0.9$, which is very close to the true distribution. Less information is available from our approach to detect the shape at a particular quantile of the distribution since there are few observations at extreme quantiles.

4 Analysis of the Tours Data

In this section, we apply our Bayesian quantile regression approach to examine the quantiles of 6 month weight loss from a recent weight management study (Perri et al., 2008?). In particular, we are interested in the effects of age and race. The response of interest is the weight loss from baseline to 6 months. The age of the subjects ranged from 50 to 75, and there were 43 people with race classifed as black (race=1) and 181 people, white (race=2). Our goal is to determine how the percentiles of weight change are affected by their age and race.

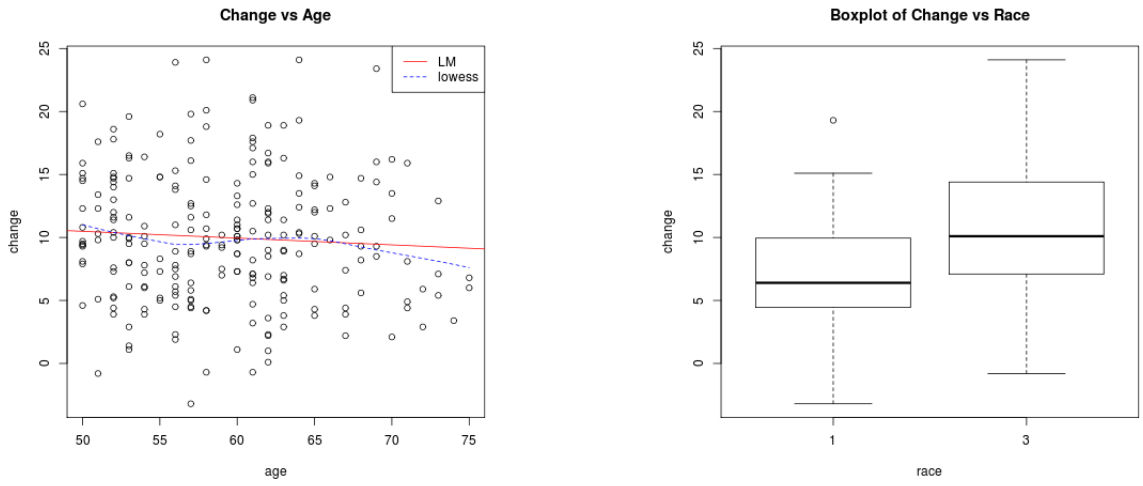

Figure 1: Scatterplots of weight change vs age and Boxplots of weight loss for each race. On the left figure, Red solid line is the fitted line from regular mean regression model with one covariate 'age', and blue dashed line is the fitted lowess line for model. The boxplots use the default settings: (0.75, 0.5, 0.25) quantile for box and $Q1 - 1.5IQR$ for lower whisker and $Q3 + 1.5IQR$ for upper whisker.

Figure 3 shows weight loss vs age and the boxplots by race. There does not appear to be heterogeneity of the response errors. Weight loss though does appear to be a bit right skewed.

For numerical stability, age was centered and standardized.

Table 2: 95% credible (confidence) intervals for tours data quantile regression parameter for the traditional frequentist approach (QReg) and the proposed PT approach.

	τ	PT	QReg	PT Estimates	QReg Estimates
Intercept	0.5	(9.73, 11.20)	(9.31, 10.80)	10.38	10.30
Age		(-1.02, 0.07)	(-1.27, 0.17)	-0.41	-0.69
Race 1		(-5.28, -2.33)	(-5.46, -2.42)	-3.88	-3.53
Intercept	0.9	(16.76, 18.16)	(16.64, 18.42)	17.41	17.38
Age		(-1.28, 0.50)	(-1.93, -0.06)	-0.24	-0.86
Race 1		(-6.70, -2.41)	(-6.85, -2.48)	-4.77	-6.08

Results appear in Table 2. Both methods show the median and 90% percentile for weight loss are significantly affected by race, which can be interpreted as whites generally tend to lose more weight than blacks and furthermore, this differential becomes larger when comparing the most successful (highest) weight losers (90% percentile). The results for 'age' parameter in 90% quantile regression differ between the two approaches. Our approach indicates the relationship between weight loss and age in terms of 90% percentile is not significant, while the traditional frequentist method shows the age did affect weight loss. **also more specifically comment on the large differences in estimates between the .5 and .9 quantile and between the qreg and pt approaches** *Minzhao: Estimates for intercept in median and 90% percentile model from both methods are almost the same. Other coefficients estimates ('age' and 'race') keep the same sign. However, when making inference of comparison between median and 90 % percentile model, Pólya tree model tends to smooth the coefficient change (-0.41 to -0.24, -3.88 to -4.77), while the 'qreg' method would 'model' the error with completely different distributions analog to asymmetric Laplace distribution.*

Figure 2 shows the residuals and posterior mean of the probability density function of ϵ, $\hat{f}(\epsilon)$. We can see our approach correctly captures the small minor mode on the right tail. Thus, if upper quantile of the response is of interest, our approach would result in more accurate estimation of the quantile regression parameter than other methods. In addition, the right skewness is also captured by our approach.

5 Discussion

This paper introduced a Bayesian approach for linear quantile regression model simultaneously by introducing mixture of Pólya tree priors and estimating heterogeneity parameters. By marginalizing the predictive density function of the Pólya tree distribution, quantiles of interest can be obtained in closed form by inverting the predictive cumulative density function. Exact posterior inference can be made via MCMC. Here, quantile lines cannot cross since quantiles are estimated through density estimation. The simulations show our method performs better than the frequentist approach especially when the error is multimodal and

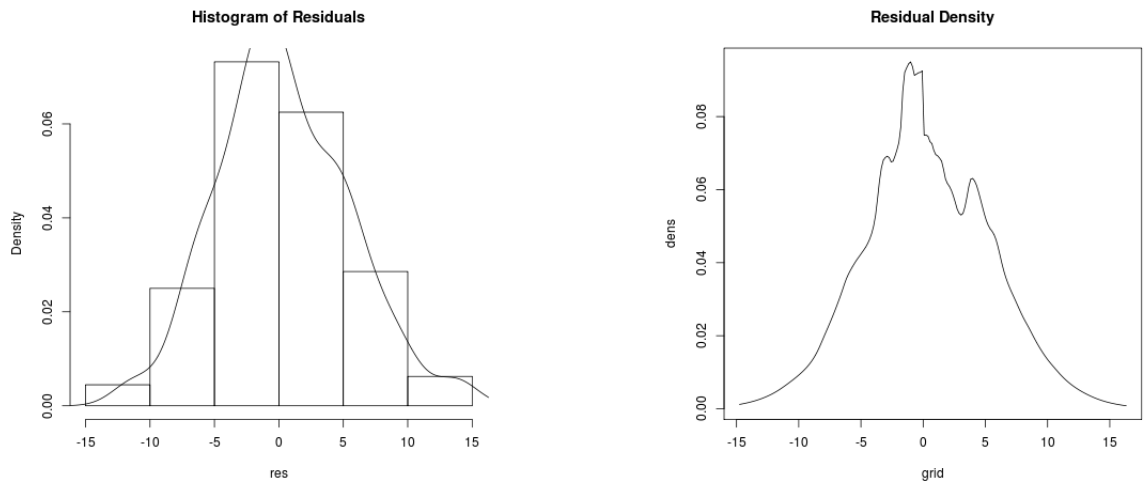

Figure 2: Estimated residuals ($r_i = (y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})/(\boldsymbol{x}_i'\hat{\gamma})$), where $\hat{\boldsymbol{\beta}}, \hat{\gamma}$ are the posterior medians. The left figure shows the histogram of the residuals, and the right one illustrates the estimated predictive density function, where the predictive density function is estimated by averaging predictive Pólya tree distribution density over MCMC iterations.

highly skewed. We also applied and illustrated our approach on the Tours data exploring the relationship between quantiles of weight loss and age and race.

Further research includes quantile regression for correlated data by modelling error as a mixture of multivariate Pólya tree distribution and shrinking the heterogeneity coefficients to zero for increased efficiency (Minzhao: just set γ prior narrow and close to zero ?; **might want spike and slab priors here**). *Minzhao: Spike and slab priors can accommodate zero-inflated situation and specialists' priors to help variable selection.* Our approach allows for quantile regression with missing data under ignorability by adding a data augmentation step. We are exploring extending our approach to allow for nonignorable missingness (**ref. Biometrics paper here**). Also spatial quantile regression (**ref. Gelfand here**) with Pólya tree might be a promising area for further exploration.

References

J.O. Berger and A. Guglielmi. Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. Journal of the American Statistical Association, 96(453):174–184, 2001.

Moshe Buchinsky. Changes in the u.s. wage structure 1963-1987: Application of quantile regression. Econometrica, 62(2):pp. 405–458, 1994. ISSN 00129682. URL http://www.jstor.org/stable/2951618.

Moshe Buchinsky. Recent advances in quantile regression models: A practical guideline for empirical research. The Journal of Human Resources, 33(1):pp. 88–126, 1998. ISSN 0022166X. URL http://www.jstor.org/stable/146316.

J. Fabius. Asymptotic behavior of bayes' estimates. The Annals of Mathematical Statistics, 35(2):846–856, 1964.

T.S. Ferguson. Prior distributions on spaces of probability measures. The Annals of Statistics, pages 615–629, 1974.

D.A. Freedman. On the asymptotic behavior of bayes' estimates in the discrete case. The Annals of Mathematical Statistics, 34(4):1386–1403, 1963.

J. Geweke and M. Keane. Smoothly mixing regressions. Journal of Econometrics, 138(1): 252–290, 2007.

T. Hanson and W.O. Johnson. Modeling regression error with a mixture of polya trees. Journal of the American Statistical Association, 97(460):1020–1033, 2002.

T.E. Hanson. Inference for mixtures of finite polya tree models. Journal of the American Statistical Association, 101(476):1548–1565, 2006.

X. He, P. Ng, and S. Portnoy. Bivariate quantile smoothing splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(3):537–550, 1998. ISSN 1467-9868. doi: 10.1111/1467-9868.00138. URL http://dx.doi.org/10.1111/1467-9868.00138.

N.L. Hjort and S. Petrone. Nonparametric quantile inference using dirichlet processes. Advances in statistical modeling and inference, pages 463–492, 2007.

N.L. Hjort and S.G. Walker. Quantile pyramids for bayesian nonparametrics. The Annals of Statistics, 37(1):105–131, 2009.

A. Jara, T.E. Hanson, and E. Lesaffre. Robustifying generalized linear mixed models using a new class of mixtures of multivariate polya trees. Journal of Computational and Graphical Statistics, 18(4):838–860, 2009.

R. Koenker. Quantile regression, volume 38. Cambridge Univ Pr, 2005.

Roger Koenker and Jr. Bassett, Gilbert. Regression quantiles. Econometrica, 46(1):pp. 33–50, 1978. ISSN 00129682. URL http://www.jstor.org/stable/1913643.

Roger Koenker and Jose A. F. Machado. Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association, 94(448):pp. 1296–1310, 1999. ISSN 01621459. URL http://www.jstor.org/stable/2669943.

A. Kottas and A.E. Gelfand. Bayesian semiparametric median regression modeling. Journal of the American Statistical Association, 96(456):1458–1468, 2001.

A. Kottas and M. Krnjajić. Bayesian semiparametric modelling in quantile regression. Scandinavian Journal of Statistics, 36(2):297–319, 2009.

M. Lavine. Some aspects of polya tree distributions for statistical modelling. The Annals of Statistics, pages 1222–1235, 1992.

M. Lavine. More aspects of polya tree distributions for statistical modelling. The Annals of Statistics, pages 1161–1176, 1994.

R.D. Mauldin, W.D. Sudderth, and SC Williams. Polya trees and random distributions. The Annals of Statistics, pages 1203–1221, 1992.

S.M. Paddock. Randomized Polya trees: Bayesian nonparametrics for multivariate data analysis. PhD thesis, Duke University, 1999.

S.M. Paddock. Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse. Biometrika, 89(3):529–538, 2002.

B.J. Reich, H.D. Bondell, and H.J. Wang. Flexible bayesian quantile regression for independent and clustered data. Biostatistics, 11(2):337–352, 2010.

L. Scaccia and P.J. Green. Bayesian growth curves using normal mixtures with nonparametric weights. Journal of Computational and Graphical Statistics, 12(2):308–331, 2003.

M.A. Taddy and A. Kottas. A bayesian nonparametric approach to inference for quantile regression. Journal of Business and Economic Statistics, 28(3):357–369, 2010.

S. Tokdar and J.B. Kadane. Simultaneous linear quantile regression: A semiparametric bayesian approach. Bayesian Analysis, 6(4):1–22, 2011.

S.G. Walker and B.K. Mallick. Hierarchical generalized linear models and frailty models with bayesian nonparametric mixing. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 59(4):845–860, 1997.

Stephen Walker and Bani K. Mallick. A bayesian semiparametric accelerated failure time model. Biometrics, 55(2):477–483, 1999. ISSN 1541-0420. doi: 10.1111/j.0006-341X.1999.00477.x. URL http://dx.doi.org/10.1111/j.0006-341X.1999.00477.x.

Ying Wei, Anneli Pere, Roger Koenker, and Xuming He. Quantile regression methods for reference growth charts. Statistics in Medicine, 25(8):1369–1382, 2006. ISSN 1097-0258. doi: 10.1002/sim.2271. URL http://dx.doi.org/10.1002/sim.2271.

Keming Yu and Rana A. Moyeed. Bayesian quantile regression. Statistics & Probability Letters, 54(4):437 – 447, 2001. ISSN 0167-7152. doi: 10.1016/S0167-7152(01)00124-9. URL http://www.sciencedirect.com/science/article/pii/S0167715201001249.

Keming Yu, Zudi Lu, and Julian Stander. Quantile regression: applications and current research areas. Journal of the Royal Statistical Society: Series D (The Statistician), 52(3): 331–350, 2003. ISSN 1467-9884. doi: 10.1111/1467-9884.00363. URL http://dx.doi.org/10.1111/1467-9884.00363.

A Small Proofs

Proof for equation (1) :

$$\begin{aligned} \mathrm{p}(Y \leq \boldsymbol{X}'\boldsymbol{\beta}(\tau)) &= \mathrm{p}\left(\boldsymbol{x}'\boldsymbol{\beta} + \epsilon \leq \boldsymbol{x}'\boldsymbol{\beta} + F_{\epsilon}^{-1}(\tau)\right) \\ &= \mathrm{p}(\epsilon \leq F_{\epsilon}^{-1}(\tau)) \\ &= \tau. \end{aligned}$$

Proof for equation (3)

$$\begin{aligned} \mathrm{p}(Y \leq \boldsymbol{x}'\boldsymbol{\beta}(\tau)) &= \mathrm{p}\left(\boldsymbol{x}'\boldsymbol{\beta} + (\boldsymbol{x}'\gamma)\epsilon \leq \boldsymbol{x}'\boldsymbol{\beta} + (\boldsymbol{x}'\boldsymbol{\gamma})F_\epsilon^{-1}(\tau)\right) \\ &= \mathrm{p}\left((\boldsymbol{x}'\boldsymbol{\gamma})\epsilon \leq (\boldsymbol{x}'\boldsymbol{\gamma})F_\epsilon^{-1}(\tau)\right) \\ &= \mathrm{p}(\epsilon \leq F_\epsilon^{-1}(\tau)) \\ &= \tau. \end{aligned}$$

B Additional Simulation

Three models with other errors are also tested in simulations to check performance of our approach. First two models were using a skewed mixture of normal distribution error (Reich 2009) $\pi N(0,1) + (1-\pi)N(3,3)$, where $\pi \sim Bern(0.8)$. One of them is homogeneous and the other is heterogeneous with $\gamma = (1, -0.5, 0.5)$. The third example uses student t distribution with degree freedom of three. Also same heterogeneity parameters $\gamma = (1, -0.5, 0.5)$ were assigned to this model.

M1 $y_i = 1 + x_{i1} + x_{i2} + \epsilon_{i1}$,

M2 $y_i = 1 + x_{i1} + x_{i2} + (1 - 0.5x_{i1} + 0.5x_{i2})\epsilon_{i1}$,

M3 $y_i = 1 + x_{i1} + x_{i2} + (1 - 0.5x_{i1} + 0.5x_{i2})\epsilon_{i2}$,

where $\epsilon_{i1} \sim \pi N(0,1) + (1-\pi)N(3,3)$, $\pi \sim Bern(0.8)$, and $\epsilon_{i2} \sim t_3$.

Table 3: Result for additional simulations: mean squared error (reported as 100*average) and standard error (reported as 100*standard error) for each quantile regression method. 'BQR', 'PT', 'rq' stand for Reich's Bayesian flexible quantile regression method, our approach with Pólya tree and the frequentist method in R package 'quantreg'.

Model	τ	BQR	PT	rq
M1	0.5	1.87(0.22)	2.24(0.27)	2.33(0.32)
	0.9	8.77(1.01)	11.30(1.32)	17.27(1.81)
M2	0.5	7.11(0.92)	8.11(0.86)	11.64(1.49)
	0.9	39.41(4.09)	44.91(4.28)	103.0(10.52)
M3	0.5	6.12(0.63)	6.93(0.69)	7.23(0.69)
	0.9	18.38(2.06)	24.16(2.32)	38.80(4.69)

From table B, Reich's method ('BQR') is generally better than our approach ('PT') and our method beats the traditional frequentist method 'rq' function in terms of MSE.