Review of Simultaneous Quantile Regression Methods

Steven Xu[1]

July 25, 2020

Abstract

Quantile regression can be used to study covariate-effect on non-central parts of the response distribution. Fitting multiple quantile regressions to a set of quantile levels can offer a more comprehensive description of the conditional response distribution. However, when estimated separately, quantile curves might cross and the results are hard to draw inference from. Simultaneous quantile regression models a given subset or all quantiles jointly by imposing monotonicity constraints. This alleviates quantile crossing and can increase overall precision since relevant information is borrowed from adjacent estimates. In this paper, recent advancement in simultaneous quantile regression is reviewed, and comparison is made to other types of method that estimate non-crossing quantile curves. Simulation studies are conducted to compare the performance of reviewed different methods on estimating the quantile process for different settings, and a discussion on future research directions is provided.

Key words: Monotonicity constraints; Non-crossing; Simultaneous estimation.

[1]Department of Statistics, North Carolina State University, Raleigh, NC, 27695, USA

1 Introduction

Quantile regression (QR) models the statistical relationship between conditional quantiles of the response distribution and a set of covariates using linear or non-linear regression equation. It has become widely used in diverse areas to complement least-square regression when investigators are interested in covariate-effect on non-central parts of the response distribution (Koenker, 2005). For example, a physician might be interested in modeling the 0.05 conditional quantile of birth-weight distribution to understand the determining factors of underweight newborns (Abrevaya, 2001); a climatologist might be interested in modeling the 0.99 conditional quantile of wind speed distribution to study the behavior of tropical cyclones that may cause major damage (Jagger and Elsner, 2009). Fitting multiple QR to a set of quantile levels can also offer a more insightful description of the conditional response distribution, especially when the observed data display strong heteroscadasticity. However, separately estimated quantile curves might cross. That is, given a set of covariates the estimated median of the response variable might be larger than the estimated 0.6 quantile which makes inference impossible. Consequently, a huge literature have emerged on estimating non-crossing quantile curves.

QR is first introduced in the seminal paper by Koenker and Bassett Jr (1978) who cast sample quantile into a minimization problem with respect to check loss. By generalising this idea to a linear regression setting, they estimated the quantile dependent coefficient vector efficiently using linear programming. Their classical work is the basis for most of the subsequent frequentist literature on QR and motivated theoretical studies on asymptotic properties of QR estimators. Bayesian QR was pioneered by Yu and Moyeed (2001). Under a linear model setting, they assumed the error terms are i.i.d and follow asymmetric laplace density (ALD) by recognizing the equivalence between maximizing the ALD likelihood and minimizing the check loss. Subsequent work relaxed the initial ALD assumption with examples including nonparameteric formulation that accommodates heteroscedastic error (Kozumi and Kobayashi, 2011; Bernardi and Petrella, 2015). In non-parametric QR, splines and Gaussian process are commonly used in estimating flexible quantile curves conditioned on a few predictors (Koenker and Ng, 1992; Koenker and Ng, 1994; Thompson and Stander, 2011; Quadrianto and Buntine, 2009; Boukouvalas et al., 2012; Abeywardana and

Ramos, 2015), whereas state-of-art machine learning algorithms such as boosting (Zheng, 2012), random forest (Meinshausen, 2006) and feedforward neural network (Taylor, 2000; Cannon, 2011) are favored for their scalability to high-dimensional regression tasks.

If the objective is to provide a reasonable estimate of a single quantile curve, all the aforementioned methods will suffice. However, the full potential of QR lies in estimating multiple parts of the conditional quantile function (QF). This is of particular interest in many applications when the investigator wants to monitor how the effect of a covariate change across different quantiles of the response. For instance, Miranda et al. (2009) studied the varying effect of lead exposure and parental education on different quantiles of children's performance in school. Estimating the conditional QF on a dense grid points also allows one to build up a simpler but proper estimate of the conditional response distribution. A naive solution to estimate multiple conditional quantiles is to fit separate QR models to each quantile level of interest. However, because the different regression equations are solved independently, the correlation structure of the true conditional quantiles will not be retained. Furthermore, because no restriction on the independent estimates were placed a prior, separately fitted QRs will not take into account the natural ordering among different quantiles and will lead to estimated quantile curves of which two or more might cross (see Figure 1 for illustration). Quantile crossing violates basic probabilistic rules, since any valid conditional quantile function should be monotonically non-decreasing in the quantile level. It might also lead to impossible interpretation of results in practice, for example based on the estimated crossing quantile curves the investigator might be forced to conclude that an observation is below the 0.8 quantile but above 0.9 quantile. Quantile crossing is also a major concern when separately estimated quantile curves are used to construct prediction intervals. It should be noted that based on the consistency property of QR estimators, separately fitted quantile curves within a frequentist framework can lead to valid inference when the sample size is large enough. Lum et al. (2012) also showed that stochastic ordering of the separate Bayesian estimates can be established under ALD assumption. However, quantile crossing is common in experiments of smaller scale. For example, if the errors follow a heavy-tailed distribution, small sample size will result in scarce observation around quantile levels at the upper (or lower) end of the response distribution. This

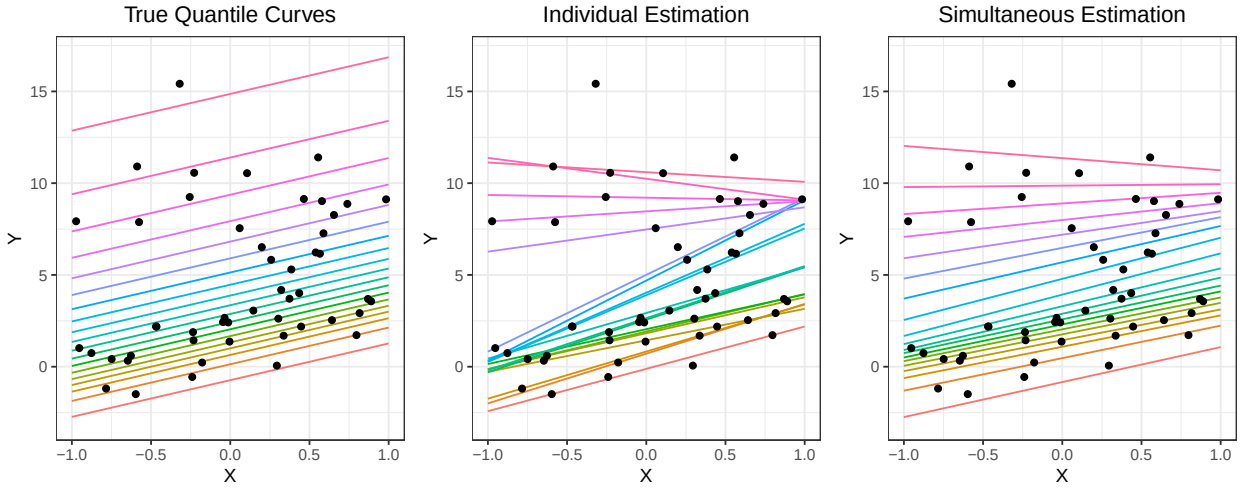
consequently increases the chance of quantile crossing at tail part of the response distribution. Separately fitted quantile curves also demonstrate a poor borrowing of information. For example, the significance of a covariate may change dramatically across neighbor quantiles, which contradicts the usual expectation that the conditional response distribution should be locally smooth. Quantile crossing persists and is more severe under a nonparametric regression setting, as the individually fitted curves become overly flexible and can easily overfit the data when the sample size is small.

Fortunately, many methods have been proposed to alleviate quantile crossing. These methods can be roughly classified into three categories: sequential estimation, post-processing and simultaneous estimation. In sequential estimation, different quantile curves are estimated sequentially under the constraint that the currently estimated curve should not cross the previous one. In post-processing, some adjustment are applied on the unconstrained QR estimates so that the monotonicity of the predicted conditional QF is enforced. Finally, in simultaneous estimation, quantile curves for a set of quantile levels are estimated jointly under the constraint that no two of them will cross each other. Instead of modeling the conditional QF, non-crossing quantile curves can also be estimated by inverting an estimated conditional cumulative distribution function (CDF).

The aim of this paper is to provide a comprehensive summary and comparison of different methods that can be used to estimate non-crossing quantile curves. Some representative methods will be presented in detail while others will be reviewed briefly. For each of the representative methods, we hope to cover details including but not limited to its motivation, computation and limitation. Through this review, we hope to summarize the similarity and distinction between different methods as well as understand the advantage and disadvantage of each type of methods under different settings. Finally, we would like to identify potential gaps within the literature so that suggestion can be made on the direction of future research in the area of quantile regression.

The remainder of the paper proceeds as follows. Notations that will be used consistently through out the paper are introduced in Section 2. Sequential estimation methods will be reviewed in Section 3 followed by post-processing methods in Section 4 and simultaneous methods in Section 5. Section 6 concludes and provides some discussion.

Figure 1: **Comparison of individually and simultaneously estimated quantile curves**. The data are generated from $y = 1 + 2x + \epsilon$ with $n = 50$, $x \sim \mathcal{U}(-1, 1)$ and $\epsilon \sim \mathcal{ALD}(\mu = 0, \sigma = 1, p = 0.2)$. Individual and simultaneous estimates are calculated using the method of Koenker and Bassett Jr (1978) and Yang and Tokdar (2017) respectively. Crossing is severe when quantile curves are estimated individually but is alleviated when they are estimated simultaneously. Simultaneous estimation also leads to significant improvement in overall precision by borrowing information across adjacent quantiles.



2 Notation

Consider a sample of n data points $\{(\boldsymbol{x}_i, y_i), i = 1, .., n\}$ where $y_i \in \mathbb{R}$ is the univariate response and $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T \in \mathcal{D}(\boldsymbol{x}) \subset \mathbb{R}^p$ is a p-dimensional covariate vector with domain $\mathcal{D}(\boldsymbol{x}) = \mathcal{X}$. Denote $(1, \boldsymbol{x}_i^T)^T$ by $\boldsymbol{z}_i$, $[\boldsymbol{x}_1|\boldsymbol{x}_2|...|\boldsymbol{x}_p]$ by $\boldsymbol{X}$ and $(y_1, ..., y_n)^T$ by $\boldsymbol{y}$. The conditional τth quantile $Q(\tau|\boldsymbol{x})$ is defined as the function satisfying the relationship Prob$(y \leq Q(\tau|\boldsymbol{x})) = \tau \in (0, 1)$, where consideration of $Q(0|\boldsymbol{x})$ and $Q(1|\boldsymbol{x})$ is omitted by assuming that y is unbounded. Given a set of unique quantile levels $0 < \tau_1 < \tau_2 < ... < \tau_K < 1$, we would like to estimate $\{Q(\tau_k|\boldsymbol{x}), 1 \leq k \leq K\}$ under the non-crossing constraint $\hat{Q}(\tau_1|\boldsymbol{x}) < \hat{Q}(\tau_2|\boldsymbol{x}) < ... < \hat{Q}(\tau_K|\boldsymbol{x})$. If the relationship between y and $\boldsymbol{x}$ is assumed to be linear, then $Q(\tau|\boldsymbol{x}) = \boldsymbol{z}^T\boldsymbol{\beta}(\tau)$ where $\boldsymbol{\beta}(\tau)$ is the coefficient vector as a function of τ. When necessary, we will also use $\boldsymbol{\beta}_{\tau_k}$ to denote the function value $\boldsymbol{\beta}(\tau_k)$ for $1 \leq k \leq K$.

3 Sequential Estimation

In their original work, Koenker and Bassett Jr (1978) proposed to solve the minimization problem

$$\min_{Q(\tau|\cdot)\in\mathcal{F}} \sum_{i=1}^{n} \rho_\tau(y_i - Q(\tau|\boldsymbol{x})) \tag{1}$$

for a pre-specified τ, where $\rho_\tau(u) = u(\tau - \mathbb{1}(u < 0))$ is the check loss function and $\mathcal{F}$ is the class of candidate functions. Solving (1) separately for $\{\tau_k, 1 \leq k \leq K\}$ will lead to quantile crossing since no constraint is imposed to ensure that $\hat{Q}(\tau_1|\boldsymbol{x}) < \hat{Q}(\tau_2|\boldsymbol{x}) < ... < \hat{Q}(\tau_K|\boldsymbol{x})$. Therefore, a straightforward solution seems to be estimating $\{Q(\tau_k|\boldsymbol{x}), 1 \leq k \leq K\}$ sequentially while ensuring that the currently estimated conditional quantile obeys its natural ordering relative to the previous one. This is the fundamental motivation of sequential estimation methods of which a representative work is Muggeo et al. (2013).

Under a linear setting, solving (1) is equivalent to solving

$$\min_{\boldsymbol{\beta}_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{z}_i^T \boldsymbol{\beta}_\tau). \tag{2}$$

For any pair of adjacent quantile levels $\tau_{k'} > \tau_k$, suppose that we have obtained $\hat{Q}(\tau_k|\boldsymbol{x}) = \boldsymbol{z}^T \hat{\boldsymbol{\beta}}_{\tau_k}$. Then estimating $Q(\tau_{k'}|\boldsymbol{x})$ under the restriction $\hat{Q}(\tau_{k'}|\boldsymbol{x}) > \hat{Q}(\tau_k|\boldsymbol{x})$ is equivalent to solving the constrained minimization problem

$$\begin{aligned} &\min_{\boldsymbol{\beta}_{\tau_{k'}}} \sum_{i=1}^{n} \rho_{\tau_{k'}}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_{\tau_{k'}}) \\ &\text{s.t. } \boldsymbol{z}^T \boldsymbol{\beta}_{\tau_{k'}} \geq \boldsymbol{z}^T \hat{\boldsymbol{\beta}}_{\tau_k}, \forall \boldsymbol{x} \in \mathcal{X}. \end{aligned} \tag{3}$$

Assume that $\mathcal{X} = [0,1]^p$, then a simple yet sufficient condition to $\boldsymbol{z}^T \boldsymbol{\beta}_{\tau_{k'}} \geq \boldsymbol{z}^T \hat{\boldsymbol{\beta}}_{\tau_k}$, $\forall \boldsymbol{x} \in [0,1]^p$ is then $\boldsymbol{\beta}_{\tau_{k'}} \geq \hat{\boldsymbol{\beta}}_{\tau_k}$ [1]. Therefore, (3) can be further simplifed to a minimization problem under

[1] The inequality is element-wise, i.e. $\beta_{j,\tau_{k'}} \geq \hat{\beta}_{j,\tau_k}$ $\forall\, 0 \leq j \leq p$

standard linear inequality constraint

$$\min_{\boldsymbol{\beta}_{\tau_{k'}}} \sum_{i=1}^{n} \rho_{\tau_{k'}}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_{\tau_{k'}}) \text{ s.t. } \boldsymbol{\beta}_{\tau_{k'}} \geq \hat{\boldsymbol{\beta}}_{\tau_k}. \tag{4}$$

It actually turns out that (4) is a special case of

$$\min_{\boldsymbol{\beta}_{\tau}} \sum_{i=1}^{n} \rho_{\tau}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_{\tau}) \text{ s.t. } \boldsymbol{R}\boldsymbol{\beta}_{\tau} \geq \boldsymbol{r} \tag{5}$$

if we let $\boldsymbol{R} = \boldsymbol{I}$ and $\boldsymbol{r} = \hat{\boldsymbol{\beta}}_{\tau_k} + \delta_0 \mathbf{1}$ where δ_0 is a non-negative number that can be set to positive to if strict monotonicity of the coefficient functions is required. Minimization problem of type (5) has been well studied (Koenker 2005) and implemented in the celebrated R package `rq`; optimization detail is not the focus of this paper and hence is omitted here. Therefore, a simple framework for estimating a set of non-crossing quantile planes $\{Q(\tau_k|\boldsymbol{x}), 1 \leq k \leq K\}$ is to fit an unconstrained quantile regression towards an initial quantile level τ_{k_0} followed by fitting (5) sequentially for each subsequent quantile level $\{\tau_k,\ k \neq k_0\}$. Detail of this framework is summarized in Algorithm 1. Confidence intervals for the estimated quantile curves can also be obtained using bootstrapped estimates.

Algorithm 1: Sequential estimation (Muggeo et al., 2013)

Given $\tau_1 < \tau_2 < ... < \tau_K$, choose τ_{k_0} and δ_0;

$\tau_1 < \tau_2 < ... < \tau_K \overset{re-index}{\longrightarrow} \tau_{k_0-K_L} < ... < \tau_{k_0} < ... < \tau_{k_0+K_U}$ s.t. $K_L + K_U = K - 1$;

$\hat{\boldsymbol{\beta}}_{\tau_{k_0}} = \arg\min_{\boldsymbol{\beta}_{\tau_{k_0}}} \sum_{i=1}^{n} \rho_{\tau_{k_0}}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_{\tau_{k_0}})$;

for $i = 1, ..., K_U$ **do**

  Set $\boldsymbol{R} = \boldsymbol{I}$, $\boldsymbol{r} = \hat{\boldsymbol{\beta}}_{k_0+i-1} + \delta_0 \mathbf{1}$;

  $\hat{\boldsymbol{\beta}}_{\tau_{k_0+i}} = \arg\min_{\boldsymbol{\beta}_{\tau_{k_0+i}}} \sum_{i=1}^{n} \rho_{\tau_{k_0+i}}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_{\tau_{k_0+i}})$ s.t. $\boldsymbol{R}\boldsymbol{\beta}(\tau_{k_0+i}) \geq \boldsymbol{r}$;

end

for $i = 1, ..., K_L$ **do**

  Set $\boldsymbol{R} = -\boldsymbol{I}$, $\boldsymbol{r} = -\hat{\boldsymbol{\beta}}_{k_0-i+1} + \delta_0 \mathbf{1}$;

  $\hat{\boldsymbol{\beta}}_{k_0-i} = \arg\min_{\boldsymbol{\beta}_{k_0-i} \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_{\tau_{k_0-i}}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_{k_0-i})$ s.t. $\boldsymbol{R}\boldsymbol{\beta}_{k_0-i} \geq \boldsymbol{r}$;

end

Since each of the subsequent estimations either directly or indirectly depends on the unconstrained quantile regression towards the initial quantile level τ_{k_0}, choosing a good τ_{k_0} is crucial for the overall performance. The original authors recommended $\tau_{k_0} = 0.5$ which is reasonable as one would expect the data to be more dense around the median than around non-central parts. Liu and Wu (2009) also showed that the asymptotic variance of the unconstrained quantile regression estimator is minimized at the median, which provides some theoretical support for this choice. One immediate drawback is that (4) is not the necessary condition to (3); (4) restricts that each coefficient function $\beta_j(\tau), 0 \leq j \leq p$ has to be a strictly increasing function and therefore puts a heavy restriction on the shape of the quantile planes. A more sophisticated minimization problem without making this assumption is presented in Liu and Wu (2009) who at each step identify the vertices that most likely violate the non-crossing constraint and iteratively finds a solution that achieves non-crossing for each of the found vertices. They also proposed a non-linear version by imposing sequential constraints on a kernel estimator. Nonetheless the straightforwardness of its computation and its effectiveness on alleviating quantile crossing makes (4) and Algorithm 1 useful in many simple settings.

The regression problem in (3) can be extended to accommodate non-linear relationship between y and $\boldsymbol{X}$. When $p = 1$, consider expanding x by B-splines of degree d and degrees of freedom J. Then the conditional quantile function $Q(\tau|x)$ can be represented by a linear combination of J spline covariates $\boldsymbol{B}_d(x)^T\boldsymbol{\alpha}(\tau)$, where $\boldsymbol{\alpha}(\tau) \in \mathbb{R}^J$ is the τ-dependent basis coefficient vector and $\boldsymbol{B}_d(\cdot) = (B_{d,1}(\cdot), ..., B_{d,J}(\cdot))$ are the J basis functions. Since $B_{d,j}(u)$ is a mapping from $\mathcal{D}(u)$ to $[0, 1]$ for $1 \leq j \leq p$, non-crossing quantile curves for any pair of adjacent quantile levels $\tau_{k'} > \tau_k$ can be estimated by a constrained minimization problem similar to that of (4)

$$\min_{\boldsymbol{\alpha}_{k'}} \sum_{i=1}^{n} \rho_{\tau_{k'}}(y_i - \boldsymbol{B}(x_i)^T\boldsymbol{\alpha}_{k'}) \text{ s.t. } \boldsymbol{\alpha}_{k'} \geq \hat{\boldsymbol{\alpha}}_k, \tag{6}$$

which can then be solved by Algorithm 1. When $p \geq 2$, one could represent the conditional quantile surface $Q(\tau|\boldsymbol{x})$ using a tensor product of B-spline basis functions

$$Q(\tau|\boldsymbol{x}) = \sum_{j_1}^{J_1} ... \sum_{j_p}^{J_p} \alpha_{j_1,...,j_p}(\tau) B_{j_1,d}(x_1)...B_{j_p,d}(x_p)$$

and follow the same steps of (4) and Algorithm 1. However, this is not recommended in general as the number of parameters is grows exponentially with p.

In summary, sequential estimation offers a simple modification of the classical estimation problem. To initialize, an unconstrained median regression is fitted; at each subsequent step, the desired conditional quantile is estimated under a non-crossing constraint formed by the previous estimate. Although this type of methods alleviates quantile crossing, an obvious drawback arises from its sequential nature: each estimated quantile curve only borrows information from its preceding estimate, but the former might also contain information that would have been helpful in estimating latter. Although in some cases the relative flexibility of each sequentially fitted quantile curve might be preferred, but when the data size is small information-borrowing across all quantile levels will improve overall performance. Moreover, although setting the initial quantile level to be the median has theoretical support in the asymptotic case, no guarantee can be made on its reliability in finite sample setting.

4 Post-processing

Rather than imposing constraint on each subsequent estimator and solve for each conditional quantile sequentially starting from the median, another approach to remedy quantile crossing is to post-process the independently fitted quantile curves such that the adjusted conditional quantile function is monotonically increasing in τ. This has the advantage that the final estimate will not depend on the order of intial estimations.

Chernozhukov et al. (2010) proposed to estimate the conditional cumulative distribution by

$$\tilde{F}(y|\boldsymbol{x}) = \int_0^1 \mathbb{1}\{\hat{Q}(u|\boldsymbol{x}) \geq y\} du \tag{7}$$

where $\hat{Q}(\tau|\boldsymbol{x})$ is a preliminary approximation of the true conditional quantile function and possibly violated the monotonicity property; one of such approximation could be found by interpolating the independent estimates $\{\hat{Q}(\tau_k|\boldsymbol{x}), 1 \leq k \leq K\}$. It is obvious that (7) is monotonically increasing in y. Therefore the monotonicity adjusted conditional quantile function $\tilde{Q}(\tau|\boldsymbol{x})$ can be obtained by inversion $\tilde{Q}(\tau|\boldsymbol{x}) = \inf\{y : \tilde{F}(y|x) \geq \tau\}$. However, the resulting estimate is not smooth which reduces its interpretability in practice. Moreover, (7) might jeopardize any assumed structure there might be between y and $\boldsymbol{x}$ – the linearity of $\hat{Q}(\tau|\boldsymbol{x})$ in $\boldsymbol{x}$ cannot be inherited by $\tilde{Q}(\tau|\boldsymbol{x})$. The idea of monotonizing the conditional quantile function is also adopted in Dette and Volgushev (2008) who proposed a Nadaraya – Watson estimator for constructing non-crossing non-parametric quantile curves. However, they only provided solution in the case of a single covariate. Recently, Rodrigues and Fan (2017) provided a two-step solution that can estimate both linear and non-linear non-crossing quantile curves for arbitrary p. In the first stage, they fit Bayesian quantile regression with ALD likelihood separately for each quantile level of interest; in the second stage, a Gaussian process regression adjustment is applied to the initial unconstrained estimates so that the final estimate of the conditional quantile function is monotonically increasing in τ. Some key results of their work is reviewed below.

Consider modeling the generic relationship $y_i = Q(\tau|\boldsymbol{x}_i) + \epsilon_i,\ 1 \leq i \leq n$ where ϵ_i are independently distributed from an error density f_ϵ with the only restriction $Q(\epsilon_i|\boldsymbol{x}_i) = 0$. Yu

and Moyeed (2001) proposed to set f_ϵ to be ALD having density $f(u) = \frac{s(1-s)}{\sigma} \exp\left\{-\rho_s\left(\frac{u-\mu}{\sigma}\right)\right\}$ where μ, σ and s are the location, scale and skewness parameter respectively. This leads to an approximate likelihood for Bayesian quantile regression

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{\mu}(\boldsymbol{X}), \sigma, s) = \frac{s^n(1-s)^n}{\sigma} \exp\left\{-\sum_{i=1}^{n} \rho_s\left(\frac{y_i - \mu(\boldsymbol{x}_i)}{\sigma}\right)\right\} \tag{8}$$

where the location parameter $\mu(\boldsymbol{x}_i)$ is simply the conditional quantile of interest $Q(\tau|\boldsymbol{x}_i)$. Following (8), $Q(\tau|\boldsymbol{x}_i)$ can be estimated by setting $s = \tau$ and obtain the posterior mode of $\mu(\boldsymbol{x}_i)$. For example, if MCMC is used to estimate the posterior distribution, a reasonable estimate for $Q(\tau|\boldsymbol{x}_i)$ could be the posterior mean

$$\hat{Q}(\tau|\boldsymbol{x}_i) = \frac{1}{M}\sum_{m=1}^{M} \tilde{Q}^{(m)}(\tau|\boldsymbol{x}_i, s = \tau) \tag{9}$$

where $\tilde{Q}^{(m)}(\tau|\boldsymbol{x}_i, s = \tau)$ is the posterior sample of $Q(\tau|\boldsymbol{x}_i)$ at mth iteration and $s = \tau$ emphasized that the likelihood being maximized is $ALD(s = \tau)$. One can repetitively carry out (8) and (9) for each of $\{\tau_k, 1 \leq k \leq K\}$ to estimate multiple quantile curves. However, since for each τ_k a different likelihood is maximized, the independently estimated quantile curves might cross.

Given that adjacent quantiles are strongly correlated, we should expect that the posterior sample $\{\tilde{Q}^{(m)}(\tau'|\boldsymbol{x}_i, s = \tau'), 1 \leq m \leq M\}$ obtained from maximizing ALD($s = \tau'$) also contain useful information for $Q(\tau|\boldsymbol{x}_i)$ if τ' and τ are close. Therefore, based on the quantile function of ALD(μ, σ, s) (Yu and Zhang, 2005)

$$Q(\tau|\mu, \sigma, s) = \begin{cases} \mu + \frac{\sigma}{1-s}\log\left(\frac{\tau}{s}\right), \text{ if } 0 \leq \tau \leq s \\ \mu - \frac{\sigma}{s}\log\left(\frac{1-\tau}{1-s}\right), \text{ if } s \leq \tau \leq 1 \end{cases}, \tag{10}$$

one can construct a $(K-1) \times M$ induced posterior sample for $Q(\tau|\boldsymbol{x}_i)$ from $\tilde{Q}^{(m)}(\tau|\boldsymbol{x}_i, s = \tau')$ for all $\tau' \neq \tau$. The resulting $K \times M$ matrix (adding the $1 \times M$ sample obtained from maximizing ALD($s = \tau$)) can be seen as M induced samples for the conditional QF $Q(\tau|\boldsymbol{x}_i)$, therefore an improved estimator for $Q(\tau|\boldsymbol{x}_i)$ can be obtained by applying Gaussian process regression to the

M samples. Following the notation of original authors, the regression problem can be set up as

$$\begin{aligned} \tilde{Q}^{(m)}(\tau|\boldsymbol{x}_i, s) &= g(s) + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \boldsymbol{\Sigma}_i) \\ g(s) &\sim \mathcal{GP}(0, \boldsymbol{K}) \end{aligned} \tag{11}$$

where $\boldsymbol{\Sigma}_i$ and $\boldsymbol{K}$ are covariance matrices both of dimension $(K \times M, K \times M)$. Given that $\tilde{Q}^{(m)}(\tau|\boldsymbol{x}_i, s)$ and $\tilde{Q}^{(m)}(\tau'|\boldsymbol{x}_i, s)$ are estimated using separate MCMC chains, their correlation is zero for all $\tau \neq \tau'$ when conditioned on the data. We can further assume $\tilde{Q}^{(m)}(\tau|\boldsymbol{x}_i, s)$ and $\tilde{Q}^{(m')}(\tau|\boldsymbol{x}_i, s)$ are uncorrelated for all $m \neq m'$ by assuming they are distant in the MCMC chain. This leads to a diagonal $\boldsymbol{\Sigma}_i$ with diagonal entries containing the variance of each induced estimate. The formulation of $\boldsymbol{K}$ should reflect the assumption that relevant information carried by nearby induced quantiles increase as the distance decrease, which can be represented by the squared exponential kernel $k(s, s') = \sigma_k^2 \exp\left\{-\frac{(s-s')^2}{2b^2}\right\}$, where b is the bandwidth and σ_k^2 is the variance hyperparameter. The predictive posterior distribution of $\tilde{Q}^{(m)}(\tau|\boldsymbol{x}_i, s)$ given (11) follows a normal distribution (see details in Rasmussen and Williams, 2006) and therefore the final adjusted estimate of $Q(\tau|\boldsymbol{x}_i)$ is simply the predictive posterior mean

$$\begin{aligned} \hat{Q}_a(\tau|\boldsymbol{x}_i) &= \sum_{k=1}^{K} \sum_{m=1}^{M} w_k \tilde{Q}^{(m)}(\tau|\boldsymbol{x}_i, \tau_k), \\ \boldsymbol{w} &= \boldsymbol{K}(., \tau)^T (\boldsymbol{K} + \boldsymbol{\Sigma}_i)^{-1} \end{aligned} \tag{12}$$

where $\boldsymbol{K}(., \tau)$ is any column vector of $\boldsymbol{K}$ with $s' = \tau$ and $\boldsymbol{w} = (w_1, ..., w_{K \times M})$ is a weight row vector. Since each of the induced QF $\tilde{Q}^{(m)}(\tau|\boldsymbol{x}_i, \tau_k)$ is monotonically increasing in τ, the monotonicity constraints on the grid points $\{\hat{Q}_a(\tau_k|\boldsymbol{x_i}), 1 \leq k \leq K\}$ depend on the weights $\boldsymbol{w}$ and is thus guarded by the bandwidth parameter b. Rodrigues and Fan (2017) showed that for any set of quantile levels $\{\tau_k, 1 \leq k \leq K\}$, there always exists b such that monotonicity constraints are satisfied. In fact, a trivial solution is to set $b \longrightarrow \infty$ so that the adjusted estimate is just the average of the induced conditional QFs. In practice, one can search for the smallest b that guarantees $\{\hat{Q}_a(\tau_1|\boldsymbol{x_i}) < ... < \hat{Q}_a(\tau_K|\boldsymbol{x_i})\}$ for every $\boldsymbol{x}_i$. Similar to other post-processing methods

mentioned before, the Gaussian regression adjustment can also jeopardize the potentially linear relationship between y and $\boldsymbol{x}$, as the weights in (12) depends on $\boldsymbol{x}$. To retain interpretability of the initial estimates, one can approximate $\boldsymbol{\Sigma}_i$ with a covairance matrix $\tilde{\boldsymbol{\Sigma}}$ that is constant with respect to i; the original authors showed that setting $\tilde{\boldsymbol{\Sigma}} = 1/n \sum_{i=1}^{n} \boldsymbol{\Sigma}_i$ performs well. As for theoretical properties, the authors proved that the adjusted estimator achieves posterior consistency as long as the unadjusted estimator is consistent, and posterior consistency of the standard Bayesian quantile regression with ALD density has been well studied (Sriram et al., 2013).

In summary, post-processing methods all start with independent estimates obtained from fitting standard QRs to each of the quantile levels of interest; the initial estimates are then rearranged so that the final estimates of the quantile curves do not cross. One general drawback of the post-processing methods is that the performance of the final estimates depend on the initial estimates, which could be poor as they do not borrow information from each other. Computation might be another concern for the method of Rodrigues and Fan (2017), as their Gaussian process adjustment in addition to the initial MCMCs can be particularly expensive when K is large.

5 Simultaneous Estimation

Both sequential estimation and post-processing methods can effectively ensure non-crossing of the resulting quantile curves. However, they still rely on the estimates obtained from unconstrained QR to some extend. In sequential estimation, all subsequent estimates depend on the unconstrained QR towards median; in post-processing, unconstrained estimates are adjusted and thus all contribute to the final estimate. In contrast, simultaneous estimation methods model all the desired quantile curves jointly. Therefore, each estimated quantile curve is somewhat regularized by its adjacent estimates, which can especially improve overall performance when the data size is small.

5.1 Composite Quantile Regression

It is worth mentioning that the idea of simultaneously fitting multiple QR models have been proposed in a relevant context. Composite QR was proposed by Zou and Yuan (2008) who considered

solving the minimization problem

$$\min_{f \in \mathcal{F}} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - f(\boldsymbol{x}_i)) \tag{13}$$

to estimate the true underlying relationship $y = f(\boldsymbol{x}) + \epsilon$. The idea is to borrow information from different quantile regression models by summing the check loss over a set of quantile levels $\{\tau_k, 1 \leq k \leq K\}$. Although the motivations of composite QR and simultaneous estimation are somewhat similar in that they both seek the information-borrowing property of joint modeling, the objectives of the two contexts genuinely differs from each other. In (13), the function to be estimated does not depend on τ. Therefore, the goal of composite QR is to come up with one τ-independent model that best characterize the true underlying relationship between the covariates and the response rather than describing the conditional quantile function. However, we will observe later that (13) can be naturally extended to estimate multiple quantile curves, and by imposing suitable constraints the jointly estimated quantile curves will not cross.

5.2 Simultaneous Quantile Regression

An imperfect attempt was made by Takeuchi et al. (2006) who enforced non-crossing constraint for all quantile levels on a subset of data points in $\mathcal{X}$. However, this only reduces the chance of quantile crossing and does not guarantee non-crossing for every $\boldsymbol{x} \in \mathcal{X}$; computation might also become infeasible when n and K are large. Later, a more concrete and general approach is proposed by Bondell et al. (2010) who applied direction correction to the classical minimization problem in (1). Since their method motivates several subsequent proposals, its key ideas will be reviewed below.

Consider the problem presented in (2). Then the point estimates $\{\hat{\boldsymbol{\beta}}_{\tau_k}, 1 \leq k \leq K\}$ satisfying the monotonicity constraint $\boldsymbol{z}^T \hat{\boldsymbol{\beta}}_{\tau_k} \geq \boldsymbol{z}^T \hat{\boldsymbol{\beta}}_{\tau_{k-1}}$ for every $\boldsymbol{x} \in \mathcal{X}$ and $2 \leq k \leq K$ is the solution to

the minimization problem

$$\begin{aligned} &\min_{\boldsymbol{\beta}_\tau} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \boldsymbol{z}_i^T \boldsymbol{\beta}_{\tau_k}) \\ &s.t.\ \boldsymbol{z}^T \boldsymbol{\beta}_{\tau_k} \geq \boldsymbol{z}^T \boldsymbol{\beta}_{\tau_{k-1}}\ \forall\ \boldsymbol{x} \in \mathcal{X} \text{ and } 2 \leq k \leq K. \end{aligned} \tag{14}$$

Notice that (14) is extended from (13) to estimate K set of parameters under non-crossing constraints. Assume that $\mathcal{X}$ is the convex hull formed by T vertices $(\tilde{\boldsymbol{z}}_1, ..., \tilde{\boldsymbol{z}}_T)$. Let $\{w_t, 1 \leq t \leq T\}$ be a set of positive weights that sum to 1. Since any point inside the convex hull can be written as a weighted sum of the T vertices, then it suffices to set $\tilde{\boldsymbol{z}}_t^T \boldsymbol{\beta}_{\tau_k} \geq \tilde{\boldsymbol{z}}_t^T \boldsymbol{\beta}_{\tau_{k-1}}$ for every $1 \leq t \leq T$ and $2 \leq k \leq K$, which implies $\sum_{t=1}^{T} w_t \tilde{\boldsymbol{z}}_t^T \boldsymbol{\beta}_{\tau_k} \geq \sum_{t=1}^{T} w_t \tilde{\boldsymbol{z}}_t^T \boldsymbol{\beta}_{\tau_{k-1}}$, which is equivalent to the constraints in (14). This is again an example of the standard linear constraint $\boldsymbol{R}\boldsymbol{\beta}_\tau \geq r$ with $\boldsymbol{\beta}_\tau = (\boldsymbol{\beta}_{\tau_1}^T, ..., \boldsymbol{\beta}_{\tau_K}^T)^T$ and can be solved via linear programming. However, there are in total $T \times (K-1)$ constraints which can be large if a fine grid of quantile levels is of interest. Fortunately, the number of constraints can be greatly reduced if we consider $\mathcal{X}$ to be the unit hypercube $[0,1]^p$ in which case only $K-1$ rather than $2^p(K-1)$ constraints are necessary. Let $\boldsymbol{\gamma_{\tau_1}} = \boldsymbol{\beta_{\tau_1}}, \boldsymbol{\gamma_{\tau_k}} = \boldsymbol{\beta_{\tau_k}} - \boldsymbol{\beta_{\tau_{k-1}}}$ for $2 \leq k \leq K$ be a reparameterization and $\boldsymbol{\gamma}_{\tau_k}^+ = \max(\boldsymbol{\gamma}_{\tau_k}, 0)$, $\boldsymbol{\gamma}_{\tau_k}^- = \max(-\boldsymbol{\gamma}_{\tau_k}, 0)$ for $1 \leq k \leq K$, where $\max(\cdot)$ is the element-wise maximum operator. Then $\boldsymbol{\gamma}_{0,\tau_k} - \sum_{j=2}^{p} \boldsymbol{\gamma}_{j,\tau_k}^- \geq 0$ enforces non-crossing on the point that is most likely to violate the monotonicity constraint (worst case), having $x_j = \mathbb{1}(\tau_{j,\tau_k} < 0)$, for $2 \leq k \leq K$ and thus is a necessary and sufficient condition to the constraints in (14). Therefore, (14) can be reduced to minimization under a standard linear constraint which can be solved efficiently via linear programming. Inference of the constrained estimators can be carried out without difficulty. In fact, it has been proved that the constrained estimators in (14) share the same asymptotic properties as the classical estimators in (2). Thus standard errors and confidence intervals for the constrained estimators can be easily calculated using known formulas (Koenker, 2005). This approach is also extendable to model non-linear quantile curves using linear splines by setting knots at the data points and analogously enforce non-crossing constraint on the points that give the worst case scenarios.

Method based on (14) can effectively enforce non-crossing constraints on any finite collec-

tion of quantile levels $\{\tau_k, 1 \leq k \leq K\}$, which is also the objective of Muggeo et al. (2013) and Rodrigues and Fan (2017). However, estimating conditional quantiles that are not within the pre-specified set requires separate model fitting. Moreover, the estimate for a specific quantile level is sensitive to the collection chosen. That is, $\hat{Q}(0.8|\boldsymbol{x})$ could be different if constraints were put on quantile levels $\{0.2, 0.35, 0.5, 0.65, 0.8\}$ instead of $\{0.2, 0.5, 0.8\}$, which is not desired (see Section 8 for a sensitivity analysis). A more ideal model should be able to estimate any $\tau \in (\delta, 1-\delta), \delta \geq 0$ simultaneously, which is equivalent to estimating the quantile process $\{Q(\tau, \boldsymbol{x}) : \tau \in (\delta, 1-\delta), \boldsymbol{x} \in \mathcal{X}\}$ under the monotonicity constraint $\frac{\partial Q(\tau, \boldsymbol{x})}{\partial \tau} \geq 0, \ \forall \boldsymbol{x}$. Quantile process of a linear model has been modeled by many authors. He (1997) gave one of the earliest solutions by assuming a heteroscedastic regression model for the response; the covariates are constructed to affect the response distribution via location-scale change of a base distribution so that the quantile process is naturally monotonically increasing in τ. However, the model coverage of this approach is limited as the covariates might affect the response in a more complex way in practice. Several works have considered expanding the coefficient function $\boldsymbol{\beta}(\tau)$ using a finite number of basis functions. This includes Reich et al. (2011) who uses Bernstein basis polynomials, Reich (2012) who uses piece-wise Gaussian basis function, Reich and Smith (2013) who extends to basis function constructed from an arbitrary base function and Yuan et al. (2017) who uses low-rank B-splines. Among them, the first three use a similar idea to that of Bondell et al. (2010) and enforced monotonicity constraints at the worst case scenarios by intoducing latent unconstrained coefficients, whereas the latter put order constraints on the spline coefficients. Despite the differences, one assumption they have in common is that $\mathcal{X}$ is a hypercube, which is also assumed in Muggeo et al. (2013) and Liu and Wu (2009). This is not a coincidence, as two planes have to be parallel to not cross in an unbounded region, and the hypercube geometry often simplifies the otherwise complicated constraints. In practice, one can always apply an affine transformation (e.g. min-max scaling) to transformed the covariates into the suitable hypercube and then back-transform them after estimation. However, this assumption might impose heavy restriction on the shape of quantile curves when the dimension is high. Data that violate this assumption will reside in a small fraction of the volume of the encompassing hypercube, and therefore estimated quantile

curves that are constrained to be non-crossing in the hypercube will appear parallel in the original domain. Recently, Yang and Tokdar (2017) offered a solution to estimate non-crossing quantile planes within convex $\mathcal{X}$ of arbitrary shape; some detail of their proposal will be reviewed below.

To ensure the ordering of quantile planes $Q(\tau|\boldsymbol{x}) < Q(\tau'|\boldsymbol{x})$ for any two quantile levels $0 < \tau < \tau' < 1$, it suffices to let $\frac{\partial\beta_0(\tau)}{\partial\tau} + \frac{\partial\boldsymbol{x}^T\boldsymbol{\beta}(\tau)}{\partial\tau} \geq 0$ for all $\tau \in (0,1)$ and $\boldsymbol{x} \in \mathcal{X}$. Without loss of generality, assume $\mathbf{0}$ is an interior point in $\mathcal{X}$. This leads to the necessary condition that $\frac{\partial\beta_0(\tau)}{\partial\tau} > 0$, which is easy to construct. However, finding a particular formulation of $\boldsymbol{\beta}(\tau)$ that satisfies the aforementioned constraint remains challenging. For a single covariate, Tokdar et al. (2012) proposed to expand the slope function using linear combination of monotonically increasing functions, but their generalization to multivariate setting through a single index model was not satisfactory. Yang and Tokdar (2017) proposed an ingenious solution by first defining a particular mapping $\boldsymbol{b} \to a(\boldsymbol{b}, \mathcal{X})$ as

$$a(\boldsymbol{b}, \mathcal{X}) = \begin{cases} \sup_{\boldsymbol{x}\in\mathcal{X}}(\frac{-\boldsymbol{x}^T\boldsymbol{b}}{||\boldsymbol{b}||}), \text{if } \boldsymbol{b} \neq 0 \\ \infty, \text{if } b = 0 \end{cases} \tag{15}$$

where $\boldsymbol{b} \in \mathbb{R} \cup \{\infty\}$ and showed that together with some suitably chosen function $\boldsymbol{v}(\cdot) \in \mathbb{R}^p$, the coefficient function with derivative defined based on this mapping

$$\frac{\partial\boldsymbol{\beta}(\tau)}{\partial\tau} = \frac{\partial\beta_0(\tau)}{\partial\tau} \times \frac{\boldsymbol{v}(\tau)}{a(\boldsymbol{v}(\tau), \mathcal{X})\sqrt{1 + ||\boldsymbol{v}(\tau)||^2}} \tag{16}$$

induce correct ordering of the quantile planes. This greatly reduced the original non-crossing constraint to the shape restriction of the slope function. Following Tokdar et al. (2012), one can construct $\beta_0(\tau)$ based on some parametric guess of the error distribution f_ϵ (e.g. Normal or t-distribution for symmetric error). Let $Q_0(\tau)$ be the quantile function of such a guess, a well experimented choice is $\beta_0(\tau) = \sigma Q_0(\xi(\tau))$ where $\xi(\cdot)$ is the logistic transformation defined by

$$\xi(\tau) = \frac{\int_0^\tau e^{w_0(u)}du}{\int_0^1 e^{w_0(u)}du}, \tau \in (0,1), \tag{17}$$

and $w_0(\cdot)$ is assumed to follow a zero mean Gaussian process with double exponential covariance structure. This assumption has two main advantages. Since $w_0(\cdot)$ is centered at zero, (17) is centered at the identity mapping; if we further set $\boldsymbol{v}(\tau) = \mathbf{0}$ then (16) leads to an ordinary linear regression model with error distribution f_0. Moreover, because (17) is a continuous mapping it will inherit the capability of $w_0(\cdot)$ on estimating all (piece-wise) continuous increasing bijection from (0,1) to itself. This ensures that $\beta_0(\tau)$ is flexible enough to model a broad range of intercept functions. If we further put a Gaussian process prior on $\boldsymbol{v}(\cdot)$, then $\boldsymbol{\beta}(\tau)$ is flexible enough to model a broad range of slope functions. The standard Bayesian treatment then proceeds by writing out the likelihood function which can be numerically calculated using the equality

$$f(y|\boldsymbol{x}) = \left. \frac{1}{\frac{\partial}{\partial \tau} Q(\tau|\boldsymbol{x})} \right|_{\tau = \tau_{\boldsymbol{x}}(y)} \tag{18}$$

where $\tau_{\boldsymbol{x}}(y)$ solves the equation $Q(\tau|\boldsymbol{x}) = y$ for τ in y. Since the coefficient functions are defined through their derivatives, numerical integration followed by root searching can be applied to solve for $\tau_{\boldsymbol{x}_i}(y_i)$ for each $1 \leq i \leq n$. The original authors proposed to work with a low rank approximation of the Gaussian process prior for $\boldsymbol{v}(\cdot)$ using a piece-wise interpolation so that integration of (16) can be discretized. After suitable hyper-priors are placed, parameters are estimated using adaptive blocked Metropolis. The computation detail is well documented in the original paper and thus omitted here. For theoretical properties, the authors proved that their Gaussian process based estimator achieves weak posterior consistency under mild smoothness and tail conditions of the true data generating distribution.

In contrast to the abundant literature in estimating quantile process under linear assumption, few non-parametric approaches have been proposed for the following possible reasons. First, non-parametric quantile curves are less interested by investigators as interpretation of the covariates' effect is less straightforward. Secondly, imposing non-crossing constraints in a complex model might lead to computation bottleneck. For example, if $Q(\tau|\boldsymbol{x})$ has a complicated structure in a Bayesian model, then solving (18) for each data point will be extremely expensive thus negatively affecting the overall efficiency of MCMC.

Das and Ghosal (2018) proposed to model the conditional quantile function using linear combination of quadratic B-splines with $L-1$ equidistant knots

$$Q(\tau|\boldsymbol{x}) = \sum_{j=1}^{L+2} \theta_j(\boldsymbol{x}) B_{j,2}(\tau), \tag{19}$$

where $\theta_j(\boldsymbol{x}), 1 \leq j \leq L+2$ are spline coefficient functions that depend on $\boldsymbol{x}$. Since (19) is a quadratic B-spline, the necessary and sufficient condition for its monotonicity is then $\theta_j(\boldsymbol{x}) < \theta_{j+1}(\boldsymbol{x})$ for $1 \leq j \leq L+1$ (De Boor 2001, Beliakov 2002); the authors chose to use $0 = \theta_1(\boldsymbol{x}) < ... < \theta_{L+2}(\boldsymbol{x}) = 1$, which is only a sufficient condition. To accommodate non-linear relationship between y and $\boldsymbol{x}$, these coefficient functions are further expanded using tensor product of quadratic splines with the same number of knots, leading to the aggregated model

$$Q(\tau|\boldsymbol{x}) = \sum_{j=1}^{L+2} \left(\sum_{k_1=1}^{L+2} ... \sum_{k_p=1}^{L+2} \alpha_{j,k_1...k_p} B_{k_1,2}(x_1)...B_{k_p,2}(x_p) \right) B_{j,2}(\tau). \tag{20}$$

Following this formulation, the constraint $0 = \theta_1(\boldsymbol{x}) < ... < \theta_{L+2}(\boldsymbol{x}) = 1$ is equivalent to the constraint $0 = \alpha_{1,k_1...k_p} < ... < \alpha_{L+2,k_1...k_p} = 1$ for $1 \leq k_1...k_d \leq (L+2)^p$. To avoid estimating the parameters directly under the orderd constraints, the authors utilized a transformation $\gamma_{l,k_1...k_p} = \alpha_{l+1,k_1...k_p} - \alpha_{l,k_1...k_p}, l = 1, ..., L+1$ and put a uniform prior on each simplex block $\{\gamma_{l,k_1...k_p}, 1 \leq l \leq L+1\}$. When $p = 1$, (20) is closely related to the model of Yuan et al. (2017), who replaced $\sum_{k_1=1}^{L+2} B_{k_1,2}(x_1)$ with $\sum_{j=0}^{p} z_j \beta_{j,L+2}$ to model non-crossing planes. By using a quadratic B-spline, (19) offers flexibility and smoothness. Moreover, (18) boils down to solving a quadratic equation which has an analytical solution. However, this approach also has immediate drawbacks. First, the constraint considered is only a sufficient condition to the original constraint, which might greatly compromise the model coverage. Secondly, as mentioned before, tensor product of spline basis functions does not scale well to multivariate setting; in fact, the number of parameters in (20) is $L(L+2)^p$. In their original work, the authors used blocked Metrapolis and updated each of the $(L+1)^p$ simplex blocks one at a time. One could expect this computation quickly becomes impractical and might lead to serious convergence problem for even a moderate p. In the next

paragraph, we will look at a scalable solution that used a neural network estimator with partial monotonicity to non-parametrically estimate the quantile process.

Consider modeling the quantile curve using neural network with one hidden layer

$$Q(\tau|\boldsymbol{x}) = \xi\left(\sum_{\ell=1}^{L} w_{\tau,\ell}\phi\left(\sum_{j=1}^{p} W_{\tau,j\ell}x_j + b_{\tau,\ell}\right) + b_{\tau}^{0}\right), \tag{21}$$

where $\xi(\cdot)$ and $\phi(\cdot)$ are monotone activation functions, $\boldsymbol{w} \in \mathbb{R}^L$ and $\boldsymbol{W} \in \mathbb{R}^{L\times p}$ are weight coefficients and $\boldsymbol{b} \in \mathbb{R}^L$ and $b^0 \in \mathbb{R}$ are bias coefficients. For a fixed τ, (21) can approximate any continuous quantile curve at any accuracy. To enforce partial monotonicity of the estimated quantile curve on some covariate $x_{j^{(m)}}$, i.e. $\frac{\partial\hat{Q}(\tau|\boldsymbol{x})}{\partial x_{jm}} > 0$. One can simply modify (21) into

$$Q(\tau|\boldsymbol{x}) = \xi\left(\sum_{\ell=1}^{L} e^{w_{\tau,\ell}}\phi\left(e^{W_{\tau,j^{(m)}\ell}}x_{j^{(m)}} + \sum_{j\neq j^{(m)}}^{p} W_{\tau,j\ell}x_j + b_{\tau,\ell}\right) + b_{\tau}^{0}\right) \tag{22}$$

which can approximate continuous quantile curve that is monotonically increasing in $x_{j^{(m)}}$ at any accuracy Zhang and Zhang (1999). In order to approximate the quantile process, one would need $\frac{\partial\hat{Q}(\tau|\boldsymbol{x})}{\partial\tau} > 0$. Then an ingenious method proposed up by Cannon (2018) is to treat τ as an additional covariate and impose partial monotonicity constraint. First, construct the stacked covariate matrix and response vector

$$\boldsymbol{X}^{(s)} = \left[\begin{array}{c|c} \boldsymbol{\tau}_1 & \boldsymbol{X} \\ \vdots & \vdots \\ \boldsymbol{\tau}_K & \boldsymbol{X} \end{array}\right], \; \boldsymbol{y}^{(s)} = \begin{bmatrix} \boldsymbol{y} \\ \vdots \\ \boldsymbol{y} \end{bmatrix}$$

where $\boldsymbol{\tau}_k = \tau_k\mathbf{1}_n$ for $1 \leq k \leq K$. Then, plug $\boldsymbol{X}^{(s)}$ and $\boldsymbol{y}^{(s)}$ back into (22), replacing $x_{j^{(m)}}$ by τ, and solve the unconstrained minimization problem

$$\min_{\boldsymbol{W},\boldsymbol{w},\boldsymbol{b},b^0} \sum_{i^{(s)}=1}^{Kn} \rho_{\tau_{i^{(s)}}}(y_{i^{(s)}}^{(s)} - Q(\tau_{i^{(s)}}|\boldsymbol{x}_{i^{(s)}}^{(s)})). \tag{23}$$

In their original work, the authors used a smooth approximation of (23) and proceeds estimation with a standard non-linear gradient based optimization algorithm (`nlm` routine in R) in which

the gradients are calculated using backpropagation. To avoid convergence to a local minima, the optimization is run for a pre-specified number of times each with different set of initial values. In terms of uncertainty analysis, confidence interval of the estimated quantile curves can be obtained through means of bootstrap (Franke and Neumann, 2000). It is obvious that (21) can be easily extended to allow more than one hidden layers, which can be useful to model extremely complex relationship under high dimension setting. A general concern of using neural network is that it can easily overfit the data. In practice, one should tune the number of layers, hidden nodes, and include weight penalties when necessary. This can be done via cross-validation or using a goodness-of-fit criteria such as the quasi-AIC. One particular drawback of this approach is that any $Q(\tau|\boldsymbol{x})$, $\tau \notin [\tau_1, \tau_K]$ has to be estimated via extrapolation. Although non-crossing is still guaranteed, the result might not be reliable since performance on that specific quantile level is not taken into account by (23).

5.3 Conditional Density Estimation

The simultaneous estimation methods reviewed above all model the conditional QF directly. It is also possible to first estimate the conditional CDF and then analytically or numerically invert it to obtain an estimate of the condition QF; as long as the constructed model represents a valid CDF, estimates for any collection of quantile curves are guaranteed to not cross. This can be particularly appealing in a Bayesian framework, where the CDF model leads to straightforward calculation of the likelihood function. However, care is needed when modeling the conditional CDF as not only does the conditional CDF have to be monotonically increasing in y, but it also needs to satisfy the boundary condition $\lim_{y\to-\infty} F(y) = 0$ and $\lim_{y\to\infty} F(y) = 1$. Yu and Jones (1998) proposed to estimate the conditional CDF using local linear fitting with double kernel smoothing, but they only considered univariate $\boldsymbol{X}$. Das and Ghosal (2018) estimated the conditional CDF based on a model motivated by (24), but their model does not represent a valid CDF or scale to high dimension. As a non-parametric approach, estimating non-crossing quantile curves through estimation of conditional CDF is more appealing in a high dimensional setting where linear relationship is less straight forward to confirm. Recently, Izbicki and Lee (2016) proposed a conditional density

estimator that not only scales well to a high dimensional regression setting, but can also adapt to the low-rank structure of $\boldsymbol{X}$. The central idea of their work is to project the conditional density function $f(y|\boldsymbol{x})$ onto a tensor product basis $f(y|\boldsymbol{x}) = \sum_{i,j} \beta_{i,j} \Psi_{i,j}(y, \boldsymbol{x}), i = 1, ..., I, j = 1, ..., J$ where $\Psi_{i,j}(y) = \phi_i(y)\psi_j(\boldsymbol{x})$ is a tensor product of Fourier basis on $\mathcal{D}(y)$ and spectral basis on $\mathcal{X}$. The authors further showed that the coefficients are simply the expected values of the tensor product bases over the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$, thus estimation of $f(y|\boldsymbol{x})$ is transformed to estimation of the spectral basis $\psi_j(\boldsymbol{x})$ which can be estimated by eigenvectors of the Gram matrix. The main advantage of this estimator over other conditional density estimators of similar kind (Efromovich, 2008; Efromovich, 2010) is that it avoids multiple tensor products in high dimension and is therefore very computationally attractive. The flexibility of choosing different kinds of kernel functions to estimate the Gram matrix also allows the estimator to model different types of $\boldsymbol{X}$. Furthermore, the authors proved that the covergence rate of their spectral series estimator only depends on the intrinsic dimension of $\boldsymbol{X}$. This provides an extra improvement in computation time since many high dimensional datasets have a low rank structure. Estimate of the conditional density based on this method can also be immediately processed to produce estimates of non-crossing quantile curves. A standard procedure would start with numerical integration (e.g. trapezoidal rule) to obtain the estimated CDF and then standard root searching to numerically solve for the quantiles of interest.

Some other works that belongs to the simultaneous estimation type are referenced here. (Hall et al., 1999; Dunson and Taylor, 2005, Taddy and Kottas, 2010; Liu and Wu, 2011; El Adlouni and Baldé, 2019; Merhi Bleik, 2019; Petrella and Raponi, 2019; Rodrigues et al., 2019). In summary, methods of this type either enforce monotonicity constraints on all desired quantile levels of interest or construct a formulation of the conditional quantile function that has partial monotonicity on τ. As such, these methods often lead to more complicated statistical modeling and estimation but provides the most comprehensive description of the conditional distribution.

6 Discussion

In this paper, we reviewed methods that aim to estimate multiple non-crossing quantile curves, which is often of interest when the investigator wants to understand how the effect of a (set of) covariate(s) change across different quantiles of the response. Three major types are compared: namely sequantile estimation, post-processing and simultaneous estimation. Sequential estimation methods start from an unconstrained median regression and fit a series of constrained quantile regression sequentially; they often results in simple estimation problem but produce estimates that are sensitive to the order of estimation. Post-processing methods apply monotonicity adjustment on the unconstrained estimates, but poor initial estimates will lead to unsatisfactory overall performance. Comparing to simultaneous estimation, these two types of methods do not fully enjoy the information-borrowing property. We also observed a significant imbalance between number of works that estimate quantile planes and number of works that estimate quantile surfaces. To date, there is scarce literature that provide an efficient solution to model non-crossing quantile surfaces of arbitrary shape in a high dimension setting. Therefore, future research can focus on developing a flexible model of such type that can approximate a broad range of quantile processes. For example, one could approximate the quantile process using a Bayesian neural network which not only offers great model flexibility but at the same time imposes regularization on the model complexity. Another possible direction is to develop a conditional density estimator that is centered around a parametric regression model so that quantile curves with known structure (e.g. linearity) can be suitably estimated; one main reason that conditional density estimators are not commonly used to estimate non-crossing quantile curves is that covariate effect is often hard to interpret.

7 Simulation study

In this section, methods that were previously reviewed will be applied on synthetic data to compare their performance on estimating non-crossing quantile curves generated from different settings. The study consists of two main parts. In the first part, methods that were devised to estimate quantile curves of a linear quantile regression model are compared. This list includes sequential

estimation by Muggeo et al. (2013) (Section 3, Algorithm 1), Gaussian process regression adjustment by Rodrigues and Fan (2017) (Section 4), simultaneous estimation for a finite collection of quantiles by Bondell et al. (2010) (Section 5.2) and linear quantile process regression by Yang and Tokdar (2017) (Section 5.2). In the second part, methods that can be used to non-parametrically estimate non-linear quantile curves are compared. This list includes Das and Ghosal (2018) (Section 5.3) who expanded the quantile process using a tensor product of B-spline basis functions, Cannon (2018) (Section 5.2) who expanded the quantile process using composite neural network with partial monotonicity constrant and Izbicki and Lee (2016) (Section 5.3) who estimated the conditional density using a spectral series estimator. As mentioned before, any linear quantile regression model can be extended to estimate non-linear curves by replacing the linear term with linear combination of B-spline basis functions. For brevity, we only include the method of Muggeo et al. (2013) in the non-parametric experiment; this is because only their method implements automatic selection of knots which is crucial for the performance of a spline based model. In both parts, univariate and multivarite settings are considered. For all univariate settings, different sample sizes are experimented to compare the methods' performance across small and moderate samples.

7.1 Software

All methods except for that of Das and Ghosal (2018) have been implemented in R: method of Muggeo et al. (2013) is implemented in the package `quantregGrowth`; codes for Rodrigues and Fan (2017) is available from the supplemental material of their online paper; codes for Bondell et al. (2010) is available from the first author's webpage; method of Yang and Tokdar (2017) is implemented in the package `qrjoint`; method of Cannon (2018) is implemented in the package `qrnn`; codes for Izbicki and Lee (2016) is available from the supplemental material of their online paper. For the method of Das and Ghosal (2018), Matlab code is available from the second author's webpage.

7.2 Linear quantile regression

Four simulation designs are considered in this part. For each design, data is generated according to the generic linear quantile regression model

$$Q(\tau|\boldsymbol{x}) = \beta_0(\tau) + \boldsymbol{x}^T\boldsymbol{\beta}(\tau)$$

by first simulating $u_i \sim \mathcal{U}(0,1)$, $\boldsymbol{x}_i \sim \mathcal{U}(\mathcal{X})$ and set $y_i = \beta_0(u_i) + \boldsymbol{x}^T\boldsymbol{\beta}(u_i)$ for $1 \leq i \leq n$. Specification of $\mathcal{X}$ will be provided in each of the four settings.

Design 1:

$$\mathcal{X} = [-1,1]; \beta_0(\tau) = \mathcal{T}_3^{-1}(\tau), \beta_1(\tau) = 2(\tau - 0.5),$$

where $\mathcal{T}_3^{-1}(\cdot)$ denotes the quantile function of a student-t distribution with 3 degrees of freedom.

Design 2:

$$\mathcal{X} = [-1,1]; \beta_0(\tau) = 3(\tau - \frac{1}{2})\log\frac{1}{\tau(1-\tau)}, \beta_1(\tau) = 4(\tau - \frac{1}{2})^2\log\frac{1}{\tau(1-\tau)}x.$$

Design 3:

$$\mathcal{X} = [-1,1]^5; \beta_0(\tau) = \Phi^{-1}(\tau), \beta_1(\tau) = 2\min(\tau - 0.5, 0),$$
$$\beta_2(\tau) = 2\tau, \beta_3(\tau) = 2, \beta_4(\tau) = 1, \beta_5(\tau) = 0.$$

Design 1 represents a standard heteoskedastic model where the slope coefficient is a linear function of τ. Design 2 was studied in Yang and Tokdar (2017); the slope coefficient is nearly a quadratic function of τ. Design 3 was studied in Reich and Smith (2013). It considered coefficent functions that are constant, linear, piece-wise linear and non-linear; a redundant covariate is also included. For Design 1 and 2 we considered $n = \{50, 100, 300\}$. For Design 3, only $n = 300$ is considered. Since the main advantage of linear quantile regression is its retained rate-of-change interpretation of the covariates' effect, we will compare the methods based on the root mean squared error of

$\beta_j(\tau)$ defined by

$$\text{RMSE}(\tau) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left[\beta_j(\tau) - \hat{\beta}_j^{(s)}(\tau)\right]^2},\ 1 \leq j \leq p$$

where S is the number of simulated datasets. For each design and each sample size, we generated $S = 250$ datasets and compare the RMSE of $\beta_j(\tau)$ for each j across $\tau = \{0.05, 0.1, ..., 0.9, 0.95\}$. We also calculated the coverage of 95% confidence intervals constructed based on each method. For Muggeo et al. (2013) and Bondell et al. (2010), this is the percentile interval based on 1000 Bootstrap samples; for Rodrigues and Fan (2017), this is the parametric confidence interval under the normal distribution; and for Yang and Tokdar (2017) this is the percentile interval obtained from the posterior samples. Finally, we calculated the root mean integrated squared error (RMISE) of each partial slope function β_j over the 19 quantile knots

$$\text{RMISE}(\beta_j) = \sqrt{\frac{1}{19}\sum_{k=1}^{19}\left[\beta_j(\tau_k) - \hat{\beta}_j(\tau_k)\right]^2}, 1 \leq j \leq p$$

to compare the overall performance of each method. The parameters of the models are all set to their default values. These include the variance hyperparameter of the Gaussian process prior in Rodrigues and Fan (2017) which is set to be 100 and number of knots for low rank approximation of the Gaussian process priors in Yang and Tokdar (2017) which is set to be 6. The original authors reported that results were not sensitive to the values of these parameters, and our sensitivity analysis later confirmed it. The method of Rodrigues and Fan (2017) assumes that the posterior samples used by the Gaussian process regression for estimating a single quantile are conditionally independent, therefore for each standard Bayesian quantile regression fitted in stage 1 we drew 31500 MCMC samples before discarding the first 1500 and only kept every 30th sample. For the method of Yang and Tokdar (2017), we fitted MCMC with 10000 draws and only used the last 1000 samples.

The RMISE for each design is shown in Table 1, 2 and 3 respectively. When the slope coefficient function is linear against τ, methods of Muggeo et al. (2013) and Yang and Tokdar (2017) performed equally well followed by those of Rodrigues and Fan (2017) and Bondell et al. (2010).

Table 1: **RMISE for Design 1**: Mean and standard deviation (in parentheses) of RMISE(β_1) for each of the four methods and three sample sizes. M1–4 denote the method of Muggeo et al. (2013), Bondell et al. (2010), Yang and Tokdar (2017) and Rodrigues and Fan (2017) respectively. Computation time (seconds) of a typical run when n = 300 is also provided.

n	M1	M2	M3	M4
50	0.43 (0.20)	0.52 (0.22)	0.41 (0.18)	0.44 (0.21)
100	0.31 (0.13)	0.39 (0.15)	0.32 (0.13)	0.34 (0.15)
300	0.19 (0.07)	0.22 (0.08)	0.19 (0.08)	0.20 (0.08)
Time	0.007	0.250	15.00	469.2

Table 2: **RMISE for Design 2**: Mean and standard deviation (in parentheses) of RMISE(β_1) for each of the four methods and three sample sizes. M1–4 denote the method of Muggeo et al. (2013), Bondell et al. (2010), Yang and Tokdar (2017) and Rodrigues and Fan (2017) respectively. Computation time (seconds) of a typical run when n = 300 is also provided.

n	M1	M2	M3	M4
50	1.06 (0.25)	0.85 (0.32)	0.73 (0.26)	0.76 (0.31)
100	0.92 (0.16)	0.63 (0.23)	0.55 (0.22)	0.58 (0.23)
300	0.80 (0.09)	0.35 (0.12)	0.30 (0.12)	0.33 (0.13)
Time	0.006	0.242	15.45	445.5

However when slope coefficient function is quadratic, the method of Muggeo et al. (2013) performs poorly as its constraint can only handle slope function that is monotonically increasing in τ. The overall better performance of Yang and Tokdar (2017) demonstrated the advantage of estimating the whole quantile process simultaneously. In high dimension setting, method of Yang and Tokdar (2017) again performs best as it does not explicitly assume that $\mathcal{X}$ is a hypercube. The RMSE and coverage probability for Design 1 and 2 are plotted in Figure 2. We see that when the assumption is met the frequentist methods achieve coverage probabilities that are closer to 95% than the Bayesian alternatives. In Design 1 the method of Rodrigues and Fan (2017) produces overly wide confidence interval. An explanation for this might be that the variance of the final estimate took into account the variance of the initial estimates which might be large since the QR models

in the first stage are estimated separately. The relatively low coverage probabilities from Yang and Tokdar (2017) might be due to the fact that 95% Bayesian credible bands does not guarantee 95% nominal coverage.

Table 3: **RMISE for Design 3**: Mean ($\times 100$) and standard deviation (in parentheses) of RMISE(β_j), $1 \leq j \leq 5$ for each of the four methods. M1–4 denote the method of Muggeo et al. (2013), Bondell et al. (2010), Yang and Tokdar (2017) and Rodrigues and Fan (2017) respectively. Computation time (seconds) of a typical run when n = 300 is also provided.

	M1	M2	M3	M4
β_1	23.9 (0.11)	25.3 (0.10)	24.6 (0.11)	24.8 (0.10)
β_2	24.7 (0.10)	24.8 (0.10)	25.4 (0.12)	26.2 (0.11)
β_3	25.2 (0.12)	24.3 (0.10)	20.7 (0.11)	21.3 (0.11)
β_4	24.9 (0.12)	24.9 (0.11)	21.5 (0.12)	22.4 (0.11)
β_5	24.1 (0.11)	24.7 (0.10)	21.3 (0.12)	21.9 (0.11)
Time	0.009	1.811	36.81	505.4

7.3 Non-parametric quantile regression

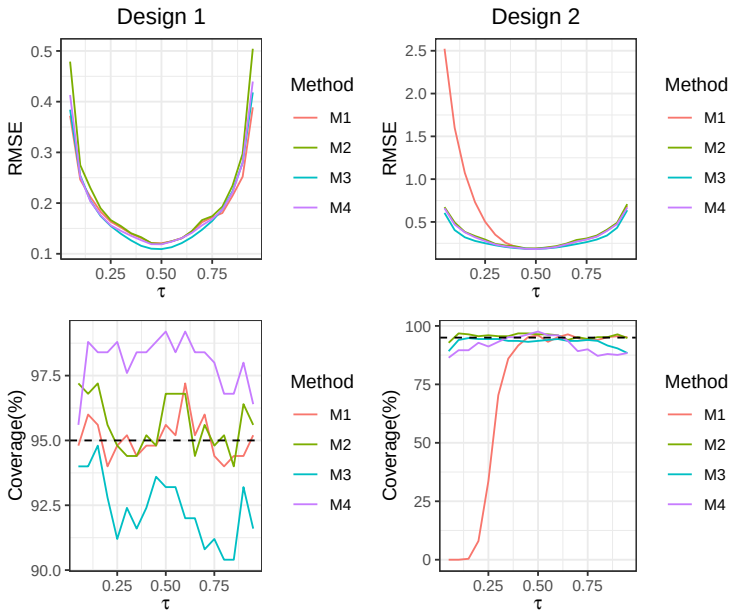
Two simulation designs are considered in this part.

Design 4:

$$\mathcal{X} = [0, 1];$$
$$y = 2x + [0.5 + 2x + \sin(2\pi x - 0.5)]\,\epsilon,$$

where $\epsilon \sim \mathcal{N}(0, 1)$.

Figure 2: **Slope estimation for Design 1 and 2**. The RMSE and coverage probability for estimating $\beta_1(\tau)$ resulted from each method are plotted against the quantile level $\tau \in [0.05, 0.95]$.



Design 5:

$$
\begin{aligned}
&\mathcal{X} = [0,1]^2; \\
&y = \sin(2\pi x_1) + \cos(2\pi x_2) \\
&\quad + \frac{\exp\{8[(x_1-0.5)^2 + (x_2-0.5)^2]\}}{(\exp\{8[(x_1-0.2)^2 + (x_2-0.7)^2]\} + \exp\{8[x_1-0.7]^2 + (x_2-0.7)^2\})} \\
&\quad + \sqrt{2(x_1^2 + x_2^2)}\epsilon
\end{aligned}
\tag{24}
$$

where $\epsilon \sim \mathcal{N}(0,1)$.

Design 4 is a location-scale model studied in Bondell et al. (2010). The quantile is linear at median and highly non-linear at non-central parts; $n = \{50, 100, 300\}$ are experimented. In Design 5, all covariates have a strongly non-linear main effect on y except for x_5 which is a redundant covariate; there is also strong interaction effect between x_3 and x_4; only n = 300 is experimented. All four candidate models are applied on Design 4. For Design 5 however, method of Muggeo et al. (2013) was omitted as it is only suitable for modeling non-linear effect of one covariate. Method of Das and Ghosal (2018) was also omitted because its model involves a tensor product of spline basis functions and lead to infeasible computation. Therefore only the method of Cannon

(2018) and Izbicki and Lee (2016) were considered since they scale well to high dimension setting. In a non-parametric regression setting, the precision rather than the interpretability of the model is of more interest. Therefore we choose our metric to be the RMISE of quantile curves across simulated data points

$$\text{RMISE}(\tau) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{Q(\tau|\boldsymbol{x}_i) - \hat{Q}(\tau|\boldsymbol{x}_i)^2\}}$$

for each of the quantile levels $\tau = \{0.05, 0.10, ..., 0.90, 0.95\}$. For Design 4, we also calculated RMISE of the quantile process

$$\text{RMISE} = \sqrt{\frac{1}{19n}\sum_{k=1}^{19}\sum_{i=1}^{n}\{Q(\tau_k|\boldsymbol{x}_i) - \hat{Q}(\tau_k|\boldsymbol{x}_i)^2\}}$$

to compare the overall performance of each model across different sample sizes.

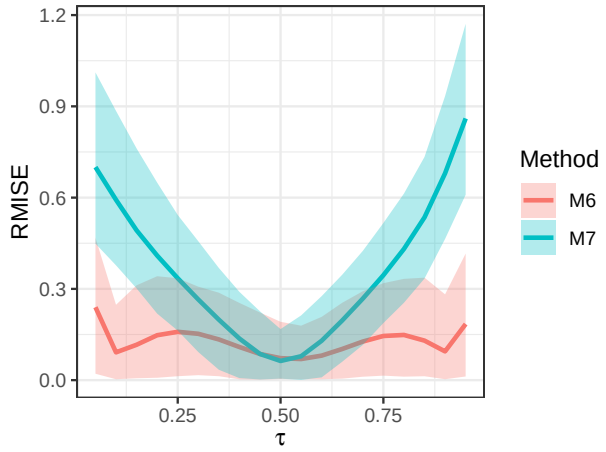
All methods in this part has at least one parameter that needs to be explicitly tuned. For Muggeo et al. (2013), the penalty coefficient for their P-spline smoothing term is tuned by 5-fold cross-validation. For Das and Ghosal (2018), the number of knots for their B-spline basis expansion is chosen based on the estimated Akaike Information Criterion (AIC) . For Izbicki and Lee (2016), the bandwidth parameter of their kernel-based operator is tuned by validation set approach. For Cannon (2018), we only considered one hidden layer and tuned the number of hidden neurons and weight penalty coefficient jointly using 5-fold cross-validation. All methods are tuned using a dataset independent to the training dataset. The parameter is tuned for each sample size and is then used through out the simulation study.

The RMISE for Design 4 is shown in Table 4. The method of Das and Ghosal (2018) performed poorly as it imposed a constraint that is only the sufficient condition of the original non-crossing constraint, thus putting too much restriction on its coverage. The method of Muggeo et al. (2013) outperformed those of Cannon (2018) and Izbicki and Lee (2016) for all three sample sizes. One explanation might be that the latter two are more suited for high dimensional settings and are more likely two overfit the data under univariate settings. In the high dimensional setting, the method of

Table 4: **RMISE for Design 4**: Averaged RMISE of each model followed by its standard deviation (in parentheses) for each of the four methods and three sample sizes. M1, M5-7 denote the method of Muggeo et al. (2013), Das and Ghosal (2018), Cannon (2018) and Izbicki and Lee (2016) respectively. Computation time (seconds) of a typical run when n = 300 is also provided.

n	M1	M5	M6	M7
50	0.63 (0.12)	1.08 (0.66)	0.69 (0.15)	0.64 (0.17)
100	0.44 (0.08)	0.78 (0.33)	0.54 (0.08)	0.59 (0.13)
300	0.29 (0.04)	0.56 (0.23)	0.37 (0.06)	0.47 (0.16)
Time	0.043	336.4	30.35	5.710

Figure 3: **RMISE for Design 5**. The RMISE for each method averaged over 250 replicated datasets as well as its 95 percentile bands are plotted against the quantile level $\tau \in [0.05, 0.95]$.



Cannon (2011) outperformed that of Izbicki and Lee (2016) almost everywhere except for around the median. This is because methods that model the conditional densities generally produce overly smooth estimations which cannot adequately capture the high non-linearity of the true conditional quantile functions. In terms of computational time, method of Izbicki and Lee (2016) took 10.96 seconds for a typical run of Design 5, whereas method of Cannon (2018) took seconds.

8 Sensitivity analysis

In this section we provide some results on the sensitivity studies of several reviewed methods. These include the sensitivity of Rodrigues and Fan (2017) to the variance hyperparameter σ_k^2 (see Section 4), the sensitivity of Yang and Tokdar (2017) to the number of knots for low rank approximation of the Gaussian process priors and the sensitivity of Bondell et al. (2010) to the set of fitted quantile levels. We compared the estimation results of each method under different settings using 250 replicated datasets generated from Design 1.

Figure 4: **Sensitivity of Rodrigues and Fan (2017) to the value of variance hyperparameter**. The RMSE and coverage probability for estimating the slope coefficient function under different settings of variance hyperparameter are plotted against the quantile level $\tau \in [0.05, 0.95]$.

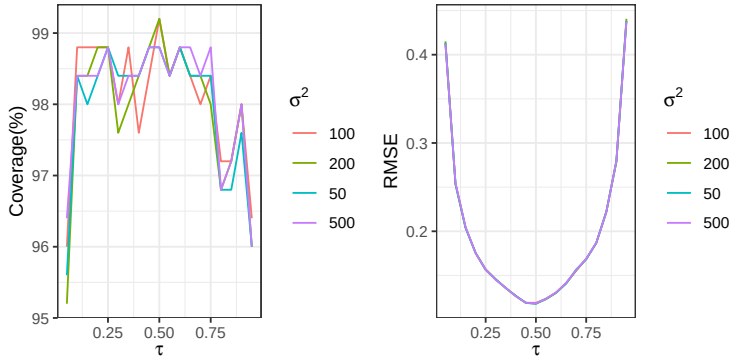
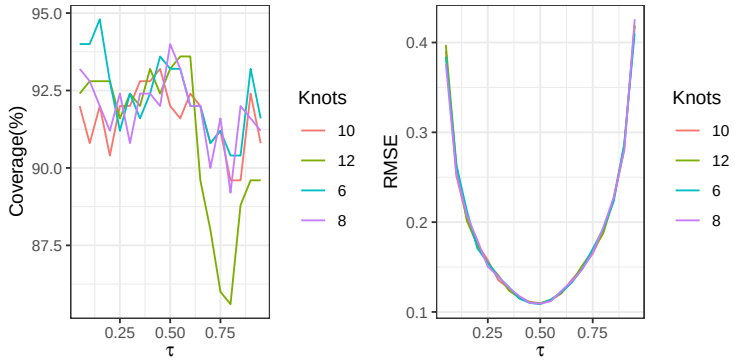


Figure 5: **Sensitivity of Yang and Tokdar (2017) to the number of knots for low rank approximation**. The RMSE and coverage probability for estimating the slope coefficient function under different settings of variance hyperparameter are plotted against the quantile level $\tau \in [0.05, 0.95]$.



For the method of Rodrigues and Fan (2017), the original authors claimed that the results were

not sensitive to moderately large values of σ_k^2, so we chose to experiment with $\sigma_k^2 = 50, 100, 200, 500$. In Figure 4, the coverage probability and RMSE under each setting are plotted against the quantile levels. We see that the results were nearly identical under these four settings. For the method of Yang and Tokdar (2017), we compared its result across four choices of number of knots $= 6, 8, 10, 12$. We see that in Figure 5 although the coverage probability dropped a little when using 12 knots, RMSE is nearly identical across the four settings. For the method of Bondell et al. (2010), we have mentioned that its result depends on the set of fitted quantile levels. Therefore we compared the estimation of the coefficient value at the deciles using two sets of equidistant quantile levels $\tau_l = \frac{l}{L+1}, l = 1, ..., L$ with $L = 9, 19$. The median and the 95 percentile interval of the 250 sets of estimated coefficients are shown in Table 5. We see that the estimates are clearly dependent on the set of fitted quantile levels, but the impact is not significant.

Table 5: **Sensitivity of Bondell et al. (2010) to the set of fitted quantile levels**: Median and 95 percentile interval for the estimated $\beta_1(\tau)$ at the deciles using method of Bondell et al. (2010) across 250 replicated datasets. The fitted quantile levels are $\tau_l = \frac{l}{L+1},\ l = 1, ..., L$.

τ	L = 19	L = 9
0.1	-0.795	-0.798
	(-1.710, 0.412)	(-1.692, 0.412)
0.2	-0.588	-0.590
	(-1.254, 0.109)	(-1.254, 0.158)
0.3	-0.398	-0.394
	(-0.915, 0.213)	(-0.918, 0.196)
0.4	-0.204	-0.213
	(-0.650, 0.315)	(-0.651, 0.321)
0.5	-0.028	-0.033
	(-0.428, 0.453)	(-0.419, 0.450)
0.6	0.160	0.159
	(-0.257, 0.643)	(-0.254, 0.642)
0.7	0.350	0.347
	(-0.211, 0.895)	(-0.206, 0.898)
0.8	0.538	0.540
	(-0.204, 1.121)	(-1.254, 1.120)
0.9	0.692	0.703
	(-0.306, 1.748)	(-0.306, 1.672)

References

Abeywardana, S. and Ramos, F. (2015) Variational inference for nonparametric bayesian quantile regression. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Abrevaya, J. (2001) The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, **26**, 247–257.

Bernardi, Mauro, G.-G. and Petrella, L. (2015) Bayesian tail risk interdependence using quantile regression. *Bayesian Analysis*, **10**, 553–603.

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

Boukouvalas, A., Barillec, R. and Cornford, D. (2012) Gaussian process quantile regression using expectation propagation. *arXiv preprint arXiv:1206.6391*.

Cannon, A. J. (2011) Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers & geosciences*, **37**, 1277–1284.

— (2018) Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic environmental research and risk assessment*, **32**, 3207–3225.

Chernozhukov, V., Fernández-Val, I. and Galichon, A. (2010) Quantile and probability curves without crossing. *Econometrica*, **78**, 1093–1125.

Das, P. and Ghosal, S. (2018) Bayesian non-parametric simultaneous quantile regression for complete and grid data. *Computational Statistics & Data Analysis*, **127**, 172–186.

Dette, H. and Volgushev, S. (2008) Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 609–627.

Dunson, D. B. and Taylor, J. A. (2005) Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics*, **17**, 385–400.

Efromovich, S. (2008) *Nonparametric curve estimation: methods, theory, and applications*. Springer Science & Business Media.

— (2010) Dimension reduction and adaptation in conditional density estimation. *Journal of the American Statistical Association*, **105**, 761–774.

El Adlouni, S. and Baldé, I. (2019) Bayesian non-crossing quantile regression for regularly varying distributions. *Journal of Statistical Computation and Simulation*, **89**, 884–898.

Franke, J. and Neumann, M. H. (2000) Bootstrapping neural networks. *Neural computation*, **12**, 1929–1949.

Hall, P., Wolff, R. C. and Yao, Q. (1999) Methods for estimating a conditional distribution function. *Journal of the American Statistical association*, **94**, 154–163.

He, X. (1997) Quantile curves without crossing. *The American Statistician*, **51**, 186–192.

Izbicki, R. and Lee, A. B. (2016) Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, **25**, 1297–1316.

Jagger, T. H. and Elsner, J. B. (2009) Modeling tropical cyclone intensity with quantile regression. *International Journal of Climatology*, **29**, 1351–1361.

Koenker, R. (2005) *Quantile Regression*. Cambridge University Press.

Koenker, R. and Bassett Jr, G. (1978) Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Koenker, Roger, S. P. and Ng, P. (1992) Nonparametric estimation of conditional quantile functions. *Dodge, Y.(Ed)*, 217–229.

— (1994) Quantile smoothing splines. *Biometrika*, **81**, 673–680.

Kozumi, H. and Kobayashi, G. (2011) Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, **81**, 1565–1578.

Liu, Y. and Wu, Y. (2009) Stepwise multiple quantile regression estimation using non-crossing constraints. *Statistics and its Interface*, **2**, 299–310.

— (2011) Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of nonparametric statistics*, **23**, 415–437.

Lum, K., Gelfand, A. E. et al. (2012) Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis*, **7**, 235–258.

Meinshausen, N. (2006) Quantile regression forests. *Journal of Machine Learning Research*, **7**, 983–999.

Merhi Bleik, J. (2019) Fully Bayesian estimation of simultaneous regression quantiles under asymmetric laplace distribution specification. *Journal of Probability and Statistics*, **2019**.

Miranda, M. L., Kim, D., Reiter, J., Galeano, M. A. O. and Maxson, P. (2009) Environmental contributors to the achievement gap. *Neurotoxicology*, **30**, 1019–1024.

Muggeo, V. M., Sciandra, M., Tomasello, A. and Calvo, S. (2013) Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. *Environmental and ecological statistics*, **20**, 519–531.

Petrella, L. and Raponi, V. (2019) Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress. *Journal of Multivariate Analysis*, **173**, 70–84.

Quadrianto, Novi, K. K. R. M. D. C. T. S. and Buntine, W. L. (2009) Kernel conditional quantile estimation via reduction revisited. In *2009 Ninth IEEE International Conference on Data Mining*, 938–943.

Rasmussen, C. and Williams, C. (2006) *Gaussian Processes for Machine Learning*. MIT Press.

Reich, B. J. (2012) Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**, 535–553.

Reich, B. J., Fuentes, M. and Dunson, D. B. (2011) Bayesian spatial quantile regression. *Journal of the American Statistical Association*, **106**, 6–20.

Reich, B. J. and Smith, L. B. (2013) Bayesian quantile regression for censored data. *Biometrics*, **69**, 651–660.

Rodrigues, T., Dortet-Bernadet, J.-L. and Fan, Y. (2019) Pyramid quantile regression. *Journal of Computational and Graphical Statistics*, **28**, 732–746.

Rodrigues, T. and Fan, Y. (2017) Regression adjustment for noncrossing bayesian quantile regression. *Journal of Computational and Graphical Statistics*, **26**, 275–284.

Sriram, K., Ramamoorthi, R., Ghosh, P. et al. (2013) Posterior consistency of bayesian quantile regression based on the misspecified asymmetric laplace density. *Bayesian Analysis*, **8**, 479–504.

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

Takeuchi, I., Le, Q. V., Sears, T. D. and Smola, A. J. (2006) Nonparametric quantile estimation. *Journal of machine learning research*, **7**, 1231–1264.

Taylor, J. W. (2000) A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, **19**, 299–311.

Thompson, Paul, Y.-C. M. R. R. D. and Stander, J. (2011) Bayesian nonparametric quantile regression using splines. *Computational Statistics & Data Analysis*, **54**, 1138–1150.

Tokdar, S. T., Kadane, J. B. et al. (2012) Simultaneous linear quantile regression: a semiparametric Bayesian approach. *Bayesian Analysis*, **7**, 51–72.

Yang, Y. and Tokdar, S. T. (2017) Joint estimation of quantile planes over arbitrary predictor spaces. *Journal of the American Statistical Association*, **112**, 1107–1120.

690 Yu, K. and Jones, M. (1998) Local linear quantile regression. *Journal of the American statistical*
691 *Association*, **93**, 228–237.

692 Yu, K. and Moyeed, R. A. (2001) Bayesian quantile regression. *Statistics & Probability Letters*,
693 **54**, 437–447.

694 Yu, K. and Zhang, J. (2005) A three-parameter asymmetric Laplace distribution and its extension.
695 *Communications in Statistics—Theory and Methods*, **34**, 1867–1879.

696 Yuan, Y., Chen, N. and Zhou, S. (2017) Modeling regression quantile process using monotone
697 B-splines. *Technometrics*, **59**, 338–350.

698 Zhang, H. and Zhang, Z. (1999) Feedforward networks with monotone constraints. In *IJCNN'99.*
699 *International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, vol. 3,
700 1820–1823. IEEE.

701 Zheng, S. (2012) QBoost: Predicting quantiles with boosting for regression and binary classifica-
702 tion. *Expert Systems with Applications*, **39**, 1687–1697.

703 Zou, H. and Yuan, M. (2008) Composite quantile regression and the oracle model selection theory.
704 *The Annals of Statistics*, **36**, 1108–1126.

Appendix A: Codes for simulation

Design 1:

```
n <- 100
b0 <- function(tau){qt(tau,df=3)}
b1 <- function(tau){2*(tau-0.5)}
u <- runif(n,0,1)
x <- runif(n,-1,1)
q <- b0(u)+b1(u)*x
```

Design 2:

```
n <- 100
b0 <- function(tau){
  3*(tau-1/2)*log(1/(tau*(1-tau)))
}
b1 <- function(tau){
  4*(tau-1/2)^2*log(1/(tau*(1-tau)))
}
u <- runif(n,0,1)
x <- runif(n,-1,1)
q <- b0(u)+b1(u)*x
```

Design 3:

```
n <- 300
bcoef <- function(tau){
  b0 <- 2*qnorm(tau)
  b1 <- 2*pmin(tau-0.5,0)
  b2 <- 2*tau
  b3 <- 2
  b4 <- 1
  b5 <- 0
  return(cbind(b0,b1,b2,b3,b4,b5))
}
u <- runif(n,0,1)
x <- matrix(runif(n*5,-1,1),nrow = n)
q <- rowSums(cbind(1,x)*bcoef(u))
```

Design 4:

```
n <- 100
```

```
f <- function(x){
  3*x
}
g <- function(x){
  0.5+2*x+sin(2*pi*x-0.5)
}
x <- runif(n)
y <- f(x)+g(x)*rnorm(n)
```

Design 5:

```
n <- 300
f1 <- function(x){
  sin(2*pi*x)
}
f2 <- function(x){
  cos(2*pi*x)
}
f3 <- function(x1,x2){
  5*exp(8*((x1-0.5)^2+(x2-0.5)^2))/
  (exp(8*((x1-0.2)^2+(x2-0.7)^2))+exp(8*((x1-0.7)^2+(x2-0.7)^2)))
}
x <- matrix(runif(n*5),ncol = 5)
y <- f1(x[,1])+f2(x[,2])+f3(x[,3],x[,4])
     +sqrt(2*(x[,1]^2+x[,2]^2))*rnorm(n)
```