# Bayesian quantile additive regression trees

Bereket P. Kindo    Hao Wang    Timothy Hanson

Edsel A. Peña

## Abstract

Ensemble of regression trees have become popular statistical tools for the estimation of conditional mean given a set of predictors. However, quantile regression trees and their ensembles have not yet garnered much attention despite the increasing popularity of the linear quantile regression model. This work proposes a Bayesian quantile additive regression trees model that shows very good predictive performance illustrated using simulation studies and real data applications. Further extension to tackle binary classification problems is also considered.

# 1    Introduction

Quantile regression gives a comprehensive picture of the relationship between a response variable and a set of predictors. It is particularly appealing when the inferential interest lies in the probabilistic properties of extreme observations conditional on a set of predictors. Such objectives arise in various disciplines: in environmental sciences, Friederichs and Hense (2007) study the probabilistic properties of extreme precipitation events, while Pedersen (2015) model the tail distribution of stock and bond returns. In an epidemiological study, Burgette et al. (2011) use penalized quantile regression to explore covariates that affect the lower tail of the distribution of birth weight of babies. When the distribution of the dependent variable is skewed, the desire for robustness to extreme observations makes quantile regression a preferred approach. Examples include the study of tourist expense patterns in Marrocu et al. (2015) and wage distribution in Buchinsky (1995).

Extensive work in the theory and application of linear quantile regression can be found in Koenker and Bassett Jr (1978); Koenker (1994); Buchinsky (1998); Tsai (2012); Cole and Green (1992). Suppose we have a data set $(y_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$, where $y_i \in \Re$ and $\mathbf{x}_i \in \Re^d$ denote the observed response and predictors for the $i^{\text{th}}$ observation, respectively. Analogous to the use of the mean function $E(y|\mathbf{x})$ used in least squares regression to explain the relationship between the response and predictors, quantile regression uses the $\tau^{\text{th}}$ quantile function $Q(y|\mathbf{x}, \tau)$, where $\tau \in (0, 1)$. The $\tau^{\text{th}}$ quantile of a random variable $Y$ with distribution $F$ is defined as $Q(\tau) = \inf \{y : F(y) \geq \tau\}$, where $F(\cdot)$ denotes the cumulative distribution function. Thus, for a given quantile value $\tau$, quantile regression seeks to estimate $Q(\mathbf{x}, \tau) = \inf \{y : F(y|\mathbf{x}) \geq \tau\}$. The linear quantile regression problem in particular is described as the minimization problem

$$\hat{\beta}_\tau = \arg \min_\beta \sum_{i=1}^n \rho_\tau \left(y_i - \mathbf{x}_i^{\mathrm{T}} \beta\right), \tag{1}$$

where $\rho_\tau(\omega) = \omega (\tau - I \{\omega < 0\})$ is usually termed as the "check loss" function. The error distribution is left largely unspecified except that its $\tau^{\text{th}}$ quantile equals zero. The work in Koenker and Bassett Jr (1978) spearheaded the use of quantile regression as a robust alternative to mean regression. More recently, $l_1$ regularized quantile regression with simultaneous variable selection and parameter estimation is studied in Zou and Yuan (2008); Belloni and Chernozhukov (2011).

An alternative, yet equivalent, formulation of (1) assumes that the random errors follow the asymmetric Laplace distribution (Yu and Moyeed, 2001; Kozumi and Kobayashi, 2011; Sriram et al., 2013). If a random variable $Y$ follows an asymmetric Laplace distribution $\text{ALD}(y; \tau, \mu)$ with location parameter $\mu \in \Re$, its density function is given by

$$f_\tau(y; \mu) = \tau(1 - \tau) \exp\{-\rho_\tau(y - \mu)\}, \tag{2}$$

where $\tau \in (0, 1), \rho_\tau(\omega) = \omega(\tau - I\{\omega < 0\})$ for $\omega \in \Re$. A special case of (2) with $\tau = 0.5$ is the Laplace double exponential distribution. Figure 1 shows the plots of the probability density functions of asymmetric Laplace distributions for fixed location parameter $\mu = 0$, and values of $\tau \in \{0.25, 0.50, 0.83\}$. The expectation and variance of $Y \sim \text{ALD}(\tau, \mu = 0)$ are

$$\mathrm{E}(Y) = \frac{1 - 2\tau}{\tau(1 - \tau)} \text{ and } \mathrm{Var}(Y) = \frac{1 - 2\tau + 2\tau^2}{\tau^2(1 - \tau)^2}, \tag{3}$$
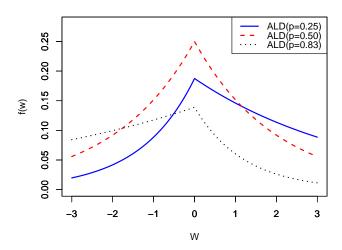
2

**Plot of ALD PDFs**

Figure 1: Asymmetric Laplace distribution with $\mu = 0$ and values of $\tau \in \{0.25, 0.50, 0.83\}$.

while its characteristic function is $\psi_Y(t) = \left[\frac{1}{2}\vartheta_2^2 t^2 - \vartheta_1 ti + 1\right]^{-1}$, where $\vartheta_1 = \frac{1-2\tau}{\tau(1-\tau)}$ and $\vartheta_2^2 = \frac{2}{\tau(1-\tau)}$.

Some Bayesian approaches to the quantile regression problem in general and median regression in particular have been considered in Yu and Moyeed (2001); Dunson and Taylor (2005); Taddy and Kottas (2012); Hanson and Johnson (2002); Kozumi and Kobayashi (2011); Kottas and Gelfand (2001); Reich et al. (2010) either by assuming asymmetric Laplace, Dirichlet process mixtures, Polya trees, or Gaussian mixture approximations as the distribution of the random error term. In particular, Kozumi and Kobayashi (2011) outline a Gibbs sampler for Bayesian quantile regression based on a mixture representation of the asymmetric Laplace distribution. With the intention of utilizing their approach, we paraphrase their finding which they show using the equality of characteristic functions. If the random variables $V$ and $Z$ which follow the standard exponential and Gaussian distributions respectively are mutually independent, then $W = \vartheta_1 V + \vartheta_2\sqrt{V}Z$ is equal in distribution to the asymmetric Laplace distribution $\text{ALD}(\tau, \mu = 0)$, where $\vartheta_1 = \frac{1-2\tau}{\tau(1-\tau)}$ and $\vartheta_2^2 = \frac{2}{\tau(1-\tau)}$. Such representation allows a formulation of an efficient algorithm to estimate regression quantiles in a Bayesian frame-

work that involves simulations from the Gaussian and Generalized Inverse Gaussian distributions.

In comparison to least squares regression trees, quantile regression trees or their ensembles have not yet garnered much attention. However, sporadic works in the literature exist including the single tree quantile regression model of Chaudhuri and Loh (2002) and the quantile regression forests model in Meinshausen (2006) which extends on the idea of random forests (Breiman, 2001). In quantile regression forests model of Meinshausen (2006), all of the observations that lie in a regression tree terminal node are used for estimation while a summary statistic (typically the average) of the observations in a terminal node are used by random forests. At the core of the quantile regression forests is the empirical estimation of the conditional cumulative density function $F(y|\mathbf{x}) = P(Y \leq y|\mathbf{x})$ so that $\hat{Q}(\mathbf{x}, \tau) = \inf\left\{y : \hat{F}(y|\mathbf{x}) \geq \tau\right\}$, where $\hat{F}$ is an estimator of $F$.

Bayesian regression trees and their ensembles are shown to have enhanced predictive performance in the framework of least squares regression, and binary and multiclass classification (Chipman et al., 1998, 2010; Abu-Nimeh et al., 2007; Zhang and Härdle, 2010; Pratola et al., 2014; Kapelner and Bleich, 2013; Kindo et al., 2016) . In particular, BART - Bayesian additive regression trees (Chipman et al., 2010) estimates the conditional mean of a response given a set of predictors by using a sum of regression trees model

$$y = \sum_{j=1}^{n_T} g\left(\mathbf{x}; \mathrm{T}_j, \mathrm{M}_j\right) + \epsilon, \text{ where } \epsilon \sim \mathrm{N}\left(0, \sigma^2\right). \tag{4}$$

BART is specified through priors on the regression trees via a "tree generating stochastic process" that favors shallow trees and prior specifications on terminal node parameters that strategically shrink the influence of individual trees. BART has been utilized in many applications with great predictive performance (Abu-Nimeh et al., 2007; Zhang and Härdle, 2010; Wu et al., 2010; He et al., 2009; Liu et al., 2015). In this article we explore the utility of ensemble of Bayesian regression trees to garner a comprehensive view of the dependence between a response and predictors. Thus, we propose a fully Bayesian framework for construction of quantile regression trees and their ensembles to complement the linear Bayesian quantile regression of Kozumi and Kobayashi (2011); Yu and Moyeed (2001) and quantile regression forests of Meinshausen (2006). We note that, at the time of this writing, we are not

aware of a Bayesian counterpart in the literature to the frequentist quantile regression tree.

The remaining parts of this article are outlined as follows. Section 2 sets the framework for Bayesian quantile additive regression trees including the prior specifications on all the parameters of the model and posterior computations. Section 3 delves into the implementation of the model with simulation studies and real data applications. Section 4 extends Bayesian quantile additive trees to tackle binary classification problems along with a simulation study and real data application. Section 5 provides concluding remarks.

# 2   Bayesian quantile additive regression trees

In this section we outline the model specifications for Bayesian quantile additive regression trees. Specifically, let the observable data be $(y_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$, where $y_i \in \Re$ and $\mathbf{x}_i \in \Re^d$ denoting the response and predictors for the $i^{\text{th}}$ observation. Consider the model

$$y_i = \mathbf{G}\left(\mathbf{x}_i; \mathbf{T}, \mathbf{M}\right) + \vartheta_1 \nu_i + \vartheta_2 \phi^{\frac{1}{2}} \sqrt{\nu_i} z_i,$$

$$\mathbf{G}\left(\mathbf{x}_i; \mathbf{T}, \mathbf{M}\right) = \sum_{j=1}^{n_{\text{T}}} g\left(\mathbf{x}_i; \mathrm{T}_j, \mathrm{M}_j\right)$$

$$p\left(\nu_i | \phi\right) = \frac{1}{\phi} \exp\left\{-\frac{\nu_i}{\phi}\right\},$$

$$p\left(z_i\right) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} z_i^2\right\},$$

(5)

where $\mathrm{T}_j$ and $\mathrm{M}_j$ are the $j^{\text{th}}$ tree in the sum and its associated terminal node parameters, and $(\mathbf{T}, \mathbf{M}) = \{(\mathrm{T}_j, \mathrm{M}_j) ; j = 1, \ldots, n_{\text{T}}\}$. Note that $\vartheta_1 \nu_i + \vartheta_2 \phi^{\frac{1}{2}} \sqrt{\nu_i} z_i = \phi \left[\vartheta_1 \tilde{\nu}_i + \vartheta_2 \sqrt{\tilde{\nu}_i} z_i\right]$, where $\tilde{\nu}_i \sim Exp\left(1\right)$ and the quantity in the square brackets is the mixture representation of the asymmetric Laplace distribution.

## 2.1   Prior specifications

We assume that the priors on any two distinct trees in the sum are independent and the prior on $\phi$ is independent of the tree priors. That is,

$(T_j, M_j) \perp (T_{j'}, M_{j'})$ for $j \neq j'$, and $(\mathbf{M}, \mathbf{T}) \perp \phi$. Further assuming that given a tree, say the $j^{\text{th}}$ tree $T_j$, the priors on its $m_j$ terminal node parameters are independent enables writing the prior distribution on $(\mathbf{T}, \mathbf{M}, \phi)$ as

$$
\begin{aligned}
p(\mathbf{T}, \mathbf{M}, \phi) &= \left[ \prod_{j=1}^{n_\text{T}} p(T_j, M_j) \right] p(\phi) \\
&= \left[ \prod_{j=1}^{n_\text{T}} [p(T_j) \, p(M_j | T_j)] \right] p(\phi) \\
&= \left[ \prod_{j=1}^{n_\text{T}} \left[ p(T_j) \prod_{k=1}^{m_j} p(\mu_{jk} | T_j) \right] \right] p(\phi),
\end{aligned}
\tag{6}
$$

where $n_\text{T}$ is the number of trees in the sum and $m_j$ is the number of terminal nodes of tree $T_j$ (i.e., $M_j = (\mu_{j1}, \ldots, \mu_{jm_j})$).

The prior $p(T_j)$ is specified through a "tree generating stochastic process" of Chipman et al. (1998). This process is governed by tree splitting rule that creates non-overlapping partitions of the predictor space by selecting a splitting variable followed by a splitting value given the selected variable. Once a terminal node is randomly selected for use in binary partitioning of the predictor space, a splitting variable is randomly chosen followed by a random selection of a value in the range of the selected predictor with condition that no empty partition is created. Furthermore, the probability that a terminal node $\eta$ with depth $d_\eta$ (number of ancestor nodes) splits is given by

$$
p_{\text{SPLIT}}(\eta) = \begin{cases} 1 & \text{if } d_\eta = 0 \\ \frac{\psi_1}{(1+d_\eta)^{\psi_2}}, & \text{if } d_\eta > 0, \end{cases}
\tag{7}
$$

where $\psi_1 \in (0, 1), \psi_2 \in [0, \infty)$. The splitting probability in (7), and the choice of $\psi_1$ and $\psi_2$ play a crucial role of regulating the influence of individual trees in the sum. For example, higher values of $\psi_2$ and lower values of $\psi_1$ result in shallow trees in general.

Given a tree $T_j$, the prior on the terminal node parameters is a Gaussian distribution $\mu_{jk} | T_j \sim N(\mu_0, \sigma_0^2)$ for $k = 1, \ldots, m_j$. In the model representation given in (5), the overall contribution of the prior distributions of the terminal node parameters on $\text{E}(y|\mathbf{x})$ and $\text{Var}(y|\mathbf{x})$ are $n_\text{T}\mu_0$ and $n_\text{T}\sigma_0^2$. The hyper-parameters $\mu_0$ and $\sigma_0^2$ are selected so that the overall effect induced by the prior distributions is in the interval $(\min(y), \max(y))$ with high probabil-

ity. A convenient aspect of the quantile function is its invariance to a monotone transformation. In particular, we use the transformation $\tilde{y} = h(y) = \frac{y - \min(y)}{(\max(y) - \min(y))} - 0.5$ for which we have $Q(y, \tau) = h^{-1}(Q(\tilde{y}, \tau))$. Taking $\tilde{y}$ as the dependent variable in (5) along with priors $\mu_{jk}|T_j \sim N\left(\mu_0 = 0, \sigma_0^2 = \frac{1}{2\kappa\sqrt{n_T}}\right)$, we ensure that the transformed response is in the interval $(-0.5, 0.5)$. This choice of the hyper-parameters also ensures that the effect of the prior distributions on the terminal nodes places high probability to the same interval. We find that a value of $\kappa$ between 2 and 3 gives reasonable results. Note that the larger the number of trees in the sum, the smaller the prior variance placed on the terminal node parameters effectively shrinking the influence of individual trees to zero. Finally, the prior on $\phi$ is specified as an Inverse-Gamma distribution $\phi \sim IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$.

## 2.2 Posterior updating scheme

The posterior updating scheme cycles through the following three posterior draws: a draw from

$$p\left(\mathbf{V}|\mathbf{T}, \mathbf{M}, \mathbf{Y}, \phi\right) \tag{8}$$

followed by consecutive updates of the $j^{\text{th}}$ tree and its terminal node parameters for $j = 1, \ldots, n_T$ accomplished by a draw from

$$p\left\{(T_j, M_j)\,|\mathbf{M}_{(-j)}, \mathbf{T}_{(-j)}, \phi, \mathbf{X}, \mathbf{Y}\right\}, \tag{9}$$

with $\left(\mathbf{T}_{(-j)}, \mathbf{M}_{(-j)}\right)$ denoting all the trees and their terminal node parameters in the sum excluding the $j^{th}$ tree; and finally a draw from

$$p\left\{\phi|\mathbf{M}, \mathbf{T}, \mathbf{X}, \mathbf{Y}\right\}, \tag{10}$$

where $\mathbf{V} = (\nu_1, \ldots, \nu_n)^{\mathsf{T}}$, $\mathbf{Y} = (y_1, \ldots, y_n)^{\mathsf{T}}$, and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathsf{T}}$. The posterior draw in (8) is $n$ sequential samples from the Generalized Inverse Gaussian distribution

$$p\left(\nu_i|\mathbf{T}, \mathbf{M}, \phi, y_i, \mathbf{x}_i\right) \propto \nu_i^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\delta_{1i}\nu_i^{-1} + \delta_{2i}\nu_i\right]\right\}, \tag{11}$$

where $\delta_{1i} = \frac{(y_i - \mathbf{G}(\mathbf{x}_i; \mathbf{T}, \mathbf{M}))^2}{\vartheta_2^2\phi}$ and $\delta_2 = \frac{2\vartheta_2^2 + \vartheta_1^2}{\vartheta_2^2\phi}$. To describe the draw in (9), we re-write (5) as

$$\omega_i \equiv y_i - \sum_{l \neq j} g\left(\mathbf{x}_i; T_l, M_l\right) - \vartheta_1\nu_i = g\left(\mathbf{x}_i; T_j, M_j\right) + \phi^{\frac{1}{2}}\vartheta_2\sqrt{\nu_i}z_i \tag{12}$$

so that $\omega_i | \mathbf{x}_i, \nu_i, \mathbf{T}_{(-j)}, \mathbf{M}_{(-j)}, \phi \sim \mathrm{N}\left(g\left(\mathbf{x}_i; \mathrm{T}_j, \mathrm{M}_j\right), \phi\vartheta_2^2\nu_i\right)$. A Metropolis-Hastings algorithm is utilized to update the tree $\mathrm{T}_j$ with $\mathbf{W} = (\omega_1, \ldots, \omega_n)^{\mathbf{T}}$ considered a residual psuedo-response variable. A similar Bayesian "back-fitting" algorithm is implemented in Chipman et al. (2010); Kindo et al. (2016).

For ease of explanation of the Metropolis-Hasting algorithm, we pursue a slight modification of notation as follows. Suppose that $\mathbf{W}_k = (\omega_{k1}, \ldots, \omega_{kn_k})^{\mathbf{T}}$ is a vector of residuals that lie in the $k^{\mathrm{th}}$ terminal node of the regression tree $\mathrm{T}_j$ which has $m_j$ terminal nodes, and that $\mathbf{X}_k = (\mathbf{x}_{k1}, \ldots, \mathbf{x}_{kn_k})^{\mathbf{T}}$ denotes the corresponding set of predictors. Likewise, $\mathbf{V}_k = (\nu_{k1}, \ldots, \nu_{kn_k})^{\mathbf{T}}$ and $\mathbf{Z}_k = (z_{k1}, \ldots, z_{kn_k})^{\mathbf{T}}$ denote the components of the mixture representation of asymmetric Laplace error term corresponding to the observations in the $k^{\mathrm{th}}$ terminal node. With this notation, we write $\mathbf{W} = \left(\mathbf{W}_1, \ldots, \mathbf{W}_{m_j}\right)^{\mathbf{T}}$, $\mathbf{X} = \left(\mathbf{X}_1, \ldots, \mathbf{X}_{m_j}\right)^{\mathbf{T}}$, and $\mathbf{V} = \left(\mathbf{V}_1, \ldots, \mathbf{V}_{m_j}\right)^{\mathbf{T}}$, where $n = n_1 + \ldots + n_{m_j}$. Similar notation is used in Chipman et al. (1998). We can then write the likelihood function of the single residual tree in (12) as

$$f(\mathbf{W}|\mathbf{X}, \mathbf{V}, \phi, \mathrm{T}_j, \mathrm{M}_j) = \prod_{k=1}^{m_j} f(\mathbf{W}_k|\mathbf{X}_k, \mathbf{V}_k, \phi, \mathrm{T}_j, \mathrm{M}_j), \qquad (13)$$

where

$$f(\mathbf{W}_k|\mathbf{X}_k, \mathbf{V}_k, \phi, \mathrm{T}_j, \mathrm{M}_j) = f(\mathbf{W}_k|\mu_{jk}, \mathbf{V}_k, \phi)$$

$$= \left[\frac{1}{\sqrt{2\pi}\vartheta_2\phi^{\frac{1}{2}}}\right]^{n_k} \prod_{l=1}^{n_k} \nu_{kl}^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\vartheta_2^2\phi}\sum_{l=1}^{n_k}\frac{(\omega_{kl}-\mu_{jk})^2}{\nu_{kl}}\right\}. \qquad (14)$$

With the prior specification $\mu_{jk} \sim \mathrm{N}\left(\mu_0 = 0, \sigma_0^2 = \frac{1}{2\kappa\sqrt{n_{\mathrm{T}}}}\right)$, we have

$$\int f(\mathbf{W}_k, \mathrm{M}_j|\mathbf{X}_k, \mathrm{T}_j, \mathbf{V}_k, \phi)d\mathrm{M}_j$$

$$= \int f(\mathbf{W}_k|\mathbf{X}_k, \mathrm{T}_j, \mathbf{V}_k, \phi)p\left(\mu_{jk}\right)d\mu_{jk}$$

$$= \left[\frac{1}{\sqrt{2\pi}\vartheta_2\phi^{\frac{1}{2}}}\right]^{n_k} \left[\prod_{l=1}^{n_k}\nu_{kl}^{-\frac{1}{2}}\right] \exp\left\{-\frac{1}{2\vartheta_2^2\phi}\sum_{l=1}^{n_k}\omega_{kl}^2\nu_{kl}^{-1}\right\} \times \qquad (15)$$

$$\sqrt{\frac{\vartheta_2^2\phi}{\vartheta_2^2\phi + \sigma_0^2\sum_{l=1}^{n_k}\nu_{kl}^{-1}}} \exp\left\{\frac{\sigma_0^2\left[\sum_{l=1}^{n_k}\omega_{kl}\nu_{kl}^{-1}\right]^2}{2\vartheta_2^2\phi\left[\vartheta_2^2\phi + \sigma_0^2\sum_{l=1}^{n_k}\nu_{kl}^{-1}\right]}\right\}.$$

To draw from $p\left\{(\mathrm{T}_j, \mathrm{M}_j) | \mathbf{M}_{(-j)}, \mathbf{T}_{(-j)}, \phi, \mathbf{X}, \mathbf{Y}\right\}$, we first obtain a tree $\mathrm{T}_j^*$ as a candidate update to $\mathrm{T}_j$ accepted with a probability

$$\min\left\{1, \frac{q(\mathrm{T}_j^*, \mathrm{T}_j)p(\mathbf{W}|\mathbf{X}, \mathbf{V}, \mathrm{T}_j^*, \phi)p(\mathrm{T}_j^*)}{q(\mathrm{T}_j, \mathrm{T}_j^*)p(\mathbf{W}|\mathbf{X}, \mathbf{V}, \mathrm{T}_j, \phi)p(\mathrm{T}_j)}\right\}. \tag{16}$$

The transition kernel $q(\cdot, \cdot)$ assigns probabilities of 0.25, 0.25, 0.40 and 0.10 to the moves GROW, PRUNE, SWAP, and CHANGE respectively. The GROW move randomly selects a terminal node and proposes a binary split with probability of (7) while its reverse counterpart PRUNE move randomly selects and collapses a pair of terminal node parameters originating from the same parent node. The CHANGE move randomly selects a non-terminal node and changes the splitting variable and value. It affects terminal nodes that are descendants of the node where CHANGE move is applied. However, this move does not change the number of terminal and non-terminal nodes. The SWAP move interchanges the splitting rule of a parent and child non-terminal nodes.

For illustrative purposes, we elaborate on the calculation of the ratio

$$\frac{p(\mathbf{W}|\mathbf{X}, \mathbf{V}, \mathrm{T}_j^*, \phi)}{p(\mathbf{W}|\mathbf{X}, \mathbf{V}, \mathrm{T}_j, \phi)}, \tag{17}$$

which is a component of (16). For the fittingly named GROW move, when a terminal node with $n_p$ observation splits to left and right nodes of size $n_l$ and $n_r$ (the subscripts $p$, $l$ and $r$ denoting "parent", and "left" and "right" child nodes), (17) simplifies through cancellations since a GROW move only affects the terminal node that is being split. That is,

$$\frac{p(\mathbf{W}|\mathbf{X}, \mathbf{V}, \mathrm{T}_j^*, \phi)}{p(\mathbf{W}|\mathbf{X}, \mathbf{V}, \mathrm{T}_j, \phi)} = \frac{p(\mathbf{W}_l|\mathbf{X}_l, \mathbf{V}_l, \mathrm{T}_j^*, \phi)p(\mathbf{W}_r|\mathbf{X}_r, \mathbf{V}_r, \mathrm{T}_j^*, \phi)}{p(\mathbf{W}_p|\mathbf{X}_p, \mathbf{V}_p, \mathrm{T}_j, \phi)} \tag{18}$$

which equals

$$\sqrt{\frac{\vartheta_2^2\phi\left(\vartheta_2^2\phi + \sigma_0^2 B_p\right)}{\left(\vartheta_2^2\phi + \sigma_0^2 B_r\right)\left(\vartheta_2^2\phi + \sigma_0^2 B_l\right)}} \times$$
$$\exp\left\{\frac{\sigma_0^2}{2\vartheta_2^2\phi}\left(\frac{A_r^2}{\vartheta_2^2\phi + \sigma_0^2 B_r} + \frac{A_l^2}{\vartheta_2^2\phi + \sigma_0^2 B_l} - \frac{A_p^2}{\vartheta_2^2\phi + \sigma_0^2 B_p}\right)\right\},$$

where $B_k = \sum_{l=1}^{n_l} \nu_{kl}^{-1}$ and $A_k = \sum_{l=1}^{n_k} \omega_{kl}\nu_{kl}^{-1}$ whose dependence on $\mathbf{V}_k$ and $\mathbf{W}_k$ is suppressed for conciseness. Given an updated tree $\mathrm{T}_j$, its terminal

9

node parameters $M_j = (\mu_{jk}; k = 1, \ldots, m_j)$ are updated by drawing from $p(\mu_{jk}|T_j, \mathbf{V}, \phi, \mathbf{W}, \mathbf{X})$ which upto a proportionality constant is given by

$$\exp\left\{-\frac{1}{2}\left(\frac{\vartheta_2^2\phi + \sigma_0^2\sum_{l=1}^{n_k}\nu_{kl}^{-1}}{\vartheta_2^2\sigma_0^2\phi}\right)\left[\mu_{jk} - \frac{\sigma_0^2\sum_{l=1}^{n_k}\omega_{kl}\nu_{kl}^{-1}}{\vartheta_2^2\phi + \sigma_0^2\sum_{l=1}^{n_k}\nu_{kl}^{-1}}\right]^2\right\}, \qquad (19)$$

indicating a sample from a Gaussian distribution.

In order to update the scale parameter $\phi$, we revert to the original notation of the quantile sum of trees in (5), then draw from Inverse-Gamma distribution

$$p(\phi|\mathbf{M}, \mathbf{T}, \mathbf{Y}, \mathbf{X}, \mathbf{V}) \propto \phi^{-\frac{n}{2}-\frac{\alpha}{2}-1}\exp\left\{-\frac{1}{\phi}\left[\frac{\beta}{2} + \sum_{i=1}^{n}\frac{(y_i - \mathbf{G}(\mathbf{x}_i) - \vartheta_1\nu_i)^2}{2\vartheta_2^2\nu_i}\right]\right\}. \qquad (20)$$

# 3 Data analysis

## 3.1 Simulation study

In this subsection, two simulation studies are conducted. The first uses the function $f : \Re^{10} \to \Re$ given by $f(\mathbf{x}) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + 0(x_6 + x_7 + x_8 + x_9 + x_{10})$, where $x_j \sim \text{Unif}(0, 1)$ for $j = 1, \ldots, 10$. This benchmark data generating function is used in Friedman (1991); Chipman et al. (2010); Gramacy and Lee (2012) among others. The response variable is simulated as $y = f(\mathbf{x}) + \epsilon$, where $\epsilon \overset{d}{=} \pi\epsilon_1 + (1 - \pi)\epsilon_2$, $\pi \sim \text{Bern}(0.8)$, $\epsilon_1 \sim \text{N}(0, 1)$ and $\epsilon_2 \sim \text{N}(1, 4)$. Note that it includes non-linear, linear, interaction effects as well as predictors that do not affect the response variable. The model evaluation metric used is the mean weighted absolute deviation given by

$$\text{MWAD} = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(\hat{y}_i - y_i), \qquad (21)$$

where $\rho_\tau(\omega) = \omega(\tau - I\{\omega < 0\})$ is the "check" loss function, $\hat{y}_i$ is the estimated conditional $\tau^{\text{th}}$ quantile and $y_i$ is the actual response value of the $i$th observation in the evaluation data set. Twenty replications of training and test data sets of 100 observations each are simulated and test data set performance comparisons for Bayesian quantile additive regression trees (BayesQArt), Bayesian quantile regression with adaptive Lasso regularization

|            | $\tau = 0.25$    | $\tau = 0.50$    | $\tau = 0.75$    |
|------------|------------------|------------------|------------------|
| BayesQArt  | 0.7190 (0.0243)  | 0.9236 (0.0228)  | 0.6795 (0.0170)  |
| QRF        | 0.9215 (0.0228)  | 1.1430 (0.0243)  | 1.0123 (0.0171)  |
| BayesQR.AL | 0.8577 (0.0065)  | 1.0274 (0.0069)  | 0.8298 (0.0083)  |
| QReg.AL    | 0.8395 (0.0194)  | 1.0132 (0.0211)  | 0.8130 (0.0151)  |

Table 1: First quantile regression simulation study results: test data average mean weighted absolute deviations (MWADs) (21) and standard errors in parentheses over 20 replications.

(BayesQR.AL) in Alhamzawi et al. (2012); Li et al. (2010), linear regression quantiles with adaptive Lasso regularization (QReg.AL) in Zou and Yuan (2008) and quantile random forest (QRF) in Meinshausen (2006) are reported. Our proposed method shows very good predictive performance with lower mean weighted absolute deviation than the competing procedures in estimating the $25^{\text{th}}$, $50^{\text{th}}$ and $75^{\text{th}}$ conditional quantiles as displayed in Table 1, underscoring its robustness to the presence of intricate relationships between predictors and the dependent variable. Figure 2 displays the predicted conditional quantiles against the actual response values.

In the second simulation study, a data set with 30 predictors of which 10 do not impact the response in any form is generated. The data generating scheme is based on the heteroskedastic error model in He (1997)

$$y = \mathbf{x}^{\mathsf{T}}\beta + \left(\mathbf{x}^{\mathsf{T}}\gamma\right)\epsilon, \tag{22}$$

where $\epsilon \sim \mathrm{N}\left(0,1\right)$, $\mathbf{x} \in \Re^{30}$, $\beta = \begin{pmatrix} 1_{20\times 1} \\ 0_{10\times 1} \end{pmatrix}$, $\gamma = \begin{pmatrix} 1_{5\times 1} \\ 0_{25\times 1} \end{pmatrix}$, and $1_{m\times 1}$ and $0_{m\times 1}$ denoting column vectors of ones and zeros of size $m$. Each component of $\mathbf{x}$ is generated independently from $\mathrm{Unif}\left(0,1\right)$. The results of this simulation study for estimation of $25^{\text{th}}$, $50^{\text{th}}$ and $75^{\text{th}}$ conditional quantiles is reported in Table 2 on twenty replications of training and test data sets of 100 observations each. For the estimation of the $50^{\text{th}}$ and $75^{\text{th}}$ conditional quantiles, the linear models have better performance than our method or quantile random forest. This is expected given the underlying data generating process assumes a linear relationship between the predictors and the dependent variable. Our method performs well showing better results than quantile random forests for the estimation of the $50^{\text{th}}$ and $75^{\text{th}}$ conditional quantiles.
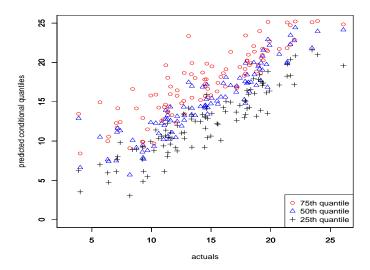
Figure 2: Predicted conditional quantiles against the actual response for the first simulation study.

|  | $\tau = 0.25$ | $\tau = 0.50$ | $\tau = 0.75$ |
|---|---|---|---|
| BayesQArt | 0.4864 (0.0137) | 0.5367 (0.0087) | 0.3619 (0.0045) |
| QRF | 0.4601 (0.0179) | 0.6327 (0.0166) | 0.5143 (0.0067) |
| BayesQR.AL | 0.7601 (0.0042) | 0.5083 (0.0028) | 0.2549 (0.0014) |
| QReg.AL | 0.7624 (0.0042) | 0.5082 (0.0028) | 0.2541 (0.0014) |

Table 2: Second quantile regression simulation study results: test data average mean weighted absolute deviations (MWADs) (21) and standard errors in parentheses based on 20 replications.

|              | $\tau = 0.25$      | $\tau = 0.50$      | $\tau = 0.75$      |
|--------------|--------------------|--------------------|--------------------|
| BayesQArt    | 4.7855 (0.4967)    | 6.7465 (0.5402)    | 6.0362 (0.7636)    |
| QRF          | 4.4835 (0.7453)    | 6.4043 (0.7731)    | 5.7270 (0.6917)    |
| BayesQR.AL   | 5.9865 (0.4938)    | 7.6987 (0.6479)    | 7.5316 (0.7150)    |
| QReg.AL      | 6.0649 (0.5171)    | 8.7799 (0.3412)    | 8.0962 (0.3104)    |

Table 3: Ozone data set: test data average mean weighted absolute deviations (MWADs) (21) and standard errors in parentheses based on 5 consecutive splits of the data as they appear in the $R$ package *datasets*.

## 3.2 Real data examples

The first data used for illustrating Bayesian quantile additive regression trees is the airquality data from the $R$ package *datasets*. This data set records the ozone levels (in parts per billion) in New York from May to September 1973. The predictors used are a measure of solar radiation level, wind speed, maximum daily temperature, and month and day of measurement. We estimate the 25th, 50th and 75th conditional quantile ozone level using competing statistical procedures in Section 3.1 and Bayesian quantile additive regression trees. After removing observations with missing records, we split the data into five nearly equal partitions. Table 3 reports the mean weighted absolute deviations and standard errors.

The second real data set considered is an auto insurance data consisting of 2,812 auto insurance policyholders with 56 predictors along with an aggregate paid claim amount. This data set is available in the $R$ package *HDtweedie* (Qian et al., 2015). Examples of the predictors are driver's age, driver's income, use of vehicle (commercial or not), vehicle type (either of 6 categories), and driver's gender. The response variable is the aggregate claim amount and it is skewed with substantial policyholders having zero claims. When the claim amounts are non-zero, larger claim amounts tend to be reported.

Insurers are often interested in understanding the distribution of claim amounts conditional on a set of policyholder and policy characteristics with added emphasis on higher quantiles. Neither the existence of claims nor the amount if a claim occurs is known at the time of the policy purchase. Hence, insurers use estimates of future claims to appropriately price the insurance

|              | $\tau = 0.90$       | $\tau = 0.95$       |
|--------------|---------------------|---------------------|
| BayesQArt    | 1.4487 (0.0690)     | 1.0440 (0.0571)     |
| QRF          | 1.4862 (0.0676)     | 1.0656 (0.0522)     |
| BayesQR.AL   | 1.4508 (0.0681)     | 1.0483 (0.0602)     |
| QReg.AL      | 1.4542 (0.0671)     | 1.0559 (0.0603)     |

Table 4: Auto insurance claims data set: test data average mean weighted absolute deviations (MWADs) (21) and standard errors in parentheses based on 10 splits.

product and also to set aside sufficient amount of monetary reserves to pay future claims. Thus, we estimate the 90th and 95th conditional quantiles by splitting the data set into 10 nearly equal partitions each time using nine-tenth of the data for training and the remaining for testing Bayesian quantile additive regression trees and the statistical procedures in Section 3.1. Table 4 displays the predictive performances of each procedure and our method performs very well. Note that we are intentionally using the regularized versions of the procedures Bayesian linear quantile regression and the classical quantile regression since variable selection is a component of these procedures.

# 4 Binary classification extension

In this section, we extend Bayesian quantile additive regression trees to tackle binary classification problems. Kordas (2006); Benoit and Van den Poel (2012); Benoit et al. (2013) among others consider the binary classification problem in a quantile regression framework. Suppose $y_i \in \{0, 1\}$ and $\mathbf{x}_i \in \Re^d$ are the binary response and the predictors for the $i$[th] observation. Suppose also that there is an unobserved latent variable $\tilde{y}_i$ for $i = 1, \ldots, n$ such that $y_i = 1$ if $\tilde{y}_i > 0$ and $y_i = 0$ otherwise. The goal of the classification problem is to obtain an estimate $\hat{P}(y_i = 1 \mid \mathbf{x}_i)$ for $P(y_i = 1 \mid \mathbf{x}_i)$ which we obtain via

the hierarchical model

$$y_i | \tilde{y}_i, \nu_i, \mathbf{G}, \mathbf{T}, \mathbf{M} \sim \mathrm{Bern}\left(\mathrm{P}\left(\tilde{y}_i > 0 \mid \nu_i, \mathbf{G}, \mathbf{T}, \mathbf{M}\right)\right)$$

$$\tilde{y}_i \mid \nu_i, \mathbf{G}, \mathbf{T}, \mathbf{M} \sim \mathrm{N}\left(\mathbf{G}\left(\mathbf{x}_i; \mathbf{T}, \mathbf{M}\right) + \vartheta_1 \nu_i, \vartheta_2^2 \phi \nu_i\right)$$

$$\mathbf{G}\left(\mathbf{x}_i; \mathbf{T}, \mathbf{M}\right) = \sum_{j=1}^{n_\mathrm{T}} g\left(\mathbf{x}_i; \mathrm{T}_j, \mathrm{M}_j\right) \tag{23}$$

$$\nu_i \mid \phi \sim \mathrm{Exp}\left(\phi\right),$$

$$z_i \sim \mathrm{N}\left(0, 1\right),$$

where $\mathbf{G}\left(\mathbf{x}_i; \mathbf{T}, \mathbf{M}\right)$ is a sum of regression trees. The prior specifications for $\mathbf{T}$, $\mathbf{M}$ are as specified in (5). The posterior computation cycles through the following MCMC steps. Sequential draws from truncated normal distributions to sample from the latent variable $\tilde{y}_i$ for $i = 1, \ldots, n$

$$\tilde{y}_i \mid y_i, \mathbf{x}_i, \nu_i, \mathbf{T}, \mathbf{M}, \phi \sim \mathrm{N}\left(\mathbf{G}\left(\mathbf{x}_i; \mathbf{T}, \mathbf{M}\right) + \vartheta_1 \nu_i, \vartheta_2^2 \phi \nu_i\right) \mathrm{I}\left(y_i = 1, \tilde{y}_i \geq 0\right)$$
$$+ \mathrm{N}\left(\mathbf{G}\left(\mathbf{x}_i; \mathbf{T}, \mathbf{M}\right) + \vartheta_1 \nu_i, \vartheta_2^2 \phi \nu_i\right) \mathrm{I}\left(y_i = 0, \tilde{y}_i < 0\right), \tag{24}$$

followed by draws from $\mathbf{V}, \left(\mathbf{T}, \mathbf{M}\right),$ and $\phi$ as described in Section 2.2 with $\tilde{\mathbf{Y}} = \left(\tilde{y}_1, \ldots, \tilde{y}_n\right)^{\mathsf{T}}$ considered as the response vector of the Bayesian quantile additive regression model.

## 4.1   Binary classification simulation study

We simulate a binary classification data set with ten predictors using the data generating scheme known as "cicle" from the $R$ package *mlbench* (Leisch and Dimitriadou, 2010) which has often been considered as a benchmark classification data set (Chung and Kim, 2015; Ishwaran, 2015; Rudnicki et al., 2015). Suppose $\mathbf{x} \in [-1, 1]^d$, with $x_j \sim \mathrm{Unif}\left(-1, 1\right)$, $j = 1, \ldots, d$, where we take $d = 10$. The goal of this classification problem is to identify if the coordinate $(x_1, \ldots, x_d)$ in a $d$ dimensional hypercube with edges at all sign permutations of the coordinates $\{\pm 1, \ldots, \pm 1\}$ lies outside of a hypersphere which lies inside the hypercube. That is, $y = 1$ if $\sum_{j=1}^d x_j^2 > r^2$, otherwise $y = 0$. The radius of the hypersphere, $r$, is chosen so that there is nearly equal representation between the two classes. The class boundaries are non-linear making it an interesting classification problem (see Figure 3 which shows the class boundary for the two dimensional case).

   We simulate training and test data sets of size 100 each and report the averages of classification error rate and area under the ROC curve over

| Procedure | Error Rate | AUC |
| --- | --- | --- |
| BayesQArt | 0.2185 (0.0100) | 0.8297 (0.0088) |
| RF | 0.2600 (0.0112) | 0.6363 (0.0130) |
| BayesQR.AL | 0.5500 (0.0133) | 0.4684 (0.0078) |

Table 5: Simulation study for binary classification: test data averages of test data classification error rate, area under the ROC curve (AUC), and their standard errors in parentheses based on 20 replications.

twenty replications for binary Bayesian quantile additive regression trees (BayesQArt), binary Bayesian linear quantile regression (BayesQR) and random forests (RF). Note that the random forest procedure used for classification in this section is one described in Breiman (2001) and not the quantile random forest in Meinshausen (2006). For an evaluation data set with $m$ observations, classification error rate is computed as

$$\text{Error Rate} = \frac{1}{m} \sum_{i=1}^{m} y_i \neq \hat{y}_i.$$

## 4.2 Real data for binary classification

We consider a binary classification real data example in which the number of predictors is much larger than the sample size to illustrate the predictive performance of the binary extension of Bayesian quantile additive regression trees. The goal for this data set is to classify whether a patient has cancer (ovarian or prostate cancer) based on 10,000 predictors of which a portion is mass-spectra data and the other portion consisting of unimportant predictors. This data set is obtained from a data set named "arcene" at the UCI machine learning repository (Bache and Lichman, 2013). The training and validation data sets combined contain 200 patient samples. Additional details on this data set are in Guyon et al. (2007, 2008). We split the data into five nearly equal partitions and report the averages of test data classification error rate and area under the ROC curve for binary Bayesian quantile additive regression trees (BayesQArt) and random forests (RF). Results in Table 6 show that our proposed method handles regression problems in which the
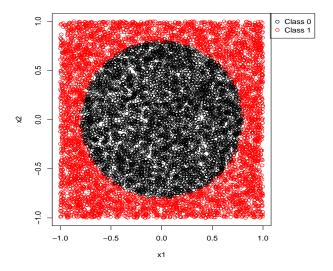
Figure 3: circle in square - two dimensional illustration of class boundaries of the binary classification simulation study for Bayesian quantile additive regression trees.

number of predictors is much larger than the number samples in the training data while exhibiting very good predictive performance. For this example, we only used the moves GROW and PRUNE to reduce to the computational cost. An average computing time of 5.42 minutes is recorded using a 64-bit Windows personal computer with the specifications: *Intel Core i5-2320, 3.0GHz and 6.0GB* installed memory.

| Procedure | Error Rate | AUC |
|-----------|------------|-----|
| BayesQArt | 0.1650 (0.0170) | 0.8712 (0.0165) |
| RF | 0.1800 (0.0094) | 0.8184 (0.0110) |

Table 6: Cancer classification results: test data averages of classification error rate, area under the ROC curve (AUC), and their standard errors.

# 5  Conclusion

This article proposed a Bayesian sum of regression trees model for estimating conditional quantiles. The asymmetric Laplace distribution likelihood is employed with its mixture representation enabling tractable posterior computation of the regression trees in the sum and their terminal node parameters.

Simulation studies with data generating schemes that included linear, non-linear, interaction effects as well as unimportant predictors illustrated that Bayesian quantile additive regression trees has very good predictive performance. Real data applications dealing with insurance claims and ozone level prediction demonstrated that the proposed method complements existing powerful statistical procedures.

We also successfully extended and tested the proposed procedure to tackle binary classification problems. The proposed method exhibited very good out-of-sample classification accuracy in a simulation study characterized by a non-linear class boundary and cancer classification example in which the number of predictors is about fifty times as much as the number of samples in the training data. The source code for the implementation of our proposed method, and the selected tuning parameters for the simulation studies and real data applications are at https://github.com/bpkindo/bayesqart.

# References

Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair. A comparison of machine learning techniques for phishing detection. In *Proceedings of the Anti-phishing Working Groups 2nd Annual eCrime Researchers Summit*, eCrime '07, pages 60–69. ACM, 2007. ISBN 978-1-59593-939-5. 1, 1

Rahim Alhamzawi, Keming Yu, and Dries F Benoit. Bayesian adaptive Lasso quantile regression. *Statistical Modelling*, 12(3):279–297, 2012. 3.1

K. Bache and M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml. 4.2

Alexandre Belloni and Victor Chernozhukov. l1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011. 1

Dries F Benoit and Dirk Van den Poel. Binary quantile regression: a Bayesian approach based on the asymmetric Laplace distribution. *Journal of Applied Econometrics*, 27(7):1174–1188, 2012. 4

Dries F. Benoit, Rahim Alhamzawi, and Keming Yu. Bayesian Lasso binary quantile regression. *Computational Statistics*, 28(6):2861–2873, 2013. 4

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 1, 4.1

Moshe Buchinsky. Quantile regression, box-cox transformation model, and the US wage structure, 1963–1987. *Journal of Econometrics*, 65(1):109–154, 1995. 1

Moshe Buchinsky. Recent advances in quantile regression models: A practical guideline for empirical research. *Journal of Human Resources*, 33(1), 1998. 1

Lane F Burgette, Jerome P Reiter, and Marie Lynn Miranda. Exploratory quantile regression with many covariates: an application to adverse birth outcomes. *Epidemiology*, 22(6):859–866, 2011. 1

Probal Chaudhuri and Wei-Yin Loh. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, 8(5):561–576, 2002. 1

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93 (443):935–948, 1998. 1, 2.1, 2.2

Hugh A Chipman, Edward I George, and Robert E McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1): 266–298, 2010. 1, 2.2, 3.1

Dongjun Chung and Hyunjoong Kim. Accurate ensemble pruning with PL-bagging. *Computational Statistics & Data Analysis*, 83:1–13, 2015. 4.1

Timothy J Cole and Pamela J Green. Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, 11(10): 1305–1319, 1992. 1

David B Dunson and Jack A Taylor. Approximate Bayesian inference for quantiles. *Nonparametric Statistics*, 17(3):385–400, 2005. 1

P Friederichs and A Hense. Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, 135 (6), 2007. 1

Jerome H Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, pages 1–67, 1991. 3.1

Robert B Gramacy and Herbert KH Lee. Cases for the nugget in modeling computer experiments. *Statistics and Computing*, 22(3):713–722, 2012. 3.1

Isabelle Guyon, Jiwen Li, Theodor Mader, Patrick A Pletscher, Georg Schneider, and Markus Uhr. Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark. *Pattern Recognition Letters*, 28(12):1438–1444, 2007. 4.2

Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008. 4.2

Timothy Hanson and Wesley O Johnson. Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, 97(460), 2002. 1

Shan He, Xiaoli Li, Mark R Viant, and Xin Yao. Profiling MS proteomics data using smoothed non-linear energy operator and Bayesian additive regression trees. *Proteomics*, 9(17):4176–4191, 2009. 1

Xuming He. Quantile curves without crossing. *The American Statistician*, 51(2):186–192, 1997. 3.1

Hemant Ishwaran. The effect of splitting on random forests. *Machine Learning*, 99(1):75–118, 2015. 4.1

Adam Kapelner and Justin Bleich. bartmachine: Machine learning with Bayesian additive regression trees. *arXiv preprint arXiv:1312.2171*, 2013. 1

Bereket P. Kindo, Hao Wang, and Edsel A. Peña. Multinomial probit Bayesian additive regression trees. *Stat*, pages n/a–n/a, 2016. ISSN 2049-1573. doi: 10.1002/sta4.110. URL http://dx.doi.org/10.1002/sta4.110. 1, 2.2

Roger Koenker. Confidence intervals for regression quantiles. In *Asymptotic Statistics*, pages 349–359. Springer, 1994. 1

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978. 1, 1

Gregory Kordas. Smoothed binary regression quantiles. *Journal of Applied Econometrics*, 21(3):387–407, 2006. 4

Athanasios Kottas and Alan E Gelfand. Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, 96 (456):1458–1468, 2001. 1

Hideo Kozumi and Genya Kobayashi. Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81 (11):1565–1578, 2011. 1, 1, 1

Friedrich Leisch and Evgenia Dimitriadou. mlbench: Machine learning benchmark problems. *R package*, 2:1–1, 2010. 4.1

Qing Li, Ruibin Xi, and Nan Lin. Bayesian regularized quantile regression. *Bayesian Analysis*, 5(3):533–556, 09 2010. doi: 10.1214/10-BA521. URL http://dx.doi.org/10.1214/10-BA521. 3.1

Yang Liu, Mikhail Traskin, Scott A Lorch, Edward I George, and Dylan Small. Ensemble of trees approaches to risk adjustment for evaluating a hospitals performance. *Health Care Management Science*, 18(1):58–66, 2015. 1

Emanuela Marrocu, Raffaele Paci, and Andrea Zara. Micro-economic determinants of tourist expenditure: A quantile regression approach. *Tourism Management*, 50:13–30, 2015. 1

Nicolai Meinshausen. Quantile regression forests. *The Journal of Machine Learning Research*, 7:983–999, 2006. 1, 1, 3.1, 4.1

Thomas Q Pedersen. Predictable return distributions. *Journal of Forecasting*, 34(2):114–132, 2015. 1

Matthew T Pratola, Hugh A Chipman, James R Gattiker, David M Higdon, Robert McCulloch, and William N Rust. Parallel Bayesian additive regression trees. *Journal of Computational and Graphical Statistics*, 23(3): 830–852, 2014. 1

Wei Qian, Yi Yang, and Hui Zou. Tweedie's compound poisson model with grouped elastic net. *Journal of Computational and Graphical Statistics*, 0(ja):0–0, 2015. doi: 10.1080/10618600.2015.1005213. URL http://dx.doi.org/10.1080/10618600.2015.1005213. 3.2

Brian J Reich, Howard D Bondell, and Huixia J Wang. Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics*, 11 (2):337–352, 2010. 1

Witold R Rudnicki, Mariusz Wrzesień, and Wiesław Paja. All relevant feature selection methods and applications. In *Feature Selection for Data and Pattern Recognition*, pages 11–28. Springer, 2015. 4.1

Karthik Sriram, RV Ramamoorthi, and Pulak Ghosh. Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Analysis*, 8(2):479–504, 2013. 1

Matthew A Taddy and Athanasios Kottas. A Bayesian nonparametric approach to inference for quantile regression. *Journal of Business & Economic Statistics*, 2012. 1

I-Chun Tsai. The relationship between stock price index and exchange rate in asian markets: A quantile regression approach. *Journal of International Financial Markets, Institutions and Money*, 22(3):609 – 621, 2012. ISSN 1042-4431. doi: http://dx.doi.org/10.1016/j.intfin.2012.04.005. URL http://www.sciencedirect.com/science/article/pii/S1042443112000297. 1

Jiansheng Wu, Liangyong Huang, and Xiaoming Pan. A novel Bayesian additive regression trees ensemble model based on linear regression and nonlinear regression for torrential rain forecasting. In *Computational Science and Optimization (CSO), 2010 Third International Joint Conference on*, volume 2, pages 466–470. IEEE, 2010. 1

Keming Yu and Rana A Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001. 1, 1, 1

Junni L Zhang and Wolfgang K Härdle. The Bayesian additive classification tree applied to credit risk modelling. *Computational Statistics & Data Analysis*, 54(5):1197–1205, 2010. 1, 1

Hui Zou and Ming Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, pages 1108–1126, 2008. 1, 3.1