Journal of Machine Learning Research 7 (2006) 983–999 Submitted 10/05; Revised 2/06; Published 6/06

Quantile Regression Forests

Nicolai Meinshausen NICOLAI@STAT.MATH.ETHZ.CH
Seminar für Statistik
ETH Zürich
8092 Zürich, Switzerland

Editor: Greg Ridgeway

Abstract

Random forests were introduced as a machine learning tool in Breiman (2001) and have since proven to be very popular and powerful for high-dimensional regression and classification. For regression, random forests give an accurate approximation of the conditional mean of a response variable. It is shown here that random forests provide information about the full conditional distribution of the response variable, not only about the conditional mean. Conditional quantiles can be inferred with quantile regression forests, a generalisation of random forests. Quantile regression forests give a non-parametric and accurate way of estimating conditional quantiles for high-dimensional predictor variables. The algorithm is shown to be consistent. Numerical examples suggest that the algorithm is competitive in terms of predictive power.

Keywords: quantile regression, random forests, adaptive neighborhood regression

1. Introduction

Let Y be a real-valued response variable and X a covariate or predictor variable, possibly high-dimensional. A standard goal of statistical analysis is to infer, in some way, the relationship between Y and X. Standard regression analysis tries to come up with an estimate $\hat{\mu}(x)$ of the conditional mean $E(Y|X = x)$ of the response variable Y, given $X = x$. The conditional mean minimizes the expected squared error loss,

$$E(Y|X = x) = \arg\min_z E\{(Y - z)^2|X = x\},$$

and approximation of the conditional mean is typically achieved by minimization of a squared error type loss function over the available data.

Beyond the Conditional Mean The conditional mean illuminates just one aspect of the conditional distribution of a response variable Y, yet neglects all other features of possible interest. This led to the development of quantile regression; for a good summary see e.g. Koenker (2005). The conditional distribution function $F(y|X = x)$ is given by the probability that, for $X = x$, Y is smaller than $y \in \mathbb{R}$,

$$F(y|X = x) = P(Y \leq y|X = x).$$

For a continuous distribution function, the α-quantile $Q_\alpha(x)$ is then defined such that the probability of Y being smaller than $Q_\alpha(x)$ is, for a given $X = x$, exactly equal to α. In general,

$$Q_\alpha(x) = \inf\{y : F(y|X = x) \geq \alpha\}. \tag{1}$$

The quantiles give more complete information about the distribution of Y as a function of the predictor variable X than the conditional mean alone.

As an example, consider the predictions of next day ozone levels, as in Breiman and Friedman (1985). Least-squares regression tries to estimate the conditional mean of ozone levels. It gives little information about the fluctuations of ozone levels around this predicted conditional mean. It might for example be of interest to find an ozone level that is -with high probability- not surpassed. This can be achieved with quantile regression, as it gives information about the spread of the response variable. For some other examples see Le et al. (2005), which is to the best of our knowledge the first time that quantile regression is mentioned in the Machine Learning literature.

Prediction Intervals How reliable is a prediction for a new instance? This is a related question of interest. Consider again the prediction of next day ozone levels. Some days, it might be possible to pinpoint next day ozone levels to a higher accuracy than on other days (this can indeed be observed for the ozone data, see the section with numerical results). With standard prediction, a single point estimate is returned for each new instance. This point estimate does not contain information about the dispersion of observations around the predicted value.

Quantile regression can be used to build prediction intervals. A 95% prediction interval for the value of Y is given by

$$I(x) = [Q_{.025}(x), Q_{.975}(x)]. \tag{2}$$

That is, a new observation of Y, for $X = x$, is with high probability in the interval $I(x)$. The width of this prediction interval can vary greatly with x. Indeed, going back to the previous example, next day ozone level can on some days be predicted five times more accurately than on other days. This effect is even more pronounced for other data sets. Quantile regression offers thus a principled way of judging the reliability of predictions.

Outlier Detection Quantile regression can likewise be used for outlier detection (for surveys on outlier detection see e.g. Barnett and Lewis, 1994; Hodge and Austin, 2004). A new observation (X, Y) would be regarded as an outlier if its observed value Y is extreme, in some sense, with regard to the predicted conditional distribution function.

There is, however, no generally applicable rule of what precisely constitutes an “extreme” observation. One could possibly flag observations as outliers if the distance between Y and the median of the conditional distribution is large; “large” being measured in comparison to some robust measure of dispersion like the conditional median absolute deviation or the conditional interquartile range (Huber, 1973). Both quantities are made available by quantile regression.

Note that only anomalies in the conditional distribution of Y can be detected in this way. Outliers of X itself cannot be detected. Other research has focused on detecting anomalies for unlabelled data (e.g. Markou and Singh, 2003; Steinwart et al., 2005).

Estimating Quantiles from Data Quantile regression aims to estimate the conditional quantiles from data. Quantile regression can be cast as an optimization problem, just as estimation of the conditional mean is achieved by minimizing a squared error loss function. Let the loss function L_α be defined for $0 < \alpha < 1$ by the weighted absolute deviations

$$L_\alpha(y, q) = \begin{cases} \alpha \, |y - q| & y > q \\ (1 - \alpha) \, |y - q| & y \leq q \end{cases} . \tag{3}$$

While the conditional mean minimizes the expected squared error loss, conditional quantiles minimize the expected loss $E(L_\alpha)$,

$$Q_\alpha(x) = \arg\min_q \; E\{L_\alpha(Y, q) | X = x\}.$$

A parametric quantile regression is solved by optimizing the parameters so that the empirical loss is minimal. This can be achieved efficiently due to the convex nature of the optimization problem (Portnoy and Koenker, 1997). Non-parametric approaches, in particular quantile Smoothing Splines (He et al., 1998; Koenker et al., 1994), involve similar ideas. Chaudhuri and Loh (2002) developed an interesting tree-based method for estimation of conditional quantiles which gives good performance and allows for easy interpretation, being in this respect similar to CART (Breiman et al., 1984).

In this manuscript, a different approach is proposed, which does not directly employ minimization of a loss function of the sort (3). Rather, the method is based on random forests (Breiman, 2001). Random forests grows an ensemble of trees, employing random node and split point selection, inspired by Amit and Geman (1997). The prediction of random forests can then be seen as an adaptive neighborhood classification and regression procedure (Lin and Jeon, 2002). For every $X = x$, a set of weights $w_i(x)$, $i = 1, \ldots, n$ for the original n observations is obtained. The prediction of random forests, or estimation of the conditional *mean*, is equivalent to the *weighted mean of the observed response variables.* For quantile regression forests, trees are grown as in the standard random forests algorithm. The conditional *distribution* is then estimated by the *weighted distribution of observed response variables*, where the weights attached to observations are identical to the original random forests algorithm.

In Section 2, necessary notation is introduced and the mechanism of random forests is briefly explained, using the interpretation of Lin and Jeon (2002), which views random forests as an adaptive nearest neighbor algorithm, a view that is later supported in Breiman (2004). Using this interpretation, quantile regression forests are introduced in Section 3 as a natural generalisation of random forests. A proof of consistency is given in Section 4, while encouraging numerical results for popular machine learning data sets are presented in Section 5.

2. Random Forests

Random forests grows an ensemble of trees, using n independent observations

$$(Y_i, X_i), \quad i = 1, \ldots, n.$$

A large number of trees is grown. For each tree and each node, random forests employs randomness when selecting a variable to split on. For each tree, a bagged version of the training data is used. In addition, only a random subset of predictor variables is considered for splitpoint selection at each node. The size of the random subset, called *mtry*, is the single tuning parameter of the algorithm, even though results are typically nearly optimal over a wide range of this parameter. The value of *mtry* can be fine-tuned on the out-of-bag samples. For regression, the prediction of random forests for a new data point $X = x$ is the averaged response of all trees. For details see Breiman (2001). The algorithm is somewhat related to boosting (Schapire et al., 1998), with trees as learners. Yet, with random forests, each tree is grown using the original observations of the response variable, while boosting tries to fit the residuals after taking into account the prediction of previously generated trees (Friedman et al., 2000).

Some Notation Following the notation of Breiman (2001), call θ the random parameter vector that determines how a tree is grown (e.g. which variables are considered for splitpoints at each node). The corresponding tree is denoted by $T(\theta)$. Let $\mathcal{B}$ be the space in which X lives, that is $X : \Omega \mapsto \mathcal{B} \subseteq \mathbb{R}^p$, where $p \in \mathbb{N}_+$ is the dimensionality of the predictor variable. Every leaf $\ell = 1, \ldots, L$ of a tree corresponds to a rectangular subspace of $\mathcal{B}$. Denote this rectangular subspace by $R_\ell \subseteq \mathcal{B}$ for every leaf $\ell = 1, \ldots, L$. For every $x \in \mathcal{B}$, there is one and only one leaf ℓ such that $x \in R_\ell$ (corresponding to the leaf that is obtained when dropping x down the tree). Denote this leaf by $\ell(x, \theta)$ for tree $T(\theta)$.

The prediction of a single tree $T(\theta)$ for a new data point $X = x$ is obtained by averaging over the observed values in leaf $\ell(x, \theta)$. Let the weight vector $w_i(x, \theta)$ be given by a positive constant if observation X_i is part of leaf $\ell(x, \theta)$ and 0 if it is not. The weights sum to one, and thus

$$w_i(x, \theta) = \frac{1_{\{X_i \in R_{\ell(x,\theta)}\}}}{\#\{j : X_j \in R_{\ell(x,\theta)}\}}. \tag{4}$$

The prediction of a single tree, given covariate $X = x$, is then the weighted average of the original observations $Y_i, i = 1, \ldots, n$,

$$\text{single tree:} \quad \hat{\mu}(x) = \sum_{i=1}^{n} w_i(x, \theta)\, Y_i.$$

Using random forests, the conditional mean $E(Y|X = x)$ is approximated by the averaged prediction of k single trees, each constructed with an i.i.d. vector θ_t, $t = 1, \ldots, k$. Let $w_i(x)$ be the average of $w_i(\theta)$ over this collection of trees,

$$w_i(x) = k^{-1} \sum_{t=1}^{k} w_i(x, \theta_t). \tag{5}$$

The prediction of random forests is then

$$\text{Random Forests:} \quad \hat{\mu}(x) = \sum_{i=1}^{n} w_i(x) Y_i.$$

The approximation of the conditional mean of Y, given $X = x$, is thus given by a weighted sum over all observations. The weights vary with the covariate $X = x$ and tend to be large for those $i \in \{1, \ldots, n\}$ where the conditional distribution of Y, given $X = X_i$, is similar to the conditional distribution of Y, given $X = x$ (Lin and Jeon, 2002).

3. Quantile Regression Forests

It was shown above that random forests approximates the conditional mean $E(Y|X = x)$ by a weighted mean over the observations of the response variable Y. One could suspect that the weighted observations deliver not only a good approximation to the conditional mean but to the full conditional distribution. The conditional distribution function of Y, given $X = x$, is given by

$$F(y|X = x) = P(Y \leq y|X = x) = E(1_{\{Y \leq y\}}|X = x).$$

The last expression is suited to draw analogies with the random forest approximation of the conditional mean $E(Y|X = x)$. Just as $E(Y|X = x)$ is approximated by a weighted mean over the observations of Y, define an approximation to $E(1_{\{Y \leq y\}}|X = x)$ by the weighted mean over the observations of $1_{\{Y \leq y\}}$,

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x)\ 1_{\{Y_i \leq y\}}, \tag{6}$$

using the same weights $w_i(x)$ as for random forests, defined in equation (5). This approximation is at the heart of the quantile regression forests algorithm.

The Algorithm The algorithm for computing the estimate $\hat{F}(y|X = x)$ can be summarized as:

a) Grow k trees $T(\theta_t)$, $t = 1, \ldots, k$, as in random forests. However, for every leaf of every tree, take note of all observations in this leaf, not just their average.

b) For a given $X = x$, drop x down all trees. Compute the weight $w_i(x, \theta_t)$ of observation $i \in \{1, \ldots, n\}$ for every tree as in (4). Compute weight $w_i(x)$ for every observation $i \in \{1, \ldots, n\}$ as an average over $w_i(x, \theta_t)$, $t = 1, \ldots, k$, as in (5).

c) Compute the estimate of the distribution function as in (6) for all $y \in \mathbb{R}$, using the weights from Step b).

Estimates $\hat{Q}_\alpha(x)$ of the conditional quantiles $Q_\alpha(x)$ are obtained by plugging $\hat{F}(y|X = x)$ instead of $F(y|X = x)$ into (1). Other approaches for estimating quantiles from empirical distribution functions are discussed in Hyndman and Fan (1996).

The key difference between quantile regression forests and random forests is as follows: for each node in each tree, random forests keeps only the mean of the observations that fall into this node and neglects all other information. In contrast, quantile regression forests keeps the value of all observations in this node, not just their mean, and assesses the conditional distribution based on this information.

Software is made available as a package `quantregForest` for R (R Development Core Team, 2005). The package builds upon the excellent R-package `randomForest` (Liaw and Wiener, 2002).

4. Consistency

Consistency of the proposed method is shown. Consistency for random forests (when approximating the conditional mean) has been shown for a simplified model of random forests in (Breiman, 2004), together with an analysis of convergence rates. The conditions for growing individual trees are less stringent in the current analysis but no attempt is made to analyze convergence rates.

Assumptions Three assumptions are needed for the proof of consistency. First, an assumption is made about the distribution of covariates.

Assumption 1 $\mathcal{B} = [0,1]^p$ *and* X *uniform on* $[0,1]^p$.

This assumption is just made for notational convenience. Alternatively, one could assume that the density of X is positive and bounded from above and below by positive constants.

Next, two assumptions are made about the construction of individual trees. Denote the node-sizes of the leaves ℓ of a tree constructed with parameter vector θ by $k_\theta(\ell)$, that is $k_\theta(\ell) = \#\{i \in \{1,\ldots,n\} : X_i \in R_{\ell(x,\theta)}\}$.

Assumption 2 *The proportion of observations in a node, relative to all observations, is vanishing for large* n, $\max_{\ell,\theta} k_\theta(\ell) = o(n)$, *for* $n \to \infty$. *The minimal number of observations in a node is growing for large* n, *that is* $1/\min_{\ell,\theta} k_\theta(\ell) = o(1)$, *for* $n \to \infty$.

The first part of this assumption is necessary. The second part could possibly be dropped, with a more involved proof of consistency.

The following assumption concerns the actual construction of trees. An attempt has been made to keep these assumptions as minimal as possible.

Assumption 3 *When finding a variable for a splitpoint, the probability that variable* $m = 1,\ldots,p$ *is chosen for the splitpoint is bounded from below for every node by a positive constant. If a node is split, the split is chosen so that each of the resulting sub-nodes contains at least a proportion* γ *of the observations in the original node, for some* $0 < \gamma \leq 0.5$.

Next, The conditional distribution function is assumed to be Lipschitz continuous.

Assumption 4 *There exists a constant* L *so that* $F(y|X=x)$ *is Lipschitz continuous with parameter* L, *that is for all* $x, x' \in \mathcal{B}$,

$$\sup_y |F(y|X=x) - F(y|X=x')| \ \leq \ L\|x - x'\|_1.$$

Lastly, positive density is assumed.

Assumption 5 *The conditional distribution function* $F(y|X=x)$ *is, for every* $x \in \mathcal{B}$, *strictly monotonously increasing in* y.

This assumption is necessary to derive consistency of quantile estimates from the consistency of distribution estimates.

Consistency Under the made assumptions, consistency of quantile regression forests is shown.

Theorem 1 *Let Assumptions 1-5 be fulfilled. It holds pointwise for every $x \in \mathcal{B}$ that*

$$\sup_{y \in \mathbb{R}} |\hat{F}(y|X = x) - F(y|X = x)| \to_p 0 \qquad n \to \infty.$$

In other words, the error of the approximation to the conditional distribution converges uniformly in probability to zero for $n \to \infty$. Quantile regression forests is thus a consistent way of estimating conditional distributions and quantile functions.

Proof Let the random variables U_i, $i = 1, \dots, n$ be defined as the quantiles of the observations Y_i, conditional on $X = X_i$,

$$U_i = F(Y_i|X = X_i).$$

Note that U_i, $i = 1, \dots, n$ are i.i.d. uniform on $[0, 1]$. For a given $X = X_i$, the event $\{Y_i \le y\}$ is identical to $\{U_i \le F(y|X = X_i)\}$ under Assumption 5. The approximation $\hat{F}(y|x) = \hat{F}(y|X = x)$ can then be written as a sum of two parts,

$$\begin{aligned}
\hat{F}(y|x) &= \sum_{i=1}^{n} w_i(x)\ 1_{\{Y_i \le y\}} = \sum_{i=1}^{n} w_i(x)\ 1_{\{U_i \le F(y|X_i)\}} \\
&= \sum_{i=1}^{n} w_i(x)\ 1_{\{U_i \le F(y|x)\}} + \\
&\quad \sum_{i=1}^{n} w_i(x)\ (1_{\{U_i \le F(y|X_i)\}} - 1_{\{U_i \le F(y|x)\}}).
\end{aligned}$$

The absolute difference between the approximation and the true value is hence bounded by

$$\begin{aligned}
|F(y|x) - \hat{F}(y|x)| &\le |F(y|x) - \sum_{i=1}^{n} w_i(x)\ 1_{\{U_i \le F(y|x)\}}| + \\
&\quad |\sum_{i=1}^{n} w_i(x)(1_{\{U_i \le F(y|X_i)\}} - 1_{\{U_i \le F(y|x)\}})|.
\end{aligned}$$

The first term is a variance-type part, while the second term reflects the change in the underlying distribution (if the distribution would be constant as a function of x, the second term would vanish). Taking supremum over y in the first part leads to

$$\sup_{y \in \mathbb{R}} |F(y|x) - \sum_{i=1}^{n} w_i(x)\ 1_{\{U_i \le F(y|x)\}}| = \sup_{z \in [0,1]} |z - \sum_{i=1}^{n} w_i(x)\ 1_{\{U_i \le z\}}|.$$

Note that $E(1_{\{U_i \le z\}}) = z$, as U_i are i.i.d. uniform random variables on $[0, 1]$. Furthermore $0 \le w_i(x) \le (\min_{\ell,\theta} k_\theta(\ell))^{-1}$. As the weights add to one, $\sum_{i=1}^{n} w_i(x) = 1$, and $(\min_{\ell,\theta} k_\theta(\ell))^{-1} = o(1)$ by Assumption 2, it follows that, for every $x \in \mathcal{B}$,

$$\sum_{i=1}^{n} w_i(x)^2 \to 0 \qquad n \to \infty. \tag{7}$$

and hence, for every $z \in [0,1]$ and $x \in \mathcal{B}$,

$$|z - \sum_{i=1}^{n} w_i(x)\ 1_{\{U_i \le z\}}| = o_p(1) \qquad n \to \infty.$$

By Bonferroni's inequality, the above still holds true if, on the left hand side, the supremum over a finite set of z-values is taken, where the cardinality of this set can grow to infinity for $n \to \infty$. By straightforward yet tedious calculations, it can be shown that the supremum can be extended not only to such a set of z-values, but also to the whole interval $z \in [0,1]$, so that

$$\sup_{z \in [0,1]} |z - \sum_{i=1}^{n} w_i(x)\ 1_{\{U_i \le z\}}| = o_p(1) \qquad n \to \infty.$$

It thus remains to be shown that, for every $x \in \mathcal{B}$,

$$|\sum_{i=1}^{n} w_i(x)(1_{\{U_i \le F(y|X_i)\}} - 1_{\{U_i \le F(y|x)\}})| \to_p 0 \qquad n \to \infty.$$

As U_i, $i = 1, \ldots, n$ are uniform over $[0,1]$, it holds that

$$E(1_{\{U_i \le F(y|X_i)\}} - 1_{\{U_i \le F(y|x)\}}) = F(y|X_i) - F(y|x).$$

Using (7) and independence of all U_i, $i = 1, \ldots, n$, for $n \to \infty$,

$$|\sum_{i=1}^{n} w_i(x)(1_{\{U_i \le F(y|X_i)\}} - 1_{\{U_i \le F(y|x)\}})| \to_p \sum_{i=1}^{n} w_i(x)\{F(y|X_i) - F(y|x)\}.$$

Using Assumption 4 about Lipschitz continuity of the distribution function, it thus remains to show that

$$\sum_{i=1}^{n} w_i(x)\ \|x - X_i\|_1 = o_p(1) \qquad n \to \infty.$$

Note that $w_i(x) = k^{-1} \sum_{t=1,\ldots,k} w_i(x, \theta_t)$, where $w_i(x,\theta)$ is defined as the weight produced by a single tree with (random) parameter θ_t, as in (4). Thus it suffices to show that, for a single tree,

$$\sum_{i=1}^{n} w_i(x,\theta)\ \|x - X_i\|_1 = o_p(1) \qquad n \to \infty. \tag{8}$$

The rectangular subspace $R_{\ell(x,\theta)} \subseteq [0,1]^p$ of leaf $\ell(x,\theta)$ of tree $T(\theta)$ is defined by the intervals $I(x,m,\theta) \subseteq [0,1]$ for $m = 1, \ldots, p$,

$$R_{\ell(x,\theta)} = \otimes_{m=1}^{p} I(x,m,\theta).$$

Note that $X_i \notin I(x,m,\theta)$ implies $w_i(x,\theta) = 0$ by (4). To show (8), it thus suffices to show that $\max_m |I(x,m,\theta)| = o_p(1)$ for $n \to \infty$, for all $x \in \mathcal{B}$. The proof is thus complete with Lemma 2. ■

Lemma 2 *Under the conditions of Theorem 1, it holds for all $x \in \mathcal{B}$ that* $\max_m |I(x, m, \theta)| = o_p(1)$ *for $n \to \infty$.*

Proof As any $x \in \mathcal{B}$ is dropped down a tree, several nodes are passed. Denote by $S(x, m, \theta)$ the number of times that these nodes contain a splitpoint on variable m; this is a function of the random parameter θ and of x. The total number of nodes that x passes through is denoted by

$$S(x, \theta) = \sum_{m=1}^{p} S(x, m, \theta).$$

Using the second part of Assumption 3, the maximal number of observations in any leaf, $\max_\ell k_\theta(\ell)$, is bounded (for every tree θ) from below by $n\gamma^{S_{\min}(\theta)}$, where

$$S_{\min}(\theta) = \min_{x \in \mathcal{B}} S(x, \theta).$$

Using the first part of Assumption 2, the maximal number of observations in any leaf, $\max_\ell k_\theta(\ell)$, is on the other hand bounded from above by an $o(n)$-term. Putting together, one can conclude that $n\gamma^{S_{\min}(\theta)} = o(n)$ for $n \to \infty$ and thus $\gamma^{S_{\min}(\theta)} = o(1)$ for $n \to \infty$. Hence there exists a sequence s_n with $s_n \to \infty$ for $n \to \infty$, such that $S_{\min}(\theta) \geq s_n$ for all n.

As the probability of splitting on variable $m \in \{1, \ldots, p\}$ is bounded from below by a positive constant, by the first part of Assumption 3, there exists a sequence g_n with $g_n \to \infty$ for $n \to \infty$ such that, for every $x \in \mathcal{B}$,

$$P\{\min_m S(x, m, \theta) > g_n\} \to 1 \qquad n \to \infty. \tag{9}$$

Using Assumption 3, the proportion of observations whose m-th component is contained in $I(x, m, \theta)$ is bounded from above by

$$n^{-1}\#\{i \in \{1, \ldots, n\} : X_{i,m} \in I(x, m, \theta)\} \;\leq\; (1-\gamma)^{S(x,m,\theta)}.$$

Using (9), it follows that

$$\max_m \; n^{-1}\#\{i \in \{1, \ldots, n\} : X_{i,m} \in I(x, m, \theta)\} \;=\; o_p(1). \tag{10}$$

Let $F_n^{(m)}$ be the empirical distribution of $X_{i,m}$, $i = 1, \ldots, n$,

$$F_n^{(m)}(t) \;=\; n^{-1}\#\{i \in \{1, \ldots, n\} : X_{i,m} \leq t\}.$$

As the predictor variables are assumed to be uniform over $[0, 1]$, it holds by a Kolmogorov-Smirnov type argument that

$$\sup_{t \in [0,1]} |F_n^{(m)}(t) - t| \to_p 0 \quad n \to \infty,$$

and (10) implies thus $\max_m |I(x, m, \theta)| = o_p(1)$ for $n \to \infty$, which completes the proof. ■

Note that the discussion of consistency here has several shortcomings, which need to be addressed in follow-up work. For one, no distinction is made between noise variables and variables that contain signal, as in Breiman (2004). This treatment would require more involved assumptions about the probability that a certain variable is selected for a splitpoint, yet might explain the robustness of quantile regression forests against inclusion of many noise variables, something that has been observed empirically for random forests and, according to some numerical experience, holds as well for quantile regression forests. Second, convergence rates are not discussed. Third, it is noteworthy that Theorem 1 holds regardless of the number k of grown trees. Empirically, however, a single random tree is performing very much worse than a large ensemble of trees. The stabilizing effect of many trees is thus neglected in the current analysis, but would certainly be of relevance when discussing convergence rates.

5. Numerical Examples

Quantile regression forests (QRF) is applied to various popular data sets from the Machine Learning literature and results are compared to four other quantile regression methods: linear quantile regression with interactions (QQR) and without interactions (LQR), and quantile regression trees with with piecewise constant (TRC), piecewise multiple linear (TRM), and piecewise second-degree polynomial form (TRP).

For quantile regression forests (QRF), bagged versions of the training data are used for each of the $k = 1000$ trees. One could use the out-of-bag predictions to determine the optimal number *mtry* of variables to consider for splitpoint selection at each node. However, to demonstrate the stability of QRF with respect to this parameter, the default value is used throughout all simulations (where *mtry* is equal to one-third of all variables). Node-sizes are restricted to have more than 10 observations in each node. It is noteworthy that different values of this latter parameter do not seem to change the results very much; nevertheless, it is pointed out in Lin and Jeon (2002) that growing trees until each node is pure (as originally suggested by Breiman) might lead to overfitting.

Linear quantile regression is very similar to standard linear regression and is extensively covered in Koenker (2005). To make linear quantile regression (LQR) more competitive, interaction terms between variables were added for QQR. Starting from the linear model, interaction terms were added by forward selection until the 5-fold cross-validation error attained a minimum.

Next, tree-based methods are considered. Quantile regression trees with piecewise polynomial form were introduced in Chaudhuri and Loh (2002). Software for quantile regression trees comes in the form of the very useful software package GUIDE, available from `www.stat.wisc.edu/~loh/guide.html`, which makes also piecewise constant and piecewise linear quantile regression trees (TRC) available. The default settings are used for both piecewise linear and piecewise second-degree polynomial approximations.

Data Sets The data sets are taken from the packages *mlbench* and *alr3* of the statistical software package R (R Development Core Team, 2005), and include the well-known *Boston Housing* ($p = 13$ variables, $n = 506$ observations), *Ozone* ($p = 12$, $n = 366$, after having removed all missing value observations) and *Abalone* data set ($p = 8$, limited to $n = 500$ randomly chosen observations) from the UCI machine learning repository. In the package *alr3* (Weisberg, 2005), the data set *BigMac* contains the minutes of labor necessary to purchase a Big Mac in $n = 69$ cities worldwide, along with $p = 9$ other variables like tax rates or primaries teacher net income; these variables are used as predictor variables. Last, the data set *Fuel* lists average gas-mileage for all $n = 51$ American states in the year 2001 (the ratio of total gallons of gasoline sold and the approximate number of miles driven), along with $p = 5$ variables such as gasoline state tax rate and per capita income; again these are used as predictor variables.

Evaluation To measure the quality of the conditional quantile approximations, loss function (3) is used in conjunction with 5-fold cross-validation. The employed loss function measures the weighted absolute deviations between observations and quantiles, instead of the more common squared error loss. The minimum of the loss would be achieved by the true conditional quantile function, as discussed previously. The empirical loss over the test data is computed for all splits of the data sets at quantiles $\alpha \in \{.005, .025, .05, .5, .95, .975, .995\}$. Additionally to the average loss for each method, one might be interested to see whether the difference in performance between quantile regression forests and the other methods is significant or not. To this end bootstrapping is used, comparing each method against quantile regression forests (QRF). The resulting 95% bootstrap confidence intervals for the difference in average loss is shown by vertical bars; if they do not cross the horizontal line (which marks the average loss of QRF), the difference in average loss is statistically significant. Results are shown in Figure 1.

There is not a single data set on which any competing method performs significantly better than quantile regression forests (QRF). However, QRF is quite often significantly better than competing methods. If the difference is not significant, QRF has most often the smaller average loss.

Aggregated trees thus seem to outperform single trees. This is despite the fact that quantile regression trees have to be grown separately for each quantile α, whereas the same set of trees can be used for all quantiles with QRF. The performance of QRF could be even marginally better when growing different set of trees for each value of α. However, this performance enhancement would come at an additional computational price. Moreover, monotonicity of the quantile estimates would not be guaranteed any longer. As it is, the α-quantile of QRF is always at least as large as the β-quantile if $\alpha \geq \beta$. This monotonicity constraint is not always fulfilled for the other considered methods.

Linear quantile regression with interaction terms works surprisingly well in comparison, especially for moderate quantiles (that is α is not too close to either 0 or 1). However, for more extremal quantiles, quantile regression forests delivers most often a better better approximation. This is even more pronounced if additional noise variables are added. To this end, every original predictor variables is permuted randomly and added to the list of predictor variables. The results are shown in Figure 2. The performance of quantile regression forests seems to be robust with respect to inclusion of noise variables.

Prediction Intervals A possible application of quantile regression forests is the construction of prediction intervals, as discussed previously. For each new data point X, a prediction interval of the form (2) gives a range that will cover the new observation of the response variable Y with high probability.

In Figure 3, some graphical results are shown for the Boston Housing data. Figure 4 shows comparable plots for the remaining data sets. There are two main observations: Firstly, as expected, about 95% of all observations are inside their 95% prediction intervals.

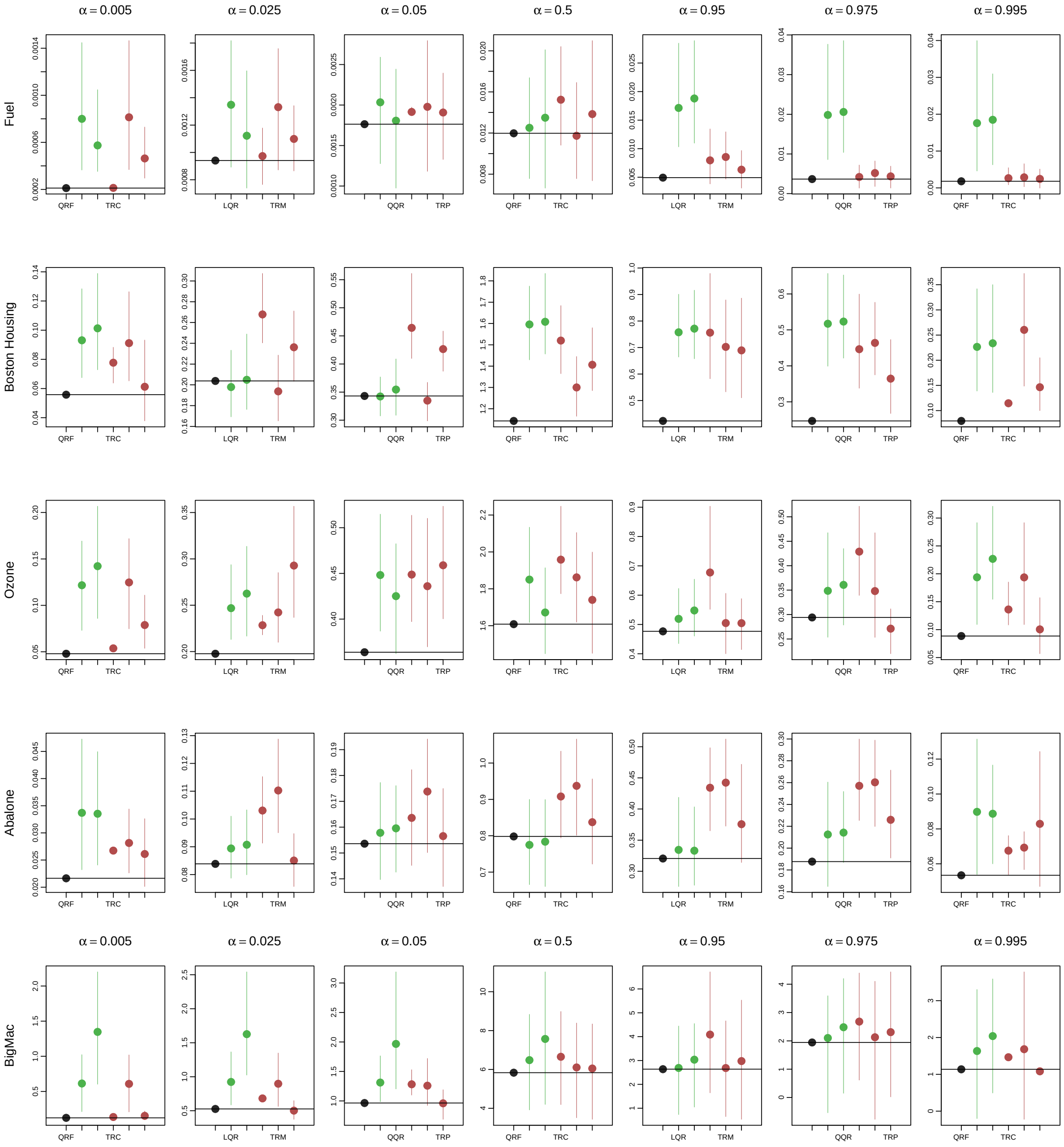

Figure 1: Average loss for various data sets (from top to bottom) and quantiles (from left to right). The average loss of quantile regression forests is shown in each plot as the leftmost dot and is indicated as well by a horizontal line for better comparison. The average losses for competing methods are shown for the linear methods in the middle and the three tree-based methods on the right of each plot. The vertical bars indicate the bootstrap confidence intervals for the difference in average loss for each method against quantile regression forests. Note that no parameters have been fine-tuned for quantile regression forests (the default settings are used throughout).

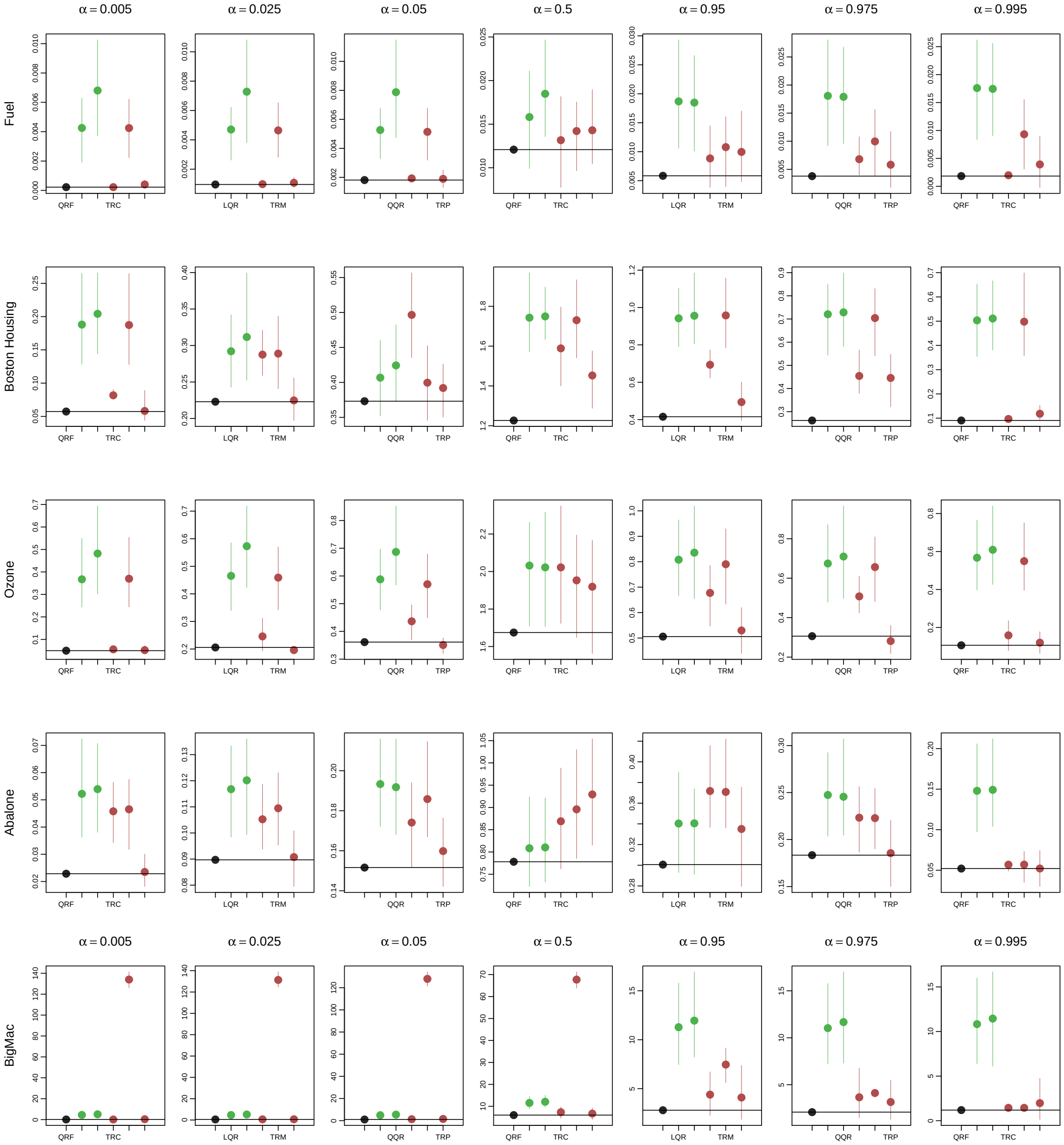

Figure 2: Same plots as in Figure 1. However, to test the performance of the methods under additional noise, each predictor variable is permuted randomly and added to the list of predictor variables.

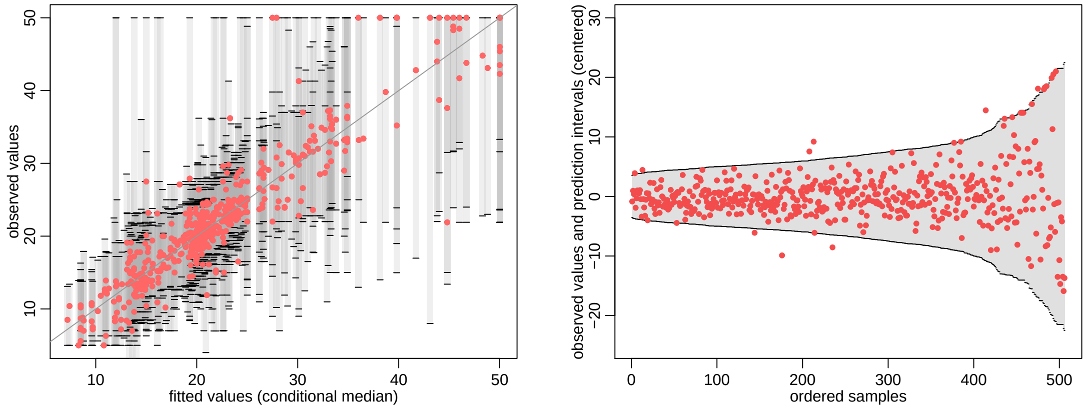

Figure 3: For each data point $i = 1, \ldots, n$ in the Boston Housing data set (with $n = 506$), conditional quantiles are estimated with QRF on a test set which does not include the i-th observation (5-fold cross-validation). Left panel: the observed values are plotted against the predicted median values. Prediction intervals are shown for each $i = 1, \ldots, n$ as transparent grey bars, with vertical black lines at the bottom and top. It can be seen that prediction intervals vary in length, some being much shorter than others. Right panel: For better visualisation, observations $i = 1, \ldots, n$ are ordered according to the length of the corresponding prediction intervals. Moreover, the mean of the upper and lower end of the prediction interval is subtracted from all observations and prediction intervals. All but 10 observations actually lie in their respective 95% prediction intervals.

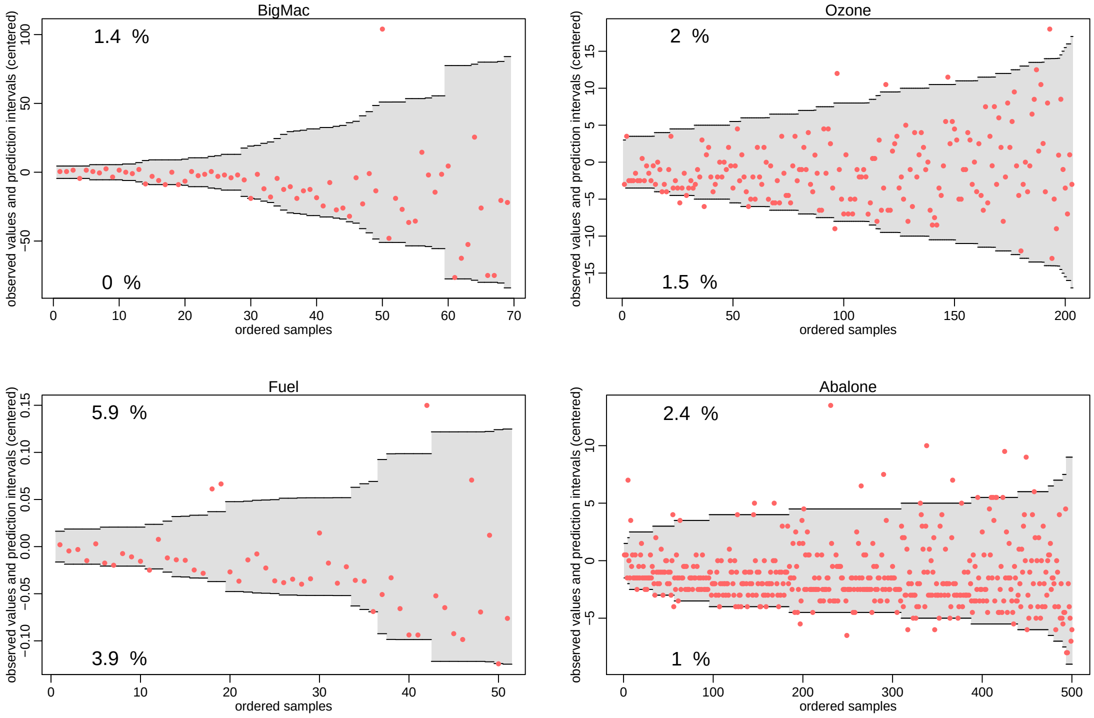

Figure 4: Same plots as in the right panel of Figure 3 for the remaining data sets. Additionally, the percentage of observations that lie above the upper end of their respective prediction intervals (below the lower end) are indicated in the upper left corner (lower left corner). As 95% prediction intervals are shown, on average 2.5% of all observations should be above (and below) their prediction intervals. For the Big Mac, Fuel, and Ozone data sets, it is particularly apparent that the lengths of the prediction intervals vary strongly (some values can thus be predicted more accurately than others).

Secondly, the lengths of prediction intervals vary greatly. Some observations can thus be predicted much more accurately than others.

With quantile regression forests, it is possible to give a range in which each observation is going to be (with high probability). The wider this range for a new instance, the less accurate any prediction is going to be. Vice versa, one knows that a prediction is reliable if the prediction interval is very short.

6. Conclusions

Quantile regression forests infer the full conditional distribution of a response variable. This information can be used to build prediction intervals and detect outliers in the data.

Prediction intervals cover new observations with high probability. The length of the prediction intervals reflect thus the variation of new observations around their predicted values. The accuracy with which new observations can be predicted varies typically quite strongly for instances in the same data set. Quantile regression forests can quantify this accuracy. The estimated conditional distribution is thus a useful addition to the commonly inferred conditional mean of a response variable.

It was shown that quantile regression forests are, under some reasonable assumptions, consistent for conditional quantile estimation. The performance of the algorithm is very competitive in comparison with linear and tree-based methods, as shown for some common Machine Learning benchmark problems.

References

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

V. Barnett and T. Lewis. *Outliers in Statistical Data.* John Wiley and Sons, 1994.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman. Consistency for a simple model of random forests. Technical Report 670, Department of Statistics, University of California, Berkeley, 2004.

L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–598, 1985.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, 1984.

P. Chaudhuri and W. Loh. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, 8:561–576, 2002.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:337–407, 2000.

X. He, P. Ng, and S. Portnoy. Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society* B, 3:537–550, 1998.

V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85 – 126, 2004.

P. Huber. Robust regression: asymptotics, conjectures, and monte carlo. *Annals of Statistics*, 1:799–821, 1973.

R. J. Hyndman and Y. Fan. Sample quantiles in statistical packages. *American Statistician*, 50:361–365, 1996.

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

R. Koenker, P. Ng, and S. Portnoy. Quantile smoothing splines. *Biometrika*, 81:673–680, 1994.

Q. V. Le, T. Sears, and A. Smola. Nonparametric quantile regression. Technical report, NICTA, 2005.

Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2:18–22, 2002.

Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. Technical Report 1055, Department of Statistics, University of Wisconsin, 2002.

M. Markou and S. Singh. Novelty detection: A review. *Signal Processing*, 83:2481–2497, 2003.

S Portnoy and R. Koenker. The gaussian hare and the laplacian tortoise: Computability of squared-error versus absolute-error estimates. *Statistical Science*, 12:279–300, 1997.

R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2005. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:1651–1686, 1998.

I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005.

S. Weisberg. *Applied Linear Regression.* Wiley, 2005.