Statistica Sinica **19** (2009), 1171-1191

COMBINING REGRESSION QUANTILE ESTIMATORS

Kejia Shan and Yuhong Yang

Amylin Pharmaceuticals and University of Minnesota

Abstract: Model selection for quantile regression is a challenging problem. In addition to the well-known general difficulty of model selection uncertainty, when quantiles at multiple probability levels are of interest, typically a single candidate does not serve all of them simultaneously. In this paper, we propose methods to combine quantile estimators. Oracle inequalities show that, at each given probability level, the combined estimators automatically perform nearly as well as the best candidate. Simulation and examples show that the proposed model combination approach often leads to a substantial gain in accuracy under global measures of performance.

Key words and phrases: Adaptive quantile regression, aggregation of estimators, model combination.

1. Introduction

Conditional quantile estimation has been used for a long time in various contexts, including agriculture, economics, and finance. Numerous methods have been proposed under different settings, including the classical linear regression, nonlinear regression, time series, and longitudinal experiment (see He, Ng and Portnoy (1998), Yu, Lu, and Stander (2003), and Koenker (2005) for some recent developments and references). In what follows, we first give a brief review of the general problem of conditional quantile estimation and model selection, and then set up our specific problem.

1.1. A background on conditional quantile estimation (CQE)

In regression, besides the conditional mean, we are often interested in other summary measures of the conditional distribution of Y given the input X. Quantile regression is used to obtain an estimate of the conditional quantile function at a given probability level τ ($\tau \in (0,1)$). When a range of τ values is considered, the quantile profile provides information much beyond the conditional mean. Conditional quantile estimation may also be used to produce confidence bands for the distribution of Y given X (see Zhou and Portnoy (1996) and Koenker (2005) for some applications). Quantile estimation also gets attention due to its robustness property, compared to the conditional mean, in case of strong skewness in the true conditional distribution (see, e.g., Yu, Lu, and Stander (2003) and Geraci and Bottai (2007)).

Koenker and Bassett (1978) introduced regression quantile estimation by minimizing an asymmetric loss function $L_\tau(\xi) = \{\tau - I_{\{\xi<0\}}\}\xi$ for $0 < \tau < 1$, known as the check or the pinball loss. It is not hard to verify that the minimizer $c(x)$ of $EL_\tau(Y - c(X)|X = x)$ is the lower-τ conditional quantile of Y given $X = x$. They considered $c(x)$ of the form $x'\beta$ with β estimated by minimizing $\sum_i L_\tau(y_i - x_i'\beta)$. This method is commonly known as **l**inear **q**uantile **r**egression (LQR). A slightly more general loss, called lin-lin loss, was considered in Granger (1969).

To reduce the impact of parametric assumptions, nonparametric and semiparametric methods have also been developed for quantile regression. For example, one might assume that the quantile function is of the semi-parametric form $q_\tau(X,T) = X'\beta + g(T)$, where both X and T are vectors of explanatory variables, β denotes a vector of unknown regression coefficients, and g represents an unparameterized smooth function to be estimated. Analogous to semiparametric mean regression, we can estimate β and g by minimizing $\sum_{i=1}^n L_\tau(y_i - x_i'\beta - g(t)) + \alpha \int g''^2 dt$, where α is a smoothing parameter to control the amount of penalty on the roughness of g. Interested readers are referred to Yu, Lu, and Stander (2003) and Koenker (2005), and the references therein for more details.

More recently, Meinshausen (2006) proposed a nonparametric method called **q**uantile **r**egression **f**orests (QRF), inspired by the random forests of Breiman (2001). As in the random forests algorithm, for each tree, one selects a random subset of all predictors to split nodes, and a large number of (random) trees are obtained in this fashion. The conditional quantile of Y given $X = x$ is then approximated by the average prediction from the collection of random trees. This method was shown to be consistent and numerical results demonstrated its good performance in problems with high-dimensional predictors, particularly at extreme values of τ (τ near zero or one).

Regression quantile is also important in areas of application other than the conventional i.i.d. setting. In longitudinal studies, Geraci and Bottai (2007) used the loss function $L_{0.5}$ to construct the Normal-Laplace joint likelihood in a mixed effect model; an interesting quantile autoregression theory is given in Koenker and Xiao (2006); Wei and He (2006) proposed a useful semi-parametric quantile regression method for constructing conditional growth charts based on longitudinal observations.

Besides the check loss, other asymmetric loss functions have also been investigated (see, e.g., Hall, Wolff and Yao (1999)), although they are used less often in the statistical literature.

1.2. Model selection and combination in CQE

The issue of model selection has been studied in connection with quantile regression. Ronchetti (1985) introduced a robust version of AIC, called AICR, which takes the form of the observed check loss plus a multiple of model size (see also Cade, Noon and Flather (2005)). Machado (1993) proposed a generalized Schwarz Information Criterion, similar to BIC except that the squared error loss is replaced by a more robust loss function. Some other model selection criteria can be found in Burman and Nolan (1995) and Ronchetti, Field and Blanchard (1997).

In an effort to combine different methods, if $\hat{q}_\tau^A(x)$ and $\hat{q}_\tau^B(x)$ are two estimates of the conditional lower-τ quantile of Y given $X = x$, Granger (1989) proposed the use of weights from $\min_{\alpha,\beta_A,\beta_B} \sum_i L_\tau(y_i - \alpha - \beta_A \hat{q}_\tau^A(x_i) - \beta_B \hat{q}_\tau^B(x_i))$. Taylor and Bunn (1998) extended this linear combination methodology by considering a number of constraints on the coefficients α, β_A, β_B, such as zero intercept, convex coefficients on the predictors, and so on. To our knowledge, theoretical results on combining quantile regression estimators have not appeared in the literature.

When the quantile profile is of interest, it is particularly important to consider model combination methods. A main reason is that the different quantile regression estimators typically have distinct relative performances that depend on the value of τ (as seen in our numerical results). Integrating the advantages of the candidates for potential global improvement is a worthy task.

A recent focus on combining or aggregating models (procedures) is the construction of methods that adaptively share the strengths of a list of arbitrary estimators (Nemirovski (2000), Yang (2001, 2004b), Catoni (2004) and Tsybakov (2003)), which allows the integration of powers of different methodologies. See Leung and Barron (2006), Birgé (2006), Bunea and Nobel (2005), Bunea, Tsybakov and Wegkamp (2006), Audibert (2006) and Lecué (2006) for some recent results in the area. We follow this spirit and present both theoretical and numerical results on combining quantile estimators.

1.3. Problem of interest

We observe (Y_i, X_i), $i = 1, \ldots, n$, where $X_i = (X_{i1}, \ldots, X_{ip})$ is a p-dimensional predictor. Assume that the true underlying relationship between Y and X is:

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i, \quad i = 1, \ldots, n,$$

where ϵ_i are i.i.d. from a distribution with mean zero and variance one, and are independent of the predictors. A time series setting which does not require that Y_i, $i = 1, \ldots, n$ are independent will be considered as well.

Based on this model, the conditional quantile of Y given $X = x$ has the form

$$q_\tau(x) = m(x) + \sigma(x)F^{-1}(\tau), \tag{1.1}$$

where F is the cumulative distribution function of the error. This provides one method for estimating $q_\tau(x)$, namely, by first obtaining $\hat{m}(x)$, $\hat{\sigma}(x)$ and $\hat{F}^{-1}(\tau)$ (if F needs to be estimated).

Based on (1.1), it can be observed that if $m(\cdot)$ is a linear function of x and $\sigma(\cdot)$ is constant, linear quantile regression (LQR) is expected to perform well. However, if either the mean function is nonlinear or the scale function is nonconstant in the predictors, bias is involved and may lead to poor performance of LQR. Also, in applications, the performance of LQR on extreme quantiles is usually impaired by insufficient extreme observations.

Now suppose we have a pool of M candidate estimators of the conditional quantile function $q_\tau(x)$, denoted by $\{\hat{q}_{\tau,j}(x)\}_{j=1}^{M}$. Our goal is to combine these estimators for an optimal performance. Specifically, at each given τ, we hope that the combined estimator performs as well as the best candidate. Since the best candidate often depends on τ, our combining approach can improve over all of the candidate procedures in terms of global performance measures over τ, as will be seen in our simulations and examples.

In the context of conditional mean regression, Yang (2001) proposed the adaptive regression by mixing (ARM) method, in which a set of weights is adaptively calculated from the data under a specified likelihood function, such as Gaussian. Alternatively, risk bounds that relate the performance of the combined estimator to that of the best candidate (typically unknown, of course) under certain quadratic-type of loss functions are given in Catoni (2004) and Yang (2004a), without specifying the error distribution. This latter approach is useful when no obvious choice of error density is available and/or when variance estimation is difficult.

In the current context, instead of a quadratic loss, the check loss function is naturally oriented toward quantile estimation and is used in our weight construction. However, the distinct natures of the absolute-type and quadratic-type of losses present real impediments to the derivation of an oracle inequality for our quantile regression problem. Risk bounds in terms of the check loss function, under both i.i.d. and a time series settings without any assumption on the form of the error density nor requiring boundedness of the response variable, are obtained, which indeed show that at each fixed τ our combined estimator performs almost as well as the best candidate. A potential application of our method is to conditional growth chart construction (Wei and He (2006)), where different semi-parameter models can be explored.

The rest of the paper is organized as follows. In Section 2, our model combining methods for regression with i.i.d. observations are presented and oracle inequalities that show their optimal performance are given. In Section 3, model combination is considered for a time series framework. Simulation and examples that demonstrate advantages of our methods are presented in Sections 4 and 5, respectively. Concluding remarks are given in Section 6. Proofs of the theoretical results are in an appendix.

2. Adaptive Quantile Regression by Mixing (AQRM)

In this section, we consider the framework in Section 1.3 with i.i.d. observations, and take two weighting approaches, one directly based on the cumulative check loss and the other on a mixture of the check and squared losses.

2.1. Weighting based on check loss

The AQRM algorithm for conditional quantile estimation is as follows. Fix a probability level $0 < \tau < 1/$ Let $1 \leq n_0 \leq n-1$ be an integer (typically n_0 is of the same order as or slightly larger order than $n - n_0$).

1. Randomly partition the data into two parts: $Z^{(1)} = \{y_l, x_l\}_{l=1}^{n_0}$ for training, and $Z^{(2)} = \{y_l, x_l\}_{l=n_0+1}^{n}$ for evaluation.
2. Based on $Z^{(1)}$, obtain candidate estimates of the conditional quantile function $q_\tau(x)$ as $\hat{q}_{\tau,j,n_0}(x) = \hat{q}_{\tau,j,n_0}(x; Z^{(1)})$. Use $\hat{q}_{\tau,j,n_0}$ to obtain the predicted quantiles from the j^{th} candidate procedure for $Z^{(2)}$, for each $j = 1, \ldots, M$.
3. Compute the candidate weights as

$$W_j = \frac{\prod_{l=n_0+1}^{n} \exp\{-\lambda L_\tau(y_l - \hat{q}_{\tau,j,n_0}(x_l))\}}{\sum_{k=1}^{M} \prod_{l=n_0+1}^{n} \exp\{-\lambda L_\tau(y_l - \hat{q}_{\tau,k,n_0}(x_l))\}},$$

   where $\lambda > 0$ is a tuning parameter.
4. Repeat Steps 1−3 a total of $B - 1$ more times and average the weights W_j over B random permutations. Denote them by $\tilde{W}_j$. The final estimator of the conditional quantile function of Y at $X = x$ is $\hat{q}_{\tau,\cdot,n}(x) = \sum_{j=1}^{M} \tilde{W}_j \hat{q}_{\tau,j,n}(x)$.

Remark. The tuning parameter λ controls how much the weights rely on the check loss performance. In the extreme case when $\lambda \downarrow 0$, simple averaging results; when $\lambda \to \infty$, the candidate with the best historic check loss is selected.

In certain problems such as online estimation/prediction, a sequential updating mechanism is also of interest. Here, we obtain $\hat{q}_{\tau,j,n_0}$ from $\{(y_l, x_l)\}_{l=1}^{n_0}$ (the initial set of observations) and update the weights sequentially once an additional observation is made. In such a setting, we define sequential weight $W_{j,i}$

as

$$W_{j,i} = \frac{\prod_{l=n_0+1}^{i-1} \exp\{-\lambda L_\tau(y_l - \hat{q}_{\tau,j,l}(x_l))\}}{\sum_{k=1}^{M} \prod_{l=n_0+1}^{i-1} \exp\{-\lambda L_\tau(y_l - \hat{q}_{\tau,k,l}(x_l))\}},$$

and the combined estimate of $q_\tau(x)$ at time i is $\hat{q}_{\tau,\cdot,i}(x) = \sum_{j=1}^{M} W_{j,i}\hat{q}_{\tau,j,i}(x)$. Also of interest is an overall convex combination

$$\hat{q}_{\tau,\cdot,\cdot}(x) = \frac{1}{n-n_0} \sum_{j=1}^{M} \sum_{i=n_0+1}^{n} W_{j,i}\hat{q}_{\tau,j,i}(x)$$

that estimates $q_\tau(x)$ in a way that utilizes the online estimates at different sample sizes; it has a nice risk property as will be seen shortly. Note that in numerical implementation for batch learning, since a sequential updating algorithm can be much more time-consuming when the sample size is not small, we follow the earlier algorithm, and candidate quantile estimators are not updated in the weight construction.

2.2. Oracle inequalities on performance

Condition 0. The observed vectors (Y_i, X_i), $i \geq 1$ are i.i.d..

Condition 1. The quantile estimators satisfy $\sup_{j\geq 1, i\geq 1} |\hat{q}_{\tau,j,i}(x_i) - q_\tau(x_i)| \leq A_\tau$, for some positive constant A_τ with probability one. In what follows, we omit the subscript τ to simplify notation.

Condition 2. There exist a positive constant t_0 and a monotone function $0 < H(t) < \infty$ on $[-t_0, t_0]$ such that for all $n \geq 1$ and $-t_0 \leq t \leq t_0$,

$$E(|\epsilon_n|^2 + 1)\exp(t|\epsilon_n|) \leq H(t),$$

where ϵ_n is the unobservable true error for the n^{th} observation.

Condition 3. There exist positive constants C_1 (that depends on τ) and C_2 such that $|m(X) - q_\tau(X)| \leq C_1$ and $|\sigma^2(X)| \leq C_2$ with probability one.

Condition 1 requires that no candidate estimators are too far away from the true conditional quantile. This is a mild technical condition, weaker than assuming Y is bounded, and is typically assumed in the literature on combining estimators. Condition 2 is satisfied by error distributions with well-defined moment generating functions, such as the normal, shifted gamma and double exponential distributions. These error distributions are considered in our numerical study. Condition 3 requires some regularity of the underlying conditional distribution of Y given the predictors, but neither constant is required to be known for application.

Let $B(\lambda) = e^{2\lambda \max(\tau,1-\tau)(A+1)}(1 + (A+1)^2)H(2\lambda \max(\tau, 1-\tau))$ and, in Theorem 1 below, let $a_\lambda = 2\lambda(\max(\tau, 1-\tau))^2 B(t_0)$.

Theorem 1. *Under Conditions 0-3, when the tuning parameter $\lambda \le \lambda_0 = (t_0)/[2\max(\tau, 1-\tau)]$, we have*

$$\frac{1}{n-n_0}\sum_{i=n_0+1}^{n} EL_\tau(Y_i - \hat{q}_{\tau,\cdot,i}(X_i))$$
$$\le \inf_j \left\{ \frac{1}{n-n_0}\sum_{n_0+1}^{n} EL_\tau(Y_i - \hat{q}_{\tau,j,i}(X_i)) + \frac{\log(M)}{\lambda(n-n_0)} + \frac{a_\lambda(C_2 + C_1^2)}{2} \right\}.$$

In particular, when $\lambda = \{[\log(M)]/[(\max(\tau, 1-\tau))^2 B(t_0)(C_2 + C_1^2)(n - n_0)]\}^{1/2}$, we have

$$\frac{1}{n-n_0}\sum_{i=n_0+1}^{n} EL_\tau(Y_i - \hat{q}_{\tau,\cdot,i}(X_i))$$
$$\le \inf_j \left\{ \frac{1}{n-n_0}\sum_{i=n_0+1}^{n} EL_\tau(Y_i - \hat{q}_{\tau,j,i}(X_i)) + \tilde{C}\sqrt{\frac{\log(M)}{n-n_0}} \right\}, \qquad (2.1)$$

and

$$EL_\tau(Y - \hat{q}_{\tau,\cdot,\cdot}(X)) \le \inf_j \left\{ \frac{1}{n-n_0}\sum_{i=n_0+1}^{n} EL_\tau(Y_i - \hat{q}_{\tau,j,i}(X_i)) + \tilde{C}\sqrt{\frac{\log(M)}{n-n_0}} \right\},$$

where $\tilde{C}$ is a constant that depends on τ, A, C_1, C_2.

Remarks.

1. For the third display in the theorem, the risk of the combined estimator at sample size n is upper bounded in terms of the best averaged risk at different sample sizes, plus a penalty. Ideally, one would want to replace the averaged risk by the risk of the candidate at the full sample size n, but this is not obtained in this work.
2. Note that the unboundedness of the response variable makes the derivation of oracle inequalities substantially different from the earlier work on combining predictions in the machine learning literature, which typically requires that Y have a bounded support (or the loss is bounded). See Bunea and Nobel (2005) for a different way to address the issue of unbounded response under squared error loss.

The inequalities above say that the risks of the combined prediction are automatically close to the risks of the best individual, with the difference being of order $(n - n_0)^{-1/2}$ when λ is chosen properly. Note that for the L_1 type of risk for regression estimation, the rate of convergence typically is $n^{-1/2}$ for parametric cases, and is slower than $n^{-1/2}$ for nonparametric cases (e.g., Yang and Barron (1999)). Therefore, with a choice of n_0 and $n - n_0$ of the same order, the risk bounds show that the combined quantile predictions adaptively converge at the best rate offered by the candidate procedures for both parametric and nonparametric situations. Furthermore, for nonparametric quantile regression, since the extra term in the risk bound is asymptotically negligible relative to the risk of estimating $q_\tau(x)$, under some regularity conditions AQRM yields combined predictions that perform asymptotically as well as the best procedure among the candidates.

Although at each given probability level τ, our approach of combining the quantile estimators does not necessarily lead to performance improvement over the best individual candidate estimator, the results are useful for three reasons. First, for various situations (e.g., one of the candidate procedures is based on the true model), the best individual procedure simply cannot be improved upon and thus the combined estimator can only perform optimally or near optimally. Second, since the best procedure is unknown, and as is well-known, one often pays a high price in trying to find it (see, e.g., Yuan and Yang (2005) for references and simulation results on reducing model selection uncertainty by model combination), it is important to show that the combining approach, as an alternative, indeed leads to optimal performance. Third, because conditional quantile functions over a range of probability levels are often of interest while the candidate quantile estimators typically have different performance ranks, the combined estimators have the potential to beat each of the candidates in terms of global performance, as will be seen later.

2.3. Weighting using a mixture of check and squared losses

Define a surrogate loss function $L_{\tau,a}(\xi) = L_\tau(\xi) + a\xi^2$ for a given $a > 0$ (see Figure 1) and use it in the construction of weights of the candidate quantile regression procedures. The new weight is

$$W_j^a = \frac{\prod_{l=n_0+1}^{n} \exp\{-\lambda L_{\tau,a}(y_l - \hat{q}_{\tau,j,n_0}(x_l))\}}{\sum_{k=1}^{M} \prod_{l=n_0+1}^{n} \exp\{-\lambda L_{\tau,a}(y_l - \hat{q}_{\tau,k,n_0}(x_l))\}}.$$

We can then derive a similar risk upper bound for the corresponding combined estimator $\hat{q}^a_{\tau,\cdot,n_0}$.

Theorem 2. *Under the same assumptions in Theorem* 1*, when λ is chosen as for* (2.1) *and $a = 2\lambda(\max(\tau, 1-\tau))^2 B(t_0)$, we have*

$$EL_\tau(Y - \hat{q}^a_{\tau,\cdot,\cdot}(X)) \le \inf_j \left\{ \frac{1}{n-n_0} \sum_{i=n_0+1}^{n} EL_\tau(Y_i - \hat{q}^a_{\tau,j,i}(X_i)) + C' \sqrt{\frac{\log(M)}{n-n_0}} \right\},$$

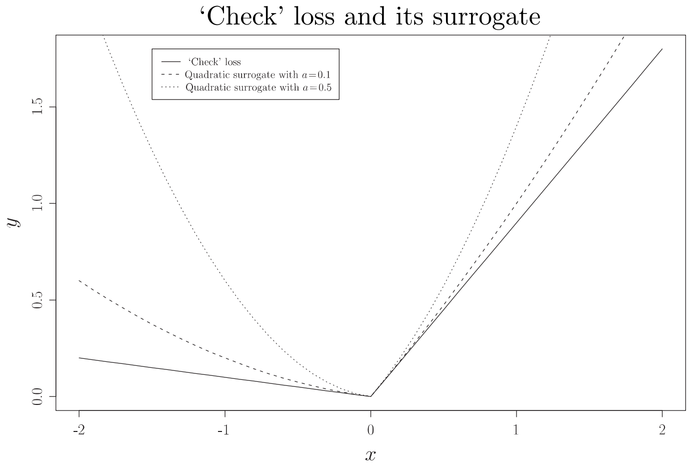


Figure 1. Check loss function with $\tau = 0.9$ and its surrogates.

where C' is a constant that depends on τ, A, C_1, C_2.

3. Combining Quantile Estimators for Time Series

For time series data, we typically have autocorrelation between observations. Consider the model

$$Y_t = m_t(X_t) + \sigma_t(X_t)\epsilon_t,$$

where X_t is the explanatory variable (which may include the past values of the response variable) at time t. We assume that the errors ϵ_t are i.i.d. from a distribution with mean zero and variance one, and that ϵ_t is independent of $\{(Y_s, X_s) : s < t\}$ and X_t. Our goal is to derive a combined (conditional) quantile estimator $\hat{q}_{\tau,\cdot,t}(x_t) = \sum_{j=1}^{M} W_{j,t}\hat{q}_{\tau,j,t}(x_t)$.

We follow an online setting which means that data come in sequentially and the candidate estimators are updated sequentially with each incoming observation. Let T be the length of the whole series. Here is the combining algorithm AQRM for the time series setting.

1. Start with T_0 observations and let $t_1 = T_0$.
2. Denote the first t_1 observations in the series by $Z^{(1)} = (y_t, x_t)_{t=1}^{t_1}$.
3. Based on $Z^{(1)}$, construct the candidate estimates of the conditional quantile function $q_\tau(x)$ as $\hat{q}_{\tau,j,t_1}(x) = \hat{q}_{\tau,j,t_1}(x; Z^{(1)})$.
4. For each j, update the candidate weight sequentially as

$$W_{j,t_1+1} = \frac{W_{j,t_1}\exp\{-\lambda L_\tau(y_{t_1} - \hat{q}_{\tau,j,t_1}(x_{t_1}))\}}{\sum_{k=1}^{M} W_{k,t_1}\exp\{-\lambda L_\tau(y_{t_1} - \hat{q}_{\tau,k,t_1}(x_{t_1}))\}},$$

   where $W_{j,T_0+1} = 1/M$.

5. Increase t_1 by 1 and repeat steps 2–4, until $t_1 = T$.

Since in the time series setting, the conditional quantiles, conditional means and conditional variances of Y_t usually depend on both the predictor and time, Conditions 1-3 need to be modified accordingly.

Theorem 3. *Under Conditions* 1–3 *on the conditional quantiles, conditional means and conditional variances, when the tuning parameter λ is chosen as for* (2.1),

$$\begin{aligned}
&\sum_{t=T_0+1}^{T} EL_\tau(Y_t - \hat{q}_{\tau,\cdot,t}(X_t)) \\
&\quad \le \inf_j \left\{ \sum_{t=T_0+1}^{T} EL_\tau(Y_i - \hat{q}_{\tau,j,t}(X_t)) + \tilde{C}\sqrt{\log(M)} \times \sqrt{T-T_0} \right\},
\end{aligned}$$

where $\tilde{C}$ is a constant that depends on τ, A, C_1, C_2.

4. Simulation Results

In this section, four cases are considered for investigating the performance of AQRM. Together with the examples in the next section, we intend to gain some insight on the differences of behaviors of the methods involved that may be more helpful than giving one or two favorable examples.

4.1. Candidate procedures and performance measures

We consider LQR (Koenker and Bassett (1978)) and QRF (Meinshausen (2006)), using *R* packages *quantreg* and *quantregForest.*

In the literature, performance of quantile regression is usually measured by the coverage probability at some fixed τ value(s), such as the 90% and 95% levels (Koenker and Bassett (1978) and Taylor and Bunn (1998). For a given quantile estimator at a given τ, its empirical coverage probability is defined as the fraction of observations which fall on or below the estimated quantile function in a new (unused) evaluation set.

Here, we focus on the overall performance of a quantile regression procedure over the full range of τ in $(0, 1)$. One reason is that quantiles at multiple levels are often of interest at the same time (e.g., for growth charts) and global measures over a range of τ are naturally relevant. Another related motivation is the fact that different regression quantile estimators often have distinct relative performances according to the value of τ, and therefore the consideration of a range of τ values yields an overall comparison of different methods. We introduce two overall performance measures below.

Let g denote a weighting function on $\tau \in (0,1)$ such that $g \geq 0$ and $\int_0^1 g(\tau)d\tau = 1$, used to differentiate the importance of τ values in different regions. We choose two different g functions here, one being the uniform weight and the other being the Beta(0.8, 0.8) density, which emphasizes extreme $\tau's$.

In simulations, considering the integrated absolute difference between the true $q_\tau(\cdot)$ and an estimator $\hat{q}_\tau(\cdot)$, under a given weight function, we define Weighted Integrated Absolute Error (WIAE) as the expectation of $\iint |\hat{q}_\tau(x) - q_\tau(x)|g(\tau)d\tau P(dx)$. For data, since we do not know the true conditional quantile function, obviously we cannot compute WIAE. Instead, we consider the discrepancy between the nominal level τ and the empirical coverage probability $\hat{\tau}$, and define Weighted Integrated Coverage Error (WICE) as $\int_0^1 |\hat{\tau}-\tau|g(\tau)d\tau$. In implementing this, we use random data splitting, which reserves part of the given data as an (artificial) evaluation set. This random partition of data is repeated 100 times and the average performance measure over these repetitions is reported.

To approximate the integrals in the definitions of WIAE and WICE, we selected a number of discrete τ values, $\tau \in \{0.01, 0.05 \times k, 0.99\}_{k=1}^{19}$. We also calculated, for each fixed τ, the simulation standard errors of both the candidate methods and AQRM.

In our investigation, we also assessed the role of λ on the performance of AQRM (automatic selection of λ will not be addressed). We define the optimal λ as the one that yields the smallest WICE (or WIAE) among all λ considered, and define the risk ratio of AQRM over the best individual candidate as

$$RR = \frac{\text{WICE (or WIAE) of AQRM under the optimal } \lambda}{\text{WICE (or WIAE) of the best individual candidate}}.$$

The simulation results in this section were based on 100 runs in each case. The sample size was 200, with equal training-testing data splitting randomly done 50 times. To compute the absolute error or coverage error loss defined above, an independent evaluation set of size 1,900 was used.

The tuning parameter λ was taken of the form $\lambda_\tau = \lambda \times \min(\tau, 1-\tau)$, where $\tau \in \{0.01, 0.05 \times k, 0.99\}_{k=1}^{19}$. Empirical evidence suggests that our combined estimator performs better with λ_τ than using a constant value for all τ. In what follows, we omit the subscript τ in λ_τ to simplify notation.

4.2. Simulation models

We considered four cases, the last two with randomly generated coefficients to reduce the reliance of the simulation results on specific choice of parameter values.

Case 1. The first model, an example used in the *R*-package *quantreg*, was

$$Y = Z + \log(X) + 0.1 \times (\log(X))^2 + 0.25 \times \log(X) \times \epsilon_2,$$

where $X \sim \chi_4^2$, $\epsilon_1 \sim N(0,1)$, $Z = X + \epsilon_1$, $\epsilon_2 \sim f$, with $\mu_f = 0$ and $\sigma_f = \sigma$, and X, ϵ_1, ϵ_2 generated independently of each other. Besides $N(0,1)$, the shifted gamma distribution was also considered to allow asymmetric error, and the error standard deviation took 6 values: $\sigma = 0.316, 0.717, 1, 2, 3, 4$. In this and other simulations in Section 4.3, unless stated otherwise, shifted gamma errors were generated from a gamma distribution with shape parameter three and scale parameter $\sigma/3$, then shifted to have zero mean. We observed two predictors X and Z, as well as the response Y. Nine λ values were considered: $\min(\tau, 1-\tau) \times \{0, 0.01, 0.1, 1, 5, 8, 10, 50, 100\}$.

We considered three candidate procedures: LQR with predictors X and Z; QRF with predictors X and Z; linear regression quantile with predictors X, Z and $\sqrt{X}$.

Case 2. Based on Case 1, to allow more complexity, we modified the scale function from $\sigma_f(x,z) \equiv \sigma$ to $\sigma_f(x,z) = \sigma\sqrt{x}$.

Case 3. To avoid "picking the best parameter setting to favor one's own method", we randomly generated coefficients $\beta = (\beta_1, \cdots, \beta_6)$, with $\beta_i \overset{i.i.d.}{\sim}$ Unif$[0.5, 2.5]$ for $i = 1, \ldots, 6$, for each data set. The true model was $Y = \beta' X + \sigma\epsilon$, where $X = (X_1, \ldots, X_6)$ has independent $N(0,1)$ components, and ϵ was either from a standard normal distribution or a shifted gamma with mean zero and variance one. Two hundred sets of coefficients were generated this way for each of which the losses of the competing procedures were calculated.

Case 4. The model was $Y = \beta' X + 2\exp(-0.35X_2 - 1.1X_3) + \sigma\epsilon\sqrt{X_2^2 + 0.8X_4^2}$ and the other aspects were the same as in Case 3.

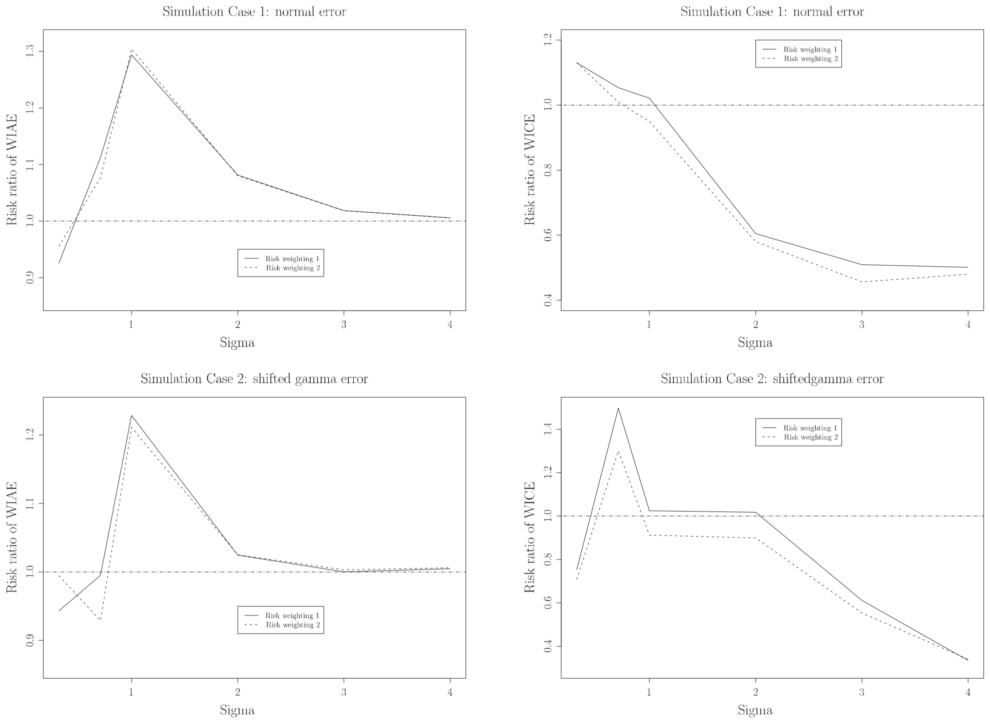
4.3. Results

We give graphical summaries of the overall performance of the combined estimator relative to the best candidate under the two loss functions. Here the best candidate refers to the one which has the smallest mean WICE (or WIAE) under the corresponding weighting function. The plots are in Figures 2 and 3.

We are also interested in performance at fixed values of τ. We compute the L_1 risks in estimating q_τ (along with the simulation standard errors) for several values of τ. The results are given in Table 1 for Case 1 (as an example). Note that the optimal λ given in the tables is τ-dependent.

The results are summarized as follows (although some are not given in this paper due to space limitation).

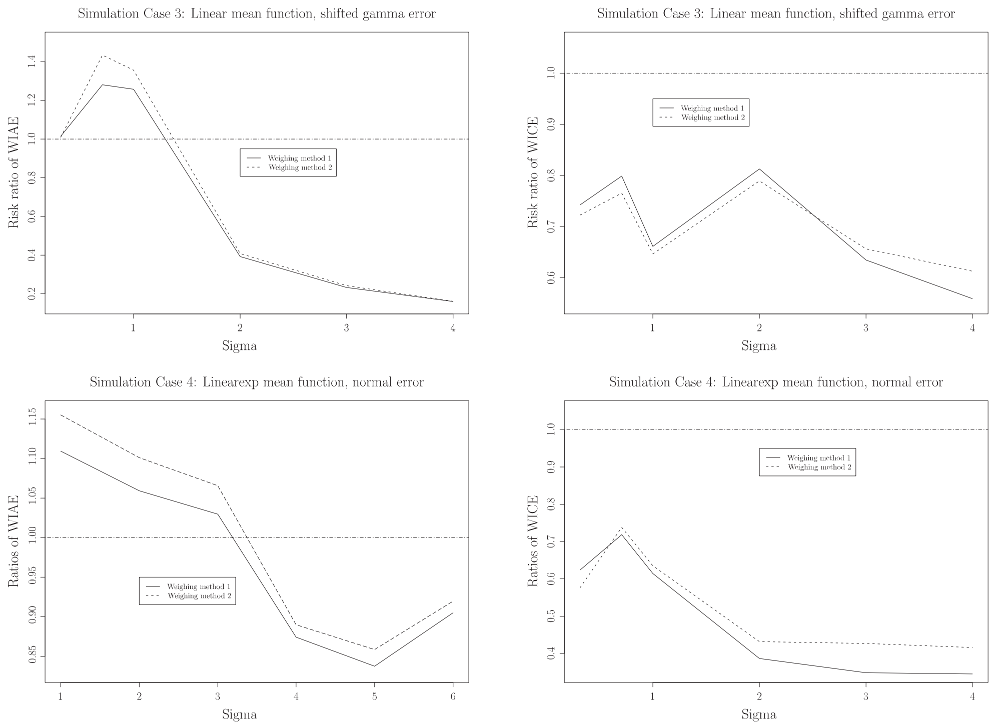
1. For the σ and error distributions considered, when τ was near either zero or one, QRF had observed coverage probability closer to the nominal level τ than LQR. But LQR performed better in the middle range of τ.



(Case 1 and Case 2) The plots give risk ratios of AQRM to that of the best candidate.

Figure 2. Summary plot for Case 1 and Case 2.

2. The L_1 risk of QRF for estimating $q_\tau(x)$ was often the largest, compared to the other two candidates, when σ was small ($\sigma \leq 0.707$), and often the smallest when σ was large ($\sigma \geq 2$). This and the item above indicate that it is unwise to use a single quantile regression estimator for all τ values.
3. AQRM performed well. For the error distributions considered and almost all τ, when $\sigma \geq 2$, AQRM basically tied with or performed better than the best candidate, both in terms of observed coverage probability and in L_1 risks (see Table 1).
4. The two performance measures are quite different. For example, the best candidate estimator under the coverage error was not the same as that under the L_1 error. Also, AQRM did not improve over the candidates under the L_1 error, but did so significantly under the coverage error when σ was not small.
5. The random coefficient cases reveal substantial advantages of AQRM. At the noise levels considered, the coverage errors of AQRM were consistently smaller than those of the candidates. Because the coefficients were randomly generated, the ranking of LQR versus QRF changed as well, in such cases the



(Case 3 and Case 4) The plots give the risk ratios of AQRM to that of the best candidate.

Figure 3. Summary plot for Case 3 and Case 4.

Table 1. L_1 risks at fixed τ for Case 1 under normal errors.

$\sigma = 2$	$\tau = 0.05$	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.95$
Best candidate	1.1594(0.0136)	0.9566(0.0130)	0.6845(0.0081)	1.0022(0.0153)	1.2619(0.0048)
Combined with optimal λ	0.8132(0.0055)	0.8077(0.0042)	0.8149(0.0026)	0.8683(0.0041)	0.9216(0.0088)

combined estimator can be much better than any fixed choice of the candidates.

5. Examples

5.1. Two regression data sets

The data set *Autoprice* is from the UCI machine learning repository. There are $n = 159$ observations with 14 continuous variables and one nominal variable. After inspecting the data, we decided to take logarithmic transformation on the response variable *price* and removed three outliers: #149, #151 and #153. The two candidate quantile regression methods were the LQR and QRF. In LQR, the best submodel selected by AIC via backward elimination was used.

In Table 2, we choose six distinct values of λ to assess its influence. To

Table 2. Weighted Integrated Coverage Errors ($\times 10^{-2}$) for *Autoprice* data.

Method	LQR	QRF	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 3$	$\lambda = 6$	$\lambda = 10$
Uniform	6.20	1.40	1.03	1.05	1.00	1.11	1.41	1.56
Beta(0.8,0.8)	7.09	1.36	1.09	1.11	1.06	1.15	1.44	1.56

Table 3. Mis-coverages at fixed τ for *Autoprice* data.

τ	0.05	0.1	0.25	0.5	0.75	0.9	0.95
Best candidate	0.001(0.003)	0.021(0.003)	0.024(0.006)	0.009(0.006)	0.011(0.005)	0.038(0.003)	0.019(0.002)
Combined with $\lambda = 1$	0.007(0.003)	0.011(0.004)	0.001(0.006)	0.014(0.006)	0.001(0.005)	0.009(0.003)	0.007(0.003)

Table 4. Weighted Integrated Coverage Errors ($\times 10^{-2}$) for *Landrent* data.

Method	LQR	QRF	Plug-in	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 3$	$\lambda = 6$	$\lambda = 10$
Uniform	2.88	2.44	2.11	2.96	2.03	1.83	1.61	1.62	1.64
Beta(0.8,0.8)	3.32	2.29	2.05	2.78	1.96	1.75	1.53	1.54	1.57

Table 5. Mis-coverage at fixed τ for *Landrent* data.

τ	0.05	0.1	0.25	0.5	0.75	0.9	0.95
Best candidate	0.008(0.004)	0.016(0.007)	0.012(0.009)	0.009(0.010)	0.029(0.008)	0.001(0.006)	0.007(0.005)
Combined with $\lambda = 3$	0.008(0.004)	0.011(0.006)	0.002(0.009)	0.014(0.010)	0.013(0.008)	0.010(0.006)	0.005(0.005)

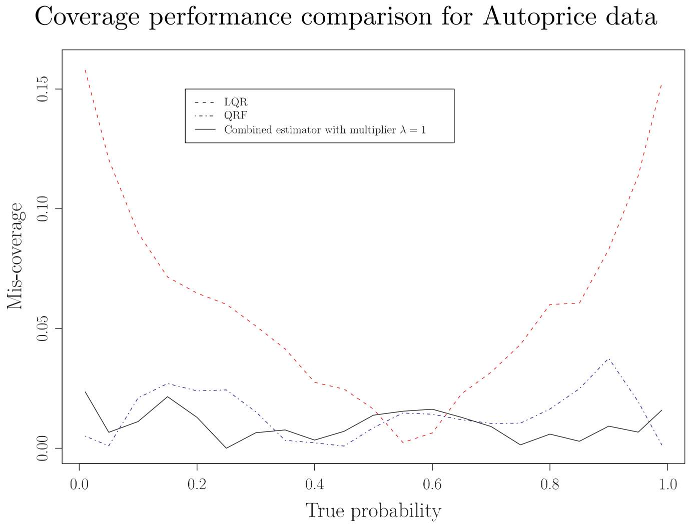
compare the quantile regression procedures, we randomly chose 75% of all data for training (including weight construction for combining the procedures), and the remaining 25% for final performance evaluation. This was repeated another 199 times through random partitioning of the data, and the final coverage performance (WICE) in Table 2 is the average over all 200 repetitions.

We also report mis-coverages along with permutation standard errors for several τ values in Table 3. The numerical results are summarized below.

1. QRF performed better than LQR under both weighting functions, although LQR was slightly more accurate in terms of coverage probability than QRF when τ was near 0.5 (see Figure 4).
2. The combined estimators achieved better accuracy under both weighting functions as long as λ was not too large.
3. AQRM had good performance under almost all τ.

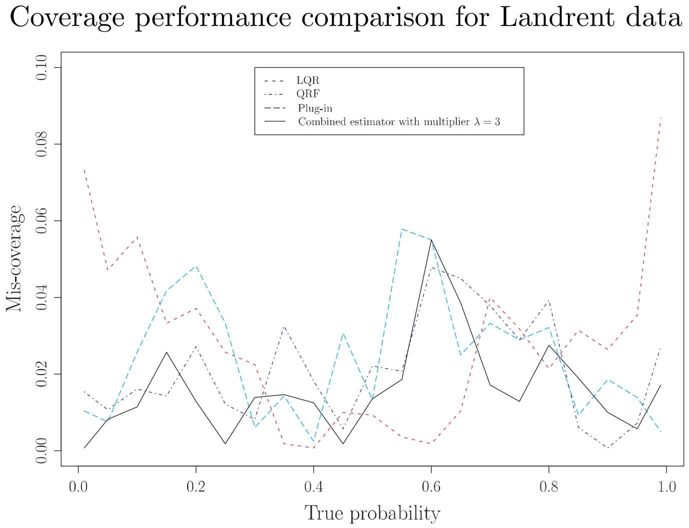
Our second data set, *Landrent* (see Weisberg (2005)), has 67 observations. The response Y is the average rent per acre planted to alfalfa. There are four predictors.

Besides LQR and QRF, we also included a plug-in estimate (see, e.g., Cai (2002)), which is based on linear regression of Y on $X_1, \ldots, X_4$, with stepwise selection of the variables based on AIC. We used an estimate of the form



Mis-coverages for AQRM and candidate estimators at different probability levels for *Autoprice* data.

Figure 4. Summary plot for *Autoprice* data.



Mis-coverages for AQRM and candidate estimators at different probability levels for *Landrent* data.

Figure 5. Summary plot for *Landrent* data.

$\hat{q}_\tau(x) = \hat{m}(x) + \Phi^{-1}(\tau) \times \hat{\sigma}$, where both $\hat{m}(x)$ and $\hat{\sigma}$ were obtained from the selected model. The graphical diagnostics on the residuals did not provide strong evidence against the normality assumption. We used 80% of all data for training (including weight construction), and the remaining 20% was reserved for performance evaluation. The final coverage performance (WICE) in Table 4 is the average over 200 repetitions.

We reach the conclusions that follow:

1. Among the candidates, the plug-in method performed the best under both weighting functions, possibly because the normal linear model describes the data very well.
2. LQR performed the best only when high weights were put on moderate τ values (results are not presented here), where it has an advantage over the other two competing methods.
3. The combined estimators achieved better estimation accuracy under both weighting functions for all $\lambda' s$ that were not too small.
4. Simple averaging did not produce better coverage probability over the best candidate under either weighting function.

Figures 4 and 5 present the coverage performance for each candidate and our combined estimator as a function of τ for the two data sets. They show that our method had good performance under most τ, especially under very large or very small values of τ.

5.2. A time series

Consider the following methods (for details, see e.g., Allen, Boudoukh and Saunders (2004)).

1. The standard GARCH(p, q) model with Gaussian innovations:

$$y_t = \beta_0 + \epsilon_t, \quad \epsilon_t = \sigma_t z_t, \quad z_t \overset{i.i.d.}{\sim} N(0,1),$$
$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \theta_j \sigma_{t-j}^2 .$$

2. The "historical simulation" method that simply uses the sample quantile of a given number (such as 100 or 500) of the most recent observations. A critique on this method can be found in Pritsker (2001).

In the financial markets, Value-at-Risk (VaR) is defined as the predicted worst-case loss with a specific confidence level (for example, 95%) over a period of time (for example, 1 day). Here we consider VaR estimation of the daily index distribution for the S&P500 energy sector with data from January 3, 2000 to November 10, 2006 (available at `http://www.globalfinancialdata.com`). By examining the autocorrelation plot of this series, we decided to apply our candidate procedures to the differenced series. The candidates were the GARCH(1,

Table 6. Observed coverage probabilities for *S&P500 energy* series.

	GARCH(1,1)	HS(100)	HS(250)	$\lambda = 0$	
$\tau = .01$	0.029	0.012	0.000	0.000	
$\tau = .05$	0.070	0.029	0.000	0.000	
$\tau = .10$	0.116	0.047	0.000	0.006	
$\tau = .90$	0.872	0.761	0.552	0.720	
$\tau = .95$	0.953	0.837	0.727	0.855	
$\tau = .99$	0.988	0.930	0.901	0.971	
Avg mis-coverage	0.0147	0.0645	0.1366	0.0746	

	$\lambda = 0.5$	$\lambda = \mathbf{1}$	$\lambda = 3$	$\lambda = 6$	$\lambda = 10$
$\tau = .01$	0.000	0.000	0.000	0.006	0.006
$\tau = .05$	0.058	0.058	0.064	0.070	0.070
$\tau = .10$	0.110	0.110	0.110	0.116	0.116
$\tau = .90$	0.890	0.883	0.878	0.884	0.884
$\tau = .95$	0.959	0.953	0.953	0.953	0.953
$\tau = .99$	0.971	0.983	0.988	0.988	0.988
Avg mis-coverage	0.0112	**0.0093**	0.0102	0.0103	0.0103

1) model, historical simulation with up to 100 (HS100) and 250 (HS250) most recent observations. The historical simulation method with more than 250 observations was tried (not shown here), but gave much worse coverage performance. For constructing the combined estimate, we initialized the estimation of the candidate methods with $T_0 = 200$, and updated both the estimators and weights sequentially. We estimated VaR at $\tau = 0.01, 0.05, 0.1, 0.9, 0.95, 0.99$, since VaR with moderate τ is of little interest to the market analysts. We used the last 10% of the series for evaluation.

In Table 6, we report the observed coverage probabilities at each chosen τ value for all procedures.

1. HS100 performed the best at 1% but, for all other quantiles, GARCH(1, 1) performed the best among the candidates.
2. Our combined estimate with a tuning parameter of $\min(\tau, 1-\tau)$ achieved the best overall performance at the elected τ values.
3. Simple average (i.e., $\lambda = 0$) performed rather poorly, suggesting the need for intelligent model combining methods such as AQRM.

6. Concluding Remarks

A lot of work has been done on the estimation of conditional quantiles. Although many of the proposed parametric methods work well asymptotically, for any realized data with a moderate sample size, insufficient extreme observations typically impair their estimation accuracy at high/low quantiles even if the assumed underlying model (e.g., linear quantile functions) is proper. Nonparametric methods can improve in some aspects, especially for extreme quantiles (as is seen in this work).

Choosing a model (or procedure) from a list for quantile regression can be very challenging. As in other contexts, model selection instability, which can substantially affect estimation/prediction accuracy, is a major issue that should not be ignored. The simulation and examples in this paper show that the candidate procedures performed very differently (relatively speaking) at moderate and extreme quantiles. Thus, selecting a single model based on a traditional model selection criterion is not a good idea for estimating multiple quantiles.

A good approach to address the aforementioned difficulties is the use of model (or procedure) combining as an alternative to choosing one. Under mild regularity conditions, we showed that our proposed estimator performs as well as the best individual candidate in terms of the asymmetric linear risk, with a cost that vanishes at $O(n^{-1/2})$ rate. Simulation examples clearly demonstrated that our method yields improved performance in terms of better overall coverage probability when error standard deviation is not small. The example of the financial series *S&P500 energy* demonstrates that our approach can be very useful for Value-at-Risk estimation.

In summary, for the reasons of model selection uncertainty and the typical dependence of the best candidate quantile regression method on the probability level, model combination methods have a great potential for reliable performance. Our proposed method AQRM can integrate the advantages of general candidate procedures that occur at different probability levels, and thus globally improve over them.

Acknowledgements

The authors thank the referees and the AE for helpful comments on improving the paper. The work of the second author is supported by NSF grant DMS-0706850.

References

Allen, L., Boudoukh, J. and Saunders, A. (2004). *Understanding Market, Credit, And Operational Risk: The Value At Risk Approach.* Blackwell Publishing Limited.

Audibert, J. Y. (2006). A randomized online learning algorithm for better variance control. *Lecture Notes in Comput. Sci.*, 4005 LNAI, 392-407.

Birgé, L. (2006). Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Annales de l'institut Henri Poincare* (*B*) *Probability and Statistics* **42**, 273-325.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5-32.

Bunea, F. and Nobel, A. (2005). Sequential procedures for aggregating arbitrary estimators of a conditional mean. Manuscript.

Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2006). Aggregation and sparsity via l_1 penalized least squares. *Lecture Notes in Comput. Sci.*, 4005 LNAI, 379-391.

Burman, P. and Nolan, D. (1995). A general Akaike-type criterion for model selection in robust regression. *Biometrika* **82**, 877-886.

Cade, B. S., Noon, B. R. and Flather, C. H. (2005). Quantile regression reveals hidden bias and uncertainty in habitat models. *Ecology* **86**, 786-800.

Cai, Z. (2002). Regression quantiles for time series. *Econom. Theory* **18**, 169-192.

Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization.* Springer, New York.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8**, 140-154.

Granger, C. W. J. (1969). Prediction with a generalized cost of error function. *Operational Research Quarterly* **20**, 199-207.

Granger, C. W. J. (1989). Combining forecasts - twenty years later. *Journal of Forecasting* **8**, 167-173.

Hall, P., Wolff, R. C. L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.* **94**, 154-163.

He, X., Ng, P. and Portnoy, S. L. (1998). Bivariate quantile smoothing splines. *J. Roy. Statist. Soc. Ser. B* **60**, 537-550.

Koenker, R. W. and Bassett, G. W. (1978). Regression quantiles. *Econometrica* **46**, 33-50.

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

Koenker, R. W., and Xiao, Z. (2006). Quantile autoregression. *J. Amer. Statist. Assoc.* **101**, 980-990.

Lecué, G. (2006). Lower bounds and aggregation in density estimation. *J. Mach. Learn. Res.* **7**, 971-981.

Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52**, 3396-3410.

Machado, J. A. F. (1993). Robust model selection and M-estimation. *Econom. Theory* **9** 478-493.

Meinshausen, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7**, 983-999.

Nemirovski, A. (2000). *Topics in Non-parametric Statistics. Lecture Notes in Mathematics*, Vol. 1738. Springer, New York.

Pritsker, M. (2001). The hidden dangers of historical simulation. Finance and Economics Discussion Series 2001-27, Board of Governors of the Federal Reserve System (U.S.).

Ronchetti, E. (1985). Robust model selection in regression. *Statist. Probab. Lett.* **3**, 21-23.

Ronchetti, E., Field, C. and Blanchard, W. (1997). Robust linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **92**, 1017-1023.

Taylor, J. W. and Bunn, D. W. (1998). Combining forecast quantiles using quantile regression: Investigating the derived weights, estimator bias and imposing constraints. *J. Appl. Stat.* **25**, 193-206.

Tsybakov, A. B. (2003). Optimal rates of aggregation. In *Proceedings of* 16*th Annual Conference on Learning Theory* (*COLT*) *and* 7*th Annual Workshop on Kernel Machines. Lecture Notes in Artificial Intelligence*, 303-313, Springer-Verlag, Heidelberg.

Wei, Y. and He, X. (2006). Conditional growth charts (with discussion). *Ann. Statist.* **34**, 2069-2131.

Weisberg, S. (2005). *Applied Linear Regression*. Wiley-Interscience, New York.

Yang, Y. (2001). Adaptive regression by mixing. *J. Amer. Statist. Assoc.* **96**, 574-588.

Yang, Y. (2004a). Combining forecasting procedures: some theoretical results. *Econom. Theory* **20**, 176-222.

Yang, Y. (2004b). Aggregating regression procedures to improve performance. *Bernoulli* **10**, 25-47.

Yang, Y. and Barron, A. R. (1999). Information theoretic determination of minimax rates of convergence. *Ann. Statist.* **27**, 1564-1599.

Yu, K., Lu, Z. and Stander, J. (2003). Quantile regression: applications and current research areas. *The Statistician* **52**, 331-350.

Yuan, Z. and Yang, Y. (2005). Combining linear regression models: when and how? *J. Amer. Statist. Assoc.* **100**, 1202-1204.

Zhou, K. Q. and Portnoy, S. L. (1996). Direct use of regression quantiles to construct confidence sets in linear models. *Ann. Statist.* **24**, 287-306.

Amylin Pharmaceuticals, Inc., 9360 Towne Centre Drive, San Diego, CA 92121, U.S.A.

E-mail: kevin.shan@amylin.com

School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street S.E., Minneapolis, MN 55455, U.S.A.

E-mail: yyang@stat.umn.edu

(Received August 2007; accepted March 2008)