Model Combination for Quantile Regression

Yuhong Yang
School of Statistics
University of Minnesota

2008 Beijing Conference on Machine Learning and Data Mining

June 19, 2008

Joint work with Kejia Shan

Supported by US NSF Grant #00003096

Outline

- Quantile estimation
- Difficulty in model comparison for quantile regression
- Adaptive quantile regression by mixing (AQRM)
- Optimality of AQRM
- Numerical results
- Conclusion

Quantile regression

- Conditional quantile estimation is useful in agriculture, economics, finance, etc.

- Numerous methods have been proposed under different settings including the classical linear regression, nonlinear regression, time series, and longitudinal experiment.

- When a range of τ values are considered, the quantile profile provides information much beyond the conditional mean.

Linear quantile regression (LQR)

- Koenker and Bassett (1978) introduced regression quantile estimation by minimizing an asymmetric loss function

$$L_\tau(\xi) = \tau I_{\xi \geq 0} + (1-\tau) I_{\xi < 0}$$

  for $0 < \tau < 1$, known as the check or pinball loss.

- 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 $x'\beta$ and the coefficients β is estimated by minimizing $\sum_i L_\tau(y_i - x_i'\beta)$.

Nonparametric methods

- To increase flexibility, nonparametric and semiparametric methods have also been developed for quantile regression.

- For example, one may 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.

Quantile regression forests (QRF)

- Meinshausen (2006) proposed QRF.
- 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 grown in this fashion.
- The conditional quantile of Y given $X = x$ is then approximated by the average prediction from the collection of random trees.
- Numerical results demonstrated its good performance in problems with high-dimensional predictors, particularly at extreme values of τ (τ near zero or one).

Model selection for CQE

- Ronchetti (1985) introduced a robust version of AIC, which takes the form of the observed check loss plus a multiple of model size.
- Machado (1993) proposed a generalized Schwarz Information Criterion, which is 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).
- More research is needed.

Model combination for CQE

- 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.

- Not much theoretical understanding on combining quantile regression estimators is given in the literature.

- When the quantile profile is of interest, it is particularly important to consider model combination methods.
  - Usual model selection uncertainty exists.
  - Different quantile regression estimators typically have distinct relative performances that depend on the value of τ.
  - A true parametric model does not necessarily produce a good quantile estimator.
  - It is a proper objective to integrate the advantages of various methods and thus globally improve over them.

Statistical theories on combining arbitrary estimators

- A recent focus is on construction of methods that adaptively share strengths of a list of arbitrary estimators.
- There are two directions on combining arbitrary estimators, one being combining for adaptation and the other being combining for improvement.
- Combining for adaptation pays a price of order $1/n$ in MSE, and combining for improvement typically pays a higher price than $1/n$, depending on the aggresiveness of the goal. (See, e.g., Nemirovskii 2000; Yang 2001 and 2004; Catoni 2004; Tsybakov 2003).

- Adaptive combination can be done to achieve multi-directional combination.

- Algorithms in computer science literature typically require boundedness of the response variables or boundedness of the loss function, which is undesirable for our statistical estimation.

- We follow the spirit of Yang (2001), Catoni (2004), Yang (2004) for combining arbitrary quantile estimators.

Problem setup

- Observe (Y_i, X_i), $i = 1, \cdots, n$, where $X_i = (X_{i1}, \cdots, X_{ip})$ is a p-dimensional predictor.

- Assume the true underlying relationship between Y and X is characterized by:

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

where ϵ_i are i.i.d. from a distribution with mean zero and variance one and are independent of the predictors.

- The conditional quantile of Y given $X = x$ has the form

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

where F is the cumulative distribution function of the error.

- Natural to estimate $q_\tau(x)$ by first obtaining $\hat{m}(x)$, $\hat{\sigma}(x)$ and $\hat{F}^{-1}(\tau)$.

- If the $m(\cdot)$ is a linear function of x and $\sigma(\cdot)$ is constant, LQR is expected to perform well. However, if either the mean function is nonlinear or the scale function is non-constant in the predictors, bias will be involved.

- In real applications, the performance of LQR on extreme quantiles is usually impaired by insufficient extreme observations.

- 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 a 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.

- In the context of conditional mean regression, Yang (2001) proposed the adaptive regression by mixing (ARM) method, in which a set of weights is sequentially 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 this work, the check loss function is naturally oriented towards quantile estimation.

- However, the distinct natures of the absolute-type and quadratic-type of losses present a non-trivial work to derive 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.

Adaptive quantile regression by mixing (AQRM)

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 $n - n_0$).

- 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.
- Based on $Z^{(1)}$, obtain candidate estimates of the conditional quantile function $q_\tau(x)$ by $\hat{q}_{\tau,j,n_0}(x) = \hat{q}_{\tau,j,n_0}(x; Z^{(1)})$. Use $\hat{q}_{\tau,j,n_0}$ to obtain the predicted quantiles from the j^{th} candidate procedure for $Z^{(2)}$, for each $j = 1, \cdots, M$.
- Compute the candidate weights as follows

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

  where $\lambda > 0$ is a tuning parameter.

- Repeat steps $1-3$ a number of times and average the weights. Denote them by $\tilde{W}_j$. Our final estimator of the conditional quantile function of Y at $X = x$ is $\hat{q}_{\tau,\cdot,n}(x) = \sum_{j=1}^{M} \tilde{W}_j \hat{q}_{\tau,j,n}(x)$.

Sequential weighting

- For online prediction, sequential updating is natural.
- First obtain $\hat{q}_{\tau,j,n_0}$ from $\{(y_l, x_l)\}_{l=1}^{n_0}$ (the initial set of observations) and the weights are updated sequentially once an additional observation is made.
  - define sequential weight $W_{j,i}$ as
  $$W_{j,i} = \frac{\prod_{l=n_0+1}^{i-1} \exp\{-\lambda L_\tau(y_l - \hat{q}_{\tau,j,l}(x_l))\}}{\sum_{k=1}^{M} \prod_{l=n_0+1}^{i-1} \exp\{-\lambda L_\tau(y_l - \hat{q}_{\tau,k,l}(x_l))\}},$$
  - the combined estimate of $q_\tau(x)$ at time i is
  $$\hat{q}_{\tau,\cdot,i}(x) = \sum_{j=1}^{M} W_{j,i}\hat{q}_{\tau,j,i}(x).$$

Role of λ

- The tuning parameter λ controls how much the weights rely on the check loss performance.

- When $\lambda \downarrow 0$, simple averaging results; when $\lambda \to \infty$, the candidate with the best historic check loss is selected.

Conditions

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

Condition 1: The quantile estimators satisfy that $\sup_{j\geq 1, i\geq 1} |\hat{q}_{\tau,j,i}(x_i) - q_\tau(x_i)| \leq A_\tau$, for some positive constant A_τ with probability one.

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.

Oracle inequalities on performance

Theorem: Under Conditions 0-3, when the tuning parameter λ is small enough, the risk $\frac{1}{n-n_0}\sum_{i=n_0+1}^{n} EL_\tau(Y_i - \hat{q}_{\tau,\cdot,i}(X_i))$ is upper bounded by

$$\inf_j \left\{ \frac{1}{n-n_0} \sum_{i=n_0+1}^{n} EL_\tau(Y_i - \hat{q}_{\tau,j,i}(X_i)) + \tilde{C}\sqrt{\frac{\log(M)}{n-n_0}} \right\}, \qquad (2)$$

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

- The inequality says that the risk of the combined prediction is automatically close to the risk of the best individual (parametric or not), with the difference being of order $(n - n_0)^{-1/2}$ when λ is chosen properly.

- When n_0 and $n - n_0$ are of the same order, the risk bound shows that the combined quantile prediction converges at the best rate offered by the candidate procedures for both parametric and non-parametric situations.

- 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 may not be improved.

- Second, since the best procedure is unknown, the combining approach can reduce uncertainty of model selection.

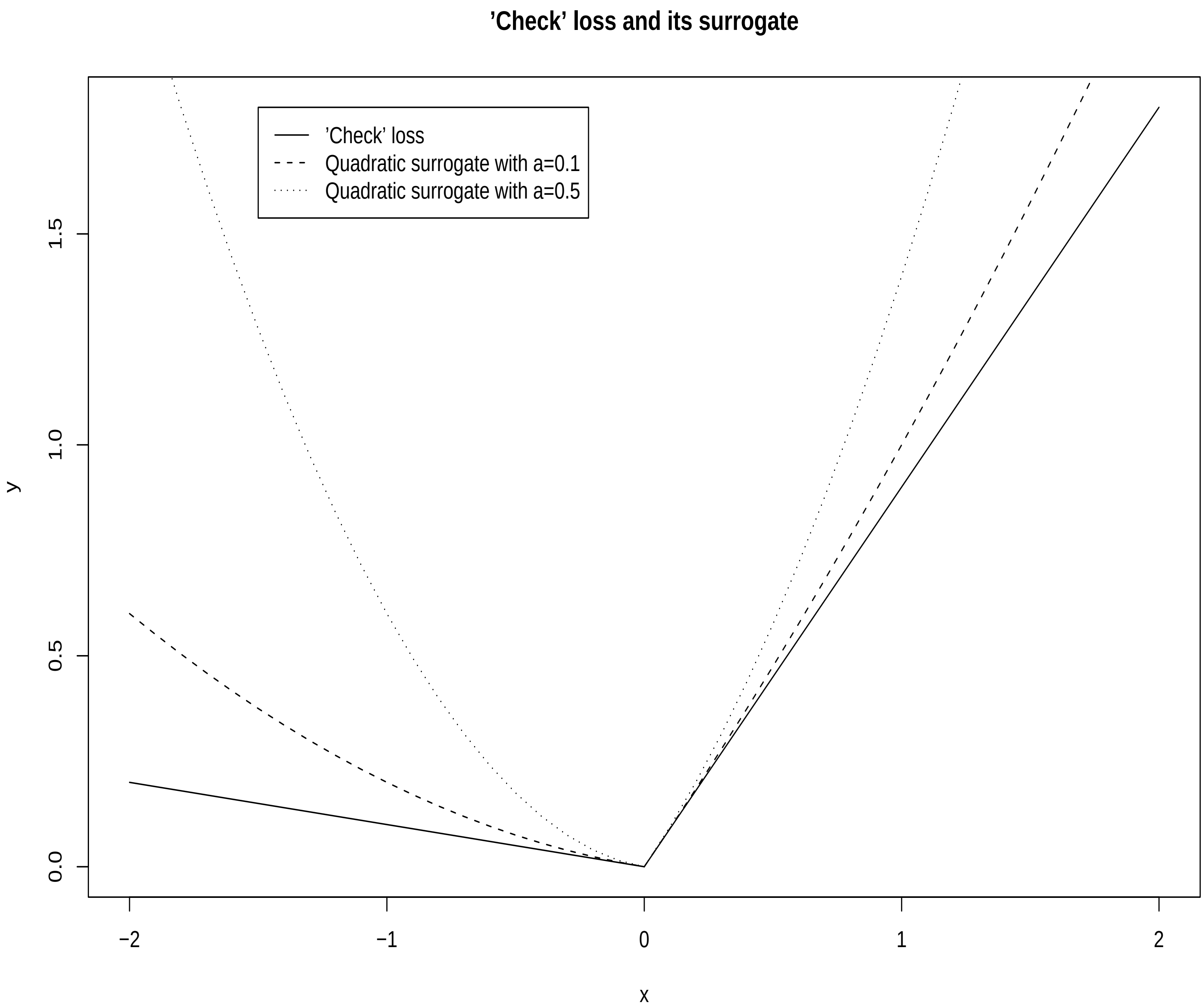
- Third, because quantiles at a range of probability level are often of interest at the same time but the candidate quantile estimators typically have different ranks in performance, the combined estimators have a good potential to beat them all globally.

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 $a > 0$ and use it in the construction of weights of the candidate quantile regression procedures. The new weight is

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

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



'Check' loss and its surrogate
'Check' loss
Quadratic surrogate with a=0.1
Quadratic surrogate with a=0.5
y
0.0
0.5
1.0
1.5
x
−2
−1
0
1
2

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 ϵ_t is independent of $\{(Y_s, X_s) : s < t\}$ and X_t.

Let T be the length of the whole series. The sequential combining algorithm AQRM for the time series setting is given below.

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

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

  where $W_{j,T_0+1} = \frac{1}{M}$.
- We 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: Under some conditions 1-3, when the tuning parameter λ is chosen properly, $\sum_{t=T_0+1}^{T} EL_\tau(Y_t - \hat{q}_{\tau,\cdot,t}(X_t))$ is upper bounded by

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

where $\tilde{C}$ is a constant.

Numerical results

Candidate procedures

- LQR (Koenker and Bassett 1978), *R* package *quantreg*
- QRF (Meinshausen 2006), *R* package *quantregForest*.
- A plug-in estimator.

Measure of performance

- In the literature, performance of quantile regression is usually measured by the coverage probability at some fixed τ value(s).

- 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.

- We focus on the overall performance of a quantile regression procedure over the full range of τ in $(0, 1)$.

- Let g denote a weighting function on $\tau \in (0, 1)$ such that $g \geq 0$ and $\int_0^1 g(\tau)d\tau = 1$, which is used to differentiate the importance of τ values in different regions.

- We choose two different g functions in this work, one being the uniform weight and the other being the Beta(0.8,0.8) density, which emphasizes extreme $\tau's$.

- Weighted Integrated Absolute Error (WIAE): the mean of

$$\int\int |\hat{q}_\tau(x) - q_\tau(x)|g(\tau)d\tau P(dx).$$

- Weighted Integrated Coverage Error (WICE):

$$\int_0^1 |\hat{\tau} - \tau|g(\tau)d\tau.$$

  In implementing this, we use a random data splitting, done 100 times.

- 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 are based on 100 runs in each case.

- The sample size is 200, with equal training-testing data splitting randomly done 50 times.

- The tuning parameter λ is taken of the form $\lambda_\tau = \lambda \times \min(\tau, 1-\tau)$, where $\tau \in \{0.01, 0.05 \times k, 0.99\}_{k=1}^{19}$.

Simulation models

Case 1.

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

where $X \sim \chi^2_4$, $\epsilon_1 \sim N(0,1)$, $Z = X + \epsilon_1$, $\epsilon_2 \sim f$, with $\mu_f = 0$ and $\sigma_f = \sigma$, and X, ϵ_1, ϵ_2 are generated independently of each other. Besides $N(0,1)$, shifted gamma distribution is also considered to allow asymmetric error.

Candidate procedures:

- LQR with predictors X and Z
- QRF with predictors X and Z
- LQR with predictors X, Z and $\sqrt{X}$.

Case 2. Same as Case 1, except that $\sigma_f(x,z) = \sigma\sqrt{x}$.

Case 3. Randomly models:

- Generate $\beta = (\beta_1, \cdots, \beta_6)$ uniformly.
- The true model is $Y = \beta' X + \sigma\epsilon$, where $X = (X_1, \cdots, X_6)$ has independent $N(0, 1)$ components, and ϵ is either from a standard normal distribution or a shifted gamma with mean zero and variance one.
- Two hundred sets of coefficients are generated.

Case 4. The model is

$$Y = \beta' X + 2\exp(-0.35X_2 - 1.1X_3) + \sigma\epsilon\sqrt{X_2^2 + 0.8X_4^2}$$

and the other aspects are the same as Case 3.

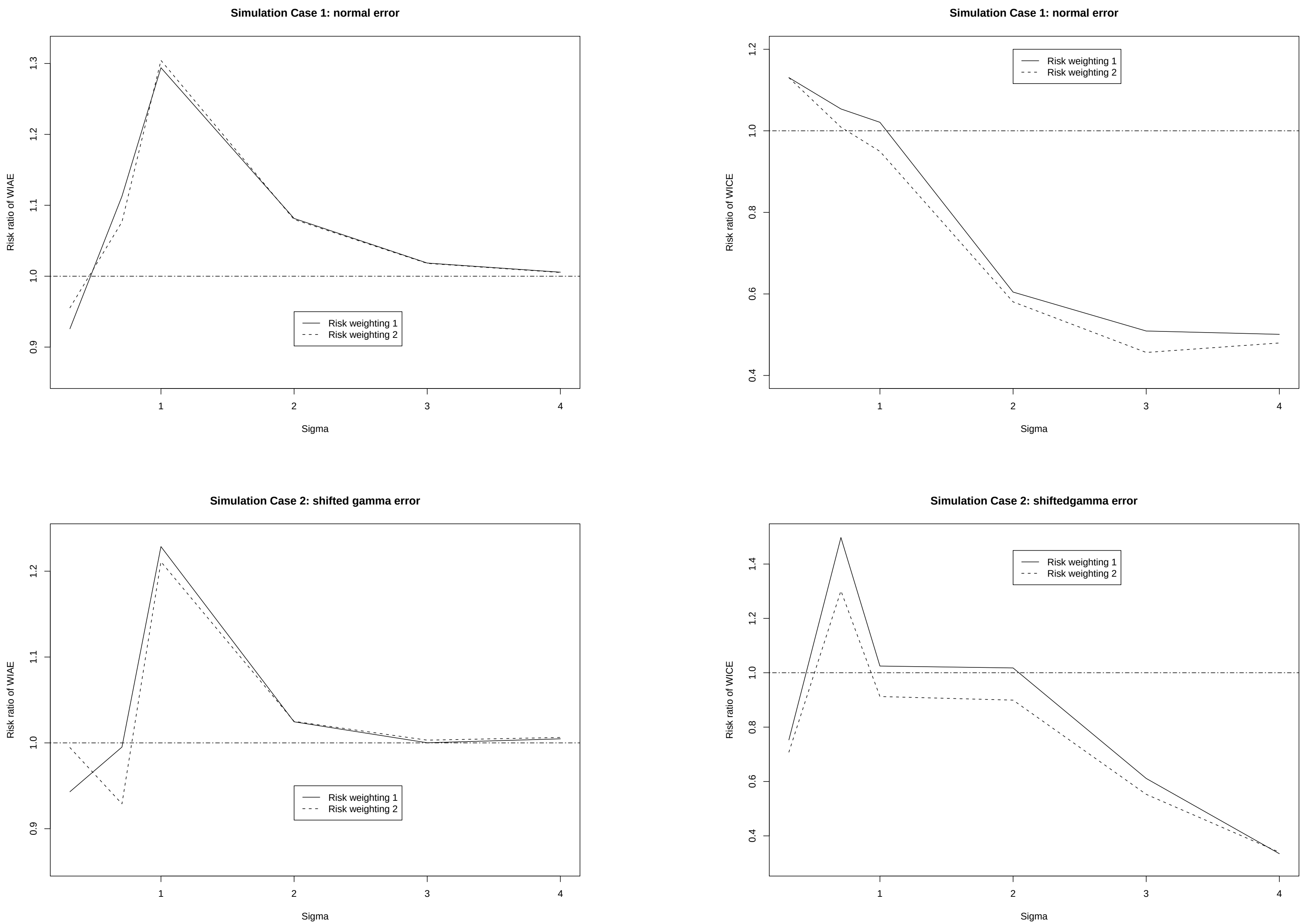


Figure 2: Risk ratios for Case 1 and Case 2.

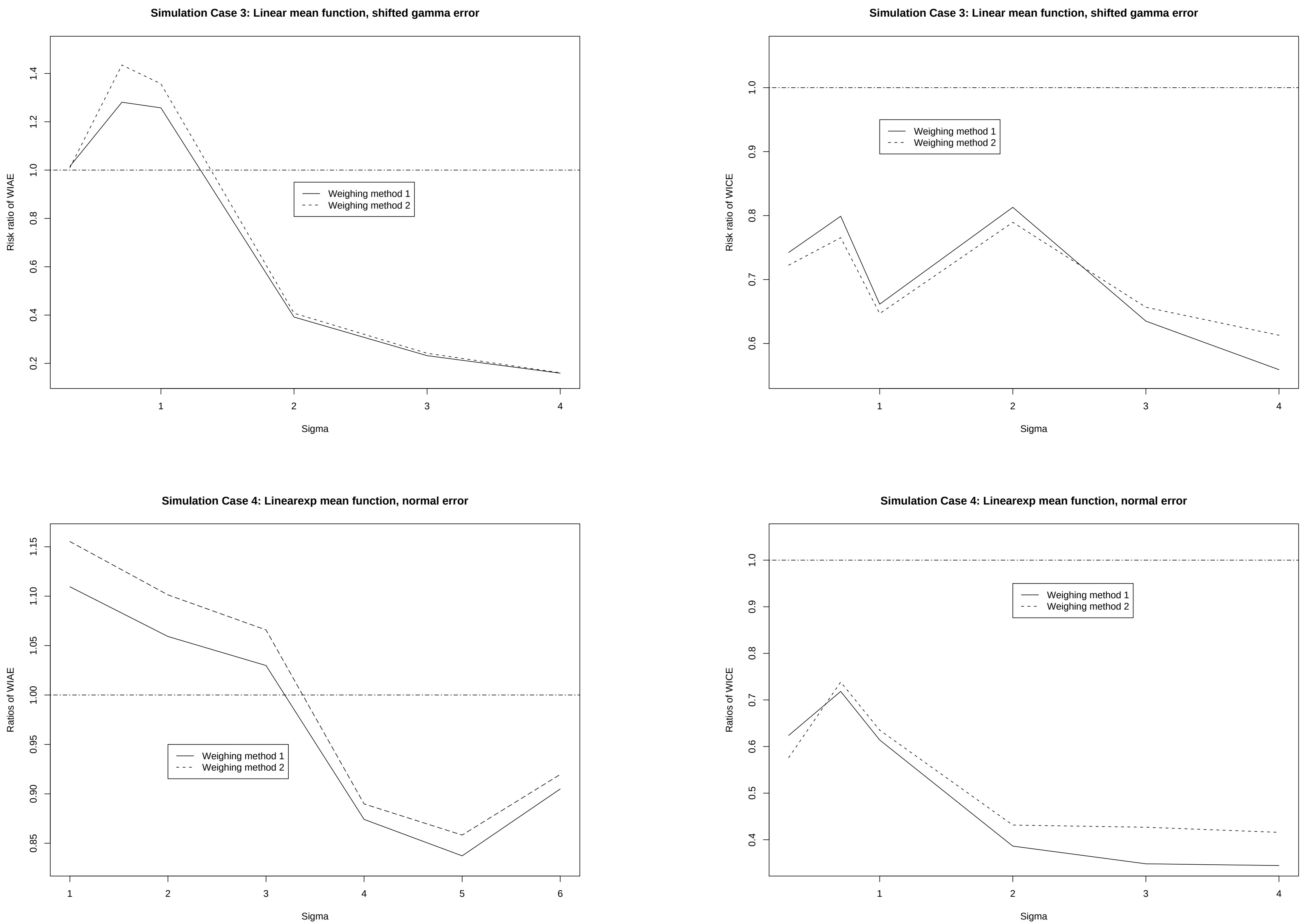


Figure 3: Risk ratios for Case 3 and Case 4

Summary of the simulations:

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

- The L_1 risk of QRF for estimating $q_\tau(x)$ is often the largest, compared to the other two candidates, when σ is small ($\sigma \leq 0.707$) and often the smallest when σ is 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.

- AQRM performed well. For the error distributions considered and almost all τ, when $\sigma \geq 2$, AQRM can basically tie with or perform better than the best candidate both in terms of observed coverage probability and in L_1 risks.

- The two performance measures are quite different.

- The random coefficient cases reveal substantial advantages of AQRM. At the noise levels considered, the coverage errors of AQRM are consistently smaller than those of the candidates. Because the coefficients were randomly generated, the ranking of LQR versus QRF can change as well, in which case the combined estimator can be much better than any fixed choice of the candidates.

Two regression data sets

The data set *Autoprice:*

- $n = 159$ observations with 14 continuous variables and one discrete variable.
- log-transformed price as the response variable *price*
- candidate quantile regression methods are the LQR (with AIC) and QRF
- 75% of data for training (including weight construction for combining the procedures), and remaining 25% for final performance evaluation.
- 200 repetitions

The data set *Landrent:*

- 67 observations
- response Y is the average rent per acre planted to alfalfa
- four predictors.
- Besides LQR and QRF, we also included a plug-in estimate, which is based on linear regression of Y on $X_1, \cdots, X_4$ with stepwise selection of the variables based on AIC.
- 80% of data for training (including weight construction), and the remaining 20% is reserved for performance evaluation.

A time series data set

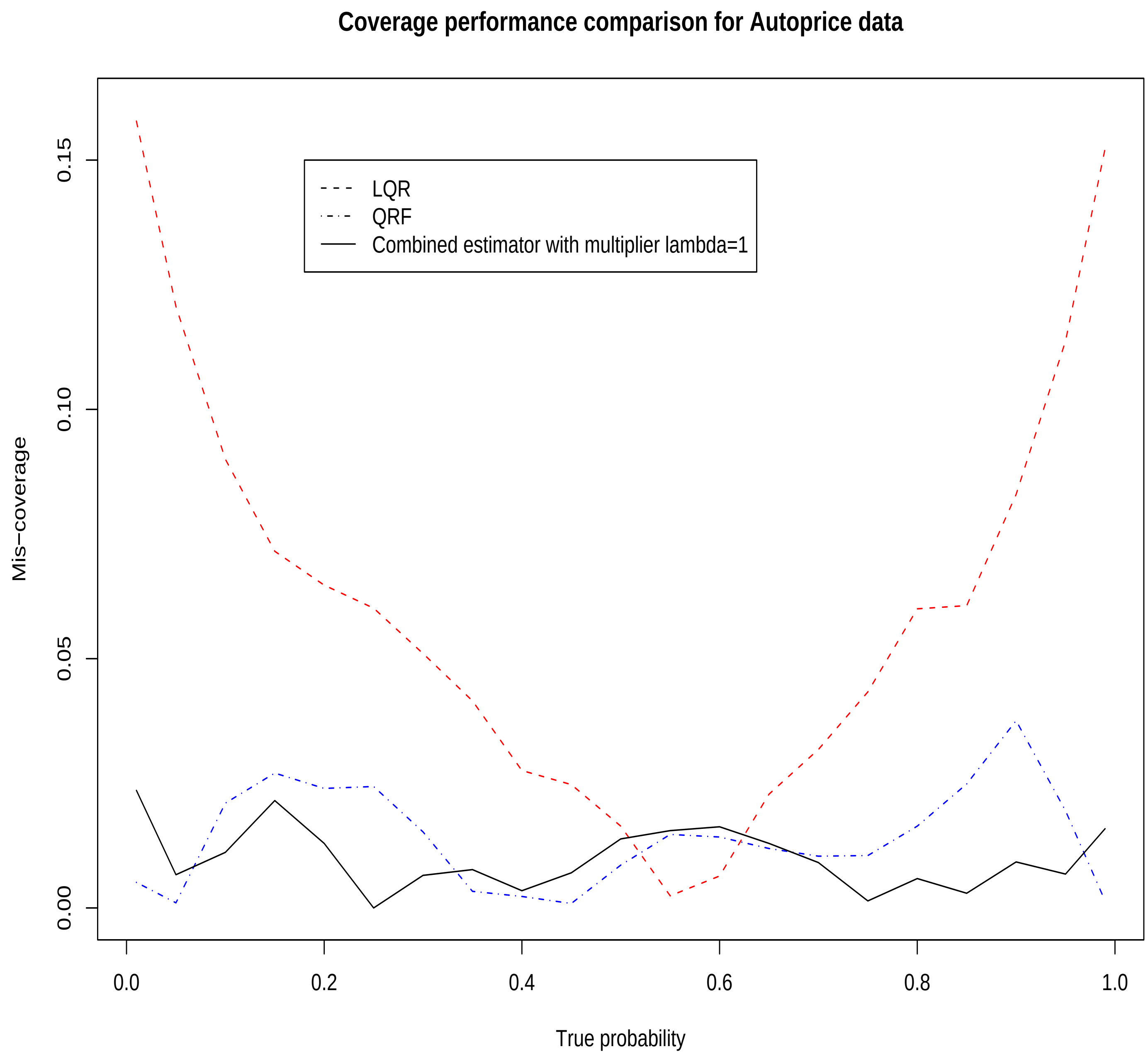
- In financial markets, Value-at-Risk (VaR) is defined as the predicted worst-case loss with a specific confidence level over a period of time.
- We consider VaR estimation of the daily index distribution for S&P500 energy sector with data from January 3, 2000 to November 10, 2006.
- The candidates are GARCH(1,1) model, historical simulation with up to 100 (HS100) and 250 (HS250) most recent observations.
- The last 10% of the series is used for evaluation.

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

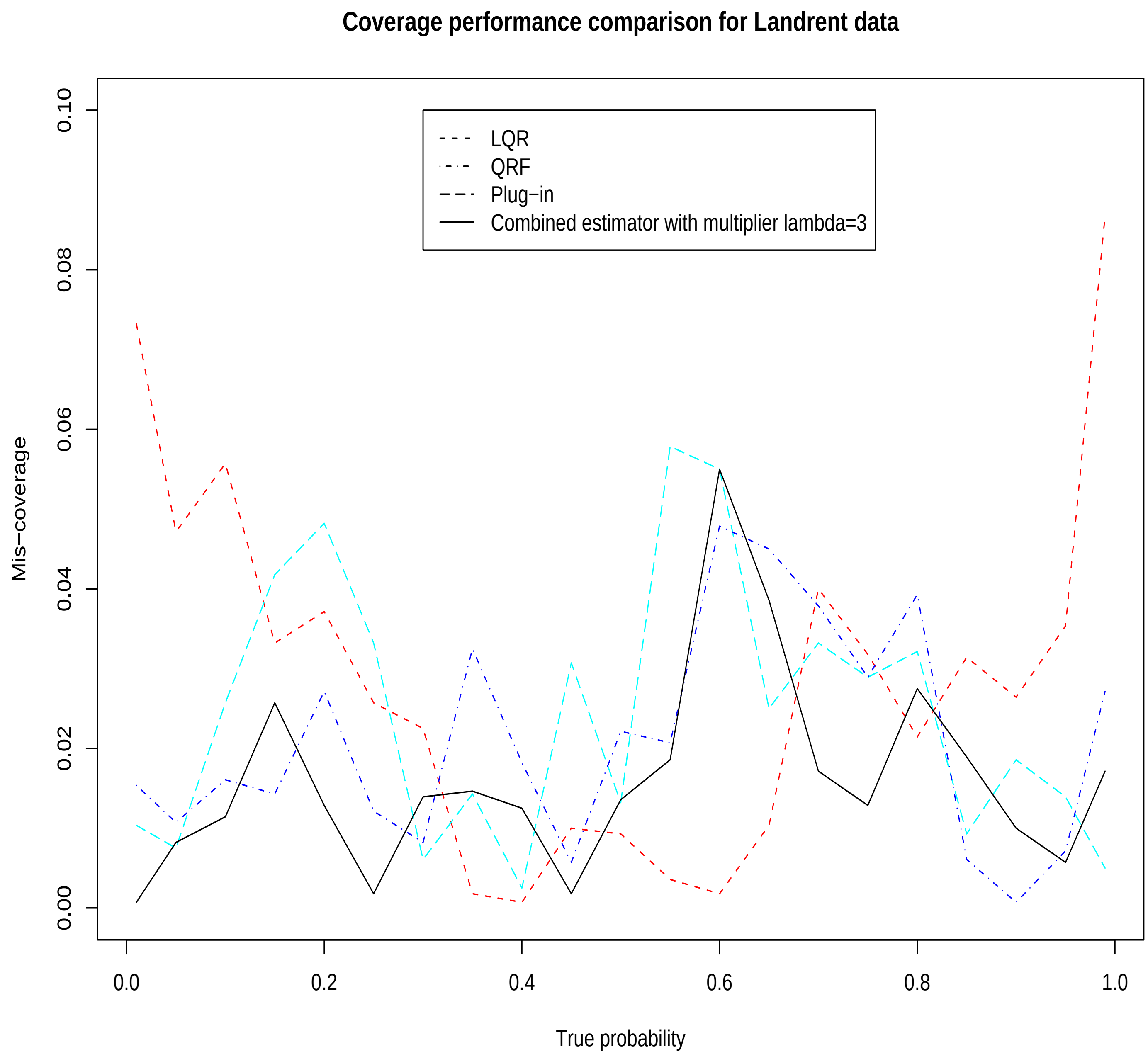
Table 1: Weighted Integrated Coverage Errors ($\times 10^{-2}$) for *Autoprice* data.

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

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



Coverage performance comparison for Autoprice data
LQR
QRF
Combined estimator with multiplier lambda=1
Mis−coverage
0.00
0.05
0.10
0.15
0.0
0.2
0.4
0.6
0.8
1.0
True probability


Coverage performance comparison for Landrent data
LQR
QRF
Plug−in
Combined estimator with multiplier lambda=3
Mis−coverage
0.00
0.02
0.04
0.06
0.08
0.10
0.0
0.2
0.4
0.6
0.8
1.0
True probability

	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

Table 3: Observed coverage probabilities for *S&P500 energy* series.

Concluding remarks

- Although correct parametric methods work well asymptotically, for a moderate sample size, insufficient extreme observations typically impair their accuracy at high/low quantiles even if the assumed underlying model is proper.
- It is desirable to consider multiple procedures.
- Choosing a model/procedure from a list for quantile regression is challenging.
  - model selection instability
  - the candidate procedures typically perform very differently at moderate and extreme quantiles
  - selecting a single model based on a traditional model selection criterion is not a good idea for estimating multiple quantiles.

- model/procedure combining can be very helpful

- AQRM performs as well as the best individual candidate in terms of the asymmetric linear risk, with a cost that vanishes at $O\left(n^{-\frac{1}{2}}\right)$ rate.

- Simulation examples clearly demonstrate that our method yields improved performance in terms of better overall coverage probability when error standard deviation is not small.

- AQRM can integrate the advantages of general candidate procedures that occur at different probability levels and thus globally improve over them.