Understanding the Literature on Model Selection and Model Combination

Yuhong Yang

School of Statistics

University of Minnesota

WORKSHOP ON CURRENT TRENDS AND CHALLENGES IN MODEL SELECTION AND RELATED AREAS

July 25, 2008

Part of the work is joint with Kejia Shan and Zheng Yuan

Supported by US NSF Grant DMS-0706850

Outline

- Some gaps/confusions/misunderstandings/controversies
- The true model or searching for it does not necessarily give the best estimator
  - A conflict between model identification and minimax estimation
  - Improving the estimator from the true model by combining with a nonparametric one (*combining quantile estimators)*
- Cross-validation for comparing regression procedures

- Model selection diagnostics
  - Can the selected model be reasonably declared the "true" model?
  - Should I use model selection or model averaging?
  - Does the model selection uncertainty matter for my specific target of estimation?
- Concluding remarks

Some gaps/confusions/misunderstandings/controversies

- Existence of a true model among candidates and consequences on estimation
- Pointwise asymptotics versus minimax
- Numerical results on model selection in the literature
  - Fairness and informativeness of the numerical results in the literature
  - Cross-validation for model/procedure comparison
- Model averaging is always better than model selection?

Existence of a true model among candidates and consequences on estimation

- Perhaps most (if not all) people agree that the models we use are convenient simplifications of the reality. But is it reasonable, sometimes, to assume the true model is among candidates?
- When one assumes that the true model is among the candidates, consistency in selection is the most sought property of a model selection criterion. Otherwise, asymptotic efficiency or minimax rate of convergence is often the goal.
- A philosophy traditionally taken by our profession: identify the best model first and then apply it for decision making.
- It makes intuitive sense, but ...

Consistency: Is it relevant and the right target to pursue?

- A conflict between model identification and minimax estimation
- Improving estimators from the true model, e.g.,
  - improving LQR by combining with a nonparametric one (*combining quantile estimators*)
  - improving plug-in MLE of extreme quantile by modifying the likelihood function (Ferrari and Yang, 2008)

- Key properties of BIC are 1) consistency in selection; 2) asymptotic efficiency for parametric cases
- Key properties of AIC are 1) minimax-rate optimality for estimating the regression function for both parametric and nonparametric cases; 2) asymptotic efficiency for nonparametric cases

Can we have these hallmark properties combined?

Theorem. (Yang, 2005, 2007) Consider two nested parametric models, model 0 and model 1.

1. No model selection criterion can be both consistent in selection and minimax-rate adaptive at the same time.

2. For any model selection criterion, if the resulting estimator is pointwise-risk adaptive, then the worst-case risk of the estimator cannot converge at the minimax optimal rate under the larger model.

3. Model averaging, BMA included, cannot solve the problem either.

4. For any model selection rule with the false selection probability under model 0 converging at order q_n for some q_n decreasing to zero, the worst case risk of the resulting estimator is at least of order $(-\log q_n)/n$.

See Leeb and Pötscher (2005) for closely related results.

- Consider quantile regression. Even if we assume that the data come from a nice and known parametric model, the resulting estimator may perform poorly for extreme quantiles, e.g., worse than a robust nonparametric one. Thus consistency may or may not lead to well-performing estimators.

- On the other hand, the estimator from the true parametric model usually performs excellently for estimating median or moderate quantiles.

- One natural approach is to combine the parametric and nonparametric estimators appropriately to have better performance that takes advantage of both of the estimators.

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

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

  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 semi-parametric methods have also been developed for quantile regression.
- For example, Meinshausen (2006) proposed Quantile regression forests (QRF).
- 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/combination for CQE

- There are model selection/combination methods for quantile regression, but not much theory is given.
- 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.

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 asymptotically. 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.
- We take the approach of Catoni that does not require specification of the error distribution (e.g., Catoni (2004)).

- The check loss function is naturally oriented towards quantile estimation and for weighting.

- However, the distinct natures of the absolute-type and quadratic-type of losses present a non-trivial work to derive an oracle inequality for the quantile regression combining problem.

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

$$\hat{q}_{\tau,.,n}(x) = \sum_{j=1}^{M} \tilde{W}_j \hat{q}_{\tau,j,n}(x).$$

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

$$\hat{q}_{\tau,.,i}(x) = \sum_{j=1}^{M} W_{j,i} \hat{q}_{\tau,j,i}(x).$$

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. (Shan and Yang, 2008) 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

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

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

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.

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):

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

- 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

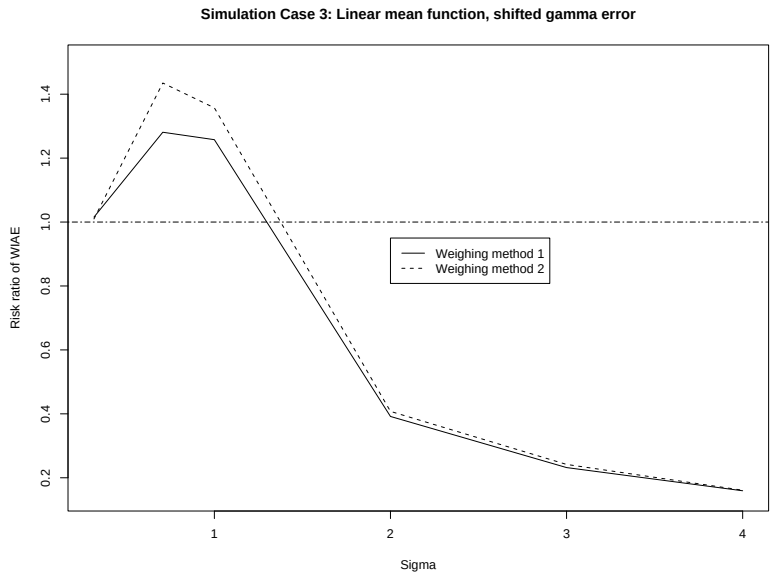
Case 1. Randomly generated models:

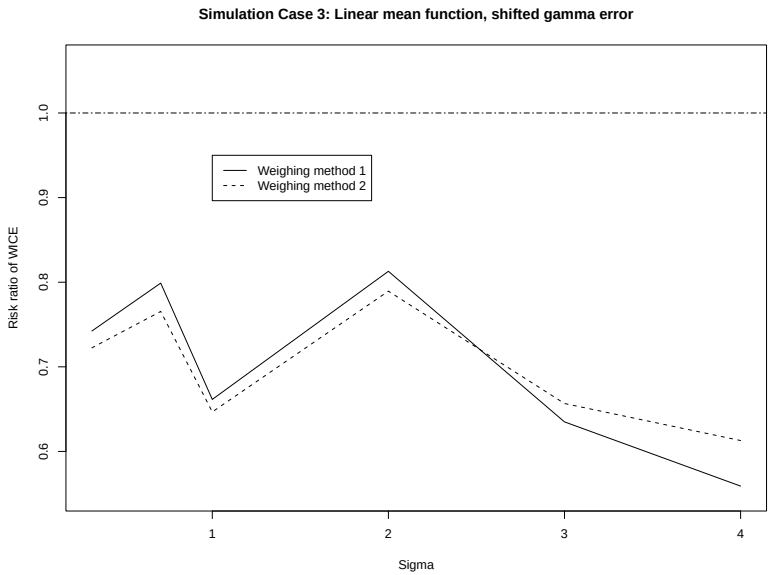
- Generate $\beta = (\beta_1, \cdots, \beta_6)$ uniformly.
- The true model is $Y = \beta' X + \sigma\epsilon$, where $X = (X_1, \cdots, X_6)$ has independent $N(0, 1)$ components, and ϵ is either from a standard normal distribution or a shifted gamma with mean zero and variance one.
- Two hundred sets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

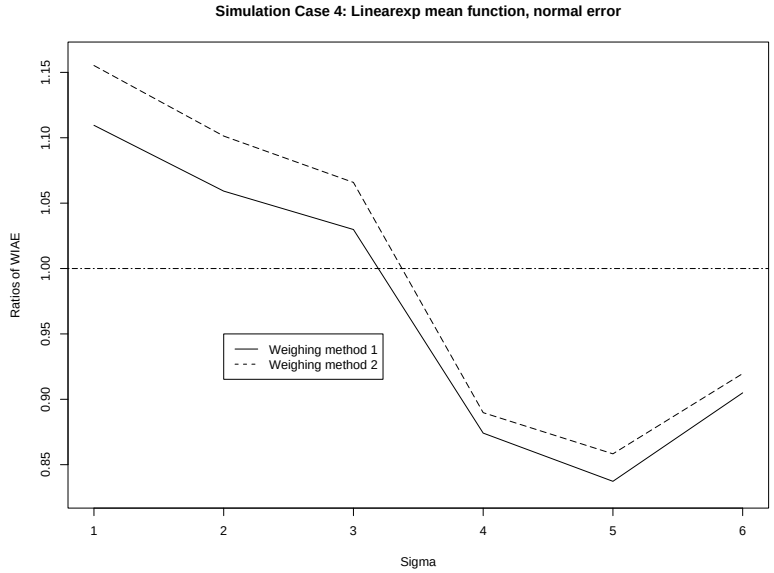
Case 2. The model is

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

and the other aspects are the same as Case 1.







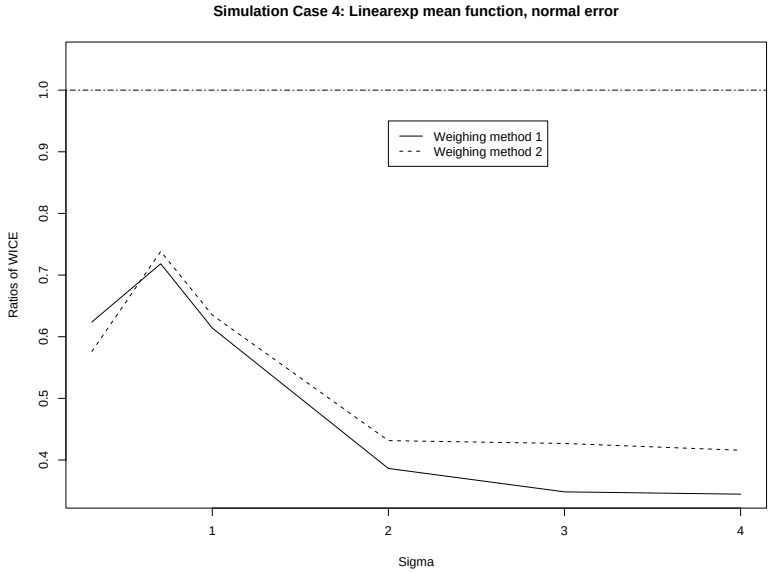


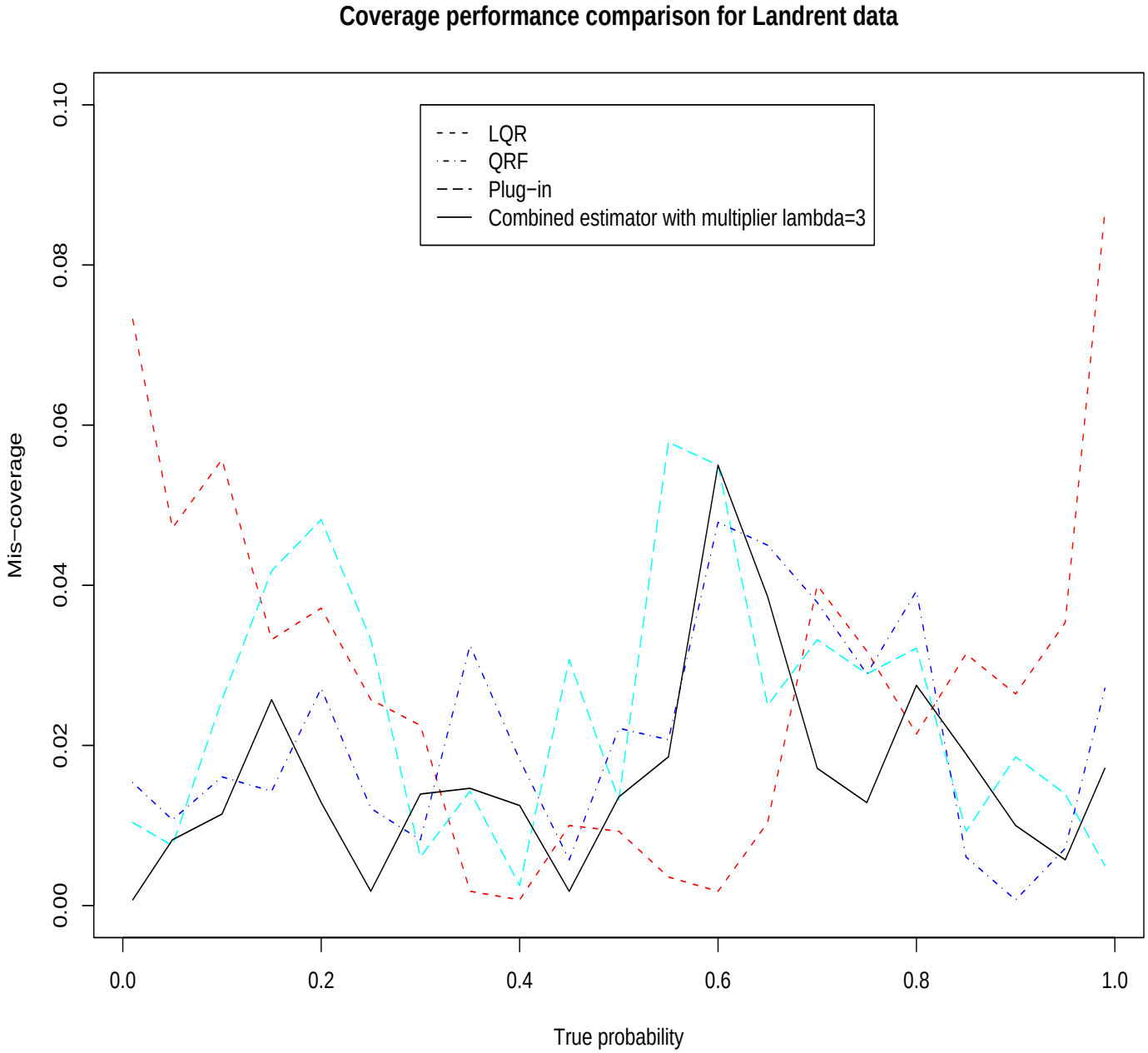
Figure 1: Risk ratios for Cases 1 and 2

A regression 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.

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


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

A summary

- Although methods based on correct parametric models work well asymptotically, for a moderate sample size, insufficient extreme observations may impair their accuracy at high/low quantiles.
- Therefore consistency in selection is not necessarily the right thing to do for quantile regression.
- Model/procedure combining can be very helpful.
- AQRM performed well by integrating the advantages of candidate procedures.

Numerical results on model selection in the literature

- Fairness and informativeness of the numerical results in the literature
- Cross-validation for model/procedure comparison

A gap between numerical results in literature and objective & informative understanding

- It is understandable for us to "sell" our own methods, but often the simulation/data examples are too narrowed
- This creates lack of understanding or mis-understanding

Insufficient numerical work

- Choosing one or two favorable simulation settings or examples
- Lack of a fair comparison with other methods and lack of a proper analysis of the outcomes
- Lack of insightful understandings: when one's method should be preferred and when it should not

Suggestions to address the issues (we are statisticians after all!)

- Design the simulation study soundly and systematically: "factorial design", randomly generate model size and parameters, etc.
- Present both idealistic and realistic (including negative) results
- Include the standard errors whenever possible and analyze the simulation outcomes formally if suitable

The use of cross validation in the literature for comparing procedures

- CV is often used to compare different candidate procedures
- It is not uncommon (e.g., in bioinformatics) that conclusions were drawn based on CV with very small evaluation size (e.g., 1)
- How reliable is the resulting conclusion?
- How to choose the data splitting ratio?

Let's have some theoretical understanding on the use of CV for procedure comparison. We focus on regression, but similar results hold for classification as well.

- CV can be used for different purposes:
  - estimating prediction error
  - tuning parameter selection
  - selecting a model which will be used for prediction
  - selecting a model for consistency
- For the first three, typically delete-one CV works optimally
- The story is totally different for the last task

Cross validation for comparing statistical procedures

CV is widely used in statistical applications.

Allen (1974), Stone (1974), Geisser (1975)

Different versions:

- delete-one
- delete-n_2
- k-fold

CV Paradox

We compare two different uses of Fisher's LDA method.

- $n = 100$
- For 40 observations with $Y = 1$, we generate three independent random variables X_1, X_2, X_3, all standard-normally distributed
- For the remaining 60 observations with $Y = 0$, we generate the three predictors with $N(0.4, 1)$, $N(0.3, 1)$ and $N(0, 1)$ distributions
- We compare LDA based on only X_1 and X_2 with LDA based on all of the three predictors.

Is MORE automatically helpful for selecting the better procedure? We evenly split the additional observations. The initial data splitting ratio is 30/70.

$n = 100$	300	500	700	900
0.835	0.825	0.803	0.768	0.772

How about maintaining the ratio of 30/70 in data splitting?

$n = 100$	300	500	700	900
0.835	0.892	0.868	0.882	0.880

How about an increasing ratio in favor of evaluation size?

Say, 70%, 75%, 80%, 85%, and 90%, respectively.

$n = 100$	300	500	700	900
0.835	0.912	0.922	0.936	0.976

When the estimation size is increased by e.g. half of the original sample size, since the estimation accuracy is improved for both of the classifiers, their difference may no longer be distinguishable with the same order of evaluation size (albeit increased).

The surprising requirement of the evaluation part in CV to be dominating in size (i.e., $n_2/n_1 \to \infty$) for differentiating nested parametric models was discovered by Shao (1993) in the context of linear regression.

What happens when comparing two general statistical procedures?

Consider the regression setting:

$$Y_i = f(X_i) + \varepsilon_i, \quad 1 \leq i \leq n,$$

- $(X_i, Y_i)_{i=1}^n$ independent observations with X_i iid (d-dimensional)
- f is the regression function
- ε_i are the random errors with $E(\varepsilon_i | X_i) = 0$ and $E(\varepsilon_i^2 | X_i)$ uniformly bounded almost surely

Two candidate regression procedures, δ_1 and δ_2. Based on a sample $(X_i, Y_i)_{i=1}^n$, they yield estimators $\widehat{f}_{n,1}(x)$ and $\widehat{f}_{n,2}(x)$ respectively.

Delete-n_2 CV:

- the estimation data $Z^1 = (X_i, Y_i)_{i=1}^{n_1}$
- the validation data $Z^2 = (X_i, Y_i)_{i=n_1+1}^{n}$. Let $n_2 = n - n_1$.
- Apply δ_1 and δ_2 on Z^1 to obtain the estimators $\widehat{f}_{n_1,1}(x)$ and $\widehat{f}_{n_1,2}(x)$ respectively.
- Compute the prediction squared errors of the two estimators on Z^2 :

$$\begin{aligned} & CV(\widehat{f}_{n_1,j}) \\ = & \sum_{i=n_1+1}^{n} \left(Y_i - \widehat{f}_{n_1,j}(X_i)\right)^2, \quad j = 1, 2. \end{aligned}$$

- If $CV(\widehat{f}_{n_1,1}) \leq CV(\widehat{f}_{n_1,2})$, δ_1 is selected and otherwise δ_2 is chosen.

Definition 1. δ_1 is asymptotically better than δ_2 if for each $0 < \epsilon < 1$, there exists a constant $c_\epsilon > 0$ such that when n is large enough,

$$P\left(\frac{\left\|f - \widehat{f}_{n,2}\right\|_2}{\left\|f - \widehat{f}_{n,1}\right\|_2} \geq (1 + c_\epsilon)\right) \geq 1 - \epsilon.$$

Definition 2. Assume that one of the candidate regression procedures, say δ^*, is asymptotically better. A selection rule is said to be consistent if the probability of selecting δ^* approaches 1 as $n \to \infty$.

Let $\{a_n\}$ be a sequence of positive numbers approaching zero.

Definition 3. A procedure δ is said to converge exactly at rate $\{a_n\}$ in probability if

$$\left\| f - \widehat{f}_n \right\|_2 = O_p(a_n),$$

and for each $0 < \epsilon < 1$, there exists $c_\epsilon > 0$ such that when n is large enough,

$$P\left(\left\| f - \widehat{f}_n \right\|_2 \geq c_\epsilon a_n \right) \geq 1 - \epsilon.$$

Condition 1. For $j = 1, 2$,

$$\left\| f - \widehat{f}_{n,j} \right\|_{\infty} = O_p(1).$$

Condition 2. Under the L_2 loss, either δ_1 is asymptotically better than δ_2, or δ_2 is asymptotically better than δ_1.

Consistency of CV

Let I^* be the better procedure. Let $\widehat{I}_n$ be the selected model. Suppose that $\widehat{f}_{n,1}$ and $\widehat{f}_{n,2}$ converge exactly at rates p_n and q_n respectively.

Theorem. (Yang, 2007). Under the earlier conditions, if the data splitting satisfies

1. $n_2 \to \infty$ and $n_1 \to \infty$;
2. $\sqrt{n_2}\max(p_{n_1}, q_{n_1}) \to \infty$,

then the delete-n_2 CV is consistent, i.e.,

$$P\left(\widehat{I}_n \neq I^*\right) \to 0 \text{ as } n \to \infty.$$

Implications: the delete-n_2 CV is consistent:

- $\max(p_n, q_n) = O(n^{-1/2})$, with the choice $n_1 \to \infty$ and $n_2/n_1 \to \infty$;
- $\max(p_n, q_n)n^{1/2} \to \infty$, with any choice such that $n_1 \to \infty$ and $n_1/n_2 = O(1)$.

Shao (1993) derived consistency of CV for linear models, and showed the surprising requirement of $n_2/n_1 \to \infty$.

The story can be very different for comparing two general estimators. The proportion of the evaluation part can even be of a smaller order.

In summary,

- Data splitting ratio is critical for cross validation to be consistent for selecting the better procedure
- Unlike parametric case, the evaluation size of CV does not have to be dominatingly large for comparing two general procedures
- Reliability of procedure comparison based on delete-one CV is questionable

Model selection diagnostics

It is difficult to choose between model selection criteria and choose between model selection and model combining. Can we construct model selection diagnostic measures that provide insight and guidance?

- Can the selected model be reasonably declared the "true" model?
- Should I use model selection or model averaging?
- Does the model selection uncertainty matter for my target of estimation?
- ...

Model selection uncertainty measures:

- bootstrap instability
- perturbation instability
- sequential instability
- ...

When should we choose model combining over model selection?

- When combining the estimates can significantly reduce bias of a small number of candidates, we should combine. When the number of candidates is large, it depends (see, e.g., Nemirovskii 2000; Yang 2001 and 2004; Catoni 2004; Tsybakov 2003).

- When there is no potential to reduce modeling bias by combining the candidates, it is not always better to do model averaging.

Instability in Model Selection

Breiman (1996) pointed out that model selection is unstable. He proposed *bagging* and other methods to stabilize an unstable procedure.

Uncertainty due to model selection has been basically ignored in most statistical applications.

Model selection instability plays an important role in choosing between model selection and model combining.

Perturbation instability in Model Selection

Consider regression models

$$Y_i = f_k(x_i, \theta_k) + \varepsilon_i, \quad i = 1, 2, ..., n; k = 1, 2, ...$$

and a model selection criterion.

- Generate new random errors W_i iid from $N(0, \theta^2 \hat{\sigma}^2)$, where θ indicates the perturbation size.
- Define $\widetilde{Y}_i = Y_i + W_i$ for $1 \leq i \leq n$.
- Apply the model selection criterion to the perturbed data set $(\widetilde{Y}_i, X_i)$, $1 \leq i \leq n$.
- Measure the change by the average squared difference between the original estimates and the new ones.

- At each θ, replicate the process and average the changes.
- Plot the average change versus perturbation size θ. The slope of the plot at zero is called the perturbation instability in estimation (PIE).

Which factors may affect instability?

- # of candidate predictors
- # of predictors in the true model
- sample size
- error variance

Simulations:

- $n = 100$ unless stated otherwise
- 10 independent candidate predictors $X_i \sim Unif(-1, 1)$.
- We report PIE for each case based on 50 replications.

The effect of sample size

- The true regression function:

  $1.0 + 1.0X_1 + 1.0X_2 + 1.0X_3 + 1.0X_4 + 1.0X_5$

- $\sigma^2 = 2$.

- A. $n = 100$: $PIE = 0.535$ (0.119).

- B. $n = 30$: $PIE = 0.756$ (0.237).

The effect of error variance σ^2

Case 1: The true regression function is

$$0.9 + 1.5X_1 + 1.6X_2 + 1.7X_3 + 1.5X_4 + 0.4X_5 + 0.3X_6 + 0.2X_7 + 0.1X_8$$

Case 2: The true regression function is

$$1 + X_1 + X_2 + X_3 + X_4 + X_5$$

	$\sigma^2 = 0.01$	0.1	1.0	2.25
Case 1	0.0322 (0.0035)	0.117 (0.023)	0.499 (0.100)	0.747 (0.223)
Case 2	0.0293 (0.0050)	0.0843 (0.0139)	0.309 (0.071)	0.535 (0.119)

A Real data example

Crime data: 15 candidate predictors and 47 observations.

$$PIE = 0.819 \text{ for BIC.}$$

Combining models reduces the instability

We use ARM (Yang, 2001) and BMA (Hoeting, et al, 1999) as model combining methods.

ARM	BMA	BIC	AIC
0.518	0.537	0.819	0.784

A data example

A 2^3 experiment with 2 replicates (Garcia-Diaz and Phillips (1995))

Parametric bootstrap Instability and Perturbation Instability in selection:

	PBI	PI
AIC	0.59	1.12
BIC	0.58	1.21

Average Squared Prediction Error:

AIC	40.0	(1.3)
BIC	41.5	(1.3)
ARM	32.5	(1.3)

Two statements

- Statisticians are good examples of people who, in their own research, do not practice what they teach others to do.
- When "promoting" one's own methods, the author should bear the burden of letting the reader know when their method does not work, especially via empirical investigations.

Concluding remarks

- Although methods based on correct parametric models work well asymptotically, for a moderate or small sample size, their performances may not be good. For example, insufficient extreme observations typically impair accuracy of LQR at high/low quantiles.
- It is desirable to consider multiple procedures
  - choosing a model/procedure from a list is challenging (especially for quantile regression)
  - finding the true model, assumed to be among the candidates, may not be the right target anyway
  - thus for purposes of reducing model selection uncertainty and improving the true-model-based estimators, model/procedures combination is important

- Difference between consistency in selecting the true model and consistency in selecting the best procedure
- Delete-one CV may not be reliable for comparing learning procedures
- Model selection diagnostics can be very useful:
  - to choose between model selection and model combining
  - to assess reliability of e.g., an identified sparse structure for a high-dimensional problem