On Improving Traditional Model Selection Methods

Yuhong Yang
School of Statistics
University of Minnesota

University of Florida

January 12, 2008

Outline

- Some challenges to traditional model selection methods
- Adaptive model selection: how far can we go?
- Localized model selection
- Is sparse linear model combining a powerful class of regression methods?
- Model combinations: gains and costs
- Conclusion and discussion

Some challenges to traditional model selection methods

- Model selection uncertainty can often be very high, which hurts both reliability of interpretation and estimation/prediction accuracy
- War between different camps of model selection principles/methods
- Global comparison may not be telling the whole truth
- ...

Rivalry Between AIC and BIC

- When the true model is among the candidates,
  - BIC is consistent in terms of selecting the true model (Nishii (1984), Haughton (1999))
  - BIC is asymptotically efficient (e.g., Shao (1997))
- For a nonparametric setting, AIC is asymptotically efficient (Shibata (1981), Li (1987), Polyak and Tsybakov (1990))
- AIC is minimax-rate optimal for both parametric and nonparametric situations
- A commonly told story is: BIC should be used for a parametric case and AIC should be used for a nonparametric case.

Let δ be a model selection rule to choose between $\widehat{f}_{n,1}(x)$ and $\widehat{f}_{n,2}(x)$. Let A_δ be the event that model 2 is selected.

The risk of the estimator based on δ at a given x_0 is

$$E\left(f(x_0) - \left(\widehat{f}_{n,1}(x_0)I_{A_\delta^c} + \widehat{f}_{n,2}(x_0)I_{A_\delta}\right)\right)^2$$

Is the story about AIC and BIC accurate?

A simple demonstration

Consider

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, ..., n,$$

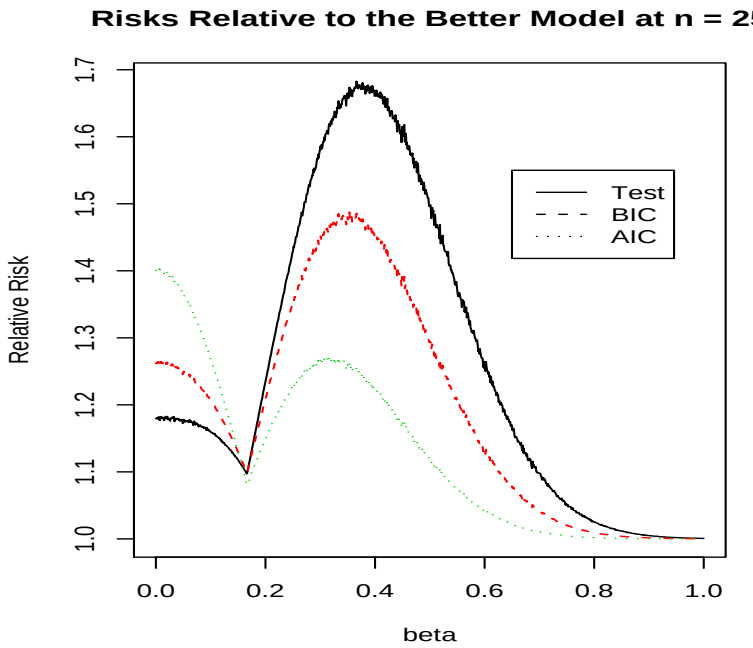
- $x \in [-1, 1]$ is the design variable with $\overline{x}_n = 0$
- $\{\varepsilon_i\}$ are Gaussian errors

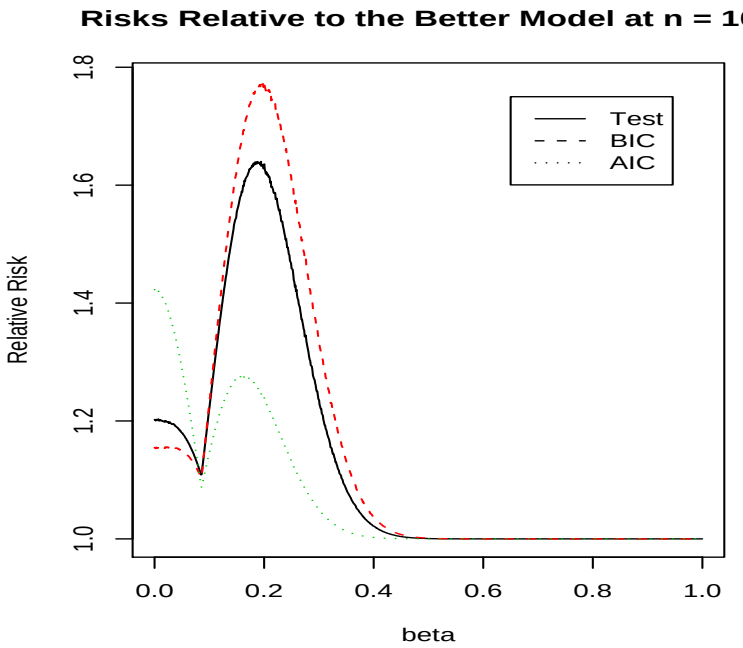
Our interest: point prediction of Y at a new value x_0 under the squared error loss

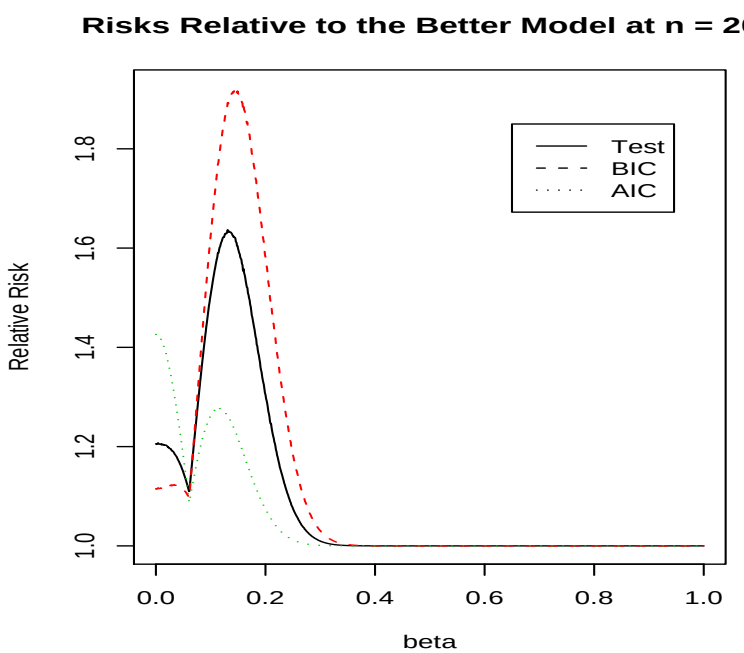
Model 0: $Y_i = \alpha + \varepsilon_i$

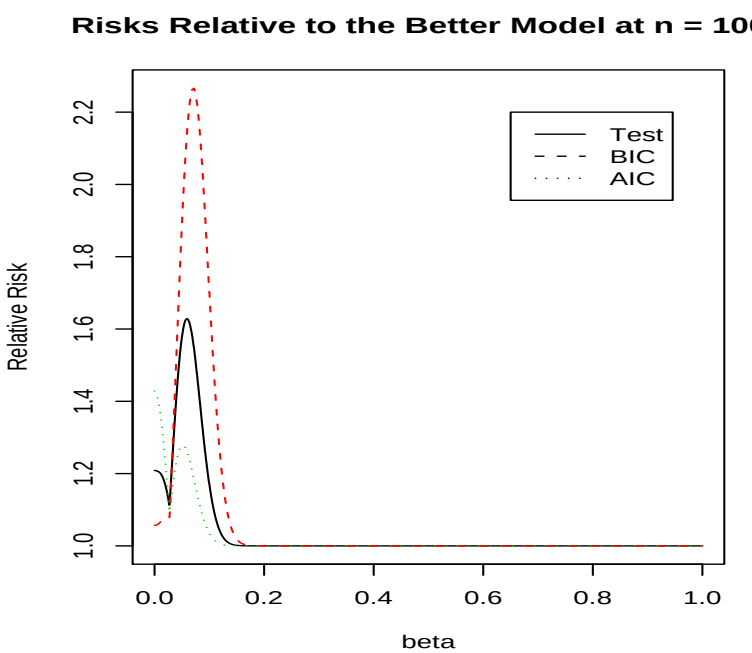
Model 1: $Y_i = \alpha + \beta x_i + \varepsilon_i$

- $n = 25, 100, 200, 1000$
- $x_0 = 0.5$
- $\sigma = 0.5$


Risks Relative to the Better Model at n = 25
Relative Risk
beta
Test
BIC
AIC
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
0.0
0.2
0.4
0.6
0.8
1.0


Risks Relative to the Better Model at n = 100
Relative Risk
beta
Test
BIC
AIC
1.0
1.2
1.4
1.6
1.8
0.0
0.2
0.4
0.6
0.8
1.0


Risks Relative to the Better Model at n = 200
Relative Risk
beta
Test
BIC
AIC
1.0
1.2
1.4
1.6
1.8
0.0
0.2
0.4
0.6
0.8
1.0


Risks Relative to the Better Model at n = 1000
Relative Risk
beta
Test
BIC
AIC
1.0
1.2
1.4
1.6
1.8
2.0
2.2
0.0
0.2
0.4
0.6
0.8
1.0

- The nice property of BIC being asymptotically efficient in this setting is not quite in line with the simulation results (cf. Foster and George (1994)).

- Pointwise convergence may not be reliable.

Moving Beyond the Debate Between AIC and BIC

Barron, Yang, Yu (1995): adaptive MDL

Hansen and Yu (1997): adaptive MDL

Tibshirani and Rao (1997): adaptive penalty based on CV

Foster and George (2000): empirical Bayes

Shen and Ye (2002): adaptive penalty in terms of generalized degree of freedom

...

What properties of AIC and BIC can be combined?

- Can an adaptive model selection criterion be asymptotically efficient for both parametric and nonparametric cases?
- Can consistency of BIC and minimax-rate optimality be achieved at the same time?
- None of the above can be obtained with a criterion of the form

$$-\text{log-likelihood} + \lambda_n \times \text{model dimension}$$

Let $\{\varphi_0 = 1, \varphi_1, ...\}$ be orthonormal trigonometric basis on $[0, 1]$ and assume that X_1 has a uniform distribution.

Consider the series expansion models:

$$Y_i = \alpha_0 + \alpha_1\varphi_1(X_i) + ... + \alpha_m\varphi_m(X_i) + \varepsilon_i,$$

for $m \geq 1$.

The projection estimator for model m

$$\widehat{f}_m(x) = \sum_{i=0}^{m} \widehat{\alpha}_i\varphi_i(x),$$

where $\widehat{\alpha}_i = \frac{1}{n}\sum_{j=1}^{n} Y_j\varphi_i(X_j)$.

Suppose $f(x) = \sum_{i \geq 0} \alpha_i \varphi_i(x)$, is bounded, differentiable, and satisfying

$$\| \sum_{i \geq k+1} \alpha_i \varphi_i \|_4 = O\left(\| \sum_{i \geq k+1} \alpha_i \varphi_i \|_2 \right)$$

Which m to use?

Selection between AIC and BIC

Let $\widehat{m}_{n,AIC}$ and $\widehat{m}_{n,BIC}$ be the models selected by AIC and BIC. Define

$$\widehat{f}_{n,BIC}(x) = \sum_{i=0}^{\widehat{m}_{n,BIC}} \widehat{\alpha}_i \varphi_i(x)$$

and define $\widehat{f}_{n,AIC}(x)$ to be

$$\begin{cases} \sum_{i=0}^{\widehat{m}_{n,AIC}} \widehat{\alpha}_i \varphi_i(x) & \text{if } \widehat{m}_{n,AIC} \neq \widehat{m}_{n,BIC} \\ \sum_{i=0}^{\widehat{m}_{n,AIC}+1} \widehat{\alpha}_i \varphi_i(x) & \text{otherwise.} \end{cases}$$

Cross validation (CV) is to be used for selection between AIC and BIC. Let

$$\widehat{\widehat{f}}(x) = \begin{cases} \widehat{f}_{n,BIC} & \text{if BIC is selected} \\ \widehat{f}_{n,AIC} & \text{if AIC is selected.} \end{cases}$$

Assumption 1: When f is not in the candidate models, we suppose

1. AIC is asymptotically efficient

$$\frac{\left\| f - \widehat{f}_{n,AIC} \right\|_2}{\inf_m \left\| f - \widehat{f}_m \right\|_2} \to 1 \text{ in prob.}$$

2. BIC is suboptimal in the sense that there exists a constant $c > 1$ such that with probability going to 1,

$$\frac{\left\| f - \widehat{f}_{n,BIC} \right\|_2}{\left\| f - \widehat{f}_{n,AIC} \right\|_2} \geq c$$

Theorem 1. Consider the delete-n_2 CV with $n_1 \to \infty$ and $n_1 = o(n_2)$. The CV method is consistent for selection between AIC and BIC. Consequently, we have

$$\frac{\left\| f - \widehat{\widehat{f}} \right\|_2^2}{\inf_m \left\| f - \widehat{f}_m \right\|_2^2} \to 1 \text{ in prob.}$$

Thus the asymptotic efficiency of AIC and BIC for exclusive situations can be integrated by adaptive model selection. Can we go further?

Recall:

- a key property of BIC is consistency in selection
- a key property of AIC is minimax-rate optimality for estimating the regression function for both parametric and nonparametric situations

Can we have these hallmark properties combined?

Theorem 2. 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.

- **An implication**: model identification and optimal rate estimation are not totally compatible

- From the results, the model selection criterion of the form

$$-\text{log-likelihood} + \lambda_n \times \text{model dimension}$$

CAN do as well as any other model selection method in terms of false selection probability and uniform rate of convergence, but *CANNOT* compete with adaptive model selection in terms of asymptotic efficiency for both parametric and nonparametric cases.

Localized cross validation to improve over global model selection

CV is widely used in statistical applications.

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

Different versions:

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

Properties of CV were investigated in:

Li (1987), Shao (1993), Zhang (1993); Wong (1983), Speckman (1985), Burman (1990), Härdle, Hall and Marron (1988), Hall and Johnstone (1992) and more

Different goals for using CV

- prediction error estimation
- model selection as the end product
- model selection as an intermediate step

CV Paradox

- Suppose a statistician's original data splitting scheme works for consistency in selection.
- The same amount of (or more) independent and identically distributed data is given to the statistician.
- He decides to add half of the new data to the estimation part and the remaining half to the evaluation part.
- He naturally thinks that with improvement in both the training and evaluation components, the comparison of the candidate procedures becomes more reliable.

Is that the case?

A simulation

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 first noted by Shao (1993) in the context of linear regression.

What happens when comparing two general statistical procedures?

Motivation for considering localized selection

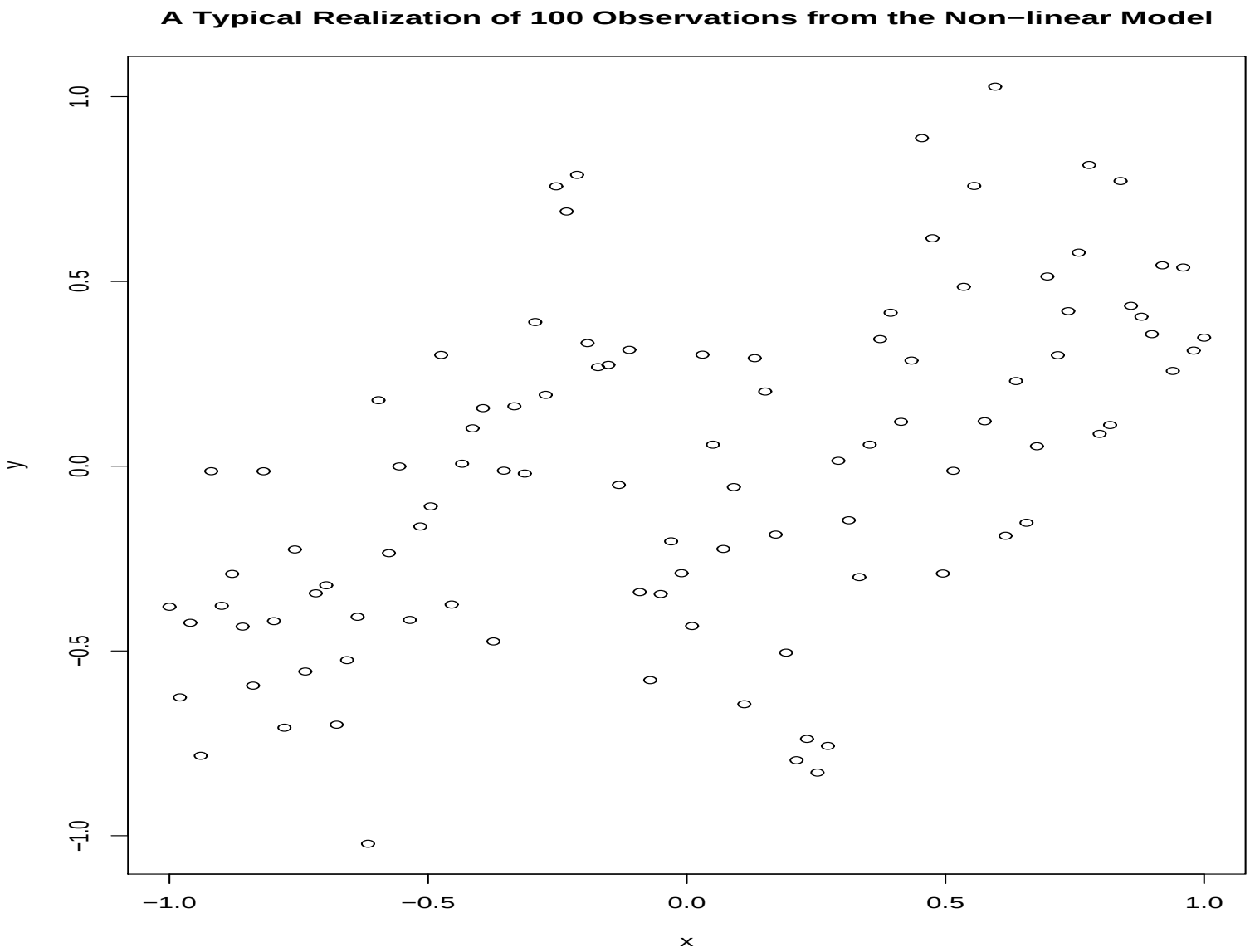
The relative performance of candidate procedures depends on x.

A numerical demonstration

Consider estimating f on $[-1, 1]$ with $n = 100$ and $\sigma = 0.3$. The true f is

$$\begin{aligned} 0.5x &+ 0.1\exp(-200(x+0.25)^2) \\ &- 0.1\exp(-200(x-0.25)^2). \end{aligned}$$

A typical realization of the data is:


A Typical Realization of 100 Observations from the Non−linear Model
y
x
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

Consider a simple linear model and also smoothing spline.

Global selection versus local selection

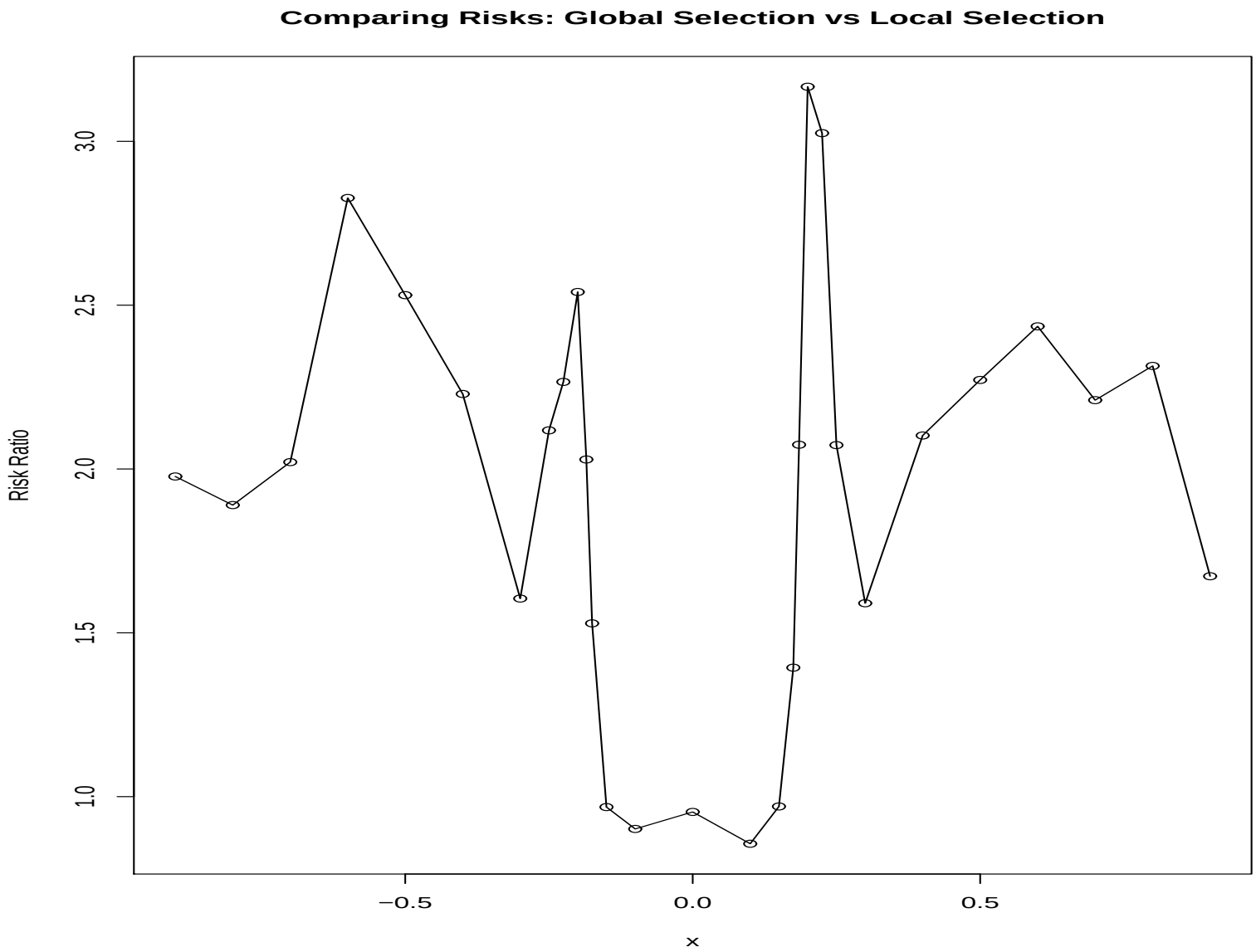
- Randomly split the data into two parts, find the linear estimate $\widehat{f}_L$ and the SS estimate $\widehat{f}_{SS}$ using the first part and compute the prediction error on the second part. Repeat 50 times to choose the linear estimate or the SS estimate (based on the full data). Let $\widehat{f}^G(x)$ be the resulting estimator.

- Consider estimators of the form

$$\widehat{f}(x;c) = \widehat{f}_L(x) I_{\{|x| \geq c\}} + \widehat{f}_{SS}(x) I_{\{|x| < c\}}.$$

  Use CV similarly to choose c in the range of $[0, 1]$ at a grid of width 0.01. Let $\widehat{f}^{NG}(x)$ be the resulting estimator.

At a given x_0, compute the risks of $\widehat{f}^G(x_0)$ and $\widehat{f}^{NG}(x_0)$ based on 200 replications. The risk ratio is:


Comparing Risks: Global Selection vs Local Selection
Risk Ratio
1.0
1.5
2.0
2.5
3.0
−0.5
0.0
0.5
x

Approaches in non-global selection

Consider two procedures δ_1 and δ_2 with risks $R(\delta_1; x; n)$ and $R(\delta_2; x; n)$ at a given x.

Let $A^* = \{x : R(\delta_1; x; n) \leq R(\delta_2; x; n)\}$. Ideally, one would use δ_1 on A^* and δ_2 on $(A^*)^c$. In reality, one can consider various sets of A of different degrees of locality and try to find the best one. Two examples:

1. At each given x_0, one considers a local neighborhood around x_0 and tries to find the candidate that performs better in the local area.

2. One considers a collection of sets of a certain mathematical form and tries to identify the one with best performance. Here the collection may depend on the sample size. The size of the collection can be pre-determined or adaptively chosen.

Implementation

When one has little idea on the form of A^*, as is often true when the input dimension is high, one can take advantage of classification methods.

- Splits the data into two parts. The first part is used to obtain the estimates from the candidate procedures, and then make predictions for the second part of the data.

- Based on the relative predictive performance in each case, we create a new variable that simply indicates which estimate is the better one.

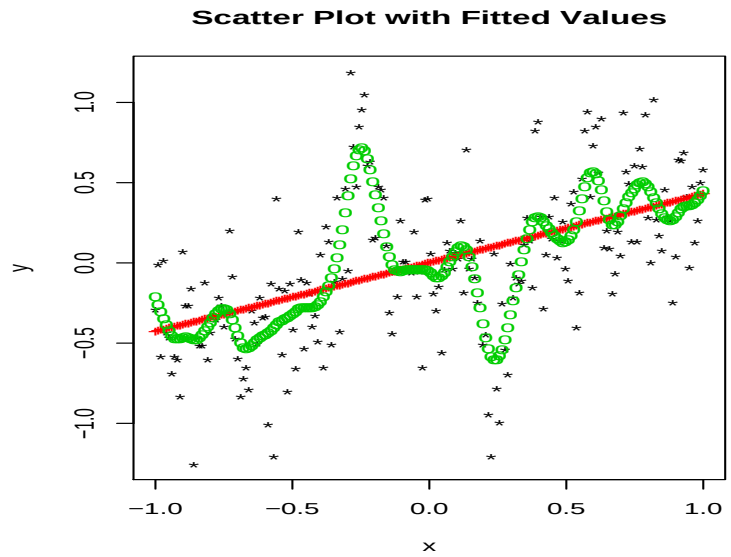
- Apply any sensible classification method to relate the performance indication variable to the covariates.

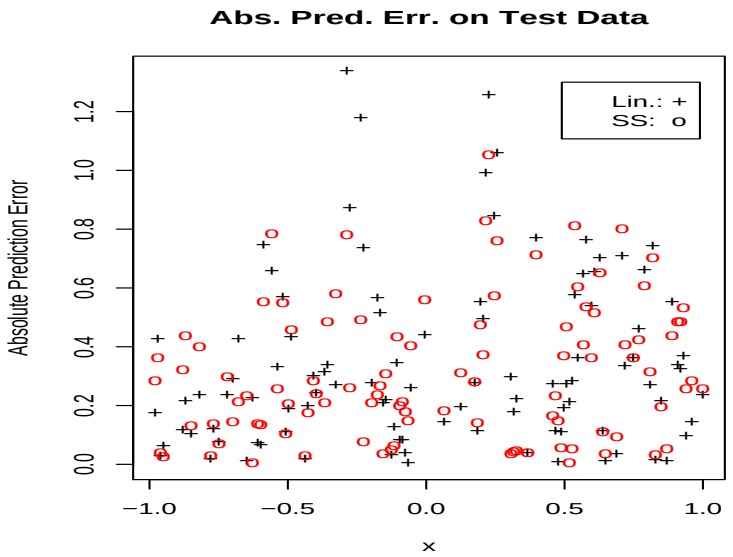
Example (continued). We focus on one realization of the data with $n = 200$. We fit a logistic regression model with three terms: 1, x and x^2 to find the region where the linear estimator performs better than the smoothing spline estimator. The estimated probability that the linear model performs better at x is

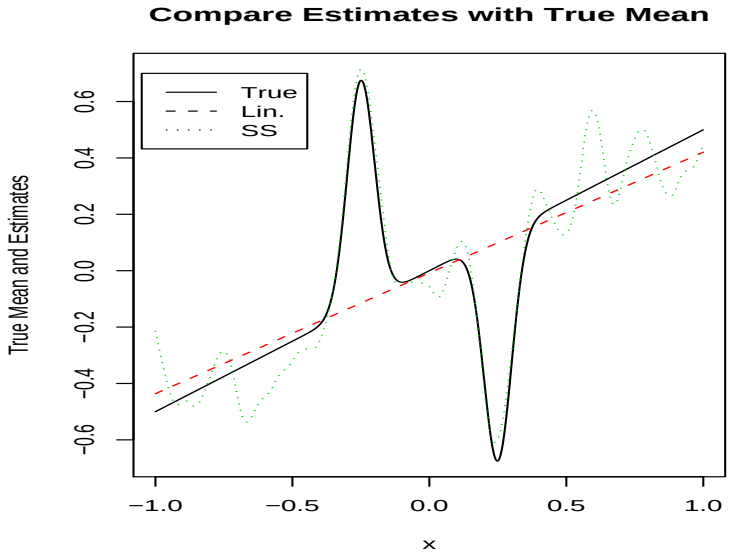
$$\widehat{p}(x) = \frac{1}{1 + \exp\left(0.379 - 0.025x - 1.497x^2\right)}.$$

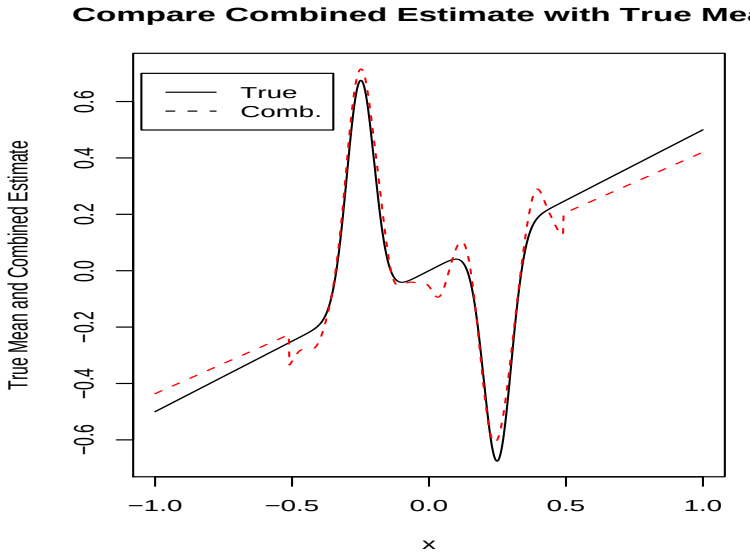
Note that $\widehat{p}(x) > 0.5$ corresponds to $x < (-0.51)$ or $x > 0.49$, which is very sensible from our knowledge of the true mean function. This suggests the following combined estimate of the regression function

$$\widehat{f}(x) = \begin{cases} \hat{f}_L(x) & \text{if } x < (-0.51) \text{ or } x > 0.49 \\ \hat{f}_{SS}(x) & \text{otherwise.} \end{cases}$$


Scatter Plot with Fitted Values
y
x
−1.0
−0.5
0.0
0.5
1.0


Abs. Pred. Err. on Test Data
Absolute Prediction Error
Lin.: +
SS: o
x
0.0
0.2
0.4
0.6
0.8
1.0
1.2


Compare Estimates with True Mean
True Mean and Estimates
True
Lin.
SS
x
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


Compare Combined Estimate with True Mean
True Mean and Combined Estimate
True
Comb.
x
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

Localized CV selection (L-CV)

For a given x_0, consider the ball centered at x_0 with radius r_n for some $r_n > 0$.

- We randomly split the data into a training set of size n_1 and a test set of size n_2.
- Fit the regression procedures on the first part of data.
- For evaluation, consider only the data points in the test set that are in the given neighborhood of x_0. Let $\hat{\jmath}(x_0) = \hat{\jmath}_n(x_0)$ be the procedure that has the smaller average squared prediction error.
- This process is repeated with a number of random splittings of the observations to avoid the splitting bias. The procedure that wins more frequently among the permutations is the final winner.

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

Question: Under what conditions, the above L-CV is consistent in selection?

Assume that $\widehat{f}_{1,n}$ and $\widehat{f}_{2,n}$ converge exactly at rate $\{p_n\}$ and $\{q_n\}$ in probability at η_n-neighborhood of x_0 respectively.

Theorem 3: Under some conditions, as long as n_1, n_2, and r_n are chosen to satisfy

$$\sqrt{n_2 r_n^d}\max(p_{n_1}, q_{n_1}) \to \infty,$$

we have that with probability going to 1, the better procedure $\widehat{f}_{j^*(x_0),n}$ will be selected.

Implications: the delete-n_2 CV is consistent:

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

Thus compared to Shao (1993), the story can be very different for comparing two general estimators (locally). The proportion of the evaluation part can even be of a smaller order.

Caution on empirical comparisons of different methods.

Is sparse linear model combination a powerful regression procedure?

Problem setup of regression learning

- **Data:** Consider the regression setting

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

  - $(X_i, Y_i)_{i=1}^n$, (X, Y) are i.i.d. copies
  - X (any dimension) has a distribution P_X
  - ε has a normal distribution with mean zero and variance $\sigma^2 > 0$

- Needs to estimate f

- Let δ be an estimation procedure producing $\widehat{f}_i(x)$ at each sample size $i \geq 1$.
- Let $\|\cdot\|$ denote the L_2 norm with respect to the distribution of X.
- Performance measure is

$$R(f; n; \delta) = E\|f - \widehat{f}_n\|^2.$$

- Rate of convergence of the risk depends on characteristics of f and nature of δ.
- Implications:
  - One should try to use a good characterization of the target function, especially for high-dimensional regression
  - A good learning procedure needs to be flexible
    * Adaptive estimation
    * Model selection and model combining

- Positive pictures
  - minimax-rate adaptivity
  - simultaneous variable selection and regression estimation
- Question: How many regression functions can be served well by any given regression procedure?

- Fix a regression procedure δ.
- Let b_n^2 be a non-increasing sequence with $b_n^2 \to 0$ as $n \to \infty$.
- Assume that the true regression function has the L_2 norm bounded by a known constant $A > 0$.

- Consider the class of regression function $\mathcal{F}(\{b_n^2\};\delta)$:

$$\{f :\| f \|\leq A \text{ and } R(f;n;\delta) \leq b_n^2 \text{ for all } n \geq 1\}.$$

  This, called maxiset, is the collection of the regression functions for which the estimation procedure δ achieves the given accuracy b_n^2 at each sample size n.

- How large can $\mathcal{F}(\{b_n^2\};\delta)$ be? For certain specific procedures (such as wavelet shrinkage), Kerkyacharian and Picard (2002), Autin, Picard and Rivoirard (2006) successfully characterized the set $\mathcal{F}(\{b_n^2\};\delta)$. But we are interested in any regression procedure.

How to measure largeness?

- Metric entropy is a proper quantity to measure largeness of a set in a metric space.
- For a given class of regression functions $\mathcal{F}$, let $M(\epsilon; \mathcal{F})$ be the logarithm of maximum number of points that are more than ϵ apart under $\|\cdot\|$.
- That how fast $M(\epsilon; \mathcal{F})$ approaches infinity as $\epsilon \to 0$ captures the massiveness of $\mathcal{F}$.
- Let $b_0^2 = A^2 + 2\log 2$ and define $B_k = \sum_{i=0}^{k} b_i^2$ for $k \geq 1$ and $B_0 = b_0^2$.

Theorem 3

- *Take* $b_n^2 = Cn^{-\gamma}$ *for some constant* $C > 0$ *and* $0 < \gamma \leq 1$. *When* $\gamma < 1$, *for every regression procedure* δ, *for* $\epsilon \leq 3C^{1/2}$, *we must have*

$$M\left(\epsilon; \mathcal{F}(\{b_n^2\}; \delta)\right) \leq C^{'} \left(\frac{1}{\epsilon}\right)^{\frac{2(1-\gamma)}{\gamma}},$$

where $C^{'}$ *is a constant depending only on* γ, C *and* A.

- *When* $\gamma = 1$, *for every regression procedure* δ, *for* $\epsilon \leq 3C^{1/2}$, *we have*

$$M\left(\epsilon; \mathcal{F}(\{b_n^2\}; \delta)\right) \leq C^{''} \log\left(\frac{1}{\epsilon}\right)$$

for some constant $C^{''}$ *depending only on* A and C.

- *For a general sequence* $\{b_n^2\}$, *we have*

$$M\left(3b_k; \mathcal{F}(\{b_n^2\}; \delta)\right) \leq \lceil B_{k-1} \rceil \text{ for all } k \geq 1.$$

Tightness of the bounds

- For smoothness function classes, the minimax rate of convergence is usually determined by a smoothness parameter α (e.g., the number of derivatives that f has) with the rate $n^{-2\alpha/(2\alpha+d)}$, where d is the dimension of f. The metric entropy order of such a class is typically $(1/\epsilon)^{d/\alpha}$ as $\epsilon \to 0$.

- For $\gamma = 2\alpha/(2\alpha+d)$, $2(1-\gamma)/\gamma = d/\alpha$ and accordingly the entropy upper bound given above for $\mathcal{F}(\{b_n^2\};\delta)$ is of order $(1/\epsilon)^{d/\alpha}$. This order matches the metric entropy of smoothness classes with convergence rates $n^{-2\alpha/(2\alpha+d)}$.

- Thus in terms of order, the upper bounds in the theorem can not be generally improved.

Sparse estimation and model combination

- For each $k \geq 1$, let $\Phi_k = \{\varphi_{k,1}, ..., \varphi_{k,L_k}\}$ be a collection of L_k linearly independent functions.

- Given k, $1 \leq m \leq L_K$ and $I = I_{k,m} = \{i_1, ..., i_m\}$ as a subset of $\{1, 2, ..., L_k\}$ with m terms in Φ_k, consider approximation of a function f by linear combinations

$$\sum_{l=1}^{m} \theta_l \varphi_{k,i_l}, \ (\theta_1, ..., \theta_m) \in R^m.$$

- When m is small compared to L_k, the terms used in the linear combination is a sparse subset of Φ_k. Such *sparse approximation* is very useful to improve estimation accuracy.

- For each choice of (k, m, I), one fits the model

$$Y_i = \sum_{l=1}^{m} \theta_l \varphi_{k,i_l}(X_i) + \varepsilon_i, \quad 1 \le i \le n.$$

- Since one does not know which subset provides a good approximation, one may select a model according to a certain appropriate criterion or do a proper model combination.

We assume σ^2 is upper bounded by a known constant $\overline{\sigma}^2 < \infty$.

Theorem 4

For any given regression procedure δ, *there exists a procedure* $\widetilde{\delta}$ *based on sparse approximation such that*:

- *for every regression function* f *with* $\| f \|_\infty < \infty$, *if* $R(f;\delta;n) \leq Cn^{-\gamma}$ *for all* n *for some constant* $C > 0$ *and* $0 < \gamma < 1$, *then* $R(f;\widetilde{\delta};n) \leq \widetilde{C}n^{-\gamma}$ *holds for all* n *for some constant* $\widetilde{C} > 0$;

- *if* $R(f;\delta;n) \leq Cn^{-1}$ *for all* n *for some constant* $C > 0$, *then* $R(f;\widetilde{\delta};n) \leq \widetilde{C}n^{-1}\log n$ *holds for some constant* $\widetilde{C} > 0$.

Discussion

- The theorem says that as far as polynomial rates of convergence are concerned, under the squared L_2 loss, estimation based on a certain sparse approximation can do as well as any given regression procedure (but loosing a logarithmic factor for the parametric rate of convergence).

- Is it possible to give a more formal complete class theorem for nonparametric estimation?

Conclusion and discussion

- Adaptive model selection can lead to a step forward, but there is a fundamental limitation
- Localized selection can much improve estimation accuracy
- In a certain sense, sparse linear model combination can be as powerful as any regression procedure
- Much more remains to be done...

Combining forecasting procedures

Problem of Interest:

Forecasting a real-valued continuous random quantity Y

Data Available:

$Y_1, ..., Y_{n-1}$ (previous realizations of Y)

$X_1, ..., X_{n-1}, X_n$ (outside information)

Mean squared error decomposition:

$$E\left(Y_n - \widehat{y}_n\right)^2 = E\left(m_n - \widehat{y}_n\right)^2 + Ev_n,$$

m_n : the conditional mean of Y_n

v_n : the conditional variance of Y_n

Two directions for combining forecasts

Suppose there are M forecasters

#1: $\widehat{y}_{1,1}, \widehat{y}_{1,2}, ..., \widehat{y}_{1,n}$

#2: $\widehat{y}_{2,1}, \widehat{y}_{2,2}, ..., \widehat{y}_{2,n}$

....

#M: $\widehat{y}_{M,1}, \widehat{y}_{M,2}, ..., \widehat{y}_{M,n}$

- **Combining for adaptation:** The goal is to combine the forecasts so as to perform automatically as well as the best forecaster.

  The best forecaster j^* (unknown) minimizes

$$\frac{1}{n}\sum_{i=1}^{n} E\left(m_i - \widehat{y}_{j,i}\right)^2$$

  over $j = 1, ..., M$

- **Combining for improvement:** A possible goal is to find a linear combination of the forecasts to beat the best individual forecaster.

  Want to find $\theta = (\theta_1, ..., \theta_M)$ such that

$$\frac{1}{n}\sum_{i=1}^{n} E\left(m_i - \sum_{j=1}^{M} \theta_j \widehat{y}_{j,i}\right)^2$$

  is much smaller than

$$\inf_{1\leq j\leq M} \frac{1}{n}\sum_{i=1}^{n} E\left(m_i - \widehat{y}_{j,i}\right)^2$$

Weights for predicting Y_n

Weight for forecaster # j, $1 \leq j \leq M$

$$w_{j,n} = \frac{\frac{1}{\left(\widehat{v}_{j,1} \cdots \widehat{v}_{j,n-1}\right)^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n-1}\left(Y_i - \widehat{y}_{j,i}\right)^2 / \widehat{v}_{j,i}\right)}{\sum_{l=1}^{M} \text{ the above quantity}}$$

The combined forecast is:

$$\widehat{y}_n^* = \sum_{l=1}^{M} w_{j,n}\widehat{y}_{j,n}.$$

We call the algorithm *Aggregated Forecasting Through Exponential Re-weighting (AFTER).*

A theoretical result on AFTER

Theorem 5: Under some conditions, the combined forecast satisfies

$$\frac{1}{n}\sum_{i=1}^{n} E\left(m_i - \widehat{y}_i^*\right)^2 \leq C\left\{\frac{\log M}{n} + \inf_{1\leq j\leq M}\left(\frac{1}{n}\sum_{i=1}^{n} E\left(m_i - \widehat{y}_{j,i}\right)^2 + \frac{1}{n}\sum_{i=1}^{n} E\left(v_i - \widehat{v}_{j,i}\right)^2\right)\right\}$$

Thus, the combined forecast performs as well as the best forecaster up to a constant factor, penalty $\log M/n$, and a penalty due to variance estimation. For nonparametric forecasting, the penalties are negligible.

A time series data example

Monthly sales of new one-family houses in US from Jan. 1987 to Nov. 1995 ($n = 107$)

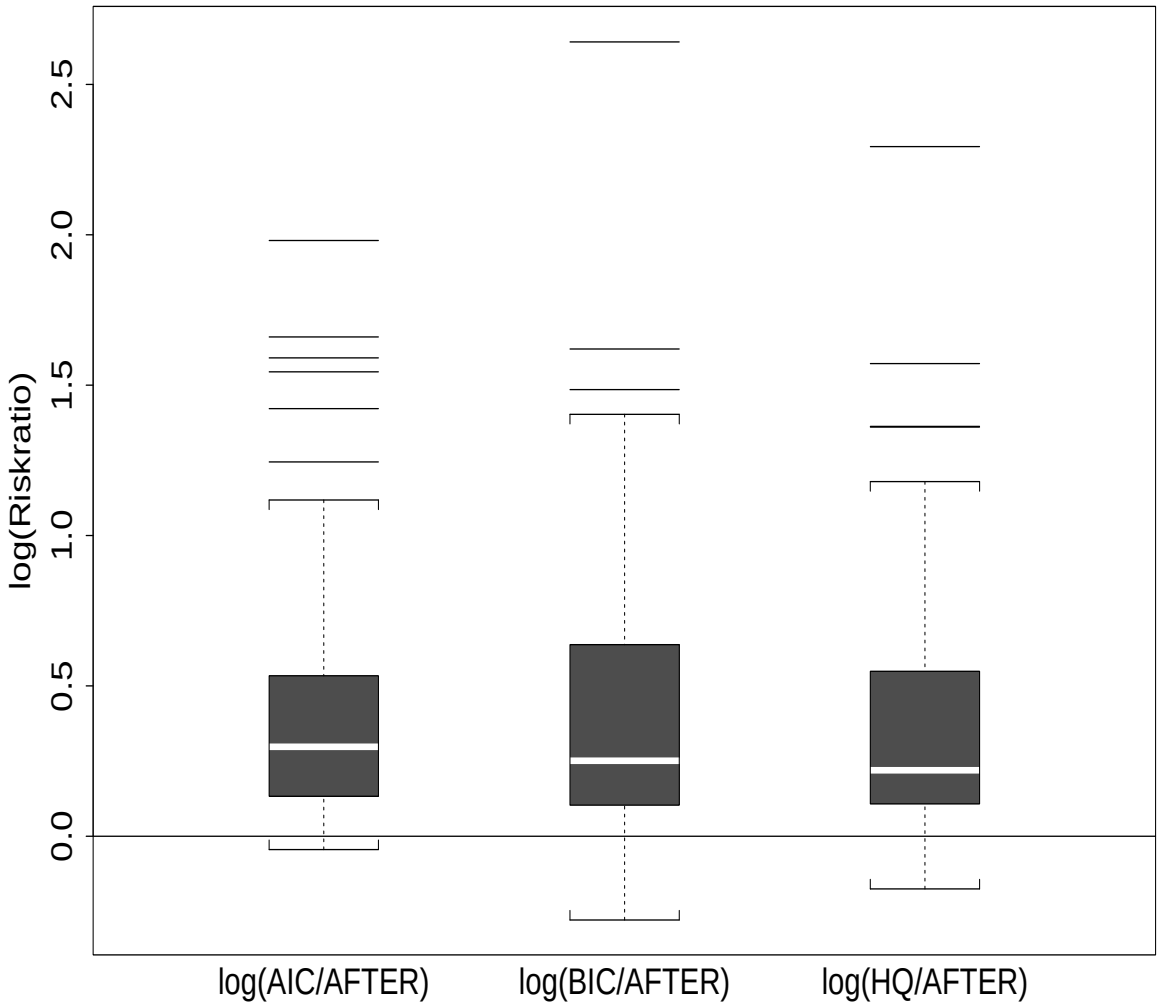
The last 20 were used for performance assessment

Consider ARIMA(p, d, q) with $p, q = 0, ..., 5$, $d = 0, 1$

ASPE of AIC, BIC, HQ, and AFTER:

	AIC	BIC	HQ	AFTER
ASPE	18.7	17.4	16.8	12.5
Reduction	33%	28%	26%	—

Random AR models: 110 random models with order uniformlly distributed between 1 and 8 and coefficients uniformly distributed between $[-10, 10]$.


log(Riskratio)
0.0
0.5
1.0
1.5
2.0
2.5
log(AIC/AFTER)
log(BIC/AFTER)
log(HQ/AFTER)

Combining for improvement: gain and price

Consider linear combinations of M forecasts:

$$\widehat{y}_n^{\theta} = \sum_{j=1}^{M} \theta_j \widehat{y}_{j,n},$$

where $\theta = (\theta_1, ..., \theta_M)$ satisfies $\sum_{j=1}^{M} |\theta_j| \leq 1$.

The corresponding average cumulative mean square error is:

$$\frac{1}{n} \sum_{i=1}^{n} E \left(m_i - \widehat{y}_i^{\theta} \right)^2 .$$

The best linear combination θ^* minimizes the above quantity over θ satisfying the constraint.

Linear combining is potentially advantageous compared to combining for adaptation if

$$\frac{1}{n}\sum_{i=1}^{n} E\left(m_i - \widehat{y}_i^{\theta^*}\right)^2 \ll \inf_{1\le j\le M} \frac{1}{n}\sum_{i=1}^{n} E\left(m_i - \widehat{y}_{j,i}\right)^2 .$$

What is the price to pay for pursuing the best linear combination?

For simplicity, consider $M = n^\tau$ for some $\tau > 0$ and the case when the conditional variances v_n are known.

Upper bound on constrained linear combining

Theorem 6: One can construct a combined forecast such that the average cumulative mean squared error is bounded above by a multiple of

$$\frac{1}{n}\sum_{i=1}^{n} E\left(m_i - \widehat{y}_i^{\theta^*}\right)^2$$
$$+ \begin{cases} \frac{\log n}{n^{1-\tau}} & \text{when } 0 \leq \tau < 1/2 \\ \left(\frac{\tau \log n}{n}\right)^{1/2} & \text{when } \tau \geq 1/2 \end{cases}$$

Similar results were obtained earlier in regression by Juditsky and Nemirovski (2000), Yang (2004), Tsybakov (2003)

Lower bounding on linear combining

Theorem 7: For every combining method, there exists a case such that the average cumulative mean squared error is lower bounded by

$$\frac{1}{n}\sum_{i=1}^{n} E\left(m_i - \widehat{y}_i^{\theta^*}\right)^2$$
$$+C\begin{cases} \frac{1}{n^{1-\tau}} & \text{when } 0 \leq \tau \leq 1/2 \\ \left(\frac{\log n}{n}\right)^{1/2} & \text{when } \tau > 1/2 \end{cases}$$

Note that the upper and lower bounds match up to at most a logarithmic factor. Thus Theorems 6 and 7 determine the price one has to pay for pursuing the best linear combination of the original forecasts.

Gain:

Reduction from $\inf_{1\le j\le M} \frac{1}{n}\sum_{i=1}^{n} E\left(m_i - \widehat{y}_{j,i}\right)^2$
to $\frac{1}{n}\sum_{i=1}^{n} E\left(m_i - \widehat{y}_i^{\theta^*}\right)^2$

Price for combining $M = n^\tau$ forecasts:

$$\begin{cases} \frac{1}{n^{1-\tau}} & \text{when } 0 \le \tau \le 1/2 \\ \left(\frac{\log n}{n}\right)^{1/2} & \text{when } \tau > 1/2 \end{cases}$$

Whether it is better to combine for adaptation or combine for improvement depends on the comparison between the *gain* and *price*.

A multi-purpose combining was carried out such that it is both *conservative* and *aggressive* whichever is better.

Bunea, Tsybakov and Wegkamp (2005) obtained aggregated estimators that are simultaneously optimal for linear aggregation and sparse aggregation.

Conclusion and discussion

- Adaptive model selection can lead to a step forward, but there is a fundamental limitation
- Localized selection can much improve estimation accuracy
- In a certain sense, sparse linear model combination can be as powerful as any regression procedure
- Combining procedures significantly improve estimation/prediction accuracy over model selection when model selection uncertainty is high
- Combining procedures can also improve over the best individual candidate in various situations (but the issue is tricky!)

Mixing on the product space

Kullback-Leibler (K-L) divergence: for densities p, q

$$D(p \parallel q) = \int p \log \frac{p}{q}$$

Some simple but key facts:

1. Under K-L, a density can be uniformly close to many densities that are far away from each other:

$$D(p \parallel \sum_j w_j q_j) \leq \inf_j \left(\log \frac{1}{w_j} + D(p \parallel q_j) \right)$$

2. Let $p^{(n)}(x_1, ..., x_n)$ and $q^{(n)}(x_1, ..., x_n)$ be two joint densities.

$$p^{(n)}(x_1, ..., x_n) = p_1(x_1) \cdot p_2(x_2|x_1) \cdots \cdot p_n(x_n|x^{n-1})$$

$$q^{(n)}(x_1, ..., x_n) = q_1(x_1) \cdot q_2(x_2|x_1) \cdots \cdot q_n(x_n|x^{n-1}).$$

Then we have

$D(p^{(n)} \parallel q^{(n)})$
$= ED(p_1 \parallel q_1) + ED(p_2 \parallel q_2) ... + ED(p_n \parallel q_n)$

3. Let $X_1, ...X_n$ be $iid \sim p(x)$. Consider an estimation procedure that produces estimators $\widehat{p}_1, \widehat{p}_2, ..., \widehat{p}_n$ based on no data, X_1, $X^2, ..., X^{n-1}$, respectively. Let

$$q^{(n)}(x_1, ..., x_n) = \widehat{p}_1(x_1)\widehat{p}_2(x_2) \cdot \cdot \cdot \cdot \cdot \widehat{p}_n(x_n).$$

Then

$$\sum_{i=1}^{n} ED(p \parallel \widehat{p}_i) = D(p^n \parallel q^{(n)}).$$