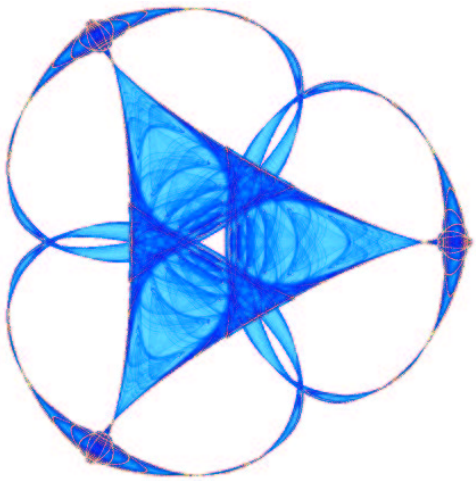
PREDICTION/ESTIMATION WITH SIMPLE LINEAR MODELS: IS IT REALLY THAT SIMPLE?

By

Yuhong Yang

IMA Preprint Series # 1974

(April 2004)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA
514 Vincent Hall
206 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612/624-6066 Fax: 612/626-7370
URL: http://www.ima.umn.edu

Prediction/Estimation with Simple Linear Models: Is it Really that Simple?

Yuhong Yang
Department of Statistics
Iowa State University
Ames, IA, 50011

April 2, 2004

Abstract

Consider the simple normal linear regression model for estimation/prediction at a new design point. When the slope parameter is not obviously nonzero, hypothesis testing and model selection methods can be used for identifying the right model. We compare performance of such methods both theoretically and empirically from different perspectives for more insight. The testing approach, in spite of being the "standard approch", performs poorly. We also found that the frequently told story "BIC is good when the true model is finite-dimensional and AIC is good when the true model is infinite-dimensional" is far from being accurate. In addition, despite some successes in the effort to go beyond the debate between AIC and BIC by adaptive model selection, it turns out that it is not possible to share the most essential properties of them by any model selection method. When model selection methods have difficulty in selection, model combining is seen to be a better alternative.

1 Introduction

Consider the simple linear regression model:

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

where x is the design variable taking values in $[-1, 1]$ and $\{\varepsilon_i\}$ are the errors assumed to be independent and normally distributed with unknown variance $\sigma^2 > 0$. Our interest is point prediction of the response variable Y at a new value $x_0 \in [-1, 1]$ of the design variable under the squared error loss.

Obviously, this is a simple statistical problem, even taught at introductory level statistical courses. Let $\hat{\alpha}$ and $\hat{\beta}$ be the least squares estimators of α and β respectively. Let $f(x) = \alpha + \beta x$ denote the true regression function. Then a predicted value of $f(x_0)$ is $\hat{f}(x_0) = \hat{\alpha} + \hat{\beta} x_0$. Under the simple linear regression model, it is the best unbiased estimator of $f(x_0)$ (under the squared error loss) and it is also a minimax estimator. Note that for a new observation Y_{n+1} at x_0, $E\left(Y_{n+1} - \hat{f}(x_0)\right)^2 = \sigma^2 + E(\hat{f}(x_0) - f(x_0))^2$. Thus, as is well known, under the squared error loss, point prediction is equivalent to point estimation of $f(x)$ at the given x value.

In many real applications, the linear relationship between x and Y is not obviously strong and it is unclear whether β is zero (we note that, of course, the issue is the same if the concern is if β is equal to another constant instead of zero). The problem seems still simple and again a solution is taught in elementary statistics courses: the familiar t-test. I imagine that this approach would be the one taken by many statisticians (if not most) in real

applications (note that in reality, of course, the normality assumption has to be assessed). Some other statisticians may prefer the use of an information criterion for assessing if $\beta = 0$.

Below we briefly review testing and model selection approaches. Without loss of generality, from now on, we assume that the design is such that $\overline{x}_n = 0$.

1.1 The testing approach

From the instruction in a typical elementary statistics textbook, when it is unclear if the slope parameter is nonzero, one should perform a hypothesis testing. A standard formulation is $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$. Then with the traditional choice of size 0.05, one conducts the well-known t test: reject H_0 if

$$\frac{|\hat{\beta}|}{\frac{s}{\sqrt{\sum_{i=1}^n x_i^2}}} \geq t_{n-2,0.025},$$

where $s^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ is an unbiased estimator of σ^2 and $t_{n-2,0.025}$ is the cutoff point of the t distribution: $P(t_{n-2} \geq t_{n-2,0.025}) = 0.025$. With the outcome of the test, we can naturally do the following prediction:

$$\hat{f}(x_0) = \begin{cases} \hat{\alpha} + \hat{\beta}x_0 & \text{when } H_0 \text{ is rejected} \\ \hat{\alpha} & \text{otherwise.} \end{cases}$$

1.2 Model selection

Alternatively from the hypothesis testing method above, one can consider using a model selection criterion. Indeed we are dealing with two models, the simple linear model in (1) and the null model: $Y_i = \alpha + \varepsilon_i, \quad i = 1, 2, ..., n,$ with the same assumptions on the errors. For convenience, call them model 1 and model 0 respectively.

Model selection based on information criteria such as AIC and BIC avoid the subjectivity of choosing the test size in the earlier hypothesis testing approach. Indeed, a valid criticism of the testing approach is that it is rather unclear how the test size influences the prediction accuracy for the problem of prediction. The strategy of controlling the probability of type I error is not intended to address the issue of prediction. Model selection criteria are motivated from different considerations, such as asymptotically maximizing the posterior model probability (BIC (Schwarz (1978)) and minimizing an estimated Kullback-Leibler divergence between the true distribution and those estimates from the models (AIC (Akaike (1973)).

Both AIC and BIC choose a model that minimizes the criterion of the form: $-\log$likelihood+penalty, where the loglikelihood is maximized within each model and the penalty is the number of parameters in the model, say, k, for AIC and is $(k \log n)/2$ for BIC. Let $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ be the MLE of σ^2 under model 1.

Simple calculations show that BIC selects model 1 when

$$\frac{|\hat{\beta}|}{\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n x_i^2}}} > \sqrt{n(n^{1/n} - 1)}.$$

Note that $\sqrt{n(n^{1/n}-1)}$ is asymptotically equivalent to $\sqrt{\log n}$. Replacing $\hat{\sigma}^2$ by the unbiased estimator s^2 (which has little effect when n is not small), we take the following slightly modified rule as BIC: select model 1 when

$$\frac{|\hat{\beta}|}{\frac{s}{\sqrt{\sum_{i=1}^n x_i^2}}} > \sqrt{\log n}.$$

Similarly, AIC selects model 1 when

$$\frac{|\hat{\beta}|}{\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n x_i^2}}} > \sqrt{n(e^{2/n}-1)}.$$

It is easy to see that $\sqrt{n(e^{2/n}-1)}$ converges to $\sqrt{2}$. Replacing $\hat{\sigma}$ by s, we take the following modified rule as AIC: select model 1 when

$$\frac{|\hat{\beta}|}{\frac{s}{\sqrt{\sum_{i=1}^n x_i^2}}} > \sqrt{2}.$$

Now AIC, BIC and the testing approach all boil down to the decision of choosing model 1 when

$$\frac{|\hat{\beta}|}{\frac{s}{\sqrt{\sum_{i=1}^n x_i^2}}} > a_n \tag{2}$$

with a_n being $\sqrt{2}$, $\sqrt{\log n}$ and $t_{n-2,0.025}$ respectively. Clearly, all of these methods are trying to assess how strong the "signal" is compared to the "noise", but they differ in the cutoff point. It is worth mentioning that $t_{n-2,0.025}$ approaches 1.96 as $n \to \infty$ and when n is large, AIC is the most aggressive in terms of choosing the larger model and BIC is the most conservative while the testing method is in between.

As is well known, the increasingly heavy penalty (or cutoff $a_n = \sqrt{\log n}$ in (2)) enables BIC to be consistent in terms of selecting the true model (i.e., model 0 if $\beta = 0$ and model 1 if $\beta \neq 0$) while AIC and testing with a fixed size cannot avoid a non-vanishing probability of selecting the wrong model when $\beta = 0$. So from this selection point of view, the penalty of AIC is too small.

Let δ_A, δ_B, δ_T denote the procedures of estimating $f(x)$ based on the outcome of AIC, BIC and testing respectively. Let $R_\beta(\delta; x_0; n)$ denote the mean squared error of the estimator $\hat{f}(x_0)$ from an estimation procedure δ at the sample size n when the true slope parameter is β, i.e.,

$$R_\beta(\delta; x_0; n) = E_\beta \left(f(x_0) - \hat{f}(x_0) \right)^2.$$

The main purpose of this paper is to examine the testing and model selection methods in terms of the above risk. As will be seen, even though the setting is very simple, complicated issues are involved. We study these issues both theoretically and via simulations.

The rest of the paper is organized as follows. In Section 2, we review the literature and give an example to illustrate some issues that we consider in this work. In Section 3, we consider some theoretical results on model

selection criteria, showing that under the squared error loss, BIC is pointwise-risk adaptive in a proper sense while AIC is not; on the other hand, BIC is rate sub-optimal from a worst-case point of view while AIC and the like are rate optimal. In addition, we theoretically address the interesting issue of whether the strengths of AIC and BIC in prediction/estimation can be combined. In Section 4, we derive an alternative criterion for choosing between model 0 and model 1 from the prediction accuracy point of view. Simulation and data examples are presented in Section 5. In Section 6, we consider the issue of what to do when model selection methods disagree. We comment on the hypothesis testing approach in Section 7. Conclusion and additional discussion are in Section 8.

2 Existing results and a motivating simulation

Given that extensive works have been done theoretically and empirically on model selection and that our problem setting is one of the simplest possible, one may expect to see a pretty clear picture here. We start with a simulation.

Consider an equally spaced design between $[-1, 1]$ with $\overline{x}_n = 0$. We are interested in the prediction/estimation accuracy at $x_0 = 0.5$ (the risks were simulated based on 2000 replications). Figure 1 gives the risk functions of δ_A, δ_B, δ_T for $\beta \geq 0$ at different sample sizes $n = 25, 100, 200$ and 1000 with $\sigma = 0.5$. Note that at $n = 25$, δ_T has a higher penalty for model 1 than the other methods but when $n \geq 100$, δ_B has the largest penalty. Typical realizations of the data at the sample size of 25 and 100 with $\beta = 0.2$ are given in Figure 2.

It is useful to note that there is no need to consider different σ values: the change of the noise level simply re-scales the risk plot. Note also that $(s^2, \hat{\alpha}, \hat{\beta})$ is a suffi cient statistic and that the design affects only the distribution of $\hat{\beta}$ (in terms of its variance $\sigma^2 / \sum x_i^2$).

Before reading Figure 1 (we intentionally put Figure 1 on a different page), let's try to predict what we would see. To that end, of course, the existing results on model selection are relevant and useful.

2.1 What do the existing results say?

We focus on the model selection criteria AIC and BIC in this brief literature review. In one sentence, a summary of the theoretical works on AIC and BIC is perhaps something like "BIC is good if the true model is fi nite-dimensional and AIC is good if the true model is infi nite-dimensional". This seems to be a frequently told story in statistics. For example, Speed and Yu (1993), Shao (1997) and Zhang (1997) quite clearly vote for BIC for estimation/prediction in a parametric setting. On the other hand, e.g., Burnham and Anderson (2002) strongly prefer AIC for its "objectivity". These views seem to be well supported by several theoretical results. To be more precise, when the true model is among the candidates (as is the case for our setting), it is well-known that BIC is consistent in terms of selection (i.e., the probability of selecting the true model approaches one as $n \to \infty$) while AIC is not (see, e.g., Nishi (1984)). Furthermore, the average squared error of the estimator from BIC is asymptotically as small as it can be from the candidate models (Shao (1997)). In addition, when cumulative prediction accuracy is the concern, it was shown that BIC achieves an accuracy lower bound while

AIC again does not (Rissanen (1986), Speed and Yu (1993)). When the true regression function is not in the candidate models, however, AIC has the so-called asymptotic efficiency property that the average squared error of the selected model is asymptotically equivalent in probability to the smallest among all the candidate models (see, e.g., Shibata (1983), Li (1987), Polyak and Tsybakov (1990) and Shao (1997)). Based on these results, Shao (1997) concluded that AIC and some other closely related criteria (such as C_p and delete-1 cross validation) are "useful when there is no fixed-dimensional correct model", while BIC and delete-d cross validation with $d/n \to 1$ are "useful in the case there exist fixed-dimensional correct models".

There has been quite a debate on AIC and BIC in the literature. One major part of the argument is on the assumptions. Some researchers in favor of AIC argue strongly that there is no finite-dimensional "true model". For example, Burnham and Anderson (2002) "do not accept the notion that there is a simple 'true model' in the biological sciences". If this view is to be taken, then consistency in terms of selecting the true model is not relevant and thus BIC is immediately deprived of its best known theoretical property. In the mean time, AIC stands out as an asymptotically efficient criterion. In defense of BIC, one can argue that in certain situations, some very simple models are clearly superior to nonparametric alternatives and are very satisfactory for practical concerns (imagine a simple linear regression with R^2 close to 1) and it makes strong sense to practically view one of the parametric models as the "true model". Even if the "true model" is infinite-dimensional, as Speed and Yu (1993, Section 4) pointed out, the relative performance between AIC and BIC depends on how fast the approximation error decays to zero. For instance, if the approximation error decreases super-exponentially fast in the model dimension (i.e., e^{-e^k}, where k is the model dimension), BIC is also asymptotically efficient. For such a case, one may prefer BIC for obtaining a more economic model. Zhang (1997, p. 255), in the discussion of Shao (1997), goes even further in this direction by stating that "An argument can be made in favor of BIC-like criteria regardless of the true model" for the reasons that the existence of a true model (regardless of the dimension) is doubtful in the first place and that "a parsimonious model often overshadows concerns over the correctness of the models". In any event, from the literature, it seems quite clear that should the debaters agree on that the true model exist and is finite-dimensional and reasonably simple, there would be little dispute: BIC is the better choice, from the consistency perspective, from the loss (average squared error) efficiency perspective, and from a sequential prediction perspective.

Now given below is what I would have predicted on the comparison between AIC and BIC based on my understanding of the literature on model selection before this study.

1. No procedure dominates any other one.

2. We expect three regions of β, β small, β large and in between, which correspond to BIC being better, AIC and BIC about the same, and AIC better. The first and third regions should shrink as n gets larger.

3. For this setting, from the literature, BIC should be favored in an overall sense.

How about you?

Now let's examine Figure 1. To help us compare the methods better, in Figure 3, we redraw Figure 1 with the risks replaced by the ratio of the original risks over the minimum of the risks of model 0 and model 1. Are there any surprises to you? To us, there are.

1. Somewhat surprisingly, at $n = 25$, BIC loses to AIC at more β values and more severely compared to its shining moments.

2. More surprisingly, for BIC, the situation does not improve much as n gets larger.

3. The worst-case performance of BIC relative to AIC does not seem to get better as n gets even to 1000.

4. At the first two sample sizes, the methods differ in risk quite a bit, indicating the potential relevance of the comparison between the methods to real applications.

5. At sample size 25, the testing approach behaves very poorly unless β is very small or quite large.

What does the example tell us? Note that when β is around zero, the advantage of BIC is relatively small (in risk) and the disadvantage of BIC is larger when β is larger. Thus it seems that BIC is paying a very high price for performing well at β around zero. Therefore, it seems fair to say that unless one has a strong reason to believe that β is around zero, the penalty of BIC seems too large and thus is too conservative for yielding a good predictive performance. The same can be said for the testing approach for the relatively small sample sizes (though it improves more than BIC when n gets larger).

The finding of the poor performance of the testing approach for prediction/estimation should not be taken lightly. The testing approach seems to be the one instructed (more or less) in many (if not most) elementary statistics textbooks. The example clearly shows that estimation/prediction is fundamentally different from assessing if model 0 or model 1 is the true model. The commonly taught and practiced approach of identifying the true model first by testing and then making inference based on the selected model is problematic.

Regarding AIC and BIC, the story that "BIC is good if the true model is finite-dimensional and AIC is good if the true model is infinite-dimensional" does not seem to be right here. Is the sample size of 1000 still not enough for BIC to perform as well as the asymptotic results seem to tell? Actually, the behavior that the worst-case risk of BIC is getting increasingly worse compared to AIC was given by Foster and George (1994). Somewhat surprisingly, even though the risk inflation criterion (RIC) proposed in their paper is quite well-known, as far as we know, this sub-optimality of BIC has rarely been discussed. Consider the whole picture, for a moderate sample size, advantage of BIC around $\beta = 0$ seems to be limited compared to its disadvantage when $|\beta|$ is larger. Even when n is large, the aforementioned optimality property of BIC do not seem to reflect the reality well.

This paper grew out of our effort in trying to understand the properties of the model selection methods. The main points of this paper are:

1. For the purpose of estimation/prediction, we compare AIC and BIC from different theoretical angles to understand their differences. The theoretical results and empirical comparisons show that the notion "BIC is good if the true model is finite-dimensional and AIC is good if the true model is infinite-dimensional" is misleading.

2. There is an unbridgeable difference between the pointwise and minimax properties of model selection. This means that no model selection methods, however sophisticated, can share the essential strengths of AIC and BIC.

3. Large (e.g., BIC) and small (e.g., AIC) penalties in model selection are both justifiable from a weighted worst-case risk point of view.

4. We compare model selection and model averaging methods.

Some of the theoretical results given in this paper are not really original in the sense that similar results (under different loss functions) were given already in the literature. Our objective is to put these results and new ones into perspectives, which hopefully will help the reader to understand the complex issues involved in model selection.

We choose to study the simple problem in this work mainly for two reasons: simplicity for better theoretical and empirical understandings and it is also a reasonably useful problem for application.

3 Theoretical comparisons between AIC and BIC

In this section, we present theoretical comparisons between AIC and BIC from several different angles all under the consideration of the prediction/estimation risk at a given point x_0.

Let $\mathcal{F}_0$ and $\mathcal{F}_1$ be two classes of functions. Consider two models:

$$Y_i = f(x_i) + \varepsilon_i, \quad f \in \mathcal{F}_0$$

and

$$Y_i = f(x_i) + \varepsilon_i, \quad f \in \mathcal{F}_1.$$

Let $\hat{f}_{0,n}$ and $\hat{f}_{1,n}$ be estimators of f under model 0 and 1 respectively.

3.1 A pointwise adaptation result

Consider a model selection rule δ. Let A_δ be the event that model 1 is selected.

Definition 1. The model selection rule δ is pointwise-risk adaptive with respect to the two estimators $\hat{f}_{0,n}$ and $\hat{f}_{1,n}$ if for all $f \in \mathcal{F}_0 \cup \mathcal{F}_1$, we have

$$\frac{E\left(f(x_0) - \left(\hat{f}_{0,n}(x_0) I_{A_\delta^c} + \hat{f}_{1,n}(x_0) I_{A_\delta}\right)\right)^2}{\min\left(E\left(f(x_0) - \hat{f}_{0,n}(x_0)\right)^2, E\left(f(x_0) - \hat{f}_{1,n}(x_0)\right)^2\right)} \to 1 \tag{3}$$

as $n \to \infty$.

In words, δ is pointwise-risk adaptive if for each given f in model 0 or 1, the estimator based on δ is asymptotically as good as the better one between $\hat{f}_{0,n}$ and $\hat{f}_{1,n}$ in risk. We use the term *pointwise-risk* to emphasize that the regression function f is fixed for the asymptotic analysis (as $n \to \infty$).

Now in our simple context, let $\hat{f}_{0,n}(x_0) = \hat{\alpha}$ and $\hat{f}_{1,n}(x_0) = \hat{\alpha} + \hat{\beta}x_0$. Note that when $x_0 = 0$, the two estimators are actually the same.

Theorem 1. Assume that the design is such that $\sum x_i^2$ is of order n as $n \to \infty$. Consider $x_0 \neq 0$. BIC is pointwise-risk adaptive with respect to $\hat{f}_{0,n}$ and $\hat{f}_{1,n}$ but AIC is not. More generally, the model selection rule in (2) with $a_n = o(\sqrt{n})$ is pointwise-risk adaptive if and only if $a_n \to \infty$.

Remarks:

1. Results similar to Theorem 1 have been well known in the literature for general linear model selection, usually under the mean average squared error at the training sites. See for example, Nishi (1984) and Shao (1997) and references therein. Speed and Yu (1993) give lower bounds for prediction risks and study achievability of the bounds by the familiar model selection criteria.

2. AIC fails to be pointwise-risk adaptive when $\beta = 0$ because it selects (wrongly) model 1 with a non-vanishing probability no matter how large the sample size is.

3. For each fixed $\beta \neq 0$, the ratio of the risks of AIC and BIC converges to 1 as $n \to \infty$.

4. If $a_n = o(\sqrt{n})$ is not satisfied, the criterion has a non-vanishing probability of under-fitting. See, e.g., Shao (1997) for a more general result on under-fitting probability.

5. The condition that $\sum x_i^2$ is of order n rules out highly irregular cases such as degeneration of the design points to a singleton.

Proof of Theorem 1: For the model selection criterion with cutoff a_n, the corresponding estimator of $f(x_0)$ is

$$\hat{f}(x_0) = \hat{\alpha} + \hat{\beta}x_0 I_{\left\{|\hat{\beta}| \geq a_n s/\sqrt{\sum_{i=1}^n {x_i}^2}\right\}}.$$

Let $A_\delta = \left\{|\hat{\beta}| \geq a_n s/\sqrt{\sum_{i=1}^n {x_i}^2}\right\}$. Then

$$R_\beta(\delta; x_0; n) = \frac{\sigma^2}{n} + x_0^2 E_\beta \left(\hat{\beta} I_{A_\delta} - \beta\right)^2.$$

Consider first the case $\beta = 0$, i.e., model 0 holds. Then clearly $\hat{f}_{0,n}(x)$ is better than $\hat{f}_{1,n}(x)$ and the their risks are $\frac{\sigma^2}{n}$ and $\frac{\sigma^2}{n} + x_0^2 E_\beta \left(\hat{\beta} - \beta\right)^2 = \frac{\sigma^2}{n} + \frac{\sigma^2 x_0^2}{\sum_{i=1}^n {x_i}^2}$, respectively. For the model selection criterion to be pointwise-risk adaptive, we must have $nE_{\beta=0}\left(\hat{\beta} I_{A_\delta} - \beta\right)^2 \to 0$ (for $x_0^2 > 0$). Since

$$A_\delta = \left\{\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\hat{\beta}}{\sigma} \geq \frac{a_n s}{\sigma}\right\} \cup \left\{\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\hat{\beta}}{\sigma} \leq -\frac{a_n s}{\sigma}\right\}$$

and with $\beta = 0$, $\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\hat{\beta}}{\sigma}$ has a standard normal distribution, we know that $P_{\beta=0}(A_\delta) \to 0$ if and only if $a_n \to \infty$. This fact is enough to rule AIC out as a pointwise-risk adaptive criterion. Since $\sqrt{\sum_{i=1}^n {x_i}^2}\hat{\beta}/\sigma$ has a standard normal distribution,

$$E_{\beta=0} n\left(\hat{\beta}-\beta\right)^2 I_{A_\delta} = \frac{\sigma^2}{\sum_{i=1}^n {x_i}^2/n} E_{\beta=0}\left(\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\hat{\beta}}{\sigma}\right)^2 I_{A_\delta} \to 0$$

as long as $P_{\beta=0}(A_\delta) \to 0$.

Now consider $\beta \neq 0$. Without loss of generality, assume $\beta > 0$. When n is large enough, $\hat{f}_{0,n}(x)$ is worse than $\hat{f}_{1,n}(x)$ and the their risks are $\frac{\sigma^2}{n} + x_0^2\beta^2$ and $\frac{\sigma^2}{n} + x_0^2 E_\beta\left(\hat{\beta}-\beta\right)^2 = \frac{\sigma^2}{n} + \frac{\sigma^2 x_0^2}{\sum_{i=1}^n {x_i}^2}$, respectively. Thus to show that the model selection criterion is pointwise-risk adaptive, we need to show that

$$E_\beta\left(\hat{\beta} I_{A_\delta} - \beta\right)^2 - E_\beta\left(\hat{\beta}-\beta\right)^2 = o(n^{-1}).$$

But the quantity is equal to

$$E_\beta\left(\hat{\beta}\left(2\beta-\hat{\beta}\right) I_{A_\delta^c}\right).$$

Note that for $\beta \neq 0$, $\hat{\beta}\left(2\beta-\hat{\beta}\right)$ converges in probability to β^2 and it is not hard to show that, under the normality assumption on the errors, $E_\beta\left(\hat{\beta}\left(2\beta-\hat{\beta}\right) I_{A_\delta^c}\right) = o(n^{-1})$ holds if $P_\beta(A_\delta^c) = o(n^{-1})$. Indeed,

$$\begin{aligned}
E_\beta|\hat{\beta}\left(\hat{\beta}-2\beta\right) I_{A_\delta^c}| &= E_\beta\left(|\hat{\beta}||\hat{\beta}-2\beta| I_{A_\delta^c} I_{\{|\hat{\beta}|\le 1.5\beta\}}\right) + E_\beta\left(|\hat{\beta}||\hat{\beta}-2\beta| I_{A_\delta^c} I_{\{|\hat{\beta}|>1.5\beta\}}\right) \\
&\le 5.25\beta^2 P_\beta(A_\delta^c) + E_\beta\left(|\hat{\beta}||\hat{\beta}-2\beta| I_{\{|\hat{\beta}|>1.5\beta\}}\right) \\
&\le 5.25\beta^2 P_\beta(A_\delta^c) + \sqrt{E_\beta\left(|\hat{\beta}|^2|\hat{\beta}-2\beta|^2\right)}\sqrt{P_\beta\left(|\hat{\beta}|>1.5\beta\right)},
\end{aligned}$$

where the last inequality follows from Cauchy-Schwaz inequality. Together with that $P_\beta\left(|\hat{\beta}|>1.5\beta\right) = o(e^{-bn})$ for some $b > 0$, the assertion follows. Now note that

$$\begin{aligned}
P_\beta(A_\delta^c) &= P_\beta\left\{-\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{\sigma} - \frac{a_n s}{\sigma} \le \frac{\sqrt{\sum_{i=1}^n {x_i}^2}\left(\hat{\beta}-\beta\right)}{\sigma} \le -\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{\sigma} + \frac{a_n s}{\sigma}\right\} \\
&\le P_\beta\left\{\frac{a_n s}{\sigma} \ge \frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{2\sigma}\right\} + P_\beta\left\{\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\left(\hat{\beta}-\beta\right)}{\sigma} \le -\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{2\sigma}\right\}.
\end{aligned}$$

With $a_n = o(\sqrt{\sum x_i^2})$, the above last two probabilities are both exponentially small in n. This completes the proof of Theorem 1.

3.2 Worst-case performance

From Theorem 1, when $\beta = 0$, BIC beats AIC, and when $\beta \neq 0$ AIC and BIC are asymptotically equivalent in risk as $n \to \infty$. Together with that BIC is consistent, it gives one a strong impression that BIC should be used for a parametric problem. Speed and Yu (1993) added further weight to BIC by showing that BIC (but not AIC)

achieves a lower bound for cumulative prediction risk. The literature does seem to say/imply that BIC should be preferred for a parametric setting, at least theoretically speaking.

Obviously, there are various ways to compare the risk functions of estimation procedures. The asymptotics of Theorem 1 are done for each fixed f in the function classes. One drawback of this notion of asymptotics is that it tells very little for the specific data at hand because in general the convergence of the risk ratio in (3) is not uniform over f. An implication is that for a data set where the distinction between the two models is not clear due to small n or large noise (that is exactly where we need a good model selection rule most), it is especially uncertain whether the asymptotics hold. A different notion to compare risk functions is in terms of worst-case risk.

Let $\mathcal{F}$ be a class of regression functions. The minimax risk of estimating $f(x_0)$ is defined to be:

$$R(\mathcal{F}; x_0; n) = \min_{\hat{f}} \max_{f \in \mathcal{F}} E\left(f(x_0) - \hat{f}(x_0)\right)^2,$$

where $\hat{f}$ is over all estimators based on the data.

Definition 2: A model selection rule δ is minimax-rate adaptive with respect to the two classes $\mathcal{F}_0$ and $\mathcal{F}_1$ if for both $j = 0, 1$, we have

$$\sup_n \frac{\sup_{f \in \mathcal{F}_j} E\left(f(x_0) - \left(\hat{f}_{0,n}(x_0) I_{A_\delta^c} + \hat{f}_{1,n}(x_0) I_{A_\delta}\right)\right)^2}{R(\mathcal{F}_j; x_0; n)} < \infty.$$

Theorem 2: Assume that the design is such that $\sum x_i^2$ is of order n as $n \to \infty$. Then AIC is minimax-rate adaptive but BIC is not. More generally, the model selection rule δ of the form (2) is minimax-rate optimal if and only if a_n is bounded above from infinity.

The result may be a little surprising. One may naturally think that since BIC is consistent and thus will eventually select the true model. Together with its pointwise-risk adaptation property, it looks like that it should perform optimally at least in rate. Another thought may be that if I simply ignore model selection and use model 1 (the larger model), I always get the optimal rate of convergence. I probably should demand BIC to do at least as well as that.

A similar result to Theorem 2 was obtained by Foster and George (1994) in a linear regression setting assuming σ^2 is known and under the average squared error at the training sites. Somewhat surprisingly, the sub-optimality of BIC did not bring in an appropriate caution on the overly simplified and overly optimistic view on BIC for a parametric setting.

Proof of Theorem 2: With the same notations as in the proof of Theorem 1, we first show that for any $a_n \to \infty$ and a fixed $c > 0$, we have

$$\frac{\sup_{|\beta| \le c} E_\beta\left(\hat{\beta} I_{A_\delta} - \beta\right)^2}{1/n} \to \infty,$$

which is sufficient to conclude that when $a_n \to \infty$, the model selection rule is minimax-rate sub-optimal. Note that the LHS above is equal to

$$\begin{aligned} & \sup_{|\beta|\leq c} E_\beta \left(\sqrt{n}\hat{\beta} I_{A_\delta} - \sqrt{n}\beta\right)^2 \\ = & \sup_{|\beta|\leq c} E_\beta \left(\sqrt{n}\left(\hat{\beta}-\beta\right) I_{A_\delta} - \sqrt{n}\beta I_{A_\delta^c}\right)^2 \\ = & \sup_{|\beta|\leq c} \left(E_\beta n\left(\hat{\beta}-\beta\right)^2 I_{A_\delta} + n\beta^2 P_\beta\left(A_\delta^c\right)\right). \end{aligned}$$

Observe that

$$A_\delta^c = \left\{ -\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{\sigma} - \frac{a_n s}{\sigma} \leq \frac{\sqrt{\sum_{i=1}^n {x_i}^2}\left(\hat{\beta}-\beta\right)}{\sigma} \leq -\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{\sigma} + \frac{a_n s}{\sigma} \right\}.$$

Take $\beta_n = \min(\sqrt{a_n/n}, c)$. It is easy to show that for any constant $c^{'} > 1$, $P(s/\sigma \geq c^{'}$ or $s/\sigma \leq 1/c^{'})$ converges to zero exponentially fast in n (i.e., $P(s/\sigma \geq c^{'}$ or $s/\sigma \leq 1/c^{'}) = o\left(e^{-bn}\right)$ for some $b > 0$. See, e.g., Yang (1999a, p. 490-491)). Note also that $\sqrt{n}\beta_n \leq \sqrt{a_n} = o\left(a_n\right)$ and thus $\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{\sigma} = o(a_n)$. Then $P_{\beta_n}\left(A_\delta^c\right)$ is asymptotically equivalent to

$$P_{\beta_n}\left(-a_n\left(c^{'}\right)^{-1} \leq N(0,1) \leq a_n\left(c^{'}\right)^{-1}\right),$$

which approaches 1. Since $n\beta_n^2 \to \infty$, we have $n\beta_n^2 P_{\beta_n}\left(A_\delta^c\right) \to \infty$. Therefore, $a_n \to \infty$ implies that the worst-case risk of the model selection estimator with cutoff a_n converges at a rate slower than n^{-1}.

It remains to show that if a_n stays bounded, the resulting estimator converges at rate n^{-1}. To that end, we need to show

$$\sup_n \sup_\beta \left(E_\beta n\left(\hat{\beta}-\beta\right)^2 I_{A_\delta} + n\beta^2 P_\beta\left(A_\delta^c\right)\right) < \infty.$$

Note that $E_\beta n\left(\hat{\beta}-\beta\right)^2 I_{A_\delta} \leq nE_\beta\left(\hat{\beta}-\beta\right)^2 = \frac{n\sigma^2}{\sum_{i=1}^n {x_i}^2}$, which is upper bounded. Consequently, it suffices to show $n\sup_{\beta\geq 0}\beta^2 P_\beta\left(A_\delta^c\right)$ is upper bounded. Observe that

$$\begin{aligned} & \left\{ -\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{\sigma} - \frac{a_n s}{\sigma} \leq \frac{\sqrt{\sum_{i=1}^n {x_i}^2}\left(\hat{\beta}-\beta\right)}{\sigma} \leq -\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{\sigma} + \frac{a_n s}{\sigma} \right\} \\ \subset & \left\{ \frac{a_n s}{\sigma} \geq \frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{2\sigma} \right\} \cup \left\{ \frac{\sqrt{\sum_{i=1}^n {x_i}^2}\left(\hat{\beta}-\beta\right)}{\sigma} \leq -\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{2\sigma} \right\}. \end{aligned}$$

Thus

$$\begin{aligned} & n\beta^2 P_\beta\left(A_\delta^c\right) \\ \leq & n\beta^2 P_\beta\left(\frac{a_n s}{\sigma} \geq \frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{2\sigma}\right) + n\beta^2 P_\beta\left(N(0,1) \leq -\frac{\sqrt{\sum_{i=1}^n {x_i}^2}\beta}{2\sigma}\right). \end{aligned}$$

For $|\beta| \leq n^{-1/2}$, clearly $n\beta^2 P_\beta\left(A_\delta^c\right)$ is upper bounded by 2. For $|\beta| > n^{-1/2}$, note that the two probabilities in the right-hand side in the above display decay exponentially fast in $\sqrt{\sum_{i=1}^n {x_i}^2}\beta$. It follows easily that $n\sup_\beta \beta^2 P_\beta\left(A_\delta^c\right)$ is indeed upper bounded. This completes the proof of Theorem 2.

3.3 Comments on Theorems 1 and 2

Theorems 1 and 2 may seem to tell opposite stories. How do we interpret them? How do we use them to guide real applications if possible? Should one take the minimax or pointwise view point?

In general, in our view, pointwise asymptotic results provide rather little guidance for real applications. For example, a universally consistent estimator in nonparametric regression with i.i.d. observations (i.e., it is consistent without any assumption on the common marginal distribution) may not perform well in practice even though theoretically it "works" universally. Pointwise asymptotics emphasize the positive outlook by assuming that we have a large number of observations for the current problem of interest. This is the case for some applications (think of many examples of simple linear regression in the elementary statistical textbooks) and perhaps can also be made the case sometimes when collecting more data is not a serious concern. In other situations, when one gets a larger sample size, it is usually desirable to consider more explanatory variables (which are usually available). For such a case, relevance and validity of the pointwise asymptotic analysis become less clear. It is perhaps worth pointing out that in the nonparametric world, pointwise asymptotics can be even less reliable in some sense. For example, super efficiency (i.e., pointwise convergence rate is faster than the minimax rate) can occur at every function in an infinite-dimensional class of regression functions (see, e.g., Brown, Low and Zhao (1997)). For another example, in pattern recognition, if one considers pointwise asymptotics, classification is easier (in terms of rate of convergence) than estimating the conditional probability function (Devroye, Györfi and Lugosi (1996)). But if one considers the minimax rate of convergence, the two problems are actually of the same difficulty for many familiar function classes (Yang (1999b)). In any case, in a typical application where model selection is clearly a nontrivial issue, perhaps more often than not, we are not is a situation where the pointwise asymptotic behavior has "kicked in" already (or at least it is not evident that is the case). This is particularly true when the model selection methods strongly disagree with each other. To address this concern, properties such as minimaxity can be very helpful for better guiding the model selection practice.

From a pure mathematical statistics point of view, the debate between AIC and BIC cannot be well resolved in the sense that they work well under different conditions/assumptions and the question of which condition is more appropriate cannot be answered by any theoretical analysis. For a successful application of model selection, it seems clear that one is obligated to take into account the context and background of the data and prior experience whenever possible. A simple (perhaps naive) consideration is to prefer AIC if β is more likely to be nonzero and use BIC if β is likely to be very small or zero based on experience or a theory about the subject matter. Of course, one can always do both AIC and BIC. In our simple case, there is a good chance that they agree with each other (which happens with high probability when $|\beta|$ is small or large relative to the noise level) and then there is little concern. We will come back later to discuss the situation when AIC and BIC do differ.

Another comment is that whenever possible, we should try to understand/assess whether the difference in prediction resulted from the different methods is really significant or not for the subject matter even if the statistical

significance is obvious. Even though statistical significance is usually important, its practical relevance should be kept in mind in applications.

3.4 Can the strengths of AIC and BIC be shared?

3.4.1 Adaptive asymptotic loss efficiency

One property that separates AIC and BIC is the condition under which an asymptotic loss efficiency (to be defined) holds. Shao (1997) showed that AIC (or a similar criterion) is asymptotically loss efficient when there is no fixed-dimensional correct mode and BIC (or a similar criterion) is asymptotically efficient if there is a fixed-dimensional correct model. If one considers only the model selection criteria of the form $RSS + \lambda_n \hat{\sigma}_n^2 k$, where k is the model dimension, $\hat{\sigma}_n^2$ is an estimator of σ^2 and λ_n is the penalty coefficient ($\lambda_n = 2$ for AIC and $\lambda_n = \log n$ for BIC), then from Shao (1997), it is clear that any choice of λ_n cannot yield the asymptotic loss efficiency under both of the aforementioned conditions. This brings up the interesting question: Can we construct an adaptive model selection rule that does achieve the goal of asymptotic loss efficiency for both situations?

Under some conditions, we show that the answer is yes. Even though the rest of the paper focuses on the parametric case, we feel that this result is useful and interesting enough to be included here (at least in a sketchy way) for a good understanding of the model selection criteria. Again, we will not try to be general and prefer simplicity for seeing the essence more clearly.

Consider a collection of nested linear models with model index $k \geq 1$, where k is simply the dimension of the model. The true model is

$$Y_i = f(x_i) + \varepsilon_i,$$

where $x_i = (x_{i1}, ..., x_{id})$ for some $d \geq 1$ and the errors are assumed to be normally distributed with mean zero and known variance $\sigma^2 = 1$. Let $M_k = M_{k,n}$ denote the projection matrix of model k. Let $f_n = (f(x_1), f(x_2), ..., f(x_n))'$ and $\underline{Y} = (Y_1, ..., Y_n)'$ be the vectors of the true values of the regression function at the design sites and the observations, respectively. Let $\hat{f}_k = M_k \underline{Y}$ be the least squares estimator of f_n based on model k. Let $\| \cdot \|_n$ denote the Euclidean distance on R^n.

Let δ be a model selection rule and let $\hat{k}_n$ be the selected model.

Definition 3: δ (or $\hat{k}_n$) is said to be asymptotically loss efficient if

$$\frac{\| f_n - \hat{f}_{\hat{k}_n} \|_n^2}{\inf_{k \geq 1} \| f_n - \hat{f}_k \|_n^2} \to 1 \text{ in probability as } n \to \infty.$$

As mentioned already, AIC and BIC are asymptotically loss efficient for nonparametric and parametric cases, respectively, but not both (Shao (1997)). We need certain conditions to establish adaptive asymptotic loss efficiency.

When at least one of the candidate models is correct, a model selection rule is said to be strongly consistent if the selected model is eventually equal to the smallest correct model with probability one. For strong consistency of BIC for regression (under conditions on the design matrices), see Rao and Wu (1989).

Assumption 1. BIC is strongly consistent when at least one of the candidate models is correct.

Assumption 2. AIC is asymptotically loss efficient when none of the candidate models is correct.

Sufficient conditions for Assumption 2 are given by Shao (1997).

We now construct an adaptive model selection rule. Let $\hat{k}_{AIC,n}$ and $\hat{k}_{BIC,n}$ be the model selected by AIC and BIC respectively at sample size n. Let $b_n = \max(3, \lfloor \log\log\log n \rfloor)$ be an increasing sequence of integers. Let

$$\hat{\hat{k}}_n = \begin{cases} \hat{k}_{BIC,n} & \text{if } \hat{k}_{BIC,b_n} = \hat{k}_{BIC,b_n+1} = \cdots = \hat{k}_{BIC,n} \\ \hat{k}_{AIC,n} & \text{otherwise.} \end{cases}$$

The idea is very simple: when BIC selects the same model again and again at different sample sizes, the true model is most likely finite-dimensional and hence BIC should be preferred.

For the strategy to work, we need some additional conditions for the nonparametric case. Let $k_n^* = k_{AIC,n}^*$ be the minimizer of $\frac{1}{n} \parallel f_n - M_k f_n \parallel_n^2 + \frac{k}{n}$ over $k \geq 1$. Note that k_n^* provides the best trade-off between the approximation error $\frac{1}{n} \parallel f_n - M_k f_n \parallel_n^2$ and the estimation error $\frac{k}{n}$ (recall σ^2 is assumed to be 1). We require that the approximation errors to behave regularly in the following sense.

Assumption 3.

1. If the approximation errors satisfy that for $k \geq 1$,

$$\frac{\frac{1}{n} \parallel f_n - M_k f_n \parallel_n^2}{e^{-e^k}} \geq c > 0$$

for some constant c, then

$$\inf_k \left(\frac{1}{n} \parallel f_n - M_k f_n \parallel_n^2 + \frac{k \log n}{n} \right) = o\left((\log\log n)^{-1} \right). \tag{4}$$

2. If the approximation errors satisfy that for $k \geq 1$,

$$\frac{\frac{1}{n} \parallel f_n - M_k f_n \parallel_n^2}{e^{-e^k}} \leq \overline{c} < \infty$$

for some constant $\overline{c}$, then

$$\frac{k_n^*}{k_{\lfloor n/\log n \rfloor}^*} \to 1 \text{ as } n \to \infty. \tag{5}$$

For interesting nonparametric regressions, it is usually the case that $\inf \left(\frac{1}{n} \parallel f_n - M_k f_n \parallel_n^2 + \frac{k}{n} \right)$ converges basically at (or around) a polynomial rate n^{-r} for some $0 < r < 1$ and then (4) is satisfied. Speed and Yu (1993) observed that when the approximation error rapidly decays at e^{-e^k}, AIC and BIC should perform the same in the sense

$$\frac{\parallel f_n - \hat{f}_{\hat{k}_{BIC,n}} \parallel_n^2}{\parallel f_n - \hat{f}_{\hat{k}_{AIC,n}} \parallel_n^2} \to 1 \text{ in probability as } n \to \infty.$$

Note that the condition (5) is satisfied when the approximation error is ce^{-e^k} for the models. If the approximation errors behave regularly and of order $O(e^{-e^k})$ (including $o(e^{-e^k})$), we expect it to continue to hold.

Proposition 1. For the model selection rule $\hat{\hat{k}}_n$,

1. when (at least) one of the candidate model is correct, under Assumption 1, $\hat{\hat{k}}_n$ is asymptotically loss effi - cient.

2. when none of the candidate models is correct, under Assumptions 2-3, $\hat{\hat{k}}_n$ is asymptotically loss effi cient.

The proposition says that under some reasonable conditions, an adaptive model selection rule can be asymptotically loss effi cient for both parametric and nonparametric situations.

Proof of Proposition 1: To prove the result, it is suffi cient to show:

1. when at least one of the candidate models is correct, $\hat{\hat{k}}_n$ will eventually equal to $\hat{k}_{BIC,n}$ almost surely;

2. when the fi rst condition of Assumption 3 holds, we have $P(\hat{\hat{k}}_n = \hat{k}_{AIC,n}) \to 1$;

3. when the second condition of Assumption 3 holds, we have

$$\frac{\| f_n - \hat{f}_{\hat{k}_{AIC,n}} \|_n^2}{\| f_n - \hat{f}_{\hat{k}_{BIC,n}} \|_n^2} \to 1 \text{ in probability as } n \to \infty.$$

The fi rst one above clearly holds with the strong consistency assumption on BIC. We next sketch the proof for the latter two cases.

Suppose that Assumption 3(1) holds. It suffi ces to show that $P\left(\hat{k}_{BIC,n} = \hat{k}_{BIC,b_n}\right) \to 0$. Note that at the sample size b_n, $\hat{k}_{BIC,b_n} \leq b_n$. Thus it is suffi cient to prove that $\sum_{k \leq b_n} P\left(\hat{k}_{BIC,n} = k\right) \to 0$. Let $k^*_{BIC,n}$ be the model that minimizes $\frac{1}{n} \| f_n - M_k f_n \|_n^2 + \frac{k(\log n - 1)}{n}$ over $k \geq 1$. Let $Crit(k) = \| \underline{Y} - M_k \underline{Y} \|_n^2 + k \log n$ be the BIC criterion value of model k (we only need to consider $k \leq n$). Let $e_n = (\varepsilon_1, ..., \varepsilon_n)$ and $A_k = I - M_k$, where I is the $k \times k$ identity matrix. Note that

$$Crit(k) = \| A_k f_n \|_n^2 + k(\log n - 1) + 2e_n' A_k f_n + (k - e_n' M_k e_n) + \| e_n \|_n^2, \tag{6}$$

where the last term $\| e_n \|_n^2$ is irrelevant for model comparison due to its always presence and independence of k. Under Assumption 3(1), for $k \leq b_n$, we have

$$\| A_k f_n \|_n^2 \geq \underline{c} n (\log n)^{-1} \tag{7}$$

for some $\underline{c}$ and thus the remainder terms $2e_n' A_k f_n$ and $(k - e_n' M_k e_n)$ in the expression (6) are asymptotically negligible compared to $\| A_k f_n \|_n^2$. Similarly, for $k = k^*_{BIC,n}$, since $k^*_{BIC,n} \to \infty$ as $n \to \infty$ for the nonparametric case, we have that $2e_n' A_{k^*_{BIC,n}} f_n + \left(k^*_{BIC,n} - e_n' M_{k^*_{BIC,n}} e_n\right)$ is asymptotically negligible compared to $\| A_{k^*_{BIC,n}} f_n \|_n^2 + k^*_{BIC,n}(\log n - 1)$. Together with the assumption that $\| A_{k^*_{BIC,n}} f_n \|_n^2 + k^*_{BIC,n}(\log n - 1)$ converges faster than $n (\log n)^{-1}$ for $k \leq b_n$ and (7), we know that asymptotically $Crit(k) > Crit(k^*_{BIC,n})$ with increasingly high probability for $k \leq b_n$. Thus it is unlikely for $\hat{\hat{k}}_n$ to be equal to $\hat{k}_{BIC,n}$. Under the normality assumption, we have exponential probability bounds for the two remainder terms $e_n' A_k f_n$ and $(k - e_n' M_k e_n)$,

which can be used to easily make the above argument rigorous (see, e.g., the analysis in Yang (1999a, pp. 489-492)).

Now we handle the remaining case when the approximation errors decay rapidly. When the approximation error is ce^{-e^k}, it can be shown that $k^*_{BIC,n}$ satisfies that $e^{-e^{k^*_{BIC,n}}} = o\left(\frac{k^*_{BIC,n}}{n}\right)$ and the same holds for $k^*_{AIC,n}$. Consequently the ratio

$$\frac{\| f_n - M_{k^*_{BIC,n}} f_n \|_n^2 + \frac{k^*_{BIC,n}}{n}}{\| f_n - M_{k^*_{AIC,n}} f_n \|_n^2 + \frac{k^*_{AIC,n}}{n}} \tag{8}$$

is asymptotically equal to the ratio of $k^*_{BIC,n}$ and $k^*_{AIC,n}$, which can be easily verified to be approaching 1 as $n \to \infty$. When the approximation errors converge faster than e^{-e^k}, we also have $\frac{1}{n} \| f_n - M_{k^*_{BIC,n}} f_n \|_n^2 = o\left(\frac{k^*_{BIC,n}}{n}\right)$ and the same is true for AIC. Note that when the approximation error is regular, $k^*_{BIC,n}$ is basically equivalent to $k^*_{BIC,\lfloor n/\log n \rfloor}$. Hence the ratio in (8) converges to 1 under Assumption 3(2). One can also show that $\frac{\|f_n - \hat{f}_{\hat{k}_{AIC,n}}\|_n^2}{\|f_n - M_{k^*_{AIC,n}} f_n\|_n^2 + \frac{k^*_{AIC,n}}{n}} \to 1$ in probability as $n \to \infty$ and the same holds for BIC. This completes a sketched proof of the proposition.

Remark: For the adaptive model selection rule $\hat{\hat{k}}_n$, it is not clear if it will eventually agree with AIC or BIC (whichever is the right one) with probability one. It is of interest to investigate if there exists any model selection rule that will eventually take AIC when none of the candidate models is correct and take BIC when at least one of the candidate models is correct.

3.4.2 No model selection rule can be really adaptive between AIC and BIC

Proposition 1 gives some hope to resolve the competition between AIC and BIC by sharing their strengths (it indeed succeeded in one aspect). Note that the adaptive asymptotic loss efficiency is a pointwise convergence property. How about other perspectives?

Theorem 2 shows that BIC pays a somewhat high price for being pointwise-risk adaptive in selection (or one can look at the issue from another angle: AIC pays a somewhat high price for being minimax-rate adaptive, i.e., it is not consistent in terms of selection). Naturally, one may wonder if this weakness of BIC can be overcome while maintaining its pointwise-risk adaptation property. In other words, is it possible to construct a more sophisticated model selection criterion that is both pointwise-risk adaptive and minimax-rate adaptive? If this can be done, then the debate between AIC and BIC is resolved to a large extent.

There are several attempts in that direction. Barron, Yang and Yu (1995) reported that the minimum description length criterion (MDL, Rissanen (1978)), when applied in a nonstandard way, essentially yields a penalty of AIC type or BIC type, whichever is better. This implies that the resulting estimator converges at the minimax optimal rates for nonparametric cases and also optimally in rate in terms of a cumulative prediction error for parametric cases. Hansen and Yu (1997) took a different approach in the MDL approach to have a penalty basically switching between AIC type and BIC type according to the outcome of a suitable test. When the true model is finite-dimensional, the criterion is consistent and pointwise prediction-optimal (Corollary 1 of Hansen

and Yu (1997), see also Hansen and Yu (2001)). George and Foster (2000) proposed new Bayesian model selection criteria based on empirical Bayes approaches to have an adaptive penalty term that acts like BIC or RIC (note that RIC has a penalty of AIC type when the number of models does not grow in the sample size). Rao and Tibshirani (1997) suggested adaptively choosing the penalty constant based on cross validation. Shen and Ye (2002) proposed an use of generalized degree of freedom in the same direction of adaptively selecting the penalty constant and reported very promising simulation results. Yang (2003a) showed empirically that when AIC and BIC estimators are combined, the new estimator performs like the better one under the squared error loss.

Despite the above positive findings, it turns out that the most essential features of AIC and BIC cannot be combined. Consider the setting in Section 1.

Theorem 3. No model selection criterion can be both pointwise-risk adaptive and minimax-rate adaptive at the same time.

The theorem says that the main strengths of AIC and BIC cannot be combined. Thus pointwise-risk adaptation and minimax-rate adaptation are conflicting performance measures to some extent. The result is significant because it gives a clear answer to the fundamental question of how far adaptive model selection can really go.

In a closely related direction, Yang (2003b) showed that in a linear regression setting with multiple candidate models, the consistency property of BIC and the minimax-rate adaptation property of AIC cannot be shared (note that a different loss, i.e. average squared error at the design points, is considered there). Here in Theorem 3 the contrast is made between pointwise-risk adaptation and minimax-rate adaptation.

Proof of Theorem 3: Consider a pointwise-risk adaptive model selection criterion δ. Let A_n be the event that model 1 is selected. From the proof of Theorem 2, to show δ is not minimax-rate optimal, it suffices to show

$$\sup_{\beta} n\beta^2 P_\beta\left(A_n^c\right) \to \infty.$$

Since δ is pointwise-risk adaptive, we have that when $\beta = 0$, for $x_0 \in [-1, 1]$,

$$\frac{\frac{\sigma^2}{n} + x_0^2 E_\beta\left(\hat{\beta} I_{A_n} - \beta\right)^2 + 2x_0 E_\beta\left(\hat{\alpha} - \alpha\right)\left(\hat{\beta} I_{A_n} - \beta\right)}{\frac{\sigma^2}{n}} \to 1$$

and thus

$$x_0 E_{\beta=0}\left(\sqrt{n}\hat{\beta}\right)^2 I_{A_n} + 2nE_{\beta=0}\left(\hat{\alpha} - \alpha\right)\hat{\beta} I_{A_n} \to 0.$$

Consequently, we must have $E_{\beta=0}\left(\sqrt{n}\hat{\beta}\right)^2 I_{A_\delta} \to 0$ and $nE_{\beta=0}\left(\hat{\alpha} - \alpha\right)\hat{\beta} I_{A_n} \to 0$. Since under $\beta = 0$, $\sqrt{n}\hat{\beta}$ has a normal distribution with mean zero and variance σ^2, we have $P_{\beta=0}(A_n) \to 0$ as $n \to \infty$.

Consider a testing problem as follows. The observations are from the model:

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

where the errors are independent and have standard normal distribution. Note that this corresponds to the case when $\alpha = 0$. Consider testing $H_0 : \beta = 0$ versus $H_1 : \beta > 0$. If we take the rejection region A_n, δ becomes a

testing rule with probability of type I error approaching zero. We next show, via Neyman-Pearson Lemma, that any test with the probability of type I error going to zero necessarily must have $\sup_{|\beta|\leq c} n\beta^2 P_\beta\left(A_n^c\right) \to \infty$. Let $f(y_1, ..., y_n; \beta)$ denote the joint probability density function of $(Y_1, ..., Y_n)$. Note that for $\beta_1 \geq \beta_0 \geq 0$,

$$\begin{aligned}\frac{f(y_1, ..., y_n; \beta_1)}{f(y_1, ..., y_n; \beta_0)} &= \exp\left(\frac{1}{2}\sum_{i=1}^{n}\left(\left(y_i - \beta_0 x_i\right)^2 - \left(y_i - \beta_1 x_i\right)^2\right)\right) \\ &= \exp\left(\left(\beta_1 - \beta_0\right)\sum_{i=1}^{n} x_i y_i + \left(\beta_0 - \beta_1\right)\sum x_i^2\right).\end{aligned}$$

Thus the family has a monotone likelihood ratio. It follows that a uniformly most powerful (UMP) test is to reject H_0 when $\sum_{i=1}^n x_i y_i$ is larger than some constant C. Let us choose the constant $C = d_n$ so that $P_{\beta=0}\left(\sum_{i=1}^n x_i y_i \geq d_n\right) = P_{\beta=0}(A_n)$. Let $A_{n,*}$ denote the event of $\sum_{i=1}^n x_i y_i \geq d_n$. By the UMP property of $A_{n,*}$, we have for all $\beta_0 > 0$

$$P_\beta\left(A_{n,*}\right) \geq P_\beta\left(A_n\right).$$

Consequently,

$$\sup_{|\beta|\leq c} n\beta^2 P_\beta\left(A_n^c\right) \geq \sup_{|\beta|\leq c} n\beta^2 P_\beta\left(A_{n,*}^c\right).$$

Now since $\sum_{i=1}^n x_i Y_i$ has normal distribution, it is easy to get

$$P_{\beta=0}\left(\sum_{i=1}^{n} x_i Y_i \geq d_n\right) = P\left(N(0,1) \geq \frac{d_n}{\sqrt{\sum x_i^2}}\right),$$

and for $\beta > 0$

$$P_\beta\left(\sum_{i=1}^{n} x_i Y_i < d_n\right) = P\left(N(0,1) < \frac{d_n - \beta\sum x_i^2}{\sqrt{\sum x_i^2}}\right).$$

Since $P_{\beta=0}\left(\sum_{i=1}^n x_i Y_i \geq d_n\right) = P_{\beta=0}(A_n) \to 0$, we must have $\frac{d_n}{\sqrt{n}} \to \infty$ (it can be easily shown that $d_n = o(n)$ for the probability of type II error to converge to zero). Then with the choice of $\beta_n = \frac{d_n}{2\sum x_i^2}$, we have

$$\sup_{|\beta|\leq c} n\beta^2 P_\beta\left(A_{n,*}^c\right) \geq n\beta_n^2 P_{\beta_n}\left(A_{n,*}^c\right) = \frac{nd_n^2}{4\left(\sum x_i^2\right)^2} P\left(N(0,1) < \frac{d_n}{2\sqrt{\sum x_i^2}}\right).$$

Since the last probability above goes to 1 and $\frac{nd_n^2}{4\left(\sum x_i^2\right)^2} \to \infty$, we conclude that $\sup_{|\beta|\leq c} n\beta^2 P_\beta\left(A_{n,*}^c\right) \to \infty$. This completes the proof of Theorem 3.

3.5 Subjectivity of the choice of the penalty in model selection

Note that under model 1, the least squares estimator gives $\hat{f}(x) = \hat{\alpha} + \hat{\beta}x_0$ for estimating $f(x)$. It has a constant risk $\frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$ under the squared error loss. Assume σ^2 is known (or bounded above by a known constant), the estimator $\hat{f}(x)$ is in fact a minimax estimator. Since model 0 is contained in model 1, thus from a minimax point of view, this estimator cannot be improved. However, model selection comes into the picture due to the simple fact that when β is zero or small, the use of model 0 results in a much more accurate estimator/prediction. Clearly,

insisting on the rigid minimax view is inappropriate since then the smaller model is always ignored. Foster and George (1994) proposed a risk inflation criterion to address this issue for linear regression. Later, George and Foster (2000) showed that different choices of the penalty (such as in AIC, BIC and in between) can be derived from asymptotically maximizing the posterior probability under different choices of priors (allowing the priors to possibly depend on the sample size).

Here we show that different choices of the cutoff constant a_n in (2) are sensible depending on how one compares the risk functions of different methods. From this angle, different choices of penalty (or cutoff) are thus subjectively justifiable. We assign a prior probability on the models and weigh the worst-case risks under the candidate models accordingly. Let π_0 and π_1 be the prior probabilities or weights on model 0 and model 1 respectively (obviously $\pi_0 + \pi_1 = 1$). Denote $\pi = (\pi_0, \pi_1)$. Let $\mathcal{F}_0 = \{\alpha : -\infty < \alpha < \infty\}$ and $\mathcal{F}_1 = \{\alpha + \beta x : -\infty < \alpha, \beta < \infty\}$. For a model selection rule δ, let

$$R(\delta; x_0; n; \pi) = \pi_0 \sup_{f \in \mathcal{F}_0} R(f; \delta; x_0; n) + \pi_1 \sup_{f \in \mathcal{F}_1} R(f; \delta; x_0; n)$$

be the weighted worst-case risk. Now obviously, the potential gain of using model 0 is reflected in the new risk.

We have the following natural definition to compare two model selection rules δ_1 and δ_2.

Definition 4: A model selection rule δ_1 is said to be better than δ_2 in terms of the weighted worst-case risk if $R(\delta_1; x_0; n; \pi) \leq R(\delta_2; x_0; n; \pi)$.

For each choice of π_0 between 0 and 1, let $a^*(\pi)$ be the best choice of a that minimizes the weighted worst-case risk among the class of model selection criteria given by (2). For the following result, for simplicity, we assume that σ^2 is known to be 1 and assume that $\sum_{i=1}^n x_i^2 = n$. Accordingly the model selection criteria in (2) is changed to rejecting model 0 when

$$|\hat{\beta}| > \frac{a_n}{\sqrt{n}}.$$

Let $\delta^{\{a_n\}}$ denote this selection rule.

Proposition 2: Let $\pi_0 > 0$ and $\pi_1 > 0$ and $\pi_0 + \pi_1 = 1$. We have the following results under the weighted worst-case risk.

1. If π is fixed, then AIC is better than BIC when n is large enough. In fact, for each fixed π,

$$\frac{R(\delta^{\{a_n\}}; x_0; n; \pi)}{R(\delta^{\{b_n\}}; x_0; n; \pi)} \to \infty$$

   for any $\{a_n\}$ and $\{b_n\}$ with $\lim a_n = \infty$ and $\limsup b_n$ bounded away from ∞.

2. For each fixed sample size, as $\pi_0 \to 1$, $a^*(\pi) \to \infty$.

From the first part of the theorem, giving model 0 a fixed positive prior probability is not enough for BIC to perform well from the weighted worst-case risk point of view. For BIC to outperform AIC, one must have a

shrinking prior probability (to 0) on model 1 as n gets larger. An interpretation of this, again, is that BIC is the right choice if one believes that model 0 (or practically β being around zero) is much more likely to be true.

The second part of Proposition 2 says that when model 1 is less likely to be true, the penalty should be larger for model 1, which clearly makes intuitive sense.

In the general linear regression context, it is natural to consider a class of criteria: choose the model that minimizes

$$RSS + \lambda_n \hat{\sigma}_n^2 k,$$

where RSS is the residual sum of squares, $\hat{\sigma}_n^2$ is an estimate of σ^2, k is the dimension of the model, and λ_n is the penalty constant. It is called the GIC method in Rao and Wu (1989). Note that the issue of choosing λ_n is essentially the same as choosing a_n in (2) in our setting.

Clearly a choice of $\lambda_n = \lambda > 2$ is a compromise between AIC and BIC. However, under the average squared error, in terms of pointwise asymptotics, the criterion is not as good as AIC when the true model is infinite-dimensional and not as good as BIC when the true model is among the candidates (Shao (1997, Section 3)). Zhang (1997) in the discussion of Shao (1997) argues that any choice of $\lambda_n = \lambda$ deserves consideration under a loss that weighs bias and variance differently from the squared error loss. The second part of Proposition 2 justifies a compromising criterion from our minimax point of view.

Proof of Proposition 2: The first part of the result is trivial from Theorem 3. From the proof of Theorem 3, we have that under model 0, $R(f;\delta;x_0;n) = \frac{1}{n} E_{\beta=0}\left(\sqrt{n}\hat{\beta}x_0\right)^2 I_{\{\sqrt{n}|\hat{\beta}|>a_n\}}$ and under model 1, $R(f;\delta;x_0;n) = \frac{1}{n}\left(E_\beta n\left(\hat{\beta}-\beta\right)^2 I_{\{\sqrt{n}|\hat{\beta}|>a_n\}} + n\beta^2 P_\beta\left(-a_n < \sqrt{n}\hat{\beta} < a_n\right)\right)$. Since $\sqrt{n}\left(\hat{\beta}-\beta\right)$ has the standard normal distribution, letting γ denote $\sqrt{n}\beta$ and Z denote a random variable with the standard normal distribution (recall the assumptions on σ^2 and $\sum_1^n x_i^2$), the risk then becomes

$$E_\gamma Z^2 I_{\{Z>a-\gamma \text{ or } Z<-a-\gamma\}} + \gamma^2 P\left(-a-\gamma < Z < a-\gamma\right).$$

For each choice of a, let γ^* denote the worst γ that maximizes the above expression. Consequently, to minimize the weighted worst-case risk, we need to select a_n to minimize

$$\pi_0 E Z^2 I_{\{|Z|>a\}} + \pi_1\left(E_{\gamma^*} Z^2 I_{\{Z>a-\gamma^* \text{ or } Z<-a-\gamma^*\}} + \gamma^{*2} P\left(-a-\gamma^* < Z < a-\gamma^*\right)\right).$$

Obviously, as $a \to \infty$, $EZ^2 I_{\{|Z|>a\}} \to 0$ and it is easy to see that as $a \to \infty$, $E_{\gamma^*} Z^2 I_{\{Z>a-\gamma^* \text{ or } Z<-a-\gamma^*\}} + \gamma^{*2} P\left(-a-\gamma^* < Z < a-\gamma^*\right)$ approaches ∞. Therefore, as $\pi_1 \to 0$, the optimal choice a^* must approaches ∞. This completes the proof of Proposition 2.

4 An alternative penalty constant for prediction

In this section, we give a simple alternative penalty constant (or cutoff) derived directly from the purpose of having a good risk function for estimation/prediction.

Recall that model 0 and model 1 have risk $\frac{\sigma^2}{n} + x_0^2\beta^2$ and $\frac{\sigma^2}{n} + \frac{\sigma^2 x_0^2}{\sum_{i=1}^n {x_i}^2}$ respectively. Thus model 1 is better when $\beta^2 > \frac{\sigma^2}{\sum_{i=1}^n {x_i}^2}$, i.e.,

$$\frac{|\beta|}{\frac{\sigma}{\sqrt{\sum_{i=1}^n {x_i}^2}}} > 1.$$

Since β and σ are unknown, the above relationship cannot be directly used for choosing between model 0 and model 1. Obviously, we can try the plug-in approach of replacing the parameters by their estimates respectively. Let $\hat{\beta}$ be the least squares estimate of β and $s^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$. Then a sensible model selection rule is to choose model 1 when $\frac{|\hat{\beta}|}{\frac{s}{\sqrt{\sum_{i=1}^n {x_i}^2}}} > 1$. Recognizing that the ratio $\frac{\hat{\beta}}{\frac{s}{\sqrt{\sum_{i=1}^n {x_i}^2}}}$ is a biased estimator of $\frac{\beta}{\frac{\sigma}{\sqrt{\sum_{i=1}^n {x_i}^2}}}$, we may consider a modification when n is small. Since the numerator and the denominator are independent, the expectation is $\beta\sqrt{\sum_{i=1}^n {x_i}^2}E\frac{1}{s}$. Since a multiple of $\frac{s^2}{\sigma^2}$ has a chi-square distribution, we can easily find that $E\frac{\sigma}{s} = \sqrt{\frac{n-2}{2}}\frac{\Gamma(\frac{n-3}{2})}{\Gamma(\frac{n-2}{2})}$. With this modification, we come to the criterion that selects model 1 when

$$\frac{|\hat{\beta}|}{\frac{s}{\sqrt{\sum_{i=1}^n {x_i}^2}}} \geq \sqrt{\frac{n-2}{2}}\frac{\Gamma(\frac{n-3}{2})}{\Gamma(\frac{n-2}{2})}.$$

We will denote this new selection rule δ_N. Note that the criterion is even more aggressive in choosing model 1 than AIC (as $n \to \infty$, the cutoff value $\frac{n-2}{2}\frac{\Gamma(\frac{n-3}{2})}{\Gamma(\frac{n-2}{2})}$ approaches 1).

5 Empirical results for comparing the model selection methods

In this section, we give simulations to compare the model selection methods. The theoretical risks of the model selection methods are computed based on Monte Carlo simulations with 3000 replications.

5.1 Comparing the procedures in terms of sample size

Here our interest is to know that, at a given value of the slope parameter β, how AIC, BIC and the other methods compare to each other at different sample sizes. Note that the asymptotic results in Section 3 have already told us that when n is really large, BIC is better than AIC (and the like) when $\beta = 0$; and when β is nonzero, the risks of AIC and BIC are asymptotically equivalent.

Note that when $n \geq 8$, BIC begins to have a cutoff a_n larger than AIC for model 1. In our simulation below, we consider sample sizes beginning from 10. We choose $\sigma = 0.5$.

There are three scenarios, represented by three different values of β: $\beta = 0, 0.2, 0.5$. The plots of the risk functions relative to AIC are given in Figures 4-6. When $\beta = 0$, BIC is better than AIC at all the sample sizes; when $\beta = 0.2$, in the beginning, BIC is better, but AIC becomes better with more observations; when $\beta = 0.5$, AIC is better right from the beginning. The second case ($\beta = 0.2$) indicates that if the sample size is not large enough to reasonably estimate β, it is better to pretend that it is zero. Note that even though in theory the risk ratio of AIC and BIC approaches 1 for the latter two cases, it has not quite happened when n is 150 for both cases.

This again suggests that pointwise asymptotic optimality of BIC may not say much when the sample size is not really large.

Regarding the new selection rule given in Section 4, it performs very well for relatively large β but poorly when β is small.

5.2 A simulation on robustness

It is rarely (or perhaps never) the case that the underlying error distribution is perfectly normal. Here we investigate the performance of the previous model selection methods when the true error distribution is double-exponential but mistaken to be normal. We choose only two sample sizes: $n = 25$ and $n = 100$.

The true error distribution is double exponential with probability density function $g(t) = e^{-2|t|}$. Figure 7 gives the risks of δ_A, δ_B, δ_T, and δ_N (relative to the better one of model 0 and model 1) in terms of β at the sample sizes $n = 25$ and $n = 100$. The results are very much similar as before.

6 What to do when AIC and BIC disagree?

For our simple setting, from (2), the disagreement between AIC and BIC indicates that the test statistic value is neither large nor small (in absolute value). It is a case where the pointwise asymptotic behavior is questionable. The region of dispute for multiple model comparison becomes much more complicated. In any case, when the model selection criteria do disagree, the previous asymptotic results provide little hint on what to do. How should we address the issue?

Prior knowledge (hopefully not too subjective), if available, can be useful. For example, if historically β tends to be small (or there is a good reason to believe so based on a subject matter theory) one should use BIC, and if no such information is available, it is better to use AIC for protection of the worst-case performance. Note that when there are a number of explanatory variables, perhaps it is more often than not that AIC and BIC select different models.

6.1 Combining the models

When the model selection criteria do disagree on the data, as far as estimation/prediction is concerned, a solution to avoiding the pain of selecting the selection rules is a compromise: averaging the estimators from the models. Fortunately with an appropriate weighting of the models, the estimation accuracy can be substantially improved for such a situation.

Much has been said already on general model averaging or model combining, including Bayesian approaches (see Hoeting *et al.* (1999) for a review of the very rich works of Bayesian methods), frequentist approaches based on bootstrap or weighting via model selection criterion values (Breiman (1996) and Burnham *et al.* (1997), and a method called ARM (adaptive regression by mixing) (Yang (2001, 2003a), Yuan and Yang (2003a)). For example,

Yuan and Yang (2003) showed that when model selection instability is high, ARM does a better job than the better one of AIC and BIC. In a simple setting, Yang (2003) proved that model selection, no matter how sophisticated it is, is worse than a combined estimator in an appropriate sense. Yang (2004) discusses two different directions of combining models (or general estimation procedures) with theoretical characterizations of their gains and prices they have to pay.

In general, model selection and model combining are both useful. Yuan and Yang (2003) showed that when model selection instability is high, model combining by ARM tends to do better or much better. But when model selection has little instability, model combining is not necessary and can even perform very poorly.

Below we give the ARM method (Yang (2003a)) for our specific setting.

- *Step 1.* Split the data into two parts $Z^{(1)} = (x_i, Y_i)_{i=1}^{n/2}$ and $Z^{(2)} = (x_i, Y_i)_{i=n/2+1}^{n}$.
- *Step 2.* Estimate the parameters by the least squares method based on $Z^{(1)}$ for each model. Let $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ be the usual unbiased estimators of σ^2 from model 0 and model 1 respectively (again based only on $Z^{(1)}$).
- *Step 3.* For each of the fitted models, assess the accuracies of the models using the remaining half of the data $Z^{(2)}$. Let D_0 and D_1 be the prediction sum of squares from model 0 and model 1 respectively.
- *Step 4.* Compute the weight

$$W_1 = \frac{(\hat{\sigma}_1)^{-n/2} \exp\left(-\hat{\sigma}_1^{-2} D_1/2\right)}{\sum_{j=0,1} (\hat{\sigma}_j)^{-n/2} \exp\left(-\hat{\sigma}_j^{-2} D_j/2\right)}.$$

  Let $W_0 = 1 - W_1$.

- *Step 5.* The combined estimate of $f(x_0)$ is

$$\widetilde{f}_n(x) = W_0 \hat{f}_{0,n}(x_0) + W_1 \hat{f}_{1,n}(x_0).$$

We consider an additional variant of ARM: it combines model 0 and model 1 as above only when AIC and BIC disagree with each other.

6.2 Simulation and a data example

6.2.1 A simulation

We now add the two combining methods in the previous subsection in the competition described in the beginning of Section 2. The settings are kept the same. Figures 8 and 9 give the relative performance of the procedures compared to the better of modol 0 and model 1 at sample size 25 and 100 respectively.

From the graphs, we do see the performance improvement by the combining procedures when beta is small (relative to the sample size). The combining procedure 1 always combines the two models and performs really well at the β values where the model selection methods differ most, yet it pays a heavy price when beta is relatively

large (when β is really large, it is fine as seen in Figure 9). When the models are combined less aggressively by the second combining method (i.e., combining when AIC and BIC disagree), the problem of combining method 1 is substantially reduced. Overall, in our judgement, the combining method 2 is the winner.

6.2.2 A data example

We demonstrate how the sample size n influences the relative performance of the model selection and model combining methods. The data set of body fat was used, which has 252 observations and 19 variables and it was obtained from

`http://www.amstat.org/publications/jse/datasets/fat.txt.`

Several problematic cases were removed following the comments by Johnson (1996). The new sample size is 247.

Unlike Penrose, Nelson and Fisher (1985) and Johnson (1996), where the purpose of the analysis is to find predictive equations for the determination of body fat, our interest here is different and we use one predictor at a time.

The response variable is the percent body fat using Brozek's equation (the second variable in the data). The two individually used predictors are age and fat free weight (original variable number 5 and 9). The correlation between the response and age is 0.293, and the correlation between the response and fat free weight is -0.036. Roughly speaking, they correspond to two situations: the slope parameter is not small and the slop parameter is very small or zero.

Figures 10 and 11 give the average prediction squared error of the model selection and model combining methods at sample sizes from 10 to 200. Note that at each sample size, we randomly split the whole data with the first part of the size of the sample size and the second part used to compute the average prediction squared error. This is replicated 200 times. The performances of the methods given in the figures are all relative to the performance of AIC.

The results are very interesting and confirm our earlier understanding. For the first case with age as the predictor, from the whole data, we know that the slope parameter is nonzero and not very small (relatively speaking). When the sample size is around 10, BIC performs better, but soon becomes worse than AIC. It continues to be worsened and then comes back and eventually performs about the same compared to AIC. Note that the pointwise-risk adaptivity of BIC assures that in the situation above, BIC will eventually perform as well as AIC. For the second case with fat free weight as the predictor, the slope parameter is very small (relative to the sample sizes) and we expect BIC to be better and that is indeed the case. As the sample size increases, their difference become smaller in the ratio. From Figure 11, at the end, AIC and BIC performs quite similarly. From the earlier theory, if the slope parameter is really zero, then the gap between AIC and BIC will never be closed. On the other hand, if we had the sample size of 10000 (say) instead of 247, the observed correlation of -0.036 is no longer

negligible and we would see that AIC and BIC cross position with AIC being better as the sample size increases and then eventually the performance ratio goes to 1. It is worth pointing out that comparing the two figures, we see that for a much wider range in case 1, BIC is significantly worse than AIC.

For the test approach, it is between AIC and BIC, performing better than BIC in case 1 but worse for case 2. Overall, it seems that the testing approach is better than BIC.

The new selection rule goes opposite to the testing approach and BIC.

Finally, but not least, the two combining procedures work both very well, except for the first combining procedure with the large sample sizes, where the model selection procedures have little difficulty finding the better model. The approach of combining the models when AIC and BIC disagree is perhaps the overall winner among the competitors.

In our view, when there is little prior experience on the slope parameter, combining the models is a better approach for prediction when model selection methods give different answers.

7 Some thoughts on the hypothesis testing approach

As mentioned already, for our simple problem, the testing approach is perhaps the "standard" way to proceed following the statistics textbooks. This reflects the traditionally important role of hypothesis testing in statistics. Even though information criteria have gained substantial popularity for model comparison, for the simple linear regression problem with only two models, it might be thought as an overkill. Thus it is probably fair to say that even if prediction is the goal, most people would naturally take the testing approach.

The simulation at $n = 25$ quite clearly indicates that the testing approach is not quite in line with the goal of prediction. If we think about it, it is hardly surprising. The traditional Neyman-Pearson approach of hypothesis testing guards against probability of type I error, which is a rational strategy as far as evidence against H_0 is the concern. Quite naturally, one may hope that since the "true" model (if not too complicated) usually outperforms a wrong model, if we can do a good job identifying the true model, our goal of good prediction can be automatically achieved.

This is indeed a philosophy traditionally taken by our profession. For example, Cox (1958) stated that "No consideration of losses is usually involved directly in the inference". Behind this is the optimistic view that we can discover the true model from the data at hand which can then be used to answer different questions with possibly different loss functions. When the model building process is highly uncertain, however, keeping a loss function in mind during the modeling stage can be a better practice. For our simple prediction problem, if we follow the hypothesis testing approach to identify the true model first, it is unclear how the choice of test size affects our ultimate goal of accurate prediction. See Hand and Vinciotti (2003) for some references and related discussions on this issue.

From the above discussion, we feel strongly that the notion "test a hypothesis first and then make a prediction"

should not be the one students get implicitly or explicitly from our future statistics textbooks. A distinction should be made and emphasized between the goals of testing a scientific hypothesis and prediction. Of course, we are not against hypothesis testing at all, but we are concerned about its use for estimation/prediction.

8 Summary and concluding remarks

Many theoretical results have been obtained on model selection. In particular, there has been a serious debate on the issue of AIC versus BIC. Based on the literature, pointwise asymptotic theorems support the popular notion "BIC is good if the true model is finite-dimensional and AIC is good if the true model is infinite-dimensional". However, in this work, with both theoretical examinations and empirical demonstrations, it is seen that the popular notion is inaccurate and it overly simplifies the comparison between the criteria. Even when the true model is among the list of candidate models being considered, even though BIC performs asymptotically as well as if one knew the true model in a poinwise fashion (with the true regression function fixed and the sample size tends to infinity), the worst case performance of BIC (over the regression functions) gets increasingly worse relative to AIC. In general, for a problem where the model selection methods strongly disagree with each other, it is probably more likely than not that the pointwise asymptotic behavior does not reflect the reality (in our opinion, the pointwise asymptotic optimality properties are overly interpreted in the current statistical literature). It is thus not true that when one knows that he/she is dealing with a parametric situation he/she needs to perfer BIC to AIC. In the setting of simple linear regression versus the null model with only the location parameter, the simulations suggest that unless one has has a strong reason to believe that the slope parameter is most likely to be zero or small, BIC performs worse than AIC in an overall sense.

It is clear that no model selection rule can dominate all others. Thus the comparison of different criteria is inherently subjective. Prior experience/knowledge should definitely be used if possible when choosing a criterion. The fact that different penalties in model selection criteria result in different theoretical properties motivated adaptive model selection, where the penalty is adaptively obtained based on the data (instead of being deterministic as in AIC and BIC). Results in this direction show that the adaptive penalty indeed can suitably switch between the AIC and BIC types, with very encouraging empirical support. However, we showed that this cannot go too far: the pointwise optimality property of BIC and the minimax-rate optimality of AIC cannot be integrated by any model selection rule.

When model selection rules give very different answers, model combining is a better alternative approach for estimation/prediction. With a proper weighting, the large variability of the estimator from model selection can be substantially reduced. Empirical results in this work show that combining the models when AIC and BIC disagree gives a much improved overall performance.

We chose a simple setting in this work because the simple structure allows one to obtain a clear grasp of the problem theoretically and empirically and the understandings are also useful for more complicated cases. For

a general situation with multiple models, the issue of under-fitting versus over-fitting is similar, but the regions where one criterion perform better than another become much more complicated (one certainly cannot plot the risk functions easily). It seems clear that the main points made in this work still apply as far as at least two nested models are compared. For the simple regression problem, one perhaps does not lose much by always having the slope parameter. But for multiple regression, of course, the use of the full model can be much worse compared to the use of a good model selection method.

In real applications, it is often desirable to try different kinds of models. When a large number of models are in competition, however, model selection bias can be severe. For results dealing with this issue, see Barron and Cover (1991), Yang and Barron (1998), Yang (1999a), Barron, Birgé and Massart (1999), and Birgé and Massart (2001).

9 Acknowledgement

This work was completed when the author was visiting the Institute for Mathematics and its Applications (IMA) at University of Minnesota as a New Direction Visiting Professor. The funding, research atmosphere and opportunities at IMA are greatly appreciated. The work was also partly supported by NSF CAREER grant DMS0094323.

The author thanks Xiaotong Shen for very helpful discussions. The paper also benefited from the questions and comments from the participants of the statistics seminars the author gave at University of Minnesota and Duke University.

References

[1] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Info. Theory*, 267-281, eds. B.N. Petrov and F. Csaki, Akademia Kiado, Budapest.

[2] Barron, A.R., Birgé, L. and Massart, P. (1999) Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, **113**, 301-413.

[3] Barron, A.R., Yang, Y., and Yu, B. (1994) Asymptotically optimal function estimation by minimum complexity criteria. In *Proc. 1994 Int. Symp. Info. Theory*, p. 38. Trondheim, Norway.

[4] Birgé, L. and Massart, P. (2001) Gaussian model selection. *J. Eur. Math. Soc.*, **3**, 203-268.

[5] Brown, L.D., Low, M.G. and Zhao, L.H. (1997) Supereffiency in nonparametric function estimation. *Ann. Statistics*, **25**, 2607-2625.

[6] Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**, 123-140.

[7] Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A practical Information-Theoretic Approach*, Springer-Verlag Inc (Berlin; New York).

[8] Devroye, L., Györfi , L. and Lugosi, G. (1996) A *probabilistic theory of pattern recognition*. Springer-Verlag Inc (Berlin; New York)

[9] Foster, D.P. and George, E.I. (1994) The risk inflation criterion for multiple regression. *Ann. Statistics*, **22**, 1947-1975.

[10] George, E.I. and Foster, D.P. (2000) Calibration and empirical Bayes variable selection. *Biometrika*, **87**, 731-747.

[11] Hansen, M. and Yu, B. (1999) Bridging AIC and BIC: an MDL model selection criterion. In *Proceedings of IEEE Information Theory Workshop on Detection, Estimation, Classification and Imaging*, p. 63. Santa Fe, NM.

[12] Hansen, M. and Yu, B. (2001) Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, **96**, 746-774.

[13] Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science* (with discussions) **14**, 382–417.

[14] Johnson, R.W. (1996) Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, **4**, available at

`http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html.`

[15] Li, K.C. (1987) Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statistics*, **15**, 958-975.

[16] Nishii, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, **12**, 758-765.

[17] Penrose, K., Nelson, A., and Fisher, A. (1985) Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques (abstract). *Medicine and Science in Sports and Exercise*, **17(2)**, 189.

[18] Polyak, B.T. and Tsybakov, A.B. (1991) Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. *Theory of Probability and its Applications* (Transl of *Teorija Verojatnostei i ee Primenenija*), **35**, 293-306.

[19] Rao, C.R. and Wu, Y. (1989) A strongly consistent procedure for model selection in a regression problem. *Biometrika*, **76**, 369-374.

[20] Rao, J.S. and Tibshirani, R. (1997) Comment on "An asymptotic theory for linear model selection". *Statistica Sinica*, **7**, 249-251.

[21] Rissanen, J. (1978) Modeling by shortest data description. *Automatica*, **14**, 465-471.

[22] Rissanen, J. (1986) Stochastic complexity and modeling. *Annals of Statistics*, **14**, 1080-1100.

[23] Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statistics* **6**, 461-464.

[24] Shao, J. (1997) An asymptotic theory for linear model selection (with discussion). *Statistica Sinica*, **7**, 221-242.

[25] Shibata, R. (1983) Asymptotic mean efficiency of a selection of regression variables. *Ann. Inst. Statisti. Math.* **35**, 415-423.

[26] Speed, T.P. and Yu, B. (1993) Model selection and prediction: Normal regression. *Annals of the Institute of Statistical Mathematics*, **45**, 35-54.

[27] Yang, Y. (1999a) Model selection for nonparametric regression, *Statistica Sinica*, **9**, 475-499.

[28] Yang, Y. (1999b) Minimax Nonparametric Classification—Part I: Rates of Convergence. *IEEE Transaction on Information Theory*, **45**, 2271-2284.

[29] Yang, Y. (2001) Adaptive regression by mixing. *Journal of American Statistical Association*, **96**, 574-588.

[30] Yang, Y. (2003a) Regression with multiple candidate models: selecting or mixing? *Statistica Sinica*, **13**, 783-809.

[31] Yang, Y. (2003b) Can The Strengths of AIC and BIC Be Shared? Preprint 2003-10, Department of Statistics, Iowa State University.

[32] Yang, Y. (2004) Aggregating regression procedures for a better performance, to appear at *Bernoulli*.

[33] Yang, Y. and Barron, A.R. (1998) An Asymptotic Property of Model Selection Criteria. *IEEE Transaction on Information Theory*, **44**, 95-116.

[34] Yuan Z. and Yang, Y. (2003). Combining Linear Regression Models: When and How? submitted.

[35] Zhang, P. (1997) Comment on "An asymptotic theory for linear model selection". *Statistica Sinica*, **7**, 254-258.

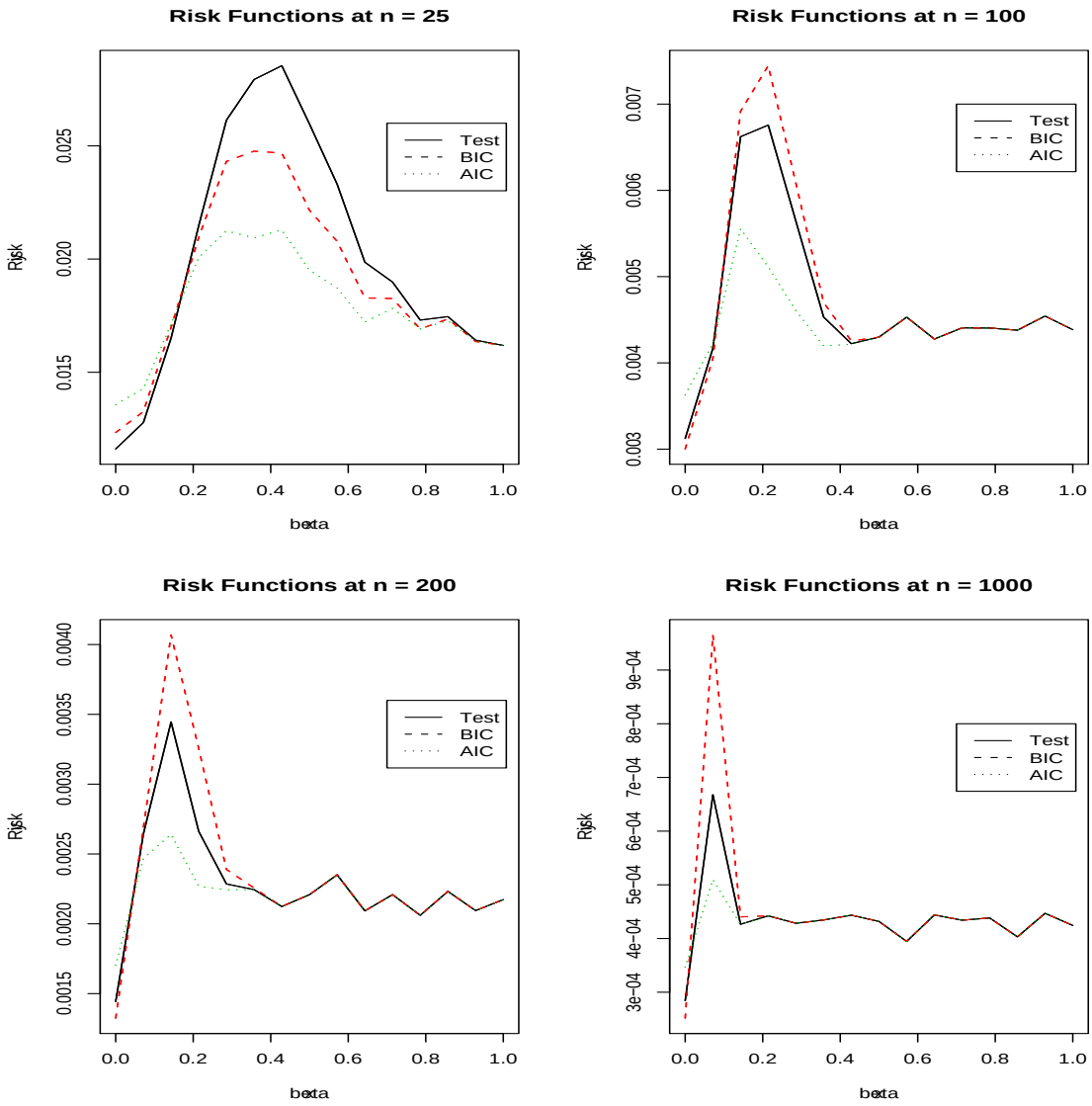


Figure 1: *Comparing the Model Selection Methods in Risk at 4 Sample Sizes*

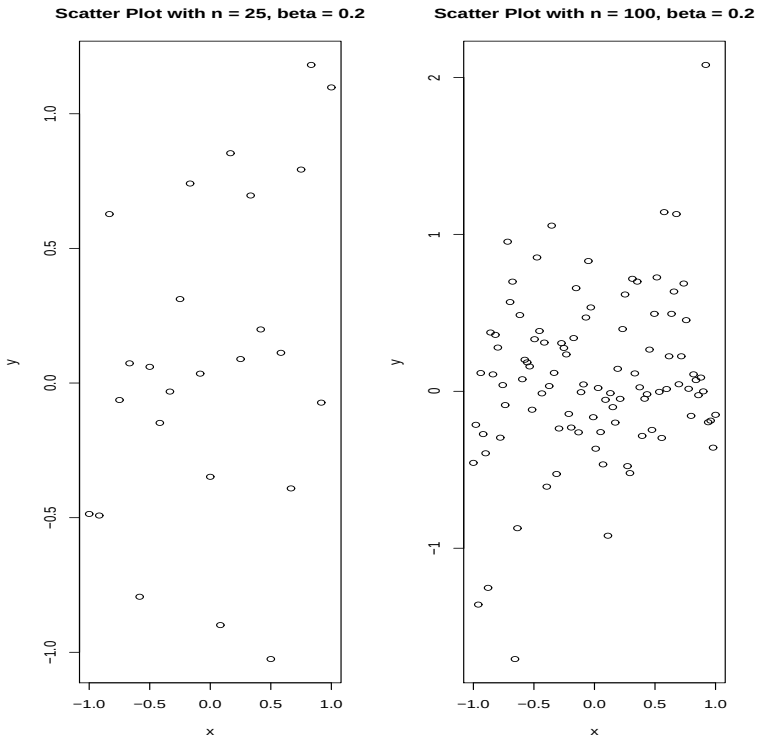


Figure 2: *Example Scatter Plots*

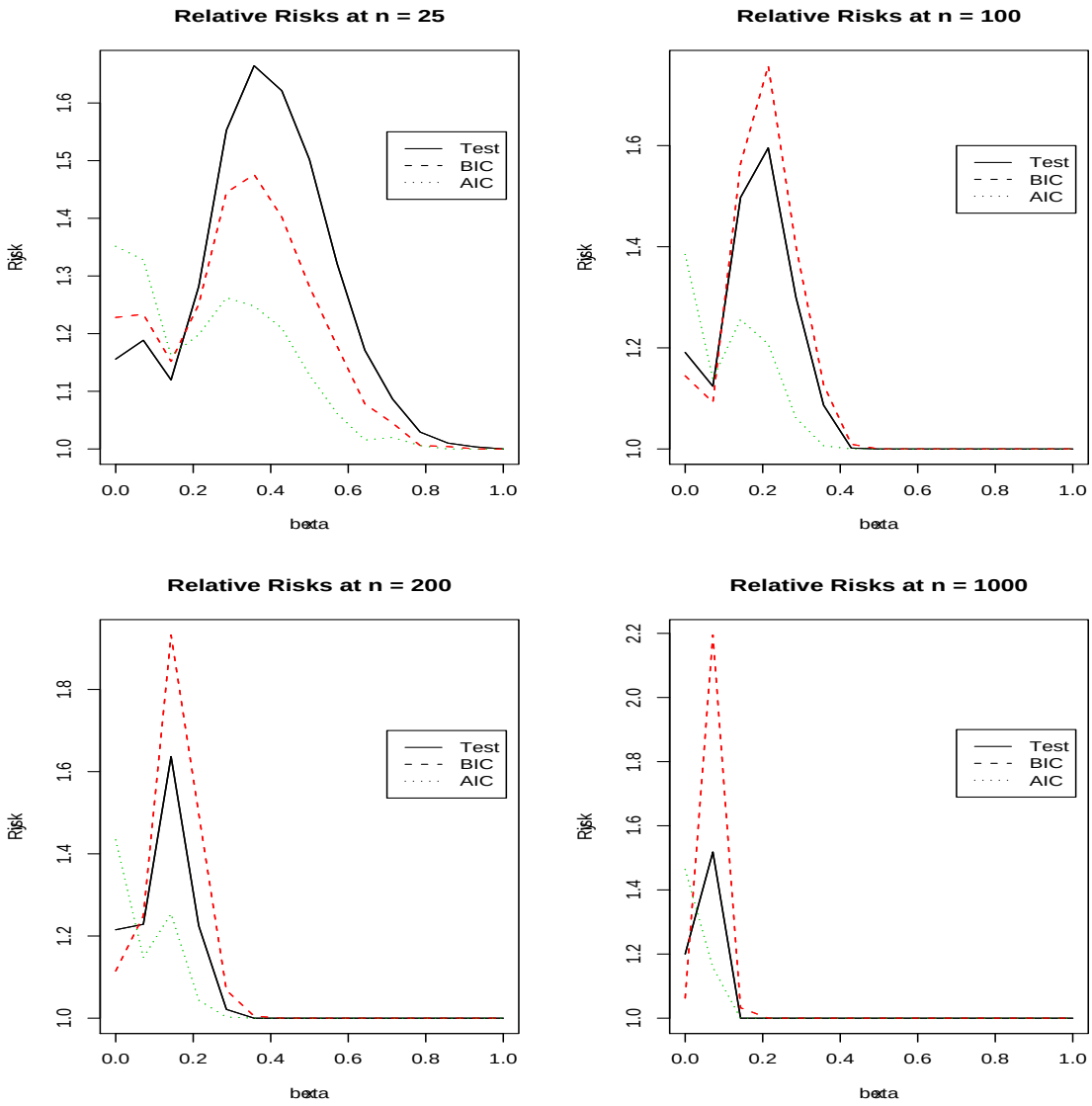


Figure 3: *Comparing the Model Selection Methods in Risk at 4 Sample Sizes*

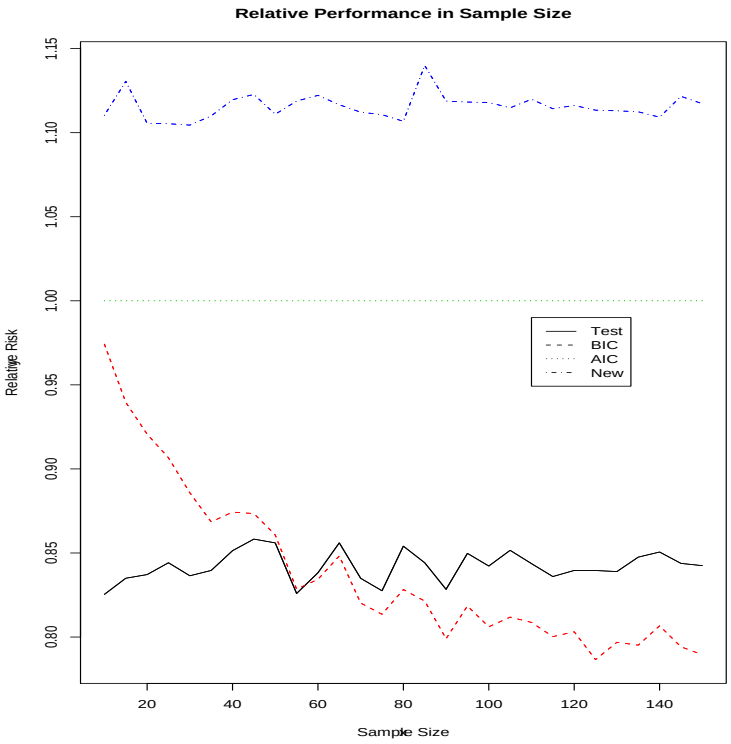


Figure 4: *Comparing the Model Selection Methods at* $\beta = 0$

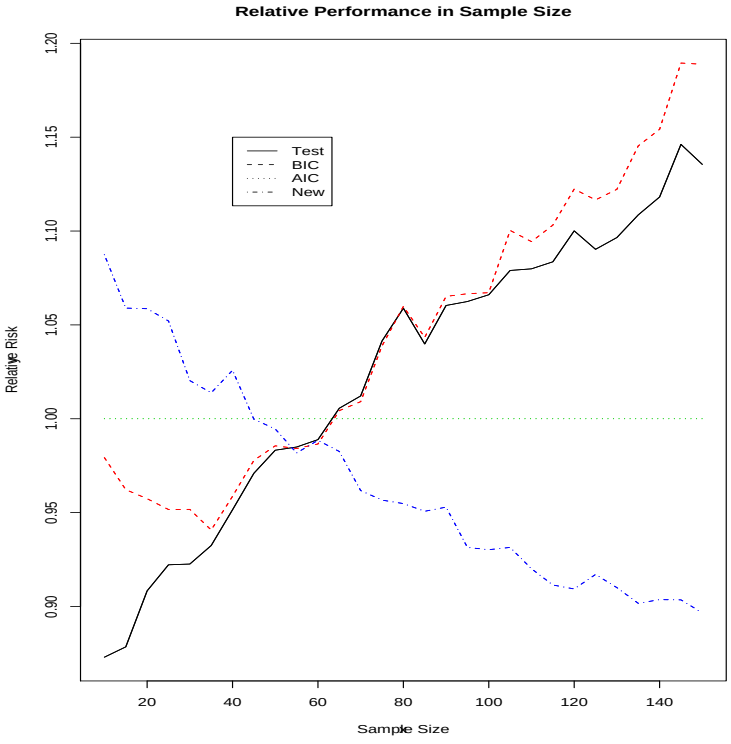


Figure 5: *Comparing the Model Selection Methods at* $\beta = 0.2$

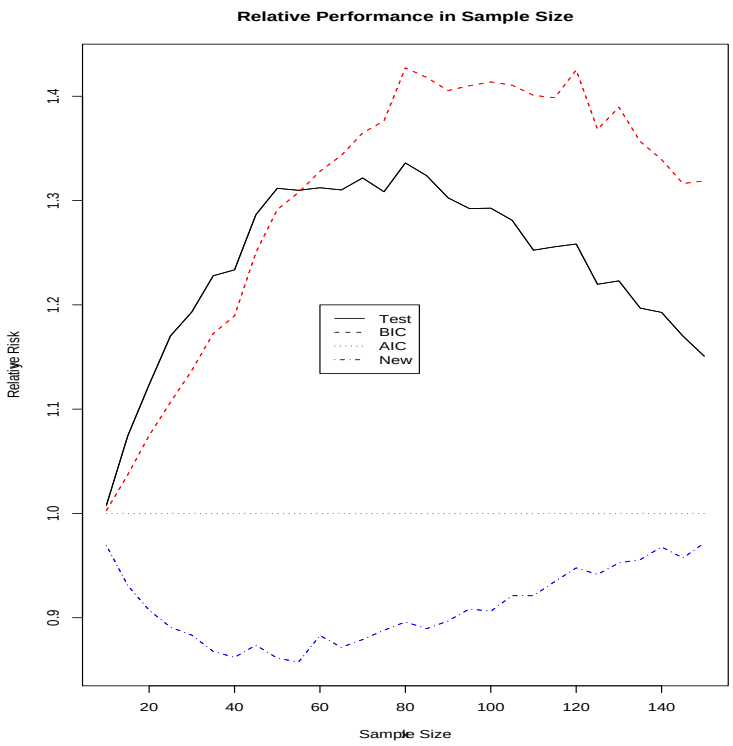


Figure 6: *Comparing the Model Selection Methods at* $\beta = 0.5$

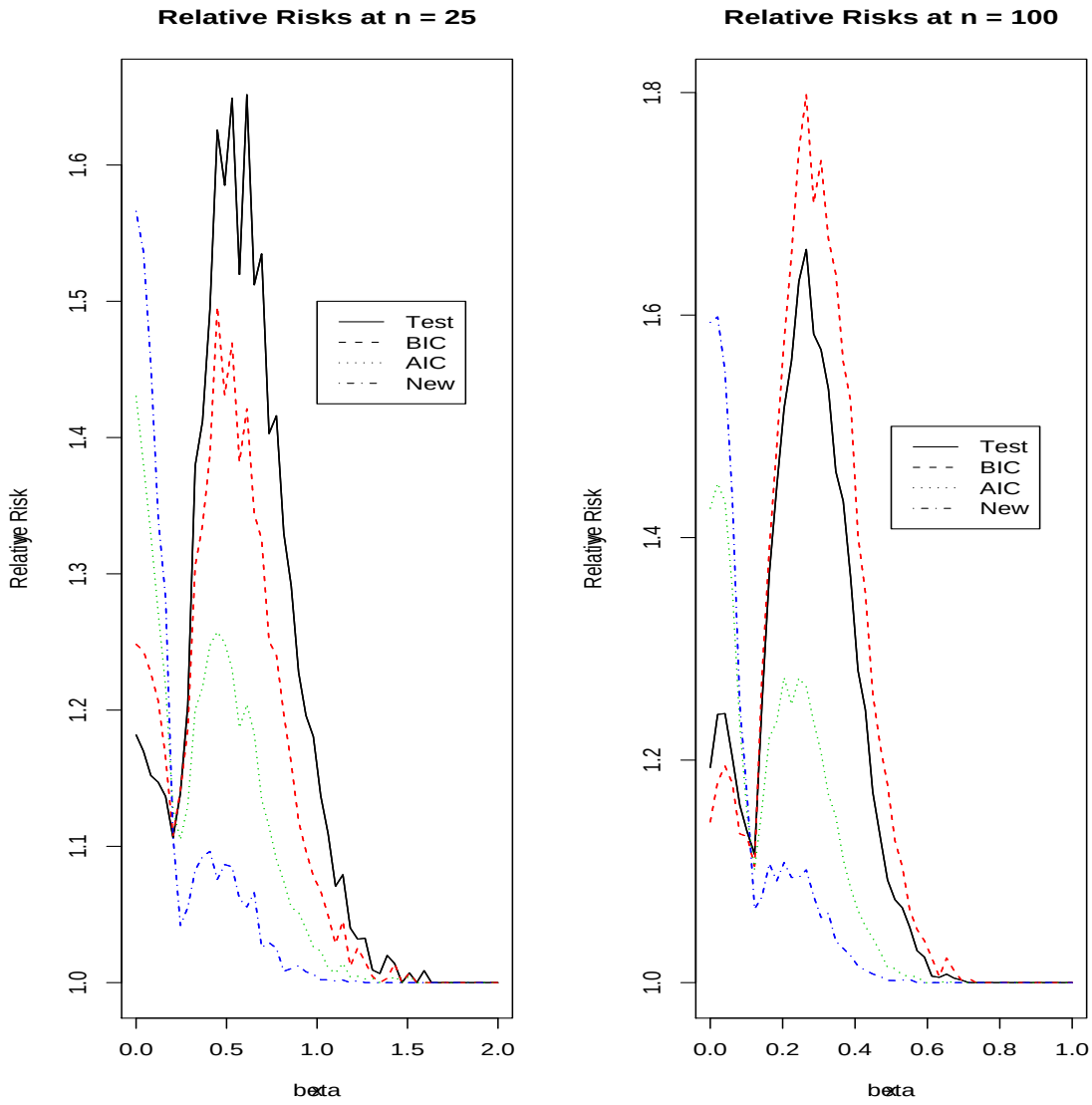


Figure 7: *Comparing the Model Selection Methods at* $n = 25, 100$ *under Laplace Error*

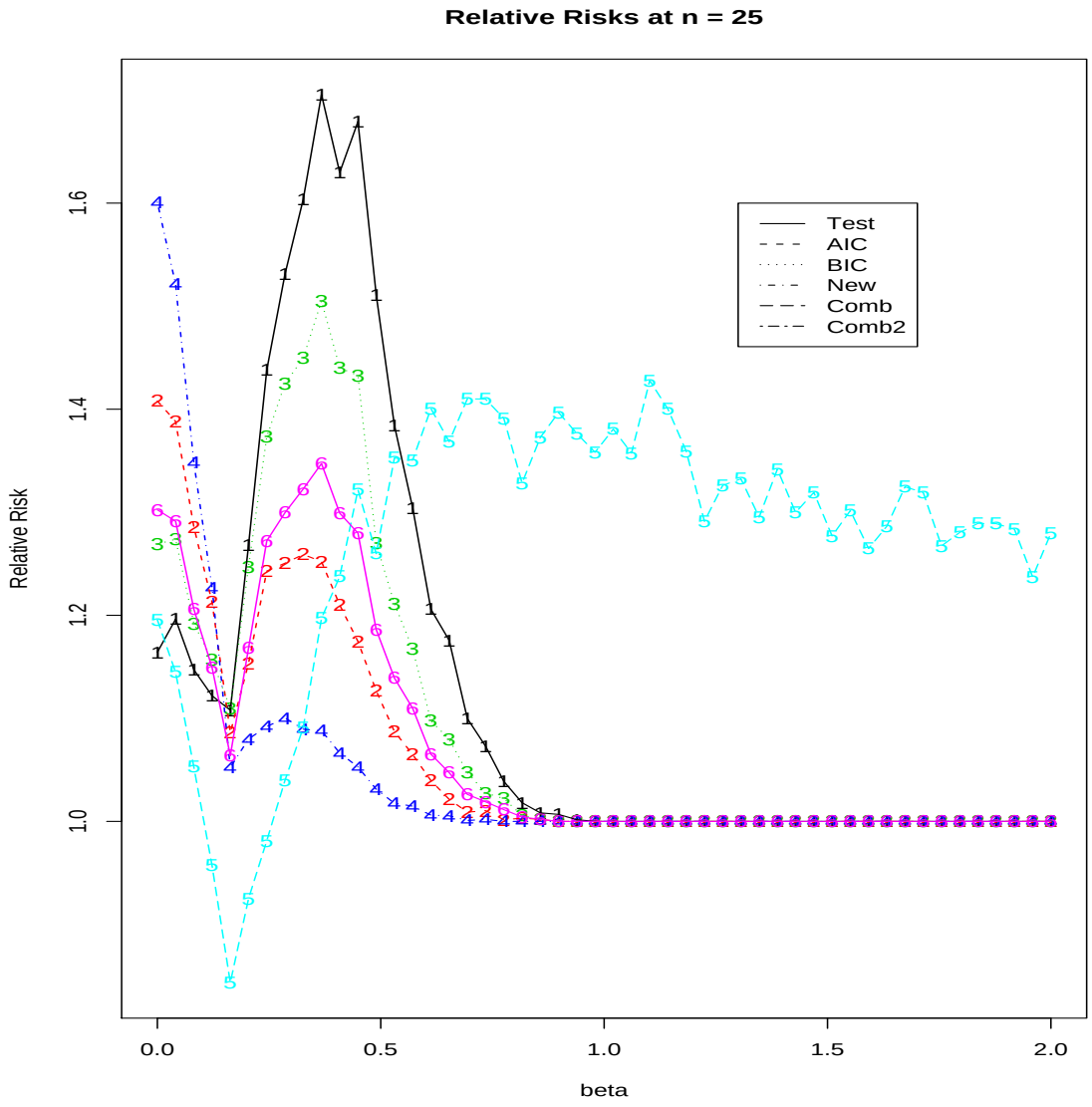


Figure 8: *Comparing the Selection/Combining Methods in β at n = 25*

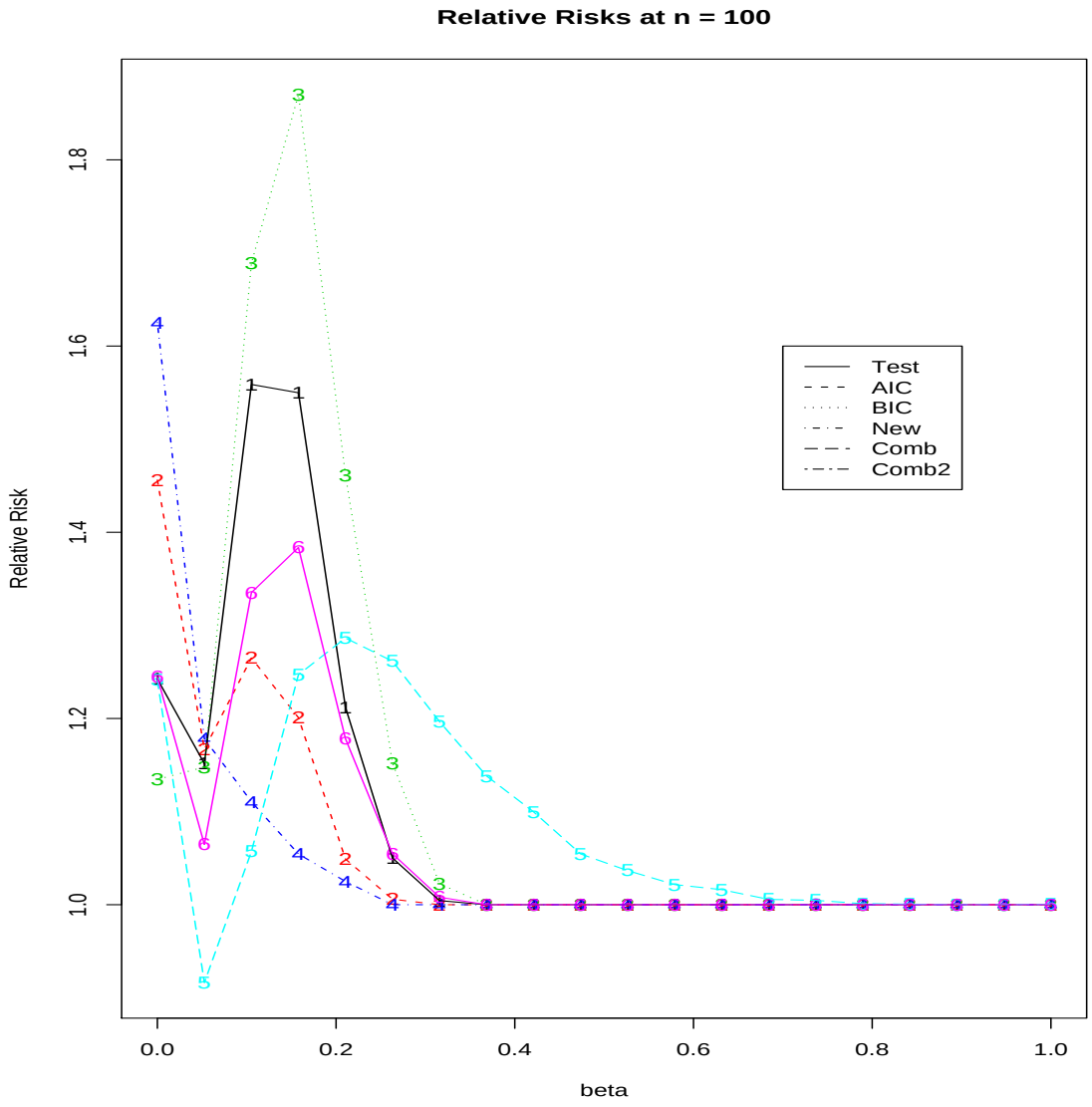


Figure 9: *Comparing the Selection/Combining Methods in β at n = 100*

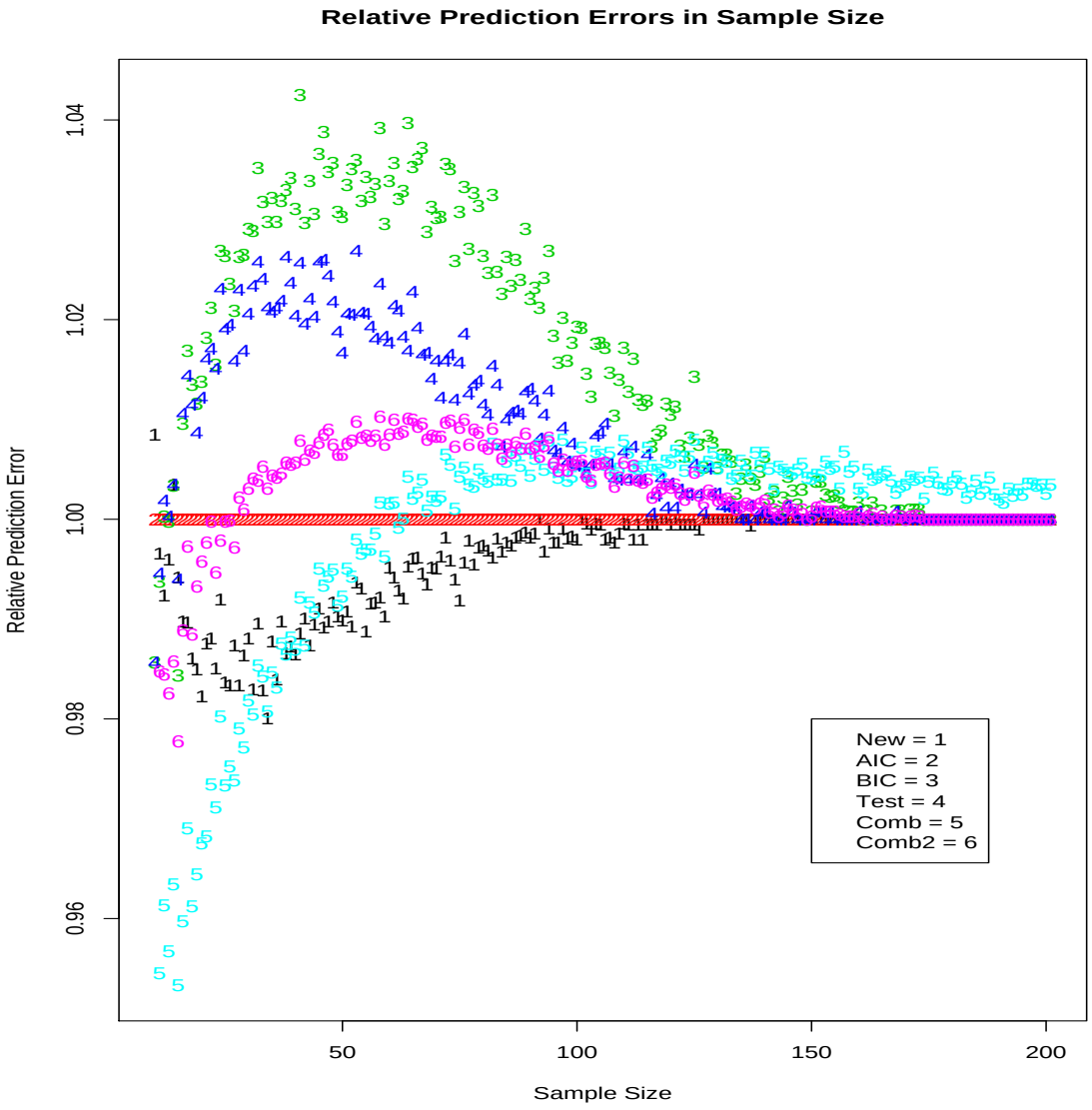


Figure 10: *Comparing the Selection/Combining Methods at Different Sample Sizes: Case 1*

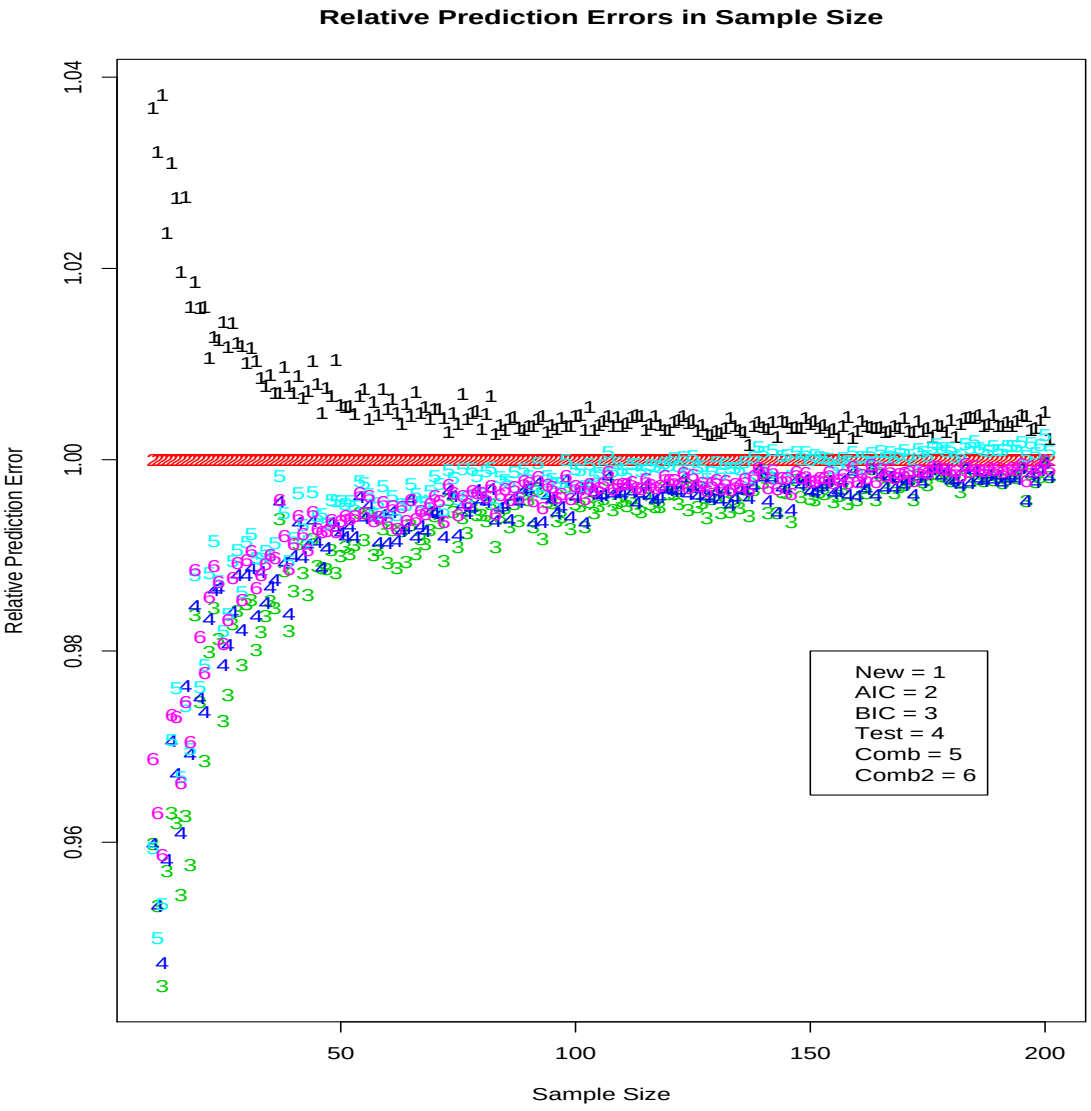


Figure 11: *Comparing the Selection/Combining Methods at Different Sample Sizes: Case 2*