Econometric Theory, **23**, 2007, 1–36. Printed in the United States of America.
DOI: 10.1017/S0266466607070016

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

YUHONG YANG
University of Minnesota

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 information criteria can be used for identifying the right model. We compare the performances of such methods both theoretically and empirically from different perspectives for more insight. The testing approach at the conventional size of 0.05, in spite of being the "standard approach," performs poorly in estimation. We also found that the frequently told story "the Bayesian information criterion (BIC) is good when the true model is finite-dimensional, and the Akaike information criterion (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 pointwise adaptation property of BIC and the minimax-rate adaptation property of AIC by any model selection method. When model selection methods have difficulty in selection, model combining is a better alternative in terms of estimation accuracy.

1. INTRODUCTION

Consider the simple linear regression model

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

where x is the design variable taking values in $(-\infty,\infty)$ and $\{\varepsilon_i\}$ are the errors assumed to be independent and normally distributed with unknown variance

This work was completed when the author was on leave from Iowa State University and was a New Direction Visiting Professor at the Institute for Mathematics and its Applications (IMA) at the University of Minnesota. The fundings from both IMA and ISU are greatly appreciated. The work was also partly supported by NSF CAREER grant DMS0094323.
The author thanks Xiaotong Shen and Hannes Leeb for very helpful discussions. The paper also benefited from the questions and comments from the participants at the statistics seminars the author gave at the University of Minnesota and Duke University.
The author is very grateful to the anonymous reviewers and the co-editor Benedikt Pötscher for carefully reading earlier versions of the paper, bringing my attention to several closely related previous and current results, and making many very valuable suggestions, which significantly improved the paper in both content and presentation. Address correspondence to Yuhong Yang, School of Statistics, University of Minnesota, 224 Church Street, Minneapolis, MN 55455; e-mail: yyang@stat.umn.edu.

© 2007 Cambridge University Press 0266-4666/07 $12.00

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

$\sigma^2 > 0$. For simplicity, we assume $\bar{x}_n = (1/n)\sum_1^n x_i = 0$. Our interest is point prediction of the response variable Y at a new value $x_0 \in (-\infty,\infty)$ of the design variable under the squared error loss.

Obviously, this is a simple statistical problem, even taught in 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 $\hat{f}_L(x_0) = \hat{\alpha} + \hat{\beta}x_0$. 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_0 at x_0, $E(Y_0 - \hat{f}_L(x_0))^2 = \sigma^2 + E(\hat{f}_L(x_0) - f(x_0))^2$. 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 $\beta = 0$.

In the rather large literature on model selection, usually optimistic stories have been told. Theoretical results showed that, with a consistent model selection method, popular asymptotic properties (e.g., asymptotic normality and asymptotic efficiency) of estimators of interest are often unaffected (see, e.g., Pötscher, 1991). In contrast to such asymptotic analyses, initiated by Bancroft (1944), relatively few researchers have seriously examined finite-sample behaviors of estimators based on a selected model. In particular, when a preliminary test is used to choose a model before estimating a parameter, the resulting bias, risk, and admissibility of such an estimator were studied (see Judge and Bock, 1978; Judge and Yancey, 1986; Giles and Giles, 1993; Magnus, 1999; Magnus and Durbin, 1999; Danilov and Magnus, 2004, and references therein). The finite-sample distributions of such post-model-selection estimators have been derived also (see Leeb, 2005; Leeb and Pötscher, 2003). It was also shown that an asymptotic optimality property of the Akaike information criterion (AIC) in a nonparametric setting does not really mean that AIC is better than the Bayesian information criterion (BIC) (Kabaila, 2002). More recently, Leeb and Pötscher (2005) provided additional results and discussions on potential inappropriateness of asymptotic analysis for point and interval estimators based on model selection.

In this work, by focusing on the simple setting for a clear understanding of certain essential issues, we compare model selection methods in both theoretical and numerical investigations in the hope of complementing and strengthening the earlier realistic depictions of model selection methods and their relative performances.

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

2. A MOTIVATING NUMERICAL EXAMPLE

2.1. Setup

From the instructions in a typical elementary statistics textbook, when it is unclear if the slope parameter is nonzero, one should perform a hypothesis test. With the almost sacred choice of size 0.05, the familiar t-test can be used as a pre-test before estimating f.

Note that here we are dealing with two models, the simple linear model in (1) and the null model, $Y_i = \alpha + \varepsilon_i$, $i = 1,2,\ldots,n$, with the same assumptions on the errors. For convenience, call them model 1 and model 0, respectively. Besides the preceding testing approach, we can use AIC or BIC to select one of them.

Let δ_A, δ_B, δ_T denote the procedures of estimating $f(x)$ based on the outcome of AIC, BIC, and the testing (at the conventional size 0.05), respectively (the details will be given later in Section 3). 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., $R_\beta(\delta;x_0;n) = E_\beta(f(x_0) - \hat{f}(x_0))^2$. The risk does not depend on α for $\hat{f}$ based on a typical model selection method because of $\bar{x}_n = 0$, and its dependence on σ^2 is suppressed because of our focus on the risk as a function of β (note that there is no need to consider different σ values: the change of the noise level simply rescales the risk plot).

Given that extensive work has 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 fairly clear picture here. We start with a numerical investigation.

Consider an equally spaced design between $[-1,1]$ with $\bar{x}_n = 0$ for prediction/estimation at $x_0 = 0.5$. Figure 1 gives the risk functions of δ_A, δ_B, δ_T relative to the smaller of the two risks resulting from models 0 and 1 for $\beta \geq 0$ at different sample sizes $n = 25$, 100, 200, and 1,000 with $\sigma = 0.5$. Note that at $n = 25$, δ_T is the least likely to select the larger model among the three methods, but when $n \geq 100$, δ_B takes that role.

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

2.2. What Do the Most Well-Known Results Say?

In one sentence, a summary of the best known theoretical works on pointwise performance of AIC and BIC is perhaps something like "BIC is good if the true model is finite-dimensional, and AIC is good if the true model is infinite-dimensional." This seems to be well supported. Indeed, when the true model is among the candidates, BIC is consistent in terms of selection (see, e.g., Nishii, 1984), and certain losses or risks of the estimator from BIC are asymptotically

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

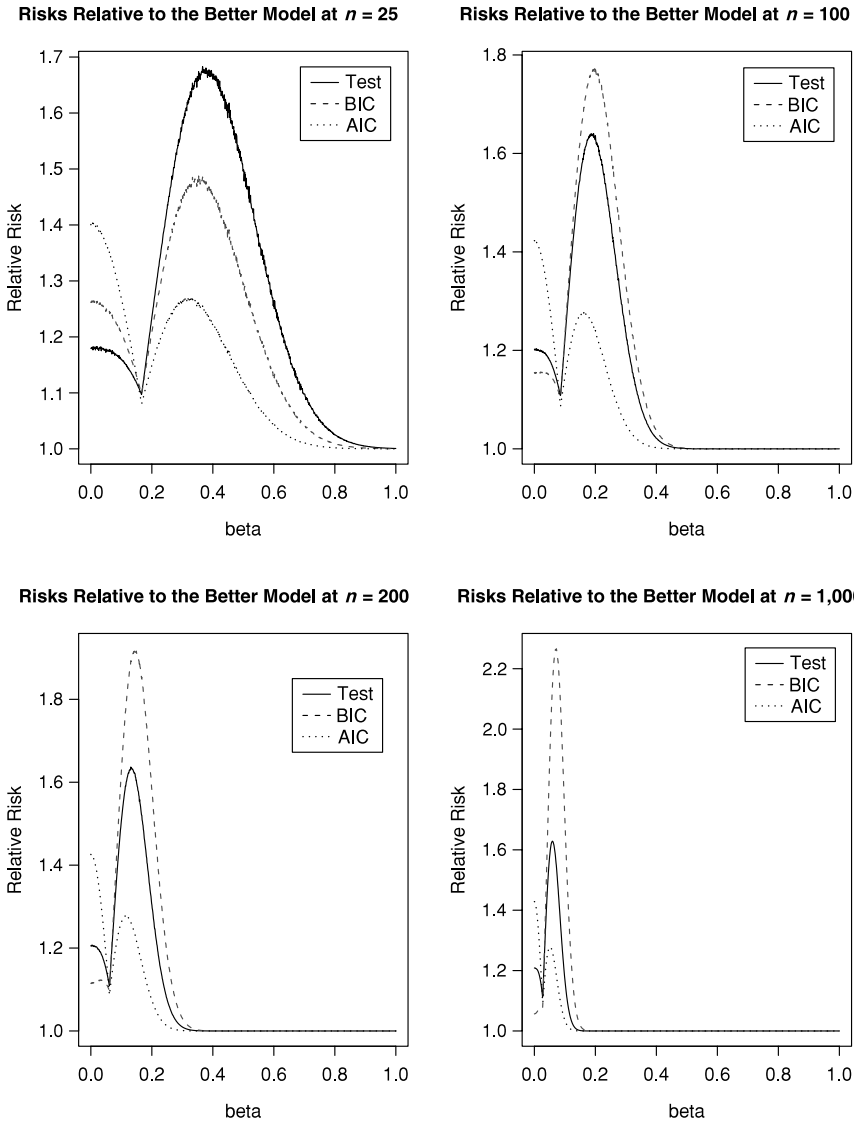


FIGURE 1. Comparing the model selection methods in risk at 4 sample sizes.

as small as they can be from the candidate models (Rissanen, 1986; Speed and Yu, 1993; Shao, 1997). When the true regression function is not in the candidate models, however, AIC has the so-called asymptotic efficiency property in estimating the regression function (see, e.g., Shibata, 1983; Li, 1987; Polyak and Tsybakov, 1991). Shao (1997) concluded that AIC and some other closely

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

related criteria (such as C_p and delete-1 cross validation) are "useful when there is no fixed-dimensional correct model," whereas BIC and delete-d cross validation with $d/n \to 1$ are "useful in the case there exist fixed-dimensional correct models."

In the debate on AIC and BIC, some researchers in favor of AIC argue strongly that there is no finite-dimensional "true model" (Burnham and Anderson, 2002). In defense of BIC, one can argue that in certain situations, some very simple models are clearly superior to nonparametric ones (imagine a simple linear regression with R^2 close to 1) and then it clearly makes sense, practically speaking, to view one of the parametric models as the "true model."

With the preceding understanding of the literature, I would have predicted the following results for the numerical experiment. (1) We expect three regions of β, β small, β large, and β in between, which correspond to BIC being better (while AIC is overfitting), AIC and BIC about the same, and AIC better (while BIC is underfitting). The first and third regions should shrink as n gets larger. (2) For this setting, from the literature, BIC should be favored in an overall sense.

Now let's examine Figure 1. Are there any surprises for you? To us, there were.

(1) Somewhat surprisingly, at $n = 25$, BIC loses to AIC at more β values and more severely compared to its shining moments.
(2) The worst case performance of BIC relative to AIC tends to get worse in n, even up to 1,000. In contrast, the worst case behavior of AIC relative to BIC seems to be confined.

What does the example tell us? 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 (as a result of being consistent in selection). Therefore, it seems fair to say that unless one has a strong reason to believe that β is around zero, the penalty of BIC is too large and thus is too conservative for yielding a good predictive performance. Considering the whole picture, even when n is large, the aforementioned optimality properties of BIC do not seem to reflect the reality well. The main reason, as will be seen later, is that the optimality properties are pointwise in nature and cannot be uniform in β.

For the testing approach, it also performs poorly, indicating that estimation/prediction is different from assessing if model 0 or model 1 is the true model at the size 0.05. We emphasize that the problem is not due to testing per se but is caused by the choice of significance level.

It should be pointed out that there have been a few underappreciated results in the literature on model selection that provide perspectives other than those given earlier. Indeed, the behavior that the worst case risk of BIC becomes increasingly worse compared to AIC was discovered by Foster and George

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

(1994). Somewhat surprisingly, even though the risk inflation criterion (RIC) proposed in their paper is quite well known, as far as we know, this suboptimality of BIC has rarely been discussed. Kabaila (2002) pointed out that the pointwise asymptotic efficiency of AIC should not be interpreted as that AIC beats BIC when the sample size is large. Further concerns over misinterpretations of pointwise asymptotic results on model selection are provided in Leeb and Pötscher (2005). In particular, they emphasized that consistent model selection procedures actually do not allow one to apply the standard asymptotic distributions obtained without model selection. It is also useful to note that in real world statistical applications, practical aspects such as parsimony (see, e.g., Zhang, 1997) rather than asymptotic properties are often taken into consideration.

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

1. For the purpose of estimation/prediction, we compare AIC and BIC from different theoretical angles to understand their differences. The theoretical results and the numerical 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 method, however sophisticated, can share the essential strengths of AIC and BIC in risk.
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 perspective, which hopefully will help the reader to understand the complex issues involved in model selection.

3. THEORETICAL COMPARISONS BETWEEN AIC AND BIC

3.1. A Brief Review of Model Selection Methods

The well-known t-test of size 0.05 is reject H_0 if

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

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

where $s^2 = [1/(n-2)]\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ 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 make the following estimation:

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

Let $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ be the maximum likelihood estimator (MLE) of σ^2 under model 1. Simple calculations show that BIC (Schwarz, 1978) selects model 1 when

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

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

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

Similarly, AIC (Akaike, 1973) selects model 1 when

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

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

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

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

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. Thus in our context, AIC and BIC in a sense are also testing procedures: AIC has a size approaching 0.16 and BIC has a size approaching zero as $n \to \infty$. Clearly, $t_{n-2,0.025}$ 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, whereas 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 $\beta \neq 0$), whereas AIC and testing with a fixed size cannot avoid a nonvanishing probability of selecting the wrong model when $\beta = 0$.

In the rest of 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.

3.2. A Pointwise Adaptation Result

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

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

and

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

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

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 $f \in \mathcal{F}_0 \cup \mathcal{F}_1$ and all $x_0 \in (-\infty, \infty)$ we have*

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

as $n \to \infty$.

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

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 $n \to \infty$. Then 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, e.g., Nishii (1984), Shao (1997), and Leeb and Pötscher (2005). 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 nonvanishing 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 nonvanishing probability of underfitting. See, e.g., Shao (1997) for a more general result on underfitting probability.
5. The condition that $\sum x_i^2$ is of order n (i.e., $\sum x_i^2/n$ is bounded between two positive constants) rules out highly irregular cases such as degeneration of the design points to a singleton.

Proof of Theorem 1. Consider $x_0 \neq 0$. 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 $A_\delta = \{|\hat{\beta}| \geq a_n s/\sqrt{\sum_{i=1}^n x_i^2}\}$. Then

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

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 their risks are σ^2/n and $(\sigma^2/n) + x_0^2 E_\beta(\hat{\beta} - \beta)^2 = (\sigma^2/n) + (\sigma^2 x_0^2/\sum_{i=1}^n x_i^2)$, respectively. For the model selection criterion to be

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

pointwise-risk adaptive, we must have $nE_{\beta=0}(\hat{\beta}I_{A_\delta} - \beta)^2 \to 0$ (for $x_0^2 > 0$). Because

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

and because $\beta = 0$, $\sqrt{\sum_{i=1}^n x_i^2}\hat{\beta}/\sigma$ 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. Because $\sqrt{\sum_{i=1}^n x_i^2}\hat{\beta}/\sigma$ has a standard normal distribution,

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

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

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 their risks are $(\sigma^2/n) + x_0^2\beta^2$ and $(\sigma^2/n) + x_0^2 E_\beta(\hat{\beta} - \beta)^2 = (\sigma^2/n) + (\sigma^2 x_0^2/\sum_{i=1}^n x_i^2)$, respectively. Thus to show that the model selection criterion is pointwise-risk adaptive, we need to show that

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

But the quantity on the left-hand side is equal to

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

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

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

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

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

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

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

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

3.3. Worst Case Performance

From Theorem 1, when $\beta = 0$, BIC beats AIC, and when $\beta \neq 0$ AIC and BIC are asymptotically equivalent in pointwise risk as $n \to \infty$. Together with the fact that BIC is consistent, it gives one a strong impression that BIC should be used for a parametric problem.

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 in f. An implication is that for a data set where the distinction between the two models is not clear because of small n or large noise (that is exactly where we need a good model selection rule most), it is especially uncertain whether the asymptotics are useful or not. 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

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

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 $j = 0,1$ and for all $x_0 \in (-\infty,\infty)$, we have*

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

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*

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

selection rule δ of the form (2) is minimax-rate optimal if and only if a_n is bounded from above by a positive constant.

The theorem implies that in the least favorable case, AIC can only be finitely many times worse than BIC but the reverse is not true. This may be a little surprising. One may naturally think that because BIC is consistent, it will eventually select the true model. Together with its pointwise-risk adaptation property, it looks as if 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 that BIC do at least as well as that.

A result similar 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. Recently, Leeb and Pötscher (2005) derived a risk expression under a slightly different loss and obtained the same conclusion without the normality assumption on the errors.

Proof of Theorem 2. Note that the minimax rates of convergence for the two classes (i.e., linear functions and constant functions) are both n^{-1}. 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

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

which is sufficient to conclude that when $a_n \to \infty$, the model selection rule is minimax-rate suboptimal. Note that the left-hand side of the preceding expression is equal to

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

Observe that

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

Take $\beta_n = \min(\sqrt{a_n/n}, c)$. It is easy to show that for any constant $c' > 1$, $P(s/\sigma \ge c' \text{ or } s/\sigma \le 1/c')$ converges to zero exponentially fast in n (i.e., $P(s/\sigma \ge c' \text{ or } s/\sigma \le 1/c') = o(e^{-bn})$ for some $b > 0$; see, e.g., Yang, 1999,

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

pp. 490–491). Note also that $\sqrt{n}\beta_n \le \sqrt{a_n} = o(a_n)$ and thus $\sqrt{\sum_{i=1}^n x_i^2}\beta/\sigma = o(a_n)$. Then $P_{\beta_n}(A_\delta^c)$ is asymptotically equivalent to

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

which approaches 1. Because $n\beta_n^2 \to \infty$, we have $n\beta_n^2 P_{\beta_n}(A_\delta^c) \to \infty$. 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

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

Note that $E_\beta n(\hat{\beta} - \beta)^2 I_{A_\delta} \le nE_\beta(\hat{\beta} - \beta)^2 = n\sigma^2/\sum_{i=1}^n x_i^2$, which is bounded from above. Consequently, it suffices to show that $n \sup_{\beta \ge 0} \beta^2 P_\beta(A_\delta^c)$ is bounded from above. Observe that

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

$$\subset \left\{ \frac{a_n s}{\sigma} \ge \frac{\sqrt{\sum_{i=1}^n x_i^2}\beta}{2\sigma} \right\} \cup \left\{ \frac{\sqrt{\sum_{i=1}^n x_i^2}(\hat{\beta} - \beta)}{\sigma} \le -\frac{\sqrt{\sum_{i=1}^n x_i^2}\beta}{2\sigma} \right\}.$$

Thus

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

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

3.4. 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 viewpoint?

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

In general, in our view, pointwise asymptotic results provide rather little guidance for real applications. For example, a theoretically elegant, universally consistent estimator in nonparametric regression with independent and identically distributed (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., the 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). In any case, in a typical application where model selection is clearly a nontrivial issue, perhaps more often than not, we are not in a situation where the pointwise asymptotic behavior has already begun operating (or at least it is not evident that this is the case). To address this concern, properties such as minimaxity can be very helpful for better guiding the model selection practice.

Another comment is that whenever possible, we should try to understand/assess whether the difference in estimation/prediction that 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.

Nonuniformity of convergence of estimators based on model selection has been investigated by Leeb and Pötscher in several papers (see, e.g., Leeb and Pötscher, 2003, 2005) in different directions (such as estimating the distribution of estimators after model selection).

3.5. Can the Strengths of AIC and BIC Be Shared?

We have now seen the very distinct properties of AIC and BIC. Can we have a more sophisticated model selection method that can automatically behave like the better one of them in risk? It turns out that the contrast between pointwise and uniform convergence again shows up in answering this question. Indeed, as we show subsequently, in a pointwise sense, the strengths of AIC and BIC in risk can be combined; but once minimax rate is also sought, the picture is totally different.

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

3.5.1. Adaptive Pointwise Asymptotic Loss Efficiency. One property that separates AIC and BIC is the condition under which a pointwise asymptotic loss efficiency (to be defined) holds. Shao (1997) showed that AIC (or a similar criterion) is pointwise asymptotically loss efficient when there is no fixed-dimensional correct model and BIC (or a similar criterion) is pointwise 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 pointwise asymptotic loss efficiency under both of the aforementioned conditions. This brings up an interesting question: Can we construct an adaptive model selection rule that does achieve the goal of pointwise 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 $x_i = (x_{i1}, \ldots, x_{id})$ for some fixed $d \geq 1$ and the errors are assumed to be independent and 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 $f_n = (f(x_1), f(x_2), \ldots, f(x_n))'$ and $\underline{Y} = (Y_1, \ldots, Y_n)'$ 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 pointwise asymptotically loss efficient if*

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

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

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

regression (under conditions on the design matrices), see Rao and Wu (1989). Related results are in Hannan and Quinn (1979), Geweke and Meese (1981), Pötscher (1989), and others.

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

Assumption 2. AIC is pointwise 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 $(1/n)\|f_n - M_k f_n\|_n^2 + k/n$ over $k \geq 1$. Note that k_n^* provides the best trade-off between the approximation error $(1/n)\|f_n - M_k f_n\|_n^2$ and the estimation error k/n (recall that σ^2 is assumed to be 1). We require the approximation errors to behave regularly in the following sense.

Assumption 3.

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

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

for some constant c, then

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

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

$$\frac{\frac{1}{n}\|f_n - M_k f_n\|_n^2}{e^{-e^k}} \leq \bar{c} < \infty$$

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

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

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

For interesting nonparametric regressions, it is usually the case that $\inf_k((1/n)\|f_n - M_k f_n\|_n^2 + (k/n))$ 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

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

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* $\hat{\hat{k}}_n$,

(i) when (at least) one of the candidate models is correct, under Assumption 1, $\hat{\hat{k}}_n$ *is pointwise asymptotically loss efficient;*
(ii) when none of the candidate models is correct, under Assumptions 2 and 3, $\hat{\hat{k}}_n$ *is pointwise asymptotically loss efficient.*

The proposition states that under some reasonable conditions, an adaptive model selection rule can be pointwise asymptotically loss efficient for both parametric and nonparametric situations. However, because the result is in terms of pointwise asymptotic behavior, it gives an overly optimistic picture on capability of model selection when viewed from a more realistic perspective, as will be seen.

Proof of Proposition 1. To prove the result, it is sufficient to show that

(a) when at least one of the candidate models is correct, $\hat{\hat{k}}_n$ will eventually equal to $\hat{k}_{BIC,n}$ almost surely;
(b) when the first condition of Assumption 3 holds, we have $P(\hat{\hat{k}}_n = \hat{k}_{AIC,n}) \to 1$;
(c) when the second condition of Assumption 3 holds, we have

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

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

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

Suppose that Assumption 3(1) holds. It suffices to show that $P(\hat{k}_{BIC,n} = \hat{k}_{BIC,b_n}) \to 0$. Note that at the sample size b_n, $\hat{k}_{BIC,b_n} \le b_n$. Thus it is sufficient to prove that $\sum_{k \le b_n} P(\hat{k}_{BIC,n} = k) \to 0$. Let $k^*_{BIC,n}$ be the model that minimizes $(1/n)\|f_n - M_k f_n\|_n^2 + [k(\log n - 1)/n]$ over $k \ge 1$. 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 $k \le n$). Let $e_n = (\varepsilon_1, \ldots, \varepsilon_n)$ 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 because of its constant presence and independence of k. Under Assumption 3(1), for $k \le b_n$, we have

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

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 $k = k^*_{BIC,n}$, because $k^*_{BIC,n} \to \infty$ as $n \to \infty$ for the nonparametric case, we have that $2e_n' A_{k^*_{BIC,n}} f_n + (k^*_{BIC,n} - e_n' M_{k^*_{BIC,n}} e_n)$ 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 $k \le b_n$ and (7), we know that asymptotically $Crit(k) > Crit(k^*_{BIC,n})$ with increasingly high probability for $k \le b_n$. 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 preceding argument rigorous (see, e.g., the analysis in Yang, 1999, 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 $e^{-e^{k^*_{BIC,n}}} = o(k^*_{BIC,n}/n)$, and the same holds for $k^*_{AIC,n}$. Consequently the ratio

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

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 $(1/n)\|f_n - M_{k^*_{BIC,n}} f_n\|_n^2 = o(k^*_{BIC,n}/n)$, 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

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

to 1 under Assumption 3(2). One can also show that $\|f_n - \hat{f}_{\hat{k}_{AIC,n}}\|_n^2 / [\|f_n - M_{k^*_{AIC,n}} f_n\|_n^2 + (k^*_{AIC,n}/n)] \to 1$ 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.

3.5.2. No Model Selection Rule Can Be Really Adaptive between AIC and BIC. Proposition 1 gives some hope of resolving the competition between AIC and BIC completely by sharing their strengths (it indeed succeeded in one aspect). Note that the adaptive pointwise 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 the direction of combining strengths of AIC and BIC with positive results, including Barron, Yang, and Yu (1994) by applying the minimum description length criterion (MDL) (Rissanen, 1978) in a novel way to yield a penalty of AIC type or BIC type, whichever is better in risk; Hansen and Yu (1999) by taking a different application of MDL based on a pre-test; George and Foster (2000) by empirical Bayes approaches to have an adaptive penalty term; Rao and Tibshirani (1997) by adaptively choosing the penalty constant based on cross validation; Shen and Ye (2002) by using a generalized degree of freedom in the same direction of adaptively selecting the penalty constant.

Despite the positive findings, it turns out that the most essential risk properties 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. Furthermore, 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* $(-\log q_n)/n$.

Remark. The order $(-\log q_n)/n$ is achieved by a model selection rule of the form (2) with a_n chosen such that the false selection probability (i.e., the prob-

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

ability of not selecting the true model) under $\beta = 0$ is of the specified order q_n. From this angle, other model selection criteria cannot improve over those of the form (2) in terms of the worst case risk rate.

The theorem states that the main strengths of AIC and BIC in terms of risk 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 (2005) 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). Relaxing the normality assumption, Leeb and Pötscher (2005, App. C) gave a similar result. Here in Theorem 3 the contrast is made between pointwise-risk adaptation and minimax-rate adaptation, and the result further characterizes the price one must pay in terms of the worst case performance for having the false selection probability go to zero. For a model selection rule of the form (2) with $a_n \to \infty$, the worst case risk is at order a_n^2/n. Note that fitting the larger model without any selection gives a risk of order $1/n$.

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 that δ is not minimax-rate optimal, it suffices to show

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

Because δ is pointwise-risk adaptive, we have that when $\beta = 0$, for each $x_0 \in (-\infty,\infty)$,

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

and thus

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

Consequently, because the convergence holds for every x_0, we must have $E_{\beta=0}(\sqrt{n}\hat{\beta})^2 I_{A_\delta} \to 0$ and $nE_{\beta=0}(\hat{\alpha} - \alpha)\hat{\beta} I_{A_n} \to 0$. Because 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 $n \to \infty$.

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

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

$$Y_i = \beta x_i + \varepsilon_i, \qquad i = 1,2,\ldots,n,$$

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 the Neyman–Pearson lemma, that any test with the probability of type I error going to zero necessarily must have $\sup_{|\beta|\le c} n\beta^2 P_\beta(A_n^c) \to \infty$. Let $f(y_1,\ldots,y_n;\beta)$ denote the joint probability density function of $(Y_1,\ldots,Y_n)$. Note that for $\beta_1 \ge \beta_0 \ge 0$,

$$\frac{f(y_1,\ldots,y_n;\beta_1)}{f(y_1,\ldots,y_n;\beta_0)} = \exp\left(\frac{1}{2}\sum_{i=1}^{n}((y_i-\beta_0 x_i)^2 - (y_i-\beta_1 x_i)^2)\right)$$

$$= \exp\left((\beta_1-\beta_0)\sum_{i=1}^{n} x_i y_i + (\beta_0-\beta_1)\sum x_i^2\right).$$

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 $P_{\beta=0}(\sum_{i=1}^n x_i y_i \ge d_n) = P_{\beta=0}(A_n)$. Let $A_{n,*}$ denote the event of $\sum_{i=1}^n x_i y_i \ge d_n$. By the UMP property of $A_{n,*}$, we have for all $\beta_0 > 0$, $P_\beta(A_{n,*}) \ge P_\beta(A_n)$. Consequently,

$$\sup_{|\beta|\le c} n\beta^2 P_\beta(A_n^c) \ge \sup_{|\beta|\le c} n\beta^2 P_\beta(A_{n,*}^c).$$

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

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

and for $\beta > 0$

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

Because $P_{\beta=0}(\sum_{i=1}^n x_i Y_i \ge d_n) = P_{\beta=0}(A_n) \to 0$, we must have $d_n/\sqrt{n} \to \infty$ (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 $\beta_n = d_n/(2\sum x_i^2)$, we have

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

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

Because the last probability goes to 1 and $nd_n^2/(4(\sum x_i^2)^2) \to \infty$, we have $\sup_{|\beta|\le c} n\beta^2 P_\beta(A_{n,*}^c) \to \infty$.

Now we prove the second part of the theorem. From the preceding discussion, we just need to show that $nd_n^2/(\sum x_i^2)^2$ is at least of order $-\log q_n$, where $q_n = P_{\beta=0}(A_n)$. Note that $d_n/\sqrt{\sum x_i^2} = \Phi^{-1}(1-q_n)$, where Φ denotes the standard normal cumulative distribution function. Let ϕ denote the standard normal probability density function. Because for $t > 0$,

$$\left(\frac{1}{t} - \frac{1}{t^3}\right)\phi(t) \le 1 - \Phi(t) \le \frac{1}{t}\,\phi(t)$$

(see e.g., Pollard, 2002, p. 317), it is easily shown that $d_n/\sqrt{\sum x_i^2}$ is of order $\sqrt{-\log q_n}$. Consequently, $nd_n^2/(\sum x_i^2)^2$ is right at order $-\log q_n$. This completes the proof of Theorem 3. ■

3.6. Subjectivity of the Choice of the Penalty in Model Selection

Note that under model 1, the least squares estimator gives $\hat{f}_L(x) = \hat{\alpha} + \hat{\beta}x$ for estimating $f(x)$. It has a constant risk $(\sigma^2/n) + x^2\sigma^2/\sum_{i=1}^n x_i^2$ under the squared error loss. Assume that σ^2 is known (or bounded from above by a known constant), so that the estimator $\hat{f}_L(x)$ is in fact a minimax estimator. Because 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 because of 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 because 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. 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.

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

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 $|\hat{\beta}| > a_n/\sqrt{n}$. Let $\delta^{\{a_n\}}$ denote this selection rule.

Proposition 2. *Let $\pi_0 > 0$ and $\pi_1 > 0$ and $\pi_0 + \pi_1 = 1$.*

(i) 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 ∞.

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

From the first part of the result, 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. The second part of Proposition 2 states 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, Sect. 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 vari-

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

ance differently from the squared error loss. The second part of Proposition 2 justifies a compromising criterion from our minimax point of view.

The proof of Proposition 2 is omitted.

4. AN ALTERNATIVE PENALTY CONSTANT FOR PREDICTION

In this section, we consider 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 $(\sigma^2/n) + x_0^2\beta^2$ and $(\sigma^2/n) + \sigma^2 x_0^2/\sum_{i=1}^n x_i^2$, respectively. Thus model 1 is better when $\beta^2 > \sigma^2/\sum_{i=1}^n x_i^2$, i.e.,

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

Because β and σ are unknown, the preceding 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. Then a sensible model selection rule is to choose model 1 when $|\hat{\beta}|/(s/\sqrt{\sum_{i=1}^n x_i^2}) > 1$. This approach was studied in a more general regression framework in Toro-Vizcarrondo and Wallace (1968) and Wallace (1972) (see also Judge and Bock, 1978, pp. 53–55). Recognizing that the ratio $\hat{\beta}/(s/\sqrt{\sum_{i=1}^n x_i^2})$ is a biased estimator of $\beta/(\sigma/\sqrt{\sum_{i=1}^n x_i^2})$, we may consider a modification when n is small. Because the numerator and the denominator are independent, the expectation is $\beta\sqrt{\sum_{i=1}^n x_i^2}E(1/s)$. Observing that a multiple of s^2/σ^2 has a chi-square distribution, we can easily find that $E(\sigma/s) = \sqrt{(n-2)/2}[(\Gamma(n-3)/2)/(\Gamma(n-2)/2)]$. With this modification, we come to the criterion that selects model 1 when

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

We will denote this new selection rule δ_N. This criterion is even more aggressive in choosing model 1 than AIC (as $n \to \infty$, the cutoff value $\sqrt{(n-2)/2}[\Gamma((n-3)/2)/\Gamma((n-2)/2)]$ approaches 1 by Stirling's formula).

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

5. NUMERICAL RESULTS FOR COMPARING THE MODEL SELECTION METHODS

In this section, we present more numerical comparisons of the model selection methods.

First note that for a model selection rule of the form (2), by conditioning on s first, we have

$$R_\beta(\delta;x_0;n) = \frac{\sigma^2}{n} + x_0^2 E_\beta E_\beta\left(\left(\hat{\beta} I_{\left\{|\hat{\beta}|\Big/\left(s\Big/\sqrt{\sum_{i=1}^n x_i^2}\right) \geq a_n\right\}} - \beta\right)^2 s\right).$$

Because $\hat{\beta}$ and s are independent, we have

$$\begin{aligned}
&E_\beta\left(\left(\hat{\beta} I_{\left\{|\hat{\beta}|\Big/\left(s\Big/\sqrt{\sum_{i=1}^n x_i^2}\right) \geq a_n\right\}} - \beta\right)^2 \Bigg| s\right) \\
&\quad = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\left(\frac{a_n s - \beta\sqrt{\sum_{i=1}^n x_i^2}}{\sigma}\phi\left(\frac{a_n s - \beta\sqrt{\sum_{i=1}^n x_i^2}}{\sigma}\right)\right. \\
&\qquad\qquad + 1 - \Phi\left(\frac{a_n s - \beta\sqrt{\sum_{i=1}^n x_i^2}}{\sigma}\right) + \frac{a_n s + \beta\sqrt{\sum_{i=1}^n x_i^2}}{\sigma} \\
&\qquad\qquad \left.\times \phi\left(\frac{a_n s + \beta\sqrt{\sum_{i=1}^n x_i^2}}{\sigma}\right) + \Phi\left(\frac{-a_n s - \beta\sqrt{\sum_{i=1}^n x_i^2}}{\sigma}\right)\right) \\
&\qquad + \beta^2\left(\Phi\left(\frac{a_n s - \beta\sqrt{\sum_{i=1}^n x_i^2}}{\sigma}\right) - \Phi\left(\frac{-a_n s - \beta\sqrt{\sum_{i=1}^n x_i^2}}{\sigma}\right)\right).
\end{aligned}$$

A similar expression is in Leeb and Pötscher (2005, Sect. 2.1). Thus as part of the computation of $R_\beta(\delta;x_0;n)$ for a model selection rule of the form (2), we can independently generate s^2 (which basically has a chi-square distribution) a large number of times (we chose 2,000 in our computation) and then average the resulting values of the preceding expression. This is what we did when all the involved model selection rules are of the form (2) (Figures 1–4).

Here our interest is in knowing 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. The asymptotic results in Section 3 have already told us that when

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

n is very large, BIC is better than AIC 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 that follows, we consider sample sizes from 10 to 1,000. 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 2–4. 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

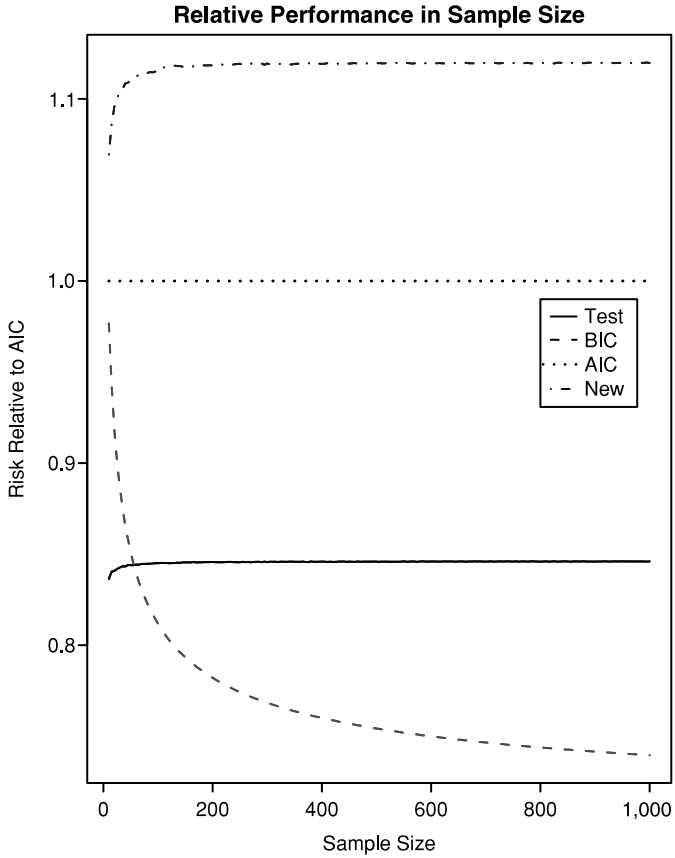


FIGURE 2. Comparing model selection methods from $n = 10$ to 1,000 at $\beta = 0$.

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

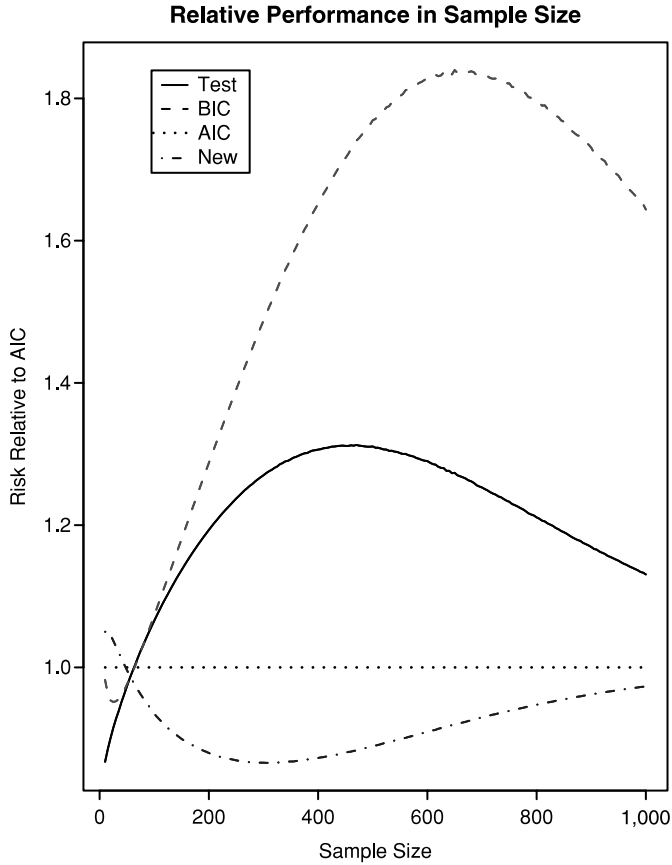


FIGURE 3. Comparing model selection methods from $n = 10$ to 1,000 at $\beta = 0.2$.

has not quite happened when n is 1,000 for $\beta = 0.2$. This again suggests that the pointwise asymptotic optimality of BIC may not say much when the sample size is not very large (relative to β and σ^2).

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

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's 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

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

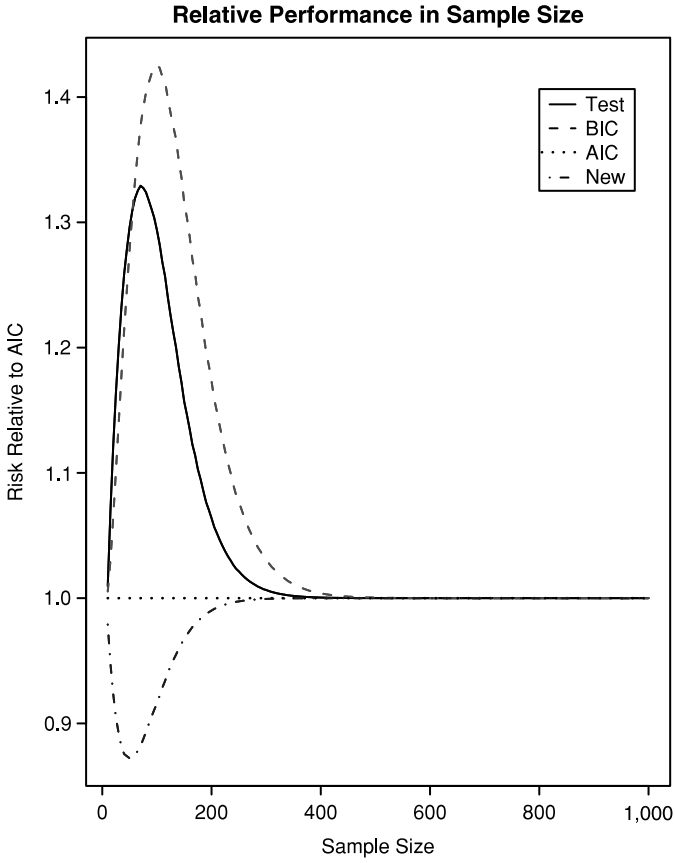


FIGURE 4. Comparing model selection methods from $n = 10$ to 1,000 at $\beta = 0.5$.

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 relatively 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 more often than not 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 choosing the selec-

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

tion 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.

Research has been done on general model averaging or model combining, including Bayesian approaches (for a review of the very rich works on Bayesian methods, see Hoeting, Madigan, Raftery, and Volinsky, 1999), frequentist approaches based on bootstrap or weighting via model selection criterion values (Breiman, 1996; Buckland, Burnham, and Augustin, 1997), and a method called ARM (adaptive regression by mixing) (Yang, 2001, 2003; Yuan and Yang, 2005). Magnus (2002) proposed the use of a Laplace prior for weighting estimates from different models in the problem of estimating the normal mean based on a single observation with known variance, which has a connection to a more general regression setting. In a simple case, Yang (2003) proved that model selection, no matter how sophisticated it is, is more than 25% 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 (2005) 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 it can even perform very poorly.

Next we give the ARM method (Yang, 2003) 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

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

Let $W_0 = 1 - W_1$.

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

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

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

We consider an additional variant of ARM: it combines model 0 and model 1 as before only when AIC and BIC disagree with each other. We call it combining method 2.

6.2. Simulation and a Data Example

6.2.1. A Simulation. We now add the two combining methods in Section 6.1 to the list of the competing model selection methods in Section 5. The settings are kept the same. However, because we do not have simple expressions for the risks of the combined procedures, we simulate the risks of all the competitors based on 2,000 replications. Figure 5 gives the relative performance of the procedures compared to the better one of model 0 and model 1 at sample size 100. Recall that the risks of the least squares estimators from the two models are $(\sigma^2/n) + x_0^2\beta^2$ and $(\sigma^2/n) + (\sigma^2 x_0^2/\sum_{i=1}^n x_i^2)$, respectively.

From the graph, we do see performance improvement by the combining procedures when β is small (relative to the sample size). The combining procedure 1 always combines the two models and performs very well at the β values where the model selection methods differ most, yet it pays a heavy price when β is relatively large (when β is very large, it is fine). 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 (although it also gives up some advantage for small β). Overall, in our judgment, the combining method 2 does well.

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 only one predictor for simple linear regression.

The response variable is the percentage of body fat using Brozek's equation (the second variable in the data). The two individually used predictors are age and fat-free weight (original variables numbers 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 slope parameter is very small or zero.

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

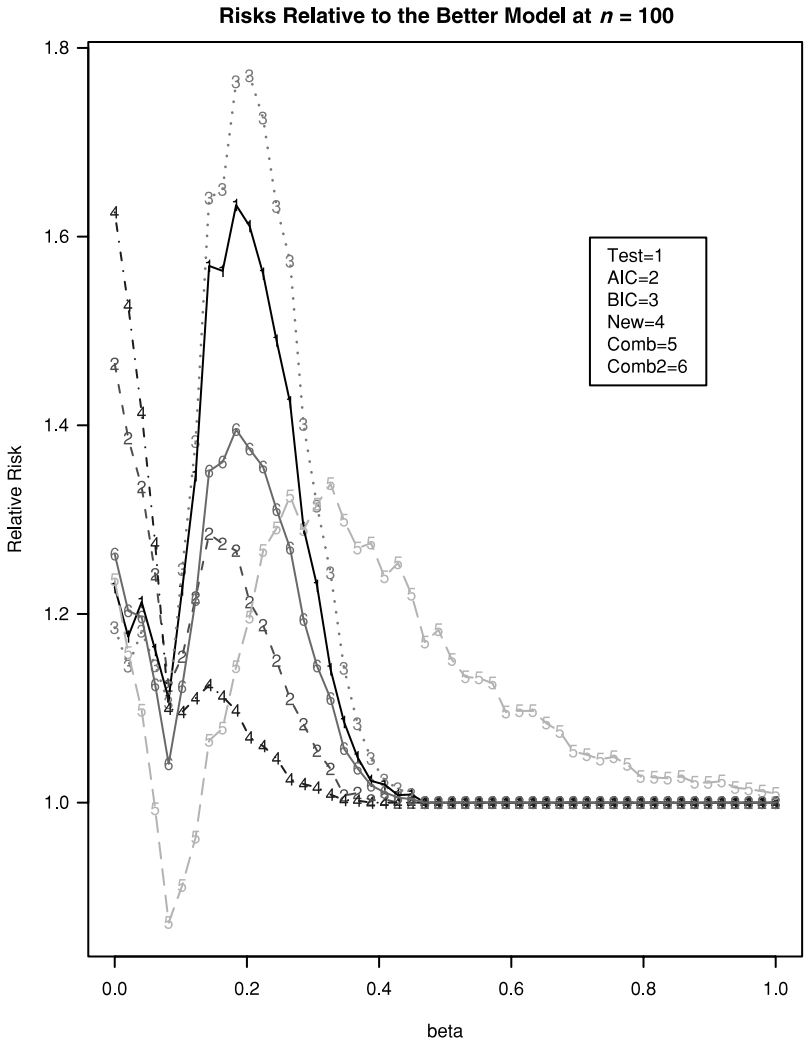


FIGURE 5. Comparing the selection/combining methods in β at $n = 100$.

Figures 6 and 7 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 set 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 entire data set, we know

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

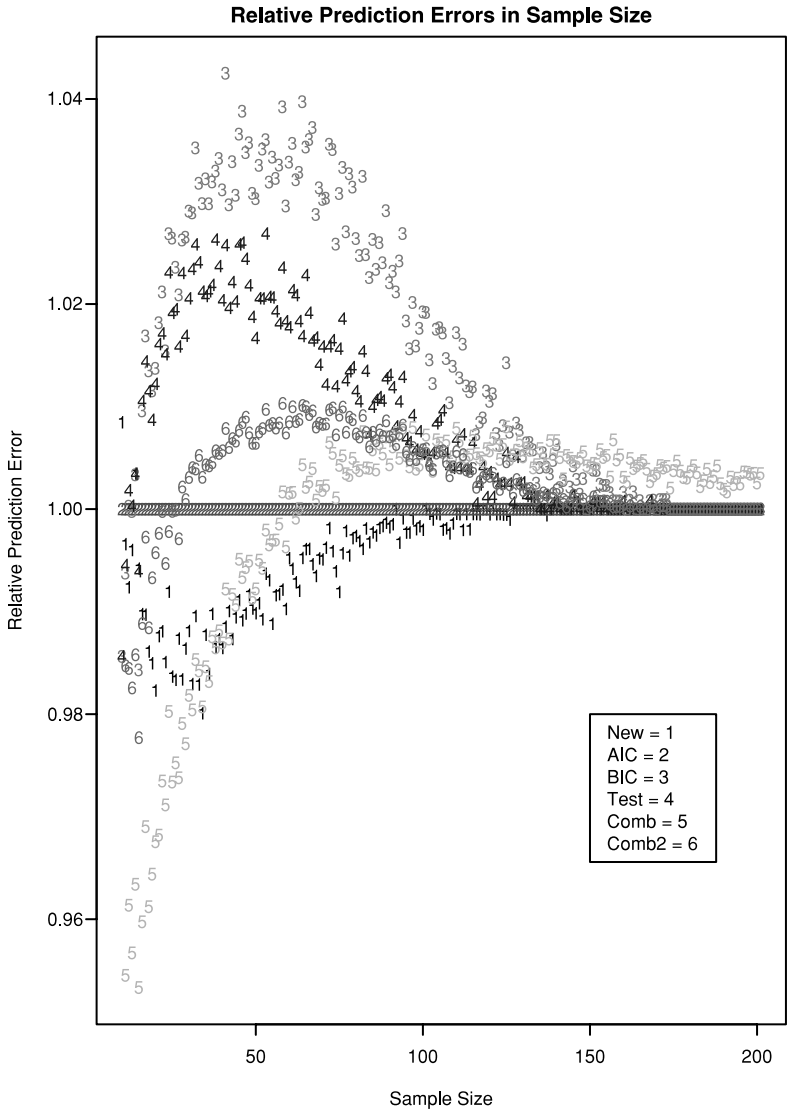


FIGURE 6. Empirical comparison of predictive performances at different sample sizes: Age as the predictor.

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 worsen 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 preceding situation, BIC will eventually per-

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

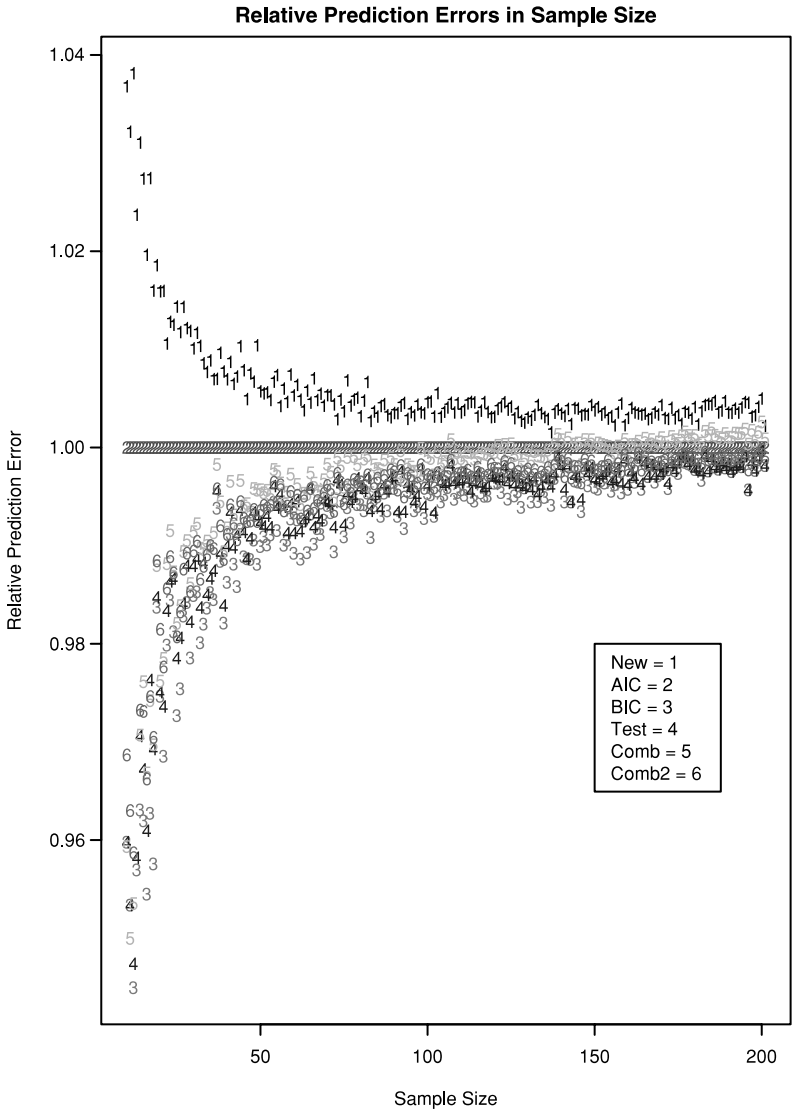


FIGURE 7. Empirical comparison of predictive performances at different sample sizes: Fat-free weight as the predictor.

form 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. That is indeed the case. As the sample size increases, their difference become smaller in the ratio. From Figure 7, at the end, AIC and BIC perform quite similarly. From the earlier theory, if the slope parameter is really

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

zero, then the gap between AIC and BIC will never be closed. On the other hand, if we had a sample size of 10,000 (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 going 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.

The test approach 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 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 good approach for prediction when model selection methods give different answers.

7. CONCLUDING REMARKS

In this work, with both theoretical examinations and numerical demonstrations, it is seen that the popular notion of preferring BIC for parametric regression is misleading. The pointwise asymptotic optimality properties of BIC (or any other consistent model selection method) are often interpreted overly optimistically in the current statistical literature. From the numerical results, it is also clear that the method of testing a hypothesis (at a conventional significance level) first and then making a prediction/estimation should not be the one students learn implicitly or explicitly from our future statistics textbooks. Although adaptive model selection with a data dependent penalty can be very helpful to reduce risk, it simply cannot integrate the pointwise optimality property of BIC and the minimax-rate optimality of AIC.

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

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 (1999), Barron, Birgé, and Massart (1999), and Birgé and Massart (2001).

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

REFERENCES

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Csaki (eds.), *Proceedings of the 2nd International Symposium on Information Theory*, pp. 267–281. Akademia Kiado.

Bancroft, T.A. (1944) On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics* 15, 190–204.

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

Barron, A.R. & T.M. Cover (1991) Minimum complexity density estimation. *IEEE Transactions on Information Theory* 37, 1034–1054.

Barron, A.R., Y. Yang, & B. Yu (1994) Asymptotically optimal function estimation by minimum complexity criteria. In *Proceedings of the 1994 International Symposium on Information Theory*, p. 38.

Birgé, L. & P. Massart (2001) Gaussian model selection. *Journal of the European Mathematical Society* 3, 203–268.

Brown, L.D., M.G. Low, & L.H. Zhao (1997) Superefficiency in nonparametric function estimation. *Annals of Statistics* 25, 2607–2625.

Breiman, L. (1996) Bagging predictors. *Machine Learning* 24, 123–140.

Buckland, S.T., K.P. Burnham, & N.H. Augustin (1997) Model selection: An integral part of inference. *Biometrics* 53, 603–618.

Burnham, K.P. & D.R. Anderson (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag.

Danilov, D. & J.R. Magnus (2004) On the harm that ignoring pretesting can cause. *Journal of Econometrics* 122, 27–46.

Foster, D.P. & E.I. George (1994) The risk inflation criterion for multiple regression. *Annals of Statistics* 22, 1947–1975.

George, E.I. & D.P. Foster (2000) Calibration and empirical Bayes variable selection. *Biometrika* 87, 731–747.

Geweke, J. & R. Meese (1981) Estimating regression models of finite but unknown order. *International Economic Review* 22, 54–70.

Giles, J.A. & D.E.A. Giles (1993) Pre-test estimation and testing in econometrics: Recent developments. *Journal of Economic Surveys* 7, 145–197.

Hannan, E.J. & B.G. Quinn (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B* 41, 190–195.

Hansen, M. & B. Yu (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. IEEE.

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

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.

Judge, G.G. & M.E. Bock (1978) *The Statistical Implications of Pre-test and Stein-Rule Estimators in Econometrics*. Elsevier/North-Holland.

Judge, G.G. & T.A. Yancey (1986) *Improved Methods of Inference in Econometrics*. Elsevier/North-Holland.

Kabaila, P. (2002) On variable selection in linear regression. *Econometric Theory* 18, 913–925.

Leeb, H. (2005) The distribution of a linear predictor after model selection: Conditional finite-sample distributions and asymptotic approximations. *Journal of Statistical Planning and Inference* 134, 64–89.

Leeb, H. & B. Pötscher (2003) The finite-sample model distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* 19, 100–142.

Leeb, H. & B. Pötscher (2005) Model selection and inference: Facts and fiction. *Econometric Theory* 21, 21–59.

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press

Li, K.C. (1987) Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics* 15, 958–975.
Magnus, J.R. (1999) The traditional pretest estimator. *Theory of Probability and Its Applications* 44, 293–308.
Magnus, J.R. (2002) Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal* 5, 225–236.
Magnus, J.R. & J. Durbin (1999) Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica* 67, 639–643.
Nishii, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics* 12, 758–765.
Penrose, K., A. Nelson, & A. Fisher (1985) Generalized body composition prediction equation for men using simple measurement techniques (abstract). *Medicine and Science in Sports and Exercise* 17, 189.
Pollard, D. (2002) *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.
Polyak, B.T. & A.B. Tsybakov (1991) Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. *Theory of Probability and Its Applications* (translation of *Teorija Verojatnostei i ee Primenenija*) 35, 293–306.
Pötscher, B. (1989) Model selection under nonstationarity: Autoregressive models and stochastic linear regression models. *Annals of Statistics* 17, 1257–1274.
Pötscher, B. (1991) Effects of model selection on inference. *Econometric Theory* 7, 163–185.
Rao, C.R. & Y. Wu (1989) A strongly consistent procedure for model selection in a regression problem. *Biometrika* 76, 369–374.
Rao, J.S. & R. Tibshirani (1997) Comment on "An asymptotic theory for linear model selection." *Statistica Sinica* 7, 249–251.
Rissanen, J. (1978) Modeling by shortest data description. *Automatica* 14, 465–471.
Rissanen, J. (1986) Stochastic complexity and modeling. *Annals of Statistics* 14, 1080–1100.
Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
Shao, J. (1997) An asymptotic theory for linear model selection (with discussion). *Statistica Sinica* 7, 221–242.
Shen, X. & J. Ye (2002) Adaptive model selection. *Journal of the American Statistical Association* 97, 210–221.
Shibata, R. (1983) Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics* 35, 415–423.
Speed, T.P. & B. Yu (1993) Model selection and prediction: Normal regression. *Annals of the Institute of Statistical Mathematics* 45, 35–54.
Toro-Vizcarrondo, C. & T.D. Wallace (1968) A test of the mean square error criterion for restriction in linear regression. *Journal of the American Statistical Association* 63, 558–572.
Wallace, T.D. (1972) Weaker criteria and tests for linear restrictions in regression. *Econometrica* 40, 689–698.
Yang, Y. (1999) Model selection for nonparametric regression. *Statistica Sinica* 9, 475–499.
Yang, Y. (2001) Adaptive regression by mixing. *Journal of the American Statistical Association* 96, 574–588.
Yang, Y. (2003) Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica* 13, 783–809.
Yang, Y. (2004) Aggregating regression procedures for a better performance. *Bernoulli* 10, 25–47.
Yang, Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92, 937–950.
Yang, Y. & A.R. Barron (1998) An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory* 44, 95–116.
Yuan, Z. & Y. Yang (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association* 100, 1202–1214.
Zhang, P. (1997) Comment on "An asymptotic theory for linear model selection." *Statistica Sinica* 7, 254–258.

https://doi.org/10.1017/S0266466607070016 Published online by Cambridge University Press