Can the Strengths of AIC and BIC Be Shared? *

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

December 30, 2003

Abstract

It is well known that AIC and BIC have different properties in model selection. BIC is consistent in the sense that if the true model is among the candidates, the probability of selecting the true model approaches 1. On the other hand, AIC is minimax-rate optimal for both parametric and nonparametric cases for estimating the regression function. There are several successful results on constructing new model selection criteria to share some strengths of AIC and BIC. However, we show that in a rigorous sense, even in the setting that the true model is included in the candidates, the above mentioned main strengths of AIC and BIC cannot be shared. That is, for any model selection criterion to be consistent, it must behave sup-optimally compared to AIC in terms of mean average squared error.

1 Introduction

1.1 Setup

Consider the regression model

$$Y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, 2, ..., n,$$

where $\mathbf{x}_i = (x_{i1}, ..., x_{id})$ is the value of a d-dimensional design variable at the i-th observation, Y_i is the response, f is the true regression function, and the random errors ε_i are assumed to be iid normally distributed with mean zero and variance σ^2.

For the purpose of statistical model identification, estimation or prediction, a number of plausible linear models are being considered:

$$Y = f_k(\mathbf{x}, \theta_k) + \varepsilon,$$

where for each k, $\mathcal{F}_k = \{f_k(\mathbf{x}, \theta_k),\ \theta_k \in \Theta_k\}$ is a linear family of regression functions with θ_k being the parameter of a finite dimension m_k.

With the candidate models given, we need to select one of them to best capture the underlying distribution of the data or best estimate the regression function f or predict the future response.

*This work was supported by the US NSF CAREER Grant DMS0094323.
AMS 2000 subject classifications. Primary 62C05; secondary 62F12, 62C20, 62G08.
Key words and phrases. Model selection, AIC, BIC, consistency, minimax-rate optimality.

The above framework includes the usual subset selection and order selection problems in linear regression. It also includes nonparametric regression based on series expansion, where the true function is approximated by linear combinations of appropriate basis functions (such as polynomials, splines, or wavelets).

1.2 Model selection criteria

Up to now, there is a rather large literature on model selection methods following different philosophies, assumptions, theoretical and/or practical considerations. The readers are referred to Shao (1997) for references. We focus on two of the most representative and widely applied model selection criteria in this work.

AIC (Akaike (1973)) and BIC (Schwarz (1978)) are derived from distinct perspectives: AIC intends to minimize the Kullback-Leibler divergence between the true distribution and the estimate from a candidate model and BIC tries to select a model that maximizes the posterior model probability. Due to the rather different motivations, it is not surprising that they have different properties.

The most well-known properties of AIC and BIC are asymptotic (loss) optimality and consistency (in selection), respectively. Simply put, when f is among the candidate families of regression functions, the probability of selecting the true model by BIC approaches 1 as $n \to \infty$ (e.g., Nishii (1984)); On the other hand, if f is not in any of the candidate families and if the number of models of the same dimension does not grow very fast in dimension, the average squared error of the selected model by AIC is asymptotically equivalent to the smallest possible offered by the candidate models (e.g., Shibata (1983), Li (1987), Polyak and Tsybakov (1990), and Shao (1997)). Note that here the true model is defined as the smallest model containing f. These two properties of BIC and AIC are respectively called consistency and asymptotic (nonparametric) optimality (under the average squared error loss). Note that in general, AIC is not consistent and BIC is not asymptotically (loss) optimal in the nonparametric case.

There has been quite a debate between AIC and BIC in the literature, centering on the assumption: Is the true model finite-dimensional or infinite-dimensional? There seems to be a consensus that for the former case, BIC should be preferred and AIC should be chosen for the latter.

1.3 The problem of interest in this work

The purpose of this paper is to investigate the possibility of uniting the rivalry model selection criteria AIC and BIC. Obviously, if possible, sharing the strengths of different statistical procedures is desirable. This is, for example, the spirit of adaptive estimation in function estimation. In that context, a large number of results have been obtained to construct estimation procedures that work optimally in rates of convergence (or even up to the right constants) over different assumptions on the true regression function or different loss functions (see, e.g., Barron, Birgé and Massart (1999) for history and some references). Following this spirit, instead of focusing on the difference between the two model selection rules, why not devise a new one that integrates their strengths together?

There have been several attempts to pursue the good qualities of AIC and BIC using a new criterion. Barron, Yang and Yu (1995) reported that the minimum description length (MDL) criterion (Rissanen (1978)), when applied in a novel way, yields a penalty of AIC type when the data are governed by a nonparametric model and of BIC type when the data are governed by a parametric model in the candidate list. A consequence is 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. Thus in an appropriate sense, the novel use of MDL indeed provides a reconciliation of the criteria AIC and BIC. Hansen and Yu (1997) took a different approach based on MDL to have a penalty term basically switching between AIC and BIC type according to a test statistic. When the true model is finite-dimensional, the criterion is consistent and prediction-optimal (Corollary 1 of Hansen and Yu (1997), see also Hansen and Yu (2001)). Foster and George (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). Yang (2003) showed empirically that when AIC and BIC estimators are properly combined, the new estimator tends to perform like the better one under the squared error loss.

Of course one can consider different aspects of the properties of AIC and BIC to be shared, if possible. The positive results in Barron, Yang and Yu (1995) and Hansen and Yu (1997) focused on the "parametric versus nonparametric" aspect. However, in this paper, from a different angle, assuming that the true model is among the candidates, we show that there is an uncompromisable difference between AIC and BIC. That is, if any model selection procedure is consistent in selection as BIC is, unlike AIC, it must be minimax rate sub-optimal. Therefore, in a strong sense, no model selection procedure can be devised to share the advantages of both AIC and BIC. It is also interesting to note that the classical hypothesis testing theory plays a fundamental role in our analysis.

The elegant asymptotic (nonparametric) optimality property of AIC is usually stated on the loss or risk of the selected model in an asymptotic expression where the limit is taken as $n \to \infty$ with the regression function held fixed. As noted by e.g., Brown, Low and Zhao (1997), in general, such an asymptotic analysis "can involve misleading conclusions" on the performance of the estimator. Indeed, the accuracy of the estimator suggested by such an asymptotic result can actually be illusionary in terms of minimax rate of convergence. Fortunately, this is not the case for AIC, as we consider in the next subsection.

1.4 An important minimax property of AIC

A key feature of an AIC type criterion (including Mallows' C_p (1973)) is that it adds a penalty of the same order as the model dimension to the negative maximized log-likelihood. The significance of this is that with the penalty added as bias correction, the criterion value (with a term common to all models removed) is of the same order as the sum of the squared bias and the estimation error (model dimension over the sample size). Consequently, when the number of the relevant models is under control, the comparison of the criterion value is pretty much similar to comparing the sum of the squared bias and the estimation error over the models. In light of the well-known fact that the best trade-off between the squared bias and the estimation error typically produces the minimax optimal rate of convergence for both parametric and nonparametric function classes (see, e.g., Yang and Barron (1999, Section 4) and the references therein), the AIC type criteria then have the property that they usually yield minimax-rate optimal estimators of the regression function under a squared error type loss. There are many results of this flavor in the literature. We mention Barron, Birgé and Massart (1999) as a source of references. Note that the minimax-rate optimality of AIC type criteria holds much more generally in terms of the assumptions on the candidate models compared to the asymptotic loss optimality.

We give an example result below.

Consider the average squared error for estimating the regression function f: for a model selection criterion δ that selects model $\hat{k}$, let $ASE(f_{\hat{k}}) = \frac{1}{n}\sum_{i=1}^{n}\left(f(\mathbf{x}_i) - f_{\hat{k}}(\mathbf{x}_i, \hat{\theta}_{\hat{k}})\right)^2$, where $\hat{\theta}_{\hat{k}}$ is the least squares estimator of the parameter in the model. It assesses the performance of the estimator at the design points. The corresponding risk is $R(f;\delta;n) = \frac{1}{n}\sum_{i=1}^{n} E\left(f(\mathbf{x}_i) - f_{\hat{k}}(\mathbf{x}_i, \hat{\theta}_{\hat{k}})\right)^2$.

Definition 1: A model selection criterion δ is said to be minimax-rate optimal over a class of regression functions $\mathcal{F}$ if $\sup_{f\in\mathcal{F}} R(f;\delta;n)$ converges at the same rate as $\inf_{\hat{f}}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} E\left(f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\right)^2$, where $\hat{f}$ is over all estimators based on the observations of $Y_1, ..., Y_n$.

Let Γ be the collection of all the models being considered. The size of Γ can be finite or countably infinite. Let N_m denote the number of models that have the same dimension m in Γ. We assume that there exists a positive constant c such that $N_m \leq e^{cm}$, i.e., the number of models of dimension m increases no faster than exponentially in m. This is certainly the case when the size of Γ is finite and also the case in the usual order selection problem in series expansion.

Let δ_{AIC} denote the estimator of f based on the outcome of AIC, i.e., the estimator is $f_{\hat{k}}(\mathbf{x}, \hat{\theta}_{\hat{k}})$, where $\hat{k}$ is the model selected by AIC. Let M_k denote the projection matrix of model k and let r_k denote the rank of M_k (note that $r_k \leq m_k$). Let $\parallel a \parallel_n$ denote the Euclidean norm of a n dimensional vector a.

For simplicity, for the following proposition, we assume that σ^2 is known and then set to be 1 to avoid unnecessary technicality for better illustrating the main point. See Barron, Birgé and Massart (1999), Birgé and Massart (2001) and references therein for more general treatments and many interesting results.

Proposition 1: There exists a constant $C > 0$ depending only on c such that for every regression function f, we have

$$R(f; \delta_{AIC}; n) \leq C \inf_{k \in \Gamma} \left(\frac{\| f - M_k f \|_n^2}{n} + \frac{r_k}{n} \right).$$

Proposition 1 follows readily from Theorem 1 of Yang (1999). A corollary is immediately available if the true model is among the candidates.

Corollary 1: Suppose that model $k^* \in \Gamma$ is the true model. Then

$$\sup_{f \in \mathcal{F}_{k^*}} R(f; \delta_{AIC}; n) \leq \frac{C m_{k^*}}{n}.$$

Thus the worst-case risk of δ_{AIC} under the true model k^* is at the parametric rate $1/n$. In other words, δ_{AIC} is minimax-rate optimal if the true model is among the candidates. When the true regression function is infinite-dimensional (relative to the candidate models), $\| f - M_k f \|_n^2 /n$ is non-zero for all k. For smoothness classes such as Sobolev balls, with an appropriate choice of the candidate models (e.g., polynomial splines), $\inf_{k \in \Gamma} \left(\frac{\| f - M_k f \|_n^2}{n} + \frac{r_k}{n} \right)$ is of the same order as the minimax rate of convergence. Therefore δ_{AIC} is automatically minimax-rate optimal over the smoothness classes without the need to know the true smoothness order.

From above, we know that δ_{AIC} is minimax-rate optimal, converging at rate $1/n$ when one of the candidate model holds; and is also minimax-rate optimal when the true regression function is infinite-dimensional in e.g., Sobolev classes (or more generally in full approximation sets, see Yang and Barron (1999, Section 4)).

It is useful to point out that many theoretical results in the literature on model selection are pointwise asymptotics in the sense that the loss or risk bound is of an asymptotic nature at a fixed f (for example, the main results in Shibata (1983), Li (1987) and Shao (1997) are of this kind). A consequence is that the results do not lend useful implications on minimax properties of the estimators. Note that the minimax view on statistical estimation has been emphasized in recent years (see, e.g., Donoho and Johnstone (1998)).

In contrast to AIC, BIC does not have the minimax-rate optimality mentioned above. Indeed, Foster and George (1994) showed that in the parametric scenario, BIC converges sub-optimally in terms of the worst-case risk performance. Therefore, even in the parametric case, BIC can perform much worse than AIC.

The rest of the paper is organized as follows. The main result is given in Section 2 and the proof is provided in Section 3. A brief summary of the paper is in Section 4.

2 Can consistency and minimax rate optimality be shared?

Assumption 1: There exist two models $k_1, k_2 \in \Gamma$ such that

1. $\mathcal{F}_{k_1} = \{f_{k_1}(\mathbf{x}, \theta_{k_1}) : \theta_{k_1} \in \Theta_{k_1}\}$ is a sub-linear space of $\mathcal{F}_{k_2} = \{f_{k_2}(\mathbf{x}, \theta_{k_2}) : \theta_{k_2} \in \Theta_{k_2}\}$;

2. There exists a function $\varphi(\mathbf{x})$ in $\mathcal{F}_{k_2}$ orthogonal to $\mathcal{F}_{k_1}$ (at the design points) with $\frac{1}{n}\sum_{i=1}^{n} \varphi^2(\mathbf{x}_i)$ being bounded between two positive constants (at least for large enough n);

3. There exists a function $f_0 \in \mathcal{F}_{k_1}$ such that f_0 is not in any family $\mathcal{F}_k$ ($k \in \Gamma$) that does not contain $\mathcal{F}_{k_1}$.

The second part of Assumption 1 is very mild and is typically satisfied for a reasonable design. The third part of the assumption always holds when one has a finite number of models or a countable list of nested models. For a general case of countably many models, the satisfaction of the third requirement is not obvious (it seems that the axiom of choice is relevant). Assumption 1 is satisfied for subset or order selection in the usual linear regression setting with a reasonable design.

Theorem 1. Under Assumption 1, if any model selection method δ is consistent in selection, then we must have

$$n \sup_{f \in \mathcal{F}_{k_2}} R(f; \delta; n) \to \infty. \tag{1}$$

Remarks:

1. Without a proper nested relationship between the models, defining consistency in model selection can be tricky. Consider any two models $k_1, k_2 \in \Gamma$ that are not nested. If $\mathcal{F}_{k_1} \cap \mathcal{F}_{k_2}$ is not degenerate and $\mathcal{F}_{k_1} \cap \mathcal{F}_{k_2}$ does not correspond to a candidate model, then for a given f in the intersection, it is unclear how to define the true model for f (especially when k_1 and k_2 have the same dimension).

2. The conclusion of (1) still holds even if one considers a compact subset of $\mathcal{F}_{k_2}$ instead of $\mathcal{F}_{k_2}$ itself in the expression (see the proof of Theorem 1 in Section 3).

3. From the proof of the theorem in Section 3, it is seen that allowing randomization in model selection (which corresponds to randomized testing there) does not help to unite AIC and BIC.

The theorem says that in the parametric case, if one is to pursue consistency in selection, one must pay a somewhat high price for estimating the regression function. Thus the strengths of AIC and BIC cannot be combined in a rigorous sense. The theorem also implies that consistency in selection and minimax-rate optimality in estimating f are somewhat conflicting performance measures on model selection.

3 Proof of Theorem 1

The key idea in the proof is to reduce the problem to a hypothesis testing problem where the classical hypothesis testing theory can be applied.

We first prove Theorem 1 in a simple case. Suppose that we have two models, the null model: $Y_i = \alpha + \varepsilon_i, \quad i = 1, 2, ..., n$ and the simple linear model below:

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

where x is a one-dimensional design variable. Without loss of generality, here we assume that the design is such that $\overline{x}_n = 0$. In addition, we assume that $\frac{1}{n}\sum_{i=1}^{n} x_i^2$ is bounded between two positive constants for all n. For convenience, call the aforementioned models model 0 and model 1 respectively.

Now consider a consistent (in selection) model selection criterion δ. Let A_n be the event that model 1 is selected. The corresponding estimator of $f(x_0)$ is

$$\hat{f}(x_0) = \hat{\alpha} + \hat{\beta} x_0 I_{A_n}.$$

Then its risk at x_0 under the squared error loss is

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

and thus the mean average squared error is

$$R(f; \delta; n) = \frac{\sigma^2}{n} + \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) E\left(\hat{\beta} I_{A_n} - \beta\right)^2.$$

Note that in the above equality, the cross-product term vanishes due to that $\overline{x}_n = 0$. We next show that for any consistent model selection method, for each $c > 0$, we must have

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

The conclusion of Theorem 1 then follows for the simple two model case. Note that the left hand side above is equal to

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

Thus to show that δ is not minimax-rate optimal at rate $1/n$, it suffices to show that for each $c > 0$,

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

Since δ is consistent, 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:

$$Y_i = \beta x_i + \varepsilon_i, \quad i = 1, 2, ..., n, \tag{3}$$

where the errors are independent and have standard normal distribution. Note that this is a sub-family of (2) with $\alpha = 0$ and $\sigma^2 = 1$. Consider the hypotheses: $H_0 : \beta = 0$ 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 for any test with the probability of type I error going to zero, it necessarily has $\sup_{|\beta|\le c} n\beta^2 P\left(\widetilde{A}_n^c\right) \to \infty$, where $\widetilde{A}_n$ is the rejection region of the test. Let $f(y_1, ..., y_n; \beta)$ denote the joint probability density function of $(Y_1, ..., Y_n)$ under (3). Note that for $\beta_1 > \beta_0 \ge 0$,

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

Thus the family has a monotone likelihood ratio in the statistic $\sum_{i=1}^n x_i Y_i$. It follows from the familiar Karlin-Rubin theorem that a uniformly most powerful (UMP) test exists, which 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 $P_{\beta=0}\left(\sum_{i=1}^n x_i Y_i \ge d_n\right) = P_{\beta=0}(A_n)$. Let $A_{n,*}$ denote the event $\{\sum_{i=1}^n x_i Y_i \ge d_n\}$. By the UMP property of $A_{n,*}$, we have that for all $\beta > 0$

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

Consequently,

$$\sup_{|\beta|\le c} n\beta^2 P_\beta\left(A_n^c\right) \ge \sup_{0\le\beta\le c} n\beta^2 P_\beta\left(A_{n,*}^c\right).$$

Now since $\sum_{i=1}^n x_i Y_i$ has a 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 $\beta > 0$

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

Since $P_{\beta=0}\left(\sum_{i=1}^n x_i Y_i \ge d_n\right) = P_{\beta=0}(A_n) \to 0$, we must have $\frac{d_n}{\sqrt{n}} \to \infty$. Then with the choice of $\beta_n = \min\left(\frac{d_n}{2\sum x_i^2}, c\right)$, we have

$$\sup_{0\le\beta\le c} n\beta^2 P_\beta\left(A_{n,*}^c\right) \ge n\beta_n^2 P_{\beta_n}\left(A_{n,*}^c\right).$$

Clearly $n\beta_n^2 \to \infty$. Also $P_{\beta_n}\left(A_{n,*}^c\right) \ge P\left(N(0,1) < \frac{d_n}{2\sqrt{\sum x_i^2}}\right)$ and thus $P_{\beta_n}\left(A_{n,*}^c\right)$ converges to 1. It follows that $\sup_{|\beta|\le c} n\beta^2 P\left(A_n^c\right) \to \infty$. This proves the result of Theorem 1 for the special case.

Now we consider the general case. Let k_1 and k_2 be two models that are nested: $\mathcal{F}_{k_1} = \{f_{k_1}(\mathbf{x}, \theta_{k_1}) : \theta_{k_1} \in \Theta_{k_1}\}$ is a sub-linear space of $\mathcal{F}_{k_2} = \{f_{k_2}(\mathbf{x}, \theta_{k_2}) : \theta_{k_2} \in \Theta_{k_2}\}$. Let $\varphi(x)$ be a function in $\mathcal{F}_{k_2}$ that is orthogonal (at the design points) to $\mathcal{F}_{k_1}$. Under Assumption 1, we can have $\frac{1}{n}\sum_{i=1}^{n} \varphi^2(\mathbf{x}_i)$ bounded between two positive constants. Also, under the third part of Assumption 1, there is a function $f_0 \in \mathcal{F}_{k_1}$ such that f_0 does not belong to any other $\mathcal{F}_k$ that does not contain $\mathcal{F}_{k_1}$ (so that the true model associated with f_0 is clearly k_1). Let B_n be the event that model k_1 is *not* selected for a model selection method δ. If δ is consistent, then

$$P_{f_0}(B_n) \to 0 \text{ as } n \to \infty.$$

Consider a simplified model:

$$Y_i = f_0(\mathbf{x}_i) + \beta\varphi(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, 2, ..., n, \tag{4}$$

and the testing problem $H_0 : \beta = 0$ versus $H_1 : \beta > 0$. Note that under H_0, the data comes from model k_1 and under H_1, the regression function is in $\mathcal{F}_{k_2}$. The model selection rule δ can be used to get a test: accept H_0 when model k_1 is selected by δ and otherwise reject H_0. Since δ is consistent, this test has probability of type I error going to zero as $n \to \infty$.

Let $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))^{'}$, $\mathbf{Y} = (Y_1, ..., Y_n)^{'}$, $\underline{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^{'}$, $\underline{\varphi} = (\varphi(\mathbf{x}_1), ..., \varphi(\mathbf{x}_n))^{'}$ and let M_{k_1} be the projection matrix of model k_1. Observe that under (4),

$$\begin{aligned} \| \quad & \mathbf{f} - M_{k_1}\mathbf{Y} \|_n^2 \\ = \quad & \| \mathbf{f} - M_{k_1}\mathbf{f} \|_n^2 + \underline{\varepsilon}^{'} M_{k_1}\underline{\varepsilon} \\ = \quad & \| \beta\underline{\varphi} - M_{k_1}\underline{\varphi} \|_n^2 + \underline{\varepsilon}^{'} M_{k_1}\underline{\varepsilon} \\ = \quad & \beta^2 \| \underline{\varphi} \|_n^2 + \underline{\varepsilon}^{'} M_{k_1}\underline{\varepsilon}, \end{aligned}$$

where the second and the third equalities follow from the fact that $(f_0(\mathbf{x}_1), ..., f_0(\mathbf{x}_n))^{'}$ is in the column space of M_{k_1} and that $\underline{\varphi}$ is orthogonal to the column space of M_{k_1}. Then under (4), the risk of the estimator associated with δ is

$$\begin{aligned} R(f; \delta; n) &= \frac{1}{n}\sum_{k\in\Gamma} E_\beta \| \mathbf{f} - M_k\mathbf{Y} \|_n^2 \, I_{\{\hat{k}=k\}} \\ &\geq \frac{1}{n} E_\beta \| \mathbf{f} - M_{k_1}\mathbf{Y} \|_n^2 \, I_{\{\hat{k}=k_1\}} \\ &\geq \frac{\beta^2}{n} E_\beta \| \underline{\varphi} \|_n^2 \, I_{\{\hat{k}=k_1\}} \\ &= \frac{\sum_{i=1}^{n} \varphi^2(\mathbf{x}_i)}{n} \beta^2 P_\beta\left(\hat{k} = k_1\right). \end{aligned}$$

Consequently, to show that $n \sup_{f\in\mathcal{F}_{k_2}} R(f; \delta; n) \to \infty$, we only need to show

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

With our setup of the testing problem, the above statement holds if we can show that for any test of the hypotheses with rejection region A_n satisfying $P_{\beta=0}(A_n) \to 0$ we must have $\sup_{|\beta|\le c} n\beta^2 P(A_n^c) \to \infty$. Let $Z_i = Y_i - f_0(\mathbf{x}_i)$. Then $Z_1, ..., Z_n$ are independent Gaussian random variables with Z_i having $N(\beta\varphi(\mathbf{x}_i), \sigma^2)$ distribution. The earlier arguments for the simple two-model case follow similarly for proving the last assertion. This completes the proof of Theorem 1.

4 Summary

Identifying the true model (when possible) and optimally estimating the regression function are both fundamental problems in regression analysis. As is well-known, BIC enjoys the consistency property in terms of selecting the true model. AIC is asymptotically optimal in terms of the average squared error when the candidate models are all incorrect. The penalty of AIC type ensures an important minimax property: it is minimax-rate optimal for both parametric and nonparametric cases.

Trying to go beyond the debate between AIC and BIC, some works in the literature have successfully combined certain aspects of the two model selection rules. In this paper, however, we have shown that the consistency aspect of BIC and the minimax-rate optimality aspect of AIC cannot be combined: no matter how one comes up with a model selection criterion, if one pursues one aspect, one must sacrifice the other. Thus the goals of model identification and minimax-rate estimation of the regression function cannot be aligned.

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] Allen, D.M. (1974) The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125-127.

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

[4] 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.

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

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

[7] Donoho, D.L. and Johnstone, I.M. (1998) Minimax estimation via wavelet shrinkage. *Ann. Statistics*, **26**, 879-921.

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

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

[10] 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.

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

[12] 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.

[13] Mallows, C.L. (1973) Some comments on C_p. *Technometrics*, **15**, 661-675.

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

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

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

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

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

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

[21] Yang, Y. and Barron, A.R. (1999) Information-theoretic determination of minimax rates of convergence. *Ann. Statistics*, **27**, 1564-1599.