Biometrika (2005), **92,** 4, *pp.* 937–950

Printed in Great Britain

Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation

BY YUHONG YANG

School of Statistics, University of Minnesota, 224 Church Street S.E., Minneapolis, Minnesota 55455, U.S.A.

yyang@stat.umn.edu

SUMMARY

A traditional approach to statistical inference is to identify the true or best model first with little or no consideration of the specific goal of inference in the model identification stage. Can the pursuit of the true model also lead to optimal regression estimation? In model selection, it is well known that BIC is consistent in selecting the true model, and AIC is minimax-rate optimal for estimating the regression function. A recent promising direction is adaptive model selection, in which, in contrast to AIC and BIC, the penalty term is data-dependent. Some theoretical and empirical results have been obtained in support of adaptive model selection, but it is still not clear if it can really share the strengths of AIC and BIC. Model combining or averaging has attracted increasing attention as a means to overcome the model selection uncertainty. Can Bayesian model averaging be optimal for estimating the regression function in a minimax sense? We show that the answers to these questions are basically in the negative: for any model selection criterion to be consistent, it must behave suboptimally for estimating the regression function in terms of minimax rate of covergence; and Bayesian model averaging cannot be minimax-rate optimal for regression estimation.

Some key words: AIC; BIC; Consistency; Minimax-rate optimality; Model averaging; Model selection.

1. INTRODUCTION

1·1. *Motivation*

In statistical data analysis, multiple competing models are often considered. Traditionally, the goal of statistical inference is often not considered until after the 'optimal' model is found. Recently, however, much work has been done on model selection uncertainty and various ways have been proposed of combining the candidate models, as opposed to selecting one of them.

In this work we address the relationship between model identification and estimation in a parametric regression context. Suppose that we must select a single model within a list of candidates, among which is the true model. We should demand that our model selection rule be consistent, in that, as the sample size goes to infinity, the probability of selecting the true model goes to 1. Then an important theoretical question is whether or not any consistent model selection rule also yields optimal performance for the goal of

estimating the regression function. If the answer is in the negative, then it provides a strong theoretical objection to the traditional practice of identifying the true or best model first and then doing the statistical inference; consequently, one should keep the specific objective of inference in mind when conducting model selection.

The above issue is closely related to the competition between BIC and AIC, in that BIC represents consistent model selection rules and AIC represents minimax-rate optimal rules for estimating the regression function. It is quite clear that no model selection criterion with a deterministic penalty can simultaneously enjoy the properties of AIC and BIC. One may then ask if an adaptive model selection rule with a data-dependent penalty can share the strengths of AIC and BIC.

Methods based on combining or averaging models are widely expected to perform better than model selection, so one may ask if model averaging can come to the rescue.

1·2. *Formulation of the problem*

Consider the regression model

$$Y_i = f(x_i) + \varepsilon_i \quad (i = 1, 2, \ldots, n),$$

where $x_i = (x_{i1}, \ldots, x_{id})$ is the value of a d-dimensional design variable at the ith observation, Y_i is the response, f is the true regression function, and the random errors ε_i are assumed to be independent and 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, from which one must be selected:

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

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

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.

1·3. *Model selection criteria*

There is a very large literature on model selection methods following different philosophies, assumptions and theoretical and/or practical considerations; see Shao (1997) and McQuarrie & Tsai (1998) for references. More recently, Leeb & Pötscher (2005) discuss problems with statistical inference after model selection. We focus on two of the most representative and widely applied model selection criteria, namely AIC (Akaike, 1973) and BIC (Schwartz, 1978), which are derived from distinct perspectives: AIC aims at minimising the Kullback–Leibler divergence between the true distribution and the estimate from a candidate model and BIC tries to select a model that maximises the posterior model probability.

When f is among the candidate families of regression functions, the probability of selecting the true model by BIC approaches 1 as $n \to \infty$ (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 model selected by AIC is asymptotically equivalent to the minimum offered by the candidate models (Shibata, 1983; Li, 1987; Polyak & Tsybakov, 1991; Shao, 1997). Note that here

the true model is defined as the smallest model containing f. Note that, in general, AIC is not consistent and BIC is not asymptotically loss-optimal in the nonparametric case.

There has been a debate between AIC and BIC in the literature, centring on the issue of whether the true model is 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.

1·4. *Adaptive model selection*

Both the AIC and BIC criteria take the form of loglikelihoods with a deterministic penalty, and the asymptotic analysis of Shao (1997) shows that, if one restricts attention to deterministic penalties, one cannot enjoy the consistency and minimax-rate optimality at the same time. However, what if the penalty is data-adaptive?

In function estimation based on series expansion, it is well known that the order of expansion should increase appropriately with the sample size according to the smoothness of the true function in order to achieve the optimal rate of convergence, and this can often be done in a data-driven fashion. In the same spirit, Barron et al. (1994) reported that the minimum description length criterion (Rissanen, 1978), when applied in a novel way, yields an AIC-like penalty when the data are governed by a nonparametric model and a BIC-like penalty when the data are governed by a parametric model in the candidate list. The resulting estimator therefore converges at the minimax optimal rate for nonparametric cases and also optimally in rate in terms of a cumulative prediction error for parametric cases. Hansen & Yu (1999) took a different approach based on a minimum description length criterion that had a penalty term which switched between AIC-like and BIC-like according to a test statistic. George & Foster (2000) proposed new Bayesian model selection criteria based on empirical Bayes approaches with an adaptive penalty term that acts like BIC or RIC, the risk inflation criterion; note that RIC has a penalty of AIC type when the number of models does not grow with the sample size. Shen & Ye (2002) proposed the use of generalised degrees of freedom for assigning an adaptive penalty term and reported promising simulation results. Yang (2003) showed numerically that, when AIC and BIC estimators are properly combined, the new estimator tends to perform like the better one under the squared-error loss.

In spite of the positive results, it is still not clear theoretically if a flexible model selection rule with a data-driven penalty can provide adaptation with respect to the consistency and minimax-rate optimality of regression estimation. If the two properties can be brought together, then adaptive model selection solves the conflict between AIC and BIC.

The asymptotic nonparametric optimality property of AIC is usually stated in terms of 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 for example by Brown et al. (1997), in general such an asymptotic analysis 'can involve misleading conclusions' about the performance of the estimator. Indeed, the accuracy of the estimator suggested by such an asymptotic result can sometimes be illusory in terms of minimax rate of convergence. Fortunately, this is not the case for AIC, as we consider in § 1·5.

1·5. *An important minimax property of* AIC

A key feature of an AIC-type criterion, including Mallow's C_p (1973), is that it adds a penalty of the same order as the model dimension to the negative maximised loglikelihood. In the regression context, it reduces to the sum of the residual sum of squares and a

multiple of the model dimension. The latter corrects the bias in the underestimation of the risk of each model by the residual sum of squares. The significance of this is that the criterion value, with a term common to all models removed, is then of the same order as the sum of the squared bias and the estimation error, the latter being of the order of model dimension over the sample size. Consequently, when the number of relevant models is under control, the comparison of the criterion values properly mimics the comparison of the unknown risks, i.e. the sum of the squared bias and the estimation error, over the models; that is, when there are not too many models, the AIC-type criteria are more or less doing what they are supposed to do, namely minimising the risk over the candidate models. In the 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 & Barron (1999, § 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 flavour in the literature. We mention Barron et al. (1999) as a source of references. Note that the minimax-rate optimality of AIC-type criteria holds under assumptions about the candidate models that are usually much less restrictive compared to those required for the asymptotic loss optimality.

We give a representative 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 $\mathrm{ASE}(f_{\hat{k}}) = n^{-1}\sum_{i=1}^{n}\{f(x_i) - f_{\hat{k}}(x_i, \hat{\theta}_{\hat{k}})\}^2$, 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 $R(f; \delta; n) = n^{-1}\sum_{i=1}^{n} E\{f(x_i) - f_{\hat{k}}(x_i, \hat{\theta}_{\hat{k}})\}^2$.

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 $\inf_{\hat{f}} \sup_{f\in\mathcal{F}} n^{-1}\sum_{i=1}^{n} E\{f(x_i) - \hat{f}(x_i)\}^2$, where $\hat{f}$ ranges over all estimators based on the observations of $Y_1, \ldots, Y_n$.*

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_0 such that $N_m \leqslant \exp(c_0 m)$; that is 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 in the usual order-selection problem in series expansion.

Let δ_{AIC} denote the estimator of f based on the outcome of AIC; that is the estimator is $f_{\hat{k}}(x, \hat{\theta}_{\hat{k}})$, 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 $r_k \leqslant m_k$. Let $\|a\|_n$ denote the Euclidean norm of an n-dimensonal vector a.

For simplicity, for the following proposition, we assume that σ^2 is known and set $\sigma^2 = 1$ to avoid unnecessary technicalities; see Barron et al. (1999), Birgé & Massart (2001) and references therein for more general treatments.

PROPOSITION 1. *There exists a constant $C > 0$ depending only on c_0 such that, for every regression function f, we have*

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

Proposition 1 follows readily from Theorem 1 of Yang (1999).

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

$$\sup_{f \in \mathscr{F}_{k^*}} R(f; \delta_{\mathrm{AIC}}; n) \leqslant \frac{Cm_{k^*}}{n}.$$

Thus the worst-case risk of δ_{AIC} under the true model k^* is at the parametric rate of $1/n$. 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 nonzero for all k. For smoothness classes such as Sobolev balls, with an appropriate choice of the candidate models, such as polynomial splines, $\inf_{k \in \Gamma} (\|f - M_k f\|_n^2/n + r_k/n)$ is of the same order as the minimax rate of convergence (Yang & Barron, 1999). 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 models holds, and is also minimax-rate optimal when the true regression function is infinite-dimensional, for example in Sobolev classes or more generally in full approximation sets (Yang & Barron, 1999, § 4).

Existing theoretical results on model selection in terms of pointwise asymptotics, i.e. the loss or risk bound is of an asymptotic nature at a fixed f, usually do not readily provide useful implications for minimax properties of the estimators; note that the minimax view of statistical estimation has been emphasised in recent years (Donoho & Johnstone, 1998).

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

With a similar derivation, Proposition 1 also holds for any AIC-type criterion which has a penalty on the residual sum of squares of the form cm_k, instead of $2m_k$ in AIC, for any constant $c > 1$; note that any choice of $c \neq 2$ does not lead to asyptotic loss optimality for the nonparametric case. The consistency property of BIC is shared by any BIC-type criterion which has a penalty of the form $c_n m_k$, instead of $m_k \log n$ in BIC, for a deterministic sequence c_n satisfying $c_n \to \infty$ and $c_n/n \to 0$ as $n \to \infty$ (Shao, 1997).

2. CAN CONSISTENCY AND MINIMAX RATE OPTIMALITY BE SHARED?

Here we show that adaptive model selection cannot be fully adaptive with respect to AIC and BIC.

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

(i) $\mathscr{F}_{k_1} = \{f_{k_1}(x, \theta_{k_1}) : \theta_{k_1} \in \Theta_{k_1}\}$ is a linear subspace of $\mathscr{F}_{k_2} = \{f_{k_2}(x, \theta_{k_2}) : \theta_{k_2} \in \Theta_{k_2}\}$;

(ii) there exists a function $\varphi(x)$ in $\mathscr{F}_{k_2}$ orthogonal to $\mathscr{F}_{k_1}$ at the design points, with $n^{-1} \sum_{i=1}^n \varphi^2(x_i)$ bounded between two positive constants, at least for large enough n;

(iii) there exists a function $f_0 \in \mathscr{F}_{k_1}$ such that f_0 is not in any family $\mathscr{F}_k$, for $k \in \Gamma$, that does not contain $\mathscr{F}_{k_1}$.

Assumption 1 (ii) is very mild and is typically satisfied for a reasonable design. Part (iii) 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 (iii) 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*

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

Remark 1. Without a proper nested relationship between the models, defining consistency in model selection can be tricky in general. 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 any of the models being considered, 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. This difficulty is not present for the all-subset-selection case nor the case with a sequence of nested models.

Remark 2. Conclusion (1) still holds even if one considers a compact subset of $\mathcal{F}_{k_2}$ of the same dimension instead of $\mathcal{F}_{k_2}$ itself in the expression; see the proof of Theorem 1 in the Appendix.

Remark 3. From the proof of the theorem, it is seen that allowing randomisation in model selection, which corresponds to randomised testing there, does not help to unite AIC and BIC.

The theorem says that if, in the parametric case, one is to pursue consistency in selection, one must pay a somewhat high price for estimating the regression function. Thus the strengths of an AIC-type criterion and a BIC-type criterion cannot be combined in a rigorous sense.

We emphasise that, although it is fairly easy to see that a nonadaptive model selection rule with a deterministic penalty cannot combine the strengths of AIC and BIC, the negative result for general model selection methods is far from trivial.

The result in this section implies that the traditional approach of 'identifying the true model first' (Hand & Vinciotti, 2003), that underlies for example the traditional Box–Jenkins approach of ARIMA modelling (Box et al., 1994), is not totally satisfactory, from a theoretical point of view. From Theorem 1, if one's goal is estimation of the regression function or prediction, it is better not to try to find the true model first.

3. COMBINING MODELS CANNOT SOLVE THE CONFLICT BETWEEN MODEL IDENTIFICATION AND REGRESSION ESTIMATION

In proposals for combining models, data-dependent weights are computed for the models and the estimators or predictions from the models are accordingly weighted; see Hoeting et al. (1999) and Yang (2003) for references.

Our interpretation of the empirical results on model combining and model selection in the literature is that, when model selection instability is high, combining the models can substantially improve the accuracy of estimation/prediction. On the other hand, when the best model can be easily identified, combining the models usually loses out to model selection.

In a sense, model selection is a special case of model combining, with the weights concentrated on a single model. We show that, even if nondegenerate weights are allowed, consistency in selection and the minimax-rate estimation of the regression function cannot be simultaneously achieved.

Consider a model combining method τ. Let W_k be the resulting data-dependent weight for model k satisfying that $W_k \geqslant 0$ and $\sum_{k \in \Gamma} W_k = 1$. With the weights, the regression estimator is $\hat{f}(x) = \sum_{k \in \Gamma} W_k f_k(x, \hat{\theta}_k)$. Let $R(f; \tau; n) = n^{-1} \sum_{i=1}^{n} E\{f(x_i) - \hat{f}(x_i)\}^2$.

DEFINITION 2. *A model combining method τ is said to be consistent in weighting if, when the true model k^* is in Γ, we have that $W_{k^*} \to 1$ in probability as $n \to \infty$.*

THEOREM 2. *Under Assumption* 1, *if any model combining method τ is consistent in weighting, then we must have*

$$n \sup_{f \in \mathcal{F}_{k_2}} R(f; \tau; n) \to \infty. \tag{2}$$

From Theorem 2, proved in the Appendix, we know that averaging the models, however it is done, cannot essentially solve the conflict between model identification and regression estimation in the sense that, if one wants the model weights to favour the true model asymptotically, then unfortunately the estimation of the regression function has to suffer in terms of the uniform rate of convergence.

4. BAYESIAN MODEL AVERAGING CANNOT BE MINIMAX-RATE OPTIMAL FOR REGRESSION

Bayesian model averaging researchers emphasise that Bayesian model averaging has optimality properties and has satisfactorily solved the model selection uncertainty problem.

In Bayesian model averaging, prior probabilities are assigned to the candidate models and prior distributions are chosen for the parameters within each model. Then the posterior probabilities of the models are derived by the Bayes rule, usually with the aid of computational techniques such as importance sampling and Markov chain Monte Carlo (Clyde et al., 1996; Hoeting et al., 1999).

Consider fixed prior distributions on the models and on the model parameters; that is the priors do not depend on the sample size. Then Bayesian model selection by maximising the posterior model probability is consistent provided that the true model is among the candidates; that is, the posterior probability of the true model converges to 1, under mild conditions (Berger & Pericchi, 2001, p. 138). Theorem 2 suggests that, if one uses the maximum posterior probability model to estimate the regression function, then it will not be minimax-rate optimal.

For the estimation of a regression model under squared-error loss the weighting of the posterior means of the candidate models obtained by Bayesian model averaging yields the Bayes estimator, and it is natural to ask if this estimator of the regression function is minimax-rate optimal or not.

It should be pointed out that Theorem 2 does not answer this question, because the least squares estimators, rather than the Bayes estimators, are combined there.

For simplicity, here we focus on the simple case of two models, namely

Model 0,

$$Y_i = \alpha + \varepsilon_i \quad (i = 1, 2, \ldots, n),$$

Model 1,

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, 2, \ldots, n), \tag{3}$$

where x is a one-dimensional design variable, and the errors ε_i are independent and normally distributed with mean zero and variance σ^2. Without loss of generality, we assume that the design is such that $n^{-1}\sum x_i = 0$. In addition, we assume that $n^{-1}\sum x_i^2$ is bounded between two positive constants for all n. Let $\hat{\beta}_B$ be the posterior mean of β.

Assumption 2. When the observations are from Model 0 or Model 1, the posterior probability of the true model approaches one in probability. In addition, when the observations are from Model 1, with $\beta \neq 0$, the posterior mean of β is such that, for any a_n converging to ∞, we have that

$$\limsup_{n\to\infty} \mathrm{pr}_{\beta = a_n/\sqrt{n}}(|\hat{\beta}_B - \beta| \geqslant \beta/2) < 1.$$

This assumption is expected to hold with reasonable priors. The second part of the assumption is sensible because, as a result of the usual consistency of $\hat{\beta}_B$ with variance of order $1/n$, $\mathrm{pr}_\beta(|\hat{\beta}_B - \beta| \geqslant \beta/2) \to 0$ usually holds as long as β is of a larger order than $1/\sqrt{n}$. One can easily verify that the assumption is met with conjugate normal priors on the parameters α and β.

Let ψ denote the procedure of estimating $f(x)$ by its posterior mean, and let $R(f; \psi; n)$ denote the average mean squared error of ψ at the design points.

THEOREM 3. *Under Assumption* 2, *we have that*, *for any constant* $c > 0$,

$$n \sup_{|\beta| \leqslant c} R(f; \psi; n) \to \infty. \tag{4}$$

From the proof in the Appendix, it is seen that, for the second requirement in Assumption 2, we only need $\mathrm{pr}_{\beta=\beta_n}(|\hat{\beta}_B - \beta| \geqslant \beta/2)$ to be bounded away from 1 for a particular sequence of β_n of higher order than $1/\sqrt{n}$.

Some researchers think that Bayesian model averaging is better than model selection. However, Theorem 3 shows that, from a frequentist point of view, this is not necessarily true. When minimax-rate estimation of the regression function is the concern, Bayesian model averaging is worse than AIC, for example.

It should be pointed out that the suboptimality in rate of Bayesian model averaging is not due to unboundedness of the parameters; the problem actually lies around $\beta = 0$. Even if the parameter spaces are compact, the Bayesian model averaging regression estimator still cannot converge at rate $1/n$ uniformly over β. This suboptimality of Bayesian model averaging is somewhat unexpected. We are quite used to asymptotically optimal performance of Bayesian procedures, such as asymptotic efficiency of the Bayes estimator. For a single parametric family, as is well known, Bayes estimators and minimax estimation are closely related. For our regression problem, under both Model 1 and Model 0, when the parameter space is compact, that is α and β are bounded, the Bayes estimator of the regression function under the squared error loss and with a reasonable prior does converge at rate $1/n$ uniformly over the parameter space. However, when multiple models are involved, Theorem 3 shows that the Bayes estimator can no longer be minimax-rate optimal. We suspect that this suboptimality also holds generally in other contexts, such as density estimation, which means that, as soon as one is willing to assign a positive prior probability to any model smaller than a model already being considered, the Bayesian model averaging estimator of the function can have risk that is arbitrarily worse than the minimax risk; that is the risk ratio goes to infinity as $n \to \infty$. In our view, this suggests that, from a function estimation or prediction standpoint, Bayesian model averaging excessively favours small models in the sense that, although its underfitting

probability goes to zero asymptotically, the finite sample behaviour can be undesirable with the underfitting probability non-negligible from the regression estimation perspective. We emphasise that this weakness is not a result of a poor usage, for example due to approximation, of the Bayesian approach. The problem exists no matter how the priors are assigned, as long as the priors do not depend on the sample size.

Empirical Bayes approaches have also been proposed for model averaging, in which the priors on the models and the parameters in the models are chosen with the help of some information in the data, such as summary statistics or just the sample size; many Bayesian model averaging methods in the literature are actually empirical Bayesian model averaging methods. Theorem 3 is not applicable to such methods and therefore it does not preclude them from being minimax-rate optimal. However, while there is no question that the empirical Bayesian model averaging methods can be very useful, once they use data-dependent priors, it becomes unclear if they share the properties of formal Bayes methods on model averaging.

Raftery & Zheng (2003) argue that Bayesian model averaging is optimal in terms of long-run performance provided that the practical distributions of the parameters and the models are equal to or close to the working prior distributions used for Bayesian model averaging. This is true, but, when empirical Bayesian model averaging is used, such optimalities no longer hold or hold approximately. It seems to us that model averaging is most useful when there is a lack of prior information about the model and parameter distributions, and thus it may not be reasonable to expect the working prior distributions to be similar to the practical or 'true' distributions of the models and parameters. When the working prior distributions are sample-dependent, one is in a sense already admitting that one does not have trustworthy prior distributions. Consequently the minimisation of the Bayes risk with respect to a set of convenient working priors on the models and the parameters by the empirical Bayesian model averaging estimator may be far from minimising the Bayes risk with respect to a set of priors on the models and parameters that reasonably represents the long-run uncertainty of the models and the parameters. Of course, one may argue that the empirical priors may eventually become reasonably close to the practical distributions. However, not only does this need to be justified rigorously but also asymptotic arguments are particularly unreliable when model selection uncertainty is high, which is a main argument for the use of model averaging. For such situations, uniform risk bounds are more meaningful ways of characterising performance of statistical procedures. In any event, if an empirical Bayesian model averaging method can be shown to be minimax-rate optimal, which we feel to be very useful, then by Theorem 2 it cannot be consistent in selection and thus it is very different from any formal Bayesian model averaging method.

Acknowledgement

The author thanks two anonymous referees and the editor for their helpful comments. This work was supported by a CAREER Grant from the U.S. National Science Foundation.

Appendix

Proofs

Proof of Theorem 1. The key idea in the proof is to reduce the problem to a hypothesis testing problem to which classical hypothesis testing theory can be applied.

We first prove Theorem 1 in a simple case. Suppose that we have Model 0 and Model 1, as given in § 4.

Now consider a consistent 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

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

and thus the mean average squared error is

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

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

$$n \sup_{|\beta| \leqslant c} E_\beta (\hat{\beta} I_{A_n} - \beta)^2 \to \infty.$$

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

$$\begin{aligned} \sup_{|\beta| \leqslant c} E_\beta (\sqrt{n}\hat{\beta} I_{A_n} - \sqrt{n}\beta)^2 &= \sup_{|\beta| \leqslant c} E_\beta \{ \sqrt{n}(\hat{\beta} - \beta) I_{A_n} - \sqrt{n}\beta I_{A_n^c} \}^2 \\ &= \sup_{|\beta| \leqslant c} \{ E_\beta n (\hat{\beta} - \beta)^2 I_{A_n} + n\beta^2 \operatorname{pr}_\beta(A_n^c) \}. \end{aligned}$$

Thus, to show that δ is not minimax-rate optimal at rate $1/n$, it suffices to show that, for each $c > 0$, $\sup_{|\beta| \leqslant c} n\beta^2 \operatorname{pr}_\beta(A_n^c) \to \infty$. Since δ is consistent, we have $\operatorname{pr}_{\beta=0}(A_n) \to 0$ as $n \to \infty$. Consider a testing problem as follows. The observations are from the model

$$Y_i = \beta x_i + \varepsilon_i \quad (i = 1, 2, \ldots, n) \tag{A1}$$

in which the errors are independent and have standard normal distributions. Note that this is a sub-family of (3) with $\alpha = 0$ and $\sigma^2 = 1$. Consider the hypotheses $H_0 : \beta = 0$ and $H_1 : \beta > 0$. If we take the rejection region A_n, δ becomes a testing rule with probability of Type I error approaching zero. We next show by the Neyman–Pearson Lemma that, for any test here with the probability of Type I error going to zero, it necessarily holds that $\sup_{|\beta| \leqslant c} n\beta^2 \operatorname{pr}(\tilde{A}_n^c) \to \infty$, where $\tilde{A}_n$ is the rejection region of the test. Let $f(y_1, \ldots, y_n; \beta)$ denote the joint probability density function of $(Y_1, \ldots, Y_n)$ under (A1). Note that, for $\beta_1 > \beta_0 \geqslant 0$,

$$\begin{aligned} \frac{f(y_1, \ldots, y_n; \beta_1)}{f(y_1, \ldots, y_n; \beta_0)} &= \exp\left[\frac{1}{2} \sum_{i=1}^{n} \{ (y_i - \beta_0 x_i)^2 - (y_i - \beta_1 x_i)^2 \} \right] \\ &= \exp\left\{ (\beta_1 - \beta_0) \sum_{i=1}^{n} x_i y_i + \frac{1}{2} (\beta_0^2 - \beta_1^2) \sum_{i=1}^{n} x_1^2 \right\}. \end{aligned}$$

Thus the family $\{ f(y_1, \ldots, y_n; \beta) : \beta \geqslant 0 \}$ 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 test exists, which is to reject H_0 when $\sum_{i=1}^n x_i Y_i$ is larger than some constant C. Let $C = d_n$ so that $\operatorname{pr}_{\beta=0}(\sum_{i=1}^n x_i Y_i \geqslant d_n) = \operatorname{pr}_{\beta=0}(A_n)$. Let $A_{n,*}$ denote the event $\{ \sum_{i=1}^n x_i Y_i \geqslant d_n \}$. By the uniformly most powerful property of $A_{n,*}$, we have that, for all $\beta > 0$, $\operatorname{pr}_\beta(A_{n,*}) \geqslant \operatorname{pr}_\beta(A_n)$. Consequently,

$$\sup_{|\beta| \leqslant c} n\beta^2 \operatorname{pr}_\beta(A_n^c) \geqslant \sup_{0 \leqslant \beta \leqslant c} n\beta^2 \operatorname{pr}_\beta(A_{n,*}^c).$$

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

$$\operatorname{pr}_{\beta=0}\left(\sum_{i=1}^{n} x_i Y_i \geqslant d_n \right) = \operatorname{pr}\left\{ N(0, 1) \geqslant \frac{d_n}{\sqrt{(\sum x_i^2)}} \right\},$$

and, for $\beta > 0$,

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

Since $\mathrm{pr}_{\beta=0}(\sum_{i=1}^n x_i Y_i \geqslant d_n) = \mathrm{pr}_{\beta=0}(A_n) \to 0$, we must have that $d_n/\sqrt{n} \to \infty$. Then, with the choice of $\beta_n = \min(2d_n/\sum x_i^2, c)$, we have

$$\sup_{0 \leqslant \beta \leqslant c} n\beta^2 \,\mathrm{pr}_\beta(A_{n,*}^c) \geqslant n\beta_n^2 \,\mathrm{pr}_{\beta_n}(A_{n,*}^c).$$

Clearly $n\beta_n^2 \to \infty$. Also $\mathrm{pr}_{\beta_n}(A_{n,*}^c) \geqslant \mathrm{pr}\{N(0,1) < 2d_n/\sqrt{(\sum x_i^2)}\}$ and thus $\mathrm{pr}_{\beta_n}(A_{n,*}^c)$ converges to 1. It follows that $\sup_{|\beta| \leqslant c} n\beta^2 \,\mathrm{pr}(A_n^c) \to \infty$. 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, so that $\mathscr{F}_{k_1} = \{f_{k_1}(x, \theta_{k_1}) : \theta_{k_1} \in \Theta_{k_1}\}$ is a linear subspace of $\mathscr{F}_{k_2} = \{f_{k_2}(x, \theta_{k_2}) : \theta_{k_2} \in \Theta_{k_2}\}$. Let $\varphi(x)$ be a function in $\mathscr{F}_{k_2}$ that is orthogonal, at the design points, to $\mathscr{F}_{k_1}$. Under Assumption 1, $n^{-1}\sum_{i=1}^n \varphi^2(x_i)$ is bounded between two positive constants. Also, under the third part of Assumption 1, there is a function $f_0 \in \mathscr{F}_{k_1}$ such that f_0 does not belong to any other $\mathscr{F}_k$ that does not contain $\mathscr{F}_{k_1}$, 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 $\mathrm{pr}_{f_0}(B_n) \to 0$ as $n \to \infty$. Consider the simplified model

$$Y_i = f_0(x_i) + \beta\varphi(x_i) + \varepsilon_i \quad (i = 1, 2, \ldots, n) \tag{A2}$$

and the problem of testing $H_0 : \beta = 0$ versus $H_1 : \beta > 0$. Note that, under H_0, the data come from Model k_1 and, under H_1, the regression function is in $\mathscr{F}_{k_2}$. The model selection rule δ can be used to construct 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 the probability of Type I error going to zero as $n \to \infty$.

Let $F = (f(x_1), \ldots, f(x_n))'$, $Y = (Y_1, \ldots, Y_n)'$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$, $\varphi = (\varphi(x_1), \ldots, \varphi(x_n))'$ and let M_{k_1} be the projection matrix of model k_1. Observe that, under (A2),

$$\|F - M_{k_1} Y\|_n^2 = \|F - M_{k_1} F\|_n^2 + \varepsilon' M_{k_1} \varepsilon = \|\beta\varphi - M_{k_1}\varphi\|_n^2 + \varepsilon' M_{k_1}\varepsilon = \beta^2 \|\varphi\|_n^2 + \varepsilon' M_{k_1}\varepsilon,$$

where the second and the third equalities follow because $(f_0(x_1), \ldots, f_0(x_n))'$ is in the column space of M_{k_1} and φ is orthogonal to the column space of M_{k_1}. Then, under (A2), the risk of the estimator associated with δ is

$$\begin{aligned} R(f; \delta; n) &= \frac{1}{n} \sum_{k \in \Gamma} E_\beta \|F - M_k Y\|_n^2 I_{\{\hat{k} = k\}} \\ &\geqslant \frac{1}{n} E_\beta \|F - M_{k_1} Y\|_n^2 I_{\{\hat{k} = k_1\}} \\ &\geqslant \frac{\beta^2}{n} E_\beta \|\varphi\|_n^2 I_{\{\hat{k} = k_1\}} \\ &= \frac{\sum_{i=1}^n \varphi^2(x_i)}{n} \beta^2 \,\mathrm{pr}_\beta(\hat{k} = k_1). \end{aligned}$$

Thus, to show that $n \sup_{f \in \mathscr{F}_{k_2}} R(f; \delta; n) \to \infty$, it suffices to show that $\sup_{|\beta| \leqslant c} n\beta^2 \,\mathrm{pr}_\beta(B_n^c) \to \infty$. With our set-up 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 $\mathrm{pr}_{\beta=0}(A_n) \to 0$ we must have that $\sup_{|\beta| \leqslant c} n\beta^2 \,\mathrm{pr}_\beta(A_n^c) \to \infty$. Let $Z_i = Y_i - f_0(x_i)$. Then $Z_1, \ldots, Z_n$ are independent Gaussian random variables with $Z_i \sim N\{\beta\varphi(x_i), \sigma^2\}$. The earlier arguments for the simple two-model case follow similarly for proving the last assertion. This completes the proof of Theorem 1. □

The approach of using two points in the parameter space for deriving lower bounds on statistical risks of estimators has a long history (Le Cam, 1973); see Brown & Low (1996) for a use of this approach for obtaining a constrained risk inequality.

Proof of Theorem 2. As in the proof of Theorem 1, it suffices to handle the case of the two simple models, Model 1 and Model 0.

Assume that a model combining procedure τ is consistent in weighting. Then we have a consistent model selection rule: choose Model 1 if $W_1 \geqslant \frac{1}{2}$ and Model 0 otherwise. Let A_n denote the event that Model 1 is selected. Again, for $\beta \neq 0$, without loss of generality, assume that $\beta > 0$. Let $\hat{f}_1(x) = \hat{\alpha} + \hat{\beta}x$ and $\hat{f}_0(x) = \hat{\alpha}$.

The risk of the combined estimator is

$$\begin{aligned} R(f; \tau; n) &= \frac{1}{n} \sum_{i=1}^{n} E\{W_1 \hat{f}_1(x_i) + W_0 \hat{f}_0(x_i) - f(x_i)\}^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} E\{\hat{\alpha} - \alpha + (W_1 \hat{\beta} - \beta)x_i\}^2 \\ &= E(\hat{\alpha} - \alpha)^2 + \frac{1}{n} \sum_{i=1}^{n} x_i^2 E(W_1 \hat{\beta} - \beta)^2 \\ &= \frac{\sigma^2}{n} + \frac{1}{n} \sum_{i=1}^{n} x_i^2 E(W_1 \hat{\beta} - \beta)^2, \end{aligned}$$

where, for the third equality, the cross-product term disappears because $\sum x_i = 0$. It follows that it suffices to show that $\sup_{|\beta| \leqslant c} nE_\beta(W_1 \hat{\beta} - \beta)^2 \to \infty$ as $n \to \infty$. Note that, when $|\hat{\beta} - \beta| \leqslant \beta/2$ and $W_1 < \frac{1}{2}$, we have $|W_1 \hat{\beta} - \beta| > \beta/4$. Thus

$$nE_\beta(W_1 \hat{\beta} - \beta)^2 \geqslant nE_\beta(W_1 \hat{\beta} - \beta)^2 I_{A_n^c} \geqslant \frac{n\beta^2}{16} \operatorname{pr}_\beta(W_1 < \tfrac{1}{2} \text{ and } |\hat{\beta} - \beta| \leqslant \beta/2).$$

Now with $\beta = \beta_n$, as in the proof of Theorem 1, we have

$$\begin{aligned} \operatorname{pr}_{\beta_n}(W_1 < \tfrac{1}{2} \text{ and } |\hat{\beta} - \beta_n| \leqslant \beta_n/2) &\geqslant \operatorname{pr}_{\beta_n}(W_1 < \tfrac{1}{2}) - \operatorname{pr}_{\beta_n}(|\hat{\beta} - \beta_n| \geqslant \beta_n/2) \\ &= \operatorname{pr}_{\beta_n}(W_1 < \tfrac{1}{2}) - \operatorname{pr}\left\{|Z| \geqslant \frac{\beta_n \sqrt{(\sum_{i=1}^n x_i^2)}}{2\sigma}\right\}, \end{aligned}$$

where, for the equality, we use the normality of $\hat{\beta}$ and Z denotes a standard normal random variable. As in the proof of Theorem 1, $\operatorname{pr}_{\beta_n}(W_1 < \frac{1}{2}) \geqslant \operatorname{pr}_{\beta_n}(A_{n,*}^c)$ and the two probabilities both converge to 1 as $n \to \infty$. Obviously, $\operatorname{pr}\{|Z| \geqslant \beta_n \sqrt{(\sum_{i=1}^n x_i^2)}/(2\sigma)\}$ goes to zero. Since also $n\beta_n^2 \to \infty$, we conclude that $\sup_{|\beta| \leqslant c} nE_\beta(W_1 \hat{\beta} - \beta)^2 \to \infty$. This completes the proof of Theorem 2. □

Proof of Theorem 3. Let $\hat{\alpha}_B$ and $\hat{\beta}_B$ be the posterior means of α and β respectively under Model 1, and let $\tilde{\alpha}_B$ be the posterior mean of α under Model 0. Let W_1 and W_0 be the posterior probabilities of Model 1 and Model 0 respectively. Then the Bayes estimator of $f(x)$ under the squared error loss is $W_1(\hat{\alpha}_B + \hat{\beta}_B x) + W_0 \tilde{\alpha}_B$. Consequently, as before, the risk of this procedure is

$$\begin{aligned} R(f; \psi; n) &= \frac{1}{n} \sum_{i=1}^{n} E_\beta\{(W_1 \hat{\alpha}_B + W_0 \tilde{\alpha}_B - \alpha) + (W_1 \hat{\beta}_B - \beta)x_i\}^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} E_\beta(W_1 \hat{\alpha}_B + W_0 \tilde{\alpha}_B - \alpha)^2 + \frac{1}{n} \sum_{i=1}^{n} x_i^2 E_\beta(W_1 \hat{\beta}_B - \beta)^2. \end{aligned}$$

Similarly to the proofs of Theorems 1 and 2, we only need to show that $\mathrm{pr}_{\beta_n}(|\hat{\beta}_B - \beta_n| \geqslant \beta_n/2)$ is bounded away from 1 as $n \to \infty$. Based on $\mathrm{pr}_{\beta=0}(\sum_{i=1}^n x_i Y_i \geqslant d_n) = \mathrm{pr}_{\beta=0}(W_1 \geqslant \frac{1}{2})$, we can easily find that

$$d_n = \sigma \sqrt{\left(\sum_{i=1}^n x_i^2 \right)} \Phi^{-1}\{1 - \mathrm{pr}_{\beta=0}(W_1 \geqslant \tfrac{1}{2})\},$$

where Φ^{-1} denotes the inverse of the cumulative distribution function of the standard normal distribution. Since $\mathrm{pr}_{\beta=0}(W_1 \geqslant \frac{1}{2})$ converges to zero, we know that d_n is of a higher order than $1/\sqrt{n}$. Clearly, with $\beta_n = \min\{d_n/(2\sum_{i=1}^n x_i^2), c\}$, β_n is of a higher order than $1/\sqrt{n}$. The conclusion follows under Assumption 2. This completes the proof of Theorem 3. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Info. Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267–81. Budapest: Akademia Kiado.

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

Barron, A. R., Yang, Y. & Yu, B. (1994). Asymptotically optimal function estimation by minimum complexity criteria. In *Proc. 1994 Int. Symp. Info. Theory*, p. 38. Trondheim, Norway: IEEE Info. Theory Soc.

Berger, J. O. & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: introduction and comparison (with Discussion). In *Model Selection*, Ed. P. Lahiri, pp. 135–207. Institute of Mathematical Statistics Lecture Notes—Monograph Series Volume 38. Beachwood, OH: Inst. Math. Statist.

Birgé, L. & Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203–68.

Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. Englewood Cliffs, NJ: Prentice-Hall.

Brown, L. D. & Low, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24**, 2524–35.

Brown, L. D., Low, M. G. & Zhao, L. H. (1997). Superefficiency in nonparametric function estimation. *Ann. Statist.* **25**, 2607–25.

Clyde, M., DeSimone, H. & Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *J. Am. Statist. Assoc.* **91**, 1197–208.

Donoho, D. L. & Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879–921.

Foster, D. P. & George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–75.

George, E. I. & Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–47.

Hand, D. J. & Vinciotti, V. (2003). Local versus global models for classification problems: fitting models where it matters. *Am. Statistician* **57**, 124–31.

Hansen, M. & 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: IEEE Info. Theory Soc.

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

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1**, 38–53.

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

Li, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15**, 958–75.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661–75.

McQuarrie, A. D. R. & Tsai, C. L. (1998). *Regression and Time Series Model Selection*. Singapore: World Scientific Publications.

Nishii, R. (1084). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12**, 758–65.

Polyak, B. T. & Tsybakov, A. B. (1991). Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. *Theory Prob. Applic.* **35**, 293–306.

Raftery, A. E. & Zheng, Y. (2003). Comment on 'The focused information criterion' by Hjort and Claeskens. *J. Am. Statist. Assoc.* **98**, 931–8.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica* **14**, 465–71.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.

SHAO, J. (1997). An asymptotic theory for linear model selection (with Discussion). *Statist. Sinica* **7**, 221–42.
SHEN, X. & YE, J. (2002). Adaptive model selection. *J. Am. Statist. Assoc.* **97**, 210–21.
SHIBATA, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Ann. Inst. Statist. Math.* **35**, 415–23.
YANG, Y. (1999). Model selection for nonparametric regression. *Statist. Sinica* **9**, 475–99.
YANG, Y. (2003). Regression with multiple candidate models: selecting or mixing? *Statist. Sinica* **13**, 783–809.
YANG, Y. & BARRON, A. R. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27**, 1564–99.

[*Received August* 2004. *Revised April* 2005]