Ann. Inst. Statist. Math.
Vol. 50, No. 1, 1-13 (1998)

AN APPLICATION OF MULTIPLE COMPARISON TECHNIQUES TO MODEL SELECTION

HIDETOSHI SHIMODAIRA*

Department of Mathematical Engineering and Information Physics, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113, Japan

(Received April 15, 1996; revised March 26, 1997)

Abstract. Akaike's information criterion (AIC) is widely used to estimate the best model from a given candidate set of parameterized probabilistic models. In this paper, considering the sampling error of AIC, a set of good models is constructed rather than choosing a single model. This set is called a confidence set of models, which includes the minimum $\mathcal{E}\{\mathrm{AIC}\}$ model at an error rate smaller than the specified significance level. The result is given as P-value for each model, from which the confidence set is immediately obtained. A variant of Gupta's subset selection procedure is devised, in which a standardized difference of AIC is calculated for every pair of models. The critical constants are computed by the Monte-Carlo method, where the asymptotic normal approximation of AIC is used. The proposed method neither requires the full model nor assumes a hierarchical structure of models, and it has higher power than similar existing methods.

Key words and phrases: Akaike's information criterion, model selection, confidence set, multiple comparison with the best, Gupta's subset selection, variable selection, multiple regression, bootstrap resampling.

1. Introduction

Since Akaike (1974) advocated the model selection criterion AIC $= -2 \times$ (maximum log-likelihood) $+ 2 \times$ (the number of parameters), the minimum AIC estimate (MAICE), the model which minimizes AIC, has been widely used as a simple and practical estimate of the best model. MAICE is said to be free from the following problems of the classical hypothesis testing: (A) The arbitrariness of the choice of significance level, and (B) the limitation on the structure of candidate models.

However, the arbitrariness in (A) may be seen as an *advantage* of testing; the significance level can be chosen according to how strictly the error should

* Now at The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan.

be controlled. On the other hand, the "significance level" of MAICE cannot be prescribed by the user, because it is implicitly adjusted as pointed out by Bozdogan (1987).

In this paper, we consider a model selection procedure which has the advantage in (A) of leading to a quantitative measure of its reliability. At the same time, it is free from the limitation in (B). The key idea is to consider tests on the magnitude of $\mathcal{E}\{\mathrm{AIC}\}$, then to find a *confidence set of models*. This set includes the (unknown) best model at a high probability while the number of models in it is kept small. MAICE will be always included in it at a reasonable significance level ($<1/2$). The confidence set is not to replace MAICE; rather it provides us supplemental information on model selection. The confidence set is regarded as an "interval" estimate of the best model, and MAICE is regarded as a "point" estimate.

Here note that the best model is defined as the minimizer of $\mathcal{E}\{\mathrm{AIC}\}$ rather than MAICE. This is natural if we remember that AIC is derived as an estimate of the *expected prediction error* unbiased up to its second term. Note also that the expectation is taken with respect to the unknown true distribution, which may not be included in any of the candidate parametric models.

To illustrate our approach, consider the simplest case where we have two (possibly nonnested) models M_0 and M_1 to be compared. First, test the null hypothesis $\mathcal{E}\{\mathrm{AIC}_0\} \leq \mathcal{E}\{\mathrm{AIC}_1\}$ against the alternative $\mathcal{E}\{\mathrm{AIC}_0\} > \mathcal{E}\{\mathrm{AIC}_1\}$, and include M_0 in the confidence set unless the null hypothesis is rejected. Next, repeat the same but the roles of M_0 and M_1 are interchanged. The confidence set will be one of $\{M_0\}$, $\{M_1\}$, and $\{M_0, M_1\}$. A standardized difference of AIC will be used as the test statistic, which is asymptotically the standard normal $N(0,1)$ if $\mathcal{E}\{\mathrm{AIC}_0\} = \mathcal{E}\{\mathrm{AIC}_1\}$ as shown in Linhart (1988), and more generally in Vuong (1989).

Our approach is different from the significance test of M_0 against M_1. Cox (1962) derived a significance test for separated two models, and it appears to be similar to the procedure described above. However, the underlying concept is very different; Cox tested the null hypothesis that M_0 includes the true distribution against the alternative that M_1 does, whereas we test $\mathcal{E}\{\mathrm{AIC}_0\} \leq \mathcal{E}\{\mathrm{AIC}_1\}$ against $\mathcal{E}\{\mathrm{AIC}_0\} > \mathcal{E}\{\mathrm{AIC}_1\}$.

It is not difficult to extend Linhart's test to the case where we have many possibly nonnested models. Denote the set of candidate models by $\{M_\alpha \mid \alpha \in \mathcal{M}\}$. For each $\alpha \in \mathcal{M}$, consider a test of the null hypothesis $H_\alpha : \mathcal{E}\{\mathrm{AIC}_\alpha\} \leq \min_{\beta \in \mathcal{M}\setminus\{\alpha\}} \mathcal{E}\{\mathrm{AIC}_\beta\}$, against the alternative $\mathcal{E}\{\mathrm{AIC}_\alpha\} > \min_{\beta \in \mathcal{M}\setminus\{\alpha\}} \mathcal{E}\{\mathrm{AIC}_\beta\}$, and include M_α in the confidence set unless H_α is rejected at a prescribed significance level. This construction makes sense since

$$\Pr\{\alpha^* \in \mathcal{T}\} = \Pr\{H_{\alpha^*} \text{ is not rejected}\} \geq 1 - \text{level}, \tag{1.1}$$

where $\mathcal{T}$ is the confidence set and α^* is the minimum $\mathcal{E}\{\mathrm{AIC}\}$ model.

Multiple comparison techniques (Gupta and Panchapakesan (1979), Hochberg and Tamhane (1987)) are used to test H_α. For every pair of models, a standardized difference of AIC is calculated, and then a variant of Gupta's subset selection is applied to the testing of H_α. Monte-Carlo calculation is used, since AIC_α's

are correlated and have different variances. We will use an asymptotic normal approximation to reduce the huge computation.

The result of our procedure will be given as P_α, the P-value of testing of H_α. The confidence set at a given significance level is easily obtained from P_α:

$$\mathcal{T} = \{\alpha \in \mathcal{M} \mid P_\alpha \geq \text{level}\}. \tag{1.2}$$

P_α is informative, since showing P_α, $\alpha \in \mathcal{M}$ is equivalent to showing all the possible values of $\mathcal{T}$ as a function of the level. P_α is regarded as a quantitative measure of the reliability that M_α is the best in the candidates.

The concept of the confidence set of models itself is not new. Several methods have been proposed in the literature, which will be discussed in Section 2. The method proposed in this paper has the following advantages in contrast to them: (i) Free from the limitation in (B); no need for the *full* model, nor the hierarchical structure. (ii) Applicable to any smooth probabilistic models. (iii) Relatively high power of the test of H_α; this is because the variance estimate of the AIC difference is computed for every pair of models. (iv) The clear probabilistic interpretation in (1.1). (v) P_α is informative.

The construction of this paper is as follows. In Section 2, the confidence set of models will be derived using multiple comparison techniques. Related methods are also discussed there. In Section 3, two examples are taken from the variable selection problem of multiple regression. In Section 4, the bootstrap estimate of P_α is given, and its normal approximation will be discussed in Section 5. In Section 6, a problem regarding higher order terms is discussed. In Section 7, a simulation result is given to illustrate the accuracy of the normal approximation. Remarks are made in Section 8.

2. Confidence set of models

Linhart (1988) derived a test of the null hypothesis $H_{\alpha\beta} : \mathcal{E}\{\text{AIC}_\alpha\} \leq \mathcal{E}\{\text{AIC}_\beta\}$, against the alternative $\mathcal{E}\{\text{AIC}_\alpha\} > \mathcal{E}\{\text{AIC}_\beta\}$ for each pair $\alpha, \beta \in \mathcal{M}$ as follows. Let $V_{\alpha\beta}$ be an estimate of $\text{var}\{\text{AIC}_\alpha - \text{AIC}_\beta\}$, where the variance is taken with respect to the true distribution, and let $T_{\alpha\beta} = (\text{AIC}_\alpha - \text{AIC}_\beta)/\sqrt{V_{\alpha\beta}}$ be the test statistic. Note that a specific form of $V_{\alpha\beta}$ will be given in (5.2). Consider a test to reject $H_{\alpha\beta}$ iff $T_{\alpha\beta} > c_{\alpha\beta}$, where $c_{\alpha\beta}$ is the critical constant. It has been shown that $S_{\alpha\beta} = T_{\alpha\beta} - \mathcal{E}\{\text{AIC}_\alpha - \text{AIC}_\beta\}/\sqrt{V_{\alpha\beta}}$ is asymptotically $N(0,1)$ under mild conditions. Thus defining $c_{\alpha\beta}$ by $\Phi(c_{\alpha\beta}) = 1 - \text{level}$, where $\Phi(x)$ is the standard normal distribution function, one has $\Pr\{\text{reject } H_{\alpha\beta}\} = \Pr\{T_{\alpha\beta} > c_{\alpha\beta}\} \leq \Pr\{S_{\alpha\beta} > c_{\alpha\beta}\} = \text{level}$ under $H_{\alpha\beta}$ asymptotically. A detailed discussion can be found in Vuong (1989).

Similarly we will have a test of H_α as follows. For each $\alpha \in \mathcal{M}$, reject H_α iff $T_\alpha > c_\alpha$, where $T_\alpha = \max_{\beta \in \mathcal{M}\setminus\{\alpha\}} T_{\alpha\beta}$ is the test statistic, and c_α is the critical constant determined from a prescribed significance level. Suppose we know the value of c_α defined by

$$\Pr\{S_\alpha \leq c_\alpha\} = 1 - \text{level}, \tag{2.1}$$

where the probability is taken with respect to the true distribution, and $S_\alpha = \max_{\beta \in \mathcal{M} \setminus \{\alpha\}} S_{\alpha\beta}$. Then it will be easily seen that $\Pr\{\text{reject } H_\alpha\} = \Pr\{T_\alpha > c_\alpha\} \leq \Pr\{S_\alpha > c_\alpha\} = \text{level under } H_\alpha$. The confidence set is $\mathcal{T} = \{\alpha \in \mathcal{M} \mid H_\alpha \text{ is not rejected at the given level}\}$. The test of H_α described here can be seen as a variant of Gupta's subset selection for unequal sample sizes case (Gupta and Huang (1976)). However, the correlations of AIC's are to be estimated in our situation, we will calculate the probability in (2.1) approximately by the Monte-Carlo method, which will be described in Sections 4–7.

The above construction of $\mathcal{T}$, denoted by Gupta's confidence set in this paper, is not the only possibility to control (1.1). If all H_α's are tested simultaneously rather than separately, we will obtain $\mathcal{T}$ which corresponds to Tukey's multiple comparison for unbalanced designs (Hochberg and Tamhane (1987), p. 85). This $\mathcal{T}$ controls the error rate to include all the tied best models simultaneously. To increase its power, the sequentially rejective procedure can be applied to it. This $\mathcal{T}$ will be denoted by Holm's confidence set in this paper. Practically, Gupta's $\mathcal{T}$ is preferable to them, since the size of $\mathcal{T}$ gets larger in the order of Gupta, Holm, and Tukey.

Besides our construction of $\mathcal{T}$, several procedures have been proposed for the variable selection of multiple regression. Spjøtvoll (1972) derived a confidence set to control (1.1), using the simultaneous confidence ellipsoid of regression coefficients for the full model. Also, a comprehensive model formed from the competing nonnested models can be used to construct the conventional statistics (Atkinson (1970), Dastoor and McAleer (1989)). On the other hand, Mallows (1973) and Aitkin (1974) derived a simultaneous significance test of all M_α's against the full model, and Spjøtvoll (1977) applied the closure method to it to increase the power. A limitation of these approaches is to require the full model; the use of large full model as a reference may lead to small power of testing, or inversely rejecting all the candidates except for the full model (Shimodaira (1997*b*)).

Arvesen and McCabe (1975) constructed $\mathcal{T}$ which controls (1.1) for the regression case. They employed a Gupta's subset selection, asymptotically equivalent to the form of $\mathcal{T} = \{\alpha \in \mathcal{M} \mid \mathrm{AIC}_\alpha - \min_\beta \mathrm{AIC}_\beta < c\sqrt{V}\}$, which is obtained from our $\mathcal{T}$ if $T_{\alpha\beta}$ is replaced by $(\mathrm{AIC}_\alpha - \mathrm{AIC}_\beta)/\sqrt{V}$. However, their $\mathcal{T}$ is larger than ours, since a pooled variance estimator V and a single critical constant c are used for all the models.

3. Examples

Before going into detailed discussion, here we look at numerical examples. The regression data set HALD is taken from Draper and Smith ((1981), p. 629), where the sample size is $n = 13$. The response variable is heat evolved in calories per gram of cement, and the four predictor variables are the amounts of four major ingredients in percentage. We are to find a good experimental formula to predict the response; each subset of predictors corresponds to a model denoted by the predictors in angle brackets. All the possible $2^4 = 16$ subsets are used as the candidates $\mathcal{M}$. See the Appendix for the statistics.

A plot of P_α against AIC is shown in Fig. 1. Gupta's P_α as well as that of Tukey and Holm are shown. These are obtained from a single Monte-Carlo run,

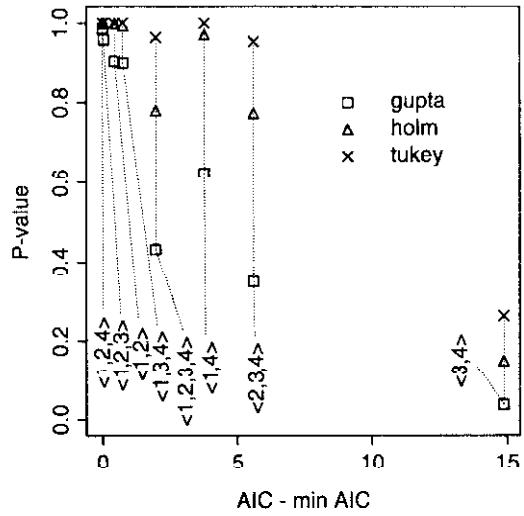


Fig. 1. AIC vs. P_α for the HALD data set. The eight models which have Gupta's P-value greater than 0.01 are shown. Seven models are selected for the confidence set of Gupta or Holm at level = 0.2; they are all the models which include one of $\langle 1,2\rangle$, $\langle 1,4\rangle$, and $\langle 2,3,4\rangle$. In addition, $\langle 3,4\rangle$ is selected by Tukey's P_α at the same level.

in which the number of replicated simulations is $N_b = 10{,}000$. Specify the level, say, 0.2. Then, Gupta's confidence set is

$$\mathcal{T} = \{\langle 1,2,4\rangle, \langle 1,2,3\rangle, \langle 1,2\rangle, \langle 1,3,4\rangle, \langle 1,4\rangle, \langle 1,2,3,4\rangle, \langle 2,3,4\rangle\}.$$

$\langle 1,2\rangle$, $\langle 1,4\rangle$, and $\langle 2,3,4\rangle$ are the minimal models of $\mathcal{T}$, which generate the seven models by inclusion. Interestingly, the same three models are chosen as the *minimal adequate sets* by Aitkin (1974) and Spjøtvoll (1977) in analyzing the same dataset.

The small sample size made $\mathcal{T}$ large. Not only MAICE $\langle 1,2,4\rangle$ is thought as a good model, but every model in $\mathcal{T}$ is possibly the best model; AIC values for the models in $\mathcal{T}$ are not significantly larger than that of MAICE. The simplest models $\langle 1,2\rangle$ and $\langle 1,4\rangle$ may be chosen, if parsimonious selection is preferred.

All the models which include three of the four predictors are found in $\mathcal{T}$. This is a consequence of the multicollinearity; the sum of the four predictors are approximately 100 in this data set. Thus any set of three predictors works fine as the full model. Simply selecting MAICE does not lead to this observation.

Next, the data set BOSTON is taken from Belsley *et al.* ((1980), p. 244), where $n = 506$. The response variable is the logarithm of the median value of owner-occupied homes for each area in Boston. There are thirteen predictor variables. For example, z_1 is per capita crime rate by town, or z_{13} is logarithm of the proportion of the population that is lower status. We consider 286 candidates; they contain three of the thirteen predictors. The result is shown in Fig. 2. At level = 0.2, every model in $\mathcal{T}$ includes $\langle 1,13\rangle$, which may be selected as a consensus model to summarize the result.

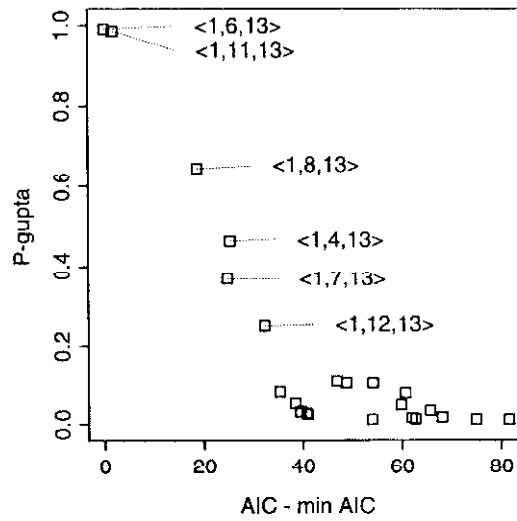


Fig. 2. AIC vs. $P_\sim$ for the BOSTON data set. Plotted for the 23 models that have Gupta's P-value greater than 0.01. Six models are selected at the level 0.2. At the same level, $|T| = 13$ with Holm, or $|T| = 25$ with Tukey.

Model $\langle 1, 12, 13\rangle$ has $P_\alpha = 0.25$, and so it is a good model relative to the candidates. But its AIC minus that of MAICE is 32, which seems quite large. Considering its variance, we have P-value of this pairwise comparison: $1 - \Phi(32/\sqrt{218}) = 0.015$, which is significant. When comparing many models simultaneously, we have to consider that some of the candidates may have small values of AIC just by chance. This probability is not reflected in AIC, nor in any single pairwise comparison.

4. Bootstrap estimate of P_α

In the subsequent four sections, we will discuss how to obtain an approximate value of the critical constant c_α defined in (2.1). This is equivalent to approximately evaluate the function

$$P_\alpha(s) = 1 - \Pr\{S_\alpha \le s\}, \tag{4.1}$$

from which $P_\alpha = P_\alpha(T_\alpha)$ is obtained.

First, we consider a straightforward application of the bootstrap method. Assume we observed n i.i.d. samples of random variable x, and denote $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$. Let $\mathrm{AIC}_\alpha[\mathcal{X}]$ be the AIC value for M_α computed from $\mathcal{X}$. Similarly $V_{\alpha\beta}[\mathcal{X}]$ denotes the estimate of $\mathrm{var}\{\mathrm{AIC}_\alpha - \mathrm{AIC}_\beta\}$ computed from $\mathcal{X}$. Having $\tilde{\mathcal{X}}$, a bootstrap resampling of size n drawn from $\mathcal{X}$, we obtain a bootstrap replication of S_α:

$$\tilde{S}_\alpha = \max_{\beta \in \mathcal{M}\setminus\{\alpha\}} \frac{\mathrm{AIC}_\alpha[\tilde{\mathcal{X}}] - \mathrm{AIC}_\beta[\tilde{\mathcal{X}}] - \mathcal{E}_{\mathcal{X}}\{\mathrm{AIC}_\alpha[\tilde{\mathcal{X}}] - \mathrm{AIC}_\beta[\tilde{\mathcal{X}}]\}}{\sqrt{V_{\alpha\beta}[\mathcal{X}]}}.$$

Note that $\mathcal{E}_{\mathcal{X}}$ denotes the expectation with respect to the resampling from $\mathcal{X}$, and $\mathcal{E}_{\mathcal{X}}\{\mathrm{AIC}_\alpha[\tilde{\mathcal{X}}]\}$ above can be replaced by $\mathrm{AIC}_\alpha[\mathcal{X}] - m_\alpha$, where m_α denotes the dimension of M_α. The bootstrap estimate of (4.1) is

$$\hat{P}_\alpha(s) = 1 - \Pr_{\mathcal{X}}\{\tilde{S}_\alpha \le s\}, \tag{4.2}$$

where $\Pr_{\mathcal{X}}$ denotes the probability with respect to the resampling.

In actual calculation, bootstrap resamples $\tilde{\mathcal{X}}_1, \ldots, \tilde{\mathcal{X}}_{N_b}$ are generated for a sufficiently large N_b, and replicates $\tilde{S}_{\alpha 1}, \ldots, \tilde{S}_{\alpha N_b}$ are computed to find an estimate of (4.2): $\hat{P}_\alpha(s) \approx 1 - \#\{\tilde{S}_\alpha \le s\}/N_b$. Unfortunately, it requires huge computation in most applications, though. In the next section, we consider a further approximation of the simple bootstrap to reduce the computation.

5. Normal approximation

Let $p_\alpha(x \mid \theta_\alpha)$ be the probabilistic model M_α parameterized by $\theta_\alpha \in \Theta_\alpha \subset \mathcal{R}^{m_\alpha}$, and $L_\alpha(\theta_\alpha) = (1/n)\sum_{t=1}^n \log p_\alpha(x_t \mid \theta_\alpha)$ be the log-likelihood function (divided by n). Let $\hat{\theta}_\alpha$ denote the maximum likelihood estimator (MLE), the maximizer of $L_\alpha(\theta_\alpha)$ over Θ_α. Then, $\mathrm{AIC}_\alpha = -2nL_\alpha(\hat{\theta}_\alpha) + 2m_\alpha$.

Under mild assumptions, $\hat{\theta}_\alpha = \theta^*_\alpha + O_p(1/\sqrt{n})$, and so $L_\alpha(\hat{\theta}_\alpha) = L_\alpha(\theta^*_\alpha) + O_p(1/n)$, where θ^*_α is the maximizer of $\mathcal{E}\{\log p_\alpha(x \mid \theta_\alpha)\}$ over Θ_α (White (1982)). Then, it follows from the central limit theorem that $L_\alpha(\hat{\theta}_\alpha)$ is asymptotic normal with mean $\mathcal{E}\{\log p_\alpha(x \mid \theta^*_\alpha)\}$ and variance $\mathrm{var}\{\log p_\alpha(x \mid \theta^*_\alpha)\}/n$, since $\log p(x_t \mid \theta^*_\alpha)$, $t = 1, \ldots, n$ are i.i.d. samples.

Quite similarly $(L_\alpha(\hat{\theta}_\alpha) : \alpha \in \mathcal{M})$ is multivariate asymptotic normal, and then

$$\begin{aligned} 1 - P_\alpha(s) &= \Pr\{\sqrt{n}(-L_\alpha(\hat{\theta}_\alpha) + L_\beta(\hat{\theta}_\beta) - \mathcal{E}\{-L_\alpha(\hat{\theta}_\alpha) + L_\beta(\hat{\theta}_\beta)\}) \\ &\qquad \le s\sqrt{V_{\alpha\beta}/4n},\ \forall \beta \in \mathcal{M} \setminus \{\alpha\}\} \\ &= \Pr\{U_\alpha - U_\beta + O_p(1/\sqrt{n}) \le s\sqrt{V^*_{\alpha\beta}/4n},\ \forall \beta \in \mathcal{M} \setminus \{\alpha\}\}, \end{aligned} \tag{5.1}$$

where $(U_\alpha : \alpha \in \mathcal{M})$ is multivariate normal with mean 0 and $\mathrm{var}\{U_\alpha - U_\beta\} = V^*_{\alpha\beta}/4n = \mathrm{var}\{\log p_\alpha(x \mid \theta^*_\alpha) - \log p_\beta(x \mid \theta^*_\beta)\}$ for $\alpha, \beta \in \mathcal{M}$. Note that $V_{\alpha\beta}/n = V^*_{\alpha\beta}/n + O_p(1/\sqrt{n})$ was assumed in the last equation.

A normal approximated $\hat{P}_\alpha(s)$ is obtained from (5.1): Replace $\tilde{S}_\alpha$ in (4.2) by $\max_{\beta \in \mathcal{M}\setminus\{\alpha\}}(\tilde{U}_\alpha - \tilde{U}_\beta)/\sqrt{V_{\alpha\beta}/4n}$, where $(\tilde{U}_\alpha : \alpha \in \mathcal{M})$ is multivariate normal with mean 0 and $\mathrm{var}\{\tilde{U}_\alpha - \tilde{U}_\beta\} = V_{\alpha\beta}/4n$. Note that the probability calculation of (5.1) does not change even if all U_α, $\alpha \in \mathcal{M}$ are replaced by $U_\alpha + W$ using any random variable $W = O_p(1)$. Thus, an estimate of the covariance matrix of $\log p_\alpha(x \mid \theta^*_\alpha)$, which has $|\mathcal{M}|(|\mathcal{M}|+1)/2$ degree of freedom, is not needed, but only $|\mathcal{M}|(|\mathcal{M}|-1)/2$ elements of $V_{\alpha\beta}$'s are needed for the Monte-Carlo sampling of $\tilde{U}_\alpha$'s.

The estimate of $\mathrm{var}\{\mathrm{AIC}_\alpha - \mathrm{AIC}_\beta\}$ will be the form of

$$V_{\alpha\beta} = 4n\,\mathrm{var}_{\mathcal{X}}\{\log p_\alpha(\tilde{x} \mid \hat{\theta}_\alpha) - \log p_\beta(\tilde{x} \mid \hat{\theta}_\beta)\} + 2v_{\alpha\beta}, \tag{5.2}$$

where $\mathrm{var}_{\mathcal{X}}$ denotes the variance with respect to the resampling from $\mathcal{X}$, or an estimate of it. Higher order terms may be included in the second term $2v_{\alpha\beta} = O_p(1)$, which is discussed in the next section. It is shown in Shimodaira (1997*a*) that the addition of $2v_{\alpha\beta}$ improves the accuracy of the normal approximation considerably, at least in the case of a single comparison. The improvement results also in the multiple comparison case as can be seen in Section 7.

6. Higher order terms

The discussion of the previous section is not complete. If $V^*_{\alpha\beta} = 0$, or equivalently $p_\alpha(x \mid \theta^*_\alpha) = p_\beta(x \mid \theta^*_\beta)$ a.e. in x, for some pairs of models, then the error term $O_p(1/\sqrt{n})$ in (5.1) will make the probability calculation meaningless (Vuong (1989)). In this paper, we assume $0 < V^*_{\alpha\beta}/n < \infty$ for all the pairs in $\mathcal{M}$. Then, the $O_p(1/\sqrt{n})$ term goes out of the probability statement in (5.1), and we obtain

$$1 - P_\alpha(s) = \Pr\{U_\alpha - U_\beta \le s\sqrt{V^*_{\alpha\beta}/4n},\ \forall \beta \in \mathcal{M} \setminus \{\alpha\}\} + O(1/\sqrt{n}), \tag{6.1}$$

which justifies the normal approximation of P_α.

In practice, it is unusual that $V^*_{\alpha\beta} = 0$ strictly holds. Thus the normal approximated $\hat{P}_\alpha(s)$ estimates $P_\alpha(s)$ consistently as $n \to \infty$. The problem is that n is finite. If $\sqrt{V^*_{\alpha\beta}/4n}$ in (5.1) is not large enough compared with the $O_p(1/\sqrt{n})$ term at a fixed n, the $O(1/\sqrt{n})$ term in (6.1) will not be negligible.

To remedy this problem, we added an estimate of $v^*_{\alpha\beta}$ to $V_{\alpha\beta}$ in (5.2), where $v^*_{\alpha\beta} = m_\alpha + m_\beta - 2\,\mathrm{tr}\, G^*_{\alpha\beta} G^{*-1}_\beta G^*_{\beta\alpha} G^{*-1}_\alpha$, and the elements of $m_\alpha \times m_\beta$ matrix $G^*_{\alpha\beta}$ are

$$G^*_{\alpha\beta\cdot ij} = \mathcal{E}\left\{ \frac{\partial \log p_\alpha(x \mid \theta^*_\alpha)}{\partial \theta^i_\alpha} \frac{\partial \log p_\beta(x \mid \theta^*_\beta)}{\partial \theta^j_\beta} \right\}.$$

To reduce the calculation, $v_{\alpha\beta}$ may be replaced by its upper bound $m_\alpha + m_\beta - 2\dim(M_\alpha \cap M_\beta)$. It is shown (Shimodaira (1997*a*)) that $\mathrm{var}\{\mathrm{AIC}_\alpha - \mathrm{AIC}_\beta\} \approx V^*_{\alpha\beta} + 2v^*_{\alpha\beta}$ under *local alternatives*, and $v_{\alpha\beta}$ in $V_{\alpha\beta}$ does work as a safeguard against the case $V^*_{\alpha\beta} \approx 0$; $v_{\alpha\beta}$ makes the testing of H_α be conservative rather than violate (1.1). For two nested models, the smaller model tends to be included in $\mathcal{T}$ than expected from the level. Note, however, that this introduction of $v_{\alpha\beta}$ is rather heuristic, because the asymptotic normality is derived under *fixed alternatives*, not local alternatives.

7. Numerical simulation

Here we see distributional behavior of MAICE and P_α for artificially generated regression data sets. The covariance matrix of HALD data set was used as that of the true distribution. The sample size for each data set is $n = 13$ unless specified.

First, Fig. 3 shows how many times each model was selected as MAICE in 1,000 replicated simulations. We observe that MAICE was scattered over the models of relatively small $\mathcal{E}\{\mathrm{AIC}\}$; MAICE is not necessarily the best model.

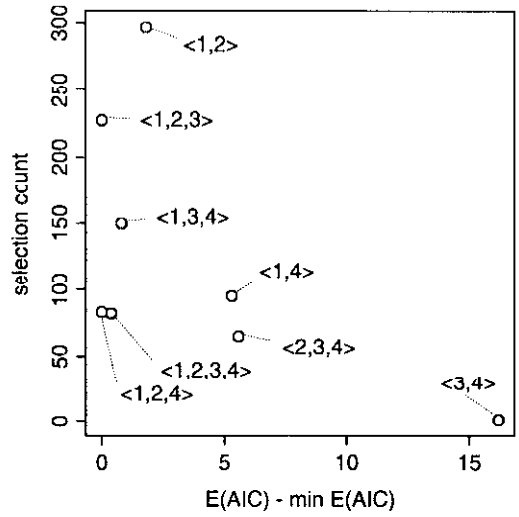


Fig. 3. $\mathcal{E}\{\text{AIC}\}$ vs. the count of selection as MAICE. Only the models selected at least once are shown. The artificial data were generated repeatedly 1,000 times. The minimum $\mathcal{E}\{\text{AIC}\}$ model is $\alpha^* = \langle 1, 2, 4\rangle$, but $\langle 1, 2\rangle$ has the largest selection count.

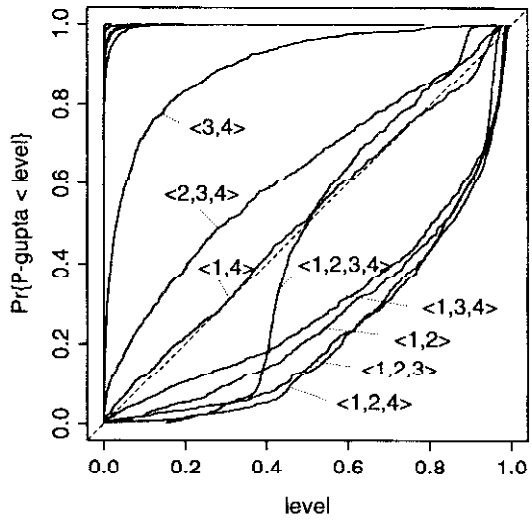
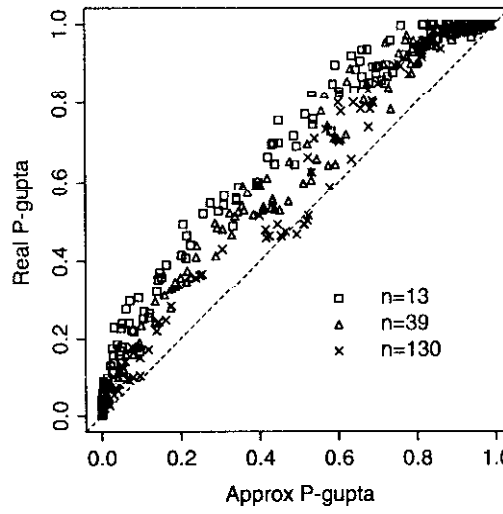
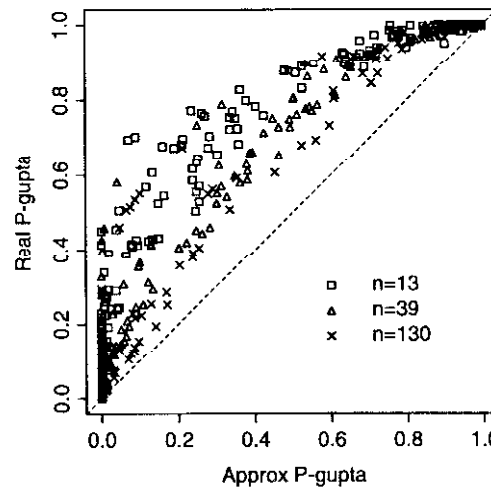


Fig. 4. Plot of $\Pr\{P_\alpha < \text{level}\}$ vs. level for the simulated data sets. $\Pr\{P_\alpha < \text{level}\}$ is the probability that M_α is excluded from $\mathcal{T}$ at the given level. $\alpha^* = \langle 1, 2, 4\rangle$ has the lowest probability of rejection, but $\langle 1, 2, 3\rangle$, $\langle 1, 2\rangle$, $\langle 1, 3, 4\rangle$, and $\langle 1, 2, 3, 4\rangle$ are also very likely included in $\mathcal{T}$. The rejection probabilities are larger than level for $\langle 1, 4\rangle$, $\langle 2, 3, 4\rangle$, $\langle 3, 4\rangle$, and the other eight models.

Next, Fig. 4 shows the distribution functions of P_α, $\alpha \in \mathcal{M}$, computed from 1,000 replicated simulations. P_α is calculated by the method of Section 5 with $N_b = 1{,}000$. We observe that $\Pr\{P_\alpha < \text{level}\} < \text{level}$ for several models other



(a)



(b)

Fig. 5. (a) Normal approximated P_α vs. real value of it for 20 replicated simulations. P_α is calculated with $N_b = 1{,}000$ for all $\mathcal{M}$. The approximated value approaches its real value as the sample size increases $n = 13$, 39, and 130. (b) Approximated P_α calculated without $v_{\alpha\beta}$.

than α^*. Note that $\Pr\{P_\alpha < \text{level}\}$ is the probability of rejecting H_α; it should be equal to or less than the level for α^*, and it is expected to be larger than level for the other models. So, the result implies that the confidence set derived from this P_α is conservative and includes many models than expected from level. One of the reasons is the normal approximation as seen in the next paragraph. Another reason is the multiple comparison itself; the critical constant is evaluated

at the *least favorable configuration*, where the $\mathcal{E}\{\mathrm{AIC}_\alpha\}$, $\alpha \in \mathcal{M}$ are assumed to be equal. This disadvantage of $\mathcal{T}$ may be overcome by making use of the information on model structures.

Last, Fig. 5 shows plots of the normal approximated P_α given in Section 5 against the P_α computed from the true distribution given in (4.1). We see the normal approximation is working fine in Fig. 5a, where $V_{\alpha\beta}$ is calculated with $v_{\alpha\beta}$. However, relatively large deviation is observed for $P_\alpha < 0.1$. This slow convergence is caused by some pairs of $V_{\alpha\beta} \approx 0$ in the HALD data set; (6.1) would be dubious if $V_{\alpha\beta} = 0$ for some pairs, as discussed in Section 6. The convergence is much slower in Fig. 5b, where $V_{\alpha\beta}$ is calculated without $v_{\alpha\beta}$. Comparing the two panels of Fig. 5, we recognize that $v_{\alpha\beta}$ in $V_{\alpha\beta}$ is working as a safeguard.

8. Concluding remarks

The shared purpose of MAICE and $\mathcal{T}$ is to find the minimum $\mathcal{E}\{\mathrm{AIC}\}$ model. It should be noted, however, that the proposed method can be easily applied to any criterion of the form IC = −(maximum log-likelihood)+(penalty term), where α^* is replaced by the minimum $\mathcal{E}\{\mathrm{IC}\}$ model. The penalty term should be $o_p(\sqrt{n})$ to make the asymptotic calculation valid.

As mentioned in Section 1, MAICE is always included in Gupta's confidence set with level $< 1/2$. This is because $T_{\mathrm{MAICE}} \leq 0$, and thus $P_{\mathrm{MAICE}}(T_{\mathrm{MAICE}}) \geq 1/2$. Note that $P_\alpha(s)$ is monotone decreasing, and that $P_\alpha(0) \geq 1/2$ since S_α is the maximum of (approximately) standard normals. Similarly, MAICE is always included in Tukey's confidence set with any level.

Models can be ranked by AIC as well as P_α, and these criteria may seem not so much different. $\mathcal{T}$ may be constructed by looking at AIC's; the difference of AIC for nested two models has a probabilistic interpretation (Bozdogan (1987)). However, this interpretation does not apply to nonnested models, and the overall error rate for many models is hard to decipher from AIC without calculating the covariance structure.

Dealing with the problem of finding the tree topology in phylogeny, Felsenstein (1985) and Felsenstein and Kishino (1993) gave a bootstrap estimate of the probability for each model (tree) to be selected as MAICE. This is another quantitative measure of the reliability. Although the confidence set they gave lacks a clear probabilistic interpretation, their "bootstrap probability" can be seen as an approximation of our P_α as mentioned in Efron *et al.* (1996).

In the regression case, "bootstrapping the residuals" method, instead of "bootstrapping pairs," can be used for resampling $\tilde{S}_\alpha$ as suggested by one of the referees; that is, resampling the residuals under the full model while the predictors are fixed (Efron and Tibshirani (1993), p. 113). This method significantly reduces the computation of Section 4, and the normal approximation discussed in Section 5 may not be needed. Although, this should be discussed further in other place, we employed the normal approximation to make the discussion general rather than special to the regression case.

The size of $\mathcal{T}$ will be large if n is small. In this case, it is important to summarize the characteristics shared by good models. Akaike (1979) gave an

answer to it within the Bayesian framework. On the other hand, a visualization of structural patterns in good models is proposed by the present author (Shimodaira (1993, 1997*b*)).

Acknowledgements

I appreciate Prof. Kaoru Nakano for the environmental support. I thank Prof. Makio Ishiguro, Dr. Avner Bar-Hen, and Prof. Ritei Shibata for their helpful suggestions and comments on earlier versions of this paper. Special thanks to Prof. Satoshi Kuriki who suggested the use of multiple comparison techniques. I also thank the referees for the suggestions to improve the manuscript. This work was supported in part by Grant-in-Aid for JSPS Fellows from the Ministry of Education, Science, Sports and Culture.

Appendix: Statistics for regression

Let $z_1, \ldots, z_m$ be the predictor variables and y be the response variable. Assume we have n i.i.d. samples of $x = (y, z_1, \ldots, z_m)$, which is unknown normal. In the log-likelihood function, $\log p(y \mid z_1, \ldots, z_m)$ is used rather than $\log p(x)$. The full model is $y = \eta_0 + \eta_1 z_1 + \cdots + \eta_m z_m + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$. The parameter is $\theta = (\sigma^2, \eta_0, \ldots, \eta_m)'$. M_α uses a subset of the predictors. Let $\hat{\sigma}_\alpha$ be the MLE of σ for M_α. Then, $\mathrm{AIC}_\alpha = 2n \log \hat{\sigma}_\alpha + 2m_\alpha$ without a constant common to all the models. The first term of $V_{\alpha\beta}$ will be obtained from $\mathrm{var}_{\mathcal{X}}\{\log p_\alpha(x \mid \hat{\theta}_\alpha) - \log p_\beta(x \mid \hat{\theta}_\beta)\} = n^{-1}\sum_{t=1}^n (\log p_\alpha(x_t \mid \hat{\theta}_\alpha) - \log p_\beta(x_t \mid \hat{\theta}_\beta))^2 - (n^{-1}\sum_{t=1}^n \{\log p_\alpha(x_t \mid \hat{\theta}_\alpha) - \log p_\beta(x_t \mid \hat{\theta}_\beta)\})^2 \approx 1 - (\sum_{t=1}^n \hat{\varepsilon}_{\alpha t}\hat{\varepsilon}_{\beta t}/(n\hat{\sigma}_\alpha\hat{\sigma}_\beta))^2$, where $\hat{\varepsilon}_{\alpha t}$ is the residual. This is easily calculated from $\hat{\theta}_\alpha$, $\hat{\theta}_\beta$, and the observed covariance matrix of x. Note that the above quantity will be $1 - (\hat{\sigma}_\alpha/\hat{\sigma}_\beta)^2$ if $M_\beta \subset M_\alpha$. The second term of $V_{\alpha\beta}$ will be obtained from

$$G^*_{\alpha\alpha\cdot\sigma^2\sigma^2} = 1/2\sigma^{*4}_\alpha, \quad G^*_{\alpha\alpha\cdot\sigma^2\eta_i} = 0, \quad G^*_{\alpha\alpha\cdot\eta_i\eta_j} = \mathcal{E}\{z_i z_j\}/\sigma^{*2}_\alpha,$$
$$G^*_{\alpha\beta\cdot\sigma^2\sigma^2} = (\mathcal{E}\{\varepsilon^*_\alpha\varepsilon^*_\beta\})^2/2\sigma^{*4}_\alpha\sigma^{*4}_\beta, \quad G^*_{\alpha\beta\cdot\sigma^2\eta_j} = \mathcal{E}\{\varepsilon^*_\alpha\varepsilon^*_\beta\}\mathcal{E}\{\varepsilon^*_\alpha z_j\}/\sigma^{*4}_\alpha\sigma^{*2}_\beta,$$

and

$$G^*_{\alpha\beta\cdot\eta_i\eta_j} = (\mathcal{E}\{\varepsilon^*_\alpha\varepsilon^*_\beta\}\mathcal{E}\{z_i z_j\} + \mathcal{E}\{\varepsilon^*_\alpha z_j\}\mathcal{E}\{\varepsilon^*_\beta z_i\})/\sigma^{*2}_\alpha\sigma^{*2}_\beta,$$

where σ^*_α is σ_α at θ^*_α, and ε^*_α is the associated residual.

References

Aitkin, M. A. (1974). Simultaneous inference and the choice of variable subsets in multiple regression, *Technometrics*, **16**, 221–227.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Control*, **19**, 716–723.

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting, *Biometrika*, **66**, 237–242.

Arvesen, J. N. and McCabe, G. P., Jr. (1975). Subset selection problems for variances with applications to regression analysis, *J. Amer. Statist. Assoc.*, **70**, 166–170.

Atkinson, A. C. (1970). A method for discriminating between models, *J. Roy. Statist. Soc. Ser. B*, **32**, 323–353.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics*, Wiley, New York.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions, *Psychometrika*, **52**, 345–370.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses, *J. Roy. Statist. Soc. Ser. B*, **24**, 406–424.

Dastoor, N. K. and McAleer, M. (1989). Some power comparisons of joint and paired tests for nonnested models under local hypotheses, *Econometric Theory*, **5**, 83–94.

Draper, N. and Smith, H. (1981). *Applied Regression Analysis* (2nd ed.), Wiley, New York.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Efron, B., Halloran, E. and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees, *Proc. Nat. Acad. Sci. U.S.A.*, **93**, 13429–13434.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap, *Evolution*, **39**, 783–791.

Felsenstein, J. and Kishino, H. (1993). Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull, *Systematic Biology*, **42**, 193–200.

Gupta, S. S. and Huang, D. Y. (1976). Selection procedures for the means and variances of normal populations: unequal sample sizes case, *Sankhyā Ser. B*, **38**, 112–128.

Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedures*, Wiley, New York.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*, Wiley, New York.

Linhart, H. (1988). A test whether two AIC's differ significantly, *South African Statistical Journal*, **22**, 153–161.

Mallows, C. L. (1973). Some comments on C_p, *Technometrics*, **15**, 661–675.

Shimodaira, H. (1993). A model search technique based on confidence set and map of models, *Proc. Inst. Statist. Math.*, **41**, 131–147 (in Japanese).

Shimodaira, H. (1997*a*). Assessing the error probability of the model selection test, *Ann. Inst. Statist. Math.*, **49**, 395–410.

Shimodaira, H. (1997*b*). A graphical technique for finding a set of good models using AIC and its variance, *Ann. Inst. Statist. Math.* (submitted).

Spjøtvoll, E. (1972). Multiple comparison of regression functions, *The Annals of Mathematical Statistics*, **43**, 1076–1088.

Spjøtvoll, E. (1977). Alternatives to plotting C_p in multiple regression, *Biometrika*, **64**, 1–8.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica*, **57**, 307–333.

White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica*, **50**, 1–25.