Ann. Inst. Statist. Math.
30 (1978), Part A, 9-14

A BAYESIAN ANALYSIS OF THE MINIMUM AIC PROCEDURE

HIROTUGU AKAIKE

(Received Oct. 15, 1977; revised Apr. 24, 1978)

Summary

By using a simple example a minimax type optimality of the minimum AIC procedure for the selection of models is demonstrated.

1. Introduction

Akaike [1] introduced an information criterion which is by definition

$$\text{AIC}=(-2)\log(\text{maximum likelihood})+2(\text{number of parameters}) \tag{1.1}$$

as an estimate of minus twice the expected log likelihood of the model whose parameters are determined by the method of maximum likelihood. Here log denotes the natural logarithm. The simple procedure which selects a model with the minimum AIC among a set of models defines the minimum AIC estimate (MAICE) (Akaike [2]). The introduction of AIC helped the recognition of the importance of modeling in statistics and many practically useful statistical procedures have been developed as minimum AIC procedures; see, for example, Akaike [2], [3].

In spite of the accumulation of successful results in practical applications the logical foundation of MAICE has been continuously questioned by theoretically minded statisticians. The purpose of the present paper is to provide a Bayesian interpretation of the MAICE procedure and show that the procedure provides a minimax type solution to the problem of model selection under the assumption of equal prior probability of the models.

Our analysis starts with a brief review of the statistic of the form

$$(-2)\log(\text{maximum likelihood})+(\log N)(\text{number of parameters})+C\,, \tag{1.2}$$

where N is the sample size used for the computation of the maximum

likelihood estimates. We will call this type of statistic by the generic name BIC. Two types of BIC have been introduced by Akaike [3] and Schwarz [4].

2. A review of BIC

Both Akaike and Schwarz based the introduction of BIC on some Bayesian arguments. Schwarz derived the statistic for models from a Koopman-Darmois family. Akaike introduced the statistic for the problem of selection of variables in linear regression. Here we will restrict our attention to the problem of selection of a multivariate Gaussian distribution. This is a special case of the model treated by Schwarz but may serve as a simplified model of the general situation where the use of the maximum likelihood estimates is contemplated.

Consider the situation where a set of observations $Y=\{y(n);\ n=1, 2, \cdots, N\}$ of L-dimensional vector random variables $y(n)=(y_1(n), y_2(n), \cdots, y_L(n))$ is given. It is assumed that $y(n)$'s are independently identically distributed as Gaussian $N(\theta, I)$, where $\theta=(\theta_1, \theta_2, \cdots, \theta_L)$ is the vector of the unknown means and I is an $L\times L$ identity matrix. We consider the set of models $N({}_k\theta, I)$ $(k=0, 1, \cdots, L)$ specified by assuming $\theta_{k+1}=\theta_{k+2}=\cdots=\theta_L=0$, i.e., ${}_k\theta=({}_k\theta_1, {}_k\theta_2, \cdots, {}_k\theta_k, 0, \cdots, 0)$ and ${}_k\theta_1, {}_k\theta_2, \cdots, {}_k\theta_k$ are unknown. Following Schwarz, we assume a prior distribution $\pi(k)$ over k, or the set of models $N({}_k\theta, I)$, $(k=0, 1, \cdots, L)$. Further we assume a prior distribution $N(0_k, \sigma^2 I_k)$ for $({}_k\theta_1, {}_k\theta_2, \cdots, {}_k\theta_k)$, where 0_k denotes a k-dimensional zero-vector and I_k a $k\times k$ identity matrix. Now the marginal posterior distribution of k is given by $C\pi(k)p(k|Y)$ $(k=0, 1, \cdots, L)$ with $p(k|Y)$ defined by

$$p(k|Y)=\exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\sum_{i=k+1}^{L} y_i^2(n)\right\}\int \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{k}(y_i(n)-{}_k\theta_i)^2\right\}$$
$$\cdot\left(\frac{1}{2\pi\sigma^2}\right)^{k/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}{}_k\theta_i^2\right\}\prod_{i=1}^{k} d_k\theta_i\ .$$

By choosing the k that maximizes the posterior probability one can maximize the probability of correct decision on k. Consider the situation where σ^2 is sufficiently large so that we get an approximate equality

$$p(k|Y)=\exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\sum_{i=k+1}^{L} y_i^2(n)\right\}\int \exp\left\{-\frac{1}{2}\sum_{i=1}^{k}\sum_{n=1}^{N}(y_i(n)-{}_k\theta_i)^2\right\}$$
$$\cdot\left(\frac{1}{2\pi\sigma^2}\right)^{k/2}\prod_{i=1}^{k} d_k\theta_i\ .$$

For this case we have

$$p(k|Y)=\exp\left\{-\frac{1}{2}S(k)\right\}\left(\frac{1}{N\sigma^2}\right)^{k/2},$$

where

$$S(k)=\sum_{i=1}^{L}\sum_{n=1}^{N}(y_i(n)-\bar{y}_i)^2+\sum_{i=k+1}^{L}N\bar{y}_i^2$$

and $\bar{y}_i$ denotes the sample mean $\sum y_i(n)/N$. The BIC statistic (1.2) of the kth model is obtained as minus twice the log posterior probability $(-2)\log\{C\pi(k)p(k|Y)\}$ and is given by

$$\text{BIC}(k)=S(k)+k\log N+R(k), \tag{2.1}$$

where $R(k)=k\log\sigma^2-2\log\{C\pi(k)\}$. The corresponding AIC statistic (1.1) of the model is given by

$$\text{AIC}(k)=S(k)+2k+R, \tag{2.2}$$

where R is independent of k and may be ignored in the following discussion.

Taking into account the relation

$$S(k)=S(L)+\sum_{i=k+1}^{L}N\bar{y}_i^2, \tag{2.3}$$

it is easy to see that the MBICE, the k that minimizes BIC(k), provides a consistent estimate of the correct model. In contrast to this the MAICE, the k that minimizes AIC(k), does not have this consistency property. This is obvious, because when the k_0th model is correct the distribution of the differences of AIC(k)'s with $k=k_0, k_0+1, \cdots, L$ tends to a non-degenerate stationary distribution as N tends to infinity. Schwarz argues in Section 4 of his paper that this is a shortcoming of MAICE. Does this mean that MAICE is generally inferior to MBICE? Schwarz carefully qualifies his statement by saying that if the assumptions made in Section 2 of his paper, which are essentially equivalent to the assumptions of our present Bayesian model, are accepted MAICE cannot be optimal. In the next section it will be shown that MAICE is optimal under some assumptions which are quite different from and often more natural than those of MBICE.

3. Minimax property of MAICE

In this section we assume that the prior distribution $p({}_k\theta|k)$ of ${}_k\theta=({}_k\theta_1, {}_k\theta_2, \cdots, {}_k\theta_L)$ of the kth model is given by

$$p({}_k\theta|k)=\left(\frac{1}{2\pi\sigma^2}\right)^{k/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}{}_k\theta_i^2\right\}\left(\frac{1}{2\pi\delta^2}\right)^{(L-k)/2}$$

$$\cdot \exp\left\{-\frac{1}{2\delta^2}\sum_{i=k+1}^{L} {}_k\theta_i^2\right\} .$$

It is assumed that $N\sigma^2$ is greater than 1 and $N\delta^2$ is smaller than 1. The logarithm of the posterior distribution $p\{({}_k\theta, k)|Y\}$ of $({}_k\theta, k)$ is then given by

$$\log p\{({}_k\theta, k)|Y\} = -\frac{1}{2}\left\{\sum_{i=1}^{L} N(\bar{y}_i - {}_k\theta_i)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{k} {}_k\theta_i^2 + \frac{1}{\delta^2}\sum_{i=k+1}^{L} {}_k\theta_i^2 + k\log\left(\frac{\sigma^2}{\delta^2}\right)\right\} + \log\pi(k) + R ,$$

where $\pi(k)$ is the prior probability of the kth model and R denotes a quantity which is independent of k. Here we assume equal probability for $\pi(k)$ and the term $\log\pi(k)$ will be ignored. The mode of the posterior distribution is then given by $({}_k\hat{\theta}, k)$ with k that maximizes $p\{({}_k\hat{\theta}, k)|Y\}$, where ${}_k\hat{\theta}_i = \{N/(N+1/\sigma^2)\}\bar{y}_i$ for $i=1, 2, \cdots, k$, $\{N/(N+1/\delta^2)\}\bar{y}_i$ for $i=k+1, k+2, \cdots, L$. Now, minus twice $\log p\{p({}_k\hat{\theta}, k)|Y\}$ is given by

$$\frac{1}{N\sigma^2+1}\sum_{i=1}^{k} N\bar{y}_i^2 + \frac{1}{N\delta^2+1}\sum_{i=k+1}^{L} N\bar{y}_i^2 + k\log\left(\frac{\sigma^2}{\delta^2}\right) , \tag{3.1}$$

where a common additive constant is ignored. This formula tells that if we allow δ^2 diminish to zero the mode of the posterior distribution can only be attained at $k=0$, which is a nonsensical result and suggests the necessity of marginalization, or the integration of the posterior distribution with respect to $d_k\theta$.

By integrating $p\{({}_k\theta, k)|Y\}$ with respect to $d_k\theta$ the posterior probability $p(k|Y)$ of the kth model is obtained and minus twice its logarithm is given by

$$\mathrm{LOG}(k) = \frac{1}{N\sigma^2+1}\sum_{i=1}^{k} N\bar{y}_i^2 + \frac{1}{N\delta^2+1}\sum_{i=k+1}^{L} N\bar{y}_i^2 + k\log\left(\frac{N\sigma^2+1}{N\delta^2+1}\right) , \tag{3.2}$$

where a common additive constant is ignored. For the purpose of comparison of models we use $\mathrm{CIC}(k) = \mathrm{LOG}(k) - \mathrm{LOG}(L)$ which is given by

$$\mathrm{CIC}(k) = \left(\frac{1}{N\delta^2+1} - \frac{1}{N\sigma^2+1}\right)\sum_{i=k+1}^{L} N\bar{y}_i^2 + k\log\left(\frac{N\sigma^2+1}{N\delta^2+1}\right) ,$$

where again a common additive constant is ignored. If N is large compared with σ^2 and $N\sigma^2$ is large but $N\delta^2$ is small compared with 1, $\mathrm{CIC}(k)$ will be approximated by $\mathrm{BIC}(k)$ of (2.1). These conditions can be satisfied by increasing the sample size N when δ^2 is sufficiently close or exactly equal to zero and represents the situation where the difference between the magnitudes of the elements in the set $({}_k\theta_1, {}_k\theta_2, \cdots, {}_k\theta_k)$ and

those in $({}_k\theta_{k+1}, {}_k\theta_{k+2}, \cdots, {}_k\theta_L)$ is clearly visible through the observations $\bar{y}_i$ $(i=1, 2, \cdots, L)$. This is the situation which we cannot expect to hold very often in ordinary exploratory data analyses. Thus it must be concluded that MBICE will find rather limited applications.

The decision on the choice of k will be difficult when the difference between $({}_k\theta_1, {}_k\theta_2, \cdots, {}_k\theta_k)$ and $({}_k\theta_{k+1}, {}_k\theta_{k+2}, \cdots, {}_k\theta_L)$ cannot be clearly recognized through the observations $\bar{y}_i$ $(i=1, 2, \cdots, L)$. The most critical will then be the situation where $N\sigma^2$ (>1) and $N\delta^2$ (<1) are both very close to 1. For this critical situation we get

$$\lim_{\substack{N\sigma^2 \downarrow 1 \\ N\delta^2 \uparrow 1}} \left(\frac{1}{N\delta^2+1}-\frac{1}{N\sigma^2+1}\right)^{-1} \mathrm{CIC}(k)=\sum_{i=k+1}^{L} N\bar{y}_i^2+2k\,. \tag{3.3}$$

Taking into account (2.2) and (2.3) one can see that the right-hand side of the above equation can be used as the definition of the statistic AIC (k) to be used in the definition of the minimum AIC procedure for the decision on k. Thus we get a proof of optimality of MAICE under this limiting condition. For the case with $\sigma^2=e^2N^{-1}$ and $\delta^2=e^{-2}N^{-1}$ we get a statistic with $2.63k$ in place of $2k$ of AIC (k). This result shows that for a fairly wide range of values of σ^2 and δ^2 the minimum AIC procedure will provide a reasonable approximation to the Bayes solution of the decision problem of k under the assumption of equal probability $\pi(k)$. Thus we have obtained a surprisingly simple proof of the minimax type optimality of MAICE and its robustness.

4. Discussion

In the discussion of Section 3, N was retained only to clarify the relation between AIC and BIC. For the discussion of MAICE N could have been put equal to 1. If we consider $\bar{y}_i$'s as the maximum likelihood estimates of the parameters of a distribution after a proper change of coordinate we can see that the result of the preceding section holds generally for MAICE and characterizes the procedure as optimal for the detection of k where the ratio of the signal, the bias squared, to the noise, the variance, crosses the critical value 1. This justifies the original intension of introduction of MAICE by Akaike [1], [2].

The formulas (3.1) and (3.2) demonstrate the close relation between the maximum and marginal, or integrated, likelihoods of each model. It is only when $N\sigma^2$ and $N\delta^2$ are both significantly greater than 1 that these two formulas become approximately equal, which is not a very interesting situation.

The optimality of MAICE discussed in the preceding section is only concerned with the probability of coincidence of the MAICE with the correct k. If an estimate of the correct value of k is required the

mean of the approximate posterior distribution $p(k|Y)=C\exp\{(-1/2)\cdot\mathrm{AIC}(k)\}$ would be useful.

Obviously the minimum AIC procedure discussed in the preceding section is a direct extension of the classical method of maximum likelihood to the multi-model situation and is not free from the defect of ordinary point estimate. Although the AIC (k)'s are defined in terms of the maximum likelihood estimates $\bar{y}_i$ $(i=1, 2, \cdots, L)$ the discussion in the preceding section does not tell which estimate of ${}_k\theta$ should eventually be used. The use of the maximum likelihood estimates of ${}_k\theta$ under the assumption of $\sigma^2=\infty$ and $\delta^2=0$ has been customary but a further analysis in necessary when there is no such clear separation of θ_i's. If the choice of one single model is not the sole purpose of the analysis of the data the average of the models with respect to the approximate posterior probability $C\exp\{(-1/2)\,\mathrm{AIC}(k)\}$ will provide a better estimate of the true distribution of Y. In this type of application the $2k$ in the definition of AIC (k) may be adaptively modified by using the model of the prior distribution defined by $\pi(k)=C\lambda^k$ and adjusting the parameter λ by the data. The feasibility of this type of procedure has been confirmed by numerical experiments and the results will be discussed in a separate paper.

Acknowledgement

The present author is grateful to Professor Gideon Schwarz of Hebrew University for providing a chance to study his paper before its publication. The paper was one of the main stimulae which lead the present author to the present investigation. The work reported in this paper was partly supported by a grant from the Ministry of Education, Science and Culture.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium of Information Theory*, B. N. Petrov and F. Csaki, eds., Akademiai Kiado, Budapest, 267-281.

[2] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Contr.*, AC-**19**, 716-723.

[3] Akaike, H. (1977). On entropy maximization principle, *Applications of Statistics*, P. R. Krishnaiah, ed., North-Holland, Amsterdam, 27-41.

[4] Schwarz, G. (1976). Estimating the dimension of a model, *Ann. Statist.*, **6**, 461-464.