

No. 276

Re-viewing on AKAIKE's Information Criterion.

by

Yoshiko Nogami

September 1985

# Re-viewing on AKAIKE's Information Criterion.

by

Yoshiko Nogami

University of Tsukuba

## 1. Introduction.

Let  $\underline{x} = (x_1, \dots, x_n)'$  be a random sample of size  $n$  with each  $x$  having the probability density function  $f(x; \theta)$ . Let  $\ell_{\underline{x}}(\theta) = \log \prod_{i=1}^n f(x_i; \theta)$ . Hereafter, we do not exhibit the subscript  $\underline{x}$  in  $\ell_{\underline{x}}(\theta)$  unless we need it. Akaike(1973, 1974) proposes the following information criterion :

$$(1) \quad \text{AIC} = -2 \ell(\hat{\theta}_n) + 2c$$

where  $\hat{\theta}_n = \hat{\theta}_n(x_1, \dots, x_n)$  is the maximum likelihood estimator for  $\theta$  and  $c$  represents the number of independently adjusted parameters within the model.

In this paper we review AKAIKE's information criterion (AIC) from three different points of view; the information theoretic observation, statistical estimation standpoint of view and statistical decision theoretic consideration for testing hypotheses. Eventually, we shall see from information theoretic observation that it is plausible that the model minimizing AIC at the same time minimizes Kullback-Leibler's mean information. We also see from statistical estimation standpoint of view that  $-(2n)^{-1}\text{AIC}$  is asymptotically unbiased estimator for  $E_Y(\log f(Y; \theta^*))$  ( $\theta^*$ : true parameter). We furthermore introduce some property between AIC and classical likelihood ratio test when they are applied to testing hypotheses.

## 2. Properties of AIC.

According to Akaike's discussion(1973), the information theoretic observation of AIC is as follows:

When past observations  $x_1, \dots, x_n$  are given, we choose  $\theta = \hat{\theta}_n$  which maximizes

$$n^{-1} \sum_{i=1}^n \log (f(x_i; \hat{\theta}_n) / f(x_i; \theta^*))$$

where  $\theta^*$  is a true parameter. Let  $\underline{y}_N = (y_1, \dots, y_N)'$  be the future observations having the likelihood  $\prod_{i=1}^N f(y_i; \theta^*)$ . By using  $\hat{\theta}_n$  and  $\underline{y}_N$ , it will be plausible for  $\hat{\theta}_n$  to maximize

$$N^{-1} \sum_{i=1}^N \log (f(y_i; \hat{\theta}_n) / f(y_i; \theta^*))$$

with (possibly) high probability. (N could be of n?) On the other hand, since as  $N \rightarrow \infty$

$$-2N^{-1} \sum_{i=1}^N \log (f(y_i; \hat{\theta}_n) / f(y_i; \theta^*)) \rightarrow -2W(\theta^*, \hat{\theta}_n)$$

where

$$W(\theta^*, \hat{\theta}_n) = \int \{ \log (f(y; \hat{\theta}_n) / f(y; \theta^*)) \} f(y; \theta^*) dy,$$

it will be plausible that for large N  $\hat{\theta}_n$  minimizes  $-2W(\theta^*, \hat{\theta}_n)$  and hence minimizes  $-2E_{\underline{X}}(W(\theta^*, \hat{\theta}_n)) = 2\{-E_{\underline{X}}(\int \{ \log (f(y; \hat{\theta}_n) / f(y; \theta^*)) \} f(y; \theta^*) dy)\} = 2(\text{Kullback-Leibler's mean information})$ . Hence, choosing the model  $f(x; \hat{\theta}_n)$  which minimizes AIC will be the same as choosing the model  $f(x; \hat{\theta}_n)$  which minimizes Kullback-Leibler's mean information.

From statistical estimation standpoint of view we can interpret AIC as follows:

Let  $\theta = (\theta_1, \dots, \theta_c)'$ . Since, as you know,  $2(\ell(\hat{\theta}_n) - \ell(\theta^*))$  is asymptotically distributed according to Chi-Square distribution with  $c$  degrees of freedom ( $\approx \chi^2(c)$ ) and hence  $E_{\underline{X}}(2(\ell(\hat{\theta}_n) - \ell(\theta^*))) \approx c$ , and since we also know that for large  $n$   $\ell(\theta^*) \approx nE_Y(\log f(Y; \theta^*))$ , it follows that for large  $n$ ,

$$(2) \quad E_{\underline{X}}(2 \ell(\hat{\theta}_n) - 2nE_Y(\log f(Y; \theta^*))) \approx c.$$

On the other hand, by Taylor expansion of  $2 \log f(y; \theta)$  for  $\theta$  about a nbd of  $\theta = \theta^*$ ,

$$2 \log f(y; \theta) \approx 2 \log f(y; \theta^*) + 2(\theta - \theta^*)' \left\| \frac{\partial}{\partial \theta} \log f(y; \theta^*) \right\| \\ + (\theta - \theta^*)' \left\| -\frac{\partial^2}{\partial \theta_p \partial \theta_q} \log f(y; \theta^*) \right\| (\theta - \theta^*)$$

where  $\|\cdot\|$  is a matrix notation.

Taking expectation wrt  $y$  leads to

$$2E_Y(\log f(Y; \theta)) \approx 2 E_Y(\log f(Y; \theta^*)) - (\theta - \theta^*)' B^2(\theta^*; \theta) (\theta - \theta^*)$$

where

$$-B^2(\theta^*; \theta) = \left\| E_Y \left( \frac{\partial^2}{\partial \theta_p \partial \theta_q} \log f(y; \theta^*) \right) \right\|.$$

Hence, putting  $\theta = \hat{\theta}_n$  gives

$$2nE_Y(\log(f(y; \theta^*)/f(y; \hat{\theta}_n))) \approx \sqrt{n}(\hat{\theta}_n - \theta^*)' B^2(\theta^*; \hat{\theta}_n) (\hat{\theta}_n - \theta^*) \sqrt{n}.$$

Since the rhs is distributed according to  $\chi^2(c)$  for large  $n$ , taking expectation wrt  $\underline{x}$  on both sides leads to

$$(3) \quad 2nE_{\underline{X}}E_Y(\log(f(y; \theta^*)/f(y; \hat{\theta}_n))) \approx c$$

for large  $n$ .

Thus, adding the both hands of (2) and (3) each other and putting  $nE_{\underline{X}}E_Y(\log f(y; \hat{\theta}_n)) = I_n$  gives

$$(4) \quad E_{\underline{X}}(2 \ell(\hat{\theta}_n) - 2 I_n) \approx 2c$$

for large  $n$ .

Therefore,  $-AIC$  is an asymptotically unbiased estimator for  $2I_n$ .

Since from (3)  $(2n)^{-1}(\text{lhs}(3)) \approx 0$  as  $n \rightarrow \infty$  and  $-W(\theta^*, \hat{\theta}_n) \geq 0$  by Akaike(1973, p. 297), it follows that  $(2n)^{-1} \text{lhs}(3) = E_{\underline{X}}(-W(\theta^*, \hat{\theta}_n)) = E_Y(\log f(Y; \theta^*)) - n^{-1} I_n (\geq 0) \approx 0$  as  $n \rightarrow \infty$  and hence  $E_{\underline{X}}(-(2n)^{-1} AIC) (\approx n^{-1} I_n) \uparrow E_Y(\log f(Y; \theta^*))$  as  $n \uparrow \infty$ . Whence,  $-(2n)^{-1} AIC$  is asymptotically unbiased estimator for  $E_Y(\log f(Y; \theta^*))$ .

Now, Sugiura(1978) proposes a corrected (c-)AIC computing the exact

bias (see (4)) directly. We notice that if  $y_n$  is iid (future) observations with pdf  $f(y; \theta^*)$ , then we obtain the equality

$$I_n = E_Y \left( \int \{ \log \prod_{i=1}^n f(y_i; \hat{\theta}_n) \} \prod_{i=1}^n f(y_i; \theta^*) dy_n \right).$$

Thus, we can propose the corrected AIC as follows:

$$(5) \quad c\text{-AIC} = -2 \ell(\hat{\theta}_n) + 2d$$

where

$$d = E_Y [\ell(\hat{\theta}_n)] - I_n.$$

Therefore, we will see from the discussion just below (4) that  $-(2n)^{-1}(c\text{-AIC})$  is also an asymptotically unbiased estimator for  $E_Y(\log f(y; \theta^*))$ .

From statistical decision theoretic consideration where we treat the null hypothesis and the alternative hypothesis equally as playing the same role, let us look at applying AIC to testing hypotheses. Let  $H_k$  ( $k=1, 2, \dots, c$ ) be the model designated by the hypothesis:  $\theta = \theta^{(k)} = (0, \dots, 0, \theta_{k+1}, \dots, \theta_c)'$  ( $k \leq c$ ). Let  $H$  be the (saturated) model with  $\theta$  having no restrictions. Suppose the problem is to choose a model out of  $H, H_1, \dots, H_c$ . Let  $\hat{\theta}$  and  $\hat{\theta}^{(k)}$  be the MLE's for  $\theta$  (in  $H$ ) and  $\theta^{(k)}$ , respectively. Then, AIC chooses the model which gives the minimum of  $AIC(H), AIC(H_1), \dots, AIC(H_c)$  according to the formula (1).

Suppose we test the (null) hypothesis  $H_k$  versus the (alternative) hypothesis  $H$ . Let  $c_k$  and  $c$  represent the numbers of parameters whose values can be changed freely under  $H_k$  and  $H$ , respectively. Since the (log-) likelihood ratio (LR) statistic

$$z = -2(\ell(\hat{\theta}^{(k)}) - \ell(\hat{\theta}))$$

has Chi-Square distribution ( $\chi^2(c-c_k)$ ) with  $c-c_k$  degrees of freedom, it follows at a  $100\alpha$  % level of significance and with  $\chi_{1-\alpha}^2(m)$  the critical

value for the upper  $100\alpha$  % portion of  $\chi^2(m)$  distribution that if

$$z > \chi_{1-\alpha}^2(c-c_k),$$

then we choose the model H, rejecting  $H_k$ .

In the same problem, since

$$(6) \quad \text{AIC}(H_k) - \text{AIC}(H) = z - 2(c-c_k),$$

noticing that when  $z \sim \chi^2(m)$ ,  $E(z)=m$  gives the fact that AIC chooses the model H if (6)  $> 0$  or equivalently

$$z > 2E(z).$$

Therefore, if  $2E(z) < \chi_{1-\alpha}^2(c-c_k)$ , then AIC is likely to choose the model designated by the alternative hypothesis H, more often than  $\chi^2$  test. Vice versa.

#### REFERENCES.

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory (B.N. Petrov and F.Csaki. eds.), Budapest: Akademia Kiado, 267-281.
- (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, Vol. AC-19, 716-723.
- Sugiura, N. (1978). Further analysis of the data by AKAIKE's information criterion and the finite corrections. Commun. Statist. Theor. Meth., A7(1), 13-26.

Notations: rhs = right hand side, nbd = neighborhood,  
 lhs = left hand side, lhs(#)= left hand side of (#),  
 wrt = with respect to.