ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
20, 4 (1990), pp. 543–549

R. ZIELIŃSKI (Warszawa)

OPTIMAL CHOICE OF A STATISTICAL MODEL: A DISCUSSION OF THE AKAIKE AND CLASSICAL APPROACHES

Abstract. Sakamoto et al. [3] have given a full account and have presented the state of art of the Akaike approach to optimal choice of a statistical model. Two features of the Akaike theory play a fundamental role: using the Kullback–Leibler information quantity (K–L quantity) to measure the loss which occurs when an unknown "true" distribution is replaced by a model distribution, and measuring the goodness-of-fit of a given statistical model by the distance (generated by K–L quantity) of the underlying family of distributions from the true distribution. In the paper we show that both ideas, when considered from the classical statistical decision theory point of view, lead to persuasive criteria of estimation and model choice.

1. Akaike loss function. Let X be a random vector with pdf $f(x|\theta^*)$ which depends on an unknown parameter $\theta^* \in \Theta$. If instead of the "true" value θ^* we adopt $\theta \in \Theta$, the loss is defined as (we use the notation from [3])

$$l^*(\theta) = -n\int [\log f(t|\theta)] f(t|\theta^*)dt. \tag{1}$$

A proper choice of θ (a choice of the model $f(\cdot|\theta)$) is that which, roughly speaking, minimizes $l^*(\theta)$.

A rationale of the above is as follows (see [3], p. 38). Consider the Kullback–Leibler information quantity

$$I(g; f) = \mathrm{E}_g \log \frac{g(X)}{f(X)} = \int \left[\log \frac{g(x)}{f(x)}\right] g(x)dx$$

with the well-known properties:

(i) $I(g; f) \geqslant 0$;

(ii) $I(g; f) = 0$ iff $g(x) = f(x)$ (a.e.).

The quantity $I(g; f)$ may be considered as a "distance" of f from g and someone who does not know the true g is interested in adopting f as close to g as possible. Given (unknown) g, this amounts to minimizing the loss $l^*(\theta)$. Rényi ([2], p. 453) calls $I(g; f)$ "a gain when the distribution f is replaced by g"; in our context a more adequate term is "a loss when an unknown distribution g is replaced by the model distribution f". Observe that, given x, $f(x|\theta)$ is the

likelihood of θ. Hence in [3] the quantity $-n^{-1}l^*(\theta)$ is called "expected log-likelihood under the true distribution $f(\cdot|\theta^*)$".

Due to the fact that θ^* is unknown and our guess $\hat{\theta}$ of θ^* is based on the observation X, i.e., $\hat{\theta} = \hat{\theta}(X)$, the classical approach is to consider the risk of $\hat{\theta}$ defined as

$$r^*(\hat{\theta}) = \mathrm{E}_{\theta^*} l^*(\hat{\theta}(x)) = \int l^*(\hat{\theta}(x)) f(x|\theta^*)dx,$$

and to find $\hat{\theta}$ which minimizes $r^*(\hat{\theta})$ uniformly in θ^* (if possible), to discuss the admissible $\hat{\theta}$, minimax $\hat{\theta}$, or Bayes $\hat{\theta}$, etc., depending on the aim of the study and on some further details of the problem under consideration. The Akaike theory has not been developed in this direction.

2. Optimal choice of the statistical model. Classical and Akaike approaches. Given an "overall" parameter space Θ, let $\Theta_k \subset \Theta$, $k = 1, 2, \ldots, K$, represent competing models. Under the loss (1), a "natural" procedure of the choice of the best model would be as follows: define

$$l_k^* = \inf_{\theta \in \Theta_k} l^*(\theta), \quad k = 1, 2, \ldots, K$$

(the "distance" of the model Θ_k from the true θ^*), and choose k which minimizes l_k^*.

But, again, θ^* is not known and our guess of k, say $\hat{k} = \hat{k}(X)$, is based on the observation X. The classical approach would be: consider the risk of $\hat{k}$ defined as

$$R^*(\hat{k}) = \mathrm{E}_{\theta^*} l^*_{\hat{k}(X)}$$

and find $\hat{k}$ which minimizes $R^*(\hat{k})$ uniformly in θ^*, or find admissible $\hat{k}$, or minimax $\hat{k}$, or Bayes $\hat{k}$, respectively. This is what one could consider as the optimal choice of a statistical model (a family of distributions) under the given Akaike loss function $l^*(\theta)$.

Observe that in the above formulation we are facing the classical problem of constructing an optimal statistical decision procedure. The Akaike approach is different. It may be presented as follows.

Given an observation $X = x$, for the k-th statistical model Θ_k ($k = 1, 2, \ldots, K$) find the MLE $\hat{\theta}_k$ defined as

$$\hat{\theta}_k(x) = \arg\max_{\theta \in \Theta_k} [\log f(x|\theta)].$$

Then calculate the "mean expected log-likelihood"

$$l^*(k) = \mathrm{E}_{\theta^*} l^*(\hat{\theta}_k(X)) = \int l^*(\hat{\theta}_k(x)) f(x|\theta^*)dx.$$

"The model with larger mean expected log-likelihood is considered to be the better one" ([3], p. 60). However, θ^* is unknown and the solution is not applicable. Some new tricks are needed and these are as follows. The Taylor expansion of $l^*(\theta)$ around the true value θ^* yields the approximation ([3], p. 65)

$$l^*(\theta) = l^*(\theta^*) - \tfrac{1}{2}\sqrt{n}(\theta - \theta^*) J_* \sqrt{n}(\theta - \theta^*)^T,$$

where J_* is the Fisher information matrix. Putting $\theta = \hat{\theta}_k(X)$ and calculating the expected value we obtain

$$l^*(k) = l^*(\theta^*) - \tfrac{1}{2}\mathrm{E}\left[\sqrt{n}(\theta - \theta^*)J_*\sqrt{n}(\theta - \theta^*)^T\right]. \tag{2}$$

On the other hand, for $l(\theta) = \log f(x|\theta)$ we get the following approximation (through the Taylor expansion around $\hat{\theta}_k = \hat{\theta}_k(X)$; see [3], pp. 67 and 68):

$$l(\theta) = l(\hat{\theta}_k) - \tfrac{1}{2}\sqrt{n}(\theta - \hat{\theta}_k)J_*\sqrt{n}(\theta - \hat{\theta}_k)^T.$$

Substituting θ^* for θ and taking expectations of both sides we get

$$\mathrm{E}_{\theta^*} l(\theta^*) = \mathrm{E}_{\theta^*} l\big(\hat{\theta}_k(X)\big) - \tfrac{1}{2}\mathrm{E}\left[\sqrt{n}(\theta^* - \hat{\theta}_k)J_*\sqrt{n}(\theta^* - \hat{\theta}_k)^T\right]. \tag{3}$$

But, by (1), $\mathrm{E}_{\theta^*} l(\theta^*) = l^*(\theta^*)$, so that combining (2) and (3) yields

$$l^*(k) = \mathrm{E}_{\theta^*} l\big(\hat{\theta}_k(X)\big) - \mathrm{E}\left[\sqrt{n}(\theta^* - \hat{\theta}_k)J_*\sqrt{n}(\theta^* - \hat{\theta}_k)^T\right].$$

Now, if $\dim_k \theta$ denotes the dimension of Θ_k, under usual regularity conditions, $\sqrt{n}(\theta^* - \hat{\theta}_k)J_*\sqrt{n}(\theta^* - \hat{\theta}_k)^T$ is asymptotically chi-square with $\dim_k \theta$ d.f., so that eventually

$$l^*(k) = \mathrm{E}_{\theta^*} l\big(\hat{\theta}_k(X)\big) - \dim_k \theta$$

or

$$l^*(k) = \mathrm{E}_{\theta^*}\big(l(\hat{\theta}_k) - \dim_k \theta\big). \tag{4}$$

Unfortunately, $l^*(k)$ still depends on the unknown θ^*, and the choice of k minimizing $l^*(k)$ is impossible again. What Akaike proposes, and what seems to be the weakest point of his approach, is to take $l(\hat{\theta}_k) - \dim_k \theta$ instead of $l^*(k)$ (a heuristic justification lies in formula (4)) and to choose the model Θ_k ($k = 1, 2, \ldots, K$) for which this quantity takes on the minimal value. More exactly,

$$\mathrm{AIC}(k) = -2\big(l(\hat{\theta}_k) - \dim_k(\theta)\big)$$

is considered and the model with the maximal value of AIC(k) is accepted. In the Akaike theory, the AIC is typically applied to a family of nested models

$$\Theta_1 \subset \Theta_2 \subset \ldots \subset \Theta_k \subset \Theta$$

such that

$$\theta = (\theta_1, \theta_2, \ldots, \theta_K) \in \Theta_k \quad \text{if } \theta_{k+1} = \theta_{k+2} = \ldots = \theta_K = 0;$$

then $\dim_k \theta = k$, and $\mathrm{AIC}(k) = -2l(\hat{\theta}_k) + 2k$.

3. A discussion of the Akaike loss function. Let us consider the loss function (1) in a more detailed way. A heuristic justification for choosing this function lies in the properties of the Kullback–Leibler information quantity

$$I(\theta; \theta^*) = \int [\log f(t|\theta^*)]\, f(t|\theta^*)\, dt + \frac{1}{n} l^*(\theta),$$

hence we return to considering $I(\theta; \theta^*)$ instead of $l^*(\theta)$.

1. Suppose that $\{f(x|\theta), \theta \in R^1\}$ is a family of densities on the real line indexed by a real-valued parameter θ such that $f(x|\theta) = f(x-\theta)$ for a given density f (location family). Then

$$I(\theta; \theta^*) = \int \left(\log \frac{f(t)}{f(t-(\theta-\theta^*))} \right) f(t)dt,$$

which depends on θ and θ^* through the difference $\theta-\theta^*$ only. This, of course, is an advantage of the Akaike criterion $I(\theta; \theta^*)$.

For example, if

$$f(x|\theta) = (\sqrt{2\pi})^{-1} \exp\{-(x-\theta)^2/2\},$$

then

$$I(\theta; \theta^*) = \tfrac{1}{2}(\theta-\theta^*)^2$$

and the Akaike optimal solution of the estimation problem is identical with that with the mean square error of estimation as the criterion.

2. Let $\{f(x|\theta), \theta \in R^1_+\}$ be a scale family of distributions

$$f(x|\theta) = \theta^{-1} f(x/\theta)$$

for a given density function f. Then

$$I(\theta; \theta^*) = \log \frac{\theta}{\theta^*} + \int \left(\log \frac{f(t)}{f(\theta^* t/\theta)} \right) f(t)dt,$$

which depends on θ and θ^* through the ratio θ/θ^* only. This should be considered as a rather natural property of the loss of the estimation of a scale parameter (see, e.g., [1], Section 3.3). For example, if

$$f(x|\theta) = [\theta^q \Gamma(q)]^{-1} x^{q-1} \exp\{-x/\theta\},$$

q known, $\theta > 0$ to be estimated, then the Akaike criterion leads to the loss

$$I(\theta; \theta^*) = q\left(\frac{\theta^*}{\theta} - \log \frac{\theta^*}{\theta} - 1 \right).$$

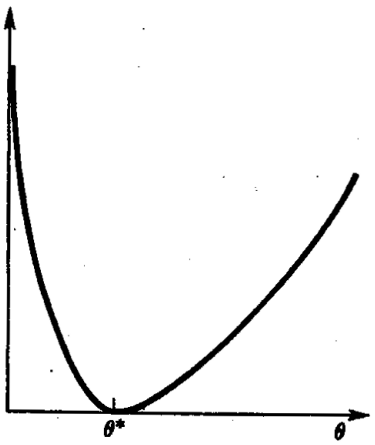

Fig. 1

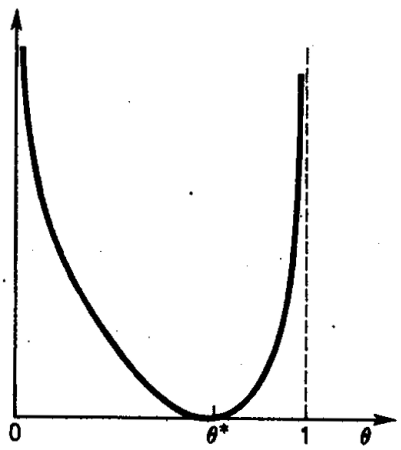

Fig. 2

This very attractive loss function in estimating a scale parameter is shown in Fig. 1, where $I(\theta; \theta^*)$ is considered as a function of the model value θ under a fixed true "value" θ^*.

3. The advantage of the Akaike loss function is not restricted to location and scale models only. Let us consider estimating θ in the binomial distribution

$$f(t|\theta) = \binom{n}{t}\theta^t(1-\theta)^{n-t}, \quad t = 0, 1, \ldots, n, \ 0 < \theta < 1.$$

Now we obtain (Fig. 2)

$$I(\theta; \theta^*) = n\left(\theta^* \log\frac{\theta^*}{\theta} + (1-\theta^*)\log\frac{1-\theta^*}{1-\theta}\right)$$

which is a convex function of θ, given θ^*, equal to zero iff $\theta = \theta^*$, and tending to infinity if θ approaches 0 or 1. This again gives a very persuasive loss function in estimating the binomial parameter, more reasonable than the quadratic loss function typically applied.

4. Choice of a model. An example. Consider the regression

$$x = at^2 + bt + c + \xi$$

with ξ distributed normally $N(0, \sigma^2)$, and suppose that one should decide if the coefficient a is "practically" zero (say, $|a| < \varepsilon$ for a suitable ε), or it is positive ($a > \varepsilon$), or negative ($a < -\varepsilon$). To avoid too many technicalities suppose that all other parameters, including σ^2, are known. Then, after a suitable normalization, the problem is to consider a random variable X distributed normally $N(\theta^*, 1)$, θ^* unknown, and to choose one of the models

$$\Theta_1 = \{\theta: \theta \leqslant -\varepsilon\}, \quad \Theta_2 = \{\theta: |\theta| < \varepsilon\}, \quad \Theta_3 = \{\theta: \theta \geqslant \varepsilon\}.$$

Under the Akaike loss function we obtain

$$l_1^* = \inf_{\theta \leqslant -\varepsilon} l^*(\theta) = \inf_{\theta \leqslant -\varepsilon} \{\tfrac{1}{2}(\theta - \theta^*)\} = \begin{cases} 0 & \text{if } \theta^* \leqslant -\varepsilon, \\ \tfrac{1}{2}(\theta^* + \varepsilon)^2 & \text{if } \theta^* > -\varepsilon, \end{cases}$$

$$l_2^* = \begin{cases} \tfrac{1}{2}(\theta^* + \varepsilon)^2 & \text{if } \theta^* \leqslant -\varepsilon, \\ 0 & \text{if } |\theta^*| < \varepsilon, \\ \tfrac{1}{2}(\theta^* - \varepsilon)^2 & \text{if } \theta^* \geqslant \varepsilon, \end{cases}$$

$$l_3^* = \begin{cases} \tfrac{1}{2}(\theta^* - \varepsilon)^2 & \text{if } \theta^* < \varepsilon, \\ 0 & \text{if } \theta^* \geqslant \varepsilon. \end{cases}$$

As decision rules, it is natural to consider the following functions of the observation X:

$$\hat{k}_c(X) = \begin{cases} 1 & \text{if } X \leqslant -c, \\ 2 & \text{if } |X| < c, \\ 3 & \text{if } X \geqslant c, \end{cases}$$

where $c > 0$. Then the risk $R^*(\hat{k})$, considered as a function of the true value θ^*, is given by the formula

$$R^*(\hat{k}_c) = \begin{cases} \frac{1}{2}(\theta^*+\varepsilon)^2[\Phi(c-\theta^*)-\Phi(-c-\theta^*)]+\frac{1}{2}(\theta^*-\varepsilon)^2[1-\Phi(c-\theta^*)] & \\ \qquad \text{if } \theta^* \leqslant -\varepsilon, & \\ \frac{1}{2}(\theta^*+\varepsilon)^2\Phi(-c-\theta^*)+\frac{1}{2}(\theta^*-\varepsilon)^2[1-\Phi(c-\theta^*)] & \\ \qquad \text{if } |\theta^*| < \varepsilon, & \\ \frac{1}{2}(\theta^*+\varepsilon)^2\Phi(-c-\theta^*)+\frac{1}{2}(\theta^*-\varepsilon)^2[\Phi(c-\theta^*)-\Phi(-c-\theta^*)] & \\ \qquad \text{if } \theta^* \geqslant \varepsilon, & \end{cases}$$

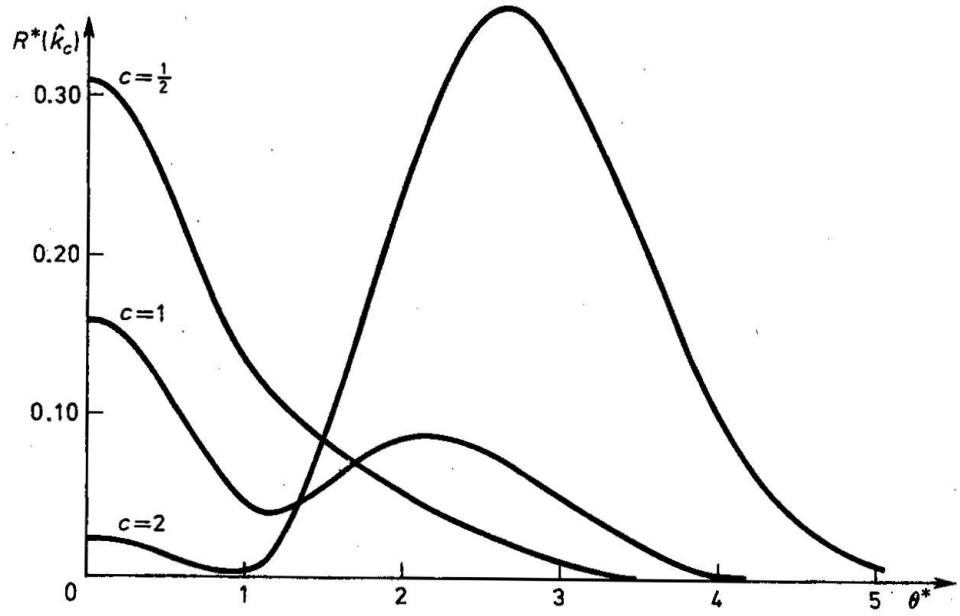

Fig. 3

where $\Phi(\cdot)$ is the cdf of $N(0, 1)$. The risk function $R^*(\hat{k}_c)$, for $\varepsilon = 1$ and $c = \frac{1}{2}$, 1 and 2, is outlined in Fig. 3 (the function is symmetrical around zero). One can easily observe that in the class $\{\hat{k}_c: c > 0\}$ of decision functions the one which minimizes, uniformly in θ^*, the risk $R^*(\hat{k}_c)$ does not exist; an "optimal" choice of c or $\hat{k}$ needs a further discussion which lies beyond the aim of this paper.

References

[1] E. L. Lehmann, *Theory of Point Estimation*, J. Wiley, New York 1983.
[2] A. A. Rényi, *Wahrscheinlichkeitsrechnung*, 6th edition, VEB Deutscher Verlag der Wissenschaften, Berlin 1979.
[3] Y. Sakamoto, M. Ishiguro and G. Kitagawa, *Akaike Information Criterion Statistics*, KTK Scientific Publisher, Tokyo, and D. Reidel Publishing Company, 1986.

RYSZARD ZIELIŃSKI
INSTYTUT MATEMATYCZNY PAN
SKRYTKA POCZTOWA 137
00-950 WARSZAWA

Received on 1988.03.17