Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle

J. deLeeuw
University of California at Los Angeles

Introduction

The problem of estimating the dimensionality of a model occurs in various forms in applied statistics: estimating the number of factors in factor analysis, estimating the degree of a polynomial describing the data, selecting the variables to be introduced in a multiple regression equation, estimating the order of an AR or MA time series model, and so on.

In factor analysis, this problem was traditionally solved by eyeballing residual eigenvalues, or by applying some other kind of heuristic procedure. When maximum likelihood factor analysis became computationally feasible, the likelihoods for different dimensionalities could be compared. Most statisticians were aware of the fact that the comparison of successive chi squares was not optimal in any well-defined decision theoretic sense. With the advent of the electronic computer, the forward and backward stepwise selection procedures in multiple regression also became quite popular, but again there were plenty of examples around showing that the procedures were not optimal and could easily lead one astray. When even more computational power became available, one could solve the best subset selection problem for up to 20 or 30 variables, but choosing an appropriate criterion on the basis of which to compare the many models remains a problem.

But exactly because of these advances in computation, finding a solution of the problem became more and more urgent. In the linear regression situation, the C_p criterion of Mallows (1973), which had already been around much longer, and the PRESS criterion of Allen (1974) were suggested. Although they seemed to work quite well, they were too limited in scope. The structural covariance models of Joreskog and others, and the log linear models of Goodman and others, made search over a much more complicated set of models necessary, and the model choice problems in those contexts could not be attacked by inherently linear methods. Three major closely related developments occurred around 1974. Akaike (1973) introduced the information criterion for model selection, generalizing his earlier work on time series analysis and factor analysis. Stone (1974) reintroduced and systematized cross-validation procedures, and Geisser (1975) discussed predictive sample reuse methods. In a sense, Stone–Geisser cross-validation is the more general procedure, but the information criterion (which rapidly became Akaike's information criterion or AIC) caught on more quickly.

There are various reasons for this. Akaike's many students and colleagues applied AIC almost immediately to a large number of interesting examples (compare Sakamoto, Ishiguro, and Kitagawa, 1986). In a sense, the AIC was more original and more daring than cross-validation, which simply seemed to amount to a lot of additional dreary computation. AIC has a close connection to the maximum likelihood method, which to many statisticians is still the ultimate in terms of rigor and precision. Moreover, the complicated structural equations and loglinear analysis programs were based on maximum likelihood theory, and the AIC criterion could be applied to the results without any additional computation. The AIC could be used to equip computerized "instant science" packages such as LISREL with an automated model search and comparison procedure, leaving even fewer decisions for the user (de Leeuw, 1989). And finally, Akaike and his colleagues succeeded in connecting the AIC effectively to the always mysterious area of the foundations of statistics. They presented the method, or at least one version of it, in a Bayesian framework (Akaike, 1977, 1978). There are many statisticians who consider the possibility of such a Bayesian presentation an advantage of the method.

Akaike's 1973 Paper

Section 1. Introduction

We start our discussion of the paper with a quotation. In the very first sentence, Akaike defines his information criterion, and the statistical principle that it implies.

> Given a set of estimates $\hat{\theta}$'s of the vector of parameters θ of a probability distribution with density $f(x|\theta)$ we adopt as our final estimate the one which will give the maximum of the expected log-likelihood, which is by definition
>
> $$\mathbf{E}(\log f(X|\hat{\theta})) = \mathbf{E}\left(\int f(x|\theta)\log f(x|\hat{\theta})\,dx\right),$$
>
> where X is a random variable following the distribution with the density function $f(x|\theta)$ and is independent of $\hat{\theta}$.

This is an impressive new principle, but its precise meaning is initially rather unclear. It is important to realize, for example, that in this definition the expected value on the left is with respect to the joint distribution of $\hat{\theta}$ and X, while the expected value on the right is with respect to the distribution of $\hat{\theta}$. It is also important that the expected log-likelihood depends both on the estimate $\hat{\theta}$ and the true value θ_0. We shall try to make this more clear by using the notation $\hat{\theta}(Z)$ for the estimate, where Z is the data, and Z is independent of X.

Akaike's principle now tells us to maximize over a class of estimates, but it does not tell us over which class, and it also does not tell us what to do about the problem when θ_0 is unknown. He points out this is certainly not the same as the principle of maximum likelihood, which adopts as the estimate the $\hat{\theta}(Z)$ that maximizes the log-likelihood $\log f(z|\theta)$ for a given realization of Z. For maximum likelihood, of course, we do not need to know θ_0.

What remains to be done is to further clarify the unclear points we mentioned above and to justify this particular choice of distance measure. This is what Akaike sets out to do in the rest of his paper.

Section 2. Information and Discrimination

In this section, Akaike justifies, or at least discusses, the choice of the information criterion. The model $f(\cdot|\theta)$ is a family of parametrized probability densities, with $\theta \in \Theta$. We shall simply refer to both θ and Θ as "models," understanding that the "model" Θ is a set of simple "models" θ. Suppose we want to compare a general model θ with the "true" model θ_0. From general decision theory, we know that comparisons can be based without loss of efficiency on the likelihood ratio $\tau(\cdot) = f(\cdot|\theta)/f(\cdot|\theta_0)$. This suggests that we define the *discrimination* between θ and θ_0 at x as $\Phi(\tau(x))$ for some function Φ, and to define the *mean discrimination* between θ and θ_0, if θ_0 is "true," as

$$\mathscr{D}(\theta, \theta_0, \Phi) = \int_{-\infty}^{+\infty} f(x|\theta_0)\Phi(\tau(x))\,dx = \mathbf{E}_X[\Phi(\tau(X))],$$

where $\mathbf{E}_X$ is the expected value over X, which has density $f(\cdot|\theta_0)$.

Now how do we choose Φ? We study $\mathscr{D}(\theta, \theta_0,\Phi)$ for θ close to θ_0. Under suitable regularity conditions, we have

$$\mathscr{D}(\theta, \theta_0; \Phi) = \Phi(1) + \tfrac{1}{2}\ddot{\Phi}(1)(\theta - \theta_0)'\mathscr{I}(\theta_0)(\theta - \theta_0) + o(\|\theta - \theta_0\|^2),$$

where

$$\mathscr{I}(\theta_0) = \int_{-\infty}^{+\infty} \left[\left(\frac{\partial \log f(x|\theta)}{\partial\theta}\right)_{\theta=\theta_0}\left(\frac{\partial \log f(x|\theta)}{\partial\theta}\right)'_{\theta=\theta_0}\right] f(x|\theta_0)\,dx$$

is the *Fisher information* at θ_0. Thus, it makes sense to require that $\Phi(1) = 0$ and $\ddot{\Phi}(1) > 0$ in order to make $\mathscr{D}$ behave like a distance. Akaike concludes, correctly, that this derivation shows the major role played by $\log f(\cdot|\theta)$, and he also concludes, somewhat mysteriously, that consequently, the choice $\Phi(t) = -2\log(t)$ makes good sense. Thus, he arrives at his entropy measure, known in other contexts as the *negentropy* or *Kullback–Leibler distance.*

$$\begin{aligned}\mathscr{D}(\theta, \theta_0) &= 2\int_{-\infty}^{+\infty} f(x|\theta_0)\log\frac{f(x|\theta_0)}{f(x|\theta)}\,dx \\ &= 2\mathbf{E}_X[\log f(X|\theta_0)] - 2\mathbf{E}_X[\log f(X|\theta)].\end{aligned}$$

It follows from the inequality $\ln t > 1 + \mathrm{t}$ that the negentropy is always nonnegative, and it is equal to zero if and only if $f(\cdot|\theta) = f(\cdot|\theta_0)$ a.e. The negentropy can consequently be interpreted as a measure of *distance* between $f(\cdot|\theta)$ and the true distribution. The Kullback-Leibler distance was introduced in statistics as early as 1951, and its use in hypothesis testing and model evaluation was propagated strongly by Kullback (1959). Akaike points out that maximizing the expected log-likelihood amounts to the same thing as minimizing $\mathbf{E}_z[\mathscr{D}(\hat{\theta}(Z), \theta_0)]$, the expected value over the data of the Kullback-Leibler distance between the estimated density $f(\cdot|\hat{\theta}(Z))$ and the true density $f(\cdot|\theta_0)$. He calls $\mathscr{D}(\hat{\theta}(Z), \theta_0)$ the *probabilistic negentropy* and uses the symbol $\mathscr{R}(\theta_0)$ for its expected value.

The justification given by Akaike for using $\Phi(t) = -2\log(t)$ may seem a bit weak, but the result is a natural distance measure between probability densities, which has strong connections with the Shannon–Wiener information criterion, Fisher information, and entropy measures used in thermodynamics. One particular reason why this measure is attractive is the situation in which we have n repeated independent trials according to $f(\cdot|\theta_0)$. This leads to densities $f_n(\cdot, \theta)$ and $f_n(\cdot, \theta_0)$ that are products of the densities of the individual observations. If $\mathscr{D}_n(\theta, \theta_0)$ is the Kullback-Leibler distance between these two product densities, then trivially $\mathscr{D}_n(\theta, \theta_0) = n\ \mathscr{D}(\theta, \theta_0)$. Obviously, the additivity of the negentropy in the case of repeated independent trials is an important point in its favour.

Section 3. Information and the Maximum Likelihood Principle

Now Akaike has to discuss what to do about the problem of the unknown θ_0. The solution he suggests is actually very similar to the approach of classical statistical large sample theory, but because of the context of the information principle, we see it in a new light.

Remember that the *entropy maximization principle* tells us to evaluate the success of our procedure, and the appropriateness of the model Θ, by computing the expectation $\mathscr{R}(\theta_0)$ of the probabilistic negentropy over the data. Also remember that

$$\mathscr{R}(\theta_0) = 2\mathbf{E}_X[\log f(X|\theta_0)] - 2\mathbf{E}_{X,Z}[\log f(X|\hat{\theta}_0(Z))],$$

which means that minimizing the expected probabilistic negentropy does indeed amount to the same thing as maximizing the expected log-likelihood mentioned in Sec. 1. Akaike's program is to estimate $\mathscr{R}(\theta_0)$, and if several models are compared, to select the model with the smallest value.

Of course, it is still not exactly easy to carry out this program. Because θ_0 is unknown we cannot really minimize the negentropy, and we cannot compute the expectation of the minimum over Z either. There is an approximate solution to this problem, however, if we have a large number of independent replications (or, more generally, if the law of large numbers applies). Minus the *mean log-likelihood ratio*

$$\hat{\mathscr{D}}_n(\theta, \theta_0) = \frac{2}{n}\sum_{i=1}^{n} \log \frac{f(x_i|\theta_0)}{f(x_i|\theta)}$$

will converge in probability to the negentropy, and under suitable regularity conditions, this convergence will be uniform in θ. This makes it plausible that maximizing the mean log- likelihood ratio (i.e., computing the *maximum likelihood estimate*) will tend to maximize the entropy, and that in the limit, the maximum likelihood estimate is the maximum entropy estimate. We do not need to know θ_0 in order to be able to compute the maximum likelihood estimate. Thus, Akaike justifies the use of maximum likelihood by deriving it from his information criterion. From now on, we will substitute the maximum likelihood estimate $\hat{\theta}(Z)$ for the unknown θ_0.

Section 4. Extension of the Maximum Likelihood Principle

This is the main theoretical section of the paper. Akaike proposes to combine point estimation and the testing of model fit into the single new principle of comparing the values of the mean log-likelihood or negentropy. This is his "extension" of the maximum likelihood principle. We have seen in the previous section that negentropy is minimized, approximately, by using the maximum likelihood estimate for $\hat{\theta}(Z)$. What must still be done is to find convenient approximations for $\mathscr{R}(\theta_0)$ at the maximum likelihood estimate.

This section is not particular easy to read. It does not have the usual proof/theorem format, expansions are given without precise regularity conditions, exact and asymptotic identities are freely mixed, stochastic and deterministic expressions are not clearly distinguished, and there are some unfortunate notational and especially typesetting choices. This is an "ideas paper," promoting a new approach to statistics, not a mathematics paper concerned with the detailed properties of a particular technique. Although we follow the paper closely, we have tried to make the notation a bit more explicit, for instance by using matrices.

Akaike analyzes the situation in which we have a number of subspaces Θ_k of Θ, with $0 \leq k \leq m$, Θ_{k+1} a subspace of Θ_k, and $\Theta_0 = \Theta$. Let $d_k = \dim(\Theta_k)$. Actually, it is convenient to simplify this, by a change of coordinates, to the problem in which $d = m$, $d_k = k$, and Θ_k is the subspace of $\Re^m$, which has the last $m - k$ elements equal to zero. We assume $\theta_0 \in \Theta_0$, and we assume we have n independent replications in Z. Let $\hat{\theta}_k(Z)$ be the corresponding maximum likelihood estimates. Akaike suggests that we estimate the expectation of the probabilistic entropy $\mathscr{R}(\theta_0)$ by using $\hat{\mathscr{D}}_n(\hat{\theta}_k(Z), \hat{\theta}_0(Z))$. But $\hat{\mathscr{D}}_n(\hat{\theta}_k(Z), \hat{\theta}_0(Z))$ will be a biased estimator of $\mathscr{R}(\theta_0)$, because of the substitution of the maximum likelihood estimator for θ_0.

It is known that $n\,\hat{\mathscr{D}}_n(\hat{\theta}_k(Z), \hat{\theta}_0(Z))$ is asymptotically chi square with $m - k$ degrees of freedom if $\theta_0 \in \Theta_k$. In general, $\hat{\mathscr{D}}_n(\hat{\theta}_k(Z), \hat{\theta}_0(Z))$ will converge in probability to $\mathscr{D}(\Theta_k, \theta_0)$, i.e., the Kullback–Leibler distance between θ_0 and the model closest to θ_0 in Θ_k. Now if $n\,\mathscr{D}(\Theta_k, \theta_0)$ is much larger than m, then the mean likelihood ratio will be very much larger than expected from the chi square appoximation. If $n\,\mathscr{D}(\Theta_k, \theta_0)$ is much smaller than m, then we can do statistics on the basis of the chi square because the model is "true." But the intermediate case, in which the two quantities are of the same order, and the model Θ_k is "not too false," is the really interesting one. This is the case Akaike sets out to study. It is, of course, similar to studying the Pitman power of large-sample tests by using sequences of alternatives converging to the null value.

First, we offer some simplifications. Instead of studying $\mathscr{D}(\theta, \theta_0)$, Akaike uses the quadratic approximation $\mathscr{W}(\theta, \theta_0) = (\theta - \theta_0)'I(\theta_0)(\theta - \theta_0)$ discussed in Sec. 2. Asymptotically, this leads to the same conclusions to the order of approximation that is used. He uses the Fisher information matrix $I(\theta_0)$ to define an inner product $\langle \cdot, \cdot \rangle_0$ and a norm $\|\cdot\|_0$ on Θ, so that $\mathscr{W}(\theta, \theta_0) = \|\theta - \theta_0\|_0^2$. Define $\theta_{0|k}$ as the projection of θ_0 on Θ_k in the information metric. Then, by Pythagoras,

$$\mathscr{W}(\hat{\theta}_k(Z), \theta_0) = \|\theta_{0|k} - \theta_0\|^2 + \|\hat{\theta}_k(Z) - \theta_{0|k}\|^2. \tag{1}$$

The idea is to use $\mathbf{E}_Z[\mathscr{W}(\hat{\theta}_k(Z), \theta_0)]$ to estimate $\mathscr{R}(\theta_0)$.

The first step in the derivation is to expand the mean log-likelihood ratio in a Taylor series. This gives

$$n\hat{\mathscr{D}}_n(\hat{\theta}_0(Z), \theta_{0|k}) = n(\hat{\theta}_0(Z) - \theta_{0|k})'\mathscr{H}[\hat{\theta}_0(Z), \theta_{0|k}](\hat{\theta}_0(Z) - \theta_{0|k}),$$
$$n\hat{\mathscr{D}}_n(\hat{\theta}_k(Z), \theta_{0|k}) = n(\hat{\theta}_k(Z) - \theta_{0|k})'\mathscr{H}[\hat{\theta}_k(Z), \theta_{0|k}](\hat{\theta}_k(Z) - \theta_{0|k}),$$

where

$$\mathscr{H}[\theta, \zeta] = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(x_i|\theta + \rho(\zeta - \theta))}{\partial\theta\partial\theta'},$$

for some $0 \le \rho \le 1$. Subtracting the two expansions gives

$$\begin{aligned} n\hat{\mathscr{D}}_n(\hat{\theta}_k(Z), \theta_0(Z)) = {} & n(\hat{\theta}_0(Z) - \theta_{0|k})'\mathscr{H}[\hat{\theta}_0(Z), \theta_{0|k}](\hat{\theta}_0(Z) - \theta_{0|k}) \\ & - n(\hat{\theta}_k(Z) - \theta_{0|k})'\mathscr{H}[\hat{\theta}_k(Z), \theta_{0|k}](\hat{\theta}_k(Z) - \theta_{0|k}). \end{aligned}$$

Let n and k tend to infinity in such a way that $n^{1/2}\,(\theta_{0|k} - \theta_0)$ stays bounded. Then, taking plims, we get

$$n\hat{\mathscr{D}}_n(\hat{\theta}_k(Z), \theta_0(Z)) \approx n\|\hat{\theta}_0(Z) - \theta_{0|k}\|_0^2 - n\|\hat{\theta}_k(Z) - \theta_{0|k}\|_0^2. \tag{2}$$

This can also be written as

$$\begin{aligned} n\hat{\mathscr{D}}_n(\hat{\theta}_k(Z), \hat{\theta}_0(Z)) \approx{}& n\|\theta_{0|k} - \theta_0\|_0^2 + n\|\hat{\theta}_0(Z) - \theta_0\|_0^2 - n\|\hat{\theta}_k(Z) - \theta_{0|k}\|_0^2 \\ & - 2n\langle\hat{\theta}_0(Z) - \theta_0, \theta_{0|k} - \theta_0\rangle \end{aligned} \tag{3}$$

In the next step, Taylor expansions are used again. For this step, we use the special symbol $=_k$, where two vectors x and y satisfy $x =_k y$ if their first k elements are equal.

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}\left[\frac{\partial \log f(x_i|\theta)}{\partial\theta}\right]_{\theta=\theta_{0|k}} &=_k n^{1/2}\mathscr{H}[\hat{\theta}_k(Z), \theta_{0|k}](\theta_{0|k} - \hat{\theta}_k(Z)) \\ &=_k n^{1/2}\mathscr{H}[\hat{\theta}_0(Z), \theta_{0|k}](\theta_{0|k} - \hat{\theta}_0(Z)) \end{aligned}$$

Then let n and k tend to infinity again in such a way that $n^{1/2}(\theta_{0|k} - \theta_0)$ stays bounded and take plims. This gives

$$n^{1/2}I(\theta_0)(\hat{\theta}_k(Z) - \theta_{0|k}) \approx_k n^{1/2}I(\theta_0)(\hat{\theta}_0(Z) - \theta_{0|k}),$$

and because of the definition of $\theta_{0|k}$ also,

$$n^{1/2}I(\theta_0)(\hat{\theta}_k(Z) - \theta_{0|k}) \approx_k n^{1/2}I(\theta_0)(\hat{\theta}_0(Z) - \theta_0). \tag{4}$$

It follows that $(\hat{\theta}_k(Z) - \theta_{0|k})$ is approximately the projection of $(\hat{\theta}_0(Z) - \theta_0)$ on Θ_k.

This implies that $n\|\hat{\theta}_0(Z) - \theta_0\|_0^2 - n\|\hat{\theta}_k(Z) - \theta_{0|k}\|_0^2$ and $n\|\hat{\theta}_k(Z) - \theta_{0|k}\|_0^2$ are asymptotically independent chi squares, with degrees of freedom $m - k$ and k. Akaike then indicates that the last (linear) term on the right-hand side of (3) is small compared to the other (quadratic) terms. If we ignore its contribution, and then subtract (3) from (1), we find

$$\begin{aligned} & n\mathscr{W}(\hat{\theta}_k(Z), \theta_0) - n\hat{\mathscr{D}}_n(\hat{\theta}_k(Z), \hat{\theta}_0(Z)) \\ &\qquad \approx n\|\hat{\theta}_k(Z) - \theta_{0|k}\|^2 - n\|\hat{\theta}_0(Z) - \theta_0\|_0^2 - n\|\hat{\theta}_k(Z) - \theta_{0|k}\|_0^2. \end{aligned}$$

Replacing the chi squares by their expectations gives

$$n\mathbf{E}_Z[\mathscr{W}(\hat{\theta}_k(Z), \theta_0)] \approx n\hat{\mathscr{D}}_n(\hat{\theta}_k(Z), \hat{\theta}_0(Z)) + 2k - m. \tag{5}$$

This defines the AIC. Of course, in actual examples, m may not be known or may be infinite (think of order estimation or log-spline density estimation), but in comparing models, we do not actually need m anyway, because it is the same for all models. Thus, in practice we simply compute $-2\sum_{i=1}^{n}\log f(x_i\,\hat{\theta}_k(Z)) + 2k$ for various values of k.

Section 5. Applications

In this section, Akaike discusses the possible applications of his principle to problems of model selection. As we pointed out in the introduction, the systematic approach to these problems and the simple answer provided by the AIC, at no additional cost, have certainly had an enormous impact. The theoretical contributions of the paper, discussed above, have been much less influential than the practical ones. The recipe has been accepted rather uncritically by many applied statisticians in the same way as the principles of least-squares or maximum likelihood or maximum posterior probability have been accepted in the past without much questioning.

Recipes for the application of the AIC to factor analysis, principal component analysis, analysis of variance, multiple regression, and autoregressive model fitting in time series analysis are discussed. It is interesting that Akaike already published applications of the general principle to time series analysis in 1969 and to factor analysis in 1971. He also points out the equivalence of the AIC to C_p proposed by Mallows in the linear model context.

Section 6. Numerical Examples

This section has two actual numerical examples, both estimating the order k of an autoregressive series. Reanalyzing data by Jenkins and Watts leads to the estimate $k = 2$, the same as that found by the orginal analysis using partial autocorrelation methods. A reanalysis of an example by Whittle leads to $k = 65$, while Whittle has decided on $k = 4$ using likelihood-ratio tests. Akaike argues that this last example illustrates dramatically that using successive log-likelihoods for testing can be quite misleading.

Section 7. Concluding Remarks

Here Akaike discusses briefly, again, the relations between maximum likelihood, the dominant paradigm in statistics, and the Shannon–Wiener entropy, the dominant paradigm in information and coding theory. As Sec. 3 shows, there are strong formal relationships, and using expected likelihood (or entropy) makes it possible to combine point-estimation and hypothesis testing in a single framework. It also gives "easy"answers to very important but very difficult multiple-decision problems.

Discussion

The reasoning behind using X, the independent replication, to estimate $\mathscr{R}(\theta_0)$, is the same as the reasoning behind *cross-validation.* We use $\hat{\theta}(Z)$ to predict X, using $f(X\,|\,\hat{\theta}(Z)$ as the criterion. If we use the maximum likelihood estimate, we systematically underestimate the distance between the data and the model, because the estimate is constructed by minimizing this distance. Thus, we need an independent replication to find out how good our fit is, and plugging in the independent replication leads to overestimation of the distance. The AIC corrects for both biases. The precise relationship between AIC and cross-validation has been discussed by Stone (1977). At a later stage, Akaike (1978) provided an asymptotic Bayesian justification of sorts. As we have indicated, AIC estimates the expected distance between the model and the true value. We could also formulate a related decision problem as estimating the dimensionality of the model, for instance by choosing from a nested sequence of models. It can be shown that the minimum AIC does not necessarily give a consistent estimate of the true dimensionality. Thus, we may want to construct better estimates, for instance choosing the model dimensionality with the highest posterior probability. This approach, however, has led to a proliferation of criteria, among them the BIC criteria of Schwartz (1978) and Akaike (1977), or the MDL principle of Rissanen (1978 and later papers). Other variations have been proposed by Shibata, Bozdogan, Hannan, and others. Compare Sclove (1987), or Hannan and Deistler (1988, Chap. 7), for a recent review. Recently, Wei (1990) proposed a new "F.I.C." criterion, in which the complexity of the selected model is penalized by its redundant Fisher informations, rather than by the dimensionality used in the conventional criteria. We do not discuss these alternative criteria here, because they would take us too far astray and entangle us in esoteric asymptotics and ad hoc inference principles. We think the justification based on cross-validation is by far the most natural one.

We have seen that the paper discussed here was an expository one, not a mathematical one. It seems safe to assume that many readers simply skipped Sec. 4 and rapidly went on to the examples. We have also seen that the arguments given by Akaike in this expository are somewhat heuristic, but in later work by him, and by his students such as Inagaki and Shibata, a rigorous version of his results has also been published. Although many people contributed to the area of model selection criteria and there are now many competing criteria, it is clear that Akaike's AIC is by far the most important contribution. This is due to the forceful presentation and great simplicity of the criterion, and it may be due partly to the important position of Akaike in Japanese and international statistics. But most of all, we like to think, the AIC caught on so quickly because of the enormous emphasis on interesting and very real practical applications that has always been an important component of Akaike's work.

Biographical Information

Hirotogu Akaike was born in 1927 in Fujinomiya-shi, Shizuoka-jen, in Japan. He completed the B.S. and D.S. degrees in mathematics at the University of Tokyo in 1952 and 1961. He started working at the Institute of Statistical

Mathematics in 1952, worked his way up through the ranks, and became its Director General in 1982. In 1976, he had already become editor of the *Annals of the Institute of Statistical Mathematics*, and he still holds both these functions, which are certainly the most important in statistics in Japan. Akaike has received many prizes and honors: He is a member of the I.S.I., Fellow of the I.M.S., Honorary Fellow of the R.S.S., and current (1990) president of the Japanese Statistical Society.

It is perhaps safe to say that Akaike's main contribution has been in the area of time series analysis. He developed in an early stage of his career the program package TIMSAC, for time series analysis and control, and he and his students have been updating TIMSAC, which is now in its fourth major revision and extension. TIMSAC has been used in many areas of science. In the course of developing TIMSAC, Akaike had to study the properties of optimization methods. He contributed the first theoretically complete study of the convergence properties of the optimum gradient (or steepest descent) method. He also analyzed and solved the identification problem for multivariate time series, using basically Kalman's state-space representation, but relating it effectively to canonical analysis. And in modeling autoregressive patterns, he came up with the FPE (or final prediction error) criterion, which later developed rapidly into the AIC.

References

Akaike, H. (1973). Information theory and the maximum likelihood principle in *2nd International Symposium on Information Theory* (B.N. Petrov and F. Csàki, eds.). Akademiai Kiàdo, Budapest.

Akaike, H. (1977). On the entropy maximization principle, in: *Applications of Statistics* (P.R. Krishnaiah, ed.). North- Holland, Amsterdam.

Akaike, H. (1978). A Bayesian analysis of the minimum A.I.C.. procedure, *Ann. Inst. Statist. Math., Tokyo*, **30**, 9–14.

Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method of prediction, Technometrics, **22**, 325–331.

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

de Leeuw, J. (1989). Review of Sakamoto et al., *Psychometrika*, **54**, 539–541.

Geisser, S. (1975). The predictive sample reuse method with applications, *J. Amer. Statist. Assoc.*, **70**, 320–328.

Hannan, E.J., and Deistler, M. (1988). *The Statistical Theory of Linear System.* Wiley, New York.

Kullback, S. (1959). Information theory and statistics, New York, Wiley.

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

Rissanen, J. (1978). Modeling by shortest data description, *Automatica*, **14**, 465–471.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike Information Criterion Statistics.* Reidel, Dordrecht, Holland.

Schwartz, G. (1978). Estimating the dimension of a model, *Ann. Statist.* **6**, 461–464.

Sclove, S.L. (1987). Application of model-selection criteria to some problems in multivariate analysis, *Psychometrika*, **52**, 333–344.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion), *J. Roy. Statist. Soc., Ser. B*, **36**, 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc., Ser, B*, **39**, 44–47.

Wei, C.Z. (1990). On predictive least squares. *Technical Report*, Department of Mathematics, University of Maryland, College Park, Md.

Source:

deLeeuw, J. 1992. Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle. Pages 599-609 *In:* Kotz, S., and N.L. Johnson, editors. *Breakthroughs in Statistics Volume 1. Foundations and Basic Theory.* Springer Series in Statistics, Perspectives in Statistics. Springer-Verlag: New York.

Information Theory and an Extension of the Maximum Likelihood Principle

Hirotogu Akaike
Institute of Statistical Mathematics

Abstract

In this paper it is shown that the classical maximum likelihood principle can be considered to be a method of asymptotic realization of an optimum estimate with respect to a very general information theoretic criterion. This observation shows an extension of the principle to provide answers to many practical problems of statistical model fitting.

1. Introduction

The extension of the maximum likelihood principle which we are proposing in this paper was first announced by the author in a recent paper [6] in the following form:

Given a set of estimates $\hat{\theta}$ of the vector of parameters θ of a probability distribution with density function $f(x|\theta)$ we adopt as our final estimate the one which will give the maximum of the expected log-likelihood, which is by definition

$$E \log f(X|\hat{\theta}) = E \int f(x|\theta) \log f(x|\hat{\theta})\, \mathrm{d}x, \tag{1.1}$$

where X is a random variable following the distribution with the density function $f(x|\theta)$ and is independent of $\hat{\theta}$.

This seems to be a formal extension of the classical maximum likelihood principle but a simple reflection shows that this is equivalent to maximizing an information theoretic quantity which is given by the definition

$$E \log\left(\frac{f(X|\hat{\theta})}{f(X|\theta)}\right) = E \int f(x|\theta) \log\left(\frac{f(x|\hat{\theta})}{f(x|\theta)}\right) \mathrm{d}x. \tag{1.2}$$

The integral in the right-hand side of the above equation gives the Kullback-Leibler's mean information for discrimination between $f(x|\hat{\theta})$ and $f(x|\theta)$ and is known to give a measure of separation or distance between the two distributions [15]. This observation makes it clear that what we are proposing here is the adoption of an information theoretic quantity of the discrepancy between the estimated and the true probability distributions to define the loss function of an estimate $\hat{\theta}$ of θ. It is well recognized that the statistical estimation theory should and can be organized within the framework of the theory of statistical decision functions [25]. The only difficulty in realizing this is the choice of a proper loss function, a point which is discussed in details in a paper by Le Cam [17].

In the following sections it will be shown that our present choice of the information theoretic loss function is a very natural and reasonable one to develop a unified asymptotic theory of estimation. We will first discuss the definition of the amount of information and make clear the relative merit, in relation to the asymptotic estimation theory, of the Kullback–Leibler type information within the infinitely many possible alternatives. The discussion will reveal that the log-likelihood is essentially a more natural quantity than the simple likelihood to be used for the definition of the maximum likelihood principle.

Our extended maximum likelihood principle can most effectively be applied for the decision of the final estimate of a finite parameter model when many alternative maximum likelihood estimates are obtained corresponding to the various restrictions of the model. The log-likelihood ratio statistics developed for the test of composite hypotheses can most conveniently be used for this purpose and it reveals the truly statistical nature of the information theoretic quantities which have often been considered to be probabilistic rather than statistical [21].

With the aid of this log-likelihood ratio statistics our extended maximum likelihood principle can provide solutions for various important practical problems which have hitherto been treated as problems of statistical hypothesis testing rather than of statistical decision or estimation. Among the possible applications there are the decisions of the number of factors in the factor analysis, of the significant factors in the analysis of variance, of the number of independent variables to be included into multiple regression and of the order of autoregressive and other finite parameter models of stationary time series.

Numerical examples are given to illustrate the difference of our present approach from the conventional procedure of successive applications of statistical tests for the determination of the order of autoregressive models. The results will convincingly suggest that our new approach will eventually be replacing many of the hitherto developed conventional statistical procedures.

2. Information and Discrimination

It can be shown [9] that for the purpose of discrimination between the two probability distributions with density functions $f_i(x)$ $(i = 0, 1)$ all the necessary information are contained in the likelihood ratio $T(x) = f_1(x)/f_0(x)$ in the sense that any decision procedure with a prescribed loss of discriminating the two distributions based on a realization of a sample point x can, if it is realizable at all, equivalently be realized through the use of $T(x)$. If we consider that the information supplied by observing a realization of a (set of) random variable(s) is essentially summarized in its effect of leading us to the discrimination of various hypotheses, it will be reasonable to assume that the amount of information obtained by observing a realization x must be a function of $T(x) = f_1(x)/f_0(x)$.

Following the above observation, the natural definition of the mean amount of information for discrimination per observation when the actual distribution is $f_0(x)$ will be given by

$$I(f_1, f_0; \Phi) = \int \Phi\left(\frac{f_1(x)}{f_0(x)}\right) f_0(x)\, \mathrm{d}x, \tag{2.1}$$

where $\Phi(r)$ is a properly chosen function of r and $\mathrm{d}x$ denotes the measure with respect to which $f_i(x)$ are defined. We shall hereafter be concerned with the parametric situation where the densities are specified by a set of parameters θ in the form

$$f(x) = f(x|\theta), \tag{2.2}$$

where it is assumed that θ is an L-dimensional vector, $\theta = (\theta_1, \theta_2, \ldots, \theta_L)'$, where $'$ denotes the transpose. We assume that the true distribution under observation is specified by $\theta = \uptheta = (\uptheta_1, \uptheta_2, \ldots, \uptheta_L)'$. We Will denote by $I(\theta, \uptheta; \Phi)$ the quantity defined by (2.1) with $f_1(x) = f(x|\theta)$ and $f_0(x) = f(x|\uptheta)$ and analyze the sensitivity of $I(\theta, \uptheta; \Phi)$ to the deviation of θ from $\uptheta$. Assuming the regularity conditions of $f(x|\theta)$ and $\Phi(r)$ which assure the following analytical treatment we get

$$\frac{\partial}{\partial\theta_1} I(\theta, \uptheta; \Phi)|_{\theta=\uptheta} = \int \left(\frac{\mathrm{d}}{\mathrm{d}r}\Phi(r)\frac{\partial r}{\partial\theta_l}\right)_{\theta=\uptheta} f_\theta\, \mathrm{d}x = \Phi(1) \int \left(\frac{\partial f_\theta}{\partial\theta_l}\right)_{\theta=\uptheta} \mathrm{d}x \tag{2.3}$$

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_l\partial\theta_m} I(\theta, \uptheta; \Phi)|_{\theta=\theta} &= \int \left[\left(\frac{\mathrm{d}^2}{\mathrm{d}r^2}\Phi(r)\right)\left(\frac{\partial r}{\partial\theta_l}\right)\left(\frac{\partial r}{\partial\theta_m}\right)\right]_{\theta=\uptheta} f_\uptheta\, \mathrm{d}x \\
&\quad + \int \left[\left(\frac{\mathrm{d}}{\mathrm{d}r}\Phi(r)\right)\left(\frac{\partial^2 r}{\partial\theta_l\partial\theta_m}\right)\right]_{\theta=\uptheta} f_\uptheta\, \mathrm{d}x \\
&= \ddot{\Phi}(1) \int \left[\left(\frac{\partial f_\theta}{\partial\theta_l}\frac{1}{f_\uptheta}\right)\left(\frac{\partial f_\theta}{\partial\theta_m}\frac{1}{f_\uptheta}\right)\right]_{\theta=\uptheta} f_\uptheta\, \mathrm{d}x \\
&\quad + \dot{\Phi}(1) \int \left(\frac{\partial^2 f_\theta}{\partial\theta_l\partial\theta_m}\right)_{\theta=\uptheta} \mathrm{d}x,
\end{aligned} \tag{2.4}$$

where r, $\dot{\Phi}(1)$, $\ddot{\Phi}(1)$ and f_θ denote $\dfrac{f(x|\theta)}{f(x|\mathrm{\theta})}$, $\left.\dfrac{d\Phi(r)}{dr}\right|_{r=1}$, $\left.\dfrac{\mathrm{d}^2\Phi(r)}{\mathrm{d}r^2}\right|_{r=1}$ and $f(x|\theta)$, respectively, and the meaning of the other quantities will be clear from the context. Taking into account that we are assuming the validity of differentiation under integral sign and that $\int f(x|\theta)\,\mathrm{d}x = 1$, we have

$$\int\left(\frac{\partial f}{\partial\theta_l}\right)\mathrm{d}x = \int\left(\frac{\partial^2 f}{\partial\theta_l\partial\theta_m}\right)\mathrm{d}x = 0. \tag{2.5}$$

Thus we get

$$I(\mathrm{\theta}, \mathrm{\theta}; \Phi) = \Phi(1) \tag{2.6}$$

$$\frac{\partial}{\partial\theta_l} I(\theta, \mathrm{\theta}; \Phi)|_{\theta=\mathrm{\theta}} = 0 \tag{2.7}$$

$$\frac{\partial^2}{\partial\theta_l\partial\theta_m} I(\theta, \mathrm{\theta}; \Phi)|_{\theta=\mathrm{\theta}} = \ddot{\Phi}(1)\int\left[\left(\frac{\partial f_\theta}{\partial\theta_l}\frac{1}{f_\theta}\right)\left(\frac{\partial f_\theta}{\partial\theta_m}\frac{1}{f_\theta}\right)\right]_{\theta=\mathrm{\theta}} f_\mathrm{\theta}\,\mathrm{d}x. \tag{2.8}$$

These relations show that $\ddot{\Phi}(1)$ must be different from zero if $I(\theta, \mathrm{\theta}; \Phi)$ ought to be sensitive to the small variations of θ. Also it is clear that the relative sensitivity of $I(\theta, \mathrm{\theta}; \Phi)$ is high when $\left|\dfrac{\ddot{\Phi}(1)}{\Phi(1)}\right|$ is large. This will be the case when $\Phi(1) = 0$. The integral on the right-hand side of (2.8) defines the (l, m)th element of Fisher's information matrix [16] and the above results show that this matrix is playing a central role in determining the behaviour of our mean information $I(\theta, \mathrm{\theta}; \Phi)$ for small variations of θ around $\mathrm{\theta}$. The possible forms of $\Phi(r)$ are e.g. $\log r$, $(r - 1)^2$ and $r^{1/2}$ and we cannot decide uniquely at this stage.

To restrict further the form of $\Phi(r)$ we consider the effect of the increase of information by N independent observations of X. For this case we have to consider the quantity

$$I_N(\theta, \mathrm{\theta}; \Phi) = \int \Phi\,\frac{\prod_{i=1}^{N} f(x_i|\theta)}{\prod_{i=1}^{N} f(x_i|\mathrm{\theta})}\,\prod_{i=1}^{N} f(x_i|\mathrm{\theta})\,\mathrm{d}x_1 \ldots \mathrm{d}x_N. \tag{2.9}$$

Corresponding to (2.5), (2.6) and (2.7) we have

$$I_N(\mathrm{\theta}, \mathrm{\theta}; \Phi) = I(\mathrm{\theta}, \mathrm{\theta}; \Phi) \tag{2.10}$$

$$\frac{\partial}{\partial\theta_l} I_N(\theta, \mathrm{\theta}; \Phi)|_{\theta=\mathrm{\theta}} = 0 \tag{2.11}$$

$$\frac{\partial^2}{\partial\theta_l\partial\theta_m} I_N(\theta, \mathrm{\theta}; \Phi)|_{\theta=\theta} = N\frac{\partial^2}{\partial\theta_l\partial\theta_m} I(\theta, \mathrm{\theta}; \Phi)|_{\theta=\mathrm{\theta}}. \tag{2.12}$$

These equations show that $I_N(\mathrm{\theta}, \mathrm{\theta}; \Phi)$ is not responsive to the increase of information and that $\frac{\partial^2}{\partial\theta_l\partial\theta_m} I_N(\theta, \uptheta; \Phi)|_{\theta=\uptheta}$ is in a linear relation with N. It can be seen that only the quantity defined by

$$\frac{\partial \prod_{i=1}^{N} f(x_i|\theta)}{\partial\theta_l} \frac{1}{\prod_{i=1}^{N} f(x_i|\uptheta)}\Bigg|_{\theta=\uptheta} = \sum_{i=1}^{N} \left(\frac{\partial f(x_i|\theta)}{\partial\theta_l}\frac{1}{f_\theta}\right)_{\theta=\uptheta} \tag{2.13}$$

is concerned with the derivation of this last relation. This shows very clearly that taking into account the relation

$$\frac{\partial f(x|\theta)}{\partial\theta_l}\frac{1}{f_\theta} = \frac{\partial \log f(x|\theta)}{\partial\theta_l}, \tag{2.14}$$

the functions $\frac{\partial}{\partial\theta_l} \log f(x|\theta)$ are playing the central role in the present definition of information. This observation suggests the adoption of $\Phi(r) = \log r$ for the definition of our amount of information and we are very naturally led to the use of Kullback-Leibler's definition of information for the purpose of our present study.

It should be noted here that at least asymptotically any other definition of $\Phi(r)$ will be useful if only $\Phi(1)$ is not vanishing. The main point of our present observation will rather be the recognition of the essential role being played by the functions $\frac{\partial}{\partial\theta_l} \log f(x|\theta)$ for the definition of the mean information for the discrimination of the distributions corresponding to the small deviations of θ from $\uptheta$.

3. Information and the Maximum Likelihood Principle

Since the purpose of estimating the parameters of $f(x|\uptheta)$ is to base our decision on $f(x|\hat{\theta})$, where $\hat{\theta}$ is an estimate of $\uptheta$, the discussion in the preceding section suggests the adoption of the following loss and risk functions:

$$W(\uptheta, \hat{\theta}) = (-2) \int f(x|\uptheta) \log\left(\frac{f(x|\hat{\theta})}{f(x|\uptheta)}\right) \mathrm{d}x \tag{3.1}$$

$$R(\uptheta, \hat{\theta}) = EW(\uptheta, \hat{\theta}), \tag{3.2}$$

where the expectation in the right-hand side of (3.2) is taken with respect to the distribution of $\hat{\theta}$. As $W(\uptheta, \hat{\theta})$ is equal to 2 times the Kullback-Leibler's information for discrimination in favour of $f(x|\uptheta)$ for $f(x|\hat{\theta})$ it is known that $W(\uptheta, \hat{\theta})$ is a non-negative quantity and is equal to zero if and only if $f(x|\uptheta) = f(x|\hat{\theta})$ almost everywhere [16]. This property is forming a basis of the proof of consistency of the maximum likelihood estimate of $\uptheta$ [24] and indicates the close relationship between the maximum likelihood principle and the information theoretic observations.

When N independent realizations x_i $(i = 1, 2, \ldots, N)$ of X are available, (-2) times the sample mean of the log-likelihood ratio

$$\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{f(x_i|\hat{\theta})}{f(x_i|\theta)}\right) \tag{3.3}$$

will be a consistent estimate of $W(\theta, \hat{\theta})$. Thus it is quite natural to expect that, at least for large N, the value of $\hat{\theta}$ which will give the maximum of (3.3) will nearly minimize $W(\theta, \hat{\theta})$. Fortunately the maximization of (3.3) can be realized without knowing the true value of θ, giving the well-known maximum likelihood estimate $\hat{\theta}$. Though it has been said that the maximum likelihood principle is not based on any clearly defined optimum consideration [18; p. 15] our present observation has made it clear that it is essentially designed to keep minimum the estimated loss function which is very naturally defined as the mean information for discrimination between the estimated and the true distributions.

4. Extension of the Maximum Likelihood Principle

The maximum likelihood principle has mainly been utilized in two different branches of statistical theories. The first is the estimation theory where the method of maximum likelihood has been used extensively and the second is the test theory where the log-likelihood ratio statistic is playing a very important role. Our present definitions of $W(\theta, \hat{\theta})$ and $R(\theta, \hat{\theta})$ suggest that these two problems should be combined into a single problem of statistical decision. Thus instead of considering a single estimate of θ we consider estimates corresponding to various possible restrictions of the distribution and instead of treating the problem as a multiple decision or a test between hypotheses we treat it as a problem of general estimation procedure based on the decision theoretic consideration. This whole idea can be very simply realized by comparing $R(\theta, \hat{\theta})$, or $W(\theta, \hat{\theta})$ if possible, for various $\hat{\theta}$'s and taking the one with the minimum of $R(\theta, \hat{\theta})$ or $W(\theta, \hat{\theta})$ as our final choice. As it was discussed in the introduction this approach may be viewed as a natural extension of the classical maximum likelihood principle. The only problem in applying this extended principle in a practical situation is how to get the reliable estimates of $R(\theta, \hat{\theta})$ or $W(\theta, \hat{\theta})$. As it was noticed in [6] and will be seen shortly, this can be done for a very interesting and practically important situation of composite hypotheses through the use of the maximum likelihood estimates and the corresponding log-likelihood ratio statistics.

The problem of statistical model identification is often formulated as the problem of the selection of $f(x|_k\theta)$ $(k = 0, 1, 2, \ldots, L)$ based on the observations of X, where $_k\theta$ is restricted to the space with $_k\theta_{k+1} = {}_k\theta_{k+2} = \cdots = {}_k\theta_L =$

0. k, or some of its equivalents, is often called the order of the model. Its decision is usually the most difficult problem in practical statistical model identification. The problem has often been treated as a subject of composite hypothesis testing and the use of the log-likelihood ratio criterion is well established for this purpose [23]. We consider the situation where the results x_i ($i = 1, 2, \ldots, N$) of N independent observations of X have been obtained. We denote by ${}_k\hat{\theta}$ the maximum likelihood estimate in the space of ${}_k\theta$, i.e., ${}_k\hat{\theta}$ is the value of ${}_k\theta$ which gives the maximum of the likelihood function $\prod_{i=1}^{N} f(x_i|{}_k\theta)$. The observation at the end of the preceding section strongly suggests the use of

$$ {}_k\omega_L = -\frac{2}{N}\sum_{i=1}^{N}\log\left(\frac{f(x_i|{}_k\hat{\theta})}{f(x_i|{}_L\hat{\theta})}\right) \tag{4.1}$$

as an estimate of $W(\theta, {}_k\hat{\theta})$. The statistics

$$ {}_k\eta_L = N \times {}_k\omega_L \tag{4.2}$$

is the familiar log-likelihood ratio test statistics which will asymptotically be distributed as a chi-square variable with the degrees of freedom equal to $L - k$ when the true parameter θ is in the space of ${}_k\theta$. If we define

$$ W(\theta, {}_k\theta) = \inf_{{}_k\theta} W(\theta, {}_k\theta), \tag{4.3}$$

then it is expected that

$$ {}_k\omega_L \rightarrow W(\theta, {}_k\theta) \text{ w.p.l.}$$

Thus when $NW(\theta, {}_k\theta)$ is significantly larger than L the value of ${}_k\eta_L$ will be very much larger than would be expected from the chi-square approximation. The only situation where a precise analysis of the behaviour of ${}_k\eta_L$ is necessary would be the case where $NW(\theta, {}_k\theta)$ is of comparable order of magnitude with L. When N is very large compared with L this means that $W(\theta, {}_k\theta)$ is very nearly equal to $W(\theta, \theta) = 0$. We shall hereafter assume that $W(\theta, \theta)$ is sufficiently smooth at $\theta = \theta$ and

$$ W(\theta, \theta) > 0 \qquad \text{for} \qquad \theta \neq \theta. \tag{4.4}$$

Also we assume that $W(\theta, {}_k\theta)$ has a unique minimum at ${}_k\theta = {}_k\theta$ and that ${}_L\theta = \theta$. Under these assumptions the maximum likelihood estimates $\hat{\theta}$ and ${}_k\hat{\theta}$ will be consistent estimates of θ and ${}_k\theta$, respectively, and since we are concerned with the situation where θ and ${}_k\theta$ are situated very near to each other, we limit our observation only up to the second-order variation of $W(\theta, {}_k\hat{\theta})$. Thus hereafter we adopt, in place of $W(\theta, {}_k\hat{\theta})$, the loss function

$$ W_2(\theta, {}_k\hat{\theta}) = \sum_{l=1}^{L}\sum_{m=1}^{L} ({}_k\hat{\theta}_l - \theta_l)({}_k\hat{\theta}_m - \theta_m)C(l, m)(\theta), \tag{4.5}$$

where $C(l, m)(\theta)$ is the (l, m)th element of Fisher's information matrix and is given by

$$C(l, m)(\theta) = \int \left(\frac{\partial f_\theta}{\partial \theta_l} \frac{1}{f_\theta}\right)\left(\frac{\partial f_\theta}{\partial \theta_m} \frac{1}{f_\theta}\right) f_\theta \, \mathrm{d}x = -\int \left(\frac{\partial^2 \log f}{\partial \theta_l \partial \theta_m}\right) f_\theta \, \mathrm{d}x. \tag{4.6}$$

We shall simply denote by $C(l, m)$ the value of $C(l, m)(\theta)$ at $\theta = \uptheta$. We denote by $\|\theta\|_c$ the norm in the space of θ defined by

$$\|\theta\|_c^2 = \sum_{i=1}^{L} \sum_{m=1}^{L} \theta_l \theta_m C(l, m). \tag{4.7}$$

We have

$$W_2(\uptheta, {}_k\hat{\theta}) = \|{}_k\hat{\theta} - \uptheta\|_c^2. \tag{4.8}$$

Also we redefine ${}_k\uptheta$ by the relation

$$\|{}_k\uptheta - \uptheta\|_c^2 = \operatorname*{Min}_{{}_k\theta} \|{}_k\theta - \uptheta\|_c^2. \tag{4.9}$$

Thus ${}_k\uptheta$ is the projection of $\uptheta$ in the space of ${}_k\theta$'s with respect to the metrics defined by $C(l, m)$ and is given by the relations

$$\sum_{m=1}^{k} C(l, m)_k\theta_m = \sum_{m=1}^{L} C(l, m)\theta_m \quad l = 1, 2, \ldots, k. \tag{4.10}$$

We get from (4.8) and (4.9)

$$W_2(\uptheta, {}_k\hat{\theta}) = \|{}_k\uptheta - \uptheta\|_c^2 + \|{}_k\hat{\theta} - {}_k\uptheta\|_c^2. \tag{4.11}$$

Since the definition of $W(\uptheta, \hat{\theta})$ strongly suggests, and is actually motivated by, the use of the log-likelihood ratio statistics we will study the possible use of this statistics for the estimation of $W_2(\uptheta, {}_k\hat{\theta})$. Taking into account the relations

$$\begin{aligned} \sum_i \frac{\partial \log f(x_i|\hat{\theta})}{\partial \theta_m} &= 0, \qquad m = 1, 2, \ldots, L, \\ \sum_i \frac{\partial \log f(x_i|{}_k\hat{\theta})}{\partial \theta_m} &= 0, \qquad m = 1, 2, \ldots, k, \end{aligned} \tag{4.12}$$

we get the Taylor expansions

$$\begin{aligned} \sum_{i=1}^{N} \log f(x_i|{}_k\uptheta) &= \sum_{i=1}^{N} \log f(x_i|\hat{\theta}) + \frac{1}{2} \sum_{m=1}^{L} \sum_{l=1}^{L} N({}_k\uptheta_m - \hat{\theta}_m)({}_k\uptheta_l - \hat{\theta}_l) \\ &\quad \times \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \log f(x_i|\hat{\theta} + \varrho({}_k\uptheta - \hat{\theta}))}{\partial \theta_m \partial \theta_l} \\ &= \sum_{i=1}^{N} \log f(x_i|{}_k\hat{\theta}) + \frac{1}{2} \sum_{m=1}^{k} \sum_{l=1}^{k} N({}_k\uptheta_m - {}_k\hat{\theta}_m)({}_k\uptheta_l - {}_k\hat{\theta}_l) \\ &\quad \times \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \log f(x_1|{}_k\hat{\theta} + \varrho_k({}_k\uptheta - {}_k\hat{\theta}))}{\partial \theta_m \partial \theta_l}, \end{aligned}$$

where the parameter values within the functions under the differential sign denote the points where the derivatives are taken and $0 \leq \varrho_k, \varrho \leq 1$, a convention which we use in the rest of this paper. We consider that, in increasing the value of N, N and k are chosen in such a way that $\sqrt{N}({}_k\theta_m - \theta_m)$ $(m = 1, 2, \ldots, L)$ are bounded, or rather tending to a set of constants for the ease of explanation. Under this circumstance, assuming the tendency towards a Gaussian distribution of $\sqrt{N}(\hat{\theta} - \theta)$ and the consistency of ${}_k\hat{\theta}$ and $\hat{\theta}$ as the estimates of ${}_k\theta$ and θ we get, from (4.6) and (4.13), an asymptotic equality in distribution for the log-likelihood ratio statistic ${}_k\eta_L$ of (4.2)

$$ {}_k\eta_L = N\|\hat{\theta} - {}_k\theta\|_c^2 - N\|{}_k\hat{\theta} - {}_k\theta\|_c^2. \tag{4.14}$$

By simple manipulation

$$ {}_k\eta_L = N\|{}_k\theta - \theta\|_c^2 + N\|\hat{\theta} - \theta\|_c^2 - N\|{}_k\hat{\theta} - {}_k\theta\|_c^2 - 2N(\hat{\theta} - \theta, \theta - \theta)_c, \tag{4.15}$$

where $(\,,)_c$ denotes the inner product defined by $C(l, m)$. Assuming the validity of the Taylor expansion up to the second order and taking into account the relations (4.12) we get for $l = 1, 2, \ldots, k$

$$\begin{aligned} &\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial}{\partial\theta_l} \log f(x_i|{}_k\theta) \\ &\quad = \sum_{m=1}^{k} \sqrt{N}({}_k\theta_m - {}_k\hat{\theta}_m) \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \log f(x_i|{}_k\hat{\theta} + \varrho_k({}_k\theta - {}_k\hat{\theta}))}{\partial\theta_m \partial\theta_l} \\ &\quad = \sum_{m=1}^{L} \sqrt{N}({}_k\theta_m - \hat{\theta}_m) \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \log f(x_i|\hat{\theta} + \varrho({}_k\theta - \hat{\theta}))}{\partial\theta_m \partial\theta_l}. \end{aligned} \tag{4.16}$$

Let C^{-1} be the inverse of Fisher's information matrix. Assuming the tendency to the Gaussian distribution $N(0, C^{-1})$ of the distribution of $\sqrt{N}(\hat{\theta} - \theta)$ which can be derived by using the Taylor expansion of the type of (4.16) at $\theta = \theta$, we can see that for N and k with bounded $\sqrt{N}({}_k\theta_m - \theta_m)$ $(m = 1, 2, \ldots, L)$ (4.16) yields, under the smoothness assumption of $C(l, m)(\theta)$ at $\theta = \theta$, the approximate equations

$$\sum_{m=1}^{k} \sqrt{N}({}_k\theta_m - {}_k\hat{\theta}_m) C(l, m) = \sum_{m=1}^{L} \sqrt{N}({}_k\theta_m - \hat{\theta}_m) C(l, m) \quad l = 1, 2, \ldots, k. \tag{4.17}$$

Taking (4.10) into account we get from (4.17), for $l = 1, 2, \ldots, k$,

$$\sum_{m=1}^{k} \sqrt{N}({}_k\theta_m - {}_k\hat{\theta}_m) C(l, m) = \sum_{m=1}^{L} \sqrt{N}(\theta_m - \hat{\theta}_m) C(l, m). \tag{4.18}$$

This shows that geometrically ${}_k\hat{\theta} - {}_k\theta$ is (approximately) the projection of $\hat{\theta} - \theta$ into the space of ${}_k\theta$'s. From this result it can be shown that $N\|\hat{\theta} - \theta\|_c^2 - N\|{}_k\hat{\theta} - {}_k\theta\|_c^2$ and $N\|{}_k\hat{\theta} - {}_k\theta\|_c^2$ are asymptotically independently distributed as chi-square variables with the degrees of freedom $L - k$ and k, respectively. It can also be shown that the standard deviation of the asymptotic distribution of $N(\hat{\theta} - \theta, {}_k\theta - \theta)_c$ is equal to $\sqrt{N}\|{}_k\theta - \theta\|_c$. Thus if $N\|{}_k\theta - \theta\|_c^2$ is of comparable magnitude with $L - k$ or k and these are large integers then the contribution of the last term in the right hand side of (4.15) remains relatively insignificant. If $N\|{}_k\theta - \theta\|_c^2$ is significantly larger than L the contribution of $N(\hat{\theta} - \theta, {}_k\theta - \theta)_c$ to ${}_k\eta_L$ will also relatively be insignificant. If $N\|{}_k\theta - \theta\|_c^2$ is significantly smaller than L and k again the contribution of $N(\hat{\theta} - \theta, {}_k\theta - \theta)_c$ will remain insignificant compared with those of other variables of chi-square type. These observations suggest that from (4.11), though $N^{-1}{}_k\eta_L$ may not be a good estimate of $W_2(\theta, {}_k\hat{\theta})$,

$$r(\hat{\theta}, {}_k\hat{\theta}) = N^{-1}({}_k\eta_L + 2k - L) \tag{4.19}$$

will serve as a useful estimate of $EW_2(\theta, {}_k\hat{\theta})$, at least for the case where N is sufficiently large and L and k are relatively large integers.

It is interesting to note that in practical applications it may sometimes happen that L is a very large, or conceptually infinite, integer and may not be defined clearly. Even under such circumstances we can realize our selection procedure of ${}_k\hat{\theta}$'s for some limited number of k's, assuming L to be equal to the largest value of k. Since we are only concerned with finding out the ${}_k\hat{\theta}$ which will give the minimum of $r(\hat{\theta}, {}_k\hat{\theta})$ we have only to compute either

$${}_k\nu_L = {}_k\eta_L + 2k \tag{4.20}$$

or

$${}_k\lambda_L = -2\sum_{i=1}^{N} \log f(x_i|{}_k\hat{\theta}) + 2k. \tag{4.21}$$

and adopt the ${}_k\hat{\theta}$ which gives the minimum of ${}_k\nu_L$ or ${}_k\lambda_L$ $(0 \leq k \leq L)$. The statistical behaviour of ${}_k\lambda_L$ is well understood by taking into consideration the successive decomposition of the chi-square variables into mutually independent components. In using ${}_k\lambda_L$ care should be taken not to lose significant digits during the computation.

5. Applications

Some of the possible applications will be mentioned here.

1. Factor Analysis

In the factor analysis we try to find the best estimate of the variance covariance matrix Σ from the sample variance covariance matrix using the model $\Sigma = AA' + D$, where Σ is a $p \times p$ dimensional matrix, A is a $p \times m$ dimensional $(m < p)$ matrix and D is a non-negative $p \times p$ diagonal matrix. The method of the maximum likelihood estimate under the assumption of normality has been extensively applied and the use of the log-likelihood ratio criterion is quite common. Thus our present procedure can readily be incorporated to help the decision of m. Some numerical examples are already given in [6] and the results are quite promising.

2. Principal Component Analysis

By assuming $D = \delta I (\delta \geq 0, I$; unit matrix) in the above model, we can get the necessary decision procedure for the principal component analysis.

3. Analysis of Variance

If in the analysis of variance model we can preassign the order in decomposing the total variance into chi-square components corresponding to some factors and interactions then we can easily apply our present procedure to decide where to stop the decomposition.

4. Multiple Regression

The situation is the same as in the case of the analysis of variance. We can make a decision where to stop including the independent variables when the order of variables for inclusion is predetermined. It can be shown that under the assumption of normality of the residual variable we have only to compare the values $s^2(k)\left(1 + \frac{2k}{N}\right)$, where $s^2(k)$ is the sample mean square of the residual after fitting the regression coefficients by the method of least squares where k is the number of fitted regression coefficients and N the sample size. k should be kept small compared with N. It is interesting to note that the use of a statistics proposed by Mallows [13] is essentially equivalent to our present approach.

5. Autoregressive Model Fitting in Time Series

Though the discussion in the present paper has been limited to the realizations of independent and identically distributed random variables, by following the approach of Billingsley [8], we can see that the same line of discussion can be extended to cover the case of finite parameter Markov processes. Thus in the case of the fitting of one-dimensional autoregressive model $X_n = \sum_{m=1}^{k} a_m X_{n-m} + \varepsilon_n$ we have, assuming the normality of the process X_n, only to adopt k which gives the minimum of $s^2(k)\left(1 + \frac{2k}{N}\right)$ or equivalently $s^2(k)\left(1 + \frac{k}{N}\right)\left(1 - \frac{k}{N}\right)^{-1}$, where $s^2(k)$ is the sample mean square of the residual after fitting the kth order model by the method of least squares or some of its equivalents. This last quantity for the decision has been first introduced by the present author and was considered to be an estimate of the quantity called the final prediction error (FPE) [1, 2]. The use of this approach for the estimation of power spectra has been discussed and recognized to be very useful [3]. For the case of the multi-dimensional process we have to replace $s^2(k)$ by the sample generalized variance or the determinant of the sample variance-covariance matrix of residuals. The procedure has been extensively used for the identification of a cement rotary kiln model [4, 5, 19].

These procedures have been originally derived under the assumption of linear process, which is slightly weaker than the assumption of normality, and with the intuitive criterion of the expected variance of the final one step prediction (FPE). Our present observation shows that these procedures are just in accordance with our extended maximum likelihood principle at least under the Gaussian assumption.

6. Numerical Examples

To illustrate the difference between the conventional test procedure and our present procedure, two numerical examples are given using published data.

The first example is taken from the book by Jenkins and Watts [14]. The original data are described as observations of yield from 70 consecutive batches of an industrial process [14, p. 142]. Our estimates of FPE are given in Table 1 in a relative scale. The results very simply suggest, without the help of statistical tables, the adoption of $k = 2$ for this case. The same conclusion has been reached by the authors of the book after a detailed analysis of significance of partial autocorrelation coefficients and by relying on a somewhat subjective judgement [14, pp. 199–200]. The fitted model produced an estimate of the power spectrum which is very much like their final choice obtained by using Blackman-Tukey type window [14, p. 292].

The next example is taken from a paper by Whittle on the analysis of a seiche record (oscillation of water level in a rock channel) [26; 27, pp. 37–38]. For this example Whittle has used the log-likelihood ratio test statistics in successively deciding the significance of increasing the order by one and adopted $k = 4$. He reports that the fitting of the power spectrum is very poor. Our procedure applied to the reported sample autocorrelation coefficients obtained from data with $N = 660$ produced a result showing that $k = 65$ should be adopted within the k's in the range $0 \leq k \leq 66$. The estimates of

Table 1. Autoregressive Model Fitting.

k	0	1	2	3	4	5	6	7
FPE_k^*	1.029	0.899	0.895	0.921	0.946	0.097	0.983	1.012

$* \ \text{FPE}_k = s^2(k)\left(1 + \frac{k+1}{N}\right)\left(1 - \frac{k+1}{N}\right)^{-1} \Big/ s^2(0)$

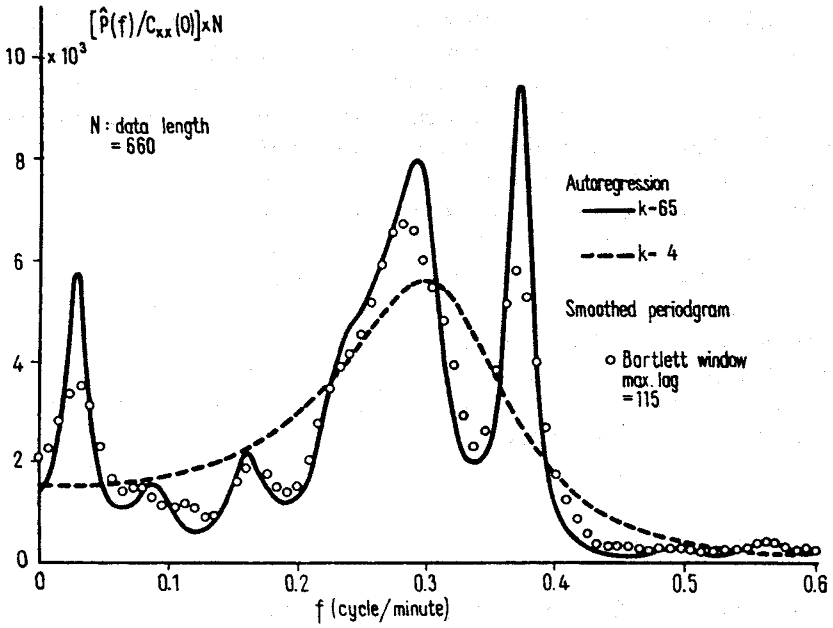

Figure 1. Estimates of the seiche spectrum. The smoothed periodgram of $x(n\,\Delta t)$ $(n = 1, 2, \ldots, N)$ is defined by

$$\Delta t \cdot \sum_{l}^{l} \left(1 - \frac{|s|}{l}\right) C_{xx}(s) \cos(2\pi f s\,\Delta t),$$

$$\text{where } l = \text{max. lag, } C_{xx}(s) = \frac{1}{N} \sum_{n=1}^{N-|s|} \tilde{x}(|s| + n)\tilde{x}(n),$$

$$\text{where } \tilde{x}(n) = x(n\,\Delta t) - \bar{x} \quad \text{and} \quad \bar{x} = \frac{1}{N} \sum_{n=1}^{N} x(n\,\Delta t).$$

the power spectrum are illustrated in Fig. 1. Our procedure suggests that $L = 66$ is not large enough, yet it produced very sharp line-like spectra at various frequencies as was expected from the physical consideration, while the fourth order model did not give any indication of them. This example dramatically illustrates the impracticality of the conventional successive test procedure depending on a subjectively chosen set of levels of significance.

7. Concluding Remarks

In spite of the early statement by Wiener [28; p. 76] that entropy; the Shannon-Wiener type definition of the amount of information, could replace Fisher's definition [11] the use of the information theoretic concepts in the statistical circle has been quite limited [10, 12, 20]; The distinction between Shannon-Wiener's entropy and Fisher's information was discussed as early as in 1950 by Bartlett [7], where the use of the Kullback-Leibler type definition of information was implicit. Since then in the theory of statistics Kullback-Leibler's or Fisher's information could not enjoy the prominent status of Shannon's entropy in communication theory, which proved its essential meaning through the source coding theorem [22, p. 28].

The analysis in the present paper shows that the information theoretic consideration can provide a foundation of the classical maximum likelihood principle and extremely widen its practical applicability. This shows that the notion of informations, which is more closely related to the mutual information in communication theory than to the entropy, will play the most fundamental role in the future developments of statistical theories and techniques.

By our present principle, the extensions of applications 3) ~ 5) of Section 5 to include the comparisons of every possible kth order models are straightforward. The analysis of the overall statistical characteristics of such extensions will be a subject of further study.

Acknowledgement

The author would like to express his thanks to Prof. T. Sugiyama of Kawasaki Medical University for helpful discussions of the possible applications

References

1. Akaike, H., Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21** (1969) 243–217.
2. Akaike., H., Statistical predictor identification. *Ann. Inst. Statist. Math.* **22** (1970) 203–217.
3. Akaike, H., On a semi-automatic power spectrum estimation procedure. *Proc. 3rd Hawaii International Conference on System Sciences,* 1970, 974–977.
4. Akaike, H., On a decision procedure for system identification, Preprints, *IFAC Kyoto Symposium on System Engineering Approach to Computer Control.* 1970, 486–490.
5. Akaike, H., Autoregressive model fitting for control. *Ann. Inst. Statist. Math.* 23 (1971) 163–180.
6. Akaike, H., Determination of the number of factors by an extended maximum likelihood principle. Research Memo. 44, Inst. Statist. Math. March, 1971.
7. Bartlett, M. S., The statistical approach to the analysis of time-series. *Symposium on Information Theory* (mimeographed Proceedings), Ministry of Supply, London, 1950, 81–101.
8. Billingsley, P., *Statistical Inference for Markov Processes.* Univ. Chicago Press, Chicago 1961.
9. Blackwell, D., Equivalent comparisons of experiments. *Ann. Math. Statist.* 24 (1953) 265–272.
10. Campbell, L.L., Equivalence of Gauss's principle and minimum discrimination information estimation of probabilities. *Ann. Math. Statist.* 41 (1970) 1011–1015.

11. Fisher, R.A., Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22** (1925) 700–725, *Contributions to Mathematical Statistics.* John Wiley & Sons, New York, 1950, paper 11.
12. Good, I.J. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.* **34** (1963) 911–934.
13. Gorman, J.W. and Toman, R.J., Selection of variables for fitting equations to data. *Technometrics* **8** (1966) 27–51.
14. Jenkins, G.M. and Watts, D.G., *Spectral Analysis and Its Applications.* Holden Day, San Francisco, 1968.
15. Kullback, S. and Leibler, R.A., On information and sufficiency. *Ann. Math Statist.* **22** (1951) 79–86.
16. Kullback, S., *Information Theory and Statistics.* John Wiley & Sons, New York 1959.
17. Le Cam, L., On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. Calif. Publ. in Stat.* 1 (1953) 277–330.
18. Lehmann, E.L., Testing Statistical Hypotheses. John Wiley & Sons, New York 1969.
19. Otomo, T., Nakagawa, T. and Akaike, H. Statistical approach to computer control of cement rotary kilns. 1971. *Automatica* **8** (1972) 35–48.
20. Rėnyi, A., Statistics and information theory. *Studia Sci. Math. Hung.* **2** (1967) 249–256.
21. Savage, L.J., The Foundations of Statistics. John Wiley & Sons, New York 1954.
22. Shannon, C.E. and Weaver, W., *The Mathematical Theory of Communication.* Univ. of Illinois Press, Urbana 1949.
23. Wald, A., Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.* **54** (1943) 426–482.
24. Wald, A., Note on the consistency of the maximum likelihood estimate. *Ann Math. Statist.* 20 (1949) 595–601.
25. Wald, A., Statistical Decision Functions. John Wiley & Sons, New York 1950.
26. Whittle, P., The statistical analysis of seiche record. *J. Marine Res.* **13** (1954) 76–100.
27. Whittle, P., *Prediction and Regulation.* English Univ. Press, London 1963.
28. Wiener, N., *Cybernetics.* John Wiley & Sons, New York, 1948.

Source:
Akaike, H. 1992. Information Theory and an Extension of the Maximum Likelihood Principle. Pages 610-624 *In:* Kotz, S., and N.L. Johnson, editors. *Breakthroughs in Statistics Volume 1. Foundations and Basic Theory.* Springer Series in Statistics, Perspectives in Statistics. Springer-Verlag: New York.