Information Theory And An Extension Of The Maximum Likelihood Principle By Hirotogu Akaike

Article · March 1994

Source: CiteSeer

CITATIONS
319

READS
3,586

1 author:

Jan De Leeuw
University of California, Los Angeles
283 PUBLICATIONS **10,294** CITATIONS

SEE PROFILE

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.*

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

It follows from the inequality $\ln t > 1 + t$ 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 $\mathfrak{R}^m$, 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 $0 \leq \rho \leq 1$. Subtracting the two expansions gives

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

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

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

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.

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

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

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

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 sys-

tematic 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.