Reprinted from *Breakthroughs in Statistics, Vol.I, Foundations and Basic Theory*, S. Kotz and N.L.Johnson, eds., Springer-Verlag, New York, 1992, 610-624 by permission of Akademia Kiado and Springer-Verlag, New York. (Originally published in Proceeding of the Second International Symposium on Information Theory, B.N. Petrov and F. Caski, eds., Akademiai Kiado, Budapest, 1973, 267-281)

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

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

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. Parzen et al. (eds.), *Selected Papers of Hirotugu Akaike*

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

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 $\theta = \theta = (\theta_1, \theta_2, \ldots, \theta_L)'$. We Will denote by $I(\theta, \theta; \Phi)$ the quantity defined by (2.1) with $f_1(x) = f(x|\theta)$ and $f_0(x) = f(x|\theta)$ and analyze the sensitivity of $I(\theta, \theta; \Phi)$ to the deviation of θ from θ. Assuming the regularity conditions of $f(x|\theta)$ and $\Phi(r)$ which assure the following analytical treatment we get

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

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

where r, $\dot{\Phi}(1)$, $\ddot{\Phi}(1)$ and f_θ denote $\dfrac{f(x|\theta)}{f(x|\theta)}$, $\left.\dfrac{d\Phi(r)}{dr}\right|_{r=1}$, $\left.\dfrac{d^2\Phi(r)}{dr^2}\right|_{r=1}$ 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 $\int f(x|\theta)\,dx = 1$, we have

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

Thus we get

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

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

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

These relations show that $\ddot{\Phi}(1)$ must be different from zero if $I(\theta, \theta; \Phi)$ ought to be sensitive to the small variations of θ. Also it is clear that the relative sensitivity of $I(\theta, \theta; \Phi)$ 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 $I(\theta, \theta; \Phi)$ for small variations of θ around θ. 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

$$I_N(\theta, \theta; \Phi) = \int \Phi\left(\frac{\prod_{i=1}^{N} f(x_i|\theta)}{\prod_{i=1}^{N} f(x_i|\theta)}\right) \prod_{i=1}^{N} f(x_i|\theta)\,dx_1 \ldots dx_N. \tag{2.9}$$

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

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

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

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

These equations show that $I_N(\theta, \theta; \Phi)$ is not responsive to the increase of

information and that $\frac{\partial^2}{\partial\theta_l\partial\theta_m} I_N(\theta, \theta; \Phi)|_{\theta=\theta}$ is in a linear relation with N. It can be seen that only the quantity defined by

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

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 θ.

3. Information and the Maximum Likelihood Principle

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

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

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

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

$$ {}_k\omega_L \to W(\theta, {}_k\theta) \text{ w.p.1.} $$

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

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

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

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

We have

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

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

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

Thus ${}_k\theta$ is the projection of θ 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)

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

Since the definition of $W(\theta, \hat{\theta})$ 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 $W_2(\theta, {}_k\hat{\theta})$. 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

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

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 conven-

tion 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

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

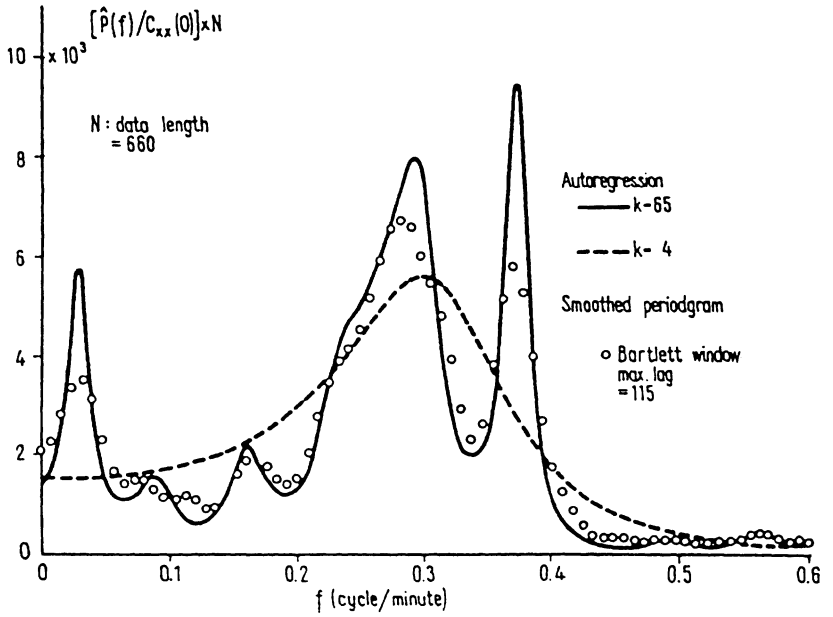

Figure 1. Estimates of the seiche spectrum. The smoothed periodgram of $x(n\,\Delta t)$ $(n = 1, 2, \ldots, N)$ is defined by

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

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

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

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) $\sim$ 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 multi-dimensional 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. Rényi, 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.