PSYCHOMETRIKA—VOL. 52, NO. 3, 345–370
SEPTEMBER 1987
SPECIAL SECTION

MODEL SELECTION AND AKAIKE'S INFORMATION CRITERION (AIC): THE GENERAL THEORY AND ITS ANALYTICAL EXTENSIONS

HAMPARSUM BOZDOGAN

UNIVERSITY OF VIRGINIA

During the last fifteen years, Akaike's entropy-based Information Criterion (AIC) has had a fundamental impact in statistical model evaluation problems. This paper studies the general theory of the AIC procedure and provides its analytical extensions in two ways without violating Akaike's main principles. These extensions make AIC asymptotically consistent and penalize overparameterization more stringently to pick only the simplest of the "true" models. These selection criteria are called CAIC and CAICF. Asymptotic properties of AIC and its extensions are investigated, and empirical performances of these criteria are studied in choosing the correct degree of a polynomial model in two different Monte Carlo experiments under different conditions.

Key words: model selection, Akaike's information criterion, AIC, CAIC, CAICF, asymptotic properties.

1. Introduction and Purpose

During the last fifteen years, Akaike's (1973) entropic information criterion, which is known as AIC, has had a fundamental impact in statistical model evaluation problems. The introduction of AIC furthered the recognition of the importance of good modeling in statistics. As a result, many important statistical modeling techniques have been developed in various fields of statistics, control theory, econometrics, engineering, psychometrics, and in many other fields.

Despite the accumulation of many successful results obtained by AIC, and despite its extreme popularity and growing school of adherents, the entropic criterion AIC has been almost universally accepted in some areas of statistics, while in other areas it is still not known or well understood.

Since most of the material on AIC is scattered in a wide range of journals and proceedings, and Akaike's (1973) original paper is not readily available, the main purpose of this paper is to study the general theory of his imaginative work, discuss its meaning and the basic philosophy, introduce its analytical extensions using the standard results established in mathematical statistics without violating Akaike's basic principles, and show their asymptotic properties and give the results on their inferential error rates.

The author extends his deep appreciation to many people. These include Hirotugu Akaike, Donald E. Ramirez, Marvin Rosenblum, and S. James Taylor for reading and commenting on some parts of this manuscript through various stages of its development. I especially wish to thank Yoshio Takane, Jim Ramsay, and Stanley L. Sclove for critically reading the paper and making many helpful suggestions. I also wish to thank Julie Riddleberger for her excellent typing of this manuscript.

This research was partially supported by NIH Biomedical Research Support Grant (BRSG) No. 5-24867 at the University of Virginia.

Requests for reprints should be sent to the author at the Department of Mathematics, Math-Astronomy Building, University of Virginia, Charlottesville, VA 22903.

0033-3123/87/0900-SS03$00.75/0

As is well known, a fundamental difficulty in statistical analysis is the choice of an appropriate model, estimating and determining the order or dimension of a model. This is a common problem when a statistical model contains many parameters. The main purpose of model evaluation is to "understand" the observed data. According to Parzen (1982), *statistical data modeling* is a field of statistical reasoning that seeks to fit models to data without knowing what the "true" model is or might be.

Consequently, one seeks to learn the model and study the quality of the model by a process which is called *statistical model identification or evaluation.* In recent years, in the literature, the necessity of introducing the concept of *model selection* or *model evaluation* has been recognized and the problem is posed how to choose the "best approximating" model among a class of competing models with different numbers of parameters by a suitable model selection criterion given a data set. Also, there is presently a great deal of interest in simple criteria represented by parsimony of parameters for choosing one of a set of competing models to describe a given data set. As discussed in Stone (1981), parsimony can take into account a variety of attributes of the selected model. One such attribute is the cost of measuring the models required to implement the model. Measurement cost, which was emphasized by Lindley (1968), is especially relevant in certain applications. A second attribute is the complexity of the selected model. The general principle is that for a given level of accuracy, a simpler or a more parsimonious model is preferable to a more complex one, known as *Occam's Razor.* Occam's Razor emphasizes the desirability of selecting the accurate and parsimonious models of reality. Therefore, the best model is the one with least complexity, or equivalently the highest information gain. For example, in factor-analytic models, parameter parsimony requires that we choose the smallest number of factors such that the corresponding model fits the data. The selection of a parsimonious model, in general, is a nontrivial problem without the aid of model selection criteria. They are called "figures of merit" for competing models. That model, which optimizes the criterion, is chosen to be the best model.

Akaike, in a very important sequence of papers, including Akaike (1973, 1974, 1977, and 1981a), was perhaps one of the first who laid the foundation of the modern field of statistical data modeling, statistical model identification or evaluation. He developed the information-theoretic, or entropic AIC criterion for the identification of an optimal and a parsimonious model in data analysis from a class of competing models which takes model complexity into account. The AIC criterion is a simple and versatile procedure which can be viewed as a relatively logical and predictable expression of earlier work of Neyman and Pearson (1928, 1933), Wald (1943), and Kullback (1959), among others. It has enormous practical importance and is one more demonstration of the importance of the likelihood ratio criterion in statistical inference.

In the next section, section 2, we give the necessary theoretical background needed for the development of AIC.

2. Information Quantity as a Measure of Goodness of Fit

2.1 Kullback-Leibler Information Quantity or Negentropy

In any statistical problem we are given a set of observations. These observations are the values of some random variables whose probability distribution is usually unknown to us, or we have some knowledge of it. From the information provided by the data, we draw inferences about the unknown aspects of the underlying distribution, such as the unknown "true" parameter values of the distribution which govern the generation of the observed data and also govern the generation of any future observations if we adopt the predictive point of view.

We shall express a model in the form of a probability distribution and regard fitting a model to the data as estimating the true probability distribution from the data and treat the estimation and the evaluation of a model together as one entity rather than separating them. In the statistical literature during the past fifty years, there has been a meaningless separation of *estimation* and *testing* which did not help the development of a practical and successful statistical model selection and evaluation procedure (Akaike, 1974).

If we had an objective measure (a metric) of the distance between the model and the true distribution, a good inference procedure ought to make this distance as small as possible. One measure of this type is *Boltzmann's* (1877) *generalized entropy*, or the *negentropy* which is also known as the *Kullback-Leibler* (1951) *information quantity*. Hereafter, we refer to this as K-L information quantity for brevity.

Even though the development of AIC has its origins in time series modeling where its practical utility has been thoroughly studied. However, the major development of AIC lies in the direct extension of an entropic or information-theoretic interpretation of the method of maximum likelihood. Its introduction is based on the entropy maximization principle, or minimizing its negative; it is based on the minimization of the K-L information quantity.

To develop this point further, suppose $\mathbf{X}$ is an absolutely continuous random vector characterized by a probability density function $f(\mathbf{x} \mid \boldsymbol{\theta})$ which is known apart from the K-dimensional parameter vector $\boldsymbol{\theta} = \boldsymbol{\theta}_K = (\theta_1, \theta_2, \ldots, \theta_K)$, $\boldsymbol{\theta}_K \in \mathbb{R}^K$. Assume that there exists a true parameter vector $\boldsymbol{\theta}^*$ of $\boldsymbol{\theta}$ with its probability density denoted by $f(\mathbf{x} \mid \boldsymbol{\theta}^*)$. Within this setup, it is required that we select $\boldsymbol{\theta}$ "closest" to the true parameter vector $\boldsymbol{\theta}^*$. Thus, we will measure the "closeness," or the "goodness-of-fit," of $f(\mathbf{x} \mid \boldsymbol{\theta}^*)$ with respect to $f(\mathbf{x} \mid \boldsymbol{\theta})$ by the generalized entropy B of Boltzmann (1877), or K-L information quantity I:

$$B(\boldsymbol{\theta}^*; \boldsymbol{\theta}) = -I(\boldsymbol{\theta}^*; \boldsymbol{\theta}). \tag{1}$$

This is defined by

$$\begin{aligned} B(\boldsymbol{\theta}^*; \boldsymbol{\theta}) &= E[\log f(\mathbf{X} \mid \boldsymbol{\theta}) - \log f(\mathbf{X} \mid \boldsymbol{\theta}^*)] \\ &= \int f(\mathbf{x} \mid \boldsymbol{\theta}^*) \log f(\mathbf{x} \mid \boldsymbol{\theta})\, d\mathbf{x} - \int f(\mathbf{x} \mid \boldsymbol{\theta}^*) \log f(\mathbf{x} \mid \boldsymbol{\theta}^*)\, d\mathbf{x} \\ &= H(\boldsymbol{\theta}^*; \boldsymbol{\theta}) - H(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*), \end{aligned} \tag{2}$$

where E denotes the expectation with respect to the true distribution $f(\mathbf{x} \mid \boldsymbol{\theta}^*)$ of $\mathbf{x}$, $H(\boldsymbol{\theta}^*; \boldsymbol{\theta}) = \int f(\mathbf{x} \mid \boldsymbol{\theta}^*) \log f(\mathbf{x} \mid \boldsymbol{\theta})\, d\mathbf{x}$ is the *cross-entropy* which determines the goodness of fit of $f(\mathbf{x} \mid \boldsymbol{\theta})$ to $f(\mathbf{x} \mid \boldsymbol{\theta}^*)$, $H(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*) \equiv H(\boldsymbol{\theta}^*)$ the usual Shannon negative entropy which is constant for a given $f(\mathbf{x} \mid \boldsymbol{\theta}^*)$, and where "log" means natural logarithm.

Instead of maximizing the entropy criterion (2), we minimize the K-L information quantity:

$$\begin{aligned} I(\boldsymbol{\theta}^*; \boldsymbol{\theta}) &= -B(\boldsymbol{\theta}^*; \boldsymbol{\theta}) \\ &= H(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*) - H(\boldsymbol{\theta}^*; \boldsymbol{\theta}). \end{aligned} \tag{3}$$

Since $H(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*) \equiv H(\boldsymbol{\theta}^*)$ is a constant in both (2) and (3), we only have to estimate the cross-entropy or the *expected log likelihood*

$$\begin{aligned} H(\boldsymbol{\theta}^*; \boldsymbol{\theta}) &= E[\log f(\mathbf{X} \mid \boldsymbol{\theta})] \\ &= \int f(\mathbf{x} \mid \boldsymbol{\theta}^*) \log f(\mathbf{x} \mid \boldsymbol{\theta})\, d\mathbf{x}. \end{aligned} \tag{4}$$

Following Wilks (1962, p. 408), if we assume that $f(\mathbf{x} \mid \boldsymbol{\theta})$ is regular with respect to its first

and second partial derivatives for $\boldsymbol{\theta} \in \mathbb{R}^K$, then, under these conditions, $H(\boldsymbol{\theta}^*; \boldsymbol{\theta})$ in (4) can be differentiated twice under the integral sign with respect to $\boldsymbol{\theta}$ and evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, yielding

$$\begin{aligned} H'(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*) &= 0, \\ H''(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*) &= -J(\boldsymbol{\theta}^*), \end{aligned} \tag{5}$$

where $J(\boldsymbol{\theta}^*)$ is the Fisher's (1922) amount of information pertaining to $\boldsymbol{\theta}^*$ per observation from $f(\mathbf{x} \mid \boldsymbol{\theta}^*)$. From (5), we note that Fisher's information per observation is the second derivative of the K-L information quantity. To put it another way, $J(\boldsymbol{\theta}^*)$ essentially measures the curvature of the expected log likelihood $H(\boldsymbol{\theta}^*; \boldsymbol{\theta})$ at its maximum value which occurs at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. The quantity $H(\boldsymbol{\theta}^*; \boldsymbol{\theta})$ plays a crucial role in the development of AIC, and is of basic importance in statistical information theory.

The analytic properties of $I(\boldsymbol{\theta}^*; \boldsymbol{\theta})$ are extensively discussed by Kullback (1959). Here we list some of the important ones.

(i) $I(\boldsymbol{\theta}^*; \boldsymbol{\theta}) > 0$ whenever $f(\mathbf{x} \mid \boldsymbol{\theta}^*) \neq f(\mathbf{x} \mid \boldsymbol{\theta})$,

(ii) $I(\boldsymbol{\theta}^*; \boldsymbol{\theta}) = 0$, if and only if $f(\mathbf{x} \mid \boldsymbol{\theta}^*) = f(\mathbf{x} \mid \boldsymbol{\theta})$ a.e. (almost everywhere) in the possible range of $\mathbf{x}$, when the model is essentially true,

(iii) if $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables, then the K-L information quantity for the whole sample is

$$I_n(\boldsymbol{\theta}^*; \boldsymbol{\theta}) = nI(\boldsymbol{\theta}^*; \boldsymbol{\theta}).$$

This last property says that if the random variables are independent, the K-L information quantity is additive.

We note that K-L information quantity is perhaps the most general of all information measures in the sense of being derivable from minimal assumptions and it represents a relative measure.

As it stands, the K-L information quantity in (3) is not directly observable or estimable. To see this, we give the following simple example as an illustration.

Example 2.1.1. Let F be the family of normal distributions $\{N(\mu, \sigma^2): -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$. Let $f(x \mid \boldsymbol{\theta}^*) \sim N(\mu^*, \sigma^{*2})$ be the true distribution with the parameter vector $\boldsymbol{\theta}^* = (\mu^*, \sigma^{*2})$, and let $f(x \mid \boldsymbol{\theta}) \sim N(\xi, \sigma^2)$ be our model with $\boldsymbol{\theta} = (\xi, \sigma^2)$. Computing $I(\boldsymbol{\theta}^*; \boldsymbol{\theta}) \equiv I[f(x \mid \boldsymbol{\theta}^*); f(x \mid \boldsymbol{\theta})]$, we obtain

$$I(\boldsymbol{\theta}^*; \boldsymbol{\theta}) = \frac{1}{2} \log\left(\frac{\sigma^{*2}}{\sigma^2}\right) + \frac{1}{2}\left[\frac{\sigma^2}{\sigma^{*2}} - 1 + \frac{(\mu^* - \xi)^2}{\sigma^{*2}}\right]. \tag{6}$$

We see that the distance between the means contributes to the K-L information quantity quadratically; while the contribution of the variances is through the ratio σ^{*2}/σ^2. Further, we note from this example, and in general, that the K-L information quantity needs to be estimated from the observed data since it depends on the true distribution, and consequently on the unknown true and model parameters. The important question we need to answer is: How can we make the K-L information quantity operationalized, or approximated, from the observed data so that we can use it to compare the goodness of fit of various models, measure the distance or deviation of the fitted model from the "true" model?

2.2 Mean Log Likelihood as an Estimate of K-L Information Quantity or Negentropy

In this section we introduce the concept of *mean log likelihood* as a measure for the goodness of fit of a model and state entropy maximization principle (EMP) according to Akaike (1977).

Suppose that the generation of data is described by a model given by a probability density function $f(\mathbf{x} \mid \boldsymbol{\theta})$. Given n independent observations from the same distribution regarded as a function of a vector-valued parameter, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$, $k = 1, 2, \ldots, K$, the likelihood function for the set of data is

$$L(\boldsymbol{\theta}) = f(x_1, \ldots, x_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}). \tag{7}$$

The log likelihood function, $\ell(\boldsymbol{\theta})$ (often called the *support*), is the natural logarithm of $L(\boldsymbol{\theta})$ and is defined by

$$\ell(\boldsymbol{\theta}) \equiv \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i \mid \boldsymbol{\theta}), \tag{8}$$

regarded as a random variable, is the sum of i.i.d. random variables $\log f(x_i \mid \boldsymbol{\theta})$, $i = 1, 2, \ldots, n$.

We define the average or mean log likelihood of the sample by

$$\begin{aligned} \frac{1}{n}\ell(\boldsymbol{\theta}) \equiv \frac{1}{n}\log L(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \log f(x_i \mid \boldsymbol{\theta}) \\ &= \ell_n(\boldsymbol{\theta}), \end{aligned} \tag{9}$$

which can be interpreted as an estimator of the "distance" between the true probability density $f(\mathbf{x} \mid \boldsymbol{\theta}^*)$ and the model $f(\mathbf{x} \mid \boldsymbol{\theta})$.

As we discussed in the previous section, the K-L information quantity is not observable. However, it can be consistently estimated from the observed data and operationalized.

Let $\tilde{I}(\boldsymbol{\theta}^*; \boldsymbol{\theta})$ denote an estimator of the K-L information quantity $I(\boldsymbol{\theta}^*; \boldsymbol{\theta})$. Then (3) becomes

$$\tilde{I}(\boldsymbol{\theta}^*; \boldsymbol{\theta}) = \tilde{H}(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*) - \tilde{H}(\boldsymbol{\theta}^*; \boldsymbol{\theta}), \tag{10}$$

or

$$\tilde{H}(\boldsymbol{\theta}^*; \boldsymbol{\theta}) = -\tilde{I}(\boldsymbol{\theta}^*; \boldsymbol{\theta}) + \tilde{H}(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*). \tag{11}$$

This tells us that maximizing the expected log likelihood $\tilde{H}(\boldsymbol{\theta}^*; \boldsymbol{\theta})$ is asymptotically equivalent to minimizing the K-L information quantity, $\tilde{I}(\boldsymbol{\theta}^*; \boldsymbol{\theta})$, and it is not necessary to know $\tilde{H}(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*) \equiv \tilde{H}(\boldsymbol{\theta}^*)$, since it is an additive constant and can be dropped. However, we will retain it for the clarity of our exposition.

Assuming that a sample of n observations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is used to provide an estimate $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{x})$ of $\boldsymbol{\theta}$, we observe that the mean log likelihood in (9) is a natural estimator of $\tilde{H}(\boldsymbol{\theta}^*; \boldsymbol{\theta})$, the expected log likelihood. That is,

$$z(\boldsymbol{\theta}) = E\left[\frac{1}{n}\ell(\boldsymbol{\theta})\right] = \tilde{H}(\boldsymbol{\theta}^*; \boldsymbol{\theta}) = E[\log f(\mathbf{X} \mid \boldsymbol{\theta})], \tag{12}$$

where again the expectation is taken relative to the true distribution $f(\mathbf{x} \mid \boldsymbol{\theta}^*)$ of $\mathbf{x}$, and that the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ is a natural estimator for $\boldsymbol{\theta}^*$, so (10) can then

be consistently estimated by

$$\tilde{I}(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}) = \tilde{H}(\boldsymbol{\theta}^*) - \frac{1}{n}\sum_{i=1}^{n} \log f(x_i \mid \hat{\boldsymbol{\theta}})$$

$$= \tilde{H}(\boldsymbol{\theta}^*) - \frac{1}{n}\ell(\hat{\boldsymbol{\theta}}), \tag{13}$$

or

$$\frac{1}{n}\ell(\hat{\boldsymbol{\theta}}) = \tilde{I}(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}) + \tilde{H}(\boldsymbol{\theta}^*). \tag{14}$$

Certainly, one approach to measure how well the maximum likelihood model $f(x_i \mid \hat{\boldsymbol{\theta}})$ "matches" the data would be to test the hypothesis that the K-L information quantity $I(\boldsymbol{\theta}^*; \boldsymbol{\theta}) = 0$. Such a test might be based on $(n)^{1/2}\tilde{I}(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}})$, but establishing the asymptotic closed form expression of the distribution of this statistic is a nontrivial problem (White, 1982). It is for this reason that we need to appeal to the asymptotic behavior of the mean log likelihood and also to asymptotic approximations to derive AIC.

From Property (ii) in section 2.1, we have seen that $I(\boldsymbol{\theta}^*; \boldsymbol{\theta}) = 0$, if and only if, $f(\mathbf{x} \mid \boldsymbol{\theta}^*) = f(\mathbf{x} \mid \boldsymbol{\theta})$ a.e. in the possible range of $\mathbf{x}$, when the model is essentially true. Hence, from (9), we note that asymptotically the maximum of $\ell_n(\boldsymbol{\theta}) = 1/n \log L(\boldsymbol{\theta})$ (the mean log likelihood) will be $H(\boldsymbol{\theta}^*) \equiv \int f(\mathbf{x} \mid \boldsymbol{\theta}^*) \log f(\mathbf{x} \mid \boldsymbol{\theta}^*)\, d\mathbf{x}$, the negative Shannon entropy, and this will be attained when $f(\mathbf{x} \mid \boldsymbol{\theta}^*) = f(\mathbf{x} \mid \boldsymbol{\theta})$ a.e. This result can also be obtained from the property of the function $z(\boldsymbol{\theta})$ or $\tilde{H}(\boldsymbol{\theta}^*; \boldsymbol{\theta})$, the expected log likelihood in (12) which is: $z(\boldsymbol{\theta})$ or $\tilde{H}(\boldsymbol{\theta}^*; \boldsymbol{\theta})$ attains its maximum value at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, and if distributions on the sample space corresponding to different parameters are essentially different, then for no other $\boldsymbol{\theta}$ is $z(\boldsymbol{\theta})$ equal to $z(\boldsymbol{\theta}^*)$ (Silvey, 1975, p. 74).

Since our estimation of K-L information quantity is based on the mean log likelihood (which is also an estimate of the expected log likelihood), and since the maximum likelihood estimates are *biased*, then there is the inevitable risk of error of estimation of the K-L information quantity when the maximum likelihood estimators of the parameters of the model is used.

In the case where θ is a real parameter, for large n, we depict the general behavior exhibited by the mean log likelihood $1/n\ell(\theta)$ and the expected log likelihood $z(\theta) = E[1/n\ell(\theta)] = E[\log f(X \mid \theta)]$ in Figure 2.1.

From Figure 2.1, we see that $\tilde{H}(\theta^*; \theta)$ assumes its maximum value at θ^*, and that $1/n\ell(\theta)$ is uniformly near $\tilde{H}(\theta^*; \theta)$ ensures that $1/n\ell(\theta)$ assumes its maximum value at a point near θ^*, that is, $\hat{\theta}(x)$ near θ^* as discussed in Silvey (1975). In other words, for finite n, we would like to have the observed values of the mean log likelihood (dashed lines) to behave like the theoretical points (solid line) of Figure 2.1. To achieve this, the bias introduced by the maximum likelihood estimates of the parameters needs to be adjusted or corrected.

Empirical justification of the behavior of the mean log likelihood plotted against the number of parameters are illustrated in detail by Atilgan (1983), and Atilgan and Bozdogan (1987) in smoothing and density estimation under various basis functions including those of *B*-splines, normal kernels, and logistic density transform.

Since the quantity $H(\boldsymbol{\theta}^*; \boldsymbol{\theta})$ is not directly observable, maximization of the mean log likelihood is carried out, and asymptotically an unbiased estimator of the mean expected log likelihood is searched by correcting the bias of the observed mean log likelihood, $\ell_n(\hat{\boldsymbol{\theta}})$.

Indeed, in defining AIC, Akaike (1973, 1974) has exactly this consideration of the bias

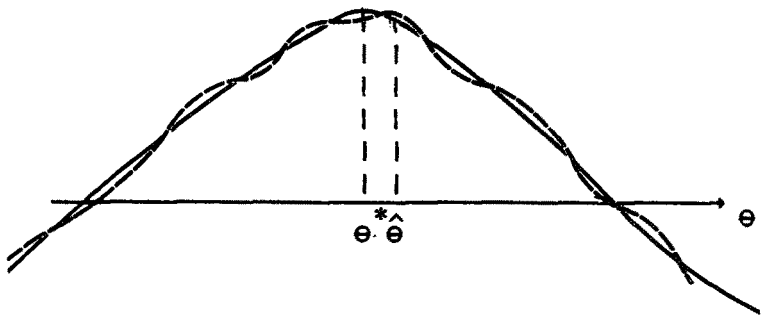

Figure 2.1. Plot of $\tilde{H}(\theta^*;\theta)$, expected log likelihood (solid line) and $\frac{1}{n}\ell(\theta)$, mean log likelihood (dashed line) (Silvey, 1975, p. 76).

by penalizing extra parameters when the maximum likelihood estimates are used in estimating the expected log likelihood by the mean log likelihood.

Now we are in a position to give the definition of entropy maximization principle emphasized by Akaike (1977) which is quite different from (and should not be confused with) Jayne's (1957) maximum entropy principle.

Definition 1: *Entropy maximization principle* (*EMP*). Formulate the object of statistical inference as the estimation of the true distribution $f(\mathbf{x}\,|\,\boldsymbol{\theta}^*)$ from the data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and try to search for an approximate model $f(\mathbf{x}\,|\,\boldsymbol{\theta})$ which will maximize expected entropy:

$$\begin{aligned} E_{\mathbf{X}}[B(\boldsymbol{\theta}^*;\boldsymbol{\theta})] &= \int B(\boldsymbol{\theta}^*;\boldsymbol{\theta}) f(\mathbf{x}\,|\,\boldsymbol{\theta}^*)\, d\mathbf{x} \\ &= E_{\mathbf{X}}[H(\boldsymbol{\theta}^*;\boldsymbol{\theta})] \\ &= E_{\mathbf{X}}\{E[\log f(\mathbf{X}\,|\,\boldsymbol{\theta})]\}. \end{aligned} \tag{15}$$

Equivalently, we minimize the expected K-L information quantity.

Large expected log likelihood $E[\log f(\mathbf{X}\,|\,\boldsymbol{\theta})]$ means large entropy, and it also means that the model $f(\mathbf{x}\,|\,\boldsymbol{\theta})$ is a good fit to $f(\mathbf{x}\,|\,\boldsymbol{\theta}^*)$, or, equivalently, the low value of $I(\boldsymbol{\theta}^*;\boldsymbol{\theta})$ means that the model $f(\mathbf{x}\,|\,\boldsymbol{\theta})$ is a good fit to $f(\mathbf{x}\,|\,\boldsymbol{\theta}^*)$. Thus, it is by the mean of $B(\boldsymbol{\theta}^*;\boldsymbol{\theta})$, $I(\boldsymbol{\theta}^*;\boldsymbol{\theta})$, or the mean expected log likelihood over the sampling distribution of the estimator of $\boldsymbol{\theta}$ that we will judge a particular model and measure our ignorance about the true structure of the model.

This definition will play an important role in that, since entropy or negentropy is nothing but a *log odds ratio* between the fitted model $f(\mathbf{x}\,|\,\boldsymbol{\theta})$ and the true distribution $f(\mathbf{x}\,|\,\boldsymbol{\theta}^*)$, it can be used as a *loss function*, and its expected value can be used as a *risk function* to measure the average estimation error of the fitted model from the true one as suggested by Akaike (1973).

Next, we derive AIC in detail as a natural sample estimate of $E[\log f(\mathbf{X}\,|\,\boldsymbol{\theta})]$, the expected log likelihood.

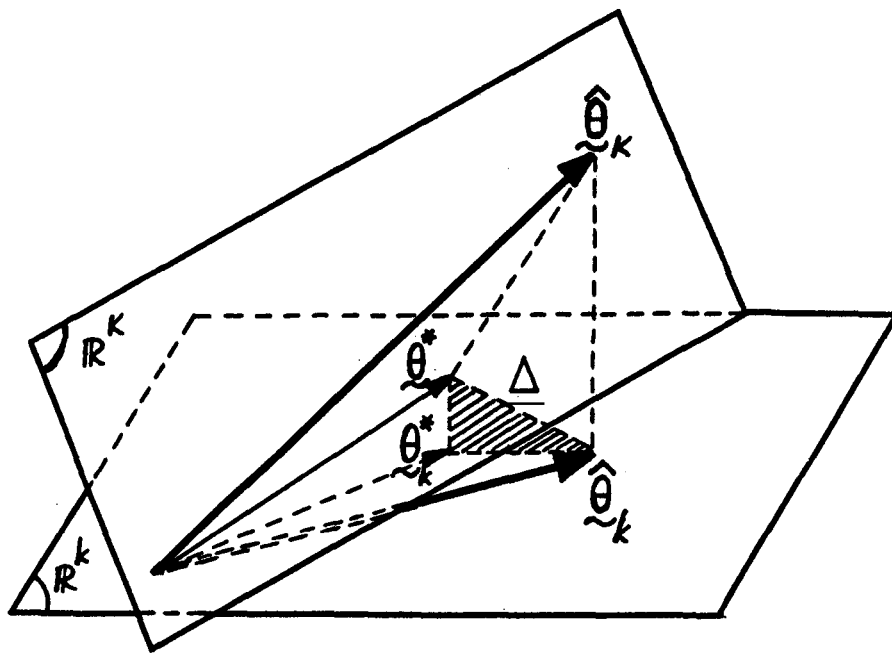

Figure 3.1. Parametric estimation interpreted as Euclidean projection from $\mathbb{R}^K$ to $\mathbb{R}^k$.

3. Akaike's Information Criterion (AIC) as an Estimate of Negentropy

Suppose that we have a general model $f(\cdot \mid \boldsymbol{\theta})$ from which all the K competing models are generated by simply restricting the general parameter vector $\boldsymbol{\theta}$. In terms of the parameters, we represent the full model with K parameters by

$$\text{MODEL}(K): \quad f(\cdot \mid \underline{\boldsymbol{\theta}}), \quad \boldsymbol{\theta} \equiv \boldsymbol{\theta}_K = (\theta_1, \theta_2, \ldots, \theta_k, \theta_{k+1}, \ldots, \theta_K). \tag{16}$$

We denote the "true" value of the parameter vector $\boldsymbol{\theta}$ by $\boldsymbol{\theta}^*$ with $\boldsymbol{\theta}^* \in \mathbb{R}^K$. Following Akaike (1973), the problem of statistical model identification can be formulated as the problem of selecting a model $f(\mathbf{x} \mid \boldsymbol{\theta}_k)$ based on n observations. In terms of parameters, $\boldsymbol{\theta}_k$ is restricted to the space with $\theta_{k+1} = \theta_{k+2} = \ldots = \theta_K = 0$, or equal to a prescribed value of θ_k, and that a particular restricted model with k parameters is given by

$$\text{MODEL}(k): \quad f(\cdot \mid \boldsymbol{\theta}_k), \quad \boldsymbol{\theta}_k = (\theta_1, \theta_2, \ldots, \theta_k, 0, 0, \ldots, 0). \tag{17}$$

Often k, the number of free parameters of MODEL(k), is called the *dimension* or *order* of the model. If we represent parametric estimation geometrically in terms of a Euclidean projection shown in Figure 3.1, we denote the maximum likelihood estimate of $\boldsymbol{\theta}_K$ by $\hat{\boldsymbol{\theta}}_K$ both of which lie in $\mathbb{R}^K$ (the K-dimensional Euclidean space). We let $\boldsymbol{\theta}^*$ denote the true parameter vector with $\boldsymbol{\theta} \in \mathbb{R}^K$, and define $\boldsymbol{\theta}_k^*$ as the parameter vector of the best fitting or approximating model with $\boldsymbol{\theta}_k^* \in \mathbb{R}^k$ (k-dimensional parametric subspace $\mathbb{R}^k$ of the original space $\mathbb{R}^K$). We note that the parameter vector $\boldsymbol{\theta}_k^*$ is the projection of the true parameter vector $\boldsymbol{\theta}^*$ in the subspace $\mathbb{R}^k$. Next, we let $\hat{\boldsymbol{\theta}}_k$ denote the restricted maximum likelihood estimator of $\boldsymbol{\theta}_k$ of MODEL(k) which lies in $\mathbb{R}^k$, and which is also the projection of $\hat{\boldsymbol{\theta}}_K$ in $\mathbb{R}^k$. Thus, geometrically $(\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k)$ will constitute our random error. It can be viewed

approximately as the projection of $(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_K)$ into the subspace $\mathbb{R}^k$. On the other hand, $(\boldsymbol{\theta}^* - \boldsymbol{\theta}_k^*)$ is deterministic and is the bias due to selecting an approximate parameter space.

To elaborate further, following Clergeot (1984), if for different values of k, subspaces $\mathbb{R}^k$ are chosen so that $\mathbb{R}^k \subset \mathbb{R}^{k+1}$, the bias, which is the distance of $\boldsymbol{\theta}^*$ to the subspace $\mathbb{R}^k$, is a nonincreasing function of k, while the random error increases monotonically with k. Thus, our goal here will be to find some optimal value of k so that the compromise between bias and random error will give the smallest estimation error.

The AIC statistic which we next state in the form of a proposition and sketch its derivation is designed to approximate the real model by a lower dimensional model so as to minimize the average estimation error.

Proposition 1: *Akaike's information criterion* (*AIC*): Let $\{M_k: k = 1, 2, \ldots, K\}$ be a set of competing models indexed by $k = 1, 2, \ldots, K$. Then the criterion

$$\text{AIC}(k) = -2 \log L(\hat{\boldsymbol{\theta}}_k) + 2k, \tag{18}$$

which is minimized to choose a model M_k over the set of models is a natural sample estimator of twice the negentropy, $2E[I(\boldsymbol{\theta}^*; \theta_k)]$, or minus twice the expected log likelihood, $-2E[\log f(\mathbf{X} \mid \boldsymbol{\theta}_k)]$, of the true distribution with respect to a model with the parameters determined by the method of maximum likelihood.

Proof. Following Akaike (1973, 1974, 1976), and Kitagawa (1979), we first assume a model which specifies a probability density function $f(\mathbf{x} \mid \boldsymbol{\theta}_k)$ of n observations with a free parameter vector $\boldsymbol{\theta}_k$ and then find the MLE $\hat{\boldsymbol{\theta}}_k$ of $\boldsymbol{\theta}_k$ with $\boldsymbol{\theta}_k \in \mathbb{R}^k$ by maximizing the likelihood function

$$L(\boldsymbol{\theta}_k \mid \mathbf{x}) = L(\boldsymbol{\theta}_k \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}_k), \tag{19}$$

with respect to $\boldsymbol{\theta}_k$. By taking the natural logarithm of the likelihood function, and dividing by the sample size n, we get

$$\frac{1}{n}\ell(\boldsymbol{\theta}_k) = \frac{1}{n} \log L(\boldsymbol{\theta}_k \mid \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i \mid \boldsymbol{\theta}_k), \tag{20}$$

the average or mean log likelihood which is a natural consistent estimator of

$$E[\log f(\mathbf{X} \mid \boldsymbol{\theta}_k)] = \int \log f(\mathbf{x} \mid \boldsymbol{\theta}_k) f(\mathbf{x} \mid \boldsymbol{\theta}_k)\, d\mathbf{x}, \tag{21}$$

the expected log likelihood.

Since the purpose of estimating the parameters of the true model $f(\mathbf{x} \mid \boldsymbol{\theta}^*)$ is to base our decision on $f(\mathbf{x} \mid \hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}} \equiv \hat{\boldsymbol{\theta}}_K$ is an estimate of $\boldsymbol{\theta}^*$, then the discussion in section 2 suggests the use of K-L information quantity

$$I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}) = -B(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}) = \int \log \left[\frac{f(\mathbf{x} \mid \boldsymbol{\theta}^*)}{f(\mathbf{x}_g \mid \hat{\boldsymbol{\theta}})} \right] f(\mathbf{x} \mid \boldsymbol{\theta}^*)\, d\mathbf{x}, \tag{22}$$

as the loss function, and

$$E_{\mathbf{X}}[I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}})] = \int I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}) f(\mathbf{x} \mid \boldsymbol{\theta}^*)\, d\mathbf{x} \tag{23}$$

as the risk function according to Definition 1 EMP of Akaike (1977).

Expanding $I(\boldsymbol{\theta}^*; \boldsymbol{\theta})$ in a Taylor series with respect to its second argument around $\boldsymbol{\theta}^*$,

we have an approximation (see, e.g., also Kullback, 1959) given as follows.

$$I(\boldsymbol{\theta}^*; \boldsymbol{\theta}^* + \Delta\boldsymbol{\theta}) \cong \tfrac{1}{2} \| \Delta\boldsymbol{\theta} \|_J^2, \tag{24}$$

where

$$\| \Delta\boldsymbol{\theta} \|_J^2 = \| \boldsymbol{\theta} - \boldsymbol{\theta}^* \|_J^2 = (\boldsymbol{\theta} - \boldsymbol{\theta}^*)' J (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \tag{25}$$

and J is the $(K \times K)$ Fisher information matrix which is positive definite and defined by

$$J = E\left\{ \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{X} \mid \boldsymbol{\theta}) \right]' \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{X} \mid \boldsymbol{\theta}) \right] \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*} \right\}. \tag{26}$$

As shown in Figure 3.1, we next restrict $\boldsymbol{\theta}$ to lie in k-dimensional restricted parameter space, Ω_k, while the true parameter vector, $\boldsymbol{\theta}^*$, lies in a K-dimensional full parameter space Ω_K, where $k < K$. Denoting by $\boldsymbol{\theta}_k^*$ the projection of $\boldsymbol{\theta}^*$ onto Ω_k and using the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_k$ of $\boldsymbol{\theta}_k^*$ in Ω_k, we have

$$2I(\boldsymbol{\theta}^*; \boldsymbol{\theta}_k) \cong 2I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k), \tag{27}$$

so that using the approximation in (3.7), we have

$$\begin{aligned} 2I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k) &\cong \| \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k \|_J^2 \\ &\cong \| \boldsymbol{\theta}^* - \boldsymbol{\theta}_k^* \|_J^2 + \| \boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k \|_J^2 \end{aligned}$$

by the Pythagorean theorem.

Thus, for large n, a measure of the average estimation error is given by the expectation of the K-L information quantity:

$$\begin{aligned} 2nE[I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k)] &\cong E[n \| \boldsymbol{\theta}^* - \boldsymbol{\theta}_k^* \|_J^2 + n \| \boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k \|_J^2] \\ &= n \| \boldsymbol{\theta}^* - \boldsymbol{\theta}_k^* \|_J^2 + E[n \| \boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k \|_J^2], \end{aligned} \tag{28}$$

where the first term in (28) is the bias and the second term is a measure of the variance of the random error $(\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k)$. For the second term, that is, for $n \| \boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k \|_J^2$ under the expectation, for sufficiently large n, we have

$$n \| \boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k \|_J^2 = \| (n)^{1/2} (\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k) \|_J^2 \overset{\text{a.d.}}{\sim} \chi^2 \tag{29}$$

with k degrees of freedom, and a.d. stands for asymptotically distributed. Since $E(\chi^2) = k$, its degrees of freedom, then (28) for large n approximately becomes

$$\begin{aligned} 2nE[I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k)] &\cong n \| \boldsymbol{\theta}^* - \boldsymbol{\theta}_k^* \|_J^2 + k \\ &\cong \delta + k. \end{aligned} \tag{30}$$

Equation (30) represents the overall risk of statistical modeling which measures the extent to which $\hat{\boldsymbol{\theta}}_k$ deviates from the true parameter vector $\boldsymbol{\theta}^*$. As we note, it is composed of two components; one which involves $\delta = n \| \boldsymbol{\theta}^* - \boldsymbol{\theta}_k^* \|_J^2$, is the error or bias due to selecting an approximate parameter space for the restricted parameter space for $\boldsymbol{\theta}_k^*$, and the other which involves k, is a measure of variance or the random error due to estimating the specified parameter vector.

Of course, it is impossible to minimize (30) directly since the bias $\delta = n \| \boldsymbol{\theta}^* - \boldsymbol{\theta}_k^* \|_J^2$ is *unknown*, but it is *deterministic*. It needs to be estimated in practice with finite samples. Akaike (1973) cleverly estimates δ using Wald's (1943) results on the asymptotic distribution of the log likelihood ratio statistic, namely, that when $\mathbf{x}$ is a vector of observations of independently identically distributed random variables under certain regularity con-

ditions the likelihood ratio statistic

$$ {}_k\eta_K = -2 \log \lambda = LR(\mathbf{x}) = -2 \sum_{i=1}^{n} \log \frac{f(x_i \mid \hat{\boldsymbol{\theta}}_k)}{f(x_i \mid \hat{\boldsymbol{\theta}}_K)} \tag{31} $$

is used to estimate $I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k)$, since the mean log likelihood is a consistent estimate of $I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k)$, and that (31) is asymptotically distributed as a noncentral χ^2 random variable. That is,

$$ {}_k\eta_K = -2 \log \lambda \overset{\text{a.d.}}{\sim} \chi'^2_v(\delta) \tag{32} $$

with $v = K - k$ degrees of freedom and the noncentrality parameter $\delta = n \| \boldsymbol{\theta}^* - \boldsymbol{\theta}_k^* \|_J^2$.

Since

$$ E[\chi'^2_v(\delta)] = \delta + v, \tag{33} $$

and since

$$ {}_k\eta_K = -2 \log \lambda \cong E[-2 \log \lambda] = E[\chi'^2_v(\delta)] = \delta + v, \tag{34} $$

solving for δ, we have

$$ \begin{aligned} \delta = n \| \boldsymbol{\theta}^* - \boldsymbol{\theta}_k^* \|_J^2 &\cong -2 \log \lambda - v \\ &= -2 \log \lambda - (K - k). \end{aligned} \tag{35} $$

It follows that (30) becomes

$$ \begin{aligned} -2\, nE[B(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k)] &= 2nE[I[\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k)] \\ &\cong -2 \log \lambda - (K - k) + k. \end{aligned} \tag{36} $$

Simplifying the right hand side of (36), we have

$$ \begin{aligned} -2 \log \lambda + 2k - K &= -2 \log \frac{L(\hat{\boldsymbol{\theta}}_k)}{L(\hat{\boldsymbol{\theta}}_K)} + 2k - K \\ &= -2[\log L(\hat{\boldsymbol{\theta}}_k) - \log L(\hat{\boldsymbol{\theta}}_K)] + 2k - K \\ &= -2 \log L(\hat{\boldsymbol{\theta}}_k) + 2k + 2 \log L(\hat{\boldsymbol{\theta}}_K) - K, \end{aligned} \tag{37} $$

so that

$$ \begin{aligned} -2nE[B(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k)] &= 2nE[I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k)] \\ &\cong -2 \log L(\hat{\boldsymbol{\theta}}_k) + 2k + 2 \log L(\hat{\boldsymbol{\theta}}_K) - K, \\ &\cong {}_k\eta_K + 2k - K. \end{aligned} \tag{38} $$

It follows from the above discussion that if the K-L information quantity, $I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k)$, is adopted as the loss function in our model building with the associated risk function (i.e., the expected loss), then from (36) we note that

$$ \mathscr{R}((\hat{\boldsymbol{\theta}}_K; \hat{\boldsymbol{\theta}}_k) = \frac{1}{n} (-2 \log \lambda + 2k - K) \tag{39} $$

serves as a useful estimate of this risk function, namely, $E[I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k)]$, at least for the case where n is sufficiently large, and K and k are relatively large integers. In practical applications, K sometimes may happen to be very large, or conceptually infinite integer, and

may not be defined clearly. Even under such circumstances we choose limited number of k's, assuming K to be equal to the larger hypothesized value of k. Since we are only concerned with finding out the $\hat{\boldsymbol{\theta}}_k$ which will give the minimum of $\mathscr{R}(\hat{\boldsymbol{\theta}}_K; \hat{\boldsymbol{\theta}}_k)$ in (39), or equivalently, the minimum of $2nE[I(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k)]$ in (38), we have only to compute either

$$ {}_k\nu_K = {}_k\eta_K + 2k = -2\sum_{i=1}^{n} \log \frac{f(x_i \mid \hat{\boldsymbol{\theta}}_k)}{f(x_i \mid \hat{\boldsymbol{\theta}}_K)} + 2k, \tag{40} $$

or, ignoring the constant terms in (38) common to every model, we reduce the form of AIC to a much simpler form

$$ \begin{aligned} \mathrm{AIC}(k) &= -2\sum_{i=1}^{n} \log f(x_i \mid \hat{\boldsymbol{\theta}}_k) + 2k \\ &= -2 \log L(\hat{\boldsymbol{\theta}}_k) + 2k, \end{aligned} \tag{41} $$

and choose the minimum of AIC over $k = 1, 2, \ldots, K$.

This completes the derivation of AIC and the sketch of the proof of the proposition. □

We note that AIC(k) in (41) is an *unbiased estimator* of minus twice the mean expected log likelihood, or equivalently $-\frac{1}{2}$ AIC(k) is asymptotically an unbiased estimator of the mean expected log likelihood. This result suggests that asymptotically a reasonable definition of the likelihood of a model is

$$ L(k) = \exp\{-\tfrac{1}{2}\,\mathrm{AIC}(k)\} \quad (k = 1, 2, \ldots, K). \tag{42} $$

In fact, with the assumption of equal prior probability for the models, the distribution of (42) defines the posterior distribution of models. If there are several models with almost equal values of AIC, it is useful to consider the averaged model by using (42) as the "likelihood" of each model as discussed in Akaike (1978, 1979).

We interpret the result in (41) as follows. The first term in (41) is a measure of inaccuracy, badness of fit, or bias when the maximum likelihood estimators of the parameters of the model are used. The second term, on the other hand, is a measure of complexity or the penalty due to the increased unreliability or compensation for the bias in the first term which depends upon the number of parameters used to fit the data.

Thus, when there are several competing models the parameters within the models are estimated by the method of maximum likelihood and the values of the AIC's are computed and compared to find a model with the minimum value of AIC. This procedure is called the *minimum AIC procedure* and the model with the minimum AIC is called the *minimum AIC estimate* (MAICE) and is chosen to be the best model. Therefore, for us the best model is the one with least complexity, or equivalently, the highest information gain. In applying AIC, the emphasis is on comparing the goodness of fit of various models with an allowance made for parsimony.

4. Consistent Akaike's Information Criterion: CAIC(k)

As we saw in the derivation of AIC in the previous section, one of the important virtues of AIC is the penalty represented by the term $2\times$ (number of free parameters) clearly demonstrates the necessity of choosing a class of models, at least one of which will be able to provide a good approximation to the distribution of the data without adjusting too many parameters. However, in the literature, the particular specifications put on the crucial structure-dependent term in AIC, that is, the so-called "magic number" 2, has been

questioned unfairly as being coincidental or arbitrary (Rissanen, 1978). If one follows the derivation of AIC in section 3, the emergence of the magic number 2 is hardly debatable. The debating questions should be: Is the magic number 2 enough, should it be greater than 2, how do we choose such a number, or on what does it depend, and so forth?

Based on the frequency of choice of the correct model from a simulation study, many authors, including Bhansali and Downham (1977), in a time series model, arbitrarily considered the range of the magic number to be between 1 and 4. The propriety of such a choice has been criticized by both Akaike (1979) and Atkinson (1980) on the grounds of lack of objectivity and the need for more explicit analytical formulation.

Objections have been raised that minimizing AIC does not produce an asymptotically consistent estimate of model order (Bhansali and Downham, 1977; Schwarz, 1978; Woodroofe, 1982; and others). However, consistency is an asymptotic property, and any real problem has a finite sample size n as stated in Sclove (1987). Certainly, from a mathematical point of view, consistency is an attractive asymptotic property to expect from a model selection procedure, but any consideration of consistency presupposes that there exist "true" order of a model. In the case of real data, the concept of the true order is not known and is suspect.

Even though the AIC procedure is not claimed to be consistent and is not designed to be such (Akaike, 1981b; Quinn, 1980), its inconsistency is not necessarily a defect in the method, and the virtue of consistency should not be exaggerated (Hannan, 1986). Shibata (1983) extensively studied the asymptotic behavior of AIC procedure and its variants in terms of asymptotic efficiency under a quadratic loss function. He found that AIC actually achieves the implicit goal that motivated Akaike—it does the best that any model-selection procedure can do in minimizing negentropy, or the expected log likelihood, at least for a large sample size n (Larimore & Mehra, 1985).

In general, the major dilemma here is how to balance optimally the underfitting and overfitting risks, or how to optimally adjust the bias in the log likelihood ratio when the maximum likelihood estimates are used.

Without violating Akaike's principles, using the established results in mathematical statistics, we improve and extend AIC analytically in two ways. These extensions make AIC asymptotically consistent, and that we penalize overparameterization more stringently to pick the simplest of the true models whenever there is nothing to be lost in doing so.

In section 2.2, we discussed the fact that when the mean log likelihood is used to estimate the K-L information quantity, the bias introduced by the maximum likelihood estimates of the parameters needs to be adjusted or corrected. In the derivation of AIC, this bias comes out as a noncentrality parameter, δ, which is a very large unknown but deterministic constant. It depends not only on the number of observations but also the specific estimation method used. For example, in our case this method is the maximum likelihood (ML) method. Moreover, the distributional change in δ for different sample sizes is also crucial to justify the correction of the bias further. That is, we do not want to have a very large δ, since it varies with the basic model. As is well known, noncentrality parameters determine the power of test procedures, and the estimation of δ on the basis of preliminary data may be necessary to choose among competing models.

We note from (35) that one such correction in δ is already given in deriving AIC, that is, $\delta \cong -2 \log \lambda - (K - k)$. Also we note that this correction factor $v = (K - k)$ is independent of the sample size n. However, in testing a null hypothesis (or a model) distinguished from the alternative hypothesis (or hypotheses) by the value of a parameter, if the test statistic has a noncentral chi-square distribution which is the case here, then the

degrees of freedom is an increasing function of the sample size n (Kendall & Stuart, 1967). This suggests that to make AIC consistent, the multiplier of the number of free parameters in the penalty term must be made to depend on the sample size by setting

$$v = a(n)(K - k), \tag{43}$$

where $a(n)$ is an increasing function of n. In AIC, we note that $a(n) = 1$. As discussed in Davis and Vinter (1985), the selection of the function $a(n)$ is important, and it should be chosen so that it has various desirable properties for the corresponding estimates. Recapitulating Davis and Vinter (1985), the properties of $a(n)$ which might be required are:

(i) high probability of choosing the correct dimension of the model, or order for finite data sets; and

(ii) consistency, that is, asymptotically correct choice of k as $n \to \infty$.

Presently, in the literature there is no theory available as to how to choose the correct dimension of a model with high probability. On the other hand, consistency holds almost surely for certain choices of $a(n)$ including the choice $a(n) = \log n$, where "log" denotes the natural logarithm. Therefore, we choose $a(n) = \log n$ since it provides a monotonically increasing function of the sample size n, such that $a(n)/n \to 0$ as $n \to \infty$ (see also, e.g., Akaike, 1978). We shall denote this type of modified AIC by a generic name, CAIC. So if we take $v = (K - k) \log n$, then we have the following proposition.

Proposition 2.

$$\text{CAIC}(k) = -2 \log L(\hat{\theta}_k) + k[(\log n) + 1]. \tag{44}$$

Proof. Recall from (30) that

$$2nE[I(\theta^*; \hat{\theta}_k)] \underline{\cong} \delta + k. \tag{45}$$

Since

$$\delta \underline{\cong} -2 \log \lambda - v, \tag{46}$$

and since $v = (K - k) \log n$, then

$$\delta \cong -2 \log \lambda - (K - k) \log n. \tag{47}$$

Substituting this into (46) and simplifying, we get

$$2nE[I(\theta^*; \hat{\theta}_k)] \cong -2 \log L(\hat{\theta}_k) + k \log n + k + 2 \log L(\hat{\theta}_K) - K \log n. \tag{48}$$

Since CAIC estimates the quantity $2nE[I]$, that is, twice the expected Kullback information, then

$$\text{CAIC}(k) = -2 \log L(\hat{\theta}_k) + k \log n + k + 2 \log L(\hat{\theta}_K) - K \log n. \tag{49}$$

Since the constants do not effect the results of comparison of models, we drop the additive terms and reduce the form of CAIC(k) to a much simpler form, namely

$$\text{CAIC}(k) = -2 \log L(\hat{\theta}_k) + k[(\log n) + 1]. \quad \square \tag{50}$$

This shows that AIC can be fairly easily extended to make it consistent, even though a practical difficulty is that consistency is a *weak property* (Atkinson, 1980). Note that CAIC(k) is similar to the Schwarz's (1978) Criterion of $k \log n$, and that the term $[k \log n + k]$ has the effect of increasing the "penalty term." Consequently, the mini-

mization of CAIC leads, in general, to lower dimensional models than those obtained by minimizing AIC.

In the literature, Hannan and Quinn (1979) proposed $a(n) = c \log \log n$, where $c > 2$. Their objective was to provide a consistent criterion in which $a(n)$ increases with n but at as slow a rate as possible. If we let $\underline{k} \leq \bar{k}(n)$, where $\bar{k}$ is an upper bound of k, possibly depending on n, other choices of $a(n)$ are:

$$a(n) = (\log n)^{1+b} \quad \text{and} \quad \bar{k}(n) = (\log n)^c, \tag{51}$$

where b and c are arbitrary strictly positive constants.

Next, we give yet another analytical extension of AIC which will unify all these procedures.

5. Consistent AIC with Fisher Information: CAICF(k)

In this section, exploiting the large sample asymptotic distributional properties of the maximum likelihood estimators, we propose a different estimator for minus twice the expected entropy to extend AIC analytically to make it consistent without deviating from Akaike's original premise. In this manner we penalize overparameterization more strongly, in particular, for large samples. Therefore, instead of dropping the term $2 \log L(\hat{\boldsymbol{\theta}}_K)$ in (38) as we do in simplifying the expression to obtain a simple form for AIC, we retain this term and propose a different approximation to $L(\hat{\boldsymbol{\theta}}_K)$ which estimates the likelihood function $L(\boldsymbol{\theta}^*)$ of the true model. This means that one can specify the true model to be the most general of the models to be selected and consider the fact that the observations are generated by the true density $f(\mathbf{x} \mid \boldsymbol{\theta}^*)$ instead of $f(\mathbf{x} \mid \boldsymbol{\theta}_k^*)$. This way we can estimate the unknown parameters of the true and approximate models by using the maximum likelihood estimators and their properties, and obtain the following proposition.

Proposition 3.

$$\begin{aligned} \text{CAICF}(k) &= -2 \log L(\hat{\boldsymbol{\theta}}_k) + k[(\log n) + 2] + \log |J(\hat{\boldsymbol{\theta}}_k)| \\ &= \text{AIC}(k) + k \log n + \log |J(\hat{\boldsymbol{\theta}}_k)|. \end{aligned} \tag{52}$$

To show the derivation of this proposition, we consider $\mathbf{x}$ i.i.d., and we assume that the true parameter vector $\boldsymbol{\theta}^*$ satisfies the restrictions set by the following model in terms of the parameters:

$$\text{MODEL}(k): \boldsymbol{\theta}_k = (\theta_1, \theta_2, \ldots, \theta_k, \quad 0, \ldots, 0), \tag{53}$$

and that $\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k = O(n^{-1/2})$, where O denotes the "order of."

Using the properties of the maximum likelihood estimates for regular models, it is well known that the MLE $\hat{\boldsymbol{\theta}}_k$ of $\boldsymbol{\theta}_k^*$ is, at least asymptotically, a sufficient statistic for $\boldsymbol{\theta}_k^*$. This can be shown easily by the factorization theorem of the likelihood of the kind to establish sufficiency (Cox & Hinkley, 1974). Assuming that $\hat{\boldsymbol{\theta}}_k$ is at least asymptotically sufficient for $\boldsymbol{\theta}_k^*$, under this assumption, it is easy to establish the following theorem for asymptotic normality.

Theorem 1. A maximum likelihood estimator $\hat{\boldsymbol{\theta}}_k$ is asymptotically distributed as multivariate normal with mean vector $\boldsymbol{\theta}_k^*$ and covariance matrix $(nJ)^{-1}$. That is,

$$\hat{\boldsymbol{\theta}}_k \overset{\text{a.d.}}{\sim} N_k(\boldsymbol{\theta}_k^*; (nJ(\boldsymbol{\theta}_k^*))^{-1}). \tag{54}$$

So the asymptotic multivariate normal density of $\hat{\boldsymbol{\theta}}_k$ is given by

$$g(\hat{\boldsymbol{\theta}}_k) = \frac{|nJ(\boldsymbol{\theta}_k^*)|^{1/2}}{(2\pi)^{k/2}} \exp\{-\tfrac{1}{2}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)' nJ(\boldsymbol{\theta}_k^*)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)\}, \tag{55}$$

where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimate of $\boldsymbol{\theta}_k^*$ and inverse covaraince matrix is

$$C(\boldsymbol{\theta}_k^*) = -\left[\frac{\partial^2 \log L(\boldsymbol{\theta}_k^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \equiv nJ(\boldsymbol{\theta}_k^*), \tag{56}$$

where $J(\boldsymbol{\theta}_k^*)$ is the Fisher information matrix at $\boldsymbol{\theta}_k^*$ with respect to one observation.

For large samples, we approximate the likelihood of $f(\mathbf{x}\,|\,\boldsymbol{\theta}^*)$ the true density at $\boldsymbol{\theta}_k^*$ by $L(\boldsymbol{\theta}_k^*; \hat{\boldsymbol{\theta}}_k) = g(\boldsymbol{\theta}_k^*)L(\boldsymbol{\theta}_k^*) = g(\boldsymbol{\theta}_k^*) \exp\{\log L(\boldsymbol{\theta}_k^*)\}$, using Taylor series expansion of $g(\boldsymbol{\theta}_k^*)$ and $\exp\{\log L(\boldsymbol{\theta}_k^*)\}$ around the ML estimate $\hat{\boldsymbol{\theta}}_k$. Taking the product of two Taylor series expansions and using the leading terms, we obtain

$$L(\boldsymbol{\theta}_k^*; \hat{\boldsymbol{\theta}}_k) = \frac{(n^k\,|\,J(\boldsymbol{\theta}_k^*)\,|)^{1/2}}{(2\pi)^{k/2}} \exp\{-\tfrac{1}{2}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)' nJ(\boldsymbol{\theta}_k^*)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)\}\,[1 + O(n^{-1/2})]. \tag{57}$$

Hence, using asymptotic sufficiency, we can write

$$L(\boldsymbol{\theta}_k^*) = h(\mathbf{x})L(\boldsymbol{\theta}_k^*; \hat{\boldsymbol{\theta}}_k) \tag{58}$$

by the factorization criterion, where $h(\mathbf{x})$ is independent of the particular parameter vector $\boldsymbol{\theta}$ we choose, and the second factor $L(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}_k)$ depends on $\mathbf{x}$ only through the value of $\hat{\boldsymbol{\theta}}_k \equiv \hat{\boldsymbol{\theta}}_k(\mathbf{x})$, which is sufficient for $\boldsymbol{\theta}_k^*$.

Now, in (38) we replace $L(\hat{\boldsymbol{\theta}}_K)$ by $L(\boldsymbol{\theta}_k^*)$ given in (58) assuming that $\boldsymbol{\theta}^*$ (the true parameter vector) is situated near $\boldsymbol{\theta}_k^*$ (the *restricted* or *pseudotrue* parameter vector) and it "almost" satisfies the restrictions set in (53) and get

$$_k\eta_K = -2 \log L(\hat{\boldsymbol{\theta}}_k) + 2 \log L(\boldsymbol{\theta}_k^*). \tag{59}$$

So, from (58) and (57), we have

$$\begin{aligned}\log L(\boldsymbol{\theta}_k^*) &= \log h(\mathbf{x}) + \log L(\boldsymbol{\theta}_k^*; \hat{\boldsymbol{\theta}}_k)\\ &= \log h(\mathbf{x}) + \frac{k}{2}\log n + \tfrac{1}{2}\log|J(\boldsymbol{\theta}_k^*)| - \frac{k}{2}\log(2\pi)\\ &\quad - \tfrac{1}{2}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)' nJ(\boldsymbol{\theta}_k^*)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) + \log[1 + O(n^{-1/2})].\end{aligned} \tag{60}$$

Now, multiplying both sides of (60) by 2, we get

$$\begin{aligned}2 \log L(\boldsymbol{\theta}_k^*) &= 2 \log h(\mathbf{x}) + k \log n + \log|J(\boldsymbol{\theta}_k^*)| - k \log(2\pi)\\ &\quad - (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)' nJ(\boldsymbol{\theta}_k^*)(\boldsymbol{\theta}_k^* - \boldsymbol{\theta}_k^*) + 2 \log[1 + O(n^{-1/2})].\end{aligned} \tag{61}$$

Since $\log[1 + O(n^{-1/2})]$ is about of order $n^{-1/2}$, since $(\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k)$ is of order $n^{-1/2}$, $nJ(\boldsymbol{\theta}_k^*)$ is of order n, which implies $(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)' nJ(\boldsymbol{\theta}_k^*)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)$ is of order $O(1)$, then (61) reduces to

$$2 \log L(\boldsymbol{\theta}_k^*) = 2 \log h(\mathbf{x}) + k \log n + \log|J(\boldsymbol{\theta}_k^*)| - k \log(2\pi) + O(n^{-1/2}). \tag{62}$$

Thus, the approximation involves an error of order $n^{-1/2}$.

In Equation (62), by ignoring the constant terms, the term involving $\mathbf{x}$, and $O(n^{-1/2})$, we have

$$2 \log L(\boldsymbol{\theta}_k^*) = k \log n + \log|J(\boldsymbol{\theta}_k^*)|. \tag{63}$$

Hence, (59) reduces to

$$ {}_k\eta_K = -2 \log L(\hat{\theta}_k) + k \log n + \log |J(\theta_k^*)| \tag{64} $$

Now estimating $J(\theta_k^*)$ by $J(\hat{\theta}_k)$ in (64) where $\hat{\theta}_k$ is the MLE of θ_k^*, and substituting the result into (38), we obtain

$$ \begin{aligned} 2nE[I(\theta^*; \hat{\theta}_k)] &\cong -2 \log L(\hat{\theta}_k) + k \log n + \log |J(\hat{\theta}_k)| + 2k - K \\ &= -2 \log L(\hat{\theta}_k) + k[(\log n) + 2] + \log |J(\hat{\theta}_k)| - K. \end{aligned} \tag{65} $$

Simplifying (65) further, we have

$$ \begin{aligned} \text{CAICF}(k) &= -2 \log L(\hat{\theta}_k) + k[(\log n) + 2] + \log |J(\hat{\theta}_k)| \\ &= \text{AIC}(k) + k \log n + \log |J(\hat{\theta}_k)|. \end{aligned} \tag{66} $$

We note that if we take the first two terms in (66), CAICF is similar to CAIC of section 4, and also Schwarz's criterion (SC, 1978).

Hence, this way, without relying on the arbitrary frequency of choice of the correct model to modify AIC heuristically, we can analytically extend AIC to make it consistent. In this manner we penalize the overparameterization more strongly, in particular, for large samples. Note that we have not deviated from Akaike's original principle: we are still estimating minus twice the expected entropy. Also, note that we have not followed a Bayesian approach.

The incorporation of the Fisher information matrix $J(\hat{\theta}_k)$ within the penalty component of CAICF has many practical and theoretical importance. Generally, when we are using model-selection criteria, we fit the models under a specified parametric probability distribution of the model. The correct specification of the probability model is a sufficient, but by no means a necessary condition. For example, even when the true distribution is not normal, we still find the maximum likelihood estimators under the assumption of normality which yields consistent estimates of the mean and the variance. But it is the consistency of the mean log likelihood which ensures us the basis for robust estimation, and that it provides us the basis for constructing specification tests. Therefore, we need to check or test first whether the probability model is misspecified or not before we actually fit and evaluate the models. This is very important in practice which is often ignored. For this reason, following White (1982), we give a simple test of *information matrix equivalence* to check the misspecification of a model. First, we define the following matrices

$$ \begin{aligned} J_n(\theta_k^*) &= -\left\{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(x_i | \theta_k^*)}{\partial \theta_r \partial \theta_s}\right\}, \\ R_n(\theta_k^*) &= \left\{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f(x_i | \theta_k^*)}{\partial \theta_r} \cdot \frac{\partial \log f(x_i | \theta_k^*)}{\partial \theta_s}\right\}, \end{aligned} \tag{67} $$

If the expectations exist, we define

$$ \begin{aligned} J(\theta_k^*) &= -\left\{E\left[\frac{\partial^2 \log f(X | \theta_k^*)}{\partial \theta_r \partial \theta_s}\right]\right\}, \\ R(\theta_k^*) &= \left\{E\left[\frac{\partial \log f(X | \theta_k^*)}{\partial \theta_r} \cdot \frac{\partial \log f(X | \theta_k^*)}{\partial \theta_s}\right]\right\}, \end{aligned} \tag{68} $$

where both in (67) and (68), $r, s = 1, 2, \ldots, k$.

When the appropriate inverses exist, we define

$$C_n(\theta_k^*) = J_n(\theta_k^*)^{-1} R_n(\theta_k^*) J_n(\theta_k^*)^{-1} \equiv J_n^{-1} R_n J_n^{-1},$$
$$C(\theta_k^*) = J(\theta_k^*)^{-1} R(\theta_k^*) J(\theta_k^*)^{-1} \equiv \underline{J^{-1} R J^{-1}}, \tag{69}$$

where $J(\theta_k^*)$ is the *Fisher's information matrix* which is positive-definite, and $C(\theta_k^*)$ is the covariance matrix. When the model is correctly specified and certain assumptions hold as in White (1982, p. 6), we have the following result.

Theorem 2: *Information matrix equivalence.* If $f(\mathbf{x}) \equiv f(\mathbf{x} \mid \theta^*)$ for θ^* in Ω_K, then $\theta^* = \theta_k^*$ and $J(\theta_k^*) = R(\theta_k^*)$, so that the covariance $C(\theta_k^*) = J(\theta_k^*)^{-1} = R(\theta_k^*)^{-1}$.

Note that, in general, $J(\theta_k^*)$ will not equal $R(\theta_k^*)$ when the model is misspecified. However, when the model is correctly specified, then this theorem says that the information matrix can be expressed in either Hessian form, $J(\theta_k^*)$, or outer product form, $R(\theta_k^*)$, giving equivalently,

$$J(\theta_k^*) - R(\theta_k^*) = 0. \tag{70}$$

The matrix in (70) is not directly observable, but we can consistently estimate it by

$$J(\hat{\theta}_k) - R(\hat{\theta}_k) = 0, \tag{71}$$

and construct an appropriate test statistic to test whether a model misspecified or not. For example, if the equality in (71) fails, then this indicates that the probability model is misspecified. In practice, this misspecification may have many serious consequences when the classical inferential procedures are used. Furthermore, the failure of information matrix equivalence might also indicate misspecifications which render the MLE to be inconsistent for particular parameters of interest. Therefore, (71) is a useful indicator of misspecifications which cause either *parameter* or *covariance matrix estimator inconsistency.*

If we find that $J(\hat{\theta}_k)$, the estimated information matrix, is singular (or nearly singular) and so indefinite, we have an indication that the K-L information quantity has no unique minimum at θ^*, the true parameter vector. Therefore, CAICF as a large sample asymptotic estimator of the mean K-L information quantity will not have a unique minimum also. Furthermore, this in practice means that there are parameter vectors $\theta \neq \theta^*$ such that the expected mean log likelihood $z(\theta) = z(\theta^*)$ given in (12), and this in turn means that there are different parameters yielding the same distribution so that θ is not identifiable. Thus, lack of identifiability of θ implies singularity of information matrix and vice versa (Silvey, 1975, p. 82). Near singularity of the information matrix will also give us an indication of high variances for the estimators which is not preferred. In either case, the parameters may not be estimated with any high degree of accuracy. In this respect, one advantage of using CAICF in (66) is that it generalizes and unifies model selection procedures, and that, whenever possible, one should compute $J(\hat{\theta}_k)$, test whether the model is correctly specified or not, and then proceed with model fitting and evaluation. Indeed, if we cannot achieve a unique minimum by using CAICF, then this should be an indication that we have problems in identifying the parameters. In that case, we should bring in a priori information in order to impose restrictions on the model or tackle the problem with reparametrizing the model.

As discussed in Sclove (1987), Kashyap (1982), taking the Bayesian approach, took the asymptotic expansion of the logarithm of the posterior probabilities a term further

than did Schwarz (1978) and obtained the criterion approximately given by

$$\mathrm{KC}(k) = -2 \log L(\hat{\boldsymbol{\theta}}_k) - \log f(\boldsymbol{\theta}_k^*) + k \log n + \log |B(\hat{\boldsymbol{\theta}}_k)|, \tag{72}$$

where $f(\boldsymbol{\theta}_k^*)$ is the prior probability density on the parameter vector $\boldsymbol{\theta}_k^*$. $B(\hat{\boldsymbol{\theta}}_k)$ is the negative of the matrix of second partials of $\log L(\boldsymbol{\theta}_k)$, evaluated at the maximum likelihood estimates. In Gaussian linear models, this is the covariance matrix of the maximum likelihood estimates of the regression coefficient; in general, the expectation of $B(\hat{\boldsymbol{\theta}}_k)$, evaluated at the true parameter values, is Fisher's information matrix. If we assume equal probabilities for $f(\boldsymbol{\theta}_k^*)$, the term $\log f(\boldsymbol{\theta}_k^*)$ in (72) will be ignored, and thus we can see the relationship between KC and our extension of AIC, i.e., the CAICF given in (66). Similar results are also given in Haughton (1983) by extending the Schwarz's criterion for exponential families a term further than did Schwarz.

6. Asymptotic Properties and Inferential Error Rates of the Criteria

It is well known that the classical theory of hypothesis testing is concerned with the problem: Is a given observation consistent with some stated hypothesis or is it not? In the hypothesis testing tradition, frequently ignoring power considerations, we choose an arbitrary significance level α, for example, the celebrated 5%, 2.5%, or 1%, and then we try to determine (at least approximately) a critical value from the standard tables of the test procedures to make our decision. Almost automatically, we also apply the hypothesis testing procedures to the situation where actually multiple decision procedures are required. However, this involves the choice of a number of dependent significance levels without knowing what the overall error rate might be. Also, test procedures do not have the provision to penalize overparameterization since usually an unstructured saturated model is always used as a reference (Akaike, 1987).

On the other hand, when we use the information-theoretic model selection criteria, we do not specify what the arbitrary significance level α should be or ought to be. This is due to the fact that in using model selection criteria, the situation is totally the opposite of the classical inferential procedures. In this case, we are concerned with: Choosing a critical value which then determines, approximately, what the significance level is or might be. Therefore, the significance level is implicitly incorporated within the model selection criteria which depends on the specific functional form of the penalty component of the criteria and on the number of observations.

In general, for model selection criteria, the inferential error rate decreases exponentially as we increase n, the number of observations. Following Efron (1967), suppose we are given n independent and identically distributed (i.i.d.) observations $x_1, x_2, \ldots, x_n$ of a random variable X having probability density function $f(x)$. Suppose we are asked to test the simple hypothesis:

$$\begin{aligned} & H_0\colon\ f(x) \equiv f_0(x) \\ \text{vs.} \qquad & \\ & H_1\colon\ f(x) \equiv f_1(x) \end{aligned} \tag{73}$$

at some significance level α, $0 < \alpha < 1$. It is well known that the most powerful test which rejects H_0 for large values of the likelihood ratio

$$\lambda = \prod_{i=1}^{n} \frac{f_1(x_i)}{f_0(x_i)} \tag{74}$$

has an "error of the second kind," that is, $\beta = P\{\text{Type II error}\}$ (probability of mistakenly

accepting the null hypothesis) satisfying

$$\lim_{n\to\infty} \frac{\beta(\alpha)}{n} = -I, \tag{75}$$

where I is the K-L information quantity

$$I = E\left[\log\left(\frac{f_0(X)}{f_1(X)}\right)\right]. \tag{76}$$

From this, we have the following:

Theorem 3. For a specified level of significance α, $0 < \alpha < 1$, under the null hypothesis in (73), we have

$$\beta(\alpha) = P\{\text{Type II error}\} = \exp\{-nI + O(1)\}. \tag{77}$$

The converse of this result is also true if we try to minimize $\alpha = P\{\text{Type I error}\}$ (probability of mistakenly rejecting the null hypothesis), keeping β only fairly small satisfying

$$\lim_{n\to\infty} \frac{\alpha(\beta)}{n} = -I. \tag{78}$$

So, following Čencov (1982, p. 122), we state the following:

Theorem 4. For a specified β, $0 < \beta < 1$, under the null hypothesis in (73), we have

$$\alpha(\beta) = P\{\text{Type I error}\} = \exp\{-nI + O(1)\}. \tag{79}$$

This means that both α and β cannot tend to zero faster than the exponential rate where $\lim_{n\to\infty} O(1) = 0$. Thus, for example, α is small when the K-L information quantity is large, and vice versa. This rate of convergence to zero of α can also be used as a criterion for evaluating the asymptotic performance of the information-theoretic procedures. Typically, for consistent criterion, this rate is exponential as stated in the above theorems.

To show the Type I error for AIC and CAIC, suppose that the "true' dimension is obtained when $k = k^*$. We consider what happens to the probability of choosing a dimension $k' > k^*$ asymptotically by these criteria.

For AIC, this means

$$\begin{aligned} &P\{\text{AIC}(k') < \text{AIC}(k^*)\} \\ &\quad = P\{2[\log L(\hat{\boldsymbol{\theta}}_{k'}) - \log L(\hat{\boldsymbol{\theta}}_{k^*})] > 2(k' - k^*)\} \\ &\quad \to P\{\chi^2_{(k'-k^*)} > 2(k' - k^*)\} \quad \text{as } n \to \infty \\ &\quad > 0. \end{aligned} \tag{80}$$

For CAIC, this means

$$\begin{aligned} &P\{\text{CAIC}(k') < \text{CAIC}(k^*)\} \\ &\quad = P\{2[\log L(\hat{\boldsymbol{\theta}}_{k'}) - \log L(\hat{\boldsymbol{\theta}}_{k^*})] > (k' - k^*)(\log n + 1)\} \\ &\quad \to P\{\chi^2_{(k'-k^*)} > \infty\} \quad \text{as } n \to \infty \\ &\quad = 0. \end{aligned} \tag{81}$$

Therefore, using AIC it is possible to choose a dimenison $k' > k^*$ (the true dimension), the probability of Type I error stays positive but decreases exponentially as $(k' - k^*)$ gets larger and the critical value (i.e., 2) of the test remains finite $n \to \infty$. So, even asymptotically, there is a positive probability of overestimating the true dimension or the size of the model. Therefore, model selection criteria, including AIC, which have this property are often called *dimension inconsistent.* On the other hand, with CAIC and CAICF, the probability of Type I error goes to zero as $n \to \infty$. It also decreases exponentially as $(k' - k^*)$ gets larger, but at a much faster rate. Therefore, asymptotically, the probability of overestimating the true dimenson or size of the model with these criteria equals zero, making them dimension consistent. (For more on the properties of model selection criteria, we refer the reader to Teräsvirta & Mellin, 1986; and Woodroofe, 1982.)

Without going into detailed proofs, we simply state the following results.

Result 1. Using CAIC or CAICF if $k' < k^*$, where k^* is the true dimension of the model, the probability of underfitting a model goes to zero as $n \to \infty$.

Result 2. Using CAIC or CAICF if $k' \geq k^*$, the probability of overfitting a model disappears as $n \to \infty$.

Note that when we use AIC, CAIC, and CAICF, the "level of significance" is adjusted in such a way that the corresponding probability of rejection of the simpler model decreases as the degrees of freedom or complexity increase. These procedures, therefore, have tendencies to adopt simpler models compared with the chi-square test procedure as the degrees of freedom increase. Comparing to AIC, for CAIC, the implied α values rapidly decrease as the degrees of freedom increase. The same is also true for CAICF. However, this rate of decrease depends on n, the number of observations. For large samples, as the degrees of freedom increase, the implied α values for these criteria decrease very sharply.

This connection between model selection criteria and the level of significance α, provides us a way to test the validity of different restrictions of a model. Also, it gives us a yardstick in comparing every possible model and choosing the model giving the smallest probability of rejection to be the best fitting model. This fact justifies the comparison of the model selection criteria in a class of models which cannot necessarily be compared by the classical goodness of fit test. We can use these results to decide what the level of significance should be per complexity or restriction when we do classical hypothesis testing rather than arbitrarily deciding what α should be on a priori grounds. Thus, in the sense of Theorem 3 and 4, these model selection procedures can be called *inferential-error-rate consistent.*

7. A Numerical Example

In this section, to demonstrate the practical utility and to show the empirical performances of the model selection criteria, we provide a Monte Carlo example in determining the degree of a polynomial model in one variable. Choosing the degree of a polynomial model is a major problem that arises when the relationship between y and x is considered to be *curvilinear.* Fitting a polynomial with the maximum degree K is a multiple decision problem which is often formulated in terms of a sequential hypothesis testing to test whether the coefficients are zero, starting with the highest specified degree.

Following Graybill (1976), we consider the polynomial model of degree K given by

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_K x_i^K + \varepsilon \qquad i = 1, 2, \ldots, n > K + 1, \tag{82}$$

TABLE 7.1

Frequency of Choosing the Correct Degree of a Polynomial Model in 100 Replications of the Monte Carlo Experiment for Varying n and Residual Variance σ^2

Experiment	Criterion	Estimated Degree 1	2	3*	4	5	6	Proportion of Overfitting	Underfitting
n = 50	AIC	0	0	86	14	0	0	.14	0
1. σ^2 = 0.25	CAIC	0	0	99	1	0	0	.01	0
	CAICF	0	0	100	0	0	0	0	0
n = 100	AIC	0	0	81	19	0	0	.19	0
2. σ^2 = 0.50	CAIC	0	0	98	2	0	0	.02	0
	CAICF	0	0	100	0	0	0	0	0
n = 200	AIC	0	0	80	20	0	0	.20	0
3. σ^2 = 1.00	CAIC	0	0	97	3	0	0	.03	0
	CAICF	0	0	100	0	0	0	0	0

NOTE: The true cubic polynomial model is:

$$y = 1+5x-1.25x^2+0.15x^3+\epsilon,\ \epsilon \sim N(0,\sigma^2).$$

*Correct degree.

where K is a *specified* positive integer. We assume that the degree of the polynomial model in (82), say k, is less than or equal to K (given), and the problem is to determine the exact degree. As we mentioned, the procedure is to test the hypothesis H_0: $\beta_K = 0$, then test $\beta_{K-1} = 0$, then test $\beta_{K-2} = 0$, and so on against the alternative of nonzero coefficient until a hypothesis is rejected. Suppose $\beta_k = 0$ is the first hypothesis that is rejected, then we declare that the model of degree k is the correct model. If no H_0 is rejected, then we declare the simple model $y_i = \beta_0 + \varepsilon_i$ is the correct model.

It has been shown by Anderson (1962) that this procedure for determining the degree of a polynomial model has some desirable optimality properties. However, as we know, application of test procedures to such multiple-decision problems involves the choice of a number of dependent significance levels. This creates the problem of how to control the overall error rate of the test procedures for determining the correct degree of a polynomial model. As an alternative to the classical inferential procedures, here we propose the use of model selection procedures to determine the degree of a polynomial model.

To demonstrate this, we carried out a Monte Carlo study under the *true cubic polynomial model* given by

$$y = 1 + 5x - 1.25x^2 + 0.15x^3 + \varepsilon, \tag{83}$$

TABLE 7.2

Frequency of Choosing the Correct Degree of a Polynomial Model in 100 Replications of the Monte Carlo Experiment for Varying n and Same Residual Variance σ^2

Experiment	Criterion	Estimated Degree 1	2	3*	4	5	6	Proportion of Overfitting	Underfitting
	AIC	3	0	73	14	10	0	.24	.03
1. n = 50	CAIC	4	0	81	10	5	0	.15	.04
	CAICF	77	0	23	0	0	0	0	.77
	AIC	0	0	80	11	9	0	.20	0
2. n = 100	CAIC	0	0	93	5	2	0	.07	0
	CAICF	0	0	100	0	0	0	0	0
	AIC	0	0	88	12	0	0	.12	0
3. n = 200	CAIC	0	0	94	6	0	0	.06	0
	CAICF	0	0	100	0	0	0	0	0

NOTE: The true cubic polynomial is:

$$y = 1+5x-1.25x^2+0.15x^3+\varepsilon,\ \varepsilon \sim N(0,\sigma^2),$$

where $\sigma^2 = 5$.

*Correct degree.

where it is assumed that $\varepsilon \sim N(0, \sigma^2)$. In the first design of the Monte Carlo study, we both varied n, the number of observations, and σ^2, the residual variance simultaneously across three different experiments in choosing the correct degree of $k = 3$, and in studying the relative performances of AIC, CAIC, and CAICF by fitting the polynomial models of degree ranging from 1 to 6. In the second design of the Monte Carlo study, we increased the residual variance σ^2, but kept it the same and varied the sample size n across three different experiments in studying the relative performances of AIC, CAIC, and CAICF.

All the computations are carried out using our POLYREG algorithm in double precision on one of the PRIME 750 computers of the University of Virginia. The results are given in Table 7.1 and 7.2 as follows.

Looking at Table 7.1, we see that AIC has the tendency to overfit the correct degree of the polynomial as the sample size n gets large and the residual variance σ^2 varies. This suggests that a stronger penalty for model complexity could be beneficial as we discussed before. For CAIC, there appears to be a slight tendency to overfit the correct degree of the polynomial model, but this is very insignificant. On the other hand, across all the three

experiments, the condition for the consistency of order determination by CAICF holds perfectly.

Looking at Table 7.2, we note that the results are different from that of Table 7.1. This is due to the fact that we increased the residual variance σ^2 to 5 in 100 replications of the Monte Carlo Experiment. For $n = 50$, AIC and CAIC are performing much better than CAICF in choosing the correct degree ($k = 3$) of the polynomial model. In fact, the proportion of underfitting for CAICF is 77%. However, for CAICF, the underfitting diminishes as n, the sample size, gets large and the condition of consistency holds. These results suggest that when n is small and residual variance σ^2 is large, CAICF might have the tendency to underfit the true order of a polynomial model. But, as we saw, this behavior of CAICF disappears as n gets large. We emphasize the fact that this was the only extreme example for CAICF when $n = 50$ among many repetitions of the Monte Carlo experiment.

Overall, we observe that AIC, CAIC, and CAICF are powerful tools to determine the best fitting model. They are superior to classical inferential methods in terms of their computational simplicity, in terms of not arbitrarily specifying a significance level α, and not worrying what the overall inferential error rate might be or ought to be in determining the degree of a polynomial model, and in general.

8. Conclusions and Discussion

In this paper, we studied the general theory of the AIC procedure, presented its mathematical derivation and showed the emergence of the magic number 2 in its derivation. We provided analytical extensions of AIC in two ways without violating Akaike's main principles which make AIC asymptotically consistent and penalize overparameterization more stringently rather than relying on the heuristic or arbitrary modifications.

We investigated the asymptotic properties of AIC and its analytical extensions. We studied the inferential error rates of these procedures for testing the validity of different complexities. The preference of one or the other of these criteria in a given situation depends on how "conservative" or "liberal" we want to be in terms of setting the level of significance α per complexity and avoid overfitting and underfitting risks.

If we want to avoid overfitting a model, then we should use the consistent criteria: CAIC and CAICF, sometimes at the cost of underfitting a model in finite samples, which leads to a significant increase in bias. Of course, as the number of observations gets large, for the consistent criteria, the probability of underfitting and overfitting a model will diminish. This suggests that one should use these consistent criteria for large samples. If we want to avoid underfitting a model, then we should use AIC.

There is no single criterion which will play the role of a panacea in model selection problems. Presently, however, AIC has become part of a general movement away from a purely inferential and restrictive approach to model selection. As a consequence, it provided us a new and modern way of thinking of how to tackle many important statistical modeling problems. For this reason, the profession is greatly in debt to Akaike for repeatedly calling our attention to the very important model evaluation and selection problem.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second International Symposium on Information Theory*, (pp. 267–281). Academiai Kiado: Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716–723.

Akaike, H. (1976). Canonical correlation analysis of time series and the use of an information criterion. In R. K. Mehra & D. G. Lainiotis (Eds.), *System identification* (pp. 27–96). New York: Academic Press.

Akaike, H. (1977). On entropy maximization principle. In P. R. Krishnaiah (Ed.), *Proceedings of the Symposium on Applications of Statistics* (pp. 27–47). Amsterdam: North-Holland.

Akaike, H. (1978). On newer statistical approaches to parameter estimation and structure determination. *International Federation of Automatic Control, 3*, 1877–1884.

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autogressive model fitting. *Biometrika, 66*, 237–242.

Akaike, H. (1981a). Likelihood of a model and information criteria. *Journal of Econometrics, 16*, 3–14.

Akaike, H. (1981b). Modern development of statistical methods. In P. Eykhoff (Ed.), *Trends and progress in system identification* (pp. 169–184). New York: Pergamon Press.

Akaike, H. (1987). Factor Analysis and AIC. *Psychometrika, 52*.

Anderson, T. W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Annals of Mathematical Statistics, 33*, 255–265.

Atilgan, T. (1983). *Parameter parsimony, model selection, and smooth density estimation.* Unpublished doctoral dissertation, Madison: University of Wisconsin, Department of Statistics.

Atilgan, T., & Bozdogan, H. (1987, June). Information-theoretic univariate density estimation under different basis functions. A paper presented at the First Conference of the International Federation of Classification Societies, Aachen, West Germany.

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika, 67*, 413–418.

Bhansali, R. J., & Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika, 64*, 547–551.

Boltzmann, L. (1877). Über die Beziehung zwischen dem zweitin Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respective den Sätzen über das Wärmegleichgewicht. *Wiener Berichte, 76*, 373–435.

Čencov, N. N. (1982). *Statistical decision rules and optimal inference.* Providence, RI: American Mathematical Society.

Clergeot, H. (1984). Filter-order selection in adaptive maximum likelihood estimation. *IEEE Transactions on Information Theory, IT-30* (2), 199–210.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics.* London: Chapman and Hall.

Davis, M. H. A., & Vinter, R. B. (1985). *Stochastic modelling and control.* New York: Chapman and Hall.

Efron, B. (1967). The power of the likelihood ratio test. *Annals of Mathematical Statistics, 38*, 802–806.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Royal Society of London. Philosophical Transactions* (Series A), *222*, 309–368.

Graybill, F. A. (1976). *Theory and application of the linear model.* Boston: Duxbury Press.

Hannan, E. J. (1986). Remembrance of things past. In J. Gani (Ed.), *The craft of probabilistic modelling.* New York: Springer-Verlag.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, (Series B), *41*, 190–195.

Haughton, D. (1983). On the choice of a model to fit data from an exponential family. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Department of Mathematics, Cambridge, MA.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review, 106*, 620–630.

Kashyap, R. L. (1982). Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 4*, 99–104.

Kendall, M. G., & Stuart, M. A. (1967). *The Advanced Theory of Statistics, Vol. 2, Second Edition.* New York: Hafner Publishing.

Kitagawa, G. (1979). On the use of AIC for the detection of outliers. *Technometrics, 21*, 193–199.

Kullback, S. (1959). *Information theory and statistics.* New York: John Wiley & Sons.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79–86.

Larimore, W. E., & Mehra, R. K. (1985, October). The problems of overfitting data. *Byte*, pp. 167–180.

Lindley, D. V. (1968). The choice of variables in multiple regression (with discussion). *Journal of the Royal Statistical Society* (Series B), *30*, 31–36.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20A*, 175–240 (Part I), 263–294 (Part II).

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Royal Society of London. Philosophical Transactions.* (Series A), *231*, 289–337.

Parzen, E. (1982). Data modeling using quantile and density-quantile functions. In J. T. de Oliveira & B. Epstein (Eds.), *Some recent advances in statistics* (pp. 23–52). London: Academic Press.

Quinn, B. G. (1980). Order determination for a multivariate autoregression. *Journal of the Royal Statistical Society* (Series B), *42*, 182–185.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

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

Shibata, R. (1983). A theoretical view of the use of AIC. In O. D. Anderson (Ed.), *Time series analysis: Theory and practice, Vol. 4* (pp. 237–244). Amsterdam: North-Holland.

Silvey, S. D. (1975). *Statistical inference*. London: Chapman and Hall.

Stone, C. J. (1981). Admissible selection of an accurate and parsimonious normal linear regression model. *Annals of Statistics, 9*, 475–485.

Teräsvirta, T., & Mellin, I. (1986). Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics, 13*, 159–171.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54*, 426–482.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*, 1–26.

Wilks, S. S. (1962). *Mathematical Statistics*. New York: John Wiley & Sons.

Woodroofe, M. (1982). On model selection and the arc sine laws. *Annals of Statistics, 10*, 1182–1194.