A Review of some Aspects of Information Criteria for Model Selection

Dominique M.A Haughton

Bentley College

Abstract

The introduction in 1973 of a model selection criterion by Akaike (AIC) initiated much research activity on the behavior and properties of a class of criteria generally known as information criteria. This paper expands on a lecture on this subject given at the March 1995 annual meeting of the Dutch Statistical Society and focuses on issues which have been the object of active discussions in the structural equation modelling literature. We briefly explain the rationale behind information criteria such as the AIC, the BIC and related criteria for choosing a model to fit a data set among a finite set of alternatives. We also briefly describe model selection criteria based on the concept of the complexity of a covariance matrix. We review some results on the asymptotic behavior of the criteria as the sample size goes to infinity both in the case when at least one of the analytic models contains the true parameter vector, and the case when none do. Finally, we illustrate the use of information criteria with an application to the selection of a harmonic regression model.

Information, Kullback-Leibler distance and information criteria.

We recall here the concept of **Kullback-Leibler distance**, following Kullback (1959). Let $f_1(x)$ and $f_2(x)$ be two probability densities. Consider a random variable X whose density is either f_1 or f_2. Call H_1 (respectively H_2) the hypothesis that the density of X is f_1 (respectively f_2). Given an observation x of X, we can apply Bayes' theorem to get the posterior probability that the hypothesis H_1 (respectively H_2) holds:

$$P(H_i|x) = \frac{f_i(x)P(H_i)}{f_1(x)P(H_1) + f_2(x)P(H_2)}, \tag{1}$$

for $i = 1, 2$. It follows that:

$$\log[\frac{f_1(x)}{f_2(x)}] = \log[\frac{P(H_1|x)}{P(H_2|x)}] - \log[\frac{P(H_1)}{P(H_2)}].$$

So the expression $\log[\frac{f_1(x)}{f_2(x)}]$ can be considered as the information in the observation x for discriminating in favor of H_1 against H_2. If x is an observation from the density f_1, then the mean information for discriminating in favor of H_1 against H_2 per observation from f_1 is:

$$I(f_1, f_2) = E_{f_1} \log[\frac{f_1(x)}{f_2(x)}],$$

where E_{f_1} denotes the expected value with respect to the density f_1. In general, the **Kullback-Leibler distance between two densities f_1 and f_2 is defined to be** $I(f_1, f_2)$.

Note that, in spite of its name, the Kullback-Leibler distance is not a distance strictly speaking because it is not symmetric: in general, $I(f_1, f_2) \neq I(f_2, f_1)$. But the two following properties hold:

- $I(f_1, f_2) \geq 0$
- $I(f_1, f_2) = 0$ if and only if $f_1 = f_2$ almost everywhere in the possible range of x.

Let us now turn to the problem of model selection. Consider a sequence $X_1, X_2, \cdots, X_n$ of independent identically distributed (i.i.d.) random variables with unknown density $f_{\theta^*} = f(x, \theta^*)$, where the unknown parameter vector θ^* belongs to d-dimensional space. In general, we will call a **model** a subset of d-dimensional space, to which the unknown parameter vector is restricted. Let us assume that we have at our disposal a finite number of competing models $m_1, m_2, \cdots, m_J$, and that our task is to select a model which fits the data well and is parsimonious. We define the **minimal true model** to be the smallest model which contains the true parameter vector θ^*. We devote our attention to the problem of trying to determine what this minimal true model is, and examine several criteria for model selection generally referred to as **information criteria**.

In 1973 Akaike introduced the *AIC* criterion, which is motivated as follows. Given a model where the vector θ of parameters is allowed to vary, a reasonable goal would be to identify the value of θ such that the distance $I(f(x,\theta^*), f(x,\theta))$ is minimum. Since

$$I(f_{\theta^*}, f_\theta) = E_{\theta^*}[\log f(x,\theta^*)] - E_{\theta^*}[\log f(x,\theta)],$$

where E_{θ^*} denotes the expected value with respect to $f(x,\theta^*)$, in order to minimize $I(f_{\theta^*}, f_\theta)$, we should maximize $E_{\theta^*}[\log f(x,\theta)]$. Of course this latter quantity depends on the unknown parameter vector θ^* so cannot be calculated directly. By the strong law of large numbers, we know that, almost surely, as the number n of observations goes to infinity, the sample mean

$$\frac{1}{n}\sum_{i=1}^{n} \log f(X_i, \theta)$$

converges to $E_{\theta^*}[\log f(x,\theta)]$. So one might think of trying to maximize

$$\sum_{i=1}^{n} \log f(X_i, \theta).$$

This idea leads directly to the concept of maximum likelihood. Let us assume that on each model m_j, $j = 1 \cdots J$ we have a maximum likelihood estimator $\hat{\theta}_j$ which maximizes the log-likelihood $\sum_{i=1}^{n} \log f(X_i, \theta)$ on m_j. Recall the problem we mentioned above of maximizing $E_{\theta^*}[\log f(x,\theta)]$. This of course is equivalent to maximizing the quantity

$$l^*(\theta) = nE_{\theta^*}[\log f(x,\theta)].$$

Consider the quantity $l^*(\hat{\theta}_j)$. This quantity depends on the random variables $X_1, \cdots, X_n$ through $\hat{\theta}_j$. Following Sakamoto, Ishiguro and Kitagawa (1986), we define the **mean expected log likelihood** on m_j to be

$$E_{\theta^*(n)}[l^*(\hat{\theta}_j)],$$

where $E_{\theta^*(n)}$ denotes the expected value with respect to the product of n densities of the form $f(x,\theta^*)$. One can then prove (Akaike (1973), Sakamoto, Ishiguro, Kitagawa (1986); see also Bozdogan (1987)) that, under some conditions, an approximately unbiased estimator of the mean expected log-likelihood on m_j when the sample size n is large is provided by:

$$AIC(m_j) = \text{maximum log likelihood on } m_j - k_j,$$

where k_j is the dimension of the model m_j (i.e. the number of free parameters in m_j). To choose a model by means of the AIC criterion, we select m_j so that $AIC(m_j)$ is largest. We note here that the objective of AIC is to find a model which maximizes the **mean expected log likelihood**, which is a somewhat different objective from searching for the **minimal true model**.

Following the introduction of the AIC by Akaike, a lot of attention has been given to the development and properties of criteria which resemble the AIC, although their motivation might be quite different. Because of the informational nature of AIC, such criteria are very often referred to as **information criteria**. Below we summarize several information criteria, with a brief description of the rationale behind each.

The BIC criterion was introduced by Schwarz in 1978 and arises as leading terms in an asymptotic expansion of the (logarithm of the) posterior probability that a model is the minimal true model. The point of view here is that each model has a prior probability (such as for example $P(H_1)$ and $P(H_2)$ in equation (1)) and defines a prior distribution for the parameters. Given those priors, a Bayesian approach to model selection would select that model with the highest posterior probability of being the minimal true model. The criteria BIC, BIC^* and $BICR$, which we are about to define, are designed to approximate this Bayesian approach. Indeed, under some regularity assumptions, as well as some assumptions on the models and the priors (see Schwarz (1978), Haughton (1988), Dudley and Haughton (1995)), asymptotic expansions of the (logarithm) of the posterior probability that a model is the true minimal model can be derived, and the expressions for the BIC, BIC^*, and $BICR$ given below appear in the expansion. Note that the BIC contains terms of order n or $\log n$, while BIC^* and $BICR$ contain terms of constant order as well. The BIC, to be maximized to select a model, is defined as:

$$BIC(m_j) = \text{maximum log likelihood on } m_j - \frac{1}{2}k_j \log n.$$

Extensions of the BIC, to more general models and including more terms in the posterior probability that a model is the minimal true model, have been developed notably by Haughton (1988), Haughton, Haughton and Izenman (1994) and by Dudley and Haughton (1995). These include:

$$BIC^*(m_j) = \text{maximum log likelihood on } m_j - \frac{1}{2}k_j \log \frac{n}{2\pi},$$

and,

$$BICR(m_j) = \text{max log lik on } m_j - \frac{1}{2}k_j \log \frac{n}{2\pi} + \log f_{prior}(\hat{\theta}_j) + \frac{1}{2} \log \det IFIM(\hat{\theta}_j),$$

where f_{prior} denotes a prior density for the unknown parameter vector on the model m_j, and $IFIM$ denotes the inverse Fisher information matrix calculated in terms of the free parameters on the model m_j (see Dudley and Haughton (1995)). Note that BIC and BIC^* do not depend on the priors. The prior density f_{prior} in $BICR$ must be non-zero on the model m_j and smooth. If not much information is known a-priori about where the parameter vector is likely to be in model m_j, a **non-informative prior** might be desirable (see, for example, Box and Tiao (1973)).

Introducing a $\log n$ term to obtain a consistent criterion (see Bozdogan (1987) p. 358), Bozdogan defined the criterion:

$$CAIC(m_j) = \text{maximum log likelihood on } m_j - \frac{1}{2}k_j(\log n + 1).$$

Note that $\log n$ could be replaced by any sequence a_n which converges to infinity as n goes to infinity, and such that a_n/n converges to zero as n goes to infinity; the consistency of the criterion would be preserved (see e.g. Bozdogan (1987), Haughton (1988)).

A criterion with $a_n = \log\log n$ was proposed by Hannan and Quinn (1979) in the context of autoregressive (AR) models. In that paper, Hannan and Quinn show that $\log\log n$ is the sequence with the slowest possible rate of increase which still makes the criterion consistent.

Geweke and Meese (1981) propose a model selection criterion BEC (Bayesian Estimation Criterion Function) for linear regression models, to be minimized to select a model, which adds a penalty function to the maximum likelihood estimator

$\hat{\sigma}^2$ of the unknown error variance (note that the BIC criterion amounts to adding a penalty function to $\log \hat{\sigma}^2$, and that $\log \hat{\sigma}^2$ equals $-2/n$ times the maximum log likelihood, plus a constant). The BEC and BIC are found to perform about equally well at the task of retrieving the minimal true model in the simulation study conducted by Geweke and Meese (the study assumes that the true parameter vector lies in at least one of the analytic models). Teräsvirta and Mellin (1986) follow up on the simulation study of Geweke and Meese in order to compare criteria such as AIC or BIC with sequences of tests of hypotheses. The BIC is found to perform well at the task of retrieving the minimal true model, and the test sequences are found to be a viable competitor. Teräsvirta and Mellin note that the choice of a significance level for the test would however be crucial. Sawa (1981) proposes a criterion for regression model selection which adds a penalty to $\log \hat{\sigma}^2$ (as AIC and BIC do), is inspired by AIC principles, and does not assume that at least one of the analysis models contains the true parameter vector.

We also note that corrections to the AIC for small sample sizes have been proposed by Hurvich and Tsai (1989) and Hurvich, Shumway and Tsai (1990) for regression and time series model selection.

A new direction in the area of information criteria is that of complexity-based information criteria, introduced by Bozdogan (1990). The idea is that the number of parameters k_j may be too simple a measure of the complexity of a model. Bozdogan (1990) introduced a new class of criteria called the $ICOMP$, of the following form:

$$ICOMP(m_j) = \text{maximum log likelihood on } m_j - \text{ complexity of the model },$$

where complexity is defined as follows. Given a k by k positive definite matrix A, the complexity of A is given by (Van Emden (1971), Bozdogan (1990)) :

$$C(A) = \frac{k}{2}\log[\frac{trace(A)}{k}] - \frac{1}{2}\log[det(A)].$$

Note that $C(A)$ is equal to the Kullback-Leibler distance between the joint density and the product of the marginal densities for the components of a normal vector with covariance matrix A, maximized over all orthogonal transformations of that vector (see Bozdogan (1990)). One can show that $C(A) \geq 0$ for every positive definite matrix A and that $C(A) = 0$ if and only if A is a multiple of the identity matrix.

A rationale for combining a maximum likelihood on a model with the complexity defined above is given in Bozdogan and Haughton (1995): basically $ICOMP$ seeks to minimize the (approximated) sum of two Kullback-Leibler distances, one of which is a measure of the **badness of fit** of the model, and the second of which measures how much dependence there is among the components of the estimated parameter vector.

Two approaches to the $ICOMP$ have been introduced and investigated in a variety of situations by Bozdogan. If a model gives rise to an estimated parameter vector $\hat{\theta}$ and estimated residuals $\hat{\epsilon}$, as for example in linear (or non-linear) regression, the $ICOMP$ can be defined as:

$$ICOMP(m_j) = \text{maximum log likelihood on } m_j - [C(\hat{\Sigma}_{\hat{\theta}}) + C(\hat{\Sigma}_{\hat{\epsilon}})],$$

where $\hat{\Sigma}_{\hat{\theta}}$ is the estimated covariance matrix of the estimated parameter vector and $\hat{\Sigma}_{\hat{\epsilon}}$ is the estimated covariance matrix of the estimated errors. Note that $C(\hat{\Sigma}_{\hat{\theta}}) + C(\hat{\Sigma}_{\hat{\epsilon}})$ measures the (approximate) amount of dependence among the components of the the vector $\hat{\theta}$ added to the (approximate) amount of dependence among the components of $\hat{\epsilon}$. Another version of the $ICOMP$ can be defined as:

$$ICOMP_{IFIM}(m_j) = \text{maximum log likelihood on } m_j - C(IF\hat{I}M),$$

where $IF\hat{I}M$ is the estimated inverse Fisher information matrix, equal to the estimated asymptotic covariance matrix of the estimated parameter vector. This second definition of the $ICOMP$ can be applied to any modelling situation where suitable regularity conditions hold to ensure that $IF\hat{I}M$ is the estimated asymptotic covariance matrix of the estimated parameter vector.

Asymptotic behavior of AIC and BIC.

We now explain what the asymptotic behaviors of AIC and BIC are as the sample size n gets large. Consider two models m_1 and m_2. The following holds (see Haughton (1989)):

If θ is in m_1 but not in m_2, then, almost surely, for n large enough, BIC selects m_1, and AIC selects m_1.

This property follows from the fact that almost surely (with probability 1), for n large enough, we have

$$\text{max log likelihood on } m_1 - \text{ max log likelihood on } m_2 > cn$$

for a positive constant c. This means that the difference in log likelihoods will overcome the AIC or BIC difference in penalty functions (see Haughton (1989)) when n goes to infinity. It implies that **for both AIC and BIC (as well as BIC^*, $BICR$, and $CAIC$), the probabilities of underfitting converge to 0 as n goes to infinity.** In fact, it is actually true that the probabilities of underfitting become less than or equal to $Ce^{-\alpha n}$ as n becomes large, where C and α are positive constants (see Haughton (1989)). So the probabilities of underfitting are likely to become small quickly as n becomes large.

We now look at the question of overfitting. Consider a situation in which θ is in both m_1 and m_2 and the dimension of m_2 is less than the dimension of m_1. The key fact here is that, by the law of the iterated logarithm, one can show (see Haughton (1989)) that almost surely for n large enough:

$$|\text{max log likelihood on } m_1 - \text{ max log likelihood on } m_2| \leq C \log \log n,$$

for a positive constant C. So the BIC penalty, because it contains a $\log n$ term, will overcome the difference in maximum log likelihoods as n becomes large. This implies that **the probability that BIC overfits converges to zero as n goes to infinity (the same is true of BIC^*, $BICR$, and $CAIC$).** The fact that the probabilities of underfitting or overfitting with BIC converge to zero as n goes to infinity is called **consistency**. Regarding overfitting, it is in fact true that the probability that BIC overfits becomes less than or equal to C/n^k as n goes to infinity, where C and k are positive constants (see Haughton (1989)).

With respect to overfitting, the AIC behaves quite differently from the BIC. The AIC penalty function, which does not depend on the sample size, may not be able to overcome the difference in maximum log likelihoods. In fact, it is known in a variety of situations that the probability that AIC overfits converges to a constant probability, as n goes to infinity. The limiting overfitting probability will depend on the competing models. For the case of regression models, we refer the reader to Nishii (1984). Limiting overfitting probabilities are also given by Woodroofe (1982), under some mild regularity conditions. There it is assumed that a set of nested models is given, with m_0 of dimension 0, m_1 of dimension 1, $\cdots$, m_J of dimension J. For example, if the true dimension is 5, and the maximum dimension is 10, the asymptotic probability that AIC overfits is .265 (see Woodroofe, 1982).

So, **the asymptotic probability that AIC overfits is in general a positive (but less than 1) probability which depends on the models.**

To summarize, when one of the entertained models is the true model, asymptotically,

- The AIC will overfit, but only some of the time,
- The BIC will overfit less and less often as the sample size increases.

A question arises: **what might the asymptotic behavior of the AIC and BIC be if none of the entertained models is the true model?** This is a common situation, for example in regression models, where one is likely to be missing a variable. It is also felt to often be the case for structural equations modelling (see, e.g., McDonald (1989)). A recent simulation study by Bozdogan and Haughton (1995), to be discussed later in this paper, involves a true linear regression model M_5 with five variables, and four analytic models M_1, M_2, M_3, and M_4 with 1, 2, 3 and 4 variables respectively. It is shown in that paper that, as the sample size goes to infinity, the probabilities approach one that the AIC (or the BIC) selects the largest model M_4. However, for finite sample sizes (even large ones, such as 1000), **sometimes the AIC (and the BIC) will try to choose larger and larger models as n increases, but sometimes they will not.**

In order to illustrate the asymptotic underfitting and overfitting properties of AIC and BIC, we give a simple example where probabilities can be evaluated analytically.

Example: Let $X_1, X_2, \cdots, X_n$ be i.i.d. according to a $N(0,1)$. Consider the two models:

$$m_0 : \mu = 0,$$

and

$$m_1 : \mu \text{ arbitrary},$$

where μ is the mean of the normal distribution of the X_i. Note that the standard deviation is one for both models m_0 and m_1. It is then easy to calculate:

$$\text{max log likelihood on } m_1 - \text{ max log likelihood on } m_0 = \frac{n}{2}\bar{X}^2,$$

where $\bar{X}$ is the sample mean of the observations $X_1, \cdots, X_n$. Then we have:

$$P(AIC \text{ overfits}) = P(\frac{n}{2}\bar{X}^2 > 1) = 2P(\sqrt{n}\bar{X} > \sqrt{2}) = 2(1 - \Phi(\sqrt{2})) \approx .158,$$

where Φ is the cumulative distribution function for the standard normal $N(0,1)$.

So, if we simulated $N(0,1)$ data and used AIC to select between m_0 and m_1, we would expect AIC to overfit (select m_1) about 15.8% of the time. Note that in this simple case the probability .158 does not depend on the sample size.

On the other hand,

$$P(\text{BIC overfits}) = P(\frac{n}{2}\bar{X}^2 > \frac{1}{2}\log n) = 2P(\sqrt{n}\bar{X} > \sqrt{\log n}) = 2(1 - \Phi(\sqrt{\log n})),$$

which converges to zero as n goes to infinity. In fact one can show by using an expression which is asymptotically equivalent to $1 - \Phi(\sqrt{\log n})$ that:

$$P(BIC(m_1) > BIC(m_0)) \sim \frac{C}{\sqrt{\log \log n}},$$

as n goes to infinity, where C is a positive constant (see Dudley, 1989, p. 352). Here $\sim$ means that the ratio of the left-hand-side to the right-hand-side converges to one as n goes to infinity.

Let us note here that the concept of consistency we have discussed represents one among several possible desirable asymptotic features of model selection criteria. Another type of analysis (see Shibata, 1980, 1981) assumes that the number of variables is infinite or increases with the sample size (in linear regression), or that the order is infinite (for autoregressive models). Shibata proposes a criterion which adds a penalty function to the estimated variance $\hat{\sigma}^2$ of the random errors in a linear regression model or an autoregressive model and is asymptotically efficient, in the sense that it asymptotically minimizes the mean square error of an estimated predictor (see Shibata, 1980, 1981). Shibata's criterion is asymptotically equivalent to AIC, so AIC is asymptotically efficient (and BIC is not) in the sense of Shibata.

Regression models when none of the entertained models is the true model: a simulation.

In a recent simulation study by Bozdogan and Haughton (1994), the following true model was considered:

$$Y_i = \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \beta_5 X_{i,5} + \epsilon_i,$$

where $i = 1, \cdots, n$. So the true model is a linear regression model with 5 variables, and the ϵ_i are assumed to be i.i.d. according to a $N(0, \sigma^2)$. In the analysis, we only entertained the following 4 models:

M_1: model with the variable X_1 only,

M_2: model with the two variables X_1 and X_2 only,

M_3: model with the three variables X_1, X_2, X_3 only,

M_4: model with the four variables X_1, X_2, X_3, and X_4.

Note that in this case there is no such thing as the smallest model which contains the true parameter vector. Instead we define as **best model** that estimated model whose Kullback-Leibler distance to the true model is smallest. So, in the same spirit as the AIC, we consider that minimizing the Kullback-Leibler distance is an ideal procedure, which of course cannot be implemented in reality since we do not know the true model. Several interesting conclusions emerge from the simulation study:

- While the probabilities converge to one (as the sample size goes to infinity) that the Kullback-Leibler distance selects model M_4 (see Bozdogan and Haughton (1995)), for **finite** sample sizes (even large, such as 1000), minimizing the Kullback-Leibler distance between each estimated model and the true model does not always choose the largest available model M_4.
- Complexity based criteria ($ICOMP$) tend to agree with Kullback-Leibler decisions more often than the AIC or BIC.
- While the probabilities converge to one (as the sample size goes to infinity) that the AIC (or BIC) selects model M_4, for **finite** sample sizes (even large, such as 1000), the AIC (or BIC) does not always select the largest model M_4.

We refer the reader to Bozdogan and Haughton (1995) for more details.

Application: harmonic regression models.

Consider observations $Y(1), Y(2), \cdots, Y(n)$ with

$$Y(t) = \alpha_0 + \sum_{j=1}^{k} [\alpha_j \cos(2\pi f_j t) + \beta_j \sin(2\pi f_j t)] + \epsilon(t),$$

where the two first terms, involving unknown parameters $\alpha_0, \alpha_1, \cdots, \alpha_k, \beta_1, \cdots, \beta_k$, $f_1, \cdots, f_k$ make up a non-random sinusoidal regression function with frequencies $f_1, \cdots, f_k$ in $[0, 1/2]$, and the random errors $\epsilon(t)$ are i.i.d. according to a $N(0, \sigma^2)$ with an unknown variance σ^2. A model such as we have just described is often referred to as a **harmonic regression model**; when, as we assume here, the frequencies are unknown, the harmonic regression model is a non-linear model. Harmonic regression models are suitable for data with some periodicity built in, such as temperature data, for example. The question of interest to us is: **How many frequencies are there really in the data? In other words, what is** k?

Techniques to fit harmonic data are presented in Haughton, Haughton and Izenman (1994), where information criteria as well as complexity based criteria are applied to the problem of deciding how many frequencies to include in models to fit two well known series - the signed sunspot numbers, the magnitudes of a variable star - as well as a series of Budapest temperatures. The paper also presents a simulation study involving the following true model:

$$Y(t) = 15 + 9\cos(2\pi.1t) + 6\sin(2\pi.1t) +$$

$$3\cos(2\pi.34t) - 2\sin(2\pi.34t) + 1.5\cos(2\pi.35t) + 5\sin(2\pi.35t) + \epsilon(t).$$

Variances σ^2 of .25, 1 and 4, sample sizes of 50, 100 and 1000 were used in the simulations, and 100 replications were run in all experiments.

Overall, the following conclusions emerged:

• The simulations show that the model fitting and model selection procedures work quite well for sample sizes of at least 100 (for all variances considered). For fewer observations (50 for example) the procedures are less successful, notably at distinguishing between the two close frequencies .34 and .35 in the simulation study.

• A parsimonious model with 6 frequencies was obtained for the sunspot data by using BIC with its penalty function multiplied by 2. If the BIC penalty function is multiplied by 4, a model with two frequencies is selected.

• Even though two frequencies, .034483, and .041667 explain 99.9% of the variability of the star magnitude data, the information criteria tend to select larger numbers of frequencies, such as 16 or 21 frequencies.

• In general, the criteria selected one frequency, the yearly cycle, to fit the Budapest temperature data. Our estimation procedure gave a very precise identification of this yearly cycle (12.0003 months).

We refer the reader to Haughton, Haughton and Izenman (1994) for more details. Details are given in that paper on an iterative procedure which allows to decide which "candidate" frequencies to consider for inclusion in harmonic regression models. One key ingredient of the procedure is an **Amplitude Density Function**, which behaves a little like a periodogram: the location of its peaks helps select "candidate" frequencies.

Conclusion.

Let us first note that, while the area of model selection is a large area of mathematical statistics with an extensive literature, this paper is not meant to give an exhaustive review of all techniques and results. Instead, one of our objectives was to help clarify certain questions, notably about the asymptotic behavior of information criteria, which have arisen in the structural equation modelling literature.

A particular focus was given to the *probabilistic* aspect of some of the results: for example, AIC has a certain asymptotic probability of overfitting which is neither zero or one in most cases where at least one of the analytic models contains the true parameter vector. We have also shed some light on what happens when none of the analytic models contain the true parameter: in that situation, while the asymptotic probability of selecting the largest available model may equal one for the AIC (or BIC) in some cases such as the regression set-up presented here, that does not imply that in practice larger sample sizes will yield larger models. It is quite likely that in many situations, even when none of the analytic models contain the true parameter vector, the asymptotic probability of selecting the largest available model will in fact not equal one (for the AIC or the BIC).

References.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Csaki (Eds.), *2nd International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

Box, G., & Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*, Wiley Classics, Wiley, New-York.

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

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics, Part A: Theory and Methods*, *19*, 221-278.

Bozdogan, H. , & Haughton, D. (1995). *Informational complexity criteria for regression models.* Preprint, Department of Mathematical Sciences, Bentley College, 175 Forest St., Waltham, MA 02154, USA.

Dudley, R. (1989). *Real Analysis and Probability*, Wadsworth, Pacific Grove, CA.

Dudley, R., & Haughton, D. (1995). Information criteria for multiple data sets and restricted parameters. Preprint, Department of Mathematical Sciences, Bentley College, 175 Forest St., Waltham, MA 02154, USA.

Geweke, J., & Meese, R. (1981). Estimating regression models of finite but unknown order. *International Economic Review*, *22*, 55-70.

Hannan, E., and B. Quinn (1979). The determination of the order of an autoregression. *J. R. Statist. Soc.*, *41*, 190-195.

Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist.*, *16*, 342-355.

Haughton, D. (1989). Size of the error in the choice of a model to fit data from an exponential family. *Sankhyā: The Indian Journal of Statistics, Series A, 51*, 45-58.

Haughton, D., Haughton, J., & Izenman, A.J. (1994). Model selection in harmonic non-linear regression. *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: an Informational Approach*, Kluwer Academic Publisher, Dordrecht, 187-207.

Hurvich, C., & Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297-307.

Hurvich, C., Shumway, R., and Tsai, C.L. (1990). Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. *Biometrika, 77*, 709-719.

Kullback, S. (1959). *Information Theory and Statistics*, Wiley, New-York.

McDonald, R.P. (1989). An index of goodness-of-fit based on noncentrality, *J. of Classification, 6*, 97-103.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist., 12*, 758-765.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike Information Criterion Statistics*, KTK Scientific Publishers, Tokyo/Kluwer Academic Publishers, USA.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica, 6*, 1273-1291.

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

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist., 8*, 147-164.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika, 68*, 45-54.

Teräsvirta, T., & Mellin, I. (1986). Model selection criteria and model selection tests in regression models. *Scand. J. Statist., 13*, 159-171.

Van Emden, M.H. (1971). An analysis of complexity. Amsterdam: Mathematical Centre Tracts 35.

Woodroofe, M. (1982). On model selection and the arc sine laws. *Ann. Statist., 10*, 1182-1194.

Acnowledgements: I would like to express sincere appreciation to Dr. Johan Oud, the Dutch Statistical Society, NWO, and the University of Nijmegen for inviting me to present this paper at one of the annual meetings of the Dutch Statistical Society, and for their kind hospitality. I would also like to thank the anonymous referee for many useful suggestions. Research for this paper was supported by a National Science Foundation Grant DMS9210131.

Ontvangen: 1 juli 1995
Geaccepteerd: 1 maart 1996