International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

# Statistical Aspects of Model Selection

**Shibata, R.**

**IIASA Working Paper**

**WP-89-077**

**October 1989**

# WORKING PAPER

## STATISTICAL ASPECTS OF MODEL SELECTION

*R. Shibata*

October 1989
WP-89-077

IIASA
International Institute
for Applied Systems Analysis

# STATISTICAL ASPECTS OF MODEL SELECTION

*R. Shibata*

October 1989
WP-89-077

Department of Mathematics, Keio University, Yokohama, Japan



INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

## FOREWORD

This is a contribution to the activity on the topic *From Data to Model* initiated at the Systems and Decision Sciences Program of IIASA by Professor J. C. Willems.

A. Kurzhanski
Program Leader
System and Decision Sciences Program.

# STATISTICAL ASPECTS OF MODEL SELECTION

RITEI SHIBATA

**Abstract**

Various aspects of statistical model selection are discussed from the view point of a statistician. Our concern here is about selection procedures based on the Kullback Leibler information number. Derivation of AIC (Akaike's Information Criterion) is given. As a result a natural extension of AIC, called TIC (Takeuchi's Information Criterion) follows. It is shown that the TIC is asymptotically equivalent to Cross Validation in a general context, although AIC is asymptotically equivalent only for the case of independent identically distributed observations. Next, the maximum penalized likelihood estimate is considered in place of the maximum likelihood estimate as an estimation of parameters after a model is selected. Then the weight of penalty is also the one to be selected. We will show that, starting from the same Kullback–Leibler information number, a useful criterion RIC (Regularization Information Criterion) is derived to select both the model and the weight of penalty. This criterion is in fact an extension of TIC as well as of AIC. Comparison of various criteria, including consistency and efficiency is summarized in Section 5. Applications of such criteria to time series models are given in the last section.

*Keywords*

Statistical modelling, model selection, information criterion, cross validation.

## 1. INTRODUCTION

In any science *modeling* is a way of approximation of reality. As far as it yields a *good* approximation, a simpler model is better than complex one both for understanding the phenomena and for various applications, for example, forecasting, control, making decision and so on. The principle is the same for selecting a statistical model. One specific point is that we, statisticians, assume that the number of observation is limited and only partial information is available through *data* which possibly involve random fluctuations. Random fluctuation means here various measurement errors as well as fluctuations of the system which generates data. By introducing randomness into a model, the model becomes much more flexible than a deterministic model and resistant to unexpected fluctuations of the system. Another advantage is that we may leave the error of approximation as a part of random fluctuations which are introduced beforehand into a model as long as the former is compatible order of magnitude to the latter. This often results in a simplification of a model. A desirable procedure of statistical model selection is, therefore, to reject a model which is far from the reality and pick up a model in which the error of approximation and the error due to random fluctuations are well balanced. There may be cases where we have to satisfy with a model, not the best one but the best possible one in a given family of models when only very poor information is available for the underlying phenomena.

Complexity of a model is restricted both by the size of observation and by the signal to noise ratio. Needless to say, complete specification might be possible if an infinite number of observations were available for a quite simple system. Otherwise a practical procedure is, starting from a simple model, to increase the complexity until a trade off between the error of approximation and the error due to random fluctuations. To do this systematically, a convenient way is to introduce a criterion to compare models. In this chapter, we discuss various criteria, some of which are based on an information measure. Although in system sciences, time series models, AR, MA or ARMA are quite common, to clarify the point, we first restrict our attention into models for independent observations. Extensions to time series models or state space models are rather technical. Some of them are explained in the last section.

## 2. INFORMATION CRITERIA

Let $\mathbf{Y}_n' = (y_1,...,y_n)$ be $n$ independent observations but not necessarily identically distributed, whose joint density is denoted by $g(\mathbf{Y}_n)$. Hereafter $'$ denotes transpose of a vector or of a matrix, and E denotes the expectation with respect to the vector of random variables, $\mathbf{Y}_n$. We mean by statistical model a parametric family of densities $\mathbf{F} \doteq \{f(\mathbf{Y}_n;\theta),\theta\in\Theta\}$. The part, usually called *model*, for example, linear or non-linear relation between input and output, is described by parametrization of densities through $\theta$ in $\mathbf{F}$. A regression equation $y = x'\beta+\varepsilon$ with explanatory variable $x$ and Gaussian

error $\varepsilon$ with mean 0 and variance $\sigma^2$ is formulated as the model

$$\mathbf{F} = \left\{ \prod_{i=1}^{n} \frac{1}{\sigma} \phi \left[ \frac{y_i - x_i'\beta}{\sigma} \right],\ \theta=(\beta,\sigma)' \in \mathbf{R}^m \times (0,\infty) \right\},$$

where $\phi$ is the standard normal density. A natural way of evaluating goodness of a model $\mathbf{F}$ is to introduce a kind of distance of the estimated density $f(\cdot;\hat{\theta})$, an approximation to the true $g(\cdot)$ based on $\mathbf{Y}_n$, from the $g(\cdot)$. For a while, to simplify the problem, $\hat{\theta}=\hat{\theta}(\mathbf{Y}_n)$ is taken to be the maximum likelihood estimate of $\theta$ under the model $\mathbf{F}$, based on $\mathbf{Y}_n$. As a distance, a natural choice is the Kullback-Leibler information number:

$$K_n(g(\cdot), f(\cdot;\hat{\theta})) = \int g(\mathbf{x}_n) \log \frac{g(\mathbf{x}_n)}{f(\mathbf{x}_n;\hat{\theta})} d\mathbf{x}_n .$$

Note that this is a pseudo-distance since the triangular inequality does not hold true. This varies with the observation $\mathbf{Y}_n$ through $\hat{\theta}(\mathbf{Y}_n)$. As a measure of closeness between two densities $g(\cdot)$ and $f(\cdot;\hat{\theta})$, the measure has been widely accepted. It is known that the measure is nonnegative, zero if two densities coincide, and additive for independent samples $\mathbf{x}_n=(x_1, \ldots, x_n)$. More importantly, as is shown below, this has a close connection with the *maximum likelihood principle* or the *minimum entropy principle* which is a basic in statistical inference. If

$$|\int g(\mathbf{x}_n) \log g(\mathbf{x}_n) d\mathbf{x}_n| < \infty,$$

then the expectation of the Kullback-Leibler information number $K_n(g(\cdot), f(\cdot;\hat{\theta}))$ can be rewritten as

$$\int g(\mathbf{x}_n) \log g(\mathbf{x}_n) d\mathbf{x}_n - \mathrm{E} \int g(\mathbf{x}_n) \log f(\mathbf{x}_n;\hat{\theta}) d\mathbf{x}_n . \tag{2.1}$$

A problem in using (2.1) as a criterion of comparing models is that the second term of (2.1) depends on unknown $g(\cdot)$. We demonstrate that a useful approximation is obtained by expanding it for a large number of observations under the following assumptions A1 to A4.

A1. The parameter space $\Theta$ is a Euclidean $p$-dimensional space $\mathbf{R}^p$ or an open subspace of it. Both the Gradient vector

$$\mathbf{g}_n(\theta)' = \left( \frac{\partial}{\partial \theta_l} l(\theta), l=1,\ldots,p \right)$$

and the Hessian matrix

$$H_n(\theta) = \left( \frac{\partial^2}{\partial \theta_l \partial \theta_m} l(\theta), 1 \le l, m \le p \right)$$

of the log-likelihood function $l(\theta) = \log f(\mathbf{Y}_n;\theta)$, are well defined with probability 1, and both continuous with respect to $\theta$.

A2. $E|\mathbf{g}_n(\theta)|<\infty$ and $E|H_n(\theta)|<\infty$, where $|\cdot|$ denotes the absolute value of each component of a vector or of a matrix.

A3. There exists a unique $\theta^*$ in $\Theta$, which is the solution of $E\,\mathbf{g}_n(\theta^*)=0$. For any $\varepsilon>0$,

$$\sup_{\|\theta-\theta^*\|>\varepsilon} l(\theta) - l(\theta^*)$$

diverges to $-\infty$ a.s..

A4. For any $\varepsilon>0$, there exists $\delta>0$ such that

$$\sup_{\|\theta-\theta^*\|<\delta} | E(\hat{\theta}-\theta^*)' J_n(\theta)(\hat{\theta}-\theta^*) - \mathrm{tr}(I_n(\theta^*)J_n(\theta^*)^{-1}) | < \varepsilon$$

for large enough $n$. Here

$$I_n(\theta^*) = E\,\mathbf{g}_n(\theta^*)\mathbf{g}_n(\theta^*)' \quad \text{and} \quad J_n(\theta) = -E\,H_n(\theta)$$

are assumed to be positive definite matrices and continuous with respect to $\theta$.

The assumption A3 assures that $\hat{\theta}-\theta^*$ converges to 0 a.s. as $n$ tends to infinity. That is, $\hat{\theta}$ is a consistent estimate of $\theta^*$. The assumption A3 together with A2 implies that $K_n(g(\cdot),f(\cdot;\theta))$ is minimized at $\theta^*$. This means that $f(\cdot;\theta^*)$ is the best approximation to $g(\cdot)$. However such a $\theta^*$ completely relies on unknown $g(\cdot)$. The situation can be further understood by looking at the decomposition,

$$K_n(g(\cdot),f(\cdot;\hat{\theta})) = K_n(g(\cdot),f(\cdot;\theta^*)) + \int g(\mathbf{x}_n)\log\frac{f(\mathbf{x}_n;\theta^*)}{f(\mathbf{x}_n;\hat{\theta})}d\mathbf{x}_n.$$

The first term on the right hand side is the least error of approximation by the model $\mathbf{F}$, and the second term is the error due to the estimation of parameter $\theta^*$ by $\hat{\theta}$.

We note that all assumptions above are commonly used regularity conditions. By expanding $\log f(\mathbf{x}_n;\hat{\theta})$ around $\theta^*$, we have

$$\log f(\mathbf{x}_n;\hat{\theta}) = \log f(\mathbf{x}_n;\theta^*) + (\hat{\theta}-\theta^*)'\frac{\partial}{\partial\theta}\log f(\mathbf{x}_n;\theta^*)$$

$$+ \frac{1}{2}(\hat{\theta}-\theta^*)'\frac{\partial^2}{\partial\theta\partial\theta'}\log f(\mathbf{x}_n;\theta^{**})\,(\hat{\theta}-\theta^*),$$

where $\theta^{**}$ is a value between $\hat{\theta}$ and $\theta^*$. We should note that the Gradient vector $\frac{\partial}{\partial\theta}\log f(\mathbf{x}_n;\theta)$ and the Hessian matrix $\frac{\partial^2}{\partial\theta\partial\theta'}\log f(\mathbf{x}_n;\theta)$ are not of the log likelihood $\log f(\mathbf{Y}_n;\theta)$ of the observations $\mathbf{Y}_n$, but of the log likelihood $\log f(\mathbf{x}_n;\theta)$ of *test sample* $\mathbf{x}_n$. Since

$$\int g(\mathbf{x}_n)\frac{\partial}{\partial\theta}\log f(\mathbf{x}_n;\theta^*)d\mathbf{x}_n = 0,$$

the assumption A3 justifies the expansion;

$$\int g(\mathbf{x}_n)\log f(\mathbf{x}_n;\hat{\theta})d\mathbf{x}_n = \int g(\mathbf{x}_n)\log f(\mathbf{x}_n;\theta^*)d\mathbf{x}_n$$

$$+\frac{1}{2}(\hat{\theta}-\theta^*)'\left\{\int g(\mathbf{x}_n)\frac{\partial^2}{\partial\theta\partial\theta'}\log f(\mathbf{x}_n;\theta^{**})\right\}(\hat{\theta}-\theta^*). \tag{2.2}$$

From the assumption A4, the expectation of (2.2) is

$$\mathrm{E}\int g(\mathbf{x}_n)\log f(\mathbf{x}_n;\hat{\theta})d\mathbf{x}_n = \int g(\mathbf{x}_n)\log f(\mathbf{x}_n;\theta^*)d\mathbf{x}_n - \frac{1}{2}\mathrm{tr}(I_n(\theta^*)J_n(\theta^*)^{-1}) + o(1)$$

$$= \mathrm{E}\,l(\theta^*) - \frac{1}{2}\mathrm{tr}(I_n(\theta^*)J_n(\theta^*)^{-1}) + o(1).$$

Furthermore, by expanding $l(\theta^*)$ around $\hat{\theta}$, from the fact that $\mathbf{g}_n(\hat{\theta})=0$ we have

$$l(\theta^*) = l(\hat{\theta}) + \frac{1}{2}(\theta^*-\hat{\theta})'H_n(\theta^{**})(\theta^*-\hat{\theta}), \tag{2.3}$$

and then

$$\mathrm{E}\int g(\mathbf{x}_n)\log f(\mathbf{x}_n;\hat{\theta})d\mathbf{x}_n = \mathrm{E}\,l(\hat{\theta}) - \mathrm{tr}(I_n(\theta^*)J_n(\theta^*)^{-1}) + o(1).$$

Thus the expected Kullback-Leibler information number (2.1), is written as

$$\mathrm{E}\,K_n(g(\cdot),f(\cdot;\hat{\theta})) = \int g(\mathbf{x}_n)\log g(\mathbf{x}_n)d\mathbf{x}_n + \mathrm{E}(-l(\hat{\theta})) + \mathrm{tr}(I_n(\theta^*)J_n(\theta^*)^{-1})) + o(1). \tag{2.4}$$

The first term on the right hand side of (2.4) is independent of any model, and we may omit it. Therefore a practical procedure for selecting a model is to compare values of

$$-l(\hat{\theta}) + \overline{t_n(\theta^*)}, \tag{2.5}$$

for various models **F**, where $\overline{t_n(\theta^*)}$ is an estimate of $t_n(\theta^*) = \mathrm{tr}(I_n(\theta^*)J_n(\theta^*)^{-1})$ which is the sum of the second term on the right hand side of (2.2), the penalty for the increasing model size and the bias correction appeared in (2.3).

There are various ways of estimating $t_n(\theta^*)$, and different criteria may follow. If $g(\cdot)$ is equal to one of densities in **F**, say $f(\cdot;\theta_0)$, then $\theta^*=\theta_0$, $I_n(\theta_0)=J_n(\theta_0)$ and $t_n(\theta^*)=p$. Therefore, for the case when $g(\cdot)$ is expected equal to or very close to one of the densities in **F**, the criterion known as Akaike's Information Criterion [2],

$$\mathrm{AIC} = -2l(\hat{\theta}) + 2p$$

follows from (2.5) since $t_n(\theta^*)=p$. Multiplication by 2 is only a convention.

A more general procedure of estimating $t_n(\theta^*)$, suggested by Takeuchi [36] is the following. An example may illustrate his idea.

**Example 2.1**

Let us consider a simple location and scale family

$$\mathbf{F} = \left\{\prod_{i=1}^{n} f(y_i;\theta),\ \theta\in\Theta\right\},$$

Here, $\theta'=(\mu,\sigma)$, $\Theta=(-\infty,\infty)\times(0,\infty)$, and $f(y_i;\theta)=\frac{1}{\sigma}\phi\left[\frac{y_i-\mu}{\sigma}\right]$ with the standard normal

density $\phi$. In other words, this is exactly the same as the observational equation $y_i=\mu+\varepsilon_i$ with normal error $\varepsilon_i \sim N(\mu,\sigma^2)$. Note that the assumptions above are only for specifying a model but not for restricting the observation generating mechanism $g(\cdot)$. We only assume that the $y_i$'s are independent observations with the same first and second moments. Since $\mu^* = \sum Ey_i/n$ and $\sigma^{*2} = \sum E(y_i-\mu^*)^2/n$, we have

$$\frac{1}{n} I_n(\theta^*) = \begin{bmatrix} 1/\sigma^{*2} & \mu(3)/\sigma^{*5} \\ \mu(3)/\sigma^{*5} & \mu(4)/\sigma^{*6}-1/\sigma^{*2} \end{bmatrix}$$

and

$$\frac{1}{n} J_n(\theta^*) = \begin{bmatrix} 1/\sigma^{*2} & 0 \\ 0 & 2/\sigma^{*2} \end{bmatrix},$$

where $\mu(l)=\sum E(y_i-\mu^*)^l/n$ for $l>1$. Then,

$$t_n(\theta^*) = 1 + \frac{1}{2}(\mu(4)/\sigma^{*4} - 1).$$

By replacing $\mu(4)$ by the 4th sample central moment $\hat{\mu}(4)=\sum(y_i-\bar{y})^4/n$ and $\sigma^{*2}$ by the maximum likelihood estimate $\hat{\sigma}^2=\sum(y_i-\bar{y})^2/n$ respectively, we have an estimate of $t_n(\theta^*)$,

$$\widehat{t_n(\theta^*)} = 1 + \frac{1}{2}(\hat{\mu}(4)/\hat{\sigma}^4 - 1).$$

A statistic which follows from (2.5) is then

$$\begin{aligned} \mathrm{TIC}_0(\mathbf{F}) &= -2l(\hat{\theta}) + 2 + (\hat{\mu}(4)/\hat{\sigma}^4 - 1) \\ &= n + n\log(2\pi\hat{\sigma}^2) + 2 + (\hat{\mu}(4)/\hat{\sigma}^4 - 1). \end{aligned}$$

Multiplication by 2 is again a convention as is in AIC. The difference between $\mathrm{TIC}_0$ and

$$\mathrm{AIC} = -2l(\hat{\theta}) + 4$$

is clear. The discrepancy of the shape of $g(\cdot)$ from the normal density is counted in $\mathrm{TIC}_0$.

By applying the same technique we can derive $\mathrm{TIC}_0$ for the problem of selecting a sample transformation $\psi$. Consider models;

$$\mathbf{F}_\psi = \left\{ \prod_{i=1}^{n} \left| \frac{\psi'(y_i)}{\sigma} \right| \phi\left[ \frac{\psi(y_i)-\mu}{\sigma} \right] \right\},$$

where $\psi'$ is the derivative of $\psi$. Then

$$\mathrm{TIC}_0(\mathbf{F}_\psi) = -2l(\hat{\theta}) + 2 + (\hat{\mu}(4)/\hat{\sigma}^4 - 1)$$

follows, where $\hat{\mu}(4)=\sum(\psi(y_i)-\hat{\mu})^4/n$ and $\hat{\sigma}^2=\sum(\psi(y_i)-\hat{\mu})^2/n$ with $\hat{\mu}=\sum\psi(y_i)/n$. Comparing

$TIC_0(F_\psi)$, we may select a transformation $\psi$.

However, such a procedure of deriving an estimate of $t_n(\theta^*)$ is not widely applicable. It is laborious to find an estimate of $t_n(\theta^*)$ model by model. And there is no definite answer, what kind of assumption is appropriate for the $y_i$'s. Before proceeding to a generalization of $TIC_0$, let us consider another example.

**Example 2.2**

A Gaussian regression model $y_i=x_i'\beta+\varepsilon_i$ with $m$-dimensional regression parameter $\beta$ is denoted by

$$\mathbf{F}_m = \left\{ \prod_{i=1}^{n} \frac{1}{\sigma}\phi\left[\frac{y_i-x_i'\beta}{\sigma}\right],\ \theta=(\beta',\sigma)'\in \mathbf{R}^m\times(0,\infty) \right\}.$$

We first only assume independence of $y_i$'s. Then

$$I_n(\theta^*) = \begin{bmatrix} \sum(\mu_i(2)-\mu_i(1)^2)x_ix_i'/\sigma^{*2} & \sum(\mu_i(3)-\mu_i(1)\mu_i(2))x_i'/\sigma^{*5} \\ \sum(\mu_i(3)-\mu_i(1)\mu_i(2))x_i/\sigma^{*5} & \sum(\mu_i(4)-\mu_i(2)^2)/\sigma^{*6} \end{bmatrix}$$

and

$$J_n(\theta^*) = \begin{bmatrix} X'X/\sigma^{*2} & 0 \\ 0 & 2n/\sigma^{*2} \end{bmatrix}.$$

Here $\mu_i(l)=E(e_i)^l$ for $l>1$ with $e_i=y_i-x_i'\beta^*$, $i=1,\ldots,n$ and $X=(x_1,\ldots,x_n)'$ is the design matrix. Denoting the hat matrix by $H=(h_{ij})=X(X'X)^{-1}X'$ we have

$$t_n(\theta^*) = \sum(\mu_i(2)-\mu_i(1)^2)h_{ii}/\sigma^{*2} + \frac{1}{2}\{\frac{1}{n}\sum(\mu_i(4)-\mu_i(2)^2)/\sigma^{*4}\}.$$

If we assume that the first and the second moments of $e_i$'s are the same, then

$$t_n(\theta^*) = m + \frac{1}{2}(\frac{1}{n}\sum \mu_i(4)/\sigma^{*4}-1) \tag{2.6}$$

and so we have

$$TIC_0(\mathbf{F}_m) = -2l(\hat{\theta}) + 2m + \frac{1}{2}(\frac{1}{n}\sum \hat{e}_i^4/\hat{\sigma}^4-1),$$

where $\hat{e}_i=y_i-x_i\hat{\beta}$, and $\hat{\beta}$ and $\hat{\sigma}$ are the maximum likelihood estimates.
One possible way to avoid such assumptions on $g(\cdot)$ that the second moments are all the same, is to make use of the following inequality.

$$t_n(\theta^*) \le \sum\mu_i(2)h_{ii}/\sigma^{*2} + \frac{1}{2}(\frac{1}{n}\sum\mu_i(4)/\sigma^{*4}-1). \tag{2.7}$$

Here the equality holds true if and only if $\mu_i(1)=E(e_i)=0$ and $\mu_i(2)=E(e_i^2)=\sigma^{*2}$ for all $i$, and the value becomes the same as in (2.6). The right hand side of (2.7) can be estimated by

$$\hat{\imath}_n = \sum \hat{e}_i^2 h_{ii}/\hat{\sigma}^2 + \frac{1}{2}(\frac{1}{n}\sum \hat{e}_i^4/\hat{\sigma}^4 - 1).$$

This estimate is possibly biased. However it is toward a safer direction. More penalty is put on models which are far from the best fitting. The resulting criterion is

$$\mathrm{TIC}(\mathbf{F}_m) = -2l(\hat{\theta}) + 2\hat{\imath}_n.$$

This example leads to a general definition of TIC. We only assume that $\mathbf{Y}_n$ is a vector of independent observations. Since we are modeling independent observations, it is natural to assume that the joint likelihood can be decomposed into

$$l(\theta) = \sum l_i(\theta),$$

where $l_i(\theta)$=log $f_i(y_i;\theta)$. Estimate $I_n(\theta^*)$ by

$$\hat{I} = \sum_i \frac{\partial}{\partial\theta} l_i(\hat{\theta}) \frac{\partial}{\partial\theta'} l_i(\hat{\theta})$$

and $J_n(\theta^*)$ by

$$\hat{J} = -H_n(\hat{\theta}) = -\sum_i \frac{\partial^2}{\partial\theta\partial\theta'} l_i(\hat{\theta}).$$

Then we have a general definition of TIC;

$$\mathrm{TIC} = -2l(\hat{\theta}) + 2\mathrm{tr}(\hat{I}\hat{J}^{-1}).$$

As noted in the previous example, since

$$\sum_i \mathrm{E}\left\{\frac{\partial}{\partial\theta} l_i(\theta^*) \frac{\partial}{\partial\theta'} l_i(\theta^*)\right\} = I_n(\theta^*) + \sum_i \mathrm{E}\left\{\frac{\partial}{\partial\theta} l_i(\theta^*)\right\} \mathrm{E}\left\{\frac{\partial}{\partial\theta'} l_i(\theta^*)\right\}, \qquad (2.8)$$

$\mathrm{tr}(\hat{I}\hat{J}^{-1})$ tends to over-estimate $\imath_n(\theta^*)$ by the last term on the right hand side of (2.8). We can not expect any stable behavior of the maximum likelihood estimate $\hat{\theta}$, as long as such an over estimation is significant. The observations contribute unequally to the Gradient of the log likelihood function at $\theta^*$, which is the solution of

$$\sum_i \mathrm{E}\frac{\partial}{\partial\theta} l_i(\theta^*) = 0.$$

Thus such a bias does not affect our objectives to select a model which well balances the approximation error and the error due to random fluctuations. It is worth noting that $\mathrm{tr}(\hat{I}\hat{J}^{-1})$ is the well known Lagrange-multiplier test statistics [15]. TIC consists of two parts, -2 log ( *maximum likelihood* ) plus twice of the test statistic.

## 3. EQUIVALENCE BETWEEN CROSS VALIDATION AND INFORMATION CRITERIA

Cross validation is one of naive methods of checking goodness of fit of a model. The observations obtained are divided into two parts. One of them is used for estimation and the other is used for goodness test of fit. Detailed analyses and discussions

can be found in Stone[32].

In this section we will show that the cross validation is asymptotically equivalent to TIC. We restrict our attention into a *simple* cross validation. By $\hat{\theta}(-i)$, we denote the maximum likelihood estimate of $\theta$ based on $\mathbf{Y}_n$ without using the ith observation $y_i$. The cross validation is then defined as

$$\mathrm{CV} = -2 \sum_i l_i(\hat{\theta}(-i)).$$

It is shown by Stone[33] that CV is asymptotically equivalent to AIC, when $y_1, \ldots, y_n$ are independent and identically distributed and $g(\cdot)$ is a member of the underlying model F. It does not hold true otherwise. However we can show an equivalence of CV to TIC. Necessary assumptions are the following A5 to A7 besides A1 to A3.

A5. For any $\varepsilon>0$,

$$\max_i \sup_{\|\theta-\theta^*\|>\varepsilon} (l_{-i}(\theta) - l_{-i}(\theta^*))$$

diverges to $-\infty$ a.s. as $n$ tends to infinity, where $l_{-i}(\theta) = l(\theta) - l_i(\theta)$. This implies that $\hat{\theta}(-i)$'s, the solutions of

$$\frac{\partial}{\partial\theta} l_{-i}(\hat{\theta}(-i)) = 0, \quad i=1,\ldots,n,$$

uniformly converge to $\theta^*$ as $n$ tends to infinity.

A6. For any $\varepsilon>0$, there exists $\delta>0$ such that

$$\sup_{\|\theta-\theta^*\|<\delta} \|I - H_n(\theta)H_n(\theta^*)^{-1}\| < \varepsilon$$

for large enough $n$, where $\|\cdot\|$ is the Euclidean norm of a vector or the operator norm of a matrix.

A7. For any $\varepsilon>0$, there exists $\delta>0$ such that

$$\max_i \sup_{\|\theta-\theta^*\|<\delta} \left\| \left\{ \frac{\partial^2}{\partial\theta\partial\theta'} l_i(\theta) \right\} H_n(\theta^*)^{-1} \right\| < \varepsilon$$

for large enough $n$.

From the definition of $\hat{\theta}(-i)$ we have

$$\begin{aligned}
\frac{\partial}{\partial\theta} l_i(\hat{\theta}(-i)) &= \frac{\partial}{\partial\theta} l(\hat{\theta}(-i)) \\
&= \frac{\partial}{\partial\theta} l(\hat{\theta}) + \left\{ \frac{\partial^2}{\partial\theta\partial\theta'} l(\hat{\theta}) \right\} (\hat{\theta}(-i) - \hat{\theta}) \ (1 + o_p(1)). \\
&= -\hat{J} \cdot (\hat{\theta}(-i) - \hat{\theta})(1 + o_p(1))
\end{aligned}$$

and

$$l_i(\hat{\theta}(-i)) = l_i(\hat{\theta}) + (\hat{\theta}(-i) - \hat{\theta})' \frac{\partial}{\partial \theta} l_i(\hat{\theta})$$

$$+ (\hat{\theta}(-i) - \hat{\theta})' \frac{\partial^2}{\partial \theta \partial \theta'} l_i(\theta^{**})(\hat{\theta}(-i) - \hat{\theta})$$

$$= l_i(\hat{\theta}) - \left\{ \frac{\partial}{\partial \theta'} l_i(\hat{\theta}(-i))\, \hat{J}^{-1} \frac{\partial}{\partial \theta} l_i(\hat{\theta}) \right\} (1 + o_p(1))$$

$$= l_i(\hat{\theta}) - \left\{ \frac{\partial}{\partial \theta'} l_i(\hat{\theta})\, \hat{J}^{-1} \frac{\partial}{\partial \theta} l_i(\hat{\theta}) \right\} (1 + o_p(1)).$$

Therefore

$$CV = -2 \sum_i l_i(\hat{\theta}(-i))$$

$$= -2l(\hat{\theta}) + 2 \sum_i \left\{ \frac{\partial}{\partial \theta'} l_i(\hat{\theta})\, \hat{J}^{-1} \frac{\partial}{\partial \theta} l_i(\hat{\theta}) \right\} (1 + o_p(1))$$

$$= -2l(\hat{\theta}) + 2\, \mathrm{tr}(\hat{I}\hat{J}^{-1})\, (1 + o_p(1))$$

is equivalent to TIC.

**Example 3.1**

Consider the same regression model as in Example 2.2. To simplify our discussion, we regard $\sigma$ as a nuisance parameter and estimate it by $\hat{\sigma}$. From the well known equality [32], we have

$$CV = n \log(2\pi\hat{\sigma}^2) + \sum_i (y_i - x_i'\hat{\beta}(-i))^2/\hat{\sigma}^2$$

$$= n \log(2\pi\hat{\sigma}^2) + \sum_i \{ \hat{e}_i/(1-h_{ii}) \}^2/\hat{\sigma}^2.$$

To assure the consistency of $\hat{\theta}$, we may assume that $\max_i(h_{ii})$ converges to 0 as $n$ tends to infinity, which is equivalent to the assumption A6. We then have

$$CV = n \log(2\pi\hat{\sigma}^2) + \sum_i \hat{e}_i^2(1 + 2h_{ii})/\hat{\sigma}^2 + o_p(1)$$

$$= n \log(2\pi\hat{\sigma}^2) + n + \sum_i \hat{e}_i^2 h_{ii}/\hat{\sigma}^2 + o_p(1),$$

which is asymptotically equivalent to TIC when $\sigma$ is regarded as a nuisance parameter. The term $(\frac{1}{n}\sum \hat{e}_i^4/\hat{\sigma}^4 - 1)$ will appear in CV, if $\hat{\sigma}(-i)$ is used in place of $\hat{\sigma}$.

GCV proposed by G. Wahba [40, 18] is a variant of cross validation. It is known that GCV is asymptotically equivalent to AIC at least in the context of regression. Actually

$$\mathrm{GCV} = \sum_i \hat{e}_i^2/(1 - m/n)^2$$

$$= \sum_i \hat{e}_i^2(1 + 2m/n) + O_p(1/n)$$

$$= n\ \{\hat{\sigma}^2(1 + 2m/n) + O_p(1/n^2)\},$$

and

$$\mathrm{AIC} = n + n\ \log 2\pi + 2 + n\ \log\{\exp(2m/n)\hat{\sigma}^2\}$$

$$= n + n\ \log 2\pi + 2 + n\ \log\{((1+2m/n)\hat{\sigma}^2) + O_p(1/n^2)\}.$$

Therefore GCV may behave differently from TIC.

Although the equivalence shown above is only for the case of large enough $n$, this allows us more freedom to choose one of two equivalent criteria, CV or TIC. An advantage of the use of TIC is that the calculation is simpler than that of CV. A simple reduction is possible for CV in the case of regression, but it is generally not true. We have to search for $n$ maximums $l_i(\hat{\theta}(-i))$ $i=1, \ldots, n$ to obtain CV. On the other hand, only one time maximization of the likelihood is necessary to obtain TIC. Another advantage of TIC is that meaning of the value is clear as an estimate of the Kullback-Leibler information number. Note that CV and TIC cover wider area than the AIC does.

## 4. FURTHER EXTENSION OF INFORMATION CRITERIA

Estimation of parameters in previous sections is always based on the maximum likelihood principle. In statistical literature, it is common to use such an estimate because of the proof of its efficiency or asymptotic efficiency. However the optimality is only valid for the class of unbiased estimators of $\hat{\theta}$. Since we are measuring the closeness of estimated density $f(\cdot;\hat{\theta})$ to $g(\cdot)$ by the Kullback-Leibler information number, there is no definite reason why we restrict our attention into such unbiased estimators. In this section, we trace the derivation of AIC or TIC for the case when a more general estimate, the maximum penalized likelihood estimate, is used for estimating $\theta$.

The penalized likelihood is defined as

$$L_\lambda(\mathbf{Y}_n;\theta) = \log f(\mathbf{Y}_n;\theta) + \lambda k(\theta),$$

or

$$L_\lambda(\mathbf{Y}_n;\theta) = \sum_i L_\lambda(y_i;\theta) = \sum_i \{\log f(y_i;\theta) + \lambda k_i(\theta)\}$$

where $k(\theta)\leq 0$ is an arbitrary penalty function which may depend on $n$ and is twice differentiable. The weight $\lambda \geq 0$ controls the amount of penalty.
The maximum penalized likelihood estimate $\hat{\theta}(\lambda)$ is the solution of

$$\frac{\partial}{\partial\theta}L_\lambda(\mathbf{Y}_n;\theta) = 0.$$

We assume that $\hat{\theta}(\lambda)$ converges to $\theta^*(\lambda)$ which is the unique solution of

$$\mathrm{E}\,\frac{\partial}{\partial\theta}L_\lambda(\mathbf{Y}_n;\theta) = 0.$$

By similar expansions as in Section 2, we can show

$$\mathrm{E}\int L_\lambda(\mathbf{x}_n;\hat{\theta}(\lambda))g(\mathbf{x}_n)d\mathbf{x}_n = \mathrm{E}\,L_\lambda(\mathbf{Y}_n;\hat{\theta}(\lambda)) - \mathrm{E}\,(\hat{\theta}(\lambda)-\theta^*(\lambda))'J_n(\lambda)(\hat{\theta}(\lambda)-\theta^*(\lambda)) + o(1), \quad (4.1)$$

where

$$J_n(\lambda) = -\,\mathrm{E}\,\frac{\partial^2}{\partial\theta\partial\theta'}L_\lambda(\mathbf{Y}_n;\theta^*(\lambda)).$$

Subtracting $\lambda k(\hat{\theta}(\lambda))$ from the both sides of (4.1), we have

$$\mathrm{E}\int g(\mathbf{x}_n)\log f(\mathbf{x}_n;\hat{\theta}(\lambda))d\mathbf{x}_n = \mathrm{E}\{\, l(\hat{\theta}(\lambda)) - (\hat{\theta}(\lambda)-\theta^*(\lambda))'J_n(\lambda)(\hat{\theta}(\lambda)-\theta^*(\lambda))\,\} + o(1).$$

Since the expansion

$$0 = \frac{\partial}{\partial\theta}L_\lambda(\mathbf{Y}_n;\theta^*(\lambda)) + \frac{\partial^2}{\partial\theta\partial\theta'}L_\lambda(\mathbf{Y}_n;\theta^*(\lambda))(\hat{\theta}(\lambda)-\theta^*(\lambda))$$

asymptotically holds true, we can rewrite the expectation of the Kullback-Leibler information number as

$$\mathrm{E}\int \log\frac{g(\mathbf{x}_n)}{f(\mathbf{x}_n;\hat{\theta}(\lambda))}\,g(\mathbf{x}_n)d\mathbf{x}_n = \int g(\mathbf{x}_n)\log g(\mathbf{x}_n)d\mathbf{x}_n - \mathrm{E}\,l(\hat{\theta}(\lambda)) + \mathrm{tr}(I_n(\lambda)J_n(\lambda)^{-1}) + o(1),$$

where

$$I_n(\lambda) = \mathrm{E}\left\{\frac{\partial}{\partial\theta}L_\lambda(\mathbf{Y}_n;\theta^*(\lambda))\,\frac{\partial}{\partial\theta'}L_\lambda(\mathbf{Y}_n;\theta^*(\lambda))\right\}.$$

Then the TIC is extended as a regularization information criterion,

$$\mathrm{RIC} = -2l(\hat{\theta}(\lambda)) + 2\mathrm{tr}(\hat{I}(\lambda)\hat{J}(\lambda)^{-1}),$$

where

$$\hat{I}(\lambda) = \sum_i\{\frac{\partial}{\partial\theta}L_\lambda(y_i;\hat{\theta}(\lambda))\,\frac{\partial}{\partial\theta'}L_\lambda(y_i;\hat{\theta}(\lambda))\}$$

and

$$\hat{J}(\lambda) = -\,\frac{\partial^2}{\partial\theta\partial\theta'}L_\lambda(\mathbf{Y}_n;\hat{\theta}(\lambda)).$$

When $\lambda$=0, RIC is reduced to TIC. Then RIC is in fact an extension of TIC. By RIC we can choose $\lambda$ as well as to select a model. One practical procedure is to choose $\lambda$ for each model so as to minimize RIC and compare the minimized value of RIC for each model. We may overcome instability of the estimate when the model happen to be overfitted.

**Example 4.1**

Consider the same regression model as in Example 2.2. To simplify the problem, we regard $\sigma$ as a nuisance parameter. As a penalty function we adopt

$$k(\theta) = -\|X\beta\|^2/2\sigma^2.$$

The maximum penalized likelihood estimate of $\beta$ is then a shrinkage estimate, $\hat{\beta}(\lambda)=\hat{\beta}(0)/(1+\lambda)$, where $\hat{\beta}(0)$ denotes the maximum likelihood estimate of $\beta$. Since

$$\hat{I}(\lambda) = \sum \hat{e}_i^2 x_i x_i'/\sigma^4$$

and

$$\hat{J}(\lambda) = (1+\lambda) X'X/\sigma^2,$$

we have

$$\begin{aligned} \mathrm{RIC}(\mathbf{F}_m, \lambda) &= n \log 2\pi\sigma^2 + \sum (y_i - x_i'\hat{\beta}(\lambda))^2/\sigma^2 + \frac{2}{1+\lambda}\sum \hat{e}_i^2 h_{ii}/\sigma^2 \\ &= n \log 2\pi\sigma^2 + \{ \sum \hat{e}_i^2 + (\frac{\lambda}{1+\lambda})^2 \sum \hat{y}_i^2 + \frac{2}{1+\lambda} \sum \hat{e}_i^2 h_{ii} \}/\sigma^2, \end{aligned}$$

where $\hat{y}_i = y_i - \hat{e}_i$. Here

$$\frac{\partial}{\partial\theta} \mathrm{RIC}(\mathbf{F}_m, \lambda) = \frac{1}{(1+\lambda)^3\sigma^2} \{ \lambda (\sum \hat{y}_i^2 - \sum \hat{e}_i^2 h_{ii}) - \sum \hat{e}_i^2 h_{ii} \}. \tag{5.2}$$

The $\hat{\lambda}$ which minimizes RIC is then

$$\hat{\lambda} = \frac{\sum \hat{e}_i^2 h_{ii}}{\sum \hat{y}_i^2 - \sum \hat{e}_i^2 h_{ii}} \quad \text{if} \quad \sum \hat{y}_i^2 > \sum \hat{e}_i^2 h_{ii},$$

$$= \infty \qquad \text{otherwise.}$$

The resulting estimate of $\beta$ is

$$\hat{\beta}(\hat{\lambda}) = (1 - \sum \hat{e}_i^2 h_{ii}/\sum \hat{y}_i^2)^+ \hat{\beta}(0),$$

where $(\alpha)^+ = \max(\alpha, 0)$. It is interesting to note that a non-negative shrinkage factor automatically follows from minimizing RIC. As a special case, for the model with a single location parameter $\mu$ as in Example 2.1,

$$\hat{\mu}(\hat{\lambda}) = (1 - \sigma^2/n\bar{y}^2)^+ \bar{y},$$

which is a natural shrinkage estimate.

The minimum value of the RIC for each model is

$$\mathrm{RIC}(\mathbf{F}_m, \hat{\lambda}) = n \log 2\pi\sigma^2 + \frac{1}{\sigma^2} \left\{ \sum \hat{e}_i^2 + 2 \left[ 1 - \frac{1}{2} \frac{\sum \hat{e}_i^2 h_{ii}}{\sum \hat{y}_i^2} \right] (\sum \hat{e}_i^2 h_{ii}) \right\} \quad \text{if } \hat{\lambda} < \infty,$$

$$= n \log 2\pi\sigma^2 + \sum y_i^2/\sigma^2 \quad \text{otherwise.}$$

Thus

$$\mathrm{RIC}(\mathbf{F}_m,\infty) \le \mathrm{RIC}(\mathbf{F}_m,\hat{\lambda}) \le \mathrm{RIC}(\mathbf{F}_m,0).$$

Therefore $\mathrm{RIC}(\mathbf{F}_m,\lambda)$ decreases as $\lambda$ increases from 0 and attains the minimum at $\hat{\lambda}$. Particularly when $\hat{\lambda}=\infty$, the complete shrinkage estimate $\hat{\beta}(\infty)=0$ follows. By using such an estimate we can always decrease the value of RIC except for the case when all $\hat{e}_i$'s are 0. We then compare such minimized value for different models $\mathbf{F}_m$, and choose one of them.

More generally if the penalty function is of the form of $k(\theta) = -\|A\beta\|^2/2\sigma^2$, then

$$\mathrm{RIC}(\mathbf{F}_m,\lambda) = n \log 2\pi\sigma^2 + \{\, \|y - X\hat{\beta}(\lambda)\|^2 + 2\sum h_{ii}(\lambda)\hat{e}_i^2 \,\}/\sigma^2,$$

where

$$H(\lambda) = (h_{ij}(\lambda)) = X(X'X+\lambda A'A)^{-1}X'$$

and

$$\hat{\beta}(\lambda) = (X'X+\lambda A'A)^{-1}X'y.$$

As a result, in this regression context, RIC is closely related to a criterion $\hat{T}(h)$ which is mentioned in Titterington [38]. A more closely related criterion is $C_L$ proposed by Mallows [19]. As far as in the context of regression, RIC is almost equivalent to $C_L$ and AIC is equivalent to the $C_p$ proposed by the same author.

The effect of introducing maximum penalized estimator and selecting both model and the $\lambda$ can be seen in Fig.1. Hundred random samples are generated from

$$y = 1 + 0.8x - 1.8x^2 + x^3 + \varepsilon$$

for $0 \le x \le 1$. Here $\varepsilon$ is a random number normally distributed with mean 0 and standard deviation 0.04. The selected order of the polynomial is 5 by TIC or AIC. By RIC, the order 4 and $\lambda=0.003$ are selected. The order 4 is still overfitting but it is compensated by choosing $\lambda$ as 0.003.

It is also possible to extend RIC for the case of more than one penalty function. Still much works have to be done for this criterion. We leave these for future investigations.

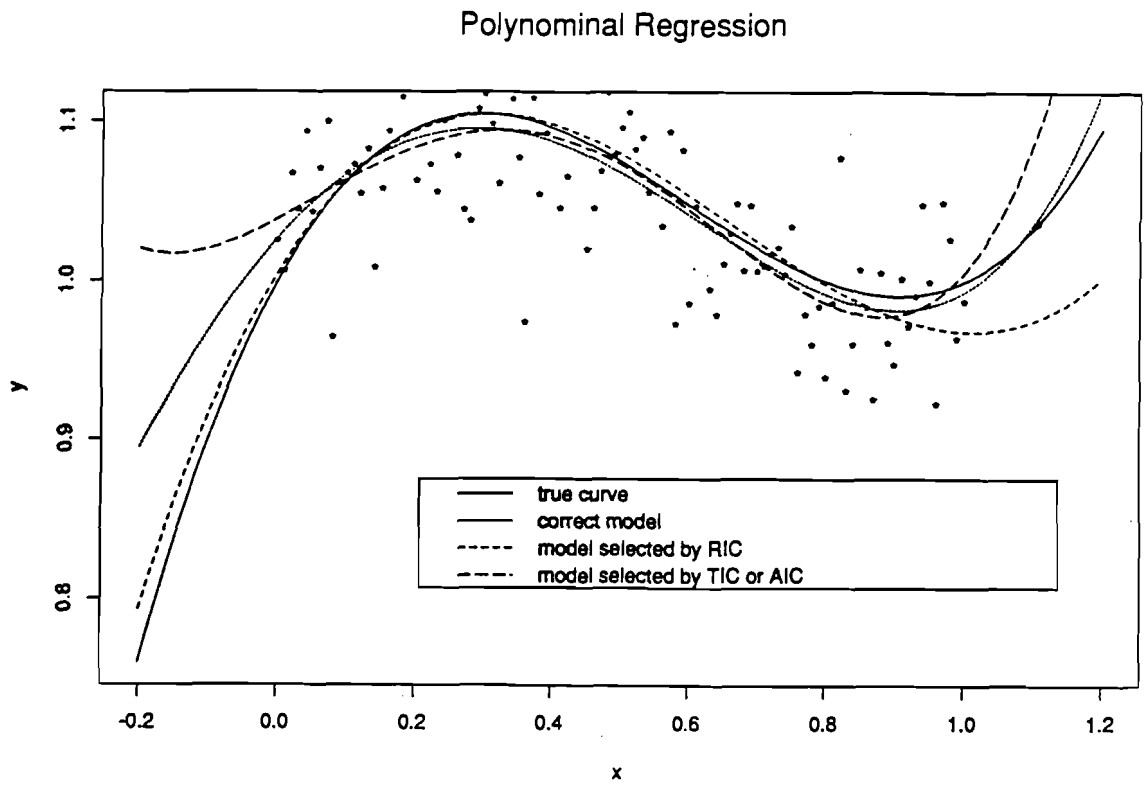


Fig.1 Comparison of three information criteria

## 5. COMPARISON OF CRITERIA

### 5.1. Consistency

A lot of papers are devoted to the consistency of various model selection procedures. Particular interest is in the *inconsistency* of the minimum AIC procedure. However, we may raise a question whether it is always meaningful to only discuss the consistency of model selection. In other words, is the correctness of the selected model is always required first? Recall that a model is only an approximation to the reality. It is a tradition of statistics to discuss correctness of the estimated parameter by assuming that the data are generated from a fixed model. Model selection is however somewhat different from typical estimation problem. We are dealing with different models and looking for a model which gives us a good approximation. Therefore, we should note that the following discussion of consistency is meaningful only when the true system is known to be quite simple and one of the underlying

model can describe the system without error. Furthermore, as is seen later, a tricky point is that consistency of the selection is not compatible with goodness of the resulting estimate of parameters.

To discuss the consistency the following generalization of AIC [5, 4, 9 pp.366-367 ] is convenient.

$$\mathrm{AIC}_\alpha = -2l(\hat{\theta}) + \alpha\, p,$$

where $\alpha$ is a pre-determined value which controls the amount of penalty for an increasing size of the model and may depend on the size $n$ of observations. The result by Hannan and Quinn [10] suggests that under suitable regularity conditions a necessary and sufficient condition for the strong consistency is , putting $\alpha = \alpha_n$,

$$\liminf_n \alpha_n/(2\log\log n) > 1$$

and

$$\limsup_n \alpha_n/n = 0.$$

That for the weak consistency is

$$\liminf_n \alpha_n = \infty$$

and

$$\limsup_n \alpha_n/n = 0.$$

The result above is not yet generally proved, but intuitively clear if we note that $2\{l(\hat{\theta})-l(\theta_0)\}$ is $\chi^2$ distributed with degree of freedom $p$ if $g(\cdot)$ is equal to $f(\cdot;\theta_0)$, a density in the underlying model, otherwise $\chi^2$ distributed with degree of freedom of the order of $n$. The condition for strong consistency comes from the law of iterated logarithm. An implication of the result above is that the AIC, TIC or CV introduced in the previous section are not consistent. For the asymptotic distribution, see Shibata [25], Bhansali & Downham[5], and Woodroofe[42]. They obtained the asymptotic distribution of the selected model by applying theorems of random walk. Some consistent criterion procedures have been proposed, BIC by Schwarz[24] and HQ by Hannan and Quinn[10], which are $\mathrm{AIC}_\alpha$ with $\alpha = \log n$ and $\alpha = c \log\log n$ for $c > 2$, respectively. It is interesting to note the result by Takada[35], that any procedure so as to minimize $\mathrm{AIC}_\alpha$ is admissible under the 0-1 loss. In other word, if our main concern is the correctness of the selection, there is no dominant selection procedure in such class of selection procedures.

### 5.2. Optimality

If we are interested in goodness of a model selection procedure, a natural way is to check the Kullback-Leibler distance of $f(\cdot;\hat{\theta})$ from $g(\cdot)$ where $\hat{\theta}$ is an estimate of the parameter $\theta$ under the selected model. Although not exactly the same, an optimality of the AIC has been shown in terms of such a distance. The key point for proving an optimality property of AIC is that the trade off between the bias and the variance remains significant even when $n$ is large enough. If we restrict our attention to the estimation of regression parameters, such trade off mechanism is rigorously formulated. The result by Shibata [26] shows that if the regression variables are selected so as to minimize one of $AIC_\alpha$ then the selection is asymptotically optimal if and only if $\alpha=2$, that is the case of AIC. Necessary assumptions for the proof are that the shape of $g(\cdot)$ is the same as that of F, and the mean vector of observations is parametrized by infinitely many regression parameters. Otherwise, AIC is not necessarily optimal. But TIC is instead expected to be optimal under a loss function like the Kullback-Leibler information number as well as under the squared loss, even when the shape of $g(\cdot)$ does not coincide with that in F. This result is not yet completely proved.

For admissibility under the squared loss with an additional penalty $p$, Stone [34] proved local asymptotic admissibility, and Kempthorne [17] proved the admissibility under the squared loss. Such results are corresponding to the result by Takada [35] in the case of 0-1 loss function.

All of the results above are in the sense of asymptotics. If the size $n$ is fixed, theoretical comparison is difficult and the only available results are by simulations. Recent paper by Hurvich [16] will help the understanding of the behavior in small samples, for example, consistency does not necessarily imply the goodness of selection. One practical procedure might be obtained by choosing $\alpha$ according to the size $n$ ( see [31] ). For more detailed discussion on incompatibility between consistency and efficiency, see [30], and for comparisons with testing procedures see [29].

## 6. SELECTION OF TIME SERIES MODELS

There have been a lot of articles on the problem of selecting a time series model. In this section, we will review some more criteria of selection of time series models and related works, in connection with the general problem of model selection. The reader can consult some other review papers on this topic [28, 8].

### 6.1. Autoregressive models

Autoregressive process with order $p$, AR($p$), is a weakly stationary process, which satisfies the equation,

$$A_p(B)z_t = \varepsilon_t, \quad t=\cdots,-1,0,1,\cdots$$

where $A_p(z)=1+a_1z+a_2z^2+\cdots+a_pz^p$ is the associate polynomial, $B$ is the backward shift operator and $\{\ \varepsilon_t\ \}$ is a sequence of innovations with mean 0 and variance $\sigma^2$. To completely specify a model, the shape of the distribution of $\varepsilon_t$ have to be specified. It is typically assumed Gaussian, but not restricted to it. A different shape of the density yields different kind of estimates.

By denoting the joint density of consecutive observations $\mathbf{z}_n=(z_1,z_2,\ldots,z_n)'$ by $f(\mathbf{z}_n;\theta)$, we can explicitly specify a model by

$$\mathbf{F}_p = \{\ f(\mathbf{z}_n;\theta);\ \theta=(\sigma,a_1,\ldots,a_p,0,\ldots,0)'\in(0,\infty)\times\mathbf{R}^P\ \}.$$

We then have a nested family of AR models $\{\ \mathbf{F}_p;\ 0\le p\le P\ \}$ for given P. Denote the maximum likelihood estimate of $\theta$ under each model $\mathbf{F}_p$ by

$$\hat{\theta}(p) = (\hat{\sigma}(p),\hat{a}_1(p),\ldots,\hat{a}_p(p),0,\ldots,0)'.$$

To obtain an estimate, the exact maximum likelihood procedure is desirable. The approximation error may affect the behavior. Hereafter, we assume that the shape of densities in the underlying model is normal. For AR models, we can replace it by the conditional maximum likelihood estimate, given $z_1,\ldots,z_P$, or the estimate which minimizes

$$f(z_{P+1},\ldots,z_n;\ \theta\mid z_1,\ldots,z_P).$$

Then $\hat{a}(p)=(\hat{a}_1(p),\ldots,\hat{a}_p(p),0,\ldots,0)'$ is the solution of the Yule-Walker equation,

$$R\,\hat{a}(p) = -\,r$$

and $\hat{\sigma}^2(p) = \frac{1}{N}\sum_{P+1}^{n}\hat{\varepsilon}_t(p)^2$ with $N=n-P$, where

$$R = \left[\ \frac{1}{N}\sum_{P+1}^{n} z_{t-l}z_{t-m};\ 1\le l,m\le P\ \right],$$

is the sample autocovariance matrix,

$$r = \left[\ \frac{1}{N}\sum_{P+1}^{n} z_t z_{t-m};\ 1\le m\le P\ \right]$$

is the vector of sample autocovariances, and

$$\hat{\varepsilon}_t(p) = z_t+\hat{a}_1(p)z_{t-1}+\cdots+\hat{a}_p(p)z_{t-p}\quad t=P+1,\ldots,n,$$

are residuals.

Similarly as in the case of multiple regression, $\mathrm{AIC}_\alpha$ for the model $\mathbf{F}_p$ is

$$\mathrm{AIC}_\alpha = N + N\ \log 2\pi\hat{\sigma}^2(p) + \alpha(p+1).$$

Note that the matrix $R$ and the vector $r$ are defined the same for any order $1\le p\le P$ since the normalization is with $N=n-P$ but not with $n-p$. This is a crucial point when a quasi-maximum likelihood estimate is used in place of the exact one. For example,

if $\tilde{\sigma}^2(p) = \frac{1}{n-p} \sum_{p+1}^{n} \hat{\varepsilon}_t(p)^2$ is used, then $AIC_\alpha$ behaves differently.

On the other hand, TIC becomes

$$TIC = N + N \log 2\pi\hat{\sigma}^2(p) + \left\{ \frac{1}{N} \sum \hat{\varepsilon}_t(p)^4 / \hat{\sigma}^4(p) - 1 \right\}$$

$$+ 2 \sum_t \left\{ \hat{\varepsilon}_t(p)^2 \frac{1}{N} \sum_{l,m} (z_{t-l} R^{lm} z_{t-m}) \right\} / \hat{\sigma}^2(p),$$

where $\{ R^{lm}; 1 \le l, m \le p \}$ is the inverse of the $p$ by $p$ principal submatrix of $R$. Although little is known about TIC, it is clear that TIC is close to AIC if the true $g(\cdot)$ is close to one of densities in $\mathbf{F}_p$, since

$$\frac{1}{N} \sum_{l,m} (z_{t-l} R^{lm} z_{t-m})$$

is corresponding to the diagonal element $h_{ii}$ of the hat matrix in the case of multiple regression and has the expectation $p/N$.

To evaluate the behavior of the selection, we need some assumptions on the true density $g(\mathbf{z}_n)$. As was mentioned before, it is meaningless to consider consistency of the selection unless $g(\mathbf{z}_n)$ is expected to be equal to one of densities in $\mathbf{F}_p$, that is, the true model $AR(p_0)$ exists with an order $p_0 \le P$. Under this assumption, the asymptotic distribution of the selected order $\hat{p}$ which minimizes AIC has been obtained by Shibata [25]. The distribution is nondegenerate and concentrated on $p \ge p_0$, so that the minimum AIC procedure is inconsistent and tends to select a higher order than $p_0$. This also holds true for $AIC_\alpha$ ( Bhansali and Downham [5] with any fixed $\alpha$. This has been extended to multiple AR models by Sakai[23], to ARMA models by Hannan [10, 11, 14], to ARIMA models by Yajima [43], and to AR models with a time dependent variance by Tjøstheim and Paulsen [39]. It is known that the minimum $AIC_\alpha$ procedure is consistent if $\alpha = \alpha_n$ is increased with $n$ at the rate that $\liminf_n (\alpha_n / 2 \log \log n) > 1$ ( see [10] ).

However, for the case when the true $g(\cdot)$ is not expected to be in $\mathbf{F}_p$ for any $p \le P$, our main concern may be about the goodness of the resulting inference rather than the correctness of the selection. In this case, one natural assumption on $g(\cdot)$ is that $\mathbf{z}_n$ comes from an autoregressive process with infinite order. That is, $\mathbf{z}_n$ is generated by the process,

$$A_\infty(B) z_t = \varepsilon_t, \tag{6.1}$$

where $\{\varepsilon_t\}$ is a sequence of innovations with variance $\sigma_\infty^2$, $A_\infty(B) = 1 + a_1 B + a_2 B^2 + \cdots$ is a nondegenerate infinite order transfer function with $\sum |a_l| < \infty$, and $A_\infty(z) \ne 0$ for $|z| \le 1$. Then we can show an optimality property of the minimum AIC procedure $\hat{p}$.

**Theorem 6.1**

Assume that $\{\varepsilon_t\}$ is a sequence of innovations which are independent and normally distributed and $z_n$ is generated by the process (6.1). If $P$ is taken to be $P_n$ which diverges to infinity with the order of $o(n^{1/2})$, then

$$\lim_{n\to\infty} \mathrm{P}\left[\frac{\|\hat{a}(\tilde{p})-a\|_c^2}{\min_p \mathrm{E}\,\|\hat{a}(p)-a\|_c^2} \geq 1-\epsilon\right] = 1.$$

for any selection $\tilde{p}$ from $1 \leq p \leq P$. Here, $\|x\|_c^2 = \sum x_l\, c_{lm}\, x_m$ is the norm with the autocovariances $c_{lm} = \mathrm{E}(z_{t+l}\, z_{t+m})$, and $a'=(a_1,a_2,\cdots)$ is the infinite dimensional vector of the coefficients of the transfer function $A_\infty(B)$. Thus $\|\hat{a}(p)-a\|_c^2+\sigma_\infty^2$ is the one step ahead prediction error of the estimated predictor $(1-\hat{A}_p(B))z_{t+1}^*$ of $z_{t+1}^*$, a realization of a process $\{z_t^*\}$ which is independent of $\{z_t\}$ but has the same covariance structure as that of $\{z_t\}$. The lower bound is attained in probability for large enough n by the selection $\hat{p}$ which minimizes $\mathrm{AIC}_\alpha$ with $\alpha=2$. Any other choice of $\alpha$ does not yield any selection which always attains the bound.

Keys for the proof of the theorem are the following facts. The prediction error is decomposed into two parts, the squared bias and the variance;

$$\begin{aligned}\|\hat{a}(p)-a\|_c^2 &= \|a(p)-a\|_c^2 + \|\hat{a}(p)-a(p)\|_c^2,\\ &= \sigma^2(p)-\sigma_\infty^2 + \|\hat{a}(p)-a(p)\|_c^2,\end{aligned} \tag{6.2}$$

where $a(p)$ is the projection of the infinite dimensional vector $a$ on the $p$-dimensional subspace $\{a=(a_1,a_2,\ldots,a_p,0,\cdots)\}$ with respect to the norm $\|\cdot\|_c$, and $\sigma^2(p)=\mathrm{E}\,(A_p(B)z_t)^2$ is the residual variance for the transfer function $A_p(B)$ with the coefficients $a(p)$. Note that

$$V = N\,\|\hat{a}(p)-a(p)\|_c^2/\sigma_\infty^2. \tag{6.3}$$

is asymptotically $\chi^2$ distributed with degree of freedom $p$. The normalized prediction error,

$$N\,\|\hat{a}(p)-a\|_c^2/\sigma_\infty^2 = N\,(\sigma^2(p)/\sigma_\infty^2-1)+V \tag{6.4}$$

is close to

$$N\,(\hat{\sigma}^2(p)/\sigma_\infty^2-1)+p.$$

For large $p$, $\hat{\sigma}^2(p)/\sigma_\infty^2$ is close to 1 and then the estimate above is approximately equal to $N\log\hat{\sigma}^2(p) - N\log\sigma_\infty^2 + p$. This means that AIC is estimating (6.4) as well as estimating the Kullback-Leibler information number for large $p$. It is enough to consider the case when $p$ is large. The $\hat{p}$ diverges to infinity and the bias term will be dominant for a fixed $p$. A remaining problem in the proof of the theorem is how well p behaves as an estimate of $V$ in (6.3) for large $p$. The estimation error is relatively small and negligible, because $V/p$ converges to 1 in probability as p tends to infinity

simultaneously with n.

In the theorem, the assumption of normality of $\{\varepsilon_t\}$ which generate $z_n$ is not essential. In fact, this theorem was extended by Taniguchi [37] to the case of ARMA model selection without the normality assumption on $\{\varepsilon_t\}$. In our case, instead of the normality it is enough to assume that

$$\sum_j j^{\beta}|a_j| < \infty \text{ for some } \beta > 1,$$

and $\{\varepsilon_t\}$ is an independent identically distributed sequence with finite moments up to the 16th. The shape of densities in each model $\mathbf{F}_p$ is assumed to be normal.

One other possible extension of the theorem is to the case of subset selection, that is, to select a model from the family of models

$$\mathbf{F}_j = \{f(z_n;\theta);\ \theta=(\sigma,0,..,a_{j_1},\ldots,0,a_{j_p},0,\ldots,0)'\in(0,\infty)\times\mathbf{R}^p\}$$

which is specified by a set of indices $j=(j_1,\ldots,j_p)$. For the case of multiple regression, Shibata [26] has already proved that the theorem still holds true. However as far as I know there is no rigorous proof for autoregressive models.

We can also show an optimal property of the minimum AIC in terms of the integrated relative squared error of autoregressive spectral estimate. A fundamental relation between two autoregressive spectral densities, $g(\lambda)=\sigma^2/|A(e^{2\pi i\lambda})|^2$ and $h(\lambda)=s^2/|B(e^{2\pi i\lambda})|^2$, is

$$\int\left\{\frac{h(\lambda)-g(\lambda)}{g(\lambda)}\right\}^2 d\lambda - 2\,\frac{\|a-b\|_h^2}{s^2} = (1-\frac{s^2}{\sigma^2})^2 + 2(2\,\frac{s^2}{\sigma^2}-\frac{\sigma^2}{s^2}-1)\,\frac{\|a-b\|_h^2}{\sigma^2}$$
$$+\frac{s^4}{\sigma^4}\int\left[\frac{|A-B|^2}{|B|^2}+2\frac{\Delta|A|^2}{|B|^2}\right]\frac{|A-B|^2}{|B|^2}\,d\lambda, \tag{6.5}$$

where $\Delta|A|^2=B(\overline{A-B})+\overline{B}(A-B)$, $a$ and $b$ are vectors of coefficients of $A$ and $B$, and $\|x\|_h^2=\sum(x_l\ h_{lm}\ x_m)$ is the norm with

$$h_{lm}=\int e^{2\pi i(l-m)\lambda}\,h(\lambda)d\lambda$$

( see [1,27] ). In (6.5) the order of transfer functions $A$ and $B$ can be infinite. Since

$$\frac{|A-B|^2}{|B|^2}\le\frac{|a-b|^2H}{s^2},\qquad\left|\frac{\Delta|A|^2}{|B|^2}\right|\le\frac{2|a-b|\,H^{1/2}}{s}$$

and

$$\int\frac{|A-B|^2}{|B|^2}d\lambda=\frac{\|a-b\|_h^2}{s^2},$$

where $|x|=\sum|x_l|$ is the absolute norm of the vector $x$ and $H=\max_{\lambda}h(\lambda)$. The last term on the right hand side of (6.5) is bounded by

$$\|a-b\|_h^2 \ (|a-b|^2 H + 4|a-b|H^{1/2}s)/\sigma^4$$

in absolute value.

Consider the autoregressive spectral estimate

$$\hat{f}_p(\lambda) = \hat{\sigma}^2(p) / |\hat{A}_p(e^{2\pi i\lambda})|^2$$

and the true spectral density

$$f_*(\lambda) = \sigma_*^2 / |A_*(e^{2\pi i\lambda})|^2.$$

Putting $h(\lambda)=\hat{f}_p(\lambda)$ and $g(\lambda)=f_*(\lambda)$ in (6.5), we can show optimality of $\hat{p}$ from Theorem 6.1, since $\hat{\sigma}^2(p)$ and $\hat{A}_p(e^{2\pi i\lambda})$ converge to $\sigma_*^2$ and $A_*(e^{2\pi i\lambda})$ respectively, as $p$ increases to infinity simultaneously with $n$.

**Theorem 6.2 ( shibata[27])**

$$\lim_{n\to\infty} \mathrm{P}\left[ \frac{\int \left[(\hat{f}_p(\lambda) - f_*(\lambda))/f_*(\lambda)\right]^2 d\lambda}{2 \min_p \mathrm{E}\, \|\hat{a}(p) - a\|_c^2/\sigma_*^2} \geq 1-\epsilon \right] = 1.$$

for any selection $\tilde{p}$. The bound is attained by $\hat{p}$ in probability for large enough $n$.

The criterion autoregressive transfer function, CAT [20] is derived from the principle to select the order $p$ so as to minimize the integrated relative squared error as above,

$$\frac{1}{2}\int \left[(\hat{f}_p(\lambda) - f_*(\lambda))/f_*(\lambda)\right]^2 d\lambda,$$

which is approximately equal to $\|\hat{a}(p) - a\|_{f_p}^2 / \hat{\sigma}^2(p)$ from (6.5). Noting the consistency of $\hat{f}_p(\lambda)$ and the decomposition (6.2), we have an estimate,

$$\mathrm{CAT}_0 = 1 - \hat{\sigma}_*^2/\hat{\sigma}^2(p) + (p/N)\,\hat{\sigma}_*^2/\hat{\sigma}^2(p),$$

provided that an estimate $\hat{\sigma}_*^2$ of $\sigma_*^2$ is available. By replacing $(p/N)\hat{\sigma}_*^2/\hat{\sigma}^2(p)$ by $p/N$, we have the criterion

$$\mathrm{CAT} = 1 - \hat{\sigma}_*^2/\hat{\sigma}^2(p) + p/N.$$

It is clear that CAT is equivalent to AIC for large $p$, so that the theorems 6.1 and 6.2 also hold true for the minimum CAT procedure. In fact,

$$\begin{aligned}
\mathrm{CAT} &= 1 - \hat{\sigma}_*^2/\hat{\sigma}^2(p) + p/N \\
&= \log \hat{\sigma}^2(p) - \log \hat{\sigma}_*^2 + p/N + O_p((1-\hat{\sigma}_*^2/\hat{\sigma}^2(p))^2) \\
&= \log \hat{\sigma}^2(p) + 2p/N + O((p/N)^2) + O_p((1-\hat{\sigma}_*^2/\hat{\sigma}^2(p))^2) - \log \hat{\sigma}_*^2.
\end{aligned}$$

As is easily seen from the derivation, $\mathrm{CAT}_0$ and CAT are more closely connected with the integrated relative squared error than AIC. As an estimate of $\hat{\sigma}_*^2$, Parzen suggested

the use of

$$\tilde{\sigma}_{-}^2 = \exp\left[\frac{2}{m}\sum_{j=1}^{m} \log I(\frac{j}{m}) + \gamma\right],$$

where $m$ is integral part of $n/2$, $\gamma$ is Euler's constant and

$$I(\lambda) = \frac{1}{n}\left|\sum_t z_t e^{2\pi i t\lambda}\right|^2$$

is the periodogram. An alternative is to use $\tilde{\sigma}^2(P)$ which does not depend on each model and goes to $\sigma_{-}^2$ as $P$ tends to infinity.

Later Parzen [21] proposed a modified CAT,

$$\mathrm{CAT}^* = \frac{1}{N}\sum_{j=1}^{P}\frac{1}{\tilde{\sigma}^2(j)} - \frac{1}{\tilde{\sigma}^2(p)}.$$

This does not require any estimate of $\sigma_{-}^2$ like $\tilde{\sigma}_{-}^2$. Note that

$$\tilde{\sigma}_{-}^2\mathrm{CAT}^* + 1 = 1 - \frac{\tilde{\sigma}_{-}^2}{\tilde{\sigma}^2(p)} + \frac{1}{N}\sum_{j=1}^{P}\frac{\tilde{\sigma}_{-}^2}{\tilde{\sigma}^2(p)}.$$

The behavior of the order which minimizes a criterion is determined only by the differences of values of the criterion. Therefore the behavior of the minimum $\mathrm{CAT}^*$ is almost equal to that of the minimum CAT for large $p$, or for $p \geq p_0$ when the true order $p_0$ is assumed. Theorems 6.1 and 6.2 will also hold true for $\mathrm{CAT}^*$.

### 6.2. Autoregressive moving average models

Autoregressive moving average process with order $p$ and $q$, ARMA($p$,$q$), is a weakly stationary process, which satisfies the equation,

$$A_p(B)z_t = B_q(B)\varepsilon_t, \qquad t=\ldots,-1,0,1,\ldots$$

where $A_p(z)$ and $\{\varepsilon_t\}$ are the same as in AR models, and $B_p(z) = 1 + b_1 z + b_2 z^2 + \cdots + b_q z^q$ is the associated polynomial for the moving average part. Similarly as in AR models, we can construct a family of models $\{\mathbf{F}_{p,q};\ 1\leq p\leq P, 1\leq q\leq Q\}$, in which $\mathbf{F}_{p,q}$ signifies the ARMA($p$,$q$) model.

In each model $\mathbf{F}_{p,q}$, densities are parametrized by $\theta' = (\sigma, a_1,\ldots,a_p,b_1,\ldots,b_q)$. Denote the covariance matrix of $\mathbf{z}_n$ by $Q(\theta)^{-1}$ or shortly $Q^{-1}$. Assuming that the shape of the densities in the model is normal, we have AIC for $\mathbf{F}_{p,q}$,

$$\mathrm{AIC} = -2l(\hat{\theta}) + 2(p+q+1)$$

with

$$-2l(\hat{\theta}) = n\ \log 2\pi\hat{\sigma}^2 - \log|\hat{Q}| + \mathbf{z}_n'\hat{Q}\ \mathbf{z}_n\ /\hat{\sigma}^2,$$

where $\hat{Q}$ and $\hat{\sigma}^2$ are the maximum likelihood estimates of $Q$ and $\sigma^2$ respectively under the model, and $|\hat{Q}|$ is the determinant of $\hat{Q}$. There are various methods for obtaining

the maximum likelihood estimate [ 22 ]. Some of them are:

a) Exact maximum likelihood [3].

b) Conditional likelihood. $z_t, t \leq 0$ are put zero or extrapolated by backward forecasting. Maximization is only for the quadratic term and the remaining terms are disregarded in the log likelihood function [6 ].

c) Whittle's approximation of the log likelihood function [ 41 ].

$$\frac{n}{2} \int \left\{ \log 2\pi f(\lambda) + \frac{I(\lambda)}{f(\lambda)} \right\} d\lambda$$

d) Three-stage approximation [13].
   i) Fit an AR($P$),
   ii) obtain initial estimates of parameters by least squares based on the innovations $\{\hat{\varepsilon}_t\}$ obtained by using the AR coefficients estimated in i),
   iii) apply a correction to the initial estimates.

We should be careful to apply an approximation like b), c) and d). Special attention should be given to the estimates $\hat{\sigma}^2$ and $|\hat{Q}|$, which should be equal to the exact ones up to the order of $O(1/n)$, except the constant which does not depend on $p$ and $q$. Otherwise, AIC will behave differently.

Simple expressions of $TIC_0$ and TIC have not been obtained. Findley [7] evaluated the bias of AIC as an estimate of the Kullback-Leibler information number for the case when the true model is an infinite order moving average process. His result suggests a simple expression of $TIC_0$.

A specific problem arises in ARMA model selection, i.e., *identifiability*. If an ARMA($p,q$) is fitted to ARMA($p_0,q_0$) with $p_0<p$ and $q_0<q$, then the transfer functions $A_p(B)$ and $B_q(B)$ have common roots, which are not identifiable. Then, the maximum likelihood estimates of parameters behave differently. In fact, Hannan [12, 14] proved that the exact maximum likelihood estimates $\hat{a}_1$ and $\hat{b}_1$ converge to $\pm 1$ if ARMA(1,1) is fitted to ARMA(0,0), and Shibata [30] proved that they are asymptotically Cauchy distributed if three-stage approximation procedure d) is employed. Therefore, as far as the true model is expected to be or close to a finite order ARMA model, inconsistency of the selection is troublesome. For example, the minimum AIC is inconsistent even when $p_0<P$ and $q_0<Q$. A modification of AIC may solve this problem [ 30 ]. The use of a consistent selection procedure like the minimum BIC or HQ may solve this problem, too. But it increases the error of the resulting parameter estimate. Whereas, under the assumption that $\mathbf{z}_n$ is generated from an infinite order moving average process which is not a degenerate finite order ARMA, such problem never arises and an optimality property holds true similarly as in AR models [37].

How to select a moving average model whose associate polynomial has roots on the unit circle is also an interesting problem which has to be investigated in the future.

## 7. REFERENCES


[1] Akaike, H., A fundamental relation between predictor identification and power spectrum estimation, *Ann. Inst. Statist. Math.*, Vol. 22, pp. 219-223, 1970.

[2] Akaike, H., Information theory and an extension of the maximum likelihood principle, pp. 267-281 in *2nd Int. Symposium on Information Theory*, Eds. B. N. Petrov and F. Csáki, Akadémia Kiado, Budapest, 1973.

[3] Ansley, C. F., An algorithm for the exact likelihood of a mixed autoregressive-moving average process, *Biometrika*, Vol. 66, pp. 59-65, 1979.

[4] Atkinson, A. C., A note on the generalized information criterion for choice of a model, *Biometrika*, Vol. 67, pp. 413-418, 1980.

[5] Bhansali, R. J. and D. Y. Downham, Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion, *Biometrika*, Vol. 64, pp. 547-551, 1977.

[6] Box, G. E. P. and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, 1976.

[7] Findley, D. F., On the unbiasedness property of AIC for exact or approximating linear stochastoic time series models, *J. Time Series Analysis*, Vol. 6, pp. 229-252, 1985.

[8] Gooijer, J. G. de, B. Abraham, A. Gould and L. Robinson, Method for determining the order of an autoregressive-moving average process: A survey, *Int. Statist. Rev.*, Vol. 53, pp. 301-329, 1985.

[9] Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel, *Robust Statistics: the Approach Based on Influence Functions*, John Wiley, 1986.

[10] Hannan, E. J. and B. G. Quinn, The determination of the order of an autoregression, *J. Roy. Statist. Soc.*, Vol. B 41, pp. 190-195, 1979.

[11] Hannan, E. J., The estimation of the order of an ARMA process, *Ann. Statist.*, Vol. 8, pp. 1071-1081, 1980.

[12] Hannan, E. J., Testing for autocorrelation and Akaike's criterion, pp. 403-412 in *Essays in Statistical Science ( Papers in honour of P.A.P. Moran )*, Eds. J. M. Gani and E. J. Hannan, Applied Probability Trust, Sheffield, 1982.

[13] Hannan, E. J. and J. Rissanen, Recursive estimation of mixed autoregressive-moving average order, *Biometrika*, Vol. 69, pp. 81-94, 1982.

[14] Hannan, E. J., Fitting multivariate ARMA models, pp. 307-316 in *Statistics and Probability ( Essays in Honor of C. R. Rao )*, Eds. G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh, North-Holland Publishing Company, Amsterdam, 1982.

[15] Hosking, J. R. M., Lagrange-multiplier tests of time-series models, *J. R. Statist. Soc.*, Vol. B42, pp. 170-181, 1980.

[16] Hurvich, C. M., Data-Driven choice of a spectraum estimate: Extending the applicability of cross-validation methods, *J. Amer. Statist. Soc.*, Vol. 80, pp. 933-940, 1985.

[17] Kempthorne, P. J., Admissible variable-selection procedures when fitting regression models by least squares for prediction, *Biometrika*, Vol. 71, pp. 593-597, 1984.

[18] Li, K. C., From Stein's unbiased estimates to the method of generalized cross validation, *Ann. Statist.*, Vol. 13, pp. 1352-1377, 1985.

[19] Mallows, C. L., Some comments on $C_p$, *Technometrics*, Vol. 15, pp. 661-675, 1973.

[20] Parzen, E., Some recent advances in time series modeling, *IEEE*, Vol. AC-19, pp. 723-730, 1974.

[21] Parzen, E., Multiple time series: determining the order of approximating autoregressive schemes, pp. 283-295 in *Multivariate Analysis-IV*, North-Holland, 1977.

[22] Priestly, M. B., *Spectral Analysis and Time Series*, Academic Press, 1981.

[23] Sakai, H., Asymptotic distribution of the order selected by AIC in multivariate autoregressive model fitting, *Int. J. Control*, Vol. 33, pp. 175-180, 1981.

[24] Schwarz, G., Estimating the dimension of a model, *Ann. Statist.*, Vol. 6, pp. 461-464, 1978.

[25] Shibata, R., Selection of the order of an autoreegressive model by Akaike's information criterion, *Biometrika*, Vol. 63, pp. 117-126, 1976.

[26] Shibata, R., An optimal selection of regression variables, *Biometrika*, Vol. 68, pp. 45-54, Correction 69, p.492, 1981.

[27] Shibata, R., An optimal autoregressive spectral estimate, *Ann. Statist. 9*, pp. 300-306, 1981.

[28] Shibata, R., Various model selection techniques in time sereis analysis, pp. 179-187 in *Handbook of Statistics*, Eds. E. J. Hannan and P. R. Krishnaiah, Elsevier, 1985.

[29] Shibata, R., Selection of regression variables, pp. 709-714 in *Encyclopedia of Statistical Sciences*, John Wiley & Sons, 1986.

[30] Shibata, R., Consistency of model selection and parameter estimateion, pp. 127-141 in *Essays in Time Series and Allied Processes*, Eds. J. M. Gani and M. B. Priestley, Applied Probability Trust, Sheffield, 1986.

[31] Shibata, R., Selection of the number of regression variables; a minimax choice of generalized FPE, *Ann. Inst. Statist. Math.*, Vol. 38 A, pp. 459-474, 1986.

[32] Stone, M., Cross-validatory choice and assessment of statistical predictions, *J. Roy. Statist. Soc.*, Vol. 36, pp. 111-133, 1974.

[33] Stone, M., An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *J. Roy. Statist. Soc.*, Vol. B 39, pp. 44-47, 1977.

[34] Stone, C. J., Local asymptotic admissibility of a generalization of Akaike's model selection rule, *Ann. Inst. Statist. Math.*, Vol. 34, pp. 123-133, 1982.

[35] Takada, Y., Admissibility of some variable selection rules in linear regression model, *J. Japan Statist. Soc.*, Vol. 12, pp. 45-49, 1982.

[36] Takeuchi, K., Distribution of information statistics and a criterion of model fitting, *Suri-Kagaku ( Mathematical Sciences )*, Vol. 153, pp. 12-18, (in Japanese), 1976.

[37] Taniguchi, M., On selection of the order of the spectral density model for a stationary process, *Ann. Inst. Statist. Math.*, Vol. 32 A, pp. 401-419, 1980.

[38] Titterington, D. M., Common structure of smoothing techniques in statistics, *Int. Statist. Rev.*, Vol. 53, pp. 141-170, 1985.

[39] Tjøstheim, D. and J. Paulsen, Least squares estimates and order determination procedures for autoregressive processes with a time dependent variance, *J. Time Series Analysis*, Vol. 6, pp. 117-133, 1985.

[40] Wahba, G., A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem, *Ann. Statist.*, Vol. 13, pp. 1378-1402, 1985.

[41] Walker, A. M., Asymptotic properties of least squares estimates of parameters of the spectrum of a stationary non-deterministic time series, *Austral. Math. Soc.*, Vol. 4, pp. 363-384, 1964.

[42] Woodroofe, M., On model selection and the arc sine laws, *Ann. Statist.*, Vol. 10, pp. 1182-1194, 1982.

[43] Yajima, Y., Estimation of the degree of differencing of an ARIMA process, *Ann. Inst. Statist. Math.*, Vol. 37, pp. 389-408, 1985.

# INDEX