Consistency of Model Selection and Parameter Estimation

RITEI SHIBATA

Abstract

The relationship between consistency of model selection and that of parameter estimation is investigated. It is shown that the consistency of model selection is achieved at the cost of a lower order of consistency of the resulting estimate of parameters in some domain. The situation is different when selecting autoregressive moving average models, since the information matrix becomes singular when overfitted. Some detailed analyses of the consistency are given in this case.

AUTOREGRESSIVE PROCESSES; AKAIKE INFORMATION CRITERION; FINITE PARAMETER ESTIMATION

1. Introduction

This paper aims to clarify a relation between the consistency of model selection and that of parameter estimation after a model is selected. Although many types of procedure have been proposed for selecting a statistical model, some of them are consistent while others are not, unless separated models are considered. For example, procedures such as the minimum BIC (Schwarz [10]) or the minimum HQ (Hannan and Quinn [7]) are consistent, while others such as the minimum AIC (Akaike [2]), or the minimum FPE (Akaike [1]) or the minimum C_p (Mallows [9]) are not consistent. Therefore, a natural question arises as to whether the consistency of model selection is really needed or not.

In this paper, we first demonstrate that if model selection is consistent, then the least order of consistency of the parameter estimate becomes lower than $\sqrt{n}$, which is the order of consistency of the original parameter estimate. This fact was partly pointed out by Shibata [13]. We shall show that the consistency of model selection is achieved at the cost of a lower order of consistency of the resulting estimate of parameters in some parameter domain.

Next, in Section 4, we consider the case where the consistency of model selection seems more important than the consistency of the parameter estimate. An example is in the case of selecting an autoregressive–moving-average

model ARMA(p, q). Since a lower-order model ARMA$(p-1, q-1)$ is not identifiable in a higher-order model ARMA(p, q), the use of an inconsistent procedure such as AIC might be troublesome. For such a case we propose a modification of AIC or FPE. With this modification the procedure correctly selects the lower model if a common root exists, but otherwise the behaviour remains unchanged and the order of consistency is retained as $\sqrt{n}$.

2. Consistency of model selection

To simplify our discussion, consider only two models, one of which, Model 1, $\{f(x, \theta); \theta \in \Theta\}$, is a family of density functions parametrised by θ, where Θ is an open subset of $\boldsymbol{R}^1$, and the other, Model 0, $\{f(x, \theta_0)\}$, consists of a density function specified by a parameter θ_0 in Θ, which is nested in Model 1. Since model selection here is very simple, it is enough to consider the likelihood ratio testing of the null hypothesis,

$$\mathrm{H}_0;\ \theta = \theta_0.$$

Given n independent samples $\boldsymbol{x}_n = (x_1, \cdots, x_n)$, the above selection is denoted by a random variable

$$\hat{m} = \begin{cases} 0, & \boldsymbol{x}_n \in A_n \\ 1, & \boldsymbol{x}_n \notin A_n, \end{cases}$$

where

$$A_n = \left\{ \boldsymbol{x}_n;\ T = 2 \sum_{i=1}^{n} \log \frac{f(x_i, \hat{\theta})}{f(x_i, \theta_0)} < \alpha_n \right\},$$

and $\hat{\theta}$ denotes the maximum likelihood estimate of θ in Model 1. By $\hat{m}$ we can denote various selection procedures, the minimum AIC with $\alpha_n = 2$, the minimum HQ with $\alpha_n = c \log \log n$ for some $c > 2$, and the minimum BIC with $\alpha_n = \log n$. We call $\hat{m}$ weakly consistent if $\hat{m}$ converges in probability to 0 under the null hypothesis H_0, and to 1 under the alternatives. We call it strongly consistent if the convergence is almost sure. We assume the following.

The Fisher information $I(\theta)$ per sample is finite and not 0, and the following condition holds true for the Kullback–Leibler information $I(\theta, \theta_0)$:

$$\inf_{\theta} \frac{I(\theta, \theta_0)}{|\theta - \theta_0|^2 I(\theta)} \geqq C_0 > 0.$$

Next assume that the asymptotic distribution of $\{nI(\theta)\}^{\frac{1}{2}} |\hat{\theta} - \theta|$ does not degenerate uniformly in θ, or, more explicitly, that

$$\lim_{M \to \infty} \limsup_{n \to \infty} \sup_{\theta} P_\theta[\{nI(\theta)\}^{\frac{1}{2}} |\hat{\theta} - \theta| > M] = 0,$$

and

$$\lim_{M \to 0} \liminf_{n \to \infty} \inf_{\theta} P_\theta[\{nI(\theta)\}^{\frac{1}{2}} |\hat{\theta} - \theta| > M] = 1.$$

We further assume that the likelihood ratio

$$T = 2 \sum_{i=1}^{n} \left\{ \log \frac{f(x_i, \hat{\theta})}{f(x_i, \theta)} + \log \frac{f(x_i, \theta)}{f(x_i, \theta_0)} \right\}$$

is well approximated by

$$nI(\theta)(\hat{\theta} - \theta)^2 + 2nI(\theta, \theta_0) \tag{2.1}$$

for large enough n when $\boldsymbol{x}_n$ is generated from a population specified by a density function $f(x, \theta)$. An example, for which the above assumptions all hold true, is the normal family with location parameter θ. The following proposition is easily obtained by noting that

$$\left(\frac{nI(\theta_0)}{2 \log \log n} \right)^{\frac{1}{2}} (\hat{\theta} - \theta_0)$$

stays between -1 and 1 for large enough n under the null hypothesis H_0, which follows from the law of the iterated logarithm (see Hannan [6]).

Proposition 2.1. A selection $\hat{m}$ is weakly consistent if and only if

$$\liminf_{n \to \infty} \alpha_n = \infty \quad \text{and} \quad \limsup_{n \to \infty} \alpha_n / n = 0,$$

and is strongly consistent if and only if

$$\liminf_{n \to \infty} \alpha_n / (2 \log \log n) > 1 \quad \text{and} \quad \limsup_{n \to \infty} \alpha_n / n = 0.$$

For uniform consistency, the following proposition holds true.

Proposition 2.2. For any α_n,

$$\liminf_{n \to \infty} \inf_{\{\theta; 2nI(\theta, \theta_0) \leqq \alpha_n\}} P_\theta(\hat{m} = 1) = 1.$$

If $\liminf_{n \to \infty} \alpha_n = \infty$, then for any $\varepsilon > 0$

$$\limsup_{n \to \infty} \sup_{\{\theta; 2nI(\theta, \theta_0) \leqq \alpha_n (1 - \varepsilon)\}} P_\theta(\hat{m} = 1) = 0,$$

otherwise, if α_n is bounded,

$$\liminf_{n \to \infty} \inf_{\theta \in \Theta} P_\theta(\hat{m} = 1) > 0.$$

3. Consistency of parameter estimate

In this section, our main concern is the order of consistency of the parameter estimate resulting from the model selection $\hat{m}$:

$$\hat{\theta}(\hat{m}) = \begin{cases} \theta_0 & \text{if} \quad \hat{m} = 0, \\ \hat{\theta} & \text{if} \quad \hat{m} = 1. \end{cases}$$

We call $\hat{\theta}(\hat{m})$ consistent if it converges to θ in probability. For such an estimate, it is necessary to introduce the concept of non-uniform order of consistency. For a consistent estimate $\tilde{\theta}$, consider a sequence of functions $c_n(\theta)$ defined on $\Theta_n \subset \Theta$ such that

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{\theta\in\Theta_n} P_\theta(c_n(\theta)\,|\tilde{\theta}-\theta|>M)=0. \tag{3.1}$$

The limit distribution of $c_n(\theta)\,|\tilde{\theta}-\theta|$ is then not degenerate at ∞. We call $c_n^*(\theta)$ an order of consistency of the estimate $\tilde{\theta}=\hat{\theta}(\hat{m})$, if

$$\liminf_{n\to\infty}\left\{\inf_{\theta\in\Theta_n}\frac{c_n^*(\theta)}{c_n(\theta)}\right\}>0$$

for any other $c_n(\theta)$ which satisfies (3.1). The order of consistency $c_n^*(\theta)$ defined here is then unique in the following sense. For any $c_n(\theta)$ which satisfies (3.1),

$$0<\liminf_{n\to\infty}\inf_{\theta\in\Theta_n}\frac{c_n^*(\theta)}{c_n(\theta)}\leqq\limsup_{n\to\infty}\sup_{\theta\in\Theta_n}\frac{c_n^*(\theta)}{c_n(\theta)}<\infty.$$

Theorem 3.1. The order of consistency $c_n^*(\theta)$ of $\hat{\theta}(\hat{m})$ depends on the rate of divergence of α_n to ∞ in the following manner.

If α_n is bounded, then $c_n^*(\theta)=\{nI(\theta)\}^{\frac{1}{2}}$ for all $\theta\in\Theta$. If $\liminf_{n\to\infty}\alpha_n=\infty$ and $\limsup_{n\to\infty}\alpha_n/n=0$, then $c_n^*(\theta)$ is defined only on $\Theta_n=\{\theta\in\Theta;\,2nI(\theta,\theta_0)\geqq\alpha_n$ or $0<2nI(\theta,\theta_0)\leqq\alpha_n(1-\varepsilon)\}$ for any $\varepsilon>0$, and

$$c_n^*(\theta)=\begin{cases}\{nI(\theta)\}^{\frac{1}{2}} & \text{if}\quad 2nI(\theta,\theta_0)\geqq\alpha_n,\\ \dfrac{1}{|\theta-\theta_0|} & \text{if } 0<2nI(\theta,\theta_0)\leqq\alpha_n(1-\varepsilon).\end{cases}$$

If $\liminf_{n\to\infty}\alpha_n/n\neq 0$, then $\hat{\theta}(\hat{m})$ is not consistent.

Proof. For the case when α_n is bounded, we have

$$\begin{aligned}&P_\theta[\{nI(\theta)\}^{\frac{1}{2}}\,|\hat{\theta}(\hat{m})-\theta|>M]\\ &\quad=P_\theta[\{nI(\theta)\}^{\frac{1}{2}}\,|\theta-\theta_0|>M,\,\hat{m}=0]+P_\theta[\{nI(\theta)\}^{\frac{1}{2}}\,|\hat{\theta}-\theta|>M,\,\hat{m}=1]\\ &\quad\leqq P_\theta[\{nI(\theta)\}^{\frac{1}{2}}\,|\theta-\theta_0|>M,\,T<\alpha_n]+P_\theta[\{nI(\theta)\}^{\frac{1}{2}}\,|\hat{\theta}-\theta|>M].\end{aligned}$$

From (2.1), the first term on the right-hand side of the above inequality is bounded by

$$P_\theta(C_0M^2<\alpha_n)+P_\theta[\{nI(\theta)\}^{\frac{1}{2}}\,|\hat{\theta}-\theta|>M].$$

Therefore we see that $c_n^*(\theta)=\{nI(\theta)\}^{\frac{1}{2}}$ satisfies (3.1). On the other hand, if there exists a function $c_n(\theta)$ such that $\lim_{n\to\infty}\{nI(\theta_n)\}^{\frac{1}{2}}/c_n(\theta_n)=0$ for a sequence θ_n in

Θ, then

$$P_{\theta_n}[c_n(\theta_n)\,|\hat{\theta}(\hat{m})-\theta_n|>M] \geqq P_{\theta_n}\left[\{nI(\theta_n)\}^{\frac{1}{2}}\,|\hat{\theta}-\theta_n|>\frac{M\{nI(\theta_n)\}^{\frac{1}{2}}}{c_n(\theta_n)}, \hat{m}=1\right]$$

$$\geqq P_{\theta_n}(\hat{m}=1)-P_{\theta_n}\left[\{nI(\theta_n)\}^{\frac{1}{2}}\,|\hat{\theta}-\theta_n|\leqq\frac{M\{nI(\theta_n)\}^{\frac{1}{2}}}{c_n(\theta_n)}\right]$$

which converges to 1 as $n\to\infty$ for any M. Therefore, such a $c_n(\theta)$ does not satisfy (3.1). This implies that $c_n^*(\theta)=\{nI(\theta)\}^{\frac{1}{2}}$ is the order of consistency.

Next, consider the case when $\liminf_{n\to\infty}\alpha_n=\infty$ and $\limsup_{n\to\infty}\alpha_n/n=0$. If $2nI(\theta,\theta_0)\geqq\alpha_n$, then

$$P_\theta[\{nI(\theta)\}^{\frac{1}{2}}\,|\hat{\theta}(\hat{m})-\theta|>M]\leqq P_\theta(\hat{m}=0)+P_\theta[\{nI(\theta)\}^{\frac{1}{2}}\,|\hat{\theta}-\theta|>M]. \tag{3.2}$$

If $0<2nI(\theta,\theta_0)\leqq\alpha_n(1-\varepsilon)$,

$$P_\theta\left\{\frac{1}{|\theta-\theta_0|}\,|\hat{\theta}(\hat{m})-\theta_0|>M\right\}\leqq P_\theta(1>M,\hat{m}=0)+P_\theta(\hat{m}=1). \tag{3.3}$$

From Proposition 2.2 together with (3.2) and (3.3), we see that $c_n^*(\theta)$ satisfies (3.1).

To show $c_n^*(\theta)$ to be the order of consistency, suppose that there exists a sequence $c_n(\theta_n)$ for which $\lim_{n\to\infty}c_n^*(\theta_n)/c_n(\theta_n)=0$. Then for any θ_n such that $2nI(\theta_n,\theta_0)>\alpha_n$,

$$P_{\theta_n}(c_n(\theta_n)\,|\hat{\theta}(\hat{m})-\theta_0|>M)$$

is bounded away from

$$P_{\theta_n}\left[\{nI(\theta_n)\}^{\frac{1}{2}}\,|\hat{\theta}-\theta_n|>M\frac{c_n^*(\theta_n)}{c_n(\theta_n)}, \hat{m}=1\right],$$

and, for any θ_n such that $0<2nI(\theta_n,\theta_0)\leqq\alpha_n(1-\varepsilon)$ it is bounded away from

$$P_{\theta_n}\left\{1>M\frac{c_n^*(\theta_n)}{c_n(\theta_n)}, \hat{m}=0\right\}+P_{\theta_n}(\hat{m}=1).$$

Therefore, from Proposition 2.2 we have

$$\lim_{M\to\infty}\liminf_{n\to\infty}P_{\theta_n}(c_n(\theta_n)\,|\hat{\theta}(\hat{m})-\theta_0|>M)=1.$$

The function $c_n(\theta)$ then does not satisfy (3.1).

If $\liminf_{n\to\infty}\alpha_n/n\neq 0$, then we can choose a small δ that

$$0<I(\theta,\theta_0)<\liminf_{n\to\infty}\alpha_n/2n \quad\text{and}\quad |\theta-\theta_0|>\delta.$$

Since $P_\theta(\hat{m}=0)$ goes to 1 for such θ, from the inequality

$$P_\theta(|\hat{\theta}(\hat{m})-\theta|>\delta)\geqq P_\theta(|\theta-\theta_0|>\delta,\ \hat{m}=0),$$

we see that $\hat{\theta}(\hat{m})$ is not consistent.

Theorem 3.1 implies that if $\alpha_n\to\infty$ and $\alpha_n/n\to\infty$ then for θ in $\{\theta;\ 0<2nI(\theta,\theta_0)\leqq\alpha_n(1-\varepsilon)\}$ the order of consistency of $\hat{\theta}(\hat{m})$ is lower than the order $\{nI(\theta)\}^{\frac{1}{2}}$ of consistency of the maximum likelihood estimate $\hat{\theta}$. To demonstrate this explicitly, we shall further assume that $I(\theta,\theta_0)$ is well approximated by $|\theta-\theta_0|^2I(\theta_0)$ in the neighbourhood of θ_0. Then for θ such that $2nI(\theta,\theta_0)=\alpha_n(1-\varepsilon)$, $c_n^*(\theta)=1/|\theta-\theta_0|$ is approximated by

$$\{2/(1-\varepsilon)\}^{\frac{1}{2}}\{nI(\theta_0)/\alpha_n\}^{\frac{1}{2}},$$

and the corollary follows.

Corollary 3.1. The estimate $\hat{\theta}(\hat{m})$ is consistent if and only if $\liminf_{n\to\infty}\alpha_n/n=0$. If, in addition, for any $\varepsilon>0$ there exists a δ such that

$$\left|\frac{I(\theta,\theta_0)}{I(\theta_0)\,|\theta-\theta_0|^2}-1\right|<\varepsilon\quad\text{for}\quad 0<|\theta-\theta_0|<\delta,$$

then the least order of consistency, $\inf_{\theta\in\Theta_n}c_n^*(\theta)$, of $\hat{\theta}(\hat{m})$ is $\{nI(\theta_0)\}^{\frac{1}{2}}/\alpha_n^{\frac{1}{2}}$.

For example, if $\alpha_n=\log\log n$ as in the HQ criterion the least order of consistency is $\{n/(\log\log n)\}^{\frac{1}{2}}$, provided that $I(\theta)$ is bounded below 0.

4. Selection of an autoregressive–moving-average model

As was seen in the previous section, the procedure specified by a bounded α_n, such as the AIC or the FPE, is safer than others specified by a divergent α_n, from the viewpoint of the consistency of the estimate $\hat{\theta}(\hat{m})$. This also holds true for the case of multidimensional θ. However, the same discussion does not follow if the Fisher information matrix $I(\theta)$ is singular at $\theta=\theta_0$. A typical example is in the case of selecting an autoregressive–moving-average model ARMA(p,q) of orders p and q. If the autoregressive and moving-average operators have a common root, the Fisher information matrix $I(\theta)$ is singular, so that the behaviour of the maximum likelihood estimates under the ARMA(p,q) is not standard. In this case, a lower-order model ARMA$(p-1,q-1)$ may yield a better estimate. From such a point of view, Hannan and Rissanen [8] and Hannan [4], [5], [6] used a consistent selection procedure such as the HQ (Hannan and Quinn [7]) or the BIC (Schwarz [10]). In what follows, we shall explicitly analyse the asymptotic behaviour of the parameter estimates when a common root exists. The results support the result of Hannan [6] that AIC and FPE are not consistent selection procedures. However, to protect ourselves in

other cases from a possible decrease of the order of consistency of parameter estimates, we propose a modification of AIC and FPE.

To simplify our discussion, we first consider the case of selecting one of the models ARMA(1, 1) and ARMA(0, 0).

Let us consider the ARMA(1, 1) model

$$z_t - \alpha z_{t-1} = a_t - \beta a_{t-1},$$

where $|\alpha|<1$, $|\beta|<1$ and $\{a_t\}$ is a sequence of independent and identically normally distributed random variables with mean 0 and variance σ^2. Clearly, if $\alpha = \beta$ the above model degenerates to the ARMA(0, 0) model. To get estimates of α, β and σ^2 under the model ARMA(1, 1), we use the following algorithm (see Hannan and Rissanen [8]). As a convention, define $z_t = 0$ for $t \leqq 0$, for given observations $z_1, \cdots, z_n$. The time t runs from 1 to n in summation unless otherwise stated.

(1) Fit a long-lag autoregressive model AR(k_n), and obtain an estimate of a_t by

$$\tilde{a}_t = \sum_{l=0}^{k_n} \hat{\varphi}_l z_{t-l}$$

for $t \geqq 1$. Here, $\hat{\varphi}_0 = 1$ and

$$\sum_{l=1}^{k_n} \hat{\varphi}_l \left(\sum z_{t-1-l} z_{t-1-m} \right) = -\sum z_{t-1} z_{t-1-m}$$

for $1 \leqq m \leqq k_n$. The order k_n should be taken as diverging rapidly to ∞ with n but with k_n^2/n converging to 0.

(2) Obtain a pair of initial estimates $\tilde{\alpha}$ and $\tilde{\beta}$ by minimising

$$\sum (z_t - \alpha z_{t-1} + \beta \tilde{a}_{t-1})^2.$$

(3) Form

$$\hat{a}_t = \tilde{\beta} \hat{a}_{t-1} + z_t - \tilde{\alpha} z_{t-1} \quad \text{for} \quad t \geqq 1$$

initialising with $\hat{a}_0 = 0$. Next, form

$$\eta_t = \tilde{\alpha} \eta_{t-1} + \hat{a}_t$$

and

$$\xi_t = \tilde{\beta} \xi_{t-1} + \hat{a}_t \quad \text{for} \quad t \geqq 1,$$

initialising with $\eta_0 = \xi_0 = 0$. Regress $\hat{a}_t$ on η_{t-1} and ξ_{t-1}. Let $\tilde{\delta}$ and $\tilde{\varepsilon}$ be the δ and the ε which minimise

$$\sum_t (\hat{a}_t - \delta(\eta_{t-1} - \xi_{t-1}) + \varepsilon \xi_{t-1})^2.$$

Our estimates of α and β are then

$$\hat{\alpha} = \tilde{\delta} + \tilde{\alpha} \quad \text{and} \quad \hat{\beta} = \tilde{\delta} + \tilde{\varepsilon} + \tilde{\beta}, \tag{4.1}$$

respectively.

It is known that the estimates $\hat{\alpha}$ and $\hat{\beta}$ are asymptotically equivalent to the corresponding maximum likelihood estimates, provided that $\alpha \neq \beta$. The proof is partly given on page 92 of Hannan and Rissanen [8]. The full proof can be found in Shibata [15] or in Chen [3]. We now proceed to evaluate the asymptotic behaviour of those estimates in the case where $\alpha = \beta$. The following proposition is for the initial estimates $\tilde{\alpha}$ and $\tilde{\beta}$.

Proposition 4.1. If $z_1, \cdots, z_n$ are generated from an ARMA(0, 0) process, then $\sqrt{n}(\tilde{\alpha} - \tilde{\beta})$ is asymptotically normally distributed with mean 0 and variance 1. The initial estimates $\sqrt{k_n}\,\tilde{\alpha}$ and $\sqrt{k_n}\,\tilde{\beta}$ are asymptotically equivalent to each other, and distributed normally with mean 0 and variance 1.

Proof. Let $\Delta_t = \tilde{a}_t - z_t$, then

$$\begin{aligned} \Delta_{t-1} &= \sum_{l=0}^{k_n} \hat{\varphi}_l z_{t-1-l} - z_{t-1} \\ &= \sum_{l=1}^{k_n} \hat{\varphi}_l z_{t-1-l}, \end{aligned}$$

so that

$$\begin{aligned} \sum \Delta_{t-1}^2 &= \sum \left[\sum_{l,m=1}^{k_n} \hat{\varphi}_l \hat{\varphi}_m z_{t-1-l} z_{t-1-m} \right] \\ &= \sum_{l,m} \hat{\varphi}_l \hat{\varphi}_m \left[\sum z_{t-1} z_{t-1-m} \right] \\ &= -\sum_{l} \hat{\varphi}_l \left[\sum z_{t-1} z_{t-1-l} \right] \\ &= -\sum \Delta_{t-1} z_{t-1}. \end{aligned}$$

We have the following explicit expression of $\tilde{\alpha}$ and $\tilde{\beta}$ by solving the least-squares equation

$$\begin{bmatrix} \alpha \\ -\beta \end{bmatrix} = \frac{1}{D} \begin{bmatrix} \sum \tilde{a}_{t-1}^2 & -\sum z_{t-1} \tilde{a}_{t-1} \\ -\sum z_{t-1} \tilde{a}_{t-1} & \sum z_{t-1}^2 \end{bmatrix} \begin{bmatrix} \sum z_t z_{t-1} \\ \sum z_t \tilde{a}_{t-1} \end{bmatrix}.$$

Here D is the determinant of the matrix on the right-hand side of the above

equation.

$$\tilde{\alpha} = -\frac{\sum z_t \Delta_{t-1}}{\sum \Delta_{t-1}^2}$$

$$\tilde{\beta} = \frac{-\sum z_t z_{t-1} - \left[\frac{\sum z_{t-1}^2}{\sum \Delta_{t-1}^2}\right](\sum z_t \Delta_{t-1})}{\sum z_{t-1}^2 - \sum \Delta_{t-1}^2},$$

and

$$\tilde{\alpha} - \tilde{\beta} = \frac{\sum z_t z_{t-1} + \sum z_t \Delta_{t-1}}{\sum z_{t-1}^2 - \sum \Delta_{t-1}^2}.$$

From the assumption that $z_t = a_t$, we see that

$$\begin{aligned} \sum \Delta_{t-1}^2 &= \sum_{l,m} \sqrt{n}\hat{\varphi}_l \sqrt{n}\hat{\varphi}_m \left[\frac{1}{n}\sum a_{t-1-l} a_{t-1-m}\right] \\ &= \sigma^2 \sum_{l=1}^{k_n} (\sqrt{n}\hat{\varphi}_l)^2 \left[1 + O_p\left(\frac{1}{\sqrt{n}}\right)\right]. \end{aligned} \tag{4.2}$$

Since it is known (Shibata [11], [12]) that $\sqrt{n}\hat{\varphi}_1, \cdots, \sqrt{n}\hat{\varphi}_{k_n}$ are asymptotically distributed as independent normal random variables with mean 0 and variance 1 under the assumption that $z_t = a_t$, the right-hand side of (4.2) is asymptotically distributed as $\sigma^2 \chi_{k_n}^2$. In the same manner, from the evaluation

$$\sum z_t \Delta_{t-1} = \sigma^2 \sum_{l=1}^{k_n} \sqrt{n}\,\hat{\varphi}_l \sqrt{n}\,\hat{\varphi}_{l+1} \left[1 + O_p\left(\frac{1}{\sqrt{n}}\right)\right],$$

with $\hat{\varphi}_{k_n+1} = \frac{1}{n}\sum z_{t-k_n-2} z_{t-1}$, we see that $(\sum z_t \Delta_{t-1})/(\sqrt{k_n}\,\sigma^2)$ is asymptotically normally distributed with mean 0 and variance 1. The first part of the theorem then follows from

$$\sqrt{n}\,(\tilde{\alpha} - \tilde{\beta}) = \frac{\frac{1}{\sqrt{n}}\sum a_t a_{t-1}}{\frac{1}{n}\sum a_{t-1}^2}\{1 + O_p((k_n/n)^{\frac{1}{2}})\}. \tag{4.3}$$

For the latter part, it is enough to note that

$$\begin{aligned} \sqrt{k_n}\,\tilde{\beta} &= -\sqrt{k_n}\frac{\sum z_t \Delta_{t-1}}{\sum \Delta_{t-1}^2}\{1 + O_p((k_n/n)^{\frac{1}{2}})\} \\ &= -\frac{\frac{1}{\sqrt{k_n}}\sum_{l=1}^{k_n} \sqrt{n}\,\hat{\varphi}_l \sqrt{n}\,\hat{\varphi}_{l+1}}{\frac{1}{k_n}\sum_{l=1}^{k_n} (\sqrt{n}\,\hat{\varphi}_l)^2}\{1 + O_p((k_n/n)^{\frac{1}{2}})\}. \end{aligned}$$

It is worth noting that the proof of consistency of $\tilde{\alpha}$ and $\tilde{\beta}$ in Proposition 4.1 owes to the error structure of the autoregressive parameter estimates $\hat{\varphi}_1, \cdots, \hat{\varphi}_{k_n}$. The consistency does not necessarily follow if other estimates are used for obtaining an estimate of a_t, or if such errors are not taken into consideration (see Hannan [5], [6]).

Proposition 4.2. If $z_1, \cdots, z_n$ are generated from an ARMA(0, 0) process, then $\tilde{\delta}$ is asymptotically distributed as Cauchy. On the other hand, $\tilde{\varepsilon}$ is of the order $n^{-\frac{1}{2}}k^{-\frac{1}{2}}$ in probability, so that $\sqrt{n}(\hat{\alpha}-\hat{\beta})$ is asymptotically equivalent to $\sqrt{n}(\tilde{\alpha}-\tilde{\beta})$ and both $\hat{\alpha}$ and $\hat{\beta}$ are distributed as Cauchy.

Proof. Since

$$\eta_t = \sum_{l=0}^{t} \tilde{\alpha}^l a_{t-l} \quad \text{and} \quad \xi_t = \sum_{l=0}^{t} \tilde{\beta}^l \hat{a}_{t-l}$$

we obtain the following evaluations:

$$\sum (\eta_{t-1}-\xi_{t-1})^2 = (\tilde{\alpha}-\tilde{\beta})^2 \left(\sum \hat{a}_{t-2}^2\right)\{1+O_p(k_n^{-\frac{1}{2}})\},$$

$$\sum \xi_{t-1}^2 = \left(\sum a_{t-1}^2\right)\{1+O_p(k_n^{-\frac{1}{2}})\},$$

$$\sum \hat{a}_t(\eta_{t-1}-\xi_{t-1}) = (\tilde{\alpha}-\tilde{\beta})\left(\sum a_t a_{t-2}\right)\{1+O_p(k_n^{-\frac{1}{2}})\},$$

$$\sum \xi_{t-1}(\eta_{t-1}-\xi_{t-1}) = (\tilde{\alpha}-\tilde{\beta})\left\{\tilde{\beta}\sum a_{t-2}^2 + O_p(k_n^{\frac{1}{2}})\right\},$$

and

$$\begin{aligned}\sum \hat{a}_t\xi_{t-1} &= \left\{\sum a_t a_{t-1} + (\tilde{\beta}-\tilde{\alpha})\sum a_{t-1}^2\right\} + O_p(k_n^{-\frac{1}{2}}n^{\frac{1}{2}}) \\ &= O_p(k_n) + O_p(k_n^{-\frac{1}{2}}n^{\frac{1}{2}}) \\ &= O_p(k_n^{-\frac{1}{2}}n^{\frac{1}{2}})\end{aligned}$$

from the formula (4.3). By solving the least-squares equation for the step (3) in the algorithm, we have, as in the proof of Proposition 4.1,

$$\begin{aligned}\tilde{\delta} &= \frac{(\sum a_{t-1}^2)(\sum a_t a_{t-2}) + \tilde{\beta}(\sum a_{t-2}^2)O_p(k_n^{-\frac{1}{2}}n^{\frac{1}{2}})}{(\tilde{\alpha}-\tilde{\beta})(\sum a_{t-2}^2)(\sum a_{t-1}^2)}\{1+O_p(k_n^{-\frac{1}{2}})\} \\ &= \frac{\sum a_t a_{t-2} + O_p(k_n^{-\frac{1}{2}}n^{\frac{1}{2}})}{\sum a_t a_{t-1}}\{1+O_p(k_n^{-\frac{1}{2}})\} \\ &= \frac{\sum a_t a_{t-2}}{\sum a_t a_{t-1}}\{1+O_p(k_n^{-\frac{1}{2}})\}.\end{aligned}$$

Therefore the asymptotic distribution of $\tilde{\delta}$ is Cauchy. On the other hand, $\tilde{\varepsilon}$ can

be evaluated as

$$\tilde{\varepsilon} = \frac{\tilde{\beta}\,(\sum a_{t-2}^2)(\sum a_t a_{t-2}) - (\sum a_{t-2}^2) O_p(k_n^{-\frac{1}{2}} n^{\frac{1}{2}})}{(\sum a_{t-2}^2)(\sum a_{t-1}^2)} \{1 + O_p(k_n^{-\frac{1}{2}})\}$$
$$= O_p(k_n^{-\frac{1}{2}} n^{-\frac{1}{2}}).$$

This proposition implies that if $\alpha = \beta$, $\hat{\alpha}$ and $\hat{\beta}$ are not consistent but their asymptotic distribution is Cauchy, while the initial estimates $\tilde{\alpha}$ and $\tilde{\beta}$ are consistent with order $k_n^{\frac{1}{2}}$. This does not coincide with the result of Hannan [6] that the maximum likelihood estimates of α and β almost surely converge to ± 1, but this is because our estimates are not necessarily equivalent to the exact maximum likelihood estimates when $\alpha = \beta$.

Now let us consider the two criteria AIC and FPE. It is already known that these criteria are asymptotically equivalent to each other in the case of autoregressive models. But this does not hold for the selection of autoregressive–moving-average models. While AIC is likelihood-oriented, FPE is prediction-error-oriented. AIC for an ARMA(p, q) is, by definition,

$$\text{AIC}(p, q) = \log \hat{\sigma}_{\text{ML}}^2(p, q) + 2(p+q)/n,$$

where $\hat{\sigma}_{\text{ML}}^2(p, q)$ is the maximum likelihood estimate of σ^2 under the model ARMA(p, q). But FPE is different from AIC.

To obtain an explicit form of FPE for the ARMA$(1, 1)$ case, consider another ARMA$(1, 1)$ process $\{z_t^*\}$ which is independent of $\{z_t\}$ but which has the same probabilistic structure as $\{z_t\}$. Since the best predictor of z_t^* is $\bar{z}_t^* = \alpha z_{t-1}^* - \beta a_{t-1}^*$, the error of the estimated predictor

$$\hat{z}_t^* = \hat{\alpha} z_{t-1}^* - \hat{\beta} a_{t-1}^*$$

is given by

$$E^*(\hat{z}_t^* - z_t^*)^2 = E^*(\bar{z}_t^* - z_t^*)^2 + E^*(\hat{z}_t^* - \bar{z}_t^*)^2$$
$$= \sigma^2\{1 + (\hat{\theta} - \theta)' Q (\hat{\theta} - \theta)\}. \tag{4.4}$$

Here E^* denotes the expectation with respect to $\{z_t^*\}$, $\hat{\theta}' = (\hat{\alpha}, \hat{\beta}')$, $\theta' = (\alpha, \beta')$,

$$Q = \begin{bmatrix} \gamma_0/\sigma^2 & -1 \\ -1 & 1 \end{bmatrix}$$

and $\gamma_k = E^*(z_t^* z_{t-k}^*) = E(z_t z_{t-k})$. From the fact that the underlying model is an ARMA$(1, 1)$, the variance γ_0 of $\{z_t\}$ is written as

$$\gamma_0 = \sigma^2 \left\{ 1 + \frac{(\alpha - \beta)^2}{1 - \alpha^2} \right\}.$$

Since $\hat{\theta}$ is asymptotically equivalent to the corresponding maximum likelihood estimates unless $\alpha = \beta$, the estimation error $\sqrt{n}(\hat{\theta} - \theta)$ is asymptotically

normally distributed with the mean 0 and the covariance matrix J^{-1}, where

$$J=\begin{bmatrix} 1/(1-\alpha^2) & -1/(1-\alpha\beta) \\ -1/(1-\alpha\beta) & 1/(1-\beta^2) \end{bmatrix}.$$

The expectation of the right-hand side of (4.3) is then asymptotically equal to

$$\sigma^2\{1+\operatorname{tr}(J^{-1}Q)/n\}=\sigma^2\{1+2(1-\alpha\beta)/n\}.$$

Replacing σ^2 by a bias-corrected estimate $\hat{\sigma}^2_{\mathrm{ML}}(1,1)/(1-2/n)$, α by $\hat{\alpha}$, and β by $\hat{\beta}$, we obtain the FPE for ARMA(1, 1),

$$\mathrm{FPE}(1,1)=\frac{\hat{\sigma}^2_{\mathrm{ML}}(1,1)\{1+2(1-\hat{\alpha}\hat{\beta})/n\}}{1-2/n}.$$

The difference between AIC(1, 1) and FPE(1, 1) is significant unless $\hat{\alpha}$ and $\hat{\beta}$ converge to 0.

Now we proceed to evaluate the behaviour of AIC and FPE. We first consider the case where

$$\hat{\sigma}^2(1,1)=\frac{1}{n}\sum(z_t-\hat{\alpha}z_{t-1}+\hat{\beta}\hat{a}_{t-1})^2 \tag{4.5}$$

is used in place of $\hat{\sigma}^2_{\mathrm{ML}}(1,1)$. We should note that the above estimate is based on $\hat{\alpha}$ and $\hat{\beta}$, but is not itself asymptotically efficient even when $\alpha\neq\beta$.

Let us define

$$\hat{\sigma}^2=\frac{1}{n}\sum a_t^2, \qquad \hat{\rho}_1=\Big(\sum a_t a_{t-1}\Big)\Big/\Big(\sum a_t^2\Big)$$

and $\hat{\rho}_2=(\sum a_t a_{t-2})/(\sum a_t^2)$, then from Propositions 4.1 and 4.2, if $\alpha=\beta$,

$$\hat{\beta}-\hat{\alpha}=-2\hat{\rho}_1(1+O_p(k_n^{-\frac{1}{2}})), \qquad \hat{\beta}=\frac{\hat{\rho}_2}{\hat{\rho}_1}(1+O_p(k_n^{-\frac{1}{2}}))$$

and

$$\begin{aligned}\hat{a}_{t-1}-a_{t-1}&=(\tilde{\beta}-\tilde{\alpha})a_{t-2}(1+O_p(k_n^{-\frac{1}{2}}))\\&=\hat{\rho}_1 a_{t-2}(1+O_p(k_n^{-\frac{1}{2}})).\end{aligned}$$

Therefore, if $\alpha=\beta$, then $z_t=a_t$ and

$$\begin{aligned}\hat{\sigma}^2(1,1)&=\frac{1}{n}\sum(a_t-\hat{\alpha}a_{t-1}+\hat{\beta}\hat{a}_{t-1})^2\\&=\frac{1}{n}\sum\{a_t+(\hat{\beta}-\hat{\alpha})a_{t-1}+\hat{\beta}(\hat{a}_{t-1}-a_{t-1})\}^2\\&=\frac{1}{n}\sum[a_t-(\hat{\rho}_1 a_{t-1}+\hat{\rho}_2 a_{t-2})\{1+O_p(k_n^{-\frac{1}{2}})\}]^2\\&=\hat{\sigma}^2-\hat{\sigma}^2(\hat{\rho}_1^2+\hat{\rho}_2^2+2\hat{\rho}_1^2\hat{\rho}_2)\{1+O_p(k_n^{-\frac{1}{2}})\}.\end{aligned}$$

Since $\hat{\sigma}^2_{\mathrm{ML}}(0,0)=\hat{\sigma}^2$, from the definition, $\mathrm{AIC}(0,0)=\log\hat{\sigma}^2_{\mathrm{ML}}(0,0)$ and $\mathrm{FPE}(0,0)=\hat{\sigma}^2_{\mathrm{ML}}(0,0)$, we have

$$\begin{aligned}\mathrm{AIC}(1,1)-\mathrm{AIC}(0,0)&=\log\{1-(\hat{\rho}_1^2+\hat{\rho}_2^2+2\hat{\rho}_1^2\hat{\rho}_2)(1+O_p(k_n^{-\frac{1}{2}}))\}+4/n\\&=\frac{1}{n}\{4-n\hat{\rho}_1^2-n\hat{\rho}_2^2+O_p(k_n^{-\frac{1}{2}})\},\end{aligned}$$

and

$$\begin{aligned}\mathrm{FPE}(1,1)-\mathrm{FPE}(0,0)&=\frac{\hat{\sigma}^2}{1-2/n}[(1-\hat{\rho}_1^2-\hat{\rho}_2^2)\{1+2(1-\hat{\rho}_2^2/\hat{\rho}_1^2-2\hat{\rho}_2)/n\}]-\hat{\sigma}^2\\&=\frac{\hat{\sigma}^2}{n}\{4-n\hat{\rho}_1^2-n\hat{\rho}_2^2-2\hat{\rho}_2^2/\hat{\rho}_1^2+O_p(k_n^{-\frac{1}{2}})\}.\end{aligned}$$

Therefore, under the assumption that $\alpha=\beta$

$$\begin{aligned}\lim_{n\to\infty}P(\mathrm{AIC}(0,0)<\mathrm{AIC}(1,1))&=\lim P(n\hat{\rho}_1^2+n\hat{\rho}_2^2<4)\\&=P(X<4)=0.86466,\end{aligned}$$

and

$$\begin{aligned}\lim_{n\to\infty}P(\mathrm{FPE}(0,0)<\mathrm{FPE}(1,1))&=\lim P(n\hat{\rho}_1^2+n\hat{\rho}_2^2+2\hat{\rho}_2^2/\hat{\rho}_1^2<4)\\&=P(X+2Y<4)=0.45120,\end{aligned}$$

where X and Y are independent, and distributed as χ^2 with degree of freedom 2, and as Cauchy, respectively. It is obvious that if $\alpha\neq\beta$, then $\mathrm{AIC}(1,1)<\mathrm{AIC}(0,0)$ and $\mathrm{FPE}(1,1)<\mathrm{FPE}(0,0)$ hold true in probability for large enough n, and we have the following theorem.

Theorem 4.1. Suppose that the competing models are only ARMA(1, 1) and ARMA(0, 0). If $\hat{\sigma}^2(1,1)$ is used in place of the exact maximum likelihood estimate, then the probability of correct selection tends to 1 if $\alpha\neq\beta$, otherwise it tends to 0.86466 and 0.45120, respectively, for AIC and FPE.

It is worth noting that in spite of non-regularity, the error probability of AIC is the same as that in the regular case, for example, in the case of AR(p) fitting (Shibata [12]), while that of FPE is different and higher.

Next we consider the case where

$$\tilde{\sigma}^2(1,1)=\frac{1}{n}\sum(z_t-\tilde{\alpha}z_{t-1}+\tilde{\beta}\hat{a}_{t-1})^2 \tag{4.6}$$

is used instead of $\hat{\sigma}^2(1,1)$.

Under the assumption that $\alpha = \beta$, it follows that

$$\begin{aligned}\tilde{\sigma}^2(1,1) - \tilde{\sigma}^2(0,0) &= \frac{1}{n}\sum \hat{a}_t^2 - \frac{1}{n}\sum a_t^2 \\ &= (\tilde{\beta}-\tilde{\alpha})^2 \frac{1}{n}\sum a_{t-1}^2 + 2(\tilde{\beta}-\tilde{\alpha})\frac{1}{n}\sum a_t a_{t-1} \\ &= \hat{\sigma}^2\{(\tilde{\beta}-\tilde{\alpha})^2 + 2(\tilde{\beta}-\tilde{\alpha})\hat{\rho}_1\} \\ &= -\hat{\sigma}^2\hat{\rho}_1^2.\end{aligned}$$

By similar discussion as in the proof of Theorem 4.1, we have

$$\lim_{n\to\infty} P(\text{AIC}(0,0) < \text{AIC}(1,1)) = P(\chi_1^2 < 4) = 0.9545,$$

provided that $\alpha = \beta$.

Theorem 4.2. Suppose that the competing models are only ARMA(1, 1) and ARMA(0, 0). If $\tilde{\sigma}^2(1,1)$ is used in place of the exact maximum likelihood estimate in the definition of AIC(1, 1), then the probability of correct selection tends to 1 if $\alpha \neq \beta$, otherwise it tends to 0.9545.

Therefore we can see that $\tilde{\sigma}^2(1,1)$ is better than $\hat{\sigma}^2(1,1)$ as a replacement of $\hat{\sigma}_{\text{ML}}^2(1,1)$ in the definition of AIC(1, 1). However, there still remains a small probability of selecting the overfitted model ARMA(1, 1), where both $\hat{\alpha}$ and $\hat{\beta}$ are not consistent. On the other hand, we have already seen that if we require consistency of selection as in HQ and BIC, then the least order of consistency becomes lower than $\sqrt{n}$. An alternative is to modify AIC as follows. For simplicity, consider the models ARMA(p, q) $0 \leqq p, q \leqq 1$. Select ARMA(0, 0) if $|\tilde{\delta}| > \alpha_n/\sqrt{n}$, otherwise select the model which minimises AIC or FPE from those models except ARMA(0, 0). Here, α_n is a divergent sequence with n as used in the definition of the HQ or the BIC, which diverges to ∞ with $\alpha_n/\sqrt{n}$ converging to 0. By this modification, AIC and FPE correctly select ARMA(0, 0) when $\alpha = \beta$, because $\tilde{\delta}$ is distributed as Cauchy. In this paper, we do not give any further proof, but it can be easily understood that by such a modification the order of consistency goes down to $O((n/\alpha_n)^{\frac{1}{2}})$ only in a domain like

$$\{(\alpha, \beta); (\alpha_n/n)^{\frac{1}{2}}(1-\varepsilon) < |\alpha - \beta| < (\alpha_n/n)^{\frac{1}{2}}\},$$

while that of BIC or HQ also goes down in a domain like

$$\{(\alpha, \beta); (\alpha_n/n)^{\frac{1}{2}}(1-\varepsilon) < |\alpha| < (\alpha_n/n)^{\frac{1}{2}}\}$$

or

$$\{(\alpha, \beta); (\alpha_n/n)^{\frac{1}{2}}(1-\varepsilon) < |\beta| < (\alpha_n/n)^{\frac{1}{2}}\}.$$

More detailed analysis and extension remain for the future.

Acknowledgements

The author would like to express his sincere thanks to the referee. His suggestions led to the correction of some serious errors in the original manuscript; and also improved the readability of the paper. Theorem 4.2 is due to the referee.

References

[1] AKAIKE, H. (1970) Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203–217.

[2] AKAIKE, H. (1973) Information theory and an extension of the maximum likelihood principle. In *2nd Int. Symposium on Information Theory*, ed. B. N. Petrov and F. Csáki, Akadémia Kiado, Budapest, 267–281.

[3] CHEN, ZHAO-GUO (1984) The asymptotic efficiency of a linear procedure of estimation for ARMA models. *J. Time Series Anal.*

[4] HANNAN, E. J. (1980) The estimation of the order of an ARMA process *Ann. Statist.* **8**, 1071–1081.

[5] HANNAN, E. J. (1982) Testing for autocorrelation and Akaike's criterion. In *Essays in Statistical Science*: *Papers in Honour of P. A. P. Moran*, ed. J. M. Gani and E. J. Hannan, Applied Probability Trust, Sheffield, 403–412.

[6] HANNAN, E. J. (1982) Fitting multivariate ARMA models. In *Statistics and Probability, Essays in Honour of C. R. Rao*, ed. G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh, North-Holland, Amsterdam, 307–316.

[7] HANNAN, E. J. AND QUINN, B. G. (1979) The determination of the order of an autoregression. *J. R. Statist. Soc.* B **41**, 190–195.

[8] HANNAN, E. J. AND RISSANEN, J. (1982) Recursive estimation of mixed autoregressive-moving average order. *Biometrika* **69**, 81–94.

[9] MALLOWS, C. L. (1973) Some comments on C_p. *Technometrics* **15**, 661–675.

[10] SCHWARZ, G. (1978) Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

[11] SHIBATA, R. (1976) Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63**, 117–126.

[12] SHIBATA, R. (1977) Convergences of least squares estimates of autoregressive parameters. *Austral. J. Statist.* **19**, 226–235.

[13] SHIBATA, R. (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8**, 147–164.

[14] SHIBATA, R. (1983) A theoretical view of the use of AIC. In *Time Series Analysis: Theory and Practice* **4**, ed. O. D. Anderson, Elsevier, Amsterdam, 237–244.

[15] SHIBATA, R. (1984) Identification and selection of ARMA models. Tech. Rep. RT-MAE-8406, Institute of Mathematics and Statistics, University of Sào Paulo.