Ann. Inst. Statist. Math.
30 (1978), Part A, 263-270

ENTROPY MAXIMIZATION PRINCIPLE AND SELECTION OF THE ORDER OF AN AUTOREGRESSIVE GAUSSIAN PROCESS*

RYOICHI SHIMIZU

(Received Aug. 16, 1977; revised July 13, 1978)

1. Introduction and summary

Let $\boldsymbol{X}=\{X_1, X_2, \cdots, X_N\}$ be a set of successive observations on a stationary system with the autoregressive structure

$$(1) \qquad X_t=a_1X_{t-1}+a_2X_{t-2}+\cdots+a_LX_{t-L}+\varepsilon_t\ , \qquad t=\cdots, -1, 0, 1, 2, \cdots,$$

where ε's are supposed to be independent normal variables with mean 0 and variance σ^2, and where the parameter $\theta=(\sigma^2, a_1, a_2, \cdots, a_L)$ is contained in the set Θ of all vectors $(c_0, c_1, \cdots, c_L)$ such that $c_0>0$ and the zeros of the polynomial $x^L-c_1x^{L-1}-c_2x^{L-2}-\cdots-c_L$ are located in the unite circle in the complex plane, which guarantees stationarity of the system.

The structure (1) is said to be a pth order autoregressive model and is denoted by AR(p) if $a_p\neq 0$ and if $a_{p+1}=a_{p+2}=\cdots=a_L=0$. We are concerned with determination of the order p as well as estimation of σ^2 and a's. If p is known, i.e., if we know in advance that $a_{p+1}=a_{p+2}=\cdots=a_L=0$, then the maximum likelihood principle will provide good estimates for σ^2 and a's. The principle does not, however, apply if we want to estimate not only σ^2 and a's, but also the order p itself.

Let $\hat{\theta}=\hat{\theta}(\boldsymbol{X})$ be an estimate of θ based on $\boldsymbol{X}$ and let $\boldsymbol{Z}=(Z_1, Z_2, \cdots, Z_N)$ be a set of observations taken from the system described by (1) independently of $\boldsymbol{X}$. The probability density $f(\boldsymbol{z}; \theta)$ for $\boldsymbol{Z}$ will be estimated by $f(\boldsymbol{z}; \hat{\theta}(\boldsymbol{X}))$ which we call a predictive density function for $\boldsymbol{Z}$.

If $\boldsymbol{X}$ is given, we can measure the distance between the predictive and true densities for $\boldsymbol{Z}$ by (twice of) the Kullback-Leibler information:

$$(2) \qquad I=I(\hat{\theta})=-2\,\mathrm{E}_{\boldsymbol{Z}} \log \frac{f(\boldsymbol{Z}; \hat{\theta})}{f(\boldsymbol{Z}; \theta)} \geqq 0\ .$$

* This research was partly supported by a National Grant in Aid for Scientific Research, 1976/77, no. 220928.

Estimating I by its sample analogue

$$(3) \qquad I^*(\hat{\theta}) = -2 \log \frac{f(X;\hat{\theta})}{f(X;\theta)} ,$$

Akaike ([2], [3]) introduced as a result of asymptotic theory a criterion called AIC for evaluating badness of the estimated distribution. He then proposed AIC for determination of the order p of an autoregressive model.

The purpose of the present paper is to study the relation between I and I^* and to look into the asymptotic behavior of the criterion AIC.

2. Entropy maximization principle and AIC

The joint probability density function of the sample X from an AR(p) is of the form

$$(4) \qquad f(x;\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^N \cdot \sigma^{-(N-p)} \cdot |\Sigma|^{-1/2} \cdot \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{l,m=1}^{p} \sigma^{l,m} x_l x_m + \sum_{j=p+1}^{N} (x_j - a_1 x_{j-1} - \cdots - a_p x_{j-p})^2\right)\right\} ,$$

where $\Sigma = (\sigma_{l-m})_{l,m=1,2,\cdots,p}$ is the covariance matrix of $(X_1, X_2, \cdots, X_p)$, $(\sigma^{l,m})_{l,m=1,2,\cdots,p}$ is the inverse matrix of $\sigma^{-2}\Sigma$, and where $\theta = (\sigma^2; a_1, a_2, \cdots, a_L)$ with $a_{p+1} = a_{p+2} = \cdots = a_L = 0$ is the parameter value specifying the density function.

Following Akaike ([3], [5]), we wish to determine $\hat{\theta}(X)$ in such a way that the expected value $\mathrm{E}_X I(\hat{\theta}(X))$ of $I(\hat{\theta})$ be minimized, or what is the same thing that the expected entropy $-(1/2)\,\mathrm{E}_X I(\hat{\theta}(X))$ of $f(z;\theta)$ with respect to $f(z;\hat{\theta})$ be maximized. We shall confine ourselves to the estimates of the form

$$\hat{\theta}_K = (\hat{\sigma}^2, \hat{a}_1, \hat{a}_2, \cdots, \hat{a}_K, 0, \cdots, 0) ,$$

where $K = K(X)$ is an estimate of the order p, and where, if $K = k$ is given, $\hat{\theta}_k = (\hat{\sigma}^2, \hat{a}_1, \cdots, \hat{a}_k, 0, \cdots, 0)$ constitutes the approximate maximum likelihood estimate of θ defined by the Yule-Walker equation,

$$(5) \qquad C_l = \hat{a}_1 C_{l-1} + \hat{a}_2 C_{l-2} + \cdots + \hat{a}_k C_{l-k} , \qquad l = 1, 2, \cdots, k ,$$

and

$$(6) \qquad \hat{\sigma}^2 = \hat{\sigma}_k^2 \equiv C_0 - \sum_{l=1}^{k} \hat{a}_l C_l ,$$

where $C_l = C_l(X) = \sum_{j=1}^{N} X_j X_{j-|l|}/N$ with the convention $X_0 = X_{-1} = X_{-2} = \cdots$

$=0$. This amounts to approximating the density (4) by

$$(7)\qquad f(\boldsymbol{x};\theta)=\text{const.}\cdot\sigma^{-N}\cdot\exp\left\{-\frac{1}{2\sigma^2}S(\boldsymbol{x};a_1,\cdots,a_p)\right\},$$

where

$$S(\boldsymbol{x};a_1,\cdots,a_p)=N\cdot\left(C_0(\boldsymbol{x})-2\sum_{l=1}^{p}a_lC_l(\boldsymbol{x})+\sum_{l,m=1}^{p}a_la_mC_{l-m}(\boldsymbol{x})\right),$$

and maximizing it with respect to θ assuming that $p=k$.

If $k\geqq p$, then both $I(\hat{\theta}_k)$ and $-I^*(\hat{\theta}_k)$ will be asymptotically distributed according to the chi-square distribution with degrees of freedom $k+1$. On the other hand if $p>k$, then $I(\hat{\theta}_k)/N$ and $I^*(\hat{\theta}_k)/N$ converge to a common positive number (see [1], [6], [7] and Proposition 2 in the next section). Therefore, for sufficiently large N, $\mathrm{E}_{\boldsymbol{X}}I(\hat{\theta}_k)$ is estimated by $\mathrm{E}_{\boldsymbol{X}}I^*(\hat{\theta}_k)+2(k+1)$ with a bias, if $k<p$, of the order $\mathrm{E}_{\boldsymbol{X}}I(\hat{\theta}_k)/N$. Thus the minimization of $\mathrm{E}_{\boldsymbol{X}}I(\hat{\theta}_k)$ reduces to that of $\mathrm{E}_{\boldsymbol{X}}I^*(\hat{\theta}_k)+2(k+1)$, or equivalently, that of $J_k\equiv-2\,\mathrm{E}_{\boldsymbol{X}}\log f(\boldsymbol{X};\hat{\theta}_k)+2k$. Akaike [3] estimates J_k, $k=0,1,\cdots,L$ by their unbiased estimates

$$(8)\qquad \mathrm{AIC}(k)=-2\log f(\boldsymbol{X};\hat{\theta}_k)+2k\,,\qquad k=0,1,\cdots,L\,.$$

He then proposes AIC (k) as a measure of badness of the estimated model

$$(9)\qquad Z_t=\hat{a}_1Z_{t-1}+\cdots+\hat{a}_kZ_{t-k}+\delta_t\,,$$

claiming that the larger AIC (k) is, the worse is the model (9).

Based on this idea he further proposes to use AIC for determining the order p. His method, called minimum AIC estimate (MAICE, for short) consists of calculating (8) for $k=0,1,\cdots,L$ and choosing k_0 as an estimate of p if AIC (k) attains its minimum at $k=k_0$.

3. The relation between $I(\hat{\theta}_k)$ and $I^*(\hat{\theta}_k)$

In this section we shall investigate the relations between $I(\hat{\theta}_k)$ and its sample analogue $I^*(\hat{\theta}_k)$ as defined in the previous section. Write $q=\max(k,p)$ and let $\Sigma=(\sigma_{l-m})_{l,m=1,2,\cdots,q}$ be the covariance matrix of $(Z_j, Z_{j+1},\cdots,Z_{j+q-1})$. Let $(b_1,\cdots,b_k)$ be the unique solution of

$$(10)\qquad \sigma_l=b_1\sigma_{l-1}+b_2\sigma_{l-2}+\cdots+b_k\sigma_{l-k}\,,\qquad l=1,2,\cdots,k\,,$$

and put $b_{k+1}=\cdots=b_p=0$ if $p>k\geqq 0$. Note that $a_l=b_l$, $l=1,2,\cdots,q$ hold if and only if $k\geqq p$, and that b's are characterized by

$$\text{(11)} \qquad \sigma^2(k) \equiv \min_{c_1,\cdots,c_k} \mathrm{E}\,(Z_j - c_1 Z_{j-1} - \cdots - c_k Z_{j-k})^2 = \mathrm{E}\,(Z_j - b_1 Z_{j-1} - \cdots - b_k Z_{j-k})^2 .$$

Also we have

$$\text{(12)} \qquad \sigma^2(k) = \sigma_0 - \sum_{l=1}^{k} b_l \sigma_l$$

$$\text{(13)} \qquad = \sigma^2 + \sum_{l,m=1}^{q} (a_l - b_l)(a_m - b_m)\sigma_{l-m} ,$$

the second term of (13) vanishing if and only if $k \geqq p$. We shall write $\hat{a}_{k+1} = \cdots = \hat{a}_p = 0$ whenever $p > k$. Then in view of (7) and (10), $I(\hat{\theta}_k)$ and $I^*(\hat{\theta}_k)$ are put in the following forms up to the terms which converge to zero with probability one:

$$\text{(14)} \qquad I(\hat{\theta}_k) = N \log \hat{\sigma}^2/\sigma^2 - N + N\sigma^2/\hat{\sigma}^2 + P/\hat{\sigma}^2$$

and

$$\text{(15)} \qquad I^*(\hat{\theta}_k) = N \log \hat{\sigma}^2/\sigma^2 + N - N\hat{\sigma}^2/\sigma^2 + Q/\sigma^2 ,$$

where

$$\text{(16)} \qquad P = P_k \equiv \mathrm{E}_{\boldsymbol{Z}}\{S(\boldsymbol{Z}; \hat{a}_1, \cdots, \hat{a}_k) - S(\boldsymbol{Z}; a_1, \cdots, a_p)\} = N \sum_{l,m=1}^{q} (\hat{a}_l - a_l)(\hat{a}_m - a_m)\sigma_{l-m} ,$$

and

$$\text{(17)} \qquad Q = Q_k = S(\boldsymbol{X}; \hat{a}_1, \cdots, \hat{a}_k) - S(\boldsymbol{X}; a_1, \cdots, a_p) = N \sum_{l,m=1}^{q} (\hat{a}_l - a_l)(\hat{a}_m - a_m) C_{l-m}(\boldsymbol{X}) - 2N \sum_{l=1}^{q} (\hat{a}_l - a_l)\left(C_l(\boldsymbol{X}) - \sum_{m=1}^{p} a_m C_{l-m}(\boldsymbol{X})\right) .$$

Now, we shall prove

PROPOSITION 1. *If $k \geqq p$, then*

$$\text{(18)} \qquad \lim_{N\to\infty} (I(\hat{\theta}_k) + I^*(\hat{\theta}_k)) = 0 \qquad \textit{in } \mathrm{P} ,$$

and

PROPOSITION 2. *For $k \geqq 0$ we have with probability one,*

$$\text{(19)} \qquad \lim_{N\to\infty} I(\hat{\theta}_k)/N = \lim_{N\to\infty} I^*(\hat{\theta}_k)/N = \log \sigma^2(k)/\sigma^2 \begin{cases} >0 , & \textit{if } k<p , \\ =0 , & \textit{if } k \geqq p , \end{cases}$$

and

$$\lim_{N\to\infty}(I^*(\hat{\theta}_k)/I(\hat{\theta}_k))=\begin{cases}1, & \text{if } k<p,\\ -1, & \text{if } k\geqq p.\end{cases} \tag{20}$$

As an illustration we give in the table below numerical values of $\hat{\sigma}_k^2$, $I(\hat{\theta}_k)$ and $I^*(\hat{\theta}_k)$ for a single observation of size $N=500$.

Values of $\hat{\sigma}_k^2$, $I(\hat{\theta}_k)$ and $I^*(\hat{\theta}_k)$ for a single observation of size $N=500$ from the AR(2) with $\theta=(1.0, 1.1, -0.5, 0.0, \cdots, 0.0)$, $\sigma^2(0)=2.885$, $\sigma^2(1)=1.335$, and $\sigma^2(2)=\sigma^2(3)=\cdots=1.0$.

k	0	1	2	3	4	5	6	7	8	9	10
σ_k^2	2.776	1.341	1.034	1.033	1.033	1.032	1.031	1.027	1.023	1.023	1.023
$I(\hat{\theta}_k)$	528.726	143.499	1.189	1.211	1.309	1.953	2.456	4.572	6.651	6.635	6.769
$I^*(\hat{\theta}_k)$	492.720	129.067	−1.225	−1.254	−1.361	−2.018	−2.453	−4.436	−6.167	−6.315	−6.410

To prove propositions we require some lemmas.

LEMMA 1. *For $l=0, 1, \cdots, q$, $C_l(X)$ converges to σ_l with probability one. If $k\geqq p$, then the distribution of $\sqrt{N}(\hat{\sigma}^2-\sigma^2)$ and the joint distribution of $\sqrt{N}(\hat{a}_l-a_l)$, $l=1, \cdots, k$ converge, respectively, to the normal distribution with mean zero and the k-dimentional normal distribution with mean vector* 0 *and covariance matrix $\sigma^2\Sigma^{-1}$, where $\Sigma=(\sigma_{l-m})$.*

For the proof see, e.g., Akaike ([1]), Doob ([6], pp. 493–498) and Hannan ([7], pp. 326–333).

LEMMA 2. *With probability one,*

$$\lim_{N\to\infty}\hat{a}_l=b_l\ , \qquad l=1, 2, \cdots, q\ . \tag{21}$$

and

$$\lim_{N\to\infty}\hat{\sigma}^2=\sigma^2(k)\begin{cases}>\sigma^2, & \text{if } k<p\\ =\sigma^2, & \text{if } k\geqq p.\end{cases} \tag{22}$$

PROOF. The assertions follow easily from Lemma 1 and the relations (5), (6), (10), (12) and (13) as well as the positive definiteness of Σ.

LEMMA 3. *For $k\geqq 0$ we have with probability one,*

$$\lim_{N\to\infty}P_k/N=\lim_{N\to\infty}Q_k/N=\sum_{l,m=1}^{q}(a_l-b_l)(a_m-b_m)\sigma_{l-m}=\sigma^2(k)-\sigma^2\ , \tag{23}$$

and

$$\lim_{N\to\infty}Q_k/P_k=\begin{cases}1, & \text{if } k<p\\ -1, & \text{if } k\geqq p.\end{cases} \tag{24}$$

If $k \geqq p$, *then*

$$\lim_{N\to\infty}(P_k+Q_k)=0 \qquad in\ \mathrm{P}. \tag{25}$$

PROOF. The assertions (23), and (24) for the case $k<p$ are simple consequences of the expressions (16)–(17) and Lemmas 1–2. Now suppose $q=k\geqq p$. Then we can use the relation (5) to reduce the expression (17) of Q_k to

$$Q_k=-N\sum_{l,m=1}^{q}(\hat{a}_l-a_l)(\hat{a}_m-a_m)C_{l-m}(X), \tag{26}$$

and the assertion (25) follows at once from (16), (26) and Lemma 1. To complete the proof of (24) let U be the orthogonal matrix of order $k(=q)$ such that $U'\Sigma U$ is a diagonal matrix with diagonal elements $\tau_l>0$, $l=1,\cdots,k$. Put $C=(C_{l-m})_{l,m=1,\cdots,k}$, and let $\tau^*_{l,m}$ be the l-m element of the matrix $U'(\Sigma-C)U$ and let B_j be the jth element of the row vector $(\hat{a}_1-a_1, \hat{a}_2-a_2,\cdots,\hat{a}_k-a_k)\cdot U$. Let, finally, B be the maximum of $|B_l|$ and τ be the minimum of τ_l respectively. Note that τ depends only on σ^2 and a's. It, then, follows from (16) and (26) that

$$\left|\frac{Q_k}{P_k}+1\right|=\left|\frac{\sum_{l,m=1}^{k}A_lA_m(\sigma_{l-m}-C_{l-m})}{\sum_{l,m=1}^{k}A_lA_m\sigma_{l-m}}\right|=\left|\frac{\sum_{l,m=1}^{k}B_lB_m\tau^*_{l,m}}{\sum_{l=1}^{k}B_l^2\tau_l^2}\right|$$

$$\leqq\frac{B^2\cdot\sum_{l,m=1}^{k}|\tau^*_{l,m}|}{\tau^2\cdot\sum_{l=1}^{k}B_l^2}\leqq\frac{k^2}{\tau^2}\max_{l,m}|\tau^*_{l,m}|,$$

and $\max|\tau^*_{l,m}|$ converges to zero by Lemma 1.

PROOF OF PROPOSITIONS. For any $k\geqq 0$, the assertions (19) of Proposition 2 are simple consequences of expressions (14)–(15) and Lemmas 2–3. They imply in turn the assertion (20) for $k<p$. Suppose next that $k\geqq p$ and let N be sufficiently large so that both $|\hat{\sigma}^2/\sigma^2-1|$ and $|\sigma^2/\hat{\sigma}^2-1|$ be less than 1/4. It follows that

$$\log\hat{\sigma}^2/\sigma^2=\log(1-(1-\hat{\sigma}^2/\sigma^2))=-(1-\hat{\sigma}^2/\sigma^2)-\frac{1}{2}(1-\hat{\sigma}^2/\sigma^2)^2+\alpha_N(\hat{\sigma}^2/\sigma^2-1)^3$$

$$=-\log(1-(1-\sigma^2/\hat{\sigma}^2))=(1-\sigma^2/\hat{\sigma}^2)+\frac{1}{2}(1-\sigma^2/\hat{\sigma}^2)^2+\beta_N(\hat{\sigma}^2/\sigma^2-1)^3$$

where $|\alpha_N|\leqq 1/2$ and $|\beta_N|\leqq 1$.

Then (14) and (15) reduce, respectively, to

$$I(\hat{\theta}_k)=N(\hat{\sigma}^2-\sigma^2)^2/2\hat{\sigma}^4+P/\hat{\sigma}^2+\beta_N N(\hat{\sigma}^2-\sigma^2)^3/\sigma^6, \tag{27}$$

and

$$(28)\qquad I^*(\hat{\theta}_k) = -N(\hat{\sigma}^2-\sigma^2)^2/2\sigma^4 + Q/\sigma^2 + \alpha_N N(\hat{\sigma}^2-\sigma^2)^3/\sigma^6\,,$$

and Proposition 1 follows from Lemmas 1–3. Also we have from (27) that

$$(29)\qquad I(\hat{\theta}_k) \geqq N(\hat{\sigma}^2-\sigma^2)^2/4\sigma^2 + P/2\sigma^2 > 0$$

and hence from (27)–(29) that

$$(30)\qquad \left|\frac{I^*(\hat{\theta}_k)}{I(\hat{\theta}_k)}+1\right| \leqq 2\sigma^2\left|\frac{1}{\hat{\sigma}^4}-\frac{1}{\sigma^4}\right| + 2\left|\frac{\sigma^2}{\hat{\sigma}^2}+\frac{Q}{P}\right| + \frac{4}{\sigma^4}|\alpha_N+\beta_N|\cdot|\hat{\sigma}^2-\sigma^2|\,.$$

In view of Lemmas 2 and 3, each term of the right-hand side of (30) converges to zero with probability one, proving (20) for $k \geqq p$.

Remark. The content of Proposition 1 was roughly stated by H. Akaike (Model selection and AIC, *Proceedings of the Symposium on Data Analyses for Natural Sciences*, Tokyo, 1976, pp. 63–67, in Japanese). His argument was based on the remark that the behavior of $I(\tau)$, as a function of $\tau \in \Theta$, in the neighbourhood of $\tau=\theta$ is well approximated by that of $I^*(\tau)$ in the neighbourhood of $\tau=\hat{\theta}$.

4. Relation between $I(\hat{\theta}_k)$ and AIC

As was stated in Section 2, E(AIC(k)) attains its minimum at $k=p$, which provides the theoretical basis of the MAICE. However, this does not imply that AIC(k)>AIC(p) even when N is sufficiently large, unless $k<p$, in which case the probability that this inequality holds tends to 1 as N. Thus as was pointed out by Akaike [1] (in terms of the FPE, which is asymptotically equivalent to the MAICE), and Shibata [8], who obtained the asymptotic distribution of the estimated order, the MAICE is apt to overestimate the order p. Now, the results of the preceding section make it possible to look deeper into this phenomena. One of the direct consequences of Propositions 1–2 is that if $k, l \geqq p$, then AIC(k)<AIC(l) is asymptotically equivalent to $I(\hat{\theta}_k) - I^*(\hat{\theta}_l) > 2(k-l)$. This means that AIC($k$) attains its minimum at $k=k_0(\geqq p)$ if and only if

$$I(\hat{\theta}_{k_0}) - I(\hat{\theta}_l) > 2(k_0 - l) \qquad \text{for all } l \geqq p\,.$$

Thus, the MAICE estimates the order to be k when $I(\hat{\theta}_k) - I(\hat{\theta}_p)$ is large, contrary to the entropy maximization principle. This is partly due to the fact that the variance of AIC(k)−AIC(p) does not diminish as N tends to infinity but approaches to $2(k-p)$, twice of its expected value.

Note that AIC(k) can be viewed as an unbiased estimate of $J_k = -2\,\mathrm{E}_X \log f(X; \hat{\theta}_k) + 2k$ *based on the sample of size* 1. This suggests that we devide the given data $X_1, \cdots, X_N$ into several parts and replace AIC(k) by the arithmetic mean of AIC's computed from each of the devided data. Results of the numerical study on the behavior of the modified procedures will be published elsewhere.

Acknowledgement

Thanks are due to Professor H. Akaike for his comments and discussions.

THE INSTITUTE OF STATISTICAL MATHEMATICS

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, *Proceedings of the Second International Symposium on Information Theory*, B. N. Petrov and F. Csari, eds., Akademiai Kiado, Budapest, 267-281.

[3] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Cont.*, AC-**19**, 716-723.

[4] Akaike, H. (1976). Canonical correlation analysis of time series and the use of an information criterion, *System Identification: Advances and Case Studies*, R. K. Mehra and D. G. Lainiotis, eds., Academic Press, New York, 27-96.

[5] Akaike, H. (1977). On entropy maximization principle, *Proc. of the Symposium on Application of Statistics*, P. R. Krishnaiah, ed., North-Holland, Amsterdam, to appear.

[6] Doob, J. L. (1953). *Stochastic Processes*, John Wiley, New York.

[7] Hannan, E. J. (1970). *Multiple Time Series*, John Wiley, New York.

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