Serdica
Bulgariacae mathematicae
publicationes

Сердика
Българско математическо
списание

ESTIMATION THE ORDER OF MARKOV CHAINS II. BAYESIAN INFORMATION CRITERION

IVA P. CANKOVA

This paper continues the information approach to the problem of how to determine the order of a Markov chain. For the most general model of TDMP it is shown that asymptotically the Bayesian and ML estimators for the transition probabilities are equivalent. An approximation of the posterior probability is obtained in the case of a fixed order multiple Markov chain. A Bayesian information criterion is proposed and the consistency of the minimum BIC estimator MBICE is proved.

0. Introduction. As shown in [9], the problem to estimate the dimension of a model by means of information theory was initiated by H. Akaike [1]. The relative merits of AIC, as discussed in [9], are significant but the inconsistency of the derived estimator MAICE is something undesirable. Using Bayesian arguments G. Schwarz [7] has obtained a modified estimator of the dimension of a model for independent and identically distributed observations with distribution from the Koopman — Darmois family. With his procedure one can derive a consistent estimator. Under some speciflc assumptions about the type of the loss function and the form of the prior distribution this estimator is asymptotically optimal in the sense of minimizing the expected loss. Akaike continues the investigation of AIC and gives a Bayesian interpretation of the MAICE procedure. He shows that this procedure provides a minimax type solution of the problem under the assumption of equal prior probability of the models [2]. Further on he proposes a Bayesian extension [3]. Some comments on those criteria have provoked the paper [8]. As an alternative to the AIC procedure — the Bayesian one (BIC) was extended to the case of Markov chains by R. Katz [6]. We shall use his definitions of BIC which is similar to AIC and is based also on penalized likelihood ratio statistics.

Under the interpretation $\Theta=D$ (see § 5 of [9]) it is easy to elucidate the relation between the order of a Markov chain and the dimension of the model. Thus it is quite natural to try to develop Schwarz propositions to cover the case of Markov chains.

First we link the Bayesian procedure to the investigated model as it was stated in the proposed minimum procedure. After that an asymptotic approximation of the posterior probabilities is obtained in § 1. The consistency of MBICE is shown. The asymptotical optimality of both the proposed procedures is also discussed in § 2.

1. Bayesian Information Criterion. We shall discuss the Bayesian information approach in the case the observed process is a Markov chain, even a multiple one under A3, which means an irreducible aperiodic chain and whose every recurrent state is nonnul ([9], p. 316). (Down some of the statements are more general to cover TDMP model).

To determine the order of the model $[X_n, P(\theta), \Theta]$ we use the minimum procedure discussed in § 3 of [9], i. e. we consider the problem in terms of decision theory. Then, Bayesian information approach means that the loss function is related to an information quantity and a Bayesian approach to the whole procedure is applied. First we have to formulate the large sample Bayes procedure in details.

SERDICA Bulgaricae mathematicae publicationes. Vol. 14, 1988, p. 316—321.

A5. Assume that the order of the chain is an integer random variable with finite values $0,1,\ldots, m$. Let β_l, $l=0,1,\ldots, m$ be the prior probability of the model with parametric space ${}_k\Theta$, $k=s^{l+1}-s^l$, i. e. l is the order of the Markov chain.

Since we are examining the asymptotic nature of the result, the prior distribution of the transition probabilities $\theta(\Theta=D)$ does not need to be known exactly. Then it sufficies to assume:

A6. The prior probability of θ is of the form $\Sigma_{l=0}^{m} \beta_l\mu_l$, where μ_l is the conditional prior distribution of θ given the l-th model and has k-dimensional density which is bounded throughout ${}_k\Theta$.

To simplify the situation we need some more assumptions about the loss function:

C7. We assume a fixed penalty for having made a wrong guess for the model, i. e. the penalty does not depend on θ or the procedure is unrandomized. Such a loss function is called 0-1 loss function.

Comment 8. Indeed a 0-1 loss function is equal to 0 or 1 over the subspaces where respectively a right or wrong decision is taken.

Actually a loss that depends on θ and the guess would yield the same asymptotic results provided the loss function remains between two fixed positive bounds for all the wrong decisions.

Considering the introduced information loss function $W(\theta;\,\theta)$ by (3.2) in [9] we realize that it satisfies C7. Indeed if the decision is right according to Th.2.2 [9] the loss is zero. But the order could be an integer not greater than m, thus there are positive bounds between which the loss remains for all wrong decisions (see Comment 8). Thus without loss of generality we can treat it as 0-1 loss function.

Under these circumstances the Bayes' procedure consists of selecting the posteriori most probable model.

Before finding out a proper approximation for the posterior probability we can recall the following general result:

Theorem 1.1. *For the model $[X_n, P(\theta), \Theta]$ of fixed dimension the MLEs are asymptotically equivalent to Bayes estimators for an arbitrary nonvanishing prior distribution $\mu(\theta)$ of the parameter θ. (Here we use Θ instead of ${}_k\Theta$. The parametric space could be of type ${}_k\Theta$ corresponding to the fixed dimension l).*

Proof. The posterior risk function has the form

$$R_n(\widetilde{\theta})=\frac{\int W(\theta;\,\widetilde{\theta})\exp\{L_n(\theta)\}\mu(d\theta)}{\int \exp\{L_n(\theta)\}\mu(d\theta)}, \tag{1.1}$$

where $\widetilde{\theta}$ is an estimator of the parameter.

In order to obtain a Bayes estimator it is necessary to minimize (1.1) or equivalently to find a minimizing $\widetilde{\theta}$ with

$$K_n(\widetilde{\theta})=\int\mu(d\theta)W(\theta;\,\widetilde{\theta})\exp\{L_n(\theta)\}. \tag{1.2}$$

Using the LLN for TDMP (Th. 1.1, [4]), we have

$$p\lim 1/n \sum_{i=1}^{n} g_{uv}(x_i,\, x_{i+1};\,\theta)=-\sigma_{uv}(\theta),$$

besides this C1.3 and MLEs are asymptotically normal (Th.2.2, [4]). Expanding each of the functions $g(x_i,\, x_{i+1};\,\theta)$ with a center $\widehat{\theta}$ and taking the sum up to n (similarly as in (4.7), [9]), we obtain

$$L_n(\theta)=L_n(\widehat{\theta})+1/2 \sum_{u=1}^{r}\sum_{v=1}^{r} \sqrt{n}(\theta_u-\widehat{\theta}_u)\sqrt{n}(\theta_v-\widehat{\theta}_v)1/n \sum_{i=1}^{n} g_{uv}(x_i,\, x_{i+1};\,\theta)+o_p(1)$$

or

$$L_n(\theta)=L_n(\widehat{\theta})-n/2\,\|\widehat{\theta}-\theta\|^2_{\Sigma(\widehat{\theta})}+o_p(1),$$

where $\widehat{\theta}\in T_\theta$.

Let $h(\theta)$ denote the prior density corresponding to $\mu(\theta)$ and

$$\omega(\theta;\ \widetilde{\theta})=W(\theta;\ \widetilde{\theta})h(\theta). \tag{1.3}$$

From the explicit relation between the loss function and the information quantity (Th.2.2 and C1, C3, [9]) it follows that $\omega(\widehat{\theta};\ \widetilde{\theta})\neq 0$ and $\omega(\theta;\ \widetilde{\theta})$ is continuously differentiable with respect to θ. Let expand it with a center $\widehat{\theta}$ up to the first order. Asymp totically we derive that

$$K_n(\widetilde{\theta})\sim\omega(\widehat{\theta};\ \widetilde{\theta})\exp\{L_n(\widehat{\theta})\}\int_{-\infty}^{+\infty}\exp\{-n/2\,\|\widehat{\theta}-\theta\|^2_{\Sigma(\widehat{\theta})}\}\,d\theta \tag{1.4}$$

$$=\sqrt{2\pi}\,\omega(\widehat{\theta};\ \widetilde{\theta})\prod_{i=1}^{n} f(x_i,\ x_{i+1};\ \widehat{\theta})\{\Sigma(\widehat{\theta})\}^{-1/2}.$$

Thus the minimization of $K_n(\widetilde{\theta})$ is asymptotically equivalent to the minimization of $\omega(\widehat{\theta};\ \widetilde{\theta})$ with respect to $\widetilde{\theta}$, but $\omega(\widehat{\theta};\ \widetilde{\theta})=0$ only when $\widetilde{\theta}=\widehat{\theta}$. Hence the Bayes estimators and MLEs are asymptotically equivalent.

A similar result is announced in [10] for the case of independent observations. From the above statement we can expect that in a large sample the leading term of the Bayes estimator is simply the maximum likelihood estimator and only the second term reflects the singularities of the prior distribution.

Although Bayes' theorem allows an arbitrary prior distribution, it is convinient theoretically to select the one which leads to simple posterior distribution or so-called conjugate distributions (§10.3, [11]).

If we fix l for the investigated model $[X_n,\ P(\theta)=\{p_{i_1}\ldots i_l:\ i_{l+1}\},\ {}_k\Theta]$, then P_θ be-ongs to the exponential family since

$$L_{n,l}=\prod_{i_1\ldots i_l} p_{i_1\ldots i_l:\,i_{l+1}}^{n_{i_1}\ldots i_{l+1}}=\exp\{\sum_{i_1\ldots i_l i_{l+1}} n_{i_1}\ldots i_{l+1}\ln p_{i_1}\ldots i_l:\ i_{l+1}\},$$

where $p_{i_1}\ldots i_l:\ i_{l+1}$ are constrained to lie in a $s^l(s-1)$ dimensional subset ${}_{\bar{k}}\Theta$ of ${}_k\Theta$.

As it is suggested in [11, p. 409] or [5, p. 96] for each set of transition probabilities $\{p_{i_1}\ldots i_l:\ i_{l+1}\}$, for fixed l-tuple $(i_1\ldots i_l)$ in ${}_k\Theta$, let assume to have an independent Dirichlet distribution $\mathsf{D}(\alpha_1,\ldots,\alpha_s)$ with density

$$\Gamma(\sum_{j=1}^{s}\alpha_j)\prod_j p_{i_1\ldots i_l:\,j}^{\alpha_j-1}/\prod_{j=1}^{s}\Gamma(\alpha_j),\quad j=1,\ 2,\ldots,\ s. \tag{1.5}$$

Thus the density of the conditional prior distribution μ_l is properly determined in accordance with A6. Then the posterior density given $n_{i_1}\ldots i_l i_{l+1},\ i_{l+1}\in\{1,\ 2,\ldots,\ s\}$ and $(i_1,\ldots,\ i_l)$ is fixed as

$$\Gamma(n_{i_1\ldots i_l}+\sum_{i=1}^{s}\alpha_j)\prod_{j=1}^{s} p_{i_1\ldots i_l:\,j}^{\alpha_j+n_{i_1}\ldots i_l j-1}(\prod_{j=1}^{s}\Gamma(\alpha_j+n_{i_1\ldots i_l j}))^{-1}. \tag{1.6}$$

If we assume all the α_j, $j=1,\ldots,s$ to be equal to 1, i. e. $\alpha_j=1$, then the transition probabilities are uniformly distributed over the simplex $P_{i_1\ldots i_l:\,i_{l+1}}\geq 0$, $\Sigma_{i_{l+1}}p_{i_1\ldots i_{l+1}}=1$, $(i_1,\ldots,\ i_l)$ fixed. We shall use the rewritten form of (1.6)

$$(1.7)\qquad (n_{i_1\ldots i_l}+\sum_{j=1}^{s}\alpha_j-1)!\prod_{j=1}^{s}p_{i_1\ldots i_l:j}^{\alpha_j+n_{i_1\ldots i_l j}-1}\Big(\prod_{j=1}^{s}(\alpha_j+n_{i_1\ldots i_l j}-1)!\Big)^{-1}$$

Let us denote by $BP_l(x_1,\ldots,x_{n+1})$ or briefly BP_l the posterior probability and by $\widehat{L}_{n,e}=\Pi_{i_1\ldots i_l i_{l+1}}\widehat{p}_{i_1\ldots i_l i_{l+1}}^{n_{i_1\ldots, i_{l+1}}}$ the likelihood function for the MLEs. Then we can generalize Th. 3. of [6].

Theorem 1.2. *Consider the model* $[X_n, P(\theta), {}_k\Theta]$ *for fixed* l *and large sample size* n, *when all the transition probabilities have independent Dirichlet prior distribution* $\mathrm{D}(\alpha_1,\ldots,\alpha_s)$. *The following approximation holds*

$$(1.8)\qquad \ln BP_l\simeq\ln\widehat{L}_{n,l}-1/2s^l(s-1)\ln n.$$

Proof. We use the usual notations for the proportionality $\propto$, i. e. $P(B_j\mid A)\propto P(A\mid B_j)P(B_j)$ reflects the variability of $P(B_j\mid A)$ under fixed A regarding j in accordance with Bayes' theorem [11]. Here A is the observation for fixed n, i. e. $(x_1,\ldots,x_{n+1})$. Then if we ignore the first term of the likelihood function (see Comment 2, [9]) and apply Bayes' formula, we obtain

$$BP_l\propto\int_{\bar{k}\Theta}\beta_l L_{n,l}d\mu_l\propto\prod_{i_1\ldots i_{l+1}}\int_{\bar{k}\Theta}\beta_l p_{i_1\ldots i_{l+1}}^{n_{i_1\ldots i_{l+1}}}d\mu_l.$$

Taking into consideration (1.7) and the properties of the Dirichlet distribution over the simplex ${}_{\bar{k}}\Theta$, we derive

$$BP_l\propto\prod_{i_1\ldots i_l}\prod_j(n_{i_1\ldots i_l j}+\alpha_j+1)!\,((n_{i_1\ldots i_l}+\sum_{j=1}^{s}\alpha_j-1)!)^{-1}\beta_l.$$

Hence, because of $n_{i_1\ldots i_l}=\Sigma_{j=1}^{s}n_{i_1\ldots i_l j}$

$$\frac{BP_l}{\widehat{L}_{n,l}}\propto\beta_l\prod_{i_1\ldots i_l}\frac{n_{i_1\ldots i_l}^{n_{i_1\ldots i_l}}}{(n_{i_1\ldots i_l}+\sum_{j=1}^{s}\alpha_j-1)\mathrm{i}}\prod_{i_{l+1}}\frac{(n_{i_1\ldots i_{l+1}}+\alpha_{i_{l+1}}-1)!}{n_{i_1\ldots i_{l+n}}^{n_{i_1\ldots i_{l+1}}}}.$$

Using now Stirling's formula $k!\sim^{k+1/2}e^{-k}\sqrt{2\pi}$, one gets

$$\frac{BP_l}{\widehat{L}_{n,l}}\simeq\beta_l\prod_{i_1\ldots i_l}\frac{n_{i_1\ldots i_l}}{n_{i_1\ldots i_l}}(2\pi)^{(s-1)/2}((n_{i_1\ldots i_l}+\sum_{j=1}^{s}\alpha_j-1)\ldots(n_{i_1\ldots i_l}+1)(n_{i_1\ldots i_l})^{1/2})^{-1}$$

$$\times\prod_{i_{l+1}}(n_{i_1\ldots i_{l+1}}+\alpha_{i_{l+1}}-1)\ldots(n_{i_1\ldots i_{l+1}}+1)(n_{i_1\ldots i_{l+1}})^{1/2},$$

and for the logarithms

$$\ln\frac{BP_l}{\widehat{L}_{n,l}}\simeq1/2\,s^l(s-1)\ln(2\pi)+\ln\beta_l-1/2\sum_{i_1\ldots i_l}\ln(n_{i_1\ldots i_l})+1/2\sum_{i_1\ldots i_{l+1}}\ln(n_{i_1\ldots i_{l+1}})$$

$$+\sum_{i_1\ldots i_{l+1}}\ln(n_{i_1\ldots i_{l+1}}+\alpha_{i_{l+1}}-1)+\ldots+\sum_{i_1\ldots i_{l+1}}\ln(n_{i_1\ldots i_{l+1}}+1)$$

$$-\sum_{i_1\ldots i_l}\ln(n_{i_1\ldots i_l}+\sum_{i=1}^{s}\alpha_j-1)-\ldots-\sum_{i_1\ldots i_l}\ln(n_{i_1\ldots i_l}+1)$$

or

$$\ln \frac{BP_l}{\widehat{L}_{n,l}} \simeq 1/2\, s^l (s-1) \ln n + (\sum_{j=1}^{s} \alpha_j - s)\, s^l \ln n - (\sum_{i=1}^{s} \alpha_j - 1) \ln n s^l = -1/2\, s^l (s-1) \ln n.$$

Above Jensen's inequality has been used and the members not depending on n have been omitted.

This approximation suggests a new risk function. In the terms of the initial model of TDMP we can formulate.

Definition 1. *For the risk function we find the statistics*

$$BIC(l) = {}_l\lambda_m - (\nabla s^m - \nabla s^l) \ln n, \tag{1.9}$$

called also Bayesian Information Criterion.

Lemma 1.1. *The following are equivalent forms of* (1.9):

f1: $R1(k) = -2 \sum_{j=1}^{n} \ln f(x_i,\ x_{i+1};\ {}_k\widehat{\theta}) + k \ln n,$

f2: $R2(k) = {}_k\eta r + k \ln n,$

f3: $R3(k) = {}_k\eta r - \ln n$ (degrees of freedom of ${}_k\eta r$).

The most proper form for the multiple Markov chains model is the third one The minimum procedure requires the minimization of the risk function (1.9).

Definition 2. *The BIC estimator of the order of a Markov chain is called. MBICE $\widetilde{l} = \widetilde{l}_{BIC}$ and is chosen in a way that*

$$BIC(\widetilde{l}) = \min_{0 \leqq l \leqq m-1} BIC(l). \tag{1.10}$$

Theorem 1.3. *The minimum BIC procedure is equivalent to the Bayes' procedure.*

Proof. Let $\widetilde{l}$ is MBICE, i. e. it satisfies (1.10). From Th.1.2. it follows that

$$BIC(l) = {}_e\lambda_m - (s^m - s^l)(s-1) \ln n$$

$$= -2 \ln \frac{BP_l}{BP_m} - (s^l - s^m)(s-1) \ln n - (s^m - s^l)(s-1) \ln n = -2 \ln (BP_l / BP_m).$$

Since $BIC(\widetilde{l})$ satisfies (1.10), it is equivalent to the event

$$\{-2 \ln BP_{\widetilde{l}} \leqq -2 \ln BP_l,\ \forall l\} \text{ or } \{BP_{\widetilde{l}} \geqq BP_l,\quad 0 \leqq l \leqq m-1\},$$

i. e. for $\widetilde{l}$ the posterior probability is maximal.

Thus we obtain that MBICE is a Bayes estimator of the order of the model. A simple reflection shows that then ${}_{\widetilde{l}}\widetilde{\theta}$ are the conditional Bayes estimators of the transition probabilities.

2. Properties of both minimum procedures

Theorem 2.1. *MBICE is a consistent estimator of the true order p of the Markov chain*: $\lim_{n\to\infty} P\{\widetilde{l} = p\} = 1$.

Proof. Let $0 \leqq l \leqq p-1$. As in the part 1 of the proof of Th.5.2. [9] it is easy to see that $\lim_{n\to\infty} P\{\widetilde{l} = l\} = 0$.

Let $p+1 \leqq l \leqq m-1$, then $P\{\widetilde{l} = l\} = P\{BIC(l) \leqq BIC(j),\ 0 \leqq j \leqq m-1\} \simeq P\{BIC(l) \leqq BIC(j),\ p \leqq j \leqq m-1\}$.

But

$$P\{\tilde{l}=l\}\cong P\{{}_l\lambda_m-\ln n(\nabla s^m-\nabla s^l)\leqq {}_j\lambda_m-\ln n(\nabla s^m-\nabla s^j)\}=P\{{}_l\lambda_j\leqq \ln n(\nabla s^l-\nabla s^j)\}.$$

Since Th.5.1. [9] holds, then ${}_l\lambda_j$ is chi-square with $(\nabla s^j-\nabla s^l)$ degrees of freedom. Thus $\lim P\{\tilde{l}=l\}=\lim_{n\to\infty}P\{{}_l\lambda_j\leqq-(\nabla s^j-\nabla s^l)\ln n\}=0$, and $\lim^n{}_{\to\infty}P\{\tilde{l}=p\}=1$.

Finally let us notice that besides the differences of the estimators MAICE and MBICE, both of them are asymptotically optimal in the sense that they minimize the chosen statistics for the risk function.

We gave two minimum procedures by which the problem of point estimation is treated in terms of decision theory. The Bayesian and Akaike information approaches are applied in solving the problem. The attractiveness of both the procedures is in their easier computer implementation especially in the form F3 (Lemma 4.4, [9]) and the form f3 Lemma 1.1).

Acknowledgments. I am deeply grateful to Dr. N. M. Yanev for his encouragement, support and criticism.

The referee's comments are most gratefully acknowledged.

REFERENCES

1. H. Akaike. Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory. (Ed. B. N. Petrov and F. Csaki). Budapest, 1972, 267-281.
2. H. Akaike. A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.*, **30**, 1978, 9-14.
3. H. Akaike. A Bayesian extension of minimum AIC procedure of autoregressive model fitting. *Biometrica*, **66**, 1979, 237-242.
4. P. Billingsley. Statistical inference for Markov processes. New York, 1961.
5. J. Hartigan. Bayes theory. New York, 1983.
6. R. Katz. On some criteria for estimating the order of a Markov chain. *Technometrics*, **23**, 1981, 243-249.
7. G. Schwarz. Estimating the dimension of the model. *Ann. Stat.*, **6**, 1978, 461-464.
8. M. Stone. Comments on model selection criteria of Akaike and Schwarz. *J. Royal Statist. Soc., Ser. B.*, **41**, 1979, 276-278.
9. I. Cankova. Estimation the order of Markov chains I. Akaike's information criterion for the case of discrete-time Markov processes. *Serdica*, **13**, 1987, 303-319.
10. Ш. Закс. Теория статистических выводов. М., 1975.
11. Д. Кокс, Д. Хинкли. Теоретическая статистика. М., 1978.

Centre for Matehematics and Mechanics
Sofia 1090 P. O. Box 373

Received 14. 5. 1985
Revised 10. 6. 1988