Serdica

Bulgariacae mathematicae publicationes

Сердика

Българско математическо списание

ESTIMATION THE ORDER OF MARKOV CHAINS I. AKAIKE'S INFORMATION CRITERION FOR THE CASE OF DISCRETE-TIME MARKOV PROCESSES

IVA P. CANKOVA

The present paper deals with an information approach to the problem of parameter estimation for the statistical model of a discrete-time Markov process. It is assumed that the process is stationary. The complete set of regularity conditions is determined. The notion of Kullback-Leibler's mean information is introduced for the case of Markovian dependence. Some of the main properties of that enformation quantity are obtained. An extention of the maximum likelihood principle and a minimum procedure of taking decision are proposed for solving the problem. A strict derivation of Akaike's information criterion is stated. This criterion is applied to determine the order of irreducible aperiodic Markov chains. A statistic AIC(l) is suggested and the asymptotic behaviour of AIC estimator is examined.

0. Introduction. A brief historical review of the problem of determination the order of a Markov chain shows that there are three important approaches to the subject, namely the likelihood ratio, the chi-square and psi-square ones, well known now. In the introduction of Tong's paper [8] is well-pointed out their close relation to the classical Karl Pearson's chi-square approach of contingency tables. Even the existence of Kulback's monograph [12] suggests that new view to the subject applying information quantities. Akaike was the first who decided to harness to a team these ideas and proposed a procedure (resulting in now wide-spread Akaike information criterion AIC) for parameter estimation, avoiding the complications of the conventional statistics' approach. In his fundamental work [1] preserving methodologically heuristic spirit Akaike stated (on p. 277) "...following the approach of Billingsley [4], we can see that the same line of discussion can be extended to cover the case of finite parameter Markov processes". On the other hand Billingsley [4] or [5] stated that discrete-time Markov processes model (TDMP) is large enough to include the Mann-Wald theory. We have to state that really most of the followers have extended the applicability of AIC to the problem of estimation the order of autoregressive processes (ARP), and autoregressive integrated moving average processes (ARIMA). For that reason we say wide-spread AIC.

The first work connecting the subject with AIC, i. e. applying that criterion to the case of Markov chain is Tong's paper [8]. Later Katz [6] arose the idea for treatment including the Bayes' approach.

Our intention is to develop the information approach for Markov process inference. We determine the precise regularity conditions for the statistical model of TDMP (§ 1, I) and define Kullback — Leibler's mean information for the case of observation with Markovian dependence (§ 2, I). Also we extend Akaike's line to cover the case of TDMP (§ 4, I). For irreducible aperiodic Markov chains AIC(l) statistic for determination the order of the chain is suggested and the inconsistency (rather the overestimating of the true order) of the AIC estimator is shown (§ 5, I).

SERDICA Bulgaricae mathematicae publicationes. Vol. 13, 1987, p. 303—319.

The second part of the paper is devoted to the Bayes' approach to the subject, the main properties of BIC estimator and to a reflection of the optimality in a certain sense of both the proposed estimators.

Our collaboration with molecular biologists in the recent few years suggest us the interpretation of a DNA molecule as an irreducible aperiodic Markov chain. Two reasons fixed our interest to the information criteria. First, the possibility of their precise mathematical examination applying them to determine the order of Markov chains, and second, their easy computer realization. The last possibility was announced earlier in [10]. The results obtained in [10] show good concurrence with the evolution theory. Moreover, the present paper illustrates the development of some techniques and problem solving necessary for the practice of another science.

1. Regularity conditions for TDMP. Let $\{X_n, n=1, 2, \dots\}$ be a stochastic process on the probability space $(\Omega, \mathscr{B}, \mathscr{F})$ with values in the measurable space $(X, \mathscr{F}_x)$. Suppose that the family of the probability measures is parametric, i. e. $\mathscr{P}=\{\boldsymbol{P}_\theta, \theta\in\Theta\}$, where in general Θ is an open subset of r-dimensional Euclidean space E^r. We assume the following specific conditions.

A1.1. For each $\theta\in\Theta$, $\{X_n\}_{n\geq 1}$ is a Markov process with stationary transition measures $p_\theta(\xi, A)=\boldsymbol{P}_\theta\{X_{n+1}\in A \mid X_n=\xi\}$, $p_\theta(\xi, A)$ is a measurable function of ξ for fixed $A\in\mathscr{F}_x$ and a probability measure on $\mathscr{F}_x$ for fixed ξ.

A1.2. There exist a unique stationary distribution $p_\theta(\cdot)$ on $\mathscr{F}_x$ such that $p_\theta(A) =\int_X p_\theta(d\xi)\,p_\theta(\xi, A)$ for all $A\in\mathscr{F}_x$.

In order that likelihood functions exist we assume some additional conditions.

A2.1. There is a measure λ on $\mathscr{F}_x$ (not necessarily finite) with respect to (w. r. t.) which all the transition measures have densities $f(\xi, \eta; \theta)$, i. e.

$$p_\theta(\xi, A)=\int_A f(\xi, \eta; \theta)\lambda(d\eta) \quad \text{for all } A\in\mathscr{F}_x.$$

A2.2. For a computational device we assume that $p_\theta(\cdot)$ is the initial distribution with a density $f(\xi; \theta)$ w. r. t. λ.

A.2.3. The functions $f(\xi; \theta)$ and $f(\xi, \eta; \theta)$ are measurable on the Cartesian products $(X\times R^r)$ and $(X\times X\times R^r)$, supplied with product σ-algebras, respectively. For R^r Borel σ-algebra is assumed, for X is $\mathscr{F}_x$.

Then we can rewrite A1.2. as

$$f(\eta; \theta)=\int_X f(\xi, \eta; \theta)\,f(\xi; \theta)\lambda(d\xi). \tag{1.1}$$

Let us use the following notations

$$g(\xi, \eta; \theta)=\ln f(\xi, \eta; \theta),\ f_u(\xi, \eta; \theta)=\partial f(\xi, \eta; \theta)/\partial\theta_u$$

for the partial derivatives of f and $\boldsymbol{E}_\theta\{\cdot\}$ for the expected value when θ is the true value of the parameter and $p_\theta(\cdot)$ is the initial distribution.

Comment 1. It is obvious that all conditional probabilities and expected values $\boldsymbol{E}_\theta\{\cdot \mid x_1\}$ are determined by the transition measures. Thus the initial distribution has no effect on them.

If some observations of the process $(x_1, x_2, \dots, x_n, x_{n+1})$ are at one's disposition it is easy to show that the likelihood function based on the observations is $L(x_1, \dots, x_{n+1}; \theta)=f(x_1; \theta)\,\Pi_{i=1}^n f(x_i, x_{i+1}; \theta)$, except on a set with measure zero w. r. t. the $(n+1)$-fold product measure of λ on the σ-algebra $\mathscr{F}_x^{n+1}$. Then the log-likelihood of the observation is $L_n(\theta)=\ln f(x_1; \theta)+\Sigma_{i=1}^n \ln f(x_i, x_{i+1}; \theta)$.

Now we are going to state the regularity conditions, local in character, well stated by Billingsley [4].

C1.1. For any ξ, the set of η for which $f(\xi, \eta; \theta)>0$ does not depend on θ.

C1.2. For any ξ and η, there exist third-order continuous partial derivatives of densities throughout Θ. Then $g(\xi, \eta; \theta)$ is well defined except on a set of $p(\xi, \cdot; \theta)$-measure zero and $g_u(\xi, \eta; \theta)$, $g_{uv}(\xi, \eta; \theta)$, $g_{uvw}(\xi, \eta; \theta)$ exist and are continuous in Θ.

C1.3. For any $\theta \in \Theta$ there exists a neighbourhood T_θ of θ such that for every u, v, $w(=1, 2, \ldots, r)$ and ξ the following conditions hold:

$$\int_X \sup_{\tilde{\theta} \in T_\theta} |f_u(\xi, \eta; \tilde{\theta})| \lambda(d\eta)<\infty,$$

$$\int_X \sup_{\tilde{\theta} \in T_\theta} |f_{uv}(\xi, \eta; \tilde{\theta})| \lambda(d\eta)<\infty, \tag{1.2}$$

$$\boldsymbol{E}_\theta \{\sup_{\tilde{\theta} \in T_\theta} |g_{uvw}(x_1, x_2; \tilde{\theta})|\}<\infty.$$

Let also for $u=1, 2, \ldots, r$ we have

$$\boldsymbol{E}_\theta \{|g_u(x_1, x_2; \theta)|^2\}<\infty \tag{1.3}$$

and the $r\times r$ matrix $\Sigma(\theta)=\|\sigma_{uv}(\theta)\|$, where

$$\sigma_{uv}(\theta)=\boldsymbol{E}_\theta \{g_u(x_1, x_2; \theta) g_v(x_1, x_2; \theta)\}, \tag{1.4}$$

is nonsingular.

We need to impose some further conditions to ensure that the law of large numbers and the central limit theorem can be applied to the random vector $n^{-1/2} \Sigma_{i=1}^n g_u(x_i, x_{i+1}; \theta)$, $u=1, 2, \ldots, r$ independently on the initial distribution.

C2.1. For any $\theta \in \Theta$, the stationary distribution $p_\theta(\cdot)$ is such that for each $\xi \in X$, $p_\theta(\xi, \cdot)$ is absolutely continuous w. r. t. $p_\theta(\cdot)$, i. e. $p_\theta(\xi, \cdot) \ll p_\theta(\cdot)$.

C2.2. There is some $\delta>0$ such that for $u=1, 2, \ldots, r$

$$\boldsymbol{E}_\theta \{|g_u(x_1, x_2; \theta)|^{2+\delta}\}<\infty \quad (\delta \text{ may depend on } \theta).$$

It is obvious that $\Sigma(\theta)$ is positive-definite for each θ, but we need also

C3. $\Sigma(\theta)$ is continuous for any $\theta \in \Theta$.

Comment 2. It is easy to realize that under the conditions (formulated such that the initial distribution pays no role) $\ln f(x_1; \theta)$ in $L_n(\theta)$ is dominated by the other members for sufficiently large n (the information about $f(\xi; \theta)$ does not increase with n). So for the purpose of the large-sample theory we could redefine the log-likelihood function as

$$L_n(\theta)=\sum_{i=1}^n g(x_i, x_{i+1}; \theta). \tag{1.5}$$

It is useful to think that for each θ all the mass of the initial distribution is concentrated at the point x_1.

The process $\{X_n\}_{n\geq 1}$, the corresponding transition densities $f(\xi, \eta; \theta)$ and the range Θ of the parameter θ form the triplet $[X_n, f(\xi, \eta; \theta), \Theta]$ which is called a model. The notion of a model will mean that $\{X_n\}_{n\geq 1}$ is subjected to conditions and assumptions described above.

Comment 3. We assume that the process is actually governed by the densities corresponding to the true value θ^0, $\theta^0 \in \Theta$. The model specifies nothing about the initial distribution so the hypothesis that θ^0 is the true value is a composite one.

Further, our attention is focussed on the inference for a Markov chain even multiple. As far as more general statements could be formulated for a discrete-time Markov processes, we would postpone the statement of the concrete problem.

2. Kullback — Leibler's mean information for TDMP. It is well known that in 1951 Kullback and Leibler [7] published a generalization of the information measure known as Shannon's and Wiener's one. It was Jeffrey's establishment for the first time that there is an analytic relationship between the generalized and Fisher's information measures. Moreover, log-likelihood ratio under fixed value of the random variable is called information for discrimination between the statistical populations. We are intended to determine Kullback's mean information for TDMP and state some lemmas analogous to those for the case of a single variable.

Let the model $[X_n, f(\xi, \eta; \theta), \Theta]$ be given. So every pair (X_i, X_{i+1}) consists of dependent variables. Using the basic idea of the mean information measure, we can determine that quantity for the dependent variables X_1 and X_2. Taking into consideration that for any θ,

$$\boldsymbol{P}_\theta\{X_1=\xi,\ X_2=\eta\}=\boldsymbol{P}_\theta\{X_2=\eta \mid X_1=\xi\}.\boldsymbol{P}_\theta\{X_1=\xi\}$$

and fixing two different elements θ^1 and θ^2 in Θ, we can express the information quantity as follows

$$I(\theta^1.\theta^2;\ X_1,\ X_2)=\int\int f(\xi,\ \eta;\ \theta^1)\ln\frac{f(\xi,\ \eta;\ \theta^1)f(\xi;\ \theta^1)}{f(\xi,\ \eta;\ \theta^2)f(\xi,\ \theta^2)}f(\xi;\ \theta^1)\lambda(d\eta)\lambda(d\xi).$$

Further for shortness we use the notations $f_i(\xi, \eta)$, $f_i(\xi)$ and $I(1:2;\ X_1,\ X_2)$ instead of $f(\xi, \eta;\theta^i)$, $f(\xi;\theta^i)$ and $I(\theta^1:\theta^2;\ X_1,\ X_2)$, respectively $(i=1, 2)$. Thus

$$\begin{aligned}I(1:2;\ X_1,\ X_2)&=\int_X\int_X f_1(\xi,\ \eta)\ln\frac{f_1(\xi,\ \eta)f_1(\xi)}{f_2(\xi,\ \eta)f_2(\xi)}f_1(\xi)\lambda(d\eta)\lambda(d\xi)\\&=\int_X\Big[\int_X f_1(\xi,\ \eta)\ln(f_1(\xi,\ \eta)/f_2(\xi,\ \eta))\lambda(d\eta)\Big]f_1(\xi)\lambda(d\xi)\\&+\int_X f_1(\xi)\ln(f_1(\xi)/f_2(\xi))\Big[\int_X f_1(\xi,\ \eta)\lambda(d\eta)\Big]\lambda(d\xi).\end{aligned}$$

Since for each $\theta\in\Theta$

$$\int_X f(\xi,\ \eta;\ \theta)\lambda(d\eta)=1 \tag{2.1}$$

we can see that the second term in $I(1:2;\ X_1,\ X_2)$ is exactly $I(1:2;\ X_1)$, which is the well-known information for discrimination between θ^1 and θ^2 for the initial state. The integral within the brackets in the first term is usually determined as a conditional information contained in X_2 when $X_1=\xi$ and is denoted by $I(1:2;\ X_2 \mid X_1=\xi)$. The mean value w. r. t. the distribution of X_1 of the conditional information is called a mean conditional information and is denoted by $I(1:2;\ X_2 \mid X_1)$. So

$$I(1:2;\ X_1,\ X_2)=I(1:2;\ X_2 \mid X_1)+I(1:2;\ X_1).$$

Since $I(1:2;\ X_2 \mid X_1=\xi)$, as a conditional expectation $\boldsymbol{E}_1\{\ln\frac{f_1(\xi,\ \eta)}{f_2(\xi,\ \eta)} \mid X_1=\xi\}$, does not depend on the initial distribution of X_1, ignoring it when $\xi=x_1$ we find $I(1:2;\ X_1)=0$ $(\forall\theta,\ f(x_1;\ \theta)=1)$ and $I(1:2;\ X_1,\ X_2)=I(1:2;\ X_2 \mid X_1=x_1)$ or generally $I(1:2;\ X_1,\ X_2)=I(1:2;\ X_2 \mid X_1)$ (see Comments 1 and 2).

Lemma 2.1. *For any pair* (X_i, X_{i+1}), *of the model* $[X_n, f(\xi, \eta; \theta), \Theta]$ *the mean (conditional) information satisfies the relation*

$$I(1:2;\ X_{i+1} \mid X_i)=I(1:2;\ X_2 \mid X_1)\ \textit{and}\ I(1:2;\ X_i)=I(1:2;\ X_1). \tag{2.2}$$

Proof. It is obvious that (2.2) holds because of the homogeneity of the process, see condition A1.1.

Thus the most important for the mean information quantity is the transition between the states of the process, not the states themselves.

Suppose, H^i is the hypothesis that the parameter θ takes a fixed value θ^i, $i=1, 2$.

Definition 2.1. *For the model* $[X_n, f(\xi, \eta; \theta), \Theta]$ *we determine information for discrimination between* H^1 *and* H^2 *when* H^1 *is true, contained in the transition from the state* ξ *to the next one, as follows*

$$I(1:2;\ X_2 \mid X_1=\xi)=\int_X f_1(\xi, \eta) \ln (f_1(\xi, \eta)/f_2(\xi, \eta))\, \lambda(d\eta)$$

and mean information for discrimination (between H^1 *and* H^2 *when* H^1 *is true) contained in the transition between two successive states (of the process)*

$$I(1:2;\ X_2 \mid X_1)=\int_X I(1:2;\ X_2 \mid X_1=\xi) f_1(\xi)\, \lambda(d\xi). \tag{2.3}$$

Analogously we can determine an information quantity for $l+1$ dependent variables, rather than for $l+1$ consequtive observations of the process. Taking into account the Markovian type of the dependence, and hence the expression of the likelihood function as a product of transition measures, we can write down

$$I(1:2;\ X_j, X_{j+1}, \ldots, X_{j+l})=\int_X \ldots \int_X \varphi_1(l) \ln \frac{\varphi_1(l)}{\varphi_2(l)}\, \lambda(d\xi_j) \ldots \lambda(d\xi_{j+l}),$$

where $\varphi_i(l)=f_i(\xi_j) f_i(\xi_j, \xi_{j+1}) \ldots f_i(\xi_{j+l-1}, \xi_{j+l})$, $i=1, 2$.

Lemma 2.2. (*Additivity*) *If* $[X_n, f(\xi, \eta; \theta), \Theta]$ *is given then*

$$I(1:2;\ X_{i-1}, X_i, X_{i+1})=2I(1:2;\ X_2 \mid X_1)+I(1:2;\ X_1)$$

holds.

Proof. Indeed, we have

$$\begin{aligned}
&I(1:2;\ X_{i-1}, X_i, X_{i+1})\\
&=\int_X \int_X \int_X f_1(\xi) f_1(\xi, \eta) f_1(\eta, \zeta) \ln \frac{f_1(\xi, \eta) f_1(\eta, \zeta) f_1(\xi)}{f_2(\xi, \eta) f_2(\eta, \zeta) f_2(\xi)}\, \lambda(d\xi)\lambda(d\eta)\lambda(d\zeta)\\
&=\int_X \Big[\int_X f_1(\xi, \eta) \ln \frac{f_1(\xi, \eta)}{f_2(\xi, \eta)}\, \lambda(d\eta)\Big] f_1(\xi)\, \lambda(d\xi)\\
&+\int_X \Big[\int_X f_1(\eta, \zeta) \ln \frac{f_1(\eta, \zeta)}{f_2(\eta, \zeta)}\, \lambda(d\zeta)\Big] f_1(\eta)\, \lambda(d\eta)+\int_X f_1(\xi) \ln \frac{f_1(\xi)}{f_2(\xi)}\, \lambda(d\xi)\\
&=I(1:2;\ X_i \mid X_{i-1})+I(1:2;\ \mid X_{i+1} \mid X_i)+I(1:2;\ X_{i-1}).
\end{aligned}$$

Thus according to Lemma 2.1 we obtain

$$I(1:2;\ X_{i-1}, X_i, X_{i+1})=2I(1:2;\ X_2 \mid X_1)+I(1:2;\ X_1).$$

The following result can easily be derived by induction.

Theorem 2.1. *For the model* $[X_n, f(\xi, \eta; \theta), \Theta]$ *we have*

$$I(1:2;\ X_1, X_2, \ldots, X_{n+1})=n\, I(1:2;\ X_2 \mid X_1)+I(1:2;\ X_1). \tag{2.4}$$

This result justifies the choice of the log-likelihood in the sense mentioned in Comment 2. It maintains it once more.

Definition 2. *For the model* $[X_n, f(\xi, \eta; \theta), \Theta]$ *we determine Kullback — Leibler's information quantity for discrimination between* H^1 *and* H^2 *when* H^1 *is true*

$$I(1:2;\ X_1, X_2)=I(1:2;\ X_2 \mid X_1). \tag{2.5}$$

Theorem 2.2. *$I(1:2;X_1,X_2)$ is almost surely positive definite w. r. t. λ and equality holds iff $f_1(\xi,\eta)=f_2(\xi,\eta)$ w. r. t. λ.*

Proof. $I(1:2;X_1,X_2)=\int_X I(1:2;X_2\mid X_1=\xi)f_1(\xi)\lambda(d\xi)=\int_X[\int_X f_2(\xi,\eta)\frac{f_1(\xi,\eta)}{f_2(\xi,\eta)}\ln(f_1(\xi,\eta)/f_2(\xi,\eta))\lambda(d\eta)]f_1(\xi)\lambda(d\xi)$. Let us use the short notations $\psi(\xi,\eta)=f_1(\xi,\eta)/f_2(\xi,\eta)$ and $J(\xi,\eta)=\int_X f_2(\xi,\eta)\psi(\xi,\eta)\ln\psi(\xi,\eta)\lambda(d\eta)$.

It is obvious that

$$\int_X f_2(\xi,\eta)\psi(\xi,\eta)\lambda(d\eta)=\int_X f_1(\xi,\eta)\lambda(d\eta)=1. \tag{2.6}$$

Take $\varphi(t)=t\ln t$ and substitute $t=\psi(\xi,\eta)$. Since $\varphi(1)=0$, $\varphi'(1)=1$, $\varphi''(t)=1/t$ and $0<\psi(\xi,\eta)<\infty$ w. r. t. λ, we obtain for the Taylor expansion of up to the second order term $\varphi(\psi(\xi,\eta))=\varphi(1)+[\psi(\xi,\eta)-1]\,\varphi'(1)+(1/2)\,[\psi(\xi,\eta)-1]^2\varphi''(h(\xi,\eta))$, where $0<h(\xi,\eta)<\infty$ w. r. t. λ and $\psi(\xi,\eta)<h(\xi,\eta)<1$. Then for the integral $J(\xi,\eta)$ we obtain

$$\begin{aligned}J(\xi,\eta)&=\int_X(f_2(\xi,\eta)[\psi(\xi,\eta)-1]\lambda(d\eta)+(1/2)\int_X f_2(\xi,\eta)[\psi(\xi,\eta)-1]^2(1/h(\xi,\eta))\lambda(d\eta)\\&=(1/2)\int_X[\psi(\xi,\eta)-1]^2f_2(\xi,\eta)/h(\xi,\eta)\,\lambda(d\eta)\geqq 0.\end{aligned}$$

Hence $J(\xi,\eta)=\int_X f_1(\xi,\eta)\ln\psi(\xi,\eta)\lambda(d\eta)\geqq 0$ and equality holds iff $\psi(\xi,\eta)=1$, i. e $f_1(\xi,\eta)=f_2(\xi,\eta)$ w. r. t. λ.

Comment 4. Since $\Theta\subset E^r$, the inner product of vectors $\langle\theta^1,\theta^2\rangle=\Sigma_{u=1}^r\theta_u^1\theta_u^2$ and the length of a vector $|\theta^1-\theta^2|=\langle\theta^1-\theta^2,\ \theta^1-\theta^2\rangle^{1/2}$ are well defined. The matrix $\Sigma(\theta)=\|\sigma_{uv}(\theta)\|$ is nonsingular and positive definite for each θ. In accordance with Comment 3 let us fix $\Sigma(\theta^0)=\Sigma$ and treat Σ as a positive definite operator. Thus the inner product caused by Σ, $\langle\theta^1,\theta^2\rangle_\Sigma=\langle\Sigma\theta^1,\theta^2\rangle$, the norm $\|\theta\|_\Sigma^2=\langle\Sigma\theta,\theta\rangle=\langle\theta,\theta\rangle_\Sigma$ and the distance $\rho(\theta^1,\theta^2)=\|\theta^1-\theta^2\|_\Sigma$ are also well defined. Further, we shall use both the inner products described above.

Theorem 2.3. (*Relationship between Kullback — Leibler's and Fisher's information measures*). *For a given model $[X_n, f(\xi,\eta;\theta),\Theta]$ the following representation holds*

$$I(\theta^0:\theta^0+\Delta\theta;X_1,X_2)=(1/2)\|\Delta\theta\|_\Sigma^2. \tag{2.7}$$

Proof. Since the model satisfies conditions C1 it is possible to differentiate under the integral sign and since (2.1) holds

$$\begin{aligned}&\int_X f_u(x_i,x_{i+1};\theta)\lambda(dx_{i+1})=0 \text{ and } \int_X f(x_i,x_{i+1};\theta)(f_u(x_i,x_{i+1};\theta)/f(x_i,x_{i+1};\theta))\lambda(dx_{i+1})\\&=\int_X f(x_i,x_{i+1};\theta)g_u(x_i,x_{i+1};\theta)\lambda(dx_{i+1})=\boldsymbol{E}_\theta\{g_u(x_i,x_{i+1};\theta)\mid x_i\}=0.\end{aligned}$$

Hence

$$\boldsymbol{E}_\theta\{g_u(x_1,x_2;\theta)\}=0 \quad \text{for each } \theta\in\Theta. \tag{2.8}$$

Differentiating (2.1) twice (permissible by C1) we obtain $\int_X f_{uv}(x_i,x_{i+1};\theta)\lambda(dx_{i+1})=0$. It follows that

$$\boldsymbol{E}_\theta\{g_{uv}(x_1,x_2;\theta)\mid x_1\}=-\boldsymbol{E}_\theta\{g_u(x_1,x_2;\theta)g_v(x_1,x_2;\theta)\mid x_1\}$$

and therefore

$$\boldsymbol{E}_\theta\{g_{uv}(x_1,x_2;\theta)\}=-\sigma_{uv}(\theta). \tag{2.9}$$

Let T_0 be the neighbourhood of θ^0 satisfying C1.3. and denote by $G(x_1,x_2)=\sup_{\tilde\theta\in T_0}|g_{uvw}(x_1,x_2;\tilde\theta)|$. Then C1.3. implies the existence of a constant M such that

$$\boldsymbol{E}_{\theta^0}G(x_1,x_2)=M<\infty. \tag{2.10}$$

Consider Taylor expansion in $\theta \in T_0$ of $g(x_1, x_2; \theta)$ around θ^0

$$\ln(g(x_1, x_2; \theta)/g(x_1, x_2; \theta^0)) = \sum_{u=1}^{r} (\theta_u - \theta_u^0) g_u(x_1, x_2; \theta^0) + (1/2) \tag{2.11}$$

$$\times \sum_{u=1}^{r} \sum_{v=1}^{r} (\theta_u - \theta_u^0)(\theta_v - \theta_v^0) g_{uv}(x_1, x_2; \theta^0) + (1/3!) \sum_{u,v,w=1}^{r} (\theta_u - \theta_u^0)(\theta_v - \theta_v^0)$$

$$\times (\theta_w - \theta_w^0) g_{uvw}(x_1, x_2; \theta^1)$$

where θ^1 is between θ and θ^0, all in T_0.

By the mean value theorem the remainder R is rewritten in the following form: $R = \alpha \,|\, \theta - \theta^0 \,|^3 G(x_1, x_2)$, $|\alpha| \leq r^3/6$. By condition C1 integrating both the sides of (2.11) w. r. t. θ^0 and using (2.10), (2.8) and (2.9), we obtain

$$E_{\theta^0}\{\ln(f(x_1, x_2; \theta^0)/f(x_1, x_2; \theta)\}$$

$$= (-1/2) \sum_{u=1}^{r} \sum_{v=1}^{r} (\theta_u - \theta_u^0)(\theta_v - \theta_v^0)(-\sigma_{uv}) + o(|\theta - \theta^0|^2).$$

Omitting the members of higher order we derive

$$E_{\theta^0}\{\ln \frac{f(x_1, x_2; \theta^0)}{f(x_1, x_2; \theta)}\} = I(\theta^0 : \theta ; X_1, X_2)$$

$$= (1/2) \sum_{u=1}^{r} \sum_{v=1}^{r} (\theta_u - \theta_u^0)(\theta_v - \theta_v^0) \sigma_{uv} = (1/2) \,\|\, \theta - \theta^0 \,\|_{\Sigma}^2.$$

Finally if put $\theta = \theta^0 + \Delta\theta$, then the last relation can be rewritten in the form $I(\theta^0 : \theta^0 + \Delta\theta ; X_1, X_2) = (1/2) \,\|\, \Delta\theta \,\|_{\Sigma}^2$.

3. Extension of the maximum likelihood principle for TDMP. It is well known that the classical maximum likelihood principle (MLP) is utilized mainly in two branches of statistics — estimation and test theory, where log-likelihood function (LLF), i. e. (1.5) instead of the simple one, is often preferable and the intention is to find out all the solutions of the equations

$$\frac{\partial}{\partial \theta_u} L_n(\theta) = \sum_{i=1}^{n} g_u(x_i, x_{i+1} : \theta) = 0, \quad u = 1, 2, \ldots, r. \tag{3.1}$$

These solutions are called maximum likelihood estimators (MLEs). For the model $[X_n$ $f(\xi, \eta; \theta), \Theta]$ Th. 2.1 in [4] quarantees the existence of a consistent MLE of the true value θ^0. Because of the local character of C1, MLE is a local maximum of $L_n(\theta)$ with probability going to 1 as $n \to \infty$ and is the only consistent solution in a neighbourhood of θ^0 with probability one as $n \to \infty$. If the dimension of the parameter is known, then MLP provides good estimators. The principle does not however apply if we want to estimate the parameter without knowing its exact dimension and if we are intended to estimate that dimension, too. It is clear that another approach is necessary. Such one is so-called extended MLP proposed by A k a i k e [1] and closely related to the information quantity.

It is well recognized that the statistical estimation theory can be organized within the framework of the decision theory by choosing a proper loss function. Further we are going to state in details the EMLP and finally to establish its main essence as a general estimation procedure based on the decision theoretic considerations.

1. According to condition C1 the function $g(X_1, X_2; \theta)$ is well defined for each θ and if $X_1=x_1$ and $X_2=x_2$ (x_1 and x_2 are arbitrary states from X), then we can examine $g(X_1, X_2; \theta)$ as a random function.

2. Let $\widehat{\Theta}=\{\widehat{\theta}\}$ be a set of the values of the estimators of θ. Since every estimator $\widehat{\theta}$ is an $\mathcal{F}_x$-measurable function, $\widehat{\theta}: X\to\Theta$, then $\widehat{\Theta}\subset\Theta$.

3. Considering the expected value of $g(X_1, X_2; \widehat{\theta})$ (where X_1 and X_2 are from the model) we can determine the expected log-likelihood, i. e. $\gamma(\widehat{\theta})=\boldsymbol{E}_\theta L_n(\widehat{\theta})$.

4. On the other hand, $\widehat{\theta}$ is a statistic with its own distribution. Then we can determine the expected value of $\gamma(\widehat{\theta})$ w. r. t. the distribution of $\widehat{\theta}$.

Akaike has stated his EMLP briefly as follows: among a lot of estimators choose one which will give the maximum of the expected log-likelihood function.

Note that the maximizing of $\gamma(\widehat{\theta})$ is equivalent to that of the information quantity $\boldsymbol{E}_\theta\{\ln(f(X_1, X_2; \widehat{\theta})/f(X_1, X_2; \theta))\}$ which is exactly $(-1)\times I(\theta: \widehat{\theta}; X_1, X_2)$ defined in part 2. By this interpretation it is natural to maximize $\gamma(\widehat{\theta})$ since it is equivalent to shorten the "distance" between the estimator and the parameter θ.

Comment 5. If we consider the following statistic $w(\theta^0; \widehat{\theta})=-(2/n)\Sigma_{i=1}^n\ln(f(X_i, X_{i+1}; \widehat{\theta})/f(X_i, X_{i+1}; \theta^0))$, then we have $\lim_{n\to\infty} w(\theta^0; \widehat{\theta})=2I(\theta^0: \widehat{\theta}; X_1, X_2)$ a. s. Here we apply the strong LLN, see Th. 1.1 in [4]. Then it is natural to expect that for large n the estimator providing the maximum of $L_n(\theta)$ will minimize the distance to θ^0.

Thus such an approach can be viewed as an extension of MLP, also taking into consideration C1.3 and (2.8).

Regarding the essence of the proposed extension it is natural to determine the information loss function

$$W(\theta^0; \widehat{\theta})=2I(\theta^0: \widehat{\theta}; X_1, X_2)=-2\boldsymbol{E}_{\theta^0}\{\ln\frac{f(X_1, X_2; \widehat{\theta})}{f(X_1, X_2; \theta^0)}\} \tag{3.2}$$

and the corresponding risk function

$$R(\theta^0; \widehat{\theta})=\boldsymbol{E}_{\widehat{\theta}}W(\theta^0; \widehat{\theta}) \tag{3.3}$$

w. r. t. the distribution of $\widehat{\theta}$.

4. Akaike's information criterion for TDMP. From the discussion above it becomes clear that there is an open problem concerning the point how to get reliable estimates for $W(\theta^0; \widehat{\theta})$ and $R(\theta^0; \widehat{\theta})$. A solution resulting in AIC unifying MLEs and the corresponding log-likelihood ratio statistics is proposed by H. Akaike [1] for the case of independent observations. We extend that technique to cover the case of TDMP.

Let ${}_k\Theta$ denote a k-dimensional subset of Θ with a generic point ${}_k\theta$ which means that ${}_k\theta_{k+1}={}_k\theta_{k+2}=\cdots={}_k\theta_r=0$.

A4.1. We assume that $\theta^0={}_r\theta^0$, i. e. we shall omit the subscript r of the value θ^0 or of any point of the original space Θ.

From the relationship (3.2), (2.7) and C3 it follows that $W(\theta^0; \theta)$ is smooth near θ^0. Also from Th.2.2 we obtain that $W(\theta^0; \theta)>0$ for $\theta\neq\theta^0$.

A4.2. Suppose that $W(\theta^0; {}_k\theta)$ has an unique minimum at ${}_k\theta^0$ given by

$$W(\theta^0; {}_k\theta^0)=\min_{{}_k\theta\in{}_k\Theta} W(\theta^0; {}_k\theta). \tag{4.1}$$

Lemma 4.1. *For an arbitrary* $\theta\in\Theta$ *we have* $W(\theta^0; \theta)=\|\theta-\theta^0\|_\Sigma^2$.

Proof. The statement follows from (3.2) and (2.7).

Thus the only minimum ${}_k\theta^0$ of the loss function over the subset ${}_k\Theta$ is determined by the relation

$$\|{}_k\theta^0-\theta^0\|^2_\Sigma=\min_{{}_k\theta\in{}_k\Omega}\|{}_k\Theta-\theta^0\|^2_\Sigma. \tag{4.2}$$

This means that the vector ${}_k\theta^0$ is the projection of θ^0 on ${}_k\Theta$ w. r. t. the metrics caused by Σ, i. e. $\langle\theta^0-{}_k\theta^0,\ {}_k\theta-{}_k\theta^0\rangle_\Sigma=0$ and consequently it is obtained from the system

$$\sum_{v=1}^{k}{}_k\theta^0_v\sigma_{uv}=\sum_{v=1}^{r}\theta^0_v\sigma_{uv},\quad u=1,\ 2,\dots,k. \tag{4.3}$$

Let us denote by ${}_k\widehat{\theta}$ and $\widehat{\theta}$ the MLEs for ${}_k\theta^0$ and θ^0, respectively. In these terms the loss function at ${}_k\widehat{\theta}$ can be written in the form

$$W(\theta^0;\ {}_k\widehat{\theta})=\|{}_k\widehat{\theta}-\theta^0\|^2_\Sigma=\|{}_k\widehat{\theta}-{}_k\theta^0+{}_k\theta^0-\theta^0\|^2_\Sigma$$

$$=\|{}_k\theta^0-\theta_0\|^2_\Sigma+\|{}_k\widehat{\theta}-{}_k\theta^0\|^2_\Sigma+2\langle{}_k\theta^0-\theta^0,\ {}_k\widehat{\theta}-{}_k\theta^0\rangle_\Sigma.$$

But the last addent is zero since

$$\langle{}_k\theta^0-\theta^0,\ {}_k\widehat{\theta}-{}_k\theta^0\rangle_\Sigma=\sum_{u=1}^{k}({}_k\widehat{\theta}_u-{}_k\theta^0_u)\Big[\sum_{v=1}^{r}({}_k\theta^0_v-\theta^0_v)\sigma_{uv}\Big]=0. \tag{4.4}$$

Lemma 4.2. *For the true value θ^0, its projection ${}_k\theta^0$ and MLEs ${}_k\widehat{\theta}$ in ${}_k\Theta$ of the model $[X_n,\ f(\xi,\ \eta;\ \theta),\ \Theta]$ the following holds:*

$$\|{}_k\widehat{\theta}-\theta^0\|^2_\Sigma=\|{}_k\theta^0-\theta^0\|^2_\Sigma+\|{}_k\widehat{\theta}-{}_k\theta^0\|^2_\Sigma. \tag{4.5}$$

Now consider two sample statistics, ${}_kw_r$ and ${}_k\eta_r$, where

$$\begin{aligned}{}_kw_r&=-2/n\sum_{i=1}^{n}\ln\frac{f(x_i,\ x_{i+1};\ {}_k\widehat{\theta})}{f(x_i,\ x_{i+1};\ \widehat{\theta})},\\ {}_k\eta_r&=n\times{}_kw_r.\end{aligned} \tag{4.6}$$

The fact that ${}_k\widehat{\theta}$ and $\widehat{\theta}$ are MLEs leads to the equations

$$\frac{\partial}{\partial\theta_u}L_n(\widehat{\theta})=\sum_{i=1}^{n}g_u(x_i,\ x_{i+1};\ \widehat{\theta})=0,$$

$$\frac{\partial}{\partial\theta_u}L_n({}_k\widehat{\theta})=\sum_{i=1}^{n}g_u(x_i,\ x_{i+1};\ {}_k\widehat{\theta})=0,$$

If we expand the function $g(x_i,\ x_{i+1};\ {}_k\theta^0)$ around ${}_k\widehat{\theta}$ and take a sum up to n, we find

$$\sum_{i=1}^{n}g(x_i,\ x_{i+1};\ {}_k\theta^0) \tag{4.7}$$

$$=\sum_{i=1}^{n}g(x_i,\ x_{i+1};\ {}_k\widehat{\theta})+(1/2)\sum_{u=1}^{r}\sum_{v=1}^{r}\sqrt{n}({}_k\widehat{\theta}_u-{}_k\theta^0_u)\sqrt{n}({}_k\widehat{\theta}_v-{}_k\theta^0_v)$$

$$\times(1/n)\sum_{i=1}^{n}g_{uv}(x_i,\ x_{i+1};\ {}_k\widehat{\theta})+R_1.$$

Analogously, around $\widehat{\theta}$,

$$\sum_{i=1}^{n} g(x_i, x_{i+1}; {}_k\theta^0)$$

$$= \sum_{i=1}^{n} g(x_i, x_{i+1}; \widehat{\theta}) + (1/2) \sum_{u=1}^{r} \sum_{v=1}^{r} \sqrt{n}(\widehat{\theta}_u - {}_k\theta_u^0)\sqrt{n}({}_k\widehat{\theta}_v - {}_k\theta_v^0)$$

$$\times (1/n) \sum_{i=1}^{n} g_{uv}(x_i, x_{i+1}; \widehat{\theta}) + R_2,$$

where

$$R_1 = (-1/3!) \sum_{u=1}^{k} \sum_{v=1}^{k} \sum_{w=1}^{k} ({}_k\widehat{\theta}_u - {}_k\theta_u^0)({}_k\widehat{\theta}_v - {}_k\theta_v^0)({}_k\widehat{\theta}_w - {}_k\theta_w^0) \sum_{i=1}^{n} g_{uvw}(x_i, x_{i+1}; \theta^1)$$

$$R_2 = (-1/3!) \sum_{u=1}^{r} \sum_{v=1}^{r} \sum_{w=1}^{r} (\widehat{\theta}_k - {}_k\theta_u^0)(\widehat{\theta}_v - {}_k\theta_v^0)(\widehat{\theta}_w - {}_k\theta_w^0) \sum_{i=1}^{n} g_{uvw}(x_i, x_{i+1}; \theta^2)$$

and ${}_k\widehat{\theta} < \theta^1 < {}_k\theta^0$, $\widehat{\theta} < \theta^2 < {}_k\theta^0$.
We want to examine the behaviour of R_1 and R_2 applying **C1.3.** It is possible to be done in case that $\widehat{\theta}$ and ${}_k\widehat{\theta}$ are in a neighbourhood ${}_kT_0$ of ${}_k\theta^0$. On the other hand, $\widehat{\theta}$ is a consistent estimator of θ^0 (Th.2.1 in [4]) and hence for sufficiently large n, $\widehat{\theta} \in T_0$ (a neighbourhood of θ^0). If $T = {}_kT_0 \cap T_0$ then for sufficiently large n the expansion under **C1.3.** is possible iff ${}_k\widehat{\theta}, \widehat{\theta} \in T$ and also ${}_k\theta^0$, $\theta^0 \in T$, i. e. it implies that ${}_k\theta^0$ and θ^0 have to be close. If we denote ${}_kG(x_1, x_{i+1}) = \sup_{\widetilde{\theta} \in T} | g_{uvw}(x_i, x_{i+1}; \widetilde{\theta})|$ and ${}_kM = E_\theta\{{}_kG(x_1, x_2)\}$, and apply the mean value theorem we obtain

$$R_1 = \sum_{i=1}^{n} {}_kG(x_i, x_{i+1})\,\alpha\,|{}_k\widehat{\theta} - {}_k\theta^0|^3,$$

$$R_2 = \sum_{i=1}^{n} {}_kG(x_i, x_{i+1})\alpha\,|\widehat{\theta} - {}_k\theta^0|^3, \text{ where } |\alpha| \leq r^3/6.$$

Since the law of large numbers holds (Th. 1.1 in [4]) then $P-\lim_{n\to\infty}\{(1/n)\Sigma_{i=1}^{n}{}_kG(X_i, X_{i+1})\} = {}_kM$ and because of Th. 2.2 in [4] quarantees that $\sqrt{n}\,|{}_k\widehat{\theta} - {}_k\theta^0|$ is bounded in probability as $u \to \infty$, or equivalently $\sqrt{n}\,|{}_k\theta - {}_k\widehat{\theta}^0|$ is of order $O_p(1)$.

$$R_1 = (1/n) \sum_{i=1}^{n} {}_kG(x_i, x_{i+1})\,\alpha\{\sqrt{n}\,|{}_k\widehat{\theta} - {}_k\theta^0|\}^3/\sqrt{n} \quad \text{i. e. } R_1 = O_p(1/\sqrt{n})$$

or also we can write that $R_1 = o_p(1)$.

C o m m e n t 6. Th. 2.2 in [4] states that if $l(n) = \sqrt{n}(\widehat{\theta} - \theta^0)$ (like a vector) then $l(n) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$. The symbol $\xrightarrow{d}$ means convergence in distribution. Therefore $l(n) = O_p(1)$.

Analogously we establish that

$$R_2 = 1/n \sum_{i=1}^{n} {}_kG(x_i, x_{i+1})\,\alpha\{\sqrt{n}\,|\widehat{\theta} - {}_k\theta^0|\}^3/\sqrt{n}, \text{ i. e. } R_2 = O_p(1/\sqrt{n}).$$

Since $\sqrt{n}\,|\widehat{\theta} - {}_k\theta^0| \leq \sqrt{n}\,|\widehat{\theta} - \theta^0| + \sqrt{n}\,|{}_k\theta^0 - \theta^0|$ is of probability order $O_p(1)$. Indeed the first term is subjected to Th. 2.2 [4] but for the second one (4.3) holds. Finally $R_2 = o_p(1)$.

Comment 7. Using the statement of Corollary 3 in [9] we verify that (since ${}_k\widehat{\theta}, \hat{\theta}, {}_k\theta^0, \theta^0 \in T$ and $\sqrt{n}({}_k\widehat{\theta}-{}_k\theta^0)=O_p(1)$, $\sqrt{n}(\widehat{\theta}-{}_k\theta^0)=O_p(1)$). Indeed, $R_1=\Sigma_{i=1}^n o_p(1/n)$ $=n\,o_p(1/n)$, $R_2=no_p(1/n)$ and hence $R_1=o_p(1)$, $R_2=o_p(1)$. Further we use "$\sim$" to indicate the asymptotic equivalence in probability.

Let us now discuss the asymptotic behaviour of the second terms in (4.7). Since $P-\lim_{n\to\infty}\widehat{\theta}=\theta^0$, C1.2 and LLN from Th. 1.1 in [4] hold, it is easy to see that $P-\lim_{n\to\infty}(1/n)\Sigma_{i=1}^n g_{uv}(x_i, x_{i+1}; \widehat{\theta})=-\sigma_{uv}$. As was just mentioned above $\sqrt{n}(\widehat{\theta}_u-{}_k\theta_u^0)$ $=O_p(1)$ for $u=1,\dots,r$. Thus

$$(1/2)\sum_{u=1}^{r}\sum_{v=1}^{r}\sqrt{n}(\widehat{\theta}_u-{}_k\theta_u^0)\sqrt{n}(\widehat{\theta}_v-{}_k\theta_v^0)(1/n)\sum_{i=1}^{n}g_{uv}(x_1, x_{i+1}; \widehat{\theta})$$

$$\sim-(n/2)\sum_{u=1}^{r}\sum_{v=1}^{r}(\widehat{\theta}_u-{}_k\theta_u^0)(\widehat{\theta}_v-{}_k\theta_v^0)\sigma_{uv}=-(n/2)\,\|\,\theta-{}_k\theta^0\,\|_\Sigma^2.$$

Analogously for the other second term we take into consideration that $P-\lim_{n\to\infty}{}_k\widehat{\theta}($ $={}_k\theta^0$, C1.2, LLN and C3 for $\Sigma(\theta)$ and conclude that for close values ${}_k\theta^0$ to θ^0, $P-\lim_{n\to\infty}(1/n)\Sigma_{i=1}^n g_{uv}(x_i, x_{i+1}; {}_k\widehat{\theta})=-\sigma_{uv}$. Thus from (4.7) we obtain

$$-2\sum_{i=1}^{n}\ln\frac{f(x_i, x_{i+1}; {}_k\widehat{\theta})}{f(x_i, x_{i+1}; \widehat{\theta})}\sim n\,\|\,\widehat{\theta}-{}_k\theta^0\,\|_\Sigma^2-n\,\|\,{}_k\widehat{\theta}-{}_k\theta^0\,\|_\Sigma^2. \tag{4.8}$$

The left hand side of (4.8) is exactly ${}_k\eta_r$ and hence the following result is valid.

Theorem 4.1. *For the model* $[X_n, f(\xi, \eta; \theta), \Theta]$ *we have the relation*

$${}_k\eta_r\sim n\,\|\,\widehat{\theta}-{}_k\theta^0\,\|_\Sigma^2-n\,\|\,{}_k\widehat{\theta}-{}_k\theta^0\,\|_\Sigma^2. \tag{4.9}$$

By a simple reflection we derive another result.

Lemma 4.3. (*Geometric interpretation*) ${}_k\widehat{\theta}-{}_k\theta^0$ *is approximately the projection* of $\widehat{\theta}-\theta^0$ *into* ${}_k\Theta$.

Proof. We have

$$(1/\sqrt{n})\sum_{i=1}^{n}g_u(x_i, x_{i+1}; {}_k\theta^0)$$

$$=-\sum_{v=1}^{k}\sqrt{n}({}_k\widehat{\theta}_v-{}_k\theta_v^0)(1/n)\sum_{i=1}^{n}g_{uv}(x_i, x_{i+1}; {}_k\widehat{\theta})+R_1'$$

$$=-\sum_{v=1}^{r}\sqrt{n}(\widehat{\theta}_v-{}_k\theta_v^0)(1/n)\sum_{i=1}^{n}g_{uv}(x_i, x_{i+1}; \widehat{\theta})+R_2'.$$

It is easy to show that the remainders R_i' are n times the term $(1/\sqrt{n})\,o_p(1/\sqrt{n})$ and therefore $R_i'=no_p(1/n)=o_p(1)$, $i=1, 2$ (see Comment 7). Hence omitting the remainders we can write

$$\sum_{v=1}^{k}\sqrt{n}({}_k\widehat{\theta}_v-{}_k\theta_v^0)\sigma_{uv}=\sum_{v=1}^{r}\sqrt{n}(\widehat{\theta}_v-{}_k\theta_v^0)\sigma_{uv}\quad u=1,\dots,r.$$

Since (4.3) holds for $u=1, 2,\dots,k$, then

$$\sum_{v=1}^{k}\sqrt{n}({}_k\widehat{\theta}_v-{}_k\theta_v^0)\sigma_{uv}=\sum_{v=1}^{r}\sqrt{n}(\widehat{\theta}_v-\theta_v^0)\sigma_{uv}. \tag{4.10}$$

We can express ${}_k\eta_r$ in another way too, namely,

$$
{}_k\eta_r \sim n \,\|\, \widehat{\theta} - {}_k\theta^0 + \theta^0 - \theta^0 \,\|_\Sigma^2 - n \,\|\, {}_k0 - {}_k\theta^0 \,\|_\Sigma^2
$$

$$
= n \,\|\, \widehat{\theta} - \theta^0 \,\|_\Sigma^2 + n \,\|\, {}_k\theta^0 - \theta^0 \,\|_\Sigma^2 - n \,\|\, {}_k\widehat{\theta} - {}_k\theta^0 \,\|_\Sigma^2 - 2n \langle \widehat{\theta} - \theta^0,\ {}_k\theta^0 - \theta^0 \rangle_\Sigma .
$$

For the inner product we have

$$
(4.11) \qquad n\langle \widehat{\theta} - \theta^0,\ {}_k\theta - \theta^0 \rangle_\Sigma = n \sum_{u=1}^{r} \sum_{v=1}^{r} (\widehat{\theta}_u - \theta_u^0)({}_k\theta_v^0 - \theta_v^0)\sigma_{uv}
$$

$$
= \sum_{u=1}^{k} [\sqrt{n} \sum_{v=1}^{r} (\widehat{\theta}_v - \theta_v^0)\,\sigma_{uv}]\,\sqrt{n}({}_k\theta_u^0 - \theta_u^0) = n\langle {}_k\widehat{\theta} - {}_k\theta^0,\ {}_k\theta^0 - \theta^0 \rangle_\Sigma = 0.
$$

It is natural to expect that (4.11) holds because of the geometric interpretation we have given. Thus

$$
(4.12) \qquad {}_k\eta_r \sim n \,\|\, \widehat{\theta} - \theta^0 \,\|_\Sigma^2 + n \,\|\, {}_k\theta^0 - \theta^0 \,\|_\Sigma^2 - n \,\|\, {}_k\widehat{\theta} - {}_k\theta^0 \,\|_\Sigma^2 .
$$

Theorem 4.2. *The log-likelihood ratio statistic ${}_k\eta_r$ has asymptotically the non-central chi-square distribution with $r-k$ degrees of freedom.*

Proof. For the first term in the right side of (4.12) we have $n\,\|\widehat{\theta} - \theta^0\|_\Sigma^2 = 2[L_n(\widehat{\theta}) - L_n(\theta^0)]$ and since Th. 2.2 in [4] holds then $n\,\|\widehat{\theta} - \theta^0\|_\Sigma^2 \xrightarrow{d} \chi_r^2$. Also applying the projection theorem Th. 11.2 in [4] we obtain that $n\,\|\,{}_k\widehat{\theta} - {}_k\theta^0\,\|_\Sigma^2 \xrightarrow{d} \chi_k^2$. These conclusions can be derived also from Th. 3.1 in [4] which quarantees the independence of the limit distributions. But that independence holds since the geometric interpretation we gave. Th. 3 in [9] implies that ${}_k\eta_r \xrightarrow{d} \chi_{r-k}^2$. The second member of (4.12) is $n\,\|\,{}_k\theta^0 - \theta^0\|_\Sigma^2$. It was mentioned above (regarding the remainder R_2) that this member is of order $O_p(1)$ (because of (4.2) or (4.3)). So it gives nothing to the asymptotic distribution of ${}_k\eta_r$ except to show that there is a shift. Thus for fixed n the noncentrality of χ_{r-k}^2 is determined by $n\,\|\,{}_k\theta^0 - \theta^0\,\|_\Sigma^2$. We suppose that Tong's quotation on p. 492 for (4.12) is unproper, the Theorem on p. 16 [4] ensists on requirements which are not satisfied.

Thus from (4.12) we derive

$$
{}_k\eta_r \sim n \,\|\, \widehat{\theta} - \theta^0 \,\|_\Sigma^2 + n \,\|\, {}_k\theta^0 - \theta^0 \,\|_\Sigma^2 + n \,\|\, {}_k\widehat{\theta} - {}_k\theta^0 \,\|_\Sigma^2 - 2n \,\|\, {}_k\widehat{\theta} - {}_k\theta^0 \,\|_\Sigma^2
$$

$$
= n \,\|\, \widehat{\theta} - \theta^0 \,\|_\Sigma^2 + n \,\|\, {}_k\widehat{\theta} - \theta^0 \,\|_\Sigma^2 - 2n \,\|\, {}_k\widehat{\theta} - {}_k\theta^0 \,\|_\Sigma^2
$$

or equivalently,

$$
W(\theta^0;\ {}_k\widehat{\theta}) \sim n^{-1}({}_k\eta_r - n \,\|\, \widehat{\theta} - \theta^0 \,\|_\Sigma^2 + 2n \,\|\, {}_k\widehat{\theta} - {}_k\theta^0 \,\|_\Sigma^2).
$$

Thus for the loss function we found asymptotically equivalent statistic in probability. But it seems easier to minimize the risk function. Indeed, let us apply the expectation operator to both the sides w. r. t. the distribution of the estimators. So we obtain a proper statistic for $R(\theta^0;\ {}_k\widehat{\theta})$, i. e.

$$
r(\widehat{\theta};\ {}_k\widehat{\theta}) = n^{-1}\ ({}_k\eta_r + 2k - r).
$$

Definition 3. *The value k which provides the minimum of the statistic $r(\widehat{\theta};\ {}_k\widehat{\theta})$ is called Minimum Akaike Information Criterion Estimate, denoted usually by MAICE.*

Since r and n are fixed for any observation then we come to the next result.

Lemma 4.4. *The following functions are equivalent forms of the risk statistic:*

F1. $R_1(k)=-2\sum_{i=1}^{n}\ln f(x_i,\ x_{i+1};\ {}_k\hat{\theta})+2k.$

F2. $R_2(k)={}_k\eta_r+2k.$

F3. $R(k)={}_k\eta_r-2\times(d.\,f.\ \text{of}\ {}_k\eta_r).$

Proof. Everywhere n is omitted. To the last form $-r$ is added. The equivalence is obvious.

For the purpose of investigating the order of a Markov chain the most convenient form is F3. We are going to discuss it in the next section. Before that, let us consider a natural extension of Th. 3.1 in [4].

Suppose the model $[X_n,\ f(\xi,\ \eta;\ \theta),\ \Theta]$ is given. Take a sequence of open subsets $\Theta_1,\ \Theta_2,\dots,\Theta_\nu$ of Euclidean spaces E^{r_i}, $i=1,\ 2,\dots,\nu$, respectively, and $r_1\leqq r_2\leqq\dots\leqq r_\nu$. For each $i=1,\ 2,\dots,\nu$ the mapping ${}^ih:\ \Theta_i\to\Theta_{i+1}$ satisfies the following regularity condition. (We denote by ${}^i\theta$ the generic point of Θ_i, i. e. if ${}^i\theta=({}^i\theta_1,\dots,{}^i\theta_{r_i})$, then ${}^ih({}^i\theta)={}^{i+1}\theta$ and ${}^ih_j({}^i\theta)={}^{i+1}\theta_j$, $j=1,\dots,r_{i+1}$; $i=1,\dots,\nu$, also $\Theta_{\nu+1}=\Theta$.)

C4. For each $j=1,\dots,r_{i+1}$, ${}^ih_j$ has continuous third-order partial derivatives and the $r_{i+1}\times r_i$ matrix $K({}^i\theta)$ with entries $\{K({}^i\theta)\}_{jl}=\dfrac{\partial^i h_j({}^i\theta)}{\partial^i\theta_l}$, $j=1,\dots,v_{i+1}$; $l=1,\dots,r_i$ has rank r_i throughout Θ_i.

By Th. 3.1 in [4] follows that all the models $[X_n,\ f(\xi,\ \eta;\ {}^i\theta),\ \Theta_i]$ satisfy conditions **C1** and **C2.** Moreover, ${}^{i+1}\theta^0={}^ih({}^i\theta^0)$ is the image of the true value of the parameter, $\theta^0={}^\nu h({}^\nu\theta^0)={}^\nu h({}^{\nu-1}h(\cdots({}^1h({}^1\theta^0))\cdots))$.

Theorem 4.3. *Suppose $\Theta_1,\ \Theta_2,\dots,\Theta_\nu$ are the subsets described above and the mappings ${}^ih$, $i=1,\ 2,\dots,\nu$ satisfy* C4. *If the true parameter point lies in Θ_1, then*

$$2[\max\nolimits_{\Theta_i}L_n-\max\nolimits_{\Theta_{i-1}}L_n]\xrightarrow{d}\chi^2_{r_i-r_{i-1}},\quad i=2,\dots,\nu \tag{4.13}$$

$$2[\max\nolimits_{\Theta_i}L_n-L_n({}^1\theta^0)]\xrightarrow{d}\chi^2_{r_i},\quad i=1,\dots,\nu. \tag{4.14}$$

Moreover, the ν statistics

$$2[\max\nolimits_{\Theta_1}L_n-L_n({}^1\theta^0)],\ 2[\max\nolimits_{\Theta_i}L_n-\max\nolimits_{\Theta_{i-1}}L_n],\quad i=2,\dots,\nu \tag{4.15}$$

are asymptotically independent.

Proof. Since Th. 3.1 in [4] holds, then iteratively for each $i=1,\dots,\nu-1$, $\Phi=\Theta_i$ and $\Theta=\Theta_{i+1}$ and hence

$$2\,[\max\nolimits_{\Theta_{i+1}}L_n-L_n^0]\xrightarrow{d}\chi^2_{r_{i+1}},\quad 2[\max\nolimits_{\Theta_i}L_n-L_n^0]\xrightarrow{d}\chi^2_{r_i},$$

$$2[\max\nolimits_{\Theta_{i+1}}L_n-\max\nolimits_{\Theta_i}L_n]\xrightarrow{d}\chi^2_{r_{i+1}-r_i}.$$

The last two statistics are asymptotically independent. Immediately it follows that statistics given by (4.13) are asymptotically independent, see Th. 11.2 in [4]. For a fixed i, $(i=2,\dots,\nu)$

$$2[\max\nolimits_{\Theta_i}L_n-L_n^0]=2[\max\nolimits_{\Theta_i}L_n-\max\nolimits_{\Theta_1}L_n]+2[\max\nolimits_{\Theta_1}L_n-L_n^0]$$

$$= \sum_{j=2}^{i} 2[\max_{\Theta_j} L_n - \max_{\Theta_{j-1}} L_n] + 2[\max_{\Theta_1} L_n - L_n^0],$$

where every addend is $\chi^2_{r_j - r_{j-1}}$ and all the addends are asymptotically independent. Therefore, $2[\max_{\Theta_1} L_n - L_n^0]$ is asymptotically independent in regard to statistics of (4.13), i. e. the statement for the statistics given by (4.15) is valid.

5. Finite discrete time Markov processes. Let us discuss the case when X is finite and TDMPs are called Markov chains. To simplify the notations we shall denote the state space by the first natural numbers $X=\{1, 2, \ldots, s\}$. Here the measure λ is the counting one. Instead of density functions w. r. t. λ we shall use the transition probabilities $p_{ij}(\theta) = \boldsymbol{P}\{X_{n+1}=j \,\|\, X_n=i\}$. They form a transition matrix $\boldsymbol{P}(\theta)$ of size $s\times s$, so the model can be rewritten as $[X_n, \boldsymbol{P}(\theta), \Theta]$. It is easy to see that conditions **C1** and **C2** are simply consequences of the following one.

C5. The set $D=\{(i, j)\colon p_{ij}(\theta)>0\}$ is independent of Θ and any $p_{ij}(\theta)$ has continuous partial derivatives of third-order throughout Θ. Moreover, if d is the number of elements in D, then the $d\times r$ matrix

(5.1) $$\frac{\partial}{\partial\theta_u} p_{ij}(\theta);\ (i, j)\in D,\ u=1, \ldots, r \text{ has rank } r \text{ throughout } \Theta.$$

A3. We shall consider only Markov chains which are irreducible and aperiodic and such that every recurrent state is nonnull.

The log-likelihood for the observation $(x_1, x_2, \ldots, x_{n+1})$ is

(5.2) $$L_n(\theta) = \Sigma_D n_{ij} \ln p_{ij}(\theta),$$

where

$$n_{ij} = \sum_{l=1}^{n} I_{X_l}(i)\, I_{X_{l+1}}(j),$$

i. e. the frequency of the transitions from state i to state j ($I_{X_l}(i)$ is the indicator function of the event that l-th member of the observation x_l takes value i). Note that $\Sigma_j p_{ij}(\theta)=1$, then $\Sigma_j \frac{\partial}{\partial\theta_u} p_{ij}(\theta)=0$ for each u, $i=1, \ldots, s$ (since the matrix (5.1) has rank r) and consequently $r\leq d-s$. The stationary transition probabilities can be estimated by maximizing (5.2) w. r. t. p_{ij} subject to restrictions $p_{ij}\geq 0$ and $\Sigma_{j=1}^{s} p_{il}=1$. Using Lagrange multipliers we obtain that MLEs are $\widehat{p}_{ij}=n_{ij}/n_i$, where $n_i=\Sigma_{j=1}^{s} n_{ij}$, i. e. the frequency of appearance of state i. The above is done under the interpretation of Θ as "equal" to D (p. 26 [4]). Thus the hypothesis $\Theta(D)$ is that the process is a Markov chain. When we concern of testing a composite null hypothesis Φ within $\Theta=D$, then **C5** is generalized to **C5.2** in [4].

The above reflections are explicitly given for simple Markov chain, i. e. for a chain of order one.

Definition 4. *A Markov chain is said to be of order t if the following equation relating the conditional probabilities is satisfied: t is the smallest positive integer such that for all n,*

$$\boldsymbol{P}\{X_n \mid X_{n-1}, X_{n-2}, \ldots\} = \boldsymbol{P}\{X_n \mid X_{n-1}, X_{n-2}, \ldots, X_{n-t}\}.$$

Definition 5. *The chain is said to be of order 0 if it is a sequence of independent random variables.*

If $\{X_n\}$ is a t-th order Markov chain, then its transition matrix is $P=(p_{\nu_1 \ldots \nu_t : \nu_{t+1}})$, where for $j=1, \ldots, t+1$, $\nu_j\in\{1, \ldots, s\}$. Here $p_{\nu_1 \ldots \nu_t : \nu_{t+1}}>0$ because of condition (A3).

It is possible to give a similar interpretation: so $d=s^{t+1}$ and MLEs are $\widehat{p}_{\nu_1 \ldots \nu_t : \nu_{t+1}} = \frac{n_{\nu_1 \ldots \nu_t \nu_{t+1}}}{n_{\nu_1 \ldots \nu_t}}$. The LLF is

$$\widehat{L}_n = \sum_{\nu_1 \ldots \nu_{t+1}} n_{\nu_1 \ldots \nu_{t+1}} \ln \frac{n_{\nu_1 \ldots \nu_t \nu_{t+1}}}{n_{\nu_1 \ldots \nu_t}},$$

where $n_{\nu_1 \ldots \nu_t} = \Sigma_{\nu_{t+1}} n_{\nu_1 \ldots \nu_t \nu_{t+1}}$.

In [4, 5] Billingsley has shown how a t-th order Markov chain can be associated with a derived process and the last one can be expressed as a first order process. So most of the theorems concerning statistical inference for simple Markov chains can easily be extended to the case of multiple ones.

If we denote by H_t the hypothesis that $\{X_n\}$ is a t-th order Markov chain with a regular transition matrix, we can generalize condition **C5.2** in [4] to the following one.

C6. For each Φ in open subset $\Phi \subset E^c$, $P(\Phi)=(p_{\nu_1 \ldots \nu_t : \nu_{t+1}}(\Phi))$ is a t-th order stochastic matrix with positive elements. Each element has continuous third-order partial derivatives and the $s^{t+1} \times c$ matrix with entries $\frac{\partial}{\partial \Phi_j} p_{\nu_1 \ldots \nu_t : \nu_{t+1}}(\Phi)$ $j=1, \ldots, c$ has rank c throughout Φ.

Then it is not difficult to see how to use Th. 4.3 and derive a generalization of Th. 6.3 in [4] for a t-th order chain.

Theorem 5.1. Suppose **C6** is satisfied and $\{X_n\}$ is t-th order Markov chain with transition matrix $P(\Phi^0)$ for some $\Phi^0 \in \Phi$. Then for any fixed m,

$$2[\max_{H_i} L_n - \max_{H_{i-1}} L_n] \xrightarrow{d} \chi^2_{(s^i - s^{i-1})(s-1)}, \quad t+1 \leq i \leq m,$$

$$2[\max_{H_t} L_n - \max_{\Phi} L_n] \xrightarrow{d} \chi^2_{s^{t+1}-s^t-c}$$

and the statistics are asymptotically independent. If Φ consists of a single point, i. e. there is no parameter to be estimated, then $c=0$. In this case the statement about the last statistic maintains that one in Th. 6.1 in [4]. If Φ is such that $\Phi = H_0$, then $c=s-1$.

Returning to Section 4, we see that the definition of ${}_k\eta_r$ as a log-likelihood ratio statistic (closely related to Neyman — Pearson criterion) implies that

$${}_l\lambda_m = {}_k\eta_r = 2[\max_{H_m} L_n - \max_{H_l} L_n] \xrightarrow{d} \chi^2_{(s^{m+1}-s^m)-(s^{l+1}-s^l)},$$

where $r=s^{m+1}-s^m$, $k=s^{l+1}-s^l$ and using the notation $\nabla s^j = s^{j+1}-s^j$, we have

$${}_l\lambda_m \xrightarrow{d} \chi^2_{(s^m - s^l)(s-1)} \text{ or } \chi^2_{(\nabla s^m - \nabla s^l)}.$$

Then using F3 form of the equivalent representations of the risk function (Lemma 4.4), we determine the statistic

$$\mathrm{AIC}(l) = {}_l\lambda_m - 2(\nabla s^m - \nabla s^l)$$

or the so-called Akaike's Information Criterion and redefine MAICE.

Definition 6. *AIC estimator $\widehat{l} = \widehat{l}_{\mathrm{AIC}}$ of the order of a Markov chain, the so-called MAICE, is chosen such that*

$$\mathrm{AIC}(\widehat{l}) = \min_{0 \leq l \leq m-1} \mathrm{AIC}(l).$$

Theorem 5.2. *Let the true order of the model* $[X_n, P(\theta), \Theta]$ *be known and denote it by p. Then MAICE* $\widehat{l}$ *is inconsistent estimator of p, i. e.*

$$\lim_{n\to\infty} P\{\widehat{l}=l\}=0, \quad if \quad 0\leqq l<p,$$

$$\lim_{n\to\infty} P\{\widehat{l}=l\}>0, \quad if \quad p\leqq l<m.$$

Proof. Let $0\leqq l<p$. Then

$$P\{\widehat{l}=l\}=P\{\text{AIC}(l)\leqq \text{AIC}(j), \quad 0\leqq j\leqq m-1\}$$
$$\leqq P\{\text{AIC}(l)\leqq \text{AIC}(p)\}.$$

Since Th. 5.1 holds, then under any alternate hypothesis H_l the test $\xi=\text{AIC}(l)-\text{AIC}(p)$ is consistent (i. e. it degenerates at ∞, see [3]). Thus the above probability can be lessen to an arbitrary positive number, or equivalently $\lim_{n\to\infty} P\{\widehat{l}=l\}=0$.

Now let $p\leqq l<m$. Then

$$P\{l=l\}=P\{\text{AIC}(l)\leqq \text{AIC}(j), \quad 0\leqq j\leqq m-1\}$$
$$\cong P\{\text{AIC}(l)\leqq \text{AIC}(j), \quad p\leqq j\leqq m-1\} \text{ (since the above)}$$
$$=P\{\text{AIC}(l)\leqq \text{AIC}(j),\ p\leqq j<l \text{ and } \text{AIC}(l)\leqq \text{AIC}(j), \quad l\leqq j\leqq m-1\}.$$

For $\xi_j=\text{AIC}(j)-\text{AIC}(l)$, $p\leqq j\leqq l-1$, since Th. 5.1 holds, we obtain that $\xi_j={}_j\lambda_l-2(s-s^j)(s-1)$ where ${}_j\lambda_l$ is chi-square distributed with $(\nabla s^l-\nabla s^j)$ degrees of freedom For $\xi_j=\text{AIC}(l)-\text{AIC}(j)$, $l\leqq j\leqq m-1$, applying Th. 5.1 again, we obtain that $\xi_j={}_l\lambda_j-2(s^j-s^l)(s-1)$ where ${}_l\lambda_j$ is chi-square distributed with $(\nabla s^j-\nabla s^l)$ degrees of freedom.

For any $p\leqq j\leqq m-1$ we can express ξ_j as a sum of independent identically chi-square distributed random variables with one degree of freedom. Indeed, let $Z_1, Z_2, \ldots$ be i. i. d. χ_1^2 and put $S_k=\Sigma_{j=1}^k(Z_j-2)$.

If $i=j-p+1$ and thus $1\leqq i\leqq m-p$, by using the notation $a_i=\nabla s^{i+p-1}\nabla s^l=a_{l-p+1}$. Then for $1\leqq i\leqq m-p$ $\nabla s^{i+p-1}=a_i$, $\nabla s^l=a_{l-p+1}$, and in details for $1\leqq i\leqq l-p$ $S_{a_{l-p+1}}-S_{a_i}\geqq 0$, for $l-p+1\leqq i\leqq m-p$ we have $S_{a_i}-S_{a_{l-p+1}}\leqq 0$. Note that $a_i<a_{l-p+1}$ for $1\leqq i\leqq l-p$ and $a_i\geqq a_{l-p+1}$, $l-p+1\leqq i\leqq m-p$.

Finally

$$P\{\widehat{l}=l\}\cong P\{\text{AIC}(l)\leqq \text{AIC}(j),\ p\leqq i<l \text{ and } \text{AIC}(l)\leqq \text{AIC}(j),\ l\leqq j\leqq m-1\}$$
$$=P\{S_{a_{l-p+1}}-S_{a_i}\geqq 0,\ 1\leqq i\leqq l-p \text{ and } S_{a_i}-S_{a_{l-p+1}}\leqq 0,\ l-p+1\leqq i\leqq m-p\}.$$

This form is similar to that one obtained for the case of independent observations, see [2]. Moreover, there it is possible to find out an explicit formula for the above probability, solving a random walk problem and hence to find explicitly the asymptotic distribution. The analogous problem arising here for not identically distributed random variables has no apparent analytical solution. Anyway, avoiding that exact form, we gave a similar representation to that one in the case of independent observations in order to point out the relation between them. We think that it is easier to realize that the left hand side probability is positive one (since the distribution of the differences of AIC(l) statistics with $l\geqq p$ tends to a non-degenerate distribution and the representations by S sums). Thus the overestimating of the true parameter is clear.

Acknowledgements. I wish to thank my research superviser Dr N. Yanev for our fruitful discussions. I am grateful to Dr D. Vandev for his suggestions. I am much obliged to the editor Dr J. Stoyanov for his recommendations.

The referee's comments are most gratefully acknowledged.

REFERENCES

1. H. Akaike. Information theory and an extension of the maximum likelihood principle. — In: Second Internat. Symp. Inform. Theory. (Eds. B. N. Petrov and F. Csaki). Budapest, 1972, 267-281.
2. J. Anděl. Fitting models in time series analysis. *Math. Operationsforsch. Statist., Ser. Statistics*. **13**, 1982, 121-143.
3. T. Anderson, L. Goodman. Statistical inference about Markov chains. *Ann. Math. Statist.*, **28**, 1957, 89-110.
4. P. Billingsley. Statistical inference for Markov processes. New York, 1961.
5. P. Billingsley. Statistical methods in Markov chains. *Ann. Math. Statist.*, **32**, 1961, 12-40.
6. R. Katz. On some criteria for estimating the order of a Markov chain. *Technometrics*, **23**, 1981, 243-249.
7. S. Kullback, R. Leibler. On information and sufficiency. *Ann. Math. Statist.*, **22**, 1951, 79-89.
8. H. Tong. Determination of the order of a Markov chain by Akaike's Information Criterion. *J. Appl Probabilty*, **12**, 1975, 488-497.
9. A. Wald, H. Mann. On stochastic limit and order relationships. *Ann. Math. Statist.*, **14**, 1943, 217-226.
10. N. Yanev, V. Nikolova, I. Tzankova, J. Yaneva, I. Ivanov. Approximation of the nucleotide sequence in DNAs with Markov chains. — In: Mathematics and Mathem. Education (Proc. 12th Spring Conf. UBM, Sunny Beach, April'83). Sofia, 1983, 268-270.
11. P. Billingsley. Convergence of probability measures. New York, 1968.
12. S. Kullback. Information theory and statistics. New York, 1959.

Faculty of Mathematics and Informatics
University of Sofia
1126 Sofia, Bulgaria

Received 28. 4. 1985