STATISTICAL INFERENCE FOR PROBABILISTIC FUNCTIONS OF FINITE STATE MARKOV CHAINS

BY LEONARD E. BAUM AND TED PETRIE

Institute for Defense Analyses, Princeton, N. J.

Let $\{X_t\}$ be an s state Markov process, generated by some $s \times s$ stochastic matrix $\{a_{ij}\}$ with positive entries. Let $\{Y_t\}$ be a *probabilistic function* of $\{X_t\}$, viz:

$$(0.1) \qquad P\{Y_t = k \mid X_t = j, Y_{t-1}, X_{t-1}, \cdots\} = b_{jk}$$

where $\{b_{jk}\}$ is an $s \times r$ matrix with positive entries and row sums $= 1$.

This paper deals with statistical estimation. We assume that the matrices $A = \{a_{ij}\}$ and $B = \{b_{jk}\}$ are unknown and we wish to recover them from an observation $\{Y_1, \cdots, Y_T\}$.

We prove that the maximum likelihood estimate converges to the correct value. We also show that the (χ^2) theory of power, estimation, and testing applies. In passing we observe that there is a per character rate of distinguishability between a correct and incorrect hypothesis. We thus carry over the standard statistical estimation theory for independent sampling or Markov chains to our case in which the processes are not generally Markov of any order.

A word about the proofs and their motivation. Let θ and θ_0 be two hypotheses as to the nature of a stochastic finite state process $\{Y_t, -\infty < t < \infty\}$. Let P_θ and P_{θ_0} be the measure on the space of infinite sequences $\{Y_t, -\infty < t < \infty\}$ determined by θ and θ_0. If θ_0 is correct how does the random variable T^{-1} log $\{P_{\theta_0}[Y_1 \cdots Y_T]/P_\theta[Y_1 \cdots Y_T]\}$ behave? By the Shannon, McMillan, Breiman theorem

$$(0.2) \qquad -T^{-1} \log P_{\theta_0}[Y_1 \cdots Y_T] \rightarrow_{\text{a.e.}} -H(\theta_0)$$

the entropy of the θ_0 process.

Let θ and θ_0 be the hypothesis that the $\{Y_t\}$ process is a probabilistic function of a Markov chain with associated matrices $((a_{ij}(x)))$ and $((b_{jk}(x)))$, $x = \theta$ or θ_0. We suppose $a_{ij}(x) > 0$, $b_{jk}(x) > 0$. Then

$$(0.3) \qquad \lim_{T\to\infty} -T^{-1} \log P_\theta[Y_1 \cdots Y_T] = -H(\theta)$$

exists a.e. P_{θ_0} and $H(\theta) < H(\theta_0)$ if P_θ is not the same measure as P_{θ_0}. The proof of this fact is strongly motivated by Khinchin's proof of (0.2), [6]. Our proof rests on the fact that $\lim_{T\to\infty} P_\theta[Y_0 \mid Y_{-1} \cdots Y_{-T}] = f[\theta, Y]$ exists for every $Y = \{Y_t - \infty < t < \infty\}$. We also show that $f[\theta, Y]$ has three partial derivatives with respect to the matrix coordinates $a_{ij}(\theta)$ and $b_{jk}(\theta)$. The continuity of $H(\theta)$ together with a slightly stronger result than (0.3) gives us our main theorems.

Received 4 April 1966.

1. Preliminaries. Our index t ranges over Z the integers. Let θ refer to an arbitrary pair $\langle\{a_{ij}(\theta)\}, \{b_{jk}(\theta)\}\rangle$. All our Markov processes $\{X_t\}: t = \cdots, -2, -1, 0, 1, 2, \cdots$ are stationary; i.e., we take stationary distributions for X_1. All the processes $\{Y_t\}$ defined in (0.1) are then also stationary.

Rather than considering probabilistic functions $\{Y_t\}$ of a Markov process $\{X_t\}$ it is convenient to reduce to (deterministic = lumping) functions of a Markov process as follows. Define a new Markov process $\{X_t'\}$ with state space $S' = \{1, \cdots, s\} \times \{1, \cdots, r\}$ and transition matrix $a_{\langle i,k\rangle, \langle i', k'\rangle} = a_{ii'}b_{i'k'}$. Let $f: S' \to \{1, \cdots, r\}$ be defined by $f\langle i, k\rangle = k$. The process $\{Y_t'\}$ where $Y_t' = f(X_t')$ is a deterministic function of the Markov process $\{X_t'\}$ which is equivalent to $\{Y_t\}$. In fact:

$$P(Y_1 = r_1, Y_2 = r_2, \cdots, Y_n = r_n) = \textstyle\sum_{i_1,i_2,\cdots,i_n=1}^{s} a_{i_1}b_{i_1r_1}a_{i_1i_2}b_{i_2r_2}\cdots a_{i_{n-1}i_n}b_{i_nr_n}$$

where a_{i_1} is the stationary probability that $x_1 = i_1$, while,

$$\begin{aligned} P(Y_1' = r_1, Y_2' = r_2, \cdots, Y_n' = r_n) &= \textstyle\sum_{\langle i_t,k_t\rangle \varepsilon f^{-1}(r_t), t=1,\cdots,n} a_{\langle i_1,k_1\rangle,\langle i_2,k_2\rangle}\cdots a_{\langle i_{n-1}\rangle,\langle i_n,k_n\rangle} \\ &= \textstyle\sum_{i_1\cdots i_n=1}^{s} a_{\langle i_1,r_1\rangle}a_{\langle i_1,r_1\rangle,\langle i_2,r_2\rangle}\cdots a_{\langle i_{n-1},r_{n-1}\rangle,\langle i_n,r_n\rangle} \\ &= \textstyle\sum_{i_1\cdots i_n=1}^{s} a_{i_1}b_{i_1r_1}a_{i_1i_2}b_{i_2r_2}\cdots a_{i_{n-1}i_n}b_{i_nr_n}. \end{aligned}$$

The mapping $\mathcal{S}: \langle\{a_{ij}\}, \{b_{ik}\}\rangle \to \{a_{\langle i,k\rangle,\langle i',k'\rangle}\}$ is a C^∞ 1–1 mapping of the $s(s-1) + s(r-1)$ dimensional set of all $s \times s$ stochastic matrices $\times$ the set of all $s \times r$ "stochastic" matrices onto an $s(s-1) + s(r-1)$ dimensional subset of the $rs(rs-1)$ dimensional set of all $rs \times rs$ stochastic matrices. We have just seen that the set of all $\{Y_t\}$ processes which are probabilistic r valued functions of an s state Markov process can be considered as a subset of the set of all r valued processes which are deterministic functions of an rs state Markov process.

In general, for the mapping function $f: f(i, k) = k, i = 1, \cdots, s, k = 1, \cdots, r$, there will be an $rs^2 - rs$ dimensional set of $rs \times rs$ stochastic matrices which yield equivalent $\{Y_t\}$ processes (see [5]). However, if only $rs \times rs$ matrices which are in the range of the above mapping $\mathcal{S}$ are allowed, since $s(s-1) + s(r-1) + rs^2 - rs < rs(rs-1)$ if $s \geqq r \geqq 2$, "in general" there will be a unique such matrix yielding the given Y process.

In the following we will restrict our discussion to matrices $\{a_{ij}\}$, $\{b_{jk}\}$ all of whose entries are >0. All the entries $a_{\langle i,k\rangle,\langle i',k'\rangle} = a_{ii'}b_{i'k'}$ are then also positive. In particular the matrix $\{a_{\langle i,k\rangle,\langle i',k'\rangle}\}$ yields an ergodic $\{X_t'\}$ process. $\{Y_t\} = \{Y_t'\}$ being a function of this ergodic process is then also ergodic.

In Section 2 we will consider the imbedding $\mathcal{S}: \langle\{a_{ij}\}, \{b_{jk}\}\rangle \to rs \times rs$ matrix space as having been made. The rs state Markov process will have state space S and be referred to as $\{X_t\}$ (without a prime), and the process $\{fX_t\}$ denoted $\{Y_t\}$. Probabilities computed according to the matrix θ in $rs \times rs$ space will be denoted by P_θ. Θ_δ will denote the set of matrices all of whose entries are $\geqq \delta > 0$.

2. Lemmas. The proofs of this section are similar to [4], p. 173.

For each $i \varepsilon Z$ let $S_i = S$ and $R_i = R$. Let $S^\infty = \prod_{i\varepsilon Z} S$ and $R^\infty = \prod_{i\varepsilon Z} R$. ω is a point in S^∞ and $Y(\omega)$ is the point in R^∞ with coordinates $\{Y_t(\omega) = fX_t(\omega), t \varepsilon Z\}$.

DEFINITION. Let C_t denote a cylinder set in R^∞, of the form $\{Y \mid Y_{t_j} = k_{t_j}, t_j \varepsilon T, t_j \geqq t\}$ or a cylinder set in S^∞ of the form $\{X \mid X_{t_j} = i_{t_j}, t_j \varepsilon T, t_j \geqq t\}$. We shall be considering random variables on R of the form $P_\theta[C_t \mid X_s = j, Y_{t_j}(\omega), t_j \varepsilon T] = W[Y(\omega)]$. Define

$$M^+[\theta, C_t, \{Y_{t_j}(\omega) \mid t_j \varepsilon T\}, d] = \max_i P_\theta[C_t \mid X_{t-d} = i, Y_{t_j}(\omega), t_j \varepsilon T]$$

$$M^-[\theta, C_t, \{Y_{t_j}(\omega) \mid t_j \varepsilon T\}, d] = \min_i P_\theta[C_t \mid X_{t-d} = i, Y_{t_j}(\omega), t_j \varepsilon T].$$

We will sometimes abbreviate these respectively by M_d^+ and M_d^-.

LEMMA 2.1. *$P_\theta(X_{t+1} = j \mid X_t = i, Y_{t_j}(\omega), t_j \varepsilon T) > \mu_\delta$ for some $\mu_\delta > 0$ independent of T, t, $Y(\omega)$ and $\theta \varepsilon \Theta_\delta$, provided that if $t + 1 \varepsilon T$, $j \varepsilon f^{-1}(Y_{t+1}(\omega))$.*

PROOF. If $t + 1 \varepsilon T$ let j', j'' be two members of $f^{-1}(Y_{t+1}(\omega))$. If $t + 1 \not\varepsilon T$, let j' and j'' be any member of S. Then

$$\frac{P_\theta[X_{t+1} = j' \mid X_t = i, Y_{t_j}(\omega), t_j \varepsilon T]}{P_\theta[X_{t+1} = j'' \mid X_t = i, Y_{t_j}(\omega), t_j \varepsilon T]}$$

$$= \frac{P_\theta[X_{t+1} = j', X_t = i, Y_{t_j}(\omega), t_j \varepsilon T]}{P_\theta[X_{t+1} = j'', X_t = i, Y_{t_j}(\omega), t_j \varepsilon T]}$$

$$= \frac{P_\theta[X_{t+1} = j', Y_{t_j}(\omega), t_j \varepsilon T \quad \text{and} \quad t_j \geqq t + 1 \mid X_t = i]}{P_\theta[X_{t+1} = j'', Y_{t_j}(\omega), t_j \varepsilon T \quad \text{and} \quad t_j \geqq t + 1 \mid X_t = i]}$$

$$= \frac{\sum_{j_0} a_{ij'} a_{j'j_0} P_\theta[Y_{t_j}(\omega), t_j \varepsilon T, t_j \geqq t + 3 \mid X_{t+2} = j_0]}{\sum_{j_0} a_{ij''} a_{j''j_0} P_\theta[Y_{t_j}(\omega), t_j \varepsilon T, t_j \geqq t + 3 \mid X_{t+2} = j_0]}$$

(where j_0 ranges over $f^{-1}(Y_{t+2}(\omega))$ if $t + 2 \varepsilon T$ and over all S otherwise)

$$\leqq \max_{i,j',j'',j_0} (a_{ij'}a_{j'j_0}/a_{ij''}a_{j''j_0}) \leqq \delta^{-2}.$$

Therefore, if either $t + 1 \not\varepsilon T$, or $t + 1 \varepsilon T$ and $f(j) = Y_{t+1}(\omega)$, then

$$P_\theta[X_{t+1} = j \mid X_t = i, Y_{t_j}(\omega), t_j \varepsilon T] \geqq [1 + (s - 1)/\delta^2]^{-1} = \mu_\delta .$$

LEMMA 2.2. *$M^+[\theta, C_t, \{Y_{t_j}(\omega), t_j \varepsilon T\}, d] - M^-[\theta, C_t, \{Y_{t_j}(\omega), t_j \varepsilon T\}, d] \leqq \rho^{d-1}$ for some $\rho < 1$ independent of t, T, $Y(\omega)$, $\theta \varepsilon \Theta_\delta$ and C_t.*

PROOF. $P_\theta[C_t \mid X_{t-d-1} = i, Y_{t_j}(\omega), t_j \varepsilon T] = \sum_{j=1}^{s} P_\theta[C_t \mid X_{t-d} = j, Y_{t_j}(\omega), t_j \varepsilon T] P_\theta[X_{t-d} = j \mid X_{t-d-1} = i, Y_{t_j}(\omega), t_j \varepsilon T] \cdots . M_{d+1}^+ \leqq (1 - \mu)M_d^+ + \mu M_d^-$ using Lemma 2.1. Similarly $M_{d+1}^- \geqq (1 - \mu)M_d^- + \mu M_d^+$; thus $M_{d+1}^+ - M_{d+1}^- \leqq (1 - 2\mu)(M_d^+ - M_d^-)$. Since $\mu > 0$ we may take $\rho = 1 - 2\mu$.

COROLLARY 2.3. *$|P_\theta[C_t \mid Y_k(\omega), Y_{k-1}(\omega) \cdots Y_n(\omega)] - P_\theta[C_t \mid Y_k(\omega), Y_{k-1}(\omega) \cdots Y_{n+1}(\omega)]| < \rho^{t-n-1}$ for any k.*

PROOF. As in the proof of Lemma 1, $M_{t-n}^- = \min_{j\varepsilon f^{-1}(Y_n(\omega))} P_\theta[C_t \mid Y_k(\omega) \cdots Y_{n+1}(\omega), X_n = j] \leqq P_\theta[C_t \mid Y_k(\omega) \cdots Y_n(\omega)] \leqq \max_{j\varepsilon f^{-1}(Y_n(\omega))} P_\theta[C_t \mid Y_k(\omega) \cdots$

$Y_{n+1}(\omega), X_n = j] = M^{+}_{t-n}$. Since $P_\theta[C_t \mid Y_k(\omega) \cdots Y_{n+1}(\omega)]$ is an average of the $P_\theta[C_t \mid Y_k(\omega) \cdots Y_{n+1}(\omega), X_n = j]$ for $j \varepsilon f^{-1}(Y_n(\omega))$, the corollary follows from Lemma 2.

Corollary 2.4. $|P_\theta[C_r \mid Y_k(\omega), Y_{k-1}(\omega) \cdots Y_n(\omega)] - P_\theta[C_r \mid Y_{k-1}(\omega) \cdots Y_n(\omega)]| \leqq \rho^{r-k-1}$.

Corollary 2.5. $\lim_{s\to-\infty} P_\theta[C_r \mid Y_k(\omega), Y_{k-1}(\omega) \cdots Y_s(\omega)] = P_\theta[C_r \mid Y_k(\omega), Y_{k-1}(\omega) \cdots]$ *exists for all* Y *and is a continuous function of* θ. *Furthermore,*

$$|P_\theta[C_r \mid Y_k(\omega), Y_{k-1}(\omega) \cdots] - P_\theta[C_r \mid Y_{k-1}(\omega) \cdots]| \leqq \rho^{r-k-1}.$$

Proof. The first statement follows from Corollary 2.3, the second from Corollary 2.4.

Corollary 2.6. $|P_\theta[C_{t-1}, X_{t-d-1} = j, X_{t-d-2} = i \mid Y_{t_j}(\omega), t_j \varepsilon T] - P_\theta[C_{t-1} \mid Y_{t_j}(\omega), t_j \varepsilon T] P_\theta[X_{t-d-1} = j, X_{t-d-2} = i \mid Y_{t_j}(\omega), t_j \varepsilon T]| < \rho^{d-1}$.

Proof. The number in question is $\leqq | P_\theta[C_{t-1} \mid X_{t-d-1} = j, X_{t-d-2} = i, Y_{t_j}(\omega), t_j \varepsilon T] - P_\theta[C_{t-1} \mid Y_{t_j}(\omega), t_j \varepsilon T]|$ which is $< \rho^{d-1}$ by Lemma 2, since $P_\theta[C_{t-1} \mid X_{t-d-1} = j, X_{t-d-2} = i, Y_{t_j}(\omega), t_j \varepsilon T] = P_\theta[C_{t-1} \mid X_{t-d-1} = j, Y_{t_j}(\omega), t_j \varepsilon T]$ and $P_\theta[C_{t-1} \mid Y_{t_j}(\omega), t_j \varepsilon T]$ is an average of these latter quantities.

Complementary to the lemmas and corollaries we have proved is a set referring to cylinder sets D_t all of whose indices are less than or equal to t. These statements and proofs will be obvious to the reader. We will refer to such lemmas and corollaries by putting primes in their lemma or corollary number.

3. Consistency of maximum likelihood estimators.

In this section we let Θ denote the space $A \times B$ where A is the space of $s \times s$ stochastic matrices with positive entries and B is the space of $r \times s$ matrices $\{b_{jk} \mid \sum_k b_{jk} = 1, b_{jk} > 0\}$. We consider $\theta, \theta_0 \varepsilon \Theta_\delta = \{\theta \varepsilon \Theta \mid a_{ij}(\theta) \geqq \delta, b_{jk}(\theta) \geqq \delta, \delta > 0\}$.

Introduce the following random variables on R^∞. Let $Y = Y(\omega) \varepsilon R^\infty$.

(1) $f_k[\theta, Y] = P_\theta[Y_0 \mid Y_{-1}, Y_{-2}, \cdots, Y_{-k+1}]$; $f[\theta, Y] = \lim_{k\to\infty} f_k[\theta, Y]$ (see Corollary 2.5).

(2) $g_k[\theta, Y] = \log f_k[\theta, Y]$; $g[\theta, Y] = \lim_{k\to\infty} g_k[\theta, Y]$.

(3) $H[\theta] = E_{\theta_0}[g[\theta, \cdot]]$. Here expected value is taken with respect to P_{θ_0} measure on R^∞.

(4) $h_n[\theta, Y] = n^{-1} \log P_\theta[Y_1 \cdots Y_n]$.

(5) $\mathfrak{g}_{k,\epsilon}[\theta', Y] = \sup_{\theta\varepsilon S(\theta',\epsilon)} g_k[\theta, Y]$. $S(\theta', \epsilon)$ is an open sphere about θ' of radius ϵ. $\mathfrak{g}_{k,\epsilon}[\theta', Y]$ is measurable in Y because the sup over a dense countable subset of $S(\theta', \epsilon) = \sup S(\theta', \epsilon)$.

(6) $\mathfrak{g}_\epsilon[\theta', Y] = \lim_{k\to\infty} \mathfrak{g}_{k,\epsilon}[\theta', Y]$. This limit exists because $|g_k[\theta, Y] - g_{k+1}[\theta, Y]| \leqq C\rho^{k-1}$ for all $\theta \varepsilon \Theta_\delta$ implies the same inequality with $\mathfrak{g}_{k,\epsilon}$ replacing g_k. $\mathfrak{g}_\epsilon[\theta', Y]$ is measurable in Y.

(7) Let $K \subset \Theta_\delta$, $\delta > 0$ and K compact. $\theta^n[Y, K] = \{\hat\theta \varepsilon K \mid h_n(\hat\theta, Y) = \max_{\theta\varepsilon K} h_n(\theta, Y)\}$.

(8) $M[\theta_0, K] = \{\theta \varepsilon K \mid H(\theta) = H(\theta_0)\}$, $\theta_0 \varepsilon K$. $M[\theta_0, K, \epsilon] = \{\theta \varepsilon K \mid d(\theta, M[\theta_0, K]) < \epsilon\}$ where d is the Euclidean distance in Θ.

A necessary and sufficient condition for the existence of a consistent test to distinguish between θ and θ_0 is that $P_\theta \perp P_{\theta_0}$ on R^∞. A necessary and sufficient condition that $P_\theta \perp P_{\theta_0}$ in our case is that $H(\theta) < H(\theta_0)$ and this is the condition which we find convenient for proving the consistency of the maximum likelihood estimate.

THEOREM 3.1. *$H_\theta < H_{\theta_0}$. $H_\theta = H_{\theta_0} \Leftrightarrow \theta$ and θ_0 define equivalent Y processes.*

PROOF.

$$\begin{aligned} H_\theta - H_{\theta_0} &= \int \lg \{P_\theta[Y_0 \mid Y_{-1}, \cdots]/P_{\theta_0}[Y_0 \mid Y_{-1}, \cdots]\}\, dP_{\theta_0}[Y_0, Y_{-1}, \cdots] \\ &\leqq \lg \int \{P_\theta[Y_0 \mid Y_{-1}, \cdots]/P_{\theta_0}[Y_0 \mid Y_{-1}, \cdots]\}\, dP_{\theta_0}[Y_0, Y_{-1}, \cdots] \\ &= \lg \int [\textstyle\sum_{Y_0=1}^n \{P_\theta[Y_0 \mid Y_{-1}, \cdots]/P_{\theta_0}[Y_0 \mid Y_{-1}, \cdots]\} \\ &\qquad\qquad \cdot P_{\theta_0}[Y_0 \mid Y_{-1}, \cdots]]\, dP_{\theta_0}[Y_{-1}, \cdots] \\ &= \lg \int 1\, dP_{\theta_0}[Y_{-1}, \cdots] = 0 \end{aligned}$$

by Jensen's inequality. The inequality is strict unless $P_\theta[Y_0 \mid Y_{-1}, \cdots]/P_{\theta_0}[Y_0 \mid Y_{-1}, \cdots] = 1$ a.e. P_{θ_0}. By stationarity this would imply

$$P_\theta[Y_0, Y_{-1}, \cdots, Y_{-l} \mid Y_{-l-1}, \cdots]/P_{\theta_0}[Y_0, Y_{-1}, \cdots, Y_{-l} \mid Y_{-l-1}, \cdots] = 1 \quad \text{a.e. } P_{\theta_0},$$

and by summation over all values of the coordinates $Y_{-k-1}, \cdots, Y_{-l}$ that $P_\theta[Y_0, Y_{-1}, \cdots, Y_{-k} \mid Y_{-l-1}, \cdots]/P_{\theta_0}[Y_0, Y_{-1}, \cdots, Y_{-k} \mid Y_{-l-1}, \cdots] = 1$ a.e. θ_0. By Corollary 2.5 we conclude $P_\theta[Y_0, Y_{-1}, \cdots, Y_{-k}] = P_{\theta_0}[Y_0, Y_{-1}, \cdots, Y_{-k}]$ for all cylinder sets; i.e., θ and θ_0 define the same Y process.

By Theorem 3.1 and the heuristic discussion of Section 1 the surface $M[\theta_0, K] = \{\theta \varepsilon K : H_\theta = H_{\theta_0}\}$ will "in general" contain the single point θ_0.

THEOREM 3.2. $-n^{-1} \lg P_\theta[Y_1, \cdots, Y_n] \rightarrow_{\text{a.e.}} -H(\theta)$.

PROOF. $h_n(\theta, Y) = n^{-1} \lg P_\theta[Y_1, \cdots, Y_n] = n^{-1} \sum_{k=1}^n g_k[\theta, T^kY]$ where $(TY)_i = Y_{i+1}$.

$$|n^{-1} \textstyle\sum_{k=1}^n g_k[\theta, T^kY] - n^{-1} \sum_{k=1}^n g[\theta, T^kY]| \leqq n^{-1} \sum_{k=1}^n |g_k[\theta, T^kY] - g[\theta, T^kY]| \rightarrow 0$$

everywhere because $|g_k[\theta, T^kY] - g[\theta, T^kY]| < C\rho^{k-1}$ for every Y in R^∞. By the ergodic theorem

$$n^{-1} \textstyle\sum_{k=1}^n g[\theta, T^kY] \rightarrow_{\text{a.e.}} E_{\theta_0}[g[\theta, \text{-}]] = H(\theta).$$

THEOREM 3.3. $n^{-1} \sum_{k=1}^n \hat{g}_{k,\epsilon}[\theta, T^kY] \rightarrow_{\text{a.e.}} E_{\theta_0}[\hat{g}_\epsilon[\theta, \text{-}]]$.

THEOREM 3.4. *For almost all Y, $\theta_n[Y, K] \rightarrow M[\theta_0, K]$*; i.e., *for almost all Y for all $\epsilon > 0$ there exists an N_ϵ such that $n > N_\epsilon$ implies $\theta_n[Y, K] \subset M[\theta_0, K, \epsilon]$.*

PROOF. We show that for each θ' in the complement of $M[\theta_0, K, \epsilon]$, there exists a sphere $S(\theta', \lambda_{\theta'})$ of radius λ_θ about θ' such that $\hat{H}(\theta') = E_{\theta_0}(\hat{g}_{\lambda_{\theta'}}[\theta', \text{-}]) < H(\theta_0)$. In fact if $H(\theta') = H(\theta_0) - \mu$, $\mu > 0$, and $\rho^{n-1} < \mu/4$ then $|g_n[\theta, Y] - g[\theta, Y]| < C\rho^{n-1} < \mu/4$ for all $\theta \varepsilon \Theta_\delta$ and for all Y. $g_n[\theta, Y]$ depends only on n

coordinates of Y. For each choice of these n coordinates by the continuity in θ of $g_n[\theta, Y]$ we can choose a sphere about $\theta' \ni g_n[\theta, Y]$ varies by less than $\mu/4$ in this sphere. Choose $\lambda_{\theta'}$ as the smallest of the radii obtained for the finitely many choices of the n coordinates of Y. Then for all Y and $\theta \in S(\theta', \lambda_{\theta'})$, $|g[\theta, Y] - g[\theta', Y]| < |g[\theta, Y] - g_n[\theta, Y]| + |g_n[\theta, Y] - g_n[\theta', Y]| + |g_n[\theta', Y] - g[\theta', Y]| < \frac{3}{4}\mu$. Thus $\hat{g}_{\lambda_{\theta'}}[\theta', Y] \leqq g[\theta', Y] + \frac{3}{4}\mu$ and $\hat{H}(\theta') \leqq H(\theta') + \frac{3}{4}\mu < H(\theta_0)$.

Cover the complement of $M[\theta_0, K, \epsilon]$ which is compact, with finitely many spheres $S(\theta_i, \lambda_i)$. For each of these finitely many i's

$$\sup_{\theta \in S(\theta_i, \lambda_i)} h_n[\theta, Y] = \sup_{\theta \in S(\theta_i, \lambda_i)} n^{-1} \sum_{i=1}^{n} g_k[\theta, T^k Y]$$
$$\leqq n^{-1} \sum_{k=1}^{n} \hat{g}_{k,\lambda_i}[\theta_i, T^k Y] \rightarrow_{\text{a.e.}} \hat{H}(\theta_i)$$

by Theorem 3.3. If $\sup \hat{H}(\theta_i) = H(\theta_0) - \alpha$, $\alpha > 0$ then for almost every Y, $\max_{K-M[\theta_0,K,\epsilon]} h_n[\theta, Y] < H(\theta_0) - \alpha/2$ for all sufficiently large n while $\sup_{M[\theta_0,K,\epsilon]} h_n[\theta, Y] > H(\theta_0) - \alpha/2$ for all sufficiently large n; hence, $\theta^n[Y, K] \subset M[\theta_0, K, \epsilon]$ for all sufficiently large n.

4. Smoothness properties of $H(\theta)$. Here Θ denotes the space of $sr \times sr$ stochastic matrices with positive entries. The aim of this section is to show that the function $H(\theta)$ is differentiable with respect to the matrix coordinates $a_{ij}(\theta)$ of θ. $\Theta_\delta = \{\theta \in \Theta: a_{ij}(\theta) \geqq \delta\}$, $\delta > 0$.

Let $g_k^{(d)}(\theta, Y)$ denote any dth order partial derivative of $g_k[\theta, Y]$ with respect to the $\{a_{ij}(\theta)\}$

LEMMA 4.1. *For all $\theta \in \Theta_\delta$ and all Y $|g_n^{(d)}[\theta, Y] - g_{n-1}^{(d)}[\theta, Y]| < \alpha_d(n)$ for $d = 0$, 1, 2, 3 where $\sum_{n=1}^{\infty} \alpha_d(n) < \infty$.*

The proof is postponed until we show the consequences we want.

COROLLARY 4.2. $\lim_{n\to\infty} g_n^{(d)}[\theta, Y] = g^d[\theta, Y]$ *exists uniformly in θ for all Y.*

CORROLLARY 4.3. $H^{(d)}[\theta]$ *exists and* $\lim_{n\to\infty} h_n^{(d)}[\theta, Y] = H^{(d)}[\theta]$ *uniformly in* θ a.e. Y.

PROOF OF COROLLARY 4.3. This follows from Lemma 4.1 and Corollary 4.2 by the line of reasoning of the proof of Theorem 3.2.

PROOF OF LEMMA 4.1. Observe that

$$\begin{aligned}
&a_{ij}(\partial/\partial a_{ij}) \log P_\theta[Y_0, Y_{-1} \cdots Y_{-n+1}] \\
&= \textstyle\sum_{t=-n+2}^{0} P_\theta[X_t = j, X_{t-1} = i \mid Y_0 \cdots Y_{-n+1}], \\
&a_{k,l} a_{ij} (\partial^2/\partial a_{k,l} \partial a_{i,j}) \log P_\theta[Y_0 \cdots Y_{-n+1}] \\
&= \textstyle\sum_{t,t'=-n+2}^{0} P_\theta[X_t = j, X_{t-1} = i, X_{t'} = k, X_{t'-1} \\
&= l \mid Y_0, Y_{-1} \cdots Y_{-n+1}] - \textstyle\sum_{t,t'=-n+2}^{0} P_\theta[X_t = j, X_{t-1} \\
&= i \mid Y_0 \cdots Y_{-n+1}] \cdot P_\theta[X_{t'=k}, X_{t'-1} = l \mid Y_0 \cdots Y_{-n+1}].
\end{aligned} \tag{4.4}$$

A similar expression in terms of conditional probabilities holds for the third partial derivatives. We prove the lemma for $d = 2$. The method carries over to the other cases. Let

$$A^n_{t,t'} = P_\theta[X_t = j, X_{t-1} = i, X_{t'} = k, X_{t'-1} = l \mid Y_0, Y_{-1} \cdots Y_{-n+1}]$$

$$B^n_{t,t'} = P_\theta[X_t = j, X_{t-1} = i \mid Y_0, Y_{-1} \cdots Y_{-n+1}] \cdot P_\theta[X_{t'} = k, X_{t'-1} = l \mid Y_0, \cdots Y_{-n+1}]$$

$$C^n_{t,t'} = P_\theta[X_t = j, X_{t-1} = i, X_{t'} = k, X_{t'-1} = l \mid Y_{-1}, Y_{-2} \cdots Y_{-n+1}]$$

$$D^n_{t,t'} = P_\theta[X_t = j, X_{t-1} = i \mid Y_{-1} \cdots Y_{-n+1}] \cdot P_\theta[X_{t'} = k, X_{t'-1} = l \mid Y_{-1} \cdots Y_{-n+1}].$$

In terms of this notation we have

$$\begin{aligned} &a_{kl}a_{ij}[(\partial/\partial a_{kl})(\partial/\partial a_{ij})g_n(\theta, Y) - (\partial/\partial a_{kl})(\partial/\partial a_{ij})g_{n-1}(\theta, Y)] \\ (4.5)\qquad &= \textstyle\sum_{t,t'=-n+2}^{0} A^{(n)}_{t,t'} - \sum_{t,t'=-n+2}^{0} B^{(n)}_{t,t'} - \sum_{t,t'=-n+2}^{-1} C^{(n)}_{t,t'} \\ &+ \textstyle\sum_{t,t'=-n+2}^{-1} D^{(n)}_{t,t'} - \sum_{t,t'=-n+3}^{0} A^{(n-1)}_{t,t'} + \sum_{t,t'=-n+3}^{0} B^{(n-1)}_{t,t'} \\ &+ \textstyle\sum_{t,t'=-n+3}^{-1} C^{(n-1)}_{t,t'} - \sum_{t,t'=-n+3}^{-1} D^{(n-1)}_{t,t'} \end{aligned}$$

We decompose the square $-n + 2 \leqq t, t' \leqq 0$ into three disjoint regions:

$$R_1 = \{(t, t') \mid |t - t'| > [n/4]\},$$

$$R_2 = \{(t, t') \mid |t - t'| \leqq [n/4], |t'| < [n/2]\},$$

$$R_3 = \{(t, t') \mid |t - t'| \leqq [n/4], |t'| \geqq [n/2]\}.$$

The idea of the proof is to pair each positive summand of (4.5) with some negative summand such that their difference is of order $\rho^{[n/4]}$. E.g., on R_1, $|A^n_{t,t'} - B^n_{t,t'}| < \rho^{[n/4]}$; on R_2, $|A^n_{t,t'} - A^{n-1}_{t,t'}| < C\rho^{[n/4]}$ on R_3, $|A^n_{t,t'} - C^n_{t,t'}| < \rho^{[n/4]}$ by Corollaries 2.6, 2.3 and 2.3$'$. In this manner we see that the absolute value of (4.5) $\leqq$

$$\begin{aligned} &\textstyle\sum_{R_1} |A^n - B^n| + \sum_{R_2} |A^n - A^{n-1}| + \sum_{R_3} |A^n - C^n| \\ &\quad + \textstyle\sum_{R_1; t,t' \leqq -1} |C^n - D^n| + \sum_{R_2; t,t' \leqq -1} |C^n - C^{n-1}| \\ &\quad + \textstyle\sum_{R_3; t,t' \geqq -n+3} |C^{n-1} - A^{n-1}| + \sum_{R_1} |A^{n-1} - B^{n-1}| \\ &\quad + \textstyle\sum_{R_2} |B^n - B^{n-1}| + \sum_{R_3} |B^n - D^n| \\ &\quad + \textstyle\sum_{R_1; t,t' \leqq -1} |C^{n-1} - D^{n-1}| + \sum_{R_2; t,t' \leqq -1} |D^n - D^{n-1}| \\ &\qquad + \textstyle\sum_{R_3; t,t' \geqq -n+3} |D^{n-1} - B^{n-1}| \leqq \rho^{[n/4]} \cdot 4n^2. \end{aligned}$$

5. Parameterization and statistical applications of differentability of $H(\theta)$. Let K be a compact subset of some Euclidean space R^m and let τ be a continuous 1–1 map of K into the θ space used in Section 3. In previous sections we have used functions of the form $W[\theta, Y]$ where $\theta \varepsilon \Theta$. By a slight abuse of notation we write $W[\lambda, Y]$ for $W[\tau(\lambda), Y]$. With this in mind we let $M[\lambda_0] = \{\lambda \varepsilon K \mid H(\lambda) = H(\lambda_0)\}$ where $\theta_0 = \tau(\lambda_0)$, $M[\lambda_0, \epsilon] = \{\lambda \varepsilon K \mid d(\lambda, M[\lambda_0] < \epsilon\}$ where d is the Euclidean metric in K and $\Lambda_n[Y] = \{\lambda \varepsilon K \mid h_n(\lambda, Y) = \max_{\lambda' \varepsilon K} h_n[\lambda', Y]\}$.

According to Theorem 3.4, for almost every Y sequence and for every $\epsilon > 0$, $\Lambda_n[Y] \subset M[\lambda_0, \epsilon]$ for sufficiently large n.

Assume moreover that interior K is open in R^m, that $\lambda_0 \varepsilon$ interior K and $\tau \varepsilon C_1$. Define $M'[\lambda_0] = [\hat{\lambda} \varepsilon K \mid \mathrm{grad}_{\lambda=\hat{\lambda}} H[\lambda] = 0\}$. Then interior $K \cap M[\lambda_0] \subset M'[\lambda_0]$. Let $\Lambda_n'[Y] = \{\hat{\lambda} \varepsilon K \mid \mathrm{grad}_{\lambda=\hat{\lambda}} h_n[\lambda, Y] = 0\}$.

THEOREM 5.1. *For almost all* Y, $\Lambda_n'[Y] \to M'[\lambda_0]$.

PROOF. This is an easy consequence of Corollary 4.3. Theorem 5.1 gives a practical method for obtaining $M[\lambda_0]$. We hope to investigate the nature of $M[\lambda_0] \subset M'[\lambda_0]$ in a future paper.

Let us assume that $\tau \varepsilon C^3$ and introduce the additional local assumption that the matrix

$$\sigma_{\lambda_0} = \{\sigma_{uv}(\lambda_0)\} = (\partial^2/\partial\lambda_u\partial\lambda_v)\,|_{\lambda=\lambda_0} H(\lambda)$$

is nonsingular. We are then able to prove

THEOREM 5.2. *There exists a consistent solution of the maximum likelihood equations.*

PROOF. The proof follows the route used in Billingsley [1] p. 11 using Corollaries 4.1, 4.2 and 4.3 and replacing Billingsley's $g_u(x_k, x_{k+1}, \theta_0)$, $g_{uv}(x_k, x_{k+1}, \theta_0)$ and $g_{uvw}(x_k, x_{k+1}, \theta_0)$ by our ${}_ug_k[\lambda_0, T^kY]$, ${}_{uv}g_k[\lambda_0, T^kY]$, ${}_{uvw}[g_k\lambda_0, T^kY]$.

We will next prove a central limit theorem for the Y process which together with the line of reasoning in [1], pp. 13–23, allow us to conclude the useful statistical theorems for the Y process which are obtained in [1] for Markov processes.

THEOREM 5.3. *The random vector whose components are* $n^{-\frac{1}{2}}(\partial/\partial\lambda_u) \log P_\lambda[Y_1 \cdots Y_n]|_{\lambda=\lambda_0}$ *converges in law to* $\mathfrak{N}(0, \sigma(\lambda_0))$.

PROOF. We will apply the following theorem of [2].

THEOREM. Let $u_1, u_2, \cdots$ be random variables with moments of order 2 and let $\mathfrak{Z}_0, \mathfrak{Z}_1, \cdots$ be a non-decreasing sequence of Borel fields such that $E[u_n \mid \mathfrak{Z}_{n-1}] = 0$ with probability 1, $n = 1, 2, \cdots$. Suppose that $\lim_{n\to\infty} n^{-1}\sum_{k=1}^n E[u_k^2 \mid \mathfrak{Z}_{k-1}] = \beta^2$ with probability 1 where β^2 is a non-negative constant. Then $n^{-\frac{1}{2}} \sum_{k=1}^n u_k \to_{\mathfrak{L}} \mathfrak{N}(0, \beta^2)$.

In order to show that the random vector $n^{-\frac{1}{2}}\sum_{k=1}^n {}_ug_k(\lambda_0, T^kY)$ converges in law to $\mathfrak{N}(0, \sigma(\lambda_0))$ it suffices to show that for any set of $t_1, \cdots, t_m$ of real numbers the random scalar $n^{-\frac{1}{2}} \sum_{k=1}^n u_k \to_{\mathfrak{L}} \mathfrak{N}(0, \beta^2)$ where

$$u_k = \sum_{v=1}^m t_v g_k(\lambda_0, T^kY) \text{ and } \beta^2 = \sum_{u,v=1}^m t_u t_v \sigma_{uv}(\lambda_0)$$

by the standard Cramér-Wold result [3].

The cited theorem of [2] is applicable to the u_k and the Borel fields $\mathfrak{Z}_k$ generated by $Y_1, \cdots, Y_k$ as follows:

(i) $E[u_n \mid \mathfrak{Z}_{n-1}] = 0$ because

$$\sum_{v=1}^{m} t_v E[{}_v g_n[\lambda_0, T^{-1}Y] \mid \mathfrak{F}_{n-1}](Y)$$
$$= \sum_{v=1}^{m} t_v \sum_{Y_n}((\partial/\partial\lambda_v)P_\lambda[Y_n \mid Y_{n-1} \cdots Y_1]_{\lambda=\lambda_0}/P_{\lambda_0}[Y_n \mid Y_{n-1} \cdots Y_1])$$
$$\cdot P_{\lambda_0}[Y_n \mid Y_{n-1} \cdots Y_1]$$
$$= \sum_{v=1}^{m} t_v \sum_{Y_n}(\partial/\partial\lambda_v)P_\lambda[Y_n \mid Y_{n-1} \cdots Y_1]_{\lambda=\lambda_0} = 0$$

since $\sum_{Y_n} P_\lambda[Y_n \mid Y_{n-1} \cdots Y_1] \equiv 1$. Now we show that

(ii) $\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} E[u_k^2 \mid \mathfrak{F}_{k-1}] = \beta^2 = \sum t_u t_v \sigma_{uv}$.

Observe that $E[[{}_u g_k \mid \lambda, T^k \text{-}] \cdot {}_v g_k[\lambda_0, T^k \text{-}] \mid \mathfrak{F}_{k-1}] = E[{}_{uv} g_k[\lambda_0, T^k \text{-}] \mid \mathfrak{F}_{k-1}]$ since

$${}_{uv}g_k[\lambda_0, T^kY] = {}_{uv}P_{\lambda_0}[Y_k \mid Y_{k-1} \cdots Y_1]\{P_{\lambda_0}[Y_k \mid Y_{k-1} \cdots Y_1]\}^{-1}$$
$$-\{{}_uP_{\lambda_0}[Y_k \mid Y_{k-1} \cdots Y_1]\, {}_vP_{\lambda_0}[Y_k \mid Y_{k-1} \cdots Y_1]\}$$
$$\cdot\{P_{\lambda_0}[Y_k \mid Y_{k-1} \cdots Y_1]P_{\lambda_0}[Y_k \mid Y_{k-1} \cdots Y_1]\}^{-1}$$

and

$$\sum_{Y_k}({}_{uv}P_{\lambda_0}[Y_k \mid Y_{k-1} \cdots Y_1]/P_{\lambda_0}[Y_k \mid Y_{k-1} \cdots Y_1])P_{\lambda_0}[Y_k \mid Y_{k-1} \cdots Y_1] = 0.$$

Hence, we need to prove that

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} \sum_{v,u=1}^{m} t_u t_v E[{}_{uv}g_k[\lambda_0, T^k -] \mid \mathfrak{F}_{k-1}] = \beta^2. \tag{5.3}$$

(We have defined $f[\lambda, Y] = P_\lambda[Y_0 \mid Y_{-1}, Y_{-2} \cdots]$.) Define $G_{uv}[\lambda, Y] = \sum_{Y_0} f[\lambda, Y]\, {}_{uv}g[\lambda, Y]$. By the ergodic theorem

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} G_{uv}[\lambda_0, T^kY] = E_{\theta_0}[G_{uv}[\lambda_0, \text{-}]]$$
$$= \int \sum_{Y_0} {}_{uv}g[\lambda_0, Y]P_{\lambda_0}[Y_0 \mid Y_{-1} \cdots Y_{-\infty}]\, dP_{\lambda_0}[Y_{-1} \cdots Y_{-\infty}]$$
$$= \int {}_{uv}g[\lambda_0, Y]\, dP_{\lambda_0}[Y_0, Y_{-1} \cdots Y_{-\infty}] = E_{\theta_0}[{}_{uv}g[\lambda_0]].$$

Then:

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} E_{\theta_0}[{}_{uv}g_k[\lambda_0, T^k \text{-}] \mid \mathfrak{F}_{k-1}](Y)$$
$$= \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} \sum_{Y_k} {}_{uv}g_k[\lambda_0, T^kY]P[Y_k \mid Y_{k-1} \cdots Y_1]$$
$$= \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} \sum_{Y_k} {}_{uv}g_k[\lambda_0, T^kY]P[Y_k \mid Y_{k-1} \cdots Y_1, Y_0 \cdots Y_{-\infty}]$$

by Corollary 2.5,

$$= \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} G_{u,v}[\lambda_0, T^kY] = E_{\theta_0}[g_{u,v}\lambda_0, \text{-}]] = \sigma_{u,v}(\lambda_0).$$

We state the principal application of Theorem 5. Define $L_n(\lambda, Y) = nH_n(\lambda, Y) = \log P_\lambda[Y_1 \cdots Y_n]$. $y(n) = \{y_u(n)\} = \{n^{-\frac{1}{2}}(\partial/\partial\lambda_u) \log P_\lambda[Y_1 \cdots Y_n] \mid_{\lambda=\lambda_0}\}$. Let $l(n)$ be the random vector with components $l_u(n) = n^{\frac{1}{2}}(\hat{\lambda}_u^n - (\lambda_0)_u)$. (For large n $\hat{\lambda}^n$ is a single point if we assume no other λ defines the same Y process as λ_0.) If u_n and v_n are random vectors $u_n \sim v_n$ means $P \lim_n(u_n - v_n) = 0$.

THEOREM. $y(n) \sim \sigma(\lambda_0)l(n)$, $l(n) \sim \sigma^{-1}(\lambda_0)y(n)$, $y(n) \to_{\mathfrak{L}} \mathfrak{N}(0, \sigma(\lambda_0))$, $l(n) \to_{\mathfrak{L}} \mathfrak{N}(0, \sigma^{-1}(\lambda_0))$, $2(L_n(\hat{\lambda}^n, Y) - L_n(\lambda_0, Y)) \sim \langle\sigma(\lambda_0)l(n), l(n)\rangle \sim \langle y(n), \sigma^{-1}(\lambda_0)y(n)\rangle$ and $2[L_n(\hat{\lambda}^n, \text{-}) - L_n(\lambda_0, \text{-})] \to_{\mathfrak{L}} \chi_m^2$.

REFERENCES

[1] Billingsley, Patrick (1961). *Statistical Inference for Markov Processes*. Univ. of Chicago Press.

[2] Billingsley, Patrick (1961). The Lindeberg-Levy theorem for martingales. *Proc. Amer. Math. Soc.* **12** 788–792.

[3] Cramér, H. and Wold, H. (1936). Some theorems on distribution functions. J. London Math. Soc. **11** 290–294.

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

[5] Gilbert, E. J. (1959). On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.* **30** 688–697.

[6] Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory*. Dover, New York.