Stochastic Processes and their Applications 40 (1992) 127-143
North-Holland

Maximum-likelihood estimation for hidden Markov models

Brian G. Leroux

Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195, USA

Received 31 January 1990
Revised 17 December 1990

Hidden Markov models assume a sequence of random variables to be conditionally independent given a sequence of state variables which forms a Markov chain. Maximum-likelihood estimation for these models can be performed using the EM algorithm. In this paper the consistency of a sequence of maximum-likelihood estimators is proved. Also, the conclusion of the Shannon-McMillan-Breiman theorem on entropy convergence is established for hidden Markov models.

AMS 1980 Subject Classifications: Primary 62M09; Secondary 62F12.

Markov chain * consistency * subadditive ergodic theorem * identifiability * entropy * Kullback-Leibler divergence * Shannon-McMillan-Breiman theorem

1. Introduction

Hidden Markov models form a large and useful class of stochastic process models, in which series of counts, proportions, or multivariate observations are described with equal ease. These models are based on a Markov chain $\{X_i\}$ which describes the evolution of the state of a system. Given a realized sequence of state variables $\{x_i\}$, observed variables $\{Y_i\}$ are conditionally independent, with the distribution of each Y_i depending on the corresponding state x_i. In estimation problems the distribution of Y_i is assumed to belong to a parametric family and the state space is assumed finite. The special case of the hidden Markov model in which the observed variables have only finitely many values is referred to as a probabilistic function of a Markov chain; this model was introduced by Baum and Petrie (1966).

There is a clear analogy between hidden Markov models and state-space models, for example the linear state-space model:

$$X_i = FX_{i-1} + V_i,$$
$$Y_i = HX_i + W_i,$$

described by sequences of unobserved state variables $\{X_i\}$, observations $\{Y_i\}$, and noise variables $\{V_i\}$ and $\{W_i\}$. In many applications of state-space models, the goal is reconstruction of a value X_i based on an observation set $Y_1, \ldots, Y_n$, i.e., filtering if $i = n$, smoothing if $i < n$, or prediction if $i > n$. In the classical model with normal errors reconstruction is performed using the Kalman filter. The analysis of non-normal and non-linear state-space models has also been considered; for example,

Kitagawa (1987) gives recursive equations for filtering, smoothing, and prediction which are valid quite generally (see also Kohn and Ansley, 1987). The key elements seem to be a state process which is Markov and an observation sequence constructed from a conditionally independent sequence, given the state process. Thus we find overlap with hidden Markov models.

Reconstruction has also been a prime concern in the study of hidden Markov models. The forward-backward algorithm contained in the iterative likelihood-maximization algorithm of Baum et al. (1970) can be used for reconstruction of the underlying Markov chain. Also, versions of the smoothing and filtering equations of Kitagawa (1987) were derived in Askar and Derin (1981) and Lindgren (1978) for hidden Markov models.

Recent applications of hidden Markov models include those of Churchill (1989) to sequences of bases of a DNA molecule, Smith (1987) to the occurrence of rainfall, and Levinson, Rabiner, and Sondhi (1983) to the modelling of a speech generating source for automatic speech recognition.

Estimation of the parameters of a hidden Markov model has most often been performed using maximum-likelihood estimation. Baum and Eagon (1967) gave an algorithm (a special case of the EM algorithm; Dempster, Laird and Rubin, 1977) for locating a local maximum of the likelihood function for a probabilistic function of a Markov chain. Baum et al. (1970) developed the EM algorithm and applied it to general hidden Markov models. The large-sample behaviour of a sequence of maximum-likelihood estimators for a probabilistic function of a Markov chain was studied in Baum and Petrie (1966) and Petrie (1969). Lindgren (1978) proved a consistency property of maximum-likelihood estimators obtained for the model which assumes that $\{Y_i\}$ is an independent sequence from a finite mixture distribution. Lindgren's result states that, in case $\{Y_i\}$ actually follows a hidden Markov model, the maximum-likelihood estimators obtained under the independence model are consistent for the stationary distribution of $\{Y_i\}$.

In this paper the consistency of maximum-likelihood estimators is proved for general hidden Markov models. The next section displays the notation and required regularity conditions and establishes an ergodicity property. Section 3 examines the identifiability of hidden Markov models. The Shannon-McMillan-Breiman theorem on entropy convergence is proved for hidden Markov models in Section 4, and Section 5 contains a more general result which provides a generalization of Kullback-Leibler divergence. The consistency proof given in the final section follows the method of Wald (1949).

2. Notation and preliminary results

Let $\{X_i\}_{-\infty}^{\infty}$ be a stationary Markov chain with state space $\{1, \ldots, m\}$ and transition probability matrix $[\alpha_{jk}]$. Let $\{f(\cdot, \theta): \theta \in \Theta\}$ be a family of densities on a Euclidean space with respect to a measure μ, and $\theta_1, \ldots, \theta_m$ elements of Θ. $\{Y_i\}_{-\infty}^{\infty}$ is a

sequence of conditionally independent random variables, given a realization $\{x_i\}$ of $\{X_i\}$, with Y_i having conditional density $f(\cdot, \theta_{x_i})$.

The characteristics of the model are parameterized by ϕ which belongs to a parameter space Φ, a subset of a Euclidean space, i.e., we have $\alpha_{jk}(\phi), j, k = 1, \ldots, m$, and $\theta_j(\phi) \in \Theta$, $j = 1, \ldots, m$. The usual case is $\phi = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{mm}, \theta_1, \ldots, \theta_m)$, and $\alpha_{jk}(\cdot)$ and $\theta_j(\cdot)$ equal to coordinate projections. The true parameter value will be denoted ϕ_0.

The likelihood function for observations $y_1, \ldots, y_n$ is

$$p_n(y_1, \ldots, y_n; \phi) = \sum_{x_1} \cdots \sum_{x_n} \alpha_{x_1}^{(1)} f(y_1, \theta_{x_1}(\phi)) \prod_{i=2}^{n} \alpha_{x_{i-1}, x_i}(\phi) f(y_i, \theta_{x_i}(\phi)),$$

and a maximum-likelihood estimate is defined to be a point $\hat{\phi}_n$ at which p_n achieves its maximum value over Φ. The initial probability distribution used in the definition of likelihood is not necessarily the stationary probability distribution for the stochastic matrix $[\alpha_{jk}(\phi)]$, but any probability vector $\{\alpha_j^{(1)}\}$ with strictly positive elements. It turns out that consistency of maximum-likelihood estimators does not depend on the choice of (positive) $\alpha_j^{(1)}$.

The mild regularity conditions to be used are stated below for future reference.

Condition 1. The stochastic matrix $[\alpha_{jk}(\phi_0)]$ is irreducible.

Condition 2. The family of mixtures of at most m elements of $\{f(y, \theta): \theta \in \Theta\}$ is identifiable (see Remark 1 below).

Condition 3. For each $y, f(y, \cdot)$ is continuous and vanishes at infinity.

Condition 4. For each j, k, $\alpha_{jk}(\cdot)$ and $\theta_j(\cdot)$ are continuous.

Condition 5. $E_{\phi_0}[|\log f(Y_1, \theta_j(\phi_0))|] < \infty$, $j = 1, \ldots, m$.

Condition 6. For every $\theta \in \Theta$, $E_{\phi_0}[\sup_{\|\theta' - \theta\| < \delta} (\log f(Y_1, \theta'))^+] < \infty$, for some $\delta > 0$, ($\|\cdot\|$ is Euclidean distance and $x^+ = \max\{x, 0\}$).

Remark 1. Condition 2 means that a finite mixture with m or fewer components determines a unique mixing distribution, i.e.,

$$\sum_{j=1}^{m} \alpha_j f(y, \theta_j) = \sum_{j=1}^{m} \alpha_j' f(y, \theta_j') \text{ a.e. } d\mu(y) \Rightarrow \sum_{j=1}^{m} \alpha_j \delta_{\theta_j} = \sum_{j=1}^{m} \alpha_j' \delta_{\theta_j'}, \tag{1}$$

where δ_θ denotes the distribution function of a point mass at θ. Notice that the parameters $\alpha_1, \ldots, \alpha_m, \theta_1, \ldots, \theta_m$ are not themselves uniquely defined. Many families, including the Poisson, normal with fixed variance, and exponential, satisfy (1) for any m (and in fact the family of arbitrary mixtures is identifiable in these cases). The binomial family with fixed index k satisfies (1) if $k \geq 2m - 1$.

Remark 2. Condition 1 seems necessary to exclude the possibility that $\{X_i\}$ enters a transient state, in which case information on certain parameters would stop accumulating. By the following lemma, the stationarity of $\{X_i\}$ and condition 1 together imply that $\{Y_i\}$ is ergodic, an essential property for the limit theorems to be presented.

Lemma 1. *If $\{X_i\}$ is stationary and irreducible, then $\{Y_i\}_{-\infty}^{\infty}$ is ergodic.*

Proof. Let $\mathscr{A}$ be a shift invariant set of sequences $\{y_i\}_{-\infty}^{\infty}$ (possible realizations of $\{Y_i\}$). This means that $y \in \mathscr{A}$ if and only if $Ty \in \mathscr{A}$ for the shift operator T which shifts each element of a sequence back one position, i.e., $Ty = y'$ where $y_i' = y_{i+1}$. We must show that $P\{Y \in \mathscr{A}\}$ is either zero or one.

According to the Kolmogorov extension theorem, there is a subsequence $\{k'\}$ and cylinder sets $\mathscr{A}_{k'}$ depending on $(y_{-k'}, \ldots, y_{k'})$ such that, for every $k \geq 1$,

$$P\{Y \in \mathscr{A} \,\Delta\, \mathscr{A}_{k'}\} < 2^{-k}, \tag{2}$$

where Y stands for $\{Y_i\}_{-\infty}^{\infty}$ and Δ is the symmetric difference operator ($E \,\Delta\, A = (E \cap A^c) \cup (E^c \cap A)$). But, since Y is stationary and $\mathscr{A}$ is invariant,

$$\begin{aligned} P\{Y \in \mathscr{A} \,\Delta\, \mathscr{A}_{k'}\} &= P\{T^{2k'} Y \in \mathscr{A} \,\Delta\, \mathscr{A}_{k'}\} \\ &= P\{Y \in \mathscr{A} \,\Delta\, T^{-2k'} \mathscr{A}_{k'}\} \\ &= P\{Y \in \mathscr{A} \,\Delta\, \tilde{\mathscr{A}}_{k'}\}, \end{aligned} \tag{3}$$

where $\tilde{\mathscr{A}}_{k'} = T^{-2k'} \mathscr{A}_{k'} = \{y\colon T^{2k'} y \in \mathscr{A}_{k'}\}$ depends on $(y_{k'}, \ldots, y_{3k'})$. Now let $\tilde{\mathscr{A}} = \{\tilde{\mathscr{A}}_{k'}, \text{i.o.}\} = \bigcap_{k \geq 1} \bigcup_{j \geq k} \tilde{\mathscr{A}}_{j'}$. Then $\mathscr{A}^c \cap \tilde{\mathscr{A}} = \mathscr{A}^c \cap \{\tilde{\mathscr{A}}_{k'}, \text{i.o.}\} = \{\mathscr{A}^c \cap \tilde{\mathscr{A}}_{k'}, \text{i.o.}\}$ and $\mathscr{A} \cap \tilde{\mathscr{A}}^c \subset \mathscr{A} \cap \{\tilde{\mathscr{A}}_{k'}^c, \text{i.o.}\} = \{\mathscr{A} \cap \tilde{\mathscr{A}}_{k'}^c, \text{i.o.}\}$, so (2), (3), and the Borel–Cantelli lemma imply $P\{Y \in \mathscr{A} \,\Delta\, \tilde{\mathscr{A}}\} = 0$. Thus it suffices to show $P\{Y \in \tilde{\mathscr{A}}\}$ is either zero or one.

Now $\tilde{\mathscr{A}}$ is in the tail σ-field, i.e., for every k it depends only on $(y_k, y_{k+1}, \ldots)$. Since the Y_i are conditionally independent given a realization $x = \{x_i\}$ of the underlying Markov chain, the zero-one law implies that $P\{Y \in \tilde{\mathscr{A}} \,|\, x\}$ is either zero or one. Let $E = \{x\colon P\{Y \in \tilde{\mathscr{A}} \,|\, x\} = 1\}$, so $P\{Y \in \tilde{\mathscr{A}}\} = E[P\{Y \in \tilde{\mathscr{A}} \,|\, X\}] = P\{X \in E\}$. Now E is an invariant set, since

$$P\{Y \in \tilde{\mathscr{A}} \,|\, x\} = P\{TY \in \tilde{\mathscr{A}} \,|\, Tx\} = P\{Y \in \tilde{\mathscr{A}} \,|\, Tx\}$$

($\tilde{\mathscr{A}}$ is invariant). But a finite irreducible Markov chain is ergodic and therefore $P\{X \in E\}$ is either zero or one; this completes the proof. □

Before developing the required probabilistic tools for the proof of consistency, we compactify the parameter space Φ by adding to it limits of Cauchy sequences and denote the resulting space Φ^c (see Kiefer and Wolfowitz, 1956, where this device was first used in the context of maximum-likelihood estimation). To explain this concept, we explicitly describe the new parameter space in the case that

$$\Phi = \left\{(\alpha_{11}, \alpha_{12}, \ldots, \alpha_{mm}, \theta_1, \ldots, \theta_m)\colon \alpha_{jk} \geq 0, \sum_k \alpha_{jk} = 1, \theta_j \in \Theta\right\}.$$

Denote by Θ^c the one-point compactification of Θ, obtained by attaching to Θ a point denoted ∞, and extend $f(y, \cdot)$ to Θ^c by defining $f(y, \infty) = 0$ (for example, if $f(y, \cdot)$ is the Poisson density with mean θ, then $\Theta^c = [0, \infty]$). The compactified space Φ^c is then

$$\left\{(\alpha_{11}, \alpha_{12}, \ldots, \alpha_{mm}, \theta_1, \ldots, \theta_m): \alpha_{jk} \geq 0, \sum_k \alpha_{jk} = 1, \theta_j \in \Theta^c\right\}. \tag{4}$$

For the general parameterization, $(\alpha_{11}(\phi), \alpha_{12}(\phi), \ldots, \alpha_{mm}(\phi), \theta_1(\phi), \ldots, \theta_m(\phi))$ will still belong to the set in (4) for all parameter values after compactification. Condition 3 ensures that $f(y, \cdot)$ is continuous on all of Θ^c; also, the continuity of $\theta(\cdot)$ and $\alpha_{jk}(\cdot)$ extends to Φ^c.

3. Identifiability

The parameters of a hidden Markov model are not strictly identifiable. For instance, as with finite mixture distributions, the indices of the states of the Markov chain can be permuted without changing the law of the process $\{\theta_{X_i}\}$ and hence also the law of $\{Y_i\}$.

Define an equivalence relation $\sim$ on Φ^c, whereby $\phi_1 \sim \phi_2$ if and only if ϕ_1 and ϕ_2 define the same law for $\{\theta_{X_i}\}$. Let $\tilde{\phi}$ denote the equivalence class to which ϕ belongs. Notice that the law of $\{\theta_{X_i}\}$ is determined by the initial distribution of $\{X_i\}$ and there may be more than one initial distribution for which the process $\{X_i\}$ is stationary. To accommodate such parameters, we extend the definition of equivalence to allow somewhat arbitrary choices of initial distributions for X; more precisely, $\phi_1 \sim \phi_2$ if and only if there are initial probability distributions α^{ϕ_1} and α^{ϕ_2} such that the following holds:

(i) for $l = 1, 2, \{\theta_{X_i}(\phi_l)\}$ is a stationary process, where $\{X_i\}$ has transition probabilities $\alpha_{jk}(\phi_l)$ and initial distribution α^{ϕ_l};

(ii) the processes $\{\theta_{X_i}(\phi_1)\}$ and $\{\theta_{X_i}(\phi_2)\}$ have the same laws.

For example, all parameters ϕ_1 with $\theta(\phi_1) = (\lambda, \lambda)^T$ are in the same equivalence class, and also in this class is the parameter ϕ_2 with

$$[\alpha_{jk}(\phi_2)] = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad \theta(\phi_2) = \begin{pmatrix} 0 \\ \lambda \end{pmatrix}.$$

If $\theta_1(\phi), \ldots, \theta_m(\phi)$ are distinct and $[\alpha_{jk}(\phi)]$ is irreducible and aperiodic, so has a unique stationary distribution (Karlin and Taylor, 1975, Chapter 3), then $\tilde{\phi}$ only contains points obtained by permutations of the indices of the states of $\{X_i\}$; this corresponds to a finite mixture distribution with distinct support points and positive mixing proportions. Baum and Petrie (1966) and Petrie (1969) consider the identifiability question for probabilistic functions of a Markov chain.

The following lemma shows that the equivalence classes are identifiable, in the sense that two parameter values in different equivalence classes produce different stationary laws for the process $\{Y_i\}$. We will later establish the consistency of the equivalence class of the maximum-likelihood estimator.

Lemma 2. *If Condition 2 holds, then ϕ_1 and ϕ_2 define the same stationary law for the process $\{Y_i\}$ if and only if $\phi_1 \sim \phi_2$.*

Proof. If ϕ_1 and ϕ_2 define the same stationary law for the process $\{Y_i\}$, then, in particular, the joint distribution of Y_1 and Y_2 is the same under ϕ_1 and ϕ_2. Now these joint distributions have densities of the following form:

$$\sum_{j=1}^{m}\sum_{k=1}^{m} \alpha_j^{\phi}\alpha_{jk}(\phi) f(y_1, \theta_j(\phi)) f(y_2, \theta_k(\phi)),$$

namely, finite mixtures of products of two densities from the family $\{f(y, \theta): \theta \in \Theta\}$. We would like to conclude that ϕ_1 and ϕ_2 define the same mixing distribution, but Condition 2 states only the identifiability of mixtures from the family $\{f(y, \theta): \theta \in \Theta\}$ itself. However, Teicher (1967) showed how the identifiability of mixtures carries over to products of densities from a specific family; this result holds also for finite mixtures with a fixed number of components. Therefore we have that ϕ_1 and ϕ_2 define the same distribution for $(\theta_{X_1}, \theta_{X_2})$, and hence the same law for $\{\theta_{X_i}\}$. □

4. Entropy

In this section we define the entropy for stationary hidden Markov models and show that the conclusion of the Shannon–McMillan–Breiman Theorem, which concerns finite-state processes, also holds for the general hidden Markov model. This result is relatively simple to prove and anticipates the more general result of the next section, but none of the development in this section is necessary for anything in the sequel.

The entropy of the stationary process $\{Y_i\}$ under the parameter ϕ_0 is defined by the following expression:

$$H(\phi_0) = -E_{\phi_0}[\log p(Y_0 | Y_{-1}, Y_{-2}, \ldots; \phi_0)]. \tag{5}$$

In order for this definition to have meaning, the conditional density $p(Y_0 | Y_{-1}, Y_{-2}, \ldots; \phi_0)$ must be shown to exist. We will construct the conditional density by considering limits of the conditional densities which depend on a finite number of past values of the process, and then allow this number to grow arbitrarily large. The term entropy is used because the above definition of $H(\phi_0)$ is a generalization of the well known entropy for a random variable Y with density p, namely $E[-\log p(Y)]$.

In order to define the conditional density of Y_0 given the infinite past, consider the following representation for the conditional densities which depend on only a finite number of past observations:

$$p_i(y_0|y_{-1},\ldots,y_{-i+1};\phi_0)=\sum_{j=1}^{m} P_{\phi_0}\{X_0=j|y_{-1},\ldots,y_{-i+1}\}f(y_0,\theta_j(\phi_0)).$$

A classical martingale convergence result says that, if Z is an integrable random variable and $\{\mathcal{G}_i\}$ is an increasing sequence of σ-fields, then $\lim_{i\to\infty} E[Z|\mathcal{G}_i]=E[Z|\mathcal{G}_\infty]$, with probability one, where $\mathcal{G}_\infty$ is the σ-field generated by $\bigcup_i \mathcal{G}_i$. Applying this result with Z equal to the indicator of the event $\{X_0=j\}$ and $\mathcal{G}_i$ equal to the σ-field generated by $Y_{-1},\ldots,Y_{-i+1}$ gives

$$\lim_{i\to\infty} P_{\phi_0}\{X_0=j|Y_{-1},\ldots,Y_{-i+1}\}=P_{\phi_0}\{X_0=j|Y_{-1},Y_{-2},\ldots\}, \tag{6}$$

with probability one. Therefore we can define the conditional density depending on the infinite past by

$$p(Y_0|Y_{-1},Y_{-2},\ldots;\phi_0)=\sum_{j=1}^{m} P_{\phi_0}\{X_0=j|Y_{-1},Y_{-2},\ldots\}f(Y_0,\theta_j(\phi_0)), \tag{7}$$

and we have

$$\lim_{i\to\infty} p_i(Y_0|Y_{-1},\ldots,Y_{-i+1};\phi_0)=p(Y_0|Y_{-1},Y_{-2},\ldots;\phi_0), \tag{8}$$

with probability one.

Theorem 1. *If Conditions* 1 *and* 5 *hold, then*

$$H(\phi_0)=E_{\phi_0}[-\log p(Y_0|Y_{-1},Y_{-2},\ldots;\phi_0)]$$

is finite and

(i) $\lim_n n^{-1}E_{\phi_0}[\log p_n(Y_1,\ldots,Y_n;\phi_0)]=-H(\phi_0)$;

(ii) $\lim_n n^{-1}\log p_n(Y_1,\ldots,Y_n;\phi_0)=-H(\phi_0)$, *with probability one, under* ϕ_0.

Proof. (i) By Condition 5, $\{\log p_i(Y_0|Y_{-1},\ldots,Y_{-i+1};\phi_0)\}$ is a uniformly integrable sequence of random variables, since

$$\min_j f(Y_0,\theta_j(\phi_0))\le p_i(Y_0|Y_{-1},\ldots,Y_{-i+1};\phi_0)\le\max_j f(Y_0,\theta_j(\phi_0)),$$

and hence (8) implies

$$H(\phi_0)=\lim_i E_{\phi_0}[-\log p_i(Y_0|Y_{-1},\ldots,Y_{-i+1};\phi_0)];$$

therefore, $H(\phi_0)$ is finite. Using

$$p_n(Y_1,\ldots,Y_n;\phi_0)=\prod_{i=1}^{n} p_i(Y_i|Y_{i-1},\ldots,Y_1;\phi_0)$$

we conclude that

$$\begin{aligned}\frac{1}{n}E_{\phi_0}[\log p_n(Y_1,\ldots,Y_n;\phi_0)] &= \frac{1}{n}\sum_{i=1}^{n} E_{\phi_0}[\log p_i(Y_i|Y_{i-1},\ldots,Y_1;\phi_0)]\\ &= \frac{1}{n}\sum_{i=1}^{n} E_{\phi_0}[\log p_i(Y_0|Y_{-1},\ldots,Y_{-i+1};\phi_0)]\\ &\to -H(\phi_0).\end{aligned}$$

(ii) The ergodic theorem implies

$$\lim\frac{1}{n}\sum_{i=1}^{n}\log p_i(Y_i|Y_{i-1},Y_{i-2},\ldots;\phi_0) = E_{\phi_0}[\log p_i(Y_0|Y_{-1},Y_{-2},\ldots;\phi_0)],$$

with probability one, under P_{ϕ_0}, and (ii) follows from approximating $\log p_n(Y_1,\ldots,Y_n;\phi_0)/n$ by

$$\frac{1}{n}\sum_{i=1}^{n}\log p(Y_i|Y_{i-1},Y_{i-2},\ldots;\phi_0)$$

as in Karlin and Taylor (1975, pp. 498–502). □

5. Generalized Kullback–Leibler divergence

Here we prove a limit theorem for the log-likelihood function, similar to Theorem 1, with the important difference that the limit of the log-likelihood at points other than ϕ_0 is identified. This limit leads to a definition of generalized Kullback–Leibler divergence; the remainder of this section is devoted to proving that this divergence distinguishes parameter points in different equivalence classes.

Theorem 2. *Assume Conditions* 1, 3 *and* 6 *hold. Then, for* $\phi\in\Phi^c$, *there is a constant* $H(\phi_0,\phi)<\infty$ (*possibly equal to* $-\infty$), *such that*

(i) $\lim_n n^{-1}E_{\phi_0}[\log p_n(Y_1,\ldots,Y_n;\phi)] = H(\phi_0,\phi)$;

(ii) $\lim_n n^{-1}\log p_n(Y_1,\ldots,Y_n;\phi) = H(\phi_0,\phi)$, *with probability one, under* ϕ_0.

These conclusions hold for any choice of positive initial probabilities $\alpha_j^{(1)}$, *and* $H(\phi_0,\phi)$ *has the same value for any choice.*

Proof. Fix the value of $\phi\in\Phi$; where no other indication is given, the parameters α_{jk} and θ_j (and joint densities defined using them) will be assumed to be evaluated at this point. Define

$$p_n(y_1,\ldots,y_n|j) = f(y_1,\theta_j)\sum_{x_2}\cdots\sum_{x_n}\alpha_{j,x_2}f(y_2,\theta_{x_2})\prod_{i=3}^{n}\alpha_{x_{i-1},x_i}f(y_i,\theta_{x_i}) \qquad (9)$$

(with $p_1(y_1|j)=f(y_1,\theta_j)$) and

$$q_n(y_1,\ldots,y_n) = \max_j p_n(y_1,\ldots,y_n|j).$$

Then the likelihood satisfies $p_n(y_1, \ldots, y_n) \leq q_n(y_1, \ldots, y_n)$ and

$$p_n(y_1, \ldots, y_n) = \sum_j \alpha_j^{(1)} p_n(y_1, \ldots, y_n | j)$$

$$\geq q_n(y_1, \ldots, y_n) \min_j \alpha_j^{(1)};$$

hence

$$\log\left(\min_j \alpha_j^{(1)}\right) \leq \log \frac{p_n(y_1, \ldots, y_n)}{q_n(y_1, \ldots, y_n)} \leq 0. \tag{10}$$

Therefore $\log p_n(Y_1, \ldots, Y_n)/n$ and $E_{\phi_0}[\log p_n(Y_1, \ldots, Y_n)]/n$ have the same limiting values as $\log q_n(Y_1, \ldots, Y_n)/n$ and $E_{\phi_0}[\log q_n(Y_1, \ldots, Y_n)]/n$, respectively, and so the conclusions of the theorem will follow from the corresponding conclusions applied to q_n. Notice that q_n does not depend on the initial probabilities, provided they are positive, so that the limit of the log-likelihood is valid for any choice. The advantage in working with q_n rather than p_n is provided by the property given in the following lemma.

Lemma 3. *For any sequence* $\{y_n\}$,

$$q_{s+t}(y_1, \ldots, y_{s+t}) \leq q_s(y_1, \ldots, y_s) q_t(y_{s+1}, \ldots, y_{s+t}), \quad s, t \geq 1.$$

Proof. By definition,

$$\begin{aligned}
&p_{s+t}(y_1, \ldots, y_{s+t} | j) \\
&= f(y_1, \theta_j) \sum_{x_2} \cdots \sum_{x_s} \alpha_{j,x_2} f(y_2, \theta_{x_2}) \prod_3^s \alpha_{x_{i-1},x_i} f(y_i, \theta_{x_i}) \\
&\quad \times \sum_k \sum_{x_{s+2}} \cdots \sum_{x_{s+t}} \alpha_{x_s,k} f(y_{s+1}, \theta_k) \alpha_{k,x_{s+2}} f(y_{s+2}, \theta_{x_{s+2}}) \prod_{s+3}^{s+t} \alpha_{x_{i-1},x_i} f(y_i, \theta_{x_i}) \\
&= f(y_1, \theta_j) \sum_{x_2} \cdots \sum_{x_s} \alpha_{j,x_2} f(y_2, \theta_{x_2}) \prod_3^s \alpha_{x_{i-1},x_i} f(y_i, \theta_{x_i}) \\
&\quad \times \sum_k \alpha_{x_s,k} p_t(y_{s+1}, \ldots, y_{s+t} | k) \\
&\leq f(y_1, \theta_j) \sum_{x_2} \cdots \sum_{x_s} \alpha_{j,x_2} f(y_2, \theta_{x_2}) \prod_3^s \alpha_{x_{i-1},x_i} f(y_i, \theta_{x_i}) q_t(y_{s+1}, \ldots, y_{s+t}) \\
&\leq q_s(y_1, \ldots, y_s) q_t(y_{s+1}, \ldots, y_{s+t}). \qquad \square
\end{aligned}$$

Proof of Theorem 2 (continued). Now define the doubly indexed sequence of random variables $\{W_{st}\}$ by $W_{st} = \log(q_{t-s}(Y_{s+1}, \ldots, Y_t))$, $s < t$. The above lemma says

$$W_{st} \leq W_{su} + W_{ut}, \quad s < u < t. \tag{11}$$

Ergodic theorems for processes satisfying this subadditivity property are given in Kingman (1976), so we consider next the other properties which were used to obtain these theorems. By the stationarity of $\{Y_n\}$,

$$\{W_{st}\} \text{ is stationary relative to the shift transformation } W_{st} \to W_{s+1,t+1}; \tag{12}$$

for example, W_{st} and $W_{s+1,t+1}$ have the same distribution. Also, the integrability condition

$$E_{\phi_0}[W_{01}^+] < \infty \tag{13}$$

is satisfied under condition 6, since $\log q_1(y_1) \leq \log(\max_j f(y_1, \theta_j))$.

Kingman (1976, Theorems 1.5 and 1.8) proved that a process $\{W_{st}\}$ satisfying (11), (12), and (13) also satisfies the conclusions of the ergodic theorem, namely, (i) $\lim_n W_{0n}/n = W < \infty$ exists with probability one; (ii) $E[W] = \lim_n E[W_{0n}/n]$; and (iii) W is degenerate if the process is ergodic, i.e., the σ-field of events invariant under the shift transformation in (12) is trivial. (These results generalize the classical ergodic theorem, which deals with additive rather than subadditive processes.) An application to $W_{0n} = \log q_n(Y_1, \ldots, Y_n)$ gives (the ergodicity carries over from the ergodicity of $\{Y_n\}$)

$$\lim_n \frac{1}{n} E_{\phi_0}[\log q_n(Y_1, \ldots, Y_n)] = H(\phi_0, \phi) < \infty$$

exists and

$$\lim_n \frac{1}{n} \log q_n(Y_1, \ldots, Y_n) = H(\phi_0, \phi), \tag{14}$$

with probability one, under ϕ_0. As demonstrated above, $\log p_n/n$ and $\log q_n/n$ have the same limiting behaviour; thus the proof of the theorem is complete. □

The divergence between ϕ_0 and ϕ is now defined as $K(\phi_0; \phi) = H(\phi_0, \phi_0) - H(\phi_0, \phi)$, where $H(\phi_0, \phi_0)$ and $H(\phi_0, \phi)$ are defined in Theorem 2 ($H(\phi_0, \phi_0)$ is the negative entropy, where the entropy $H(\phi_0)$ is defined in Section 4). The function K provides a measure of distance between parameter points; the definition of $H(\phi_0, \phi)$ in Theorem 2 shows that $K(\phi_0, \phi)$ is the large-sample average Kullback-Leibler divergence per observation between $p_n(y_1, \ldots, y_n; \phi_0)$ and $p_n(y_1, \ldots, y_n; \phi)$. Juang and Rabiner (1985) use this measure of distance between hidden Markov models in a numerical study of the effects of starting values and observation sequence length on maximum-likelihood estimates.

Next we prove a result needed for the large sample analysis of maximum-likelihood estimators, namely that the divergence between two different points is positive. Obtaining this result is surprisingly difficult and will lead to another study of the asymptotic behaviour of the log-likelihood. Kingman's subadditive ergodic theorem which was used above does not include a representation of the limit as the expected

value of some random variable, as does the classical ergodic theorem. We will directly establish the convergence of the normalized log-likelihood and, using the previous results to identify the limit random variable with the constant $H(\phi_0, \phi)$, obtain such a representation for $H(\phi_0, \phi)$.

As in Section 4, we will study the log-likelihood using the relation

$$\log p_n(y_1, \ldots, y_n; \phi) = \sum_{i=1}^{n} \log p_i(y_i | y_{i-1}, \ldots, y_1; \phi).$$

However, instead of approximating $p_i(Y_i | Y_{i-1}, \ldots, Y_1; \phi)$ by a stationary process, we define a new probability measure (on an augmented probability space), under which $\{p_i(Y_i | Y_{i-1}, \ldots, Y_1; \phi)\}$ is itself stationary. The quantities derived under this new probability space will then be related back to quantities defined in terms of the original probability space. The motivation for using this approach came from Furstenburg and Kesten (1960), who studied the convergence of products of random matrices and also from Petrie (1969) who used results from the latter study to obtain the convergence of the log-likelihood for a probabilistic function of a Markov chain. There is a connection with Kingman's theorems, namely that Kingman applied his results to obtain those of Furstenburg and Kesten (1960); on the other hand, the limit results for $\{q_n\}$ obtained using Kingman's theorems could be proved using arguments similar to those of Furstenburg and Kesten (1960).

The approach to be followed requires a careful accounting of the probability spaces and measures involved. We begin with the probability measure P_{ϕ_0} defined on the measure space $(\mathcal{Y}, \mathcal{B})$, i.e., the set $\mathcal{Y}$ of sequences $\{y_i\}$ augmented by its Borel σ-field. Let Ω be the set of sequences $\{(y_n, u^{(n)})\}$, where the $u^{(n)}$ are m-dimensional vectors. Let $P'_{\phi_0,\phi}$ be the probability measure on Ω defined as the image of P_{ϕ_0} on the subset where $u_j^{(1)} = \alpha_j(\phi_0)$, the stationary probabilities of the stochastic matrix $[\alpha_{jk}(\phi_0)]$, and

$$u_k^{(n+1)} = \frac{\sum_j u_j^{(n)} f(y_n, \theta_j)\alpha_{jk}}{\sum_j u_j^{(n)} f(y_n, \theta_j)}, \quad k = 1, \ldots, m, \ n = 1, 2, \ldots, \tag{15}$$

(0/0 is taken to be 0); $\{u^{(n)}\}$ is determined by $\{y_n\}$ on this subset, so this definition is meaningful. (Notice that $P'_{\phi_0,\phi}$ depends on ϕ through its support, which is determined by (15).) Let Y_n and $U^{(n)}$ be the coordinate mappings on Ω.

The goal is to define a probability measure on Ω, under which $\{U^{(n)}\}$ is a stationary sequence, while $\{Y_n\}$ has the same distribution as it does under P_{ϕ_0}. Let T_Ω be the shift transformation on Ω, i.e., $T_\Omega\{(y_n, u^{(n)})\} = \{(y_{n+1}, u^{(n+1)})\}$. Let $P'_{\phi_0,\phi} T_\Omega^{-k}$ be the probability measure on Ω which is the inverse image of $P'_{\phi_0,\phi}$ under the kth iterate of T_Ω, i.e.,

$$P'_{\phi_0,\phi} T_\Omega^{-k}(A) = P'_{\phi_0,\phi}\{\omega \in \Omega; \ T_\Omega^k \omega \in A\}, \quad A \in \mathcal{B}_\Omega,$$

($\mathcal{B}_\Omega$ is the Borel σ-field of Ω). Define new probability measures $\tilde{P}^{(l)}_{\phi_0,\phi} = \sum_{i=0}^{l-1} P'_{\phi_0,\phi} T_\Omega^{-i}/l$, for $l = 1, 2, \ldots$. The following lemma is essentially proved in Furstenburg and Kesten (1960).

Lemma 4. *There is a subsequence* $\{l_k\}$ *and a probability measure* $\tilde{P}_{\phi_0,\phi}$ *such that*

(i) $\tilde{P}^{(l_k)}_{\phi_0,\phi}$ *converges weakly to* $\tilde{P}_{\phi_0,\phi}$ *(in particular, for every p, the joint distribution of* $(Y_1, U^{(1)}), \ldots, (Y_p, U^{(p)})$ *under* $\tilde{P}^{(l_k)}_{\phi_0,\phi}$ *converges weakly to the corresponding joint distribution under* $\tilde{P}_{\phi_0,\phi}$*);*

(ii) $\{(Y_n, U^{(n)})\}$ *is a stationary process under* $\tilde{P}_{\phi_0,\phi}$*; and*

(iii) $\{Y_n\}$ *has the same distribution under* $\tilde{P}_{\phi_0,\phi}$ *as under* P_{ϕ_0}. □

For the case $\phi = \phi_0$, the meaning of the random vector $U^{(1)}$ under $\tilde{P}_{\phi_0,\phi_0}$ will now be explained. The recursion relations (15) for $u^{(n)}$ and the initial condition $u_j^{(1)} = \alpha_j(\phi_0)$ give $U_j^{(i)} = P_{\phi_0}\{X_i = j \mid Y_{i-1}, \ldots, Y_1\}$ under P'_{ϕ_0,ϕ_0}; hence, the operation of shifting the time scale and taking the limit to obtain $\tilde{P}_{\phi_0,\phi_0}$ has the effect of converting $U_j^{(1)}$ into a conditional probability depending on infinitely many past values of $\{Y_i\}$. More precisely, $U_j^{(1)}$ represents $P_{\phi_0}\{X_1 = j \mid Y_0, Y_{-1}, \ldots\}$ in the sense (to be proved in Lemma 6 below) that the conditional density of $Y_1, \ldots, Y_i$ given $U^{(1)}$, under $\tilde{P}_{\phi_0,\phi_0}$, is $\sum_j U_j^{(1)} p_i(y_1, \ldots, y_i \mid j; \phi_0)$. Therefore, the entropy $H(\phi_0)$, defined in Section 4 by $H(\phi_0) = E_{\phi_0}[-\log\{\sum_j P_{\phi_0}\{X_1 = j \mid Y_0, Y_{-1}, \ldots\} f(Y_1, \theta_j)\}]$, is seen to be equal to $\tilde{E}_{\phi_0,\phi_0}[-\log\{\sum_j U_j^{(1)} f(Y_1, \theta_j)\}]$, with the consequence that this representation can be extended to parameters other than ϕ_0, as in the following result.

Lemma 5. *Assume Conditions* 1, 3 *and* 6 *hold. Then, for every* $\phi \in \Phi^c$,

$$H(\phi_0, \phi) = \tilde{E}_{\phi_0,\phi}[\log\{\textstyle\sum_j U_j^{(1)} f(Y_1, \theta_j(\phi))\}].$$

Proof. The ergodic theorem implies

$$\tilde{P}_{\phi_0,\phi}\left\{\lim_n \frac{1}{n} \sum_{i=1}^{n} \log\left\{\sum_j U_j^{(i)} f(Y_i, \theta_j(\phi))\right\} = Z\right\} = 1,$$

where Z is a random variable with $\tilde{E}_{\phi_0,\phi}[Z] = \tilde{E}_{\phi_0,\phi}[\log\{\sum_j U_j^{(1)} f(Y_1, \theta_j(\phi))\}]$.

First we show $\tilde{E}_{\phi_0,\phi}[Z] \leq H(\phi_0, \phi)$. The recursion relations (15) imply

$$\sum_k U_k^{(2)} f(Y_2, \theta_k(\phi)) = \sum_j \sum_k U_j^{(1)} a_{jk} f(Y_1, \theta_j(\phi)) f(Y_2, \theta_k(\phi)) \Big/ \sum_j U_j^{(1)} f(Y_1, \theta_j(\phi))$$

$$= \sum_j U_j^{(1)} p_2(Y_1, Y_2 \mid j; \phi) \Big/ \sum_j U_j^{(1)} p_1(Y_1 \mid j; \phi) \tag{16}$$

(see (9)) and iterating gives

$$\sum_j U_j^{(i)} f(Y_i, \theta_j(\phi))$$

$$= \sum_j U_j^{(1)} p_i(Y_1, \ldots, Y_i \mid j; \phi) \Big/ \sum_j U_j^{(1)} p_{i-1}(Y_1, \ldots, Y_{i-1} \mid j; \phi). \tag{17}$$

Therefore,

$$\sum_{i=1}^{n} \log\left\{\sum_j U_j^{(i)} f(Y_i, \theta_j(\phi))\right\} = \log\left\{\sum_j U_j^{(1)} p_n(Y_1, \ldots, Y_n \mid j; \phi)\right\}$$

$$\leq \log q_n(Y_1, \ldots, Y_n; \phi),$$

and, since $\{Y_i\}$ has the same distribution under P_{ϕ_0} as under $\tilde{P}_{\phi_0,\phi}$, the proof of Theorem 2 gives $\tilde{E}_{\phi_0,\phi}[Z] \leq H(\phi_0, \phi)$.

Next we show $\tilde{E}_{\phi_0,\phi}[Z] \geq H(\phi_0, \phi)$. Assume, without loss of generality, $H(\phi_0, \phi) > -\infty$. Using the fact that the joint distribution of $(Y_1, U^{(1)})$ under $\tilde{P}^{(l_k)}_{\phi_0,\phi}$ converges weakly to the corresponding distribution under $\tilde{P}_{\phi_0,\phi}$, we get

$$\limsup_k \int_A \log\left\{\sum_j U_j^{(1)} f(Y_1, \theta_j(\phi))\right\} d\tilde{P}^{(l_k)}_{\phi_0,\phi}$$

$$\leq \int_A \log\left\{\sum_j U_j^{(1)} f(Y_1, \theta_j(\phi))\right\} d\tilde{P}_{\phi_0,\phi},$$

where $A = \{\log\{\sum_j U_j^{(1)} f(Y_1, \theta_j(\phi))\} \leq 0\}$. Also, since $(\log\{\sum_j U_j^{(1)} f(Y_1, \theta_j(\phi))\})^+$ is uniformly integrable with respect to $\tilde{P}^{(l_k)}_{\phi_0,\phi}$ by condition 6,

$$\lim_k \int_{\Omega\setminus A} \log\left\{\sum_j U_j^{(1)} f(Y_1, \theta_j(\phi))\right\} d\tilde{P}^{(l_k)}_{\phi_0,\phi}$$

$$= \int_{\Omega\setminus A} \log\left\{\sum_j U_j^{(1)} f(Y_1, \theta_j(\phi))\right\} d\tilde{P}_{\phi_0,\phi}.$$

Therefore,

$$\begin{aligned}
\tilde{E}_{\phi_0,\phi}[Z] &= \tilde{E}_{\phi_0,\phi}\left[\log\left\{\sum_j U_j^{(1)} f(Y_1, \theta_j(\phi))\right\}\right] \\
&\geq \limsup_k \tilde{E}^{(l_k)}_{\phi_0,\phi}\left[\log\left\{\sum_j U_j^{(1)} f(Y_1, \theta_j(\phi))\right\}\right] \\
&= \limsup_k \frac{1}{l_k} \sum_{i=1}^{l_k} E'_{\phi_0,\phi}\left[\log\left\{\sum_j U_j^{(i)} f(Y_i, \theta_j(\phi))\right\}\right] \\
&= \limsup_k \frac{1}{l_k} E'_{\phi_0,\phi}\left[\sum_{i=1}^{l_k} \log\left\{\sum_j U_j^{(i)} f(Y_i, \theta_j(\phi))\right\}\right] \\
&= \limsup_k \frac{1}{l_k} E_{\phi_0}\left[\log\left\{\sum_j \alpha_j(\phi_0) p_{l_k}(Y_1, \ldots, Y_{l_k} | j; \phi)\right\}\right] \\
&= H(\phi_0, \phi);
\end{aligned}$$

the second last equality follows from (17) and $U_j^{(1)} = \alpha_j(\phi_0)$ on the support of $P'_{\phi_0,\phi}$, and the last from Theorem 2 (using $\alpha_j(\phi_0) > 0$, which follows from the irreducibility of $[\alpha_{jk}(\phi_0)]$). □

The representation in the lemma is used next to prove that the divergence between two different parameter points is positive.

Lemma 6. *Assume Conditions* 1-3, 5 *and* 6 *hold. For every* $\phi \in \Phi^c$, $K(\phi_0, \phi) \geq 0$. *If* $\phi \neq \phi_0$, *then* $K(\phi_0, \phi) > 0$.

Proof. The first step is a verification of the property (described following Lemma 4) of the joint distribution of $Y_1, \ldots, Y_i, U^{(1)}$ under $\tilde{P}_{\phi_0,\phi_0}$, namely that $(Y_1, \ldots, Y_i)$ has the conditional density $\sum_j U_j^{(1)} p_i(y_1, \ldots, y_i | j; \phi_0)$, given $U^{(1)}$. The case $i=2$ will be considered; the general case is verified similarly. Let Q be the distribution of $U^{(1)}$ under $\tilde{P}_{\phi_0,\phi_0}$; then, if B is a continuity set of Q,

$$\begin{aligned}
&\tilde{P}_{\phi_0,\phi_0}\{(Y_1, Y_2) \in A,\ U^{(1)} \in B\} \\
&= \lim_k \frac{1}{l_k} \sum_{i=1}^{l_k} P'_{\phi_0,\phi_0}\{(Y_i, Y_{i+1}) \in A,\ U^{(i)} \in B\} \\
&= \lim_k \frac{1}{l_k} \sum_{i=1}^{l_k} \int_B \int_A \sum_j u_j^{(i)} p_2(y_i, y_{i+1} | j; \phi_0)\, d\mu(y_i)\, d\mu(y_{i+1})\, dQ_i(u^{(i)}) \\
&= \lim_k \int_B \int_A \sum_j u_j p_2(y_1, y_2 | j; \phi_0)\, d\mu(y_1)\, d\mu(y_2)\, dQ^{(l_k)}(u) \\
&= \int_B \int_A \sum_j u_j p_2(y_1, y_2 | j; \phi_0)\, d\mu(y_1)\, d\mu(y_2)\, dQ(u),
\end{aligned}$$

where Q_i and $Q^{(l)}$ are the distributions of $U^{(i)}$ under P'_{ϕ_0,ϕ_0} and $\sum_{i=1}^{l} P'_{\phi_0,\phi_0} T_\Omega^{-i}/l$, respectively; the second equality follows from $U_j^{(i)} = P_{\phi_0}\{X_i = j \mid Y_{i-1}, \ldots, Y_1\}$ under P'_{ϕ_0,ϕ_0}.

Now stationarity, (16), and the above property imply

$$\begin{aligned}
&2H(\phi_0, \phi_0) \\
&= \tilde{E}_{\phi_0,\phi_0}\left[\log\left\{\sum_j U_j^{(1)} f(Y_1, \theta_j(\phi_0))\right\} + \log\left\{\sum_j U_j^{(2)} f(Y_2, \theta_j(\phi_0))\right\}\right] \\
&= \tilde{E}_{\phi_0,\phi_0}\left[\log\left\{\sum_j U_j^{(1)} p_2(Y_1, Y_2 | j; \phi_0)\right\}\right] \\
&= \int\int\int \sum_j u_j p_2(y_1, y_2 | j; \phi_0) \\
&\qquad \times \log\left\{\sum_j u_j p_2(y_1, y_2 | j; \phi_0)\right\} d\mu(y_1)\, d\mu(y_2)\, dQ(u).
\end{aligned}$$

Next we extend the construction in Lemma 4 to simultaneously include two sequences, $\{U^{(i)}\}$ which satisfies (15) with the parameter value ϕ_0, and $\{V^{(i)}\}$ which satisfies (15) with the parameter value ϕ. Then, as above,

$$\begin{aligned}
&2H(\phi_0, \phi) \\
&= \tilde{E}_{\phi_0,\phi}\left[\log\left\{\sum_j V_j^{(1)} p_2(Y_1, Y_2 | j; \phi)\right\}\right] \\
&= \int\int\int\int\int \sum_j u_j p_2(y_1, y_2 | j; \phi_0) \\
&\qquad \times \log\left\{\sum_j v_j p_2(y_1, y_2) | j; \phi)\right\} d\mu(y_1)\, d\mu(y_2)\, dQ'(u, v),
\end{aligned}$$

where $Q'(\cdot,\cdot)$ is the distribution of $(U^{(1)}, V^{(1)})$ under $\tilde{P}_{\phi_0,\phi}$. Since the marginal distribution of Q' corresponding to the first coordinate is Q, we have

$$K(\phi_0, \phi) = \int\int\int\int \sum_j u_j p_2(y_1, y_2|j; \phi_0) \times \log\left\{\frac{\sum_j u_j p_2(y_1, y_2|j; \phi_0)}{\sum_j v_j p_2(y_1, y_2|j; \phi)}\right\} d\mu(y_1)\, d\mu(y_2)\, dQ'(u, v).$$

Since the inner integral, for fixed u, v, is the Kullback–Leibler divergence between two mixture densities, $K(\phi_0, \phi) \geq 0$ and, if $K(\phi_0, \phi) = 0$, then this Kullback–Leibler divergence is zero for almost every pair u, v (with respect to Q'). However m-component mixtures of products of densities from the family $\{f(\cdot, \theta): \theta \in \Theta\}$ are identifiable by Condition 1 and the result of Teicher (1967) (see Section 3). Therefore (using Jensen's inequality), we conclude

$$\sum_j \sum_k u_j \alpha_{jk}(\phi_0) \delta_{(\theta_j(\phi_0), \theta_k(\phi_0))} = \sum_j \sum_k v_j \alpha_{jk}(\phi) \delta_{(\theta_j(\phi), \theta_k(\phi))}$$

for almost every pair u, v (with respect to Q'), where δ denotes a distribution function of a point mass. Since $U_j^{(1)}$ has the distribution of $P_{\phi_0}\{X_1 = j | Y_0, Y_{-1}, \ldots\}$, $\tilde{E}_{\phi_0,\phi}[U_j^{(1)}] = \alpha_j(\phi_0)$. Therefore, $K(\phi_0, \phi) = 0$ implies

$$\sum_j \sum_k \alpha_j(\phi_0) \alpha_{jk}(\phi_0) \delta_{(\theta_j(\phi_0), \theta_k(\phi_0))} = \sum_j \sum_k \int v_j \, dQ'(u, v) \alpha_{jk}(\phi) \delta_{(\theta_j(\phi), \theta_k(\phi))};$$

hence ϕ and ϕ_0 define the same symmetric law for $(\theta_{X_1}, \theta_{X_2})$ and so $\phi \sim \phi_0$. □

6. Consistency of the maximum-likelihood estimator

We can now present the main result, which concerns the consistency of the maximum-likelihood estimator. The results of the previous sections allow the application of the basic strategy invented by Wald (1949) and further developed by Kiefer and Wolfowitz (1956).

Consistency must be stated in terms of convergence of the equivalence class of the maximum-likelihood estimate $\hat{\phi}_n$ (see Section 3). We will obtain convergence in the quotient topology defined relative to the equivalence relation $\sim$. Redner (1981) used convergence in this sense for estimators of the parameters of a finite mixture distribution. Consistency in the sense of the quotient topology simply means that any open subset of the parameter space Φ^c which contains the equivalence class $\tilde{\phi}_0$ of the true parameter must, for large n, contain the equivalence class of $\hat{\phi}_n$.

Theorem 3. *Assume conditions 1–6 hold. Let ϕ_0 be the true parameter value and let $\hat{\phi}_n$ be a maximum-likelihood estimator. Then $\hat{\phi}_n$ converges to ϕ_0 in the quotient topology, with probability one.*

Proof. Let $q_n(\phi)$ denote $q_n(Y_1, \ldots, Y_n; \phi)$, and similarly for p_n. For $\phi \neq \phi_0$, we have $\lim_n E_{\phi_0}[\log q_n(\phi)]/n = H(\phi_0, \phi) < H(\phi_0, \phi_0)$, by Lemma 6 and (14); hence there is an $\varepsilon > 0$ and integer n_ε such that $E_{\phi_0}[\log q_{n_\varepsilon}(\phi)]/n_\varepsilon < H(\phi_0, \phi_0) - \varepsilon$. Now, q_{n_ε} is continuous, and, using the integrability condition 6, $E_{\phi_0}[\{\log(\sup_{\phi' \in \mathcal{O}_\phi} q_{n_\varepsilon}(\phi'))\}^+] < \infty$, for a small enough neighborhood $\mathcal{O}_\phi$ of ϕ; therefore, there is an open neighbourhood $\mathcal{O}_\phi$ for which $E_{\phi_0}[\log(\sup_{\phi' \in \mathcal{O}_\phi} q_{n_\varepsilon}(\phi'))]/n_\varepsilon < E_{\phi_0}[\log q_{n_\varepsilon}(\phi)]/n_\varepsilon + \frac{1}{2}\varepsilon < H(\phi_0, \phi_0) - \frac{1}{2}\varepsilon$. It follows from (10) that

$$\log\left(\min_j \alpha_j^{(1)}\right) \leq \log\left(\sup_{\phi' \in \mathcal{O}_\phi} p_n(\phi') \Big/ \sup_{\phi' \in \mathcal{O}_\phi} q_n(\phi')\right) \leq 0;$$

hence $\log(\sup_{\phi' \in \mathcal{O}_\phi} p_n(\phi'))/n$ and $\log(\sup_{\phi' \in \mathcal{O}_\phi} q_n(\phi'))/n$ have the same limiting behaviour as $n \to \infty$. Also, $W_{st} = \log(\sup_{\phi' \in \mathcal{O}_\phi} q_{t-s}(\phi'))$ satisfies the conditions of Kingman's subadditive ergodic theorem (see (11), (12) and (13)), and hence

$$H(\phi_0, \phi; \mathcal{O}_\phi) \overset{\text{def}}{=} \lim_n E_{\phi_0}\left[\log\left(\sup_{\phi' \in \mathcal{O}_\phi} q_n(\phi')\right)\right] \Big/ n$$

exists and

$$\lim_n \log\left(\sup_{\phi' \in \mathcal{O}_\phi} q_n(\phi')\right) \Big/ n = H(\phi_0, \phi; \mathcal{O}_\phi),$$

with probability one. By a property of subadditive processes (Kingman, 1976, Theorem 1.1),

$$H(\phi_0, \phi; \mathcal{O}_\phi) = \inf_n E_{\phi_0}\left[\log\left(\sup_{\phi' \in \mathcal{O}_\phi} q_n(\phi')\right)\right] \Big/ n,$$

so that

$$H(\phi_0, \phi; \mathcal{O}_\phi) \leq E_{\phi_0}\left[\log\left(\sup_{\phi' \in \mathcal{O}_\phi} q_{n_\varepsilon}(\phi')\right)\right] \Big/ n_\varepsilon.$$

Thus we have proved that, with probability one,

$$\lim_n \log\left(\sup_{\phi' \in \mathcal{O}_\phi} p_n(\phi')\right) \Big/ n = H(\phi_0, \phi; \mathcal{O}_\phi) < H(\phi_0, \phi_0) - \tfrac{1}{2}\varepsilon.$$

Let C be a closed subset of Φ^c, not containing any points of the equivalence class $\tilde{\phi}_0$. Since Φ^c is compact, C is compact and so is covered by the union $\bigcup_{h=1}^d \mathcal{O}_h$, where $\{\phi_1, \ldots, \phi_d\}$ is a finite set of C and $\mathcal{O}_h = \mathcal{O}_{\phi_h}$. Therefore, with probability one,

$$\sup_{\phi \in C}\left(\log p_n(\phi) - \log p_n(\phi_0)\right) = \max_h \left\{\log\left(\sup_{\phi \in \mathcal{O}_h} p_n(\phi)\right) - \log p_n(\phi_0)\right\} \to -\infty,$$

which implies that, for any open subset $\mathcal{O}$ of Φ^c which contains the equivalence class $\tilde{\phi}_0$, $\hat{\phi}_n \in \mathcal{O}$ for large n. It follows that the maximum-likelihood estimator converges to $\tilde{\phi}_0$ in the quotient topology, with probability one. □

Acknowledgements

This work is based on a part of my Ph.D. dissertation in the Department of Statistics, The University of British Columbia. I thank my supervisor Marty Puterman and Harry Joe for many useful discussions.

References

M. Askar and H. Derin, A recursive algorithm for the Bayes solution of the smoothing problem, IEEE Trans. Automat. Control 26 (1981) 558–560.

L.E. Baum and J.A. Eagon, An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology, Bull. Amer. Math. Soc. 73 (1967) 360–363.

L.E. Baum and T.A. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, Ann. Math. Statist. 37 (1966) 1554–1563.

L.E. Baum, T. Petrie, G. Soules and N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Ann. Math. Statist. 41 (1970) 164–171.

G.A. Churchill, Stochastic models for heterogeneous DNA sequences, Bull. Math. Biol. 51 (1989) 79–94.

A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), J. Roy. Statist. Soc. Ser. B 39 (1977) 1–38.

H. Furstenberg and H. Kesten, Products of random matrices, Ann. Math. Statist. 31 (1960) 457–469.

B.-H. Juang and L.R. Rabiner, A probabilistic distance measure for hidden Markov models, AT&T Tech. J. 64 (1985) 391–408.

S. Karlin and H.M. Taylor, A First Course in Stochastic Processes (Academic Press, New York, 1975).

J. Kiefer and J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters, Ann. Math. Statist. 27 (1956) 887–906.

J.F.C. Kingman, Subadditive processes, in: P.L. Hennequin, ed., Ecole d'Eté de Probabilités de Saint-Flour V-1976. Lecture Notes in Math. No. 539 (Springer, Berlin, 1976) pp. 167–223.

G. Kitagawa, Non-Gaussian state-space modeling of nonstationary time series, J. Amer. Statist. Assoc. 82 (1987) 1032–1041.

R. Kohn and C.F. Ansley, Comment on Kitagawa (1987), J. Amer. Statist. Assoc. 82 (1987) 1041–1044.

S.E. Levinson, L.R. Rabiner and M.M. Sondhi, An introduction to the application of the theory of probabilistic functions of a Markov process in automatic speech recognition, Bell System Tech. J. 62 (1983) 1035–1074.

G. Lindgren, Markov regime models for mixed distributions and switching regressions, Scand. J. Statist. 5 (1978) 81–91.

T. Petrie, Probabilistic functions of finite state Markov chains, Ann. Math. Statist. 40 (1969) 97–115.

R. Redner, Note on the consistency of the maximum likelihood estimate for non-identifiable distributions, Ann. Statist. 9 (1981) 225–228.

J.A. Smith, Statistical modeling of rainfall occurrences, Water Resour. Res. 23 (1987) 885–893.

H. Teicher, Identifiability of mixtures of product measures, Ann. Math. Statist. 38 (1967) 1330–1302.

A. Wald, Note on the consistency of the maximum likelihood estimate, Ann. Math. Statist. 20 (1949) 595–601.