AT&T Technical Journal
Vol. 64, No. 6, July–August 1985
Printed in U.S.A.

Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains

By B.-H. JUANG*

(Manuscript received November 15, 1984)

In this paper we discuss parameter estimation by means of the reestimation algorithm for a class of multivariate mixture density functions of Markov chains. The scope of the original reestimation algorithm is expanded and the previous assumptions of log concavity or ellipsoidal symmetry are obviated, thereby enhancing the modeling capability of the technique. Reestimation formulas in terms of the well-known forward-backward inductive procedure are also derived.

I. INTRODUCTION

Hidden Markov models, which use probabilistic functions of Markov chains to model random processes, have been found to be extremely useful for stock market behavior, ecology,[1-2] and more recently, speech recognition.[3-5] The effectiveness of this model class lies in its ability to deal with nonstationarity that often appears in the observed data sequences. The general structure of such a class of models may be briefly described as follows.

Consider a first-order N-state Markov chain governed by an $N \times N$ transition probability matrix $\mathbf{A} = [a_{ij}]$, and an initial probability vector $\mathbf{u}^t = [u_1 u_2 \cdots u_N]$. Obviously, $\sum_{j=1}^{N} a_{ij} = 1$ for any $i = 1, 2, \cdots, N$, and $\sum_{j=1}^{N} u_j = 1$. a_{ij} is the probability of making a transition from state

* AT&T Bell Laboratories.

i to state j given that the current state is i. For any integer state sequence $\Theta = (\theta_0, \theta_1, \theta_2, \cdots \theta_T)$, where $\theta_t \in \{1, 2, \cdots, N\}$, the probability of Θ being generated by the Markov source can be easily calculated by

$$\Pr(\Theta \mid \mathbf{A}, \mathbf{u}) = u_{\theta_0} a_{\theta_0\theta_1} \cdots a_{\theta_{T-1}\theta_T}. \tag{1}$$

Now, suppose Θ cannot be directly observed. Instead, we assume that what we observe is a stochastic process $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_T)$, produced by an underlying (stochastic) state sequence $(\theta_1, \theta_2, \cdots \theta_T)$. Each state, say i, manifests itself through a probability density function $f_i(s)$, $\int f_i(s)ds = 1$. We use $F = \{f_i(\cdot)\}$ to denote such a set of density functions. The probability density of $\mathbf{S} = S \triangleq (s_1, s_2, \cdots, s_T)$, given a specific state sequence Θ generated by the Markov chain with transition probability matrix $\mathbf{A}$, and initial probability vector $\mathbf{u}$ is thus

$$f(S \mid \Theta, \mathbf{A}, \mathbf{u}, F) = \prod_{t=1}^{T} f_{\theta_t}(s_t). \tag{2}$$

It then follows that the density of $\mathbf{S}$, given $\mathbf{A}$, $\mathbf{u}$ and F, is

$$f(S \mid \mathbf{A}, \mathbf{u}, F) = \sum_{\text{all } \Theta} u_{\theta_0} \prod_{t=1}^{T} a_{\theta_{t-1}\theta_t} f_{\theta_t}(s_t). \tag{3}$$

The triple $(\mathbf{A}, \mathbf{u}, F) \triangleq \lambda$ is called a (hidden) Markov model source and we write $f(S \mid \lambda) \triangleq f(S \mid \mathbf{A}, \mathbf{u}, F)$, for simplicity.

Given an observation sequence S, the objective in maximum-likelihood estimation is thus to maximize $f(S \mid \lambda)$ over all parameters in λ. Such a maximization problem is clearly nontrivial. To solve this problem, a reestimation algorithm—developed by Baum et al.[1] in 1970—that guarantees monotonic increase in the likelihood as the algorithm iterates, is often used. An auxiliary function, based upon the Kullback-Leibler number,[6] serves as the basis of Baum's optimization procedure, in which parameter estimates are characterized as the critical point of the auxiliary function. However, the development in Ref. 1 encounters difficulties when the densities $\{f_i(\cdot)\}$ are not log concave. The Cauchy density $f(s) = \pi^{-1}(1 + s^2)^{-1}$, which is only concave for $3s^2 \leq 1$, was cited as one such problematic example.

More than a decade later, in an effort to obviate the log-concavity limitation in Baum's algorithm, Liporace[7] invoked a representation theorem by Fan[8] to redefine the auxiliary function and then successfully extended the reestimation algorithm to accommodate a class of elliptically symmetric, multivariate distributions. As a result, each $f_i(s) \in F$ is allowed to assume the form

$$|\mathbf{R}_i|^{-1/2} h_i(g_i(s)), \tag{4}$$

where $g_i(s)$ is positive definite quadratic,

$$g_i(s) = (s - \eta_i)^* \mathbf{R}_i^{-1} (s - \eta_i).$$

The asterisk denotes the transpose of a vector or matrix as we, following Liporace, will be dealing with vector observations from this point on. The matrix $\mathbf{R}_i$ is positive definite and symmetric, and the location vector η_i is an arbitrary point in the observation space that is d-dimensional Euclidean.

While Liporace's results are significant in expanding the scope of the reestimation algorithm, the requirements that the observation densities be elliptically symmetric are in many real situations still very restrictive. In particular, useful parametrizations of speech signals, such as reflection coefficients and autocorrelation, have been shown by Gray and Markel[9] and Rabiner et al.,[10] respectively, not to exhibit the desired symmetry. This lack of symmetry is often observed even within each state because of the arbitrariness in choosing the number of states for modeling the given process. It is thus the purpose of this paper to further obviate the ellipsoidal symmetry assumption so that an even more versatile statistical modeling technique than the previous ones is obtainable. Levinson also reported the same effort.[11]

The class of densities $F = \{f_i(\cdot)\}$ we consider in this paper is the class of mixtures of general, strictly log-concave, *and/or* elliptically symmetric densities, having the form

$$f_i(s) = \sum_{k=1}^{M} c_{ik} b_{ik}(s), \tag{5}$$

where $b_{ik}(s)$ is general strictly log concave and/or elliptically symmetric and c_{ik} satisfies

$$\sum_{k=1}^{M} c_{ik} = 1 \quad \text{for} \quad i = 1, 2, \cdots, N. \tag{6}$$

As required in Liporace's results,[7] one extra assumption for elliptically symmetric $b_{ik}(s)$ is necessary: the density $b_{ik}(s)$ also satisfies the consistency conditions of Kolmogorov (see Ref. 12, p. 10) so that $b_{ik}(\cdot)$ has the representation

$$b_{ik}(s) = \int_0^\infty \mathcal{N}(s; \eta_{ik}, \nu\mathbf{R}_{ik}) dG(\nu) \tag{7}$$

for some probability distribution G on $[0, \infty)$. In (7), the expression $\mathcal{N}(s; \eta_{ik}, \nu\mathbf{R}_{ik})$ is the multivariate Gaussian density with mean vector η_{ik} and covariance matrix $\nu\mathbf{R}_{ik}$. Clearly, $\{f_i(s)\}$, as expressed in (5), is very general and may serve better in the modeling of many complex but realistic observations than unimodal, symmetric density functions.

This paper is organized as follows. The main body of the theory is presented in Section II, where the auxiliary function is redefined and the reestimation formula for all the parameters is derived. In Section

III, applications of the theory to familiar probability densities are discussed. Furthermore, for computational convenience, parameter reestimation using the forward-backward inductive procedure is also provided.

II. REESTIMATION

2.1 Joint density

For mathematical clarity, the following definitions are necessary.

Let Λ be an open subset of Euclidean p space. A hidden Markov model λ is a point in Λ and to each $\lambda \in \Lambda$ we have a smooth assignment $\lambda \rightarrow (\mathbf{A}(\lambda), \mathbf{u}(\lambda), F(\lambda))$. One trivial assignment is that dimensions in Λ are one-to-one, corresponding to the parameters defining the triple $(\mathbf{A}, \mathbf{u}, F)$, and thus p is the total number of model parameters.

Define the state alphabet $\Omega_s \triangleq \{1, 2, \cdots, N\}$. Let Ω_s^{T+1} be the $(T+1)$th Cartesian product of Ω_s. The state sequence space is denoted by Ω_s^{T+1}, and $\Theta \in \Omega_s^{T+1}$ means $\Theta = (\theta_0, \theta_1, \cdots, \theta_T)$, where every $\theta_t \in \Omega_s$.

We further define the branch alphabet $\Omega_b \triangleq \{1, 2, \cdots, M\}$. Similarly, Ω_b^T is the set of all T-tuples $K = (k_1, k_2, \cdots, k_T)$, where every $k_t \in \Omega_b$. K is called a branch sequence.

The global density function of (3) with state density defined by (5) can be written as

$$f(S \mid \lambda) = \sum_{\text{all } \Theta \in \Omega_s^{T+1}} u_{\theta_0} \prod_{t=1}^{T} \left[a_{\theta_{t-1}\theta_t} \cdot \sum_{k=1}^{M} c_{\theta_t k} b_{\theta_t k}(s_t) \right]. \tag{8}$$

The summand in (8) over all $\Theta \in \Omega_s^{T+1}$ is, in fact, the joint density $f(S, \Theta \mid \lambda)$, which can be expressed as

$$\begin{aligned} f(S, \Theta \mid \lambda) &= u_{\theta_0} \prod_{t=1}^{T} a_{\theta_{t-1}\theta_t} \sum_{k=1}^{M} c_{\theta_t k} b_{\theta_t k}(s_t) \\ &= \sum_{k_1=1}^{M} \sum_{k_2=1}^{M} \cdots \sum_{k_T=1}^{M} \left[u_{\theta_0} \prod_{t=1}^{T} a_{\theta_{t-1}\theta_t} b_{\theta_t k_t}(s_t) \right] \\ &\quad \cdot c_{\theta_1 k_1} c_{\theta_2 k_2} \cdots c_{\theta_T k_T}. \end{aligned} \tag{9}$$

We further define

$$f(S, \Theta, K \mid \lambda) = u_{\theta_0} \prod_{t=1}^{T} a_{\theta_{t-1}\theta_t} b_{\theta_t k_t}(s_t) c_{\theta_t k_t}. \tag{10}$$

Therefore, the joint density of the truncated stochastic process $\mathbf{S}$ is

$$f(S \mid \lambda) = \sum_{\Theta \in \Omega_s^{T+1}} \sum_{K \in \Omega_b^T} f(S, \Theta, K \mid \lambda). \tag{11}$$

An interpretation of (11) is that there are N^{T+1} possible stochastic state sequences that may lead to the observation S, with each possible state sequence being a superposition of M^T branch layers.

2.2 Auxiliary function and an inequality

For more general theoretical interest, let $\Omega = \Omega_s^{T+1} \times \Omega_b^T$ be a totally finite measure space with measure $\mu(\Theta, K)$. The joint density of (11) then has the following general form:

$$f(S \mid \lambda) = \int_\Omega f(S, \Theta, K \mid \lambda) d\mu(\Theta, K).$$

Following the concept of the Kullback-Leibler statistic, we define an auxiliary function $Q(\lambda, \lambda')$ of two model points, λ and λ', in Λ, given an observation S:

$$Q(\lambda, \lambda') \triangleq \int_\Omega f(S, \Theta, K \mid \lambda) \log f(S, \Theta, K \mid \lambda') d\mu(\Theta, K). \quad (12)$$

We now have the following theorem:

Theorem 1: If $Q(\lambda, \lambda') \geq Q(\lambda, \lambda)$ *then* $f(S \mid \lambda') \geq f(S \mid \lambda)$. *The inequality is strict unless* $f(S, \Theta, K \mid \lambda) = f(S, \Theta, K \mid \lambda')$ *almost everywhere* $d\mu(\Theta, K)$.

Proof: Similar to Baum et al.,[1] $\log x$ is strictly concave for $x > 0$. Hence,

$$\begin{aligned}
\log \frac{f(S \mid \lambda')}{f(S \mid \lambda)} &= \log \int_\Omega \frac{f(S, \Theta, K \mid \lambda')}{f(S \mid \lambda)} d\mu(\Theta, K) \\
&= \log \int_\Omega \frac{f(S, \Theta, K \mid \lambda)}{f(S \mid \lambda)} d\mu(\Theta, K) \frac{f(S, \Theta, K \mid \lambda')}{f(S, \Theta, K \mid \lambda)} \\
&\geq \int_\Omega \frac{f(S, \Theta, K \mid \lambda)}{f(S \mid \lambda)} \left[\log \frac{f(S, \Theta, K \mid \lambda')}{f(S, \Theta, K \mid \lambda)}\right] d\mu(\Theta, K) \\
&= [f(S \mid \lambda)]^{-1}[Q(\lambda, \lambda') - Q(\lambda, \lambda)] \geq 0
\end{aligned}$$

by hypothesis. The inequality above is due to Jensen's inequality for the measure $d\zeta(\Theta, K \mid \lambda) = f(S, \Theta, K \mid \lambda) d\mu(\Theta, K)/f(S \mid \lambda)$. This inequality is strict unless $f(S, \Theta, K \mid \lambda')/f(S, \Theta, K \mid \lambda)$ is constant almost everywhere $d\zeta(\Theta, K \mid \lambda)$, hence unless $f(S, \Theta, K \mid \lambda') = f(S, \Theta, K \mid \lambda)$ almost everywhere $d\mu(\Theta, K)$.

The significance of Theorem 1 will be discussed below. For simplicity, we often use the expression of (11) for the joint density and define the auxiliary function as

$$Q(\lambda, \lambda') = \sum_{\Theta \in \Omega_T^{T+1}} \sum_{K \in \Omega_T^T} f(S, \Theta, K \mid \lambda) \log f(S, \Theta, K \mid \lambda') \qquad (13)$$

as long as the key result of Theorem 1 is held valid.

2.3 Reestimation algorithm

Theorem 1 is one of the bases of Baum's reestimation algorithm that is sketched below for self-containedness. For a given observation S, the reestimation algorithm starts with an initial guess of the model λ. The parameter reestimates are then defined to be those that maximize $Q(\lambda, \lambda')$ as a function of λ'; that is, the model reestimate $\bar{\lambda}$ stemming from the current model λ is $\bar{\lambda} = \mathcal{T}(\lambda) \in \{\hat{\lambda} \in \Lambda \mid Q(\lambda, \hat{\lambda}) = \max_{\lambda' \in \Lambda} Q(\lambda, \lambda')\}$. The transformation $\mathcal{T}: \Lambda \to \Lambda$ is called the reestimation transformation. If Q has a unique global maximum as a function of λ', the set $\{\hat{\lambda}\}$ has only one element $\bar{\lambda}$. Then $\bar{\lambda}$ plays the role of λ as before and new reestimates are determined. The procedure iterates until some criterion is met.

Due to Theorem 1 and the following theorem, the above iterative procedure produces a sequence of reestimates that guarantee monotonic increase in the likelihood $f(S \mid \lambda)$ unless it reaches a critical point of the likelihood.

Theorem 2: Let $f(S, \Theta, K \mid \lambda)$ be continuously differentiable in λ for almost all $(\Theta, K) \in \Omega$. Let $\mathcal{T}$ be a continuous map of $\Lambda \to \Lambda$ such that for each fixed λ, $\bar{\lambda} = \mathcal{T}(\lambda)$ is a critical point of $Q(\lambda, \lambda')$ as a function of λ'. Then all fixed points of the reestimation transformation $\mathcal{T}$ are critical points of $f(S \mid \lambda)$, and if $f(S \mid \bar{\lambda}) > f(S \mid \lambda)$, unless $\bar{\lambda} = \lambda$, all limit points of $\mathcal{T}^n(\lambda_0) \triangleq \mathcal{T}(\mathcal{T}(\mathcal{T} \cdots (\mathcal{T}(\lambda_0))) \cdots)$ are fixed points of $\mathcal{T}$ for any $\lambda_0 \in \Lambda$.

Proof: Let ∇_λ be the gradient vector.

$$\begin{aligned}
\nabla_\lambda f(S \mid \lambda) \mid_\lambda &= \nabla_\lambda \int_\Omega f(S, \Theta, K \mid \lambda) d\mu(\Theta, K) \\
&= \int_\Omega \nabla_\lambda f(S, \Theta, K \mid \lambda) d\mu(\Theta, K) \\
&= \int_\Omega f(S\Theta, K \mid \lambda)[\nabla_\lambda \log f(S, \Theta, K \mid \lambda)] d\mu(\Theta, K) \\
&= \nabla_{\lambda'} Q(\lambda, (\lambda, \lambda') \mid_{\lambda' = \lambda}.
\end{aligned}$$

Thus $\nabla_\lambda f(S \mid \lambda) \mid_{\lambda = \bar{\lambda}} = 0$ if and only if $\nabla_{\lambda'} Q(\lambda, \lambda') \mid_{\lambda' = \lambda} = 0$ at $\lambda = \bar{\lambda}$. The rest of the proof follows Baum et al.[1]

Theorems 1 and 2 thus guarantee that after each iteration, the new reestimate $\bar{\lambda}$ improves the likelihood, i.e., $f(S \mid \bar{\lambda}) > f(S \mid \lambda)$, unless $\bar{\lambda}$ is a fixed point of the transformation. On the other hand, the trans-

formation will converge to a fixed point, or equivalently, a critical point of the likelihood, if an increase in the likelihood is maintained after each iteration and if the limit $\lim_{n\to\infty} \mathscr{T}^n(\lambda_0)$ exists, regardless of what the initial guess $\lambda_0(\in \Lambda)$ is. If $f(S \mid \lambda)$ has finitely many critical points, $\mathscr{T}^n(\lambda_0)$ approaches a critical point of $f(S \mid \lambda)$ that is at least a local maximum.

The transformation as previously defined requires maximization of the auxiliary function. Difficulties encountered in the maximization process would directly translate into difficulties in obtaining the maximum-likelihood estimate. We next show that if every $b_{ij}(\cdot)$, $i = 1, 2, \cdots, N$ and $j = 1, 2, \cdots, M$ is strictly log concave or elliptically symmetric with the representation (7), $Q(\lambda, \lambda')$ has a unique global maximum as a function of λ', and thus the transformation exists and is single valued. The reestimation algorithm is thus guaranteed to work for the joint density of (8).

2.4 *Maximization of the auxiliary function*

The auxiliary function for the joint density (8) with mixture densities is defined in (12). The following decomposition can be easily seen:

$$\log f(S, \Theta, K \mid \lambda') = \log \mu'_{\theta_0} + \sum_{t=1}^{T} \log a'_{\theta_{t-1}\theta_t} + \sum_{t=1}^{T} \log b'_{\theta_t k_t}(s_t) + \sum_{t=1}^{T} \log c'_{\theta_t k_t}. \quad (14)$$

The next theorem suggests that maximization of the likelihood by way of reestimation can be accomplished on individual parameter sets due to the separability shown in (14), if the following assumptions hold. Suppose for almost all (Θ, K), $\log f(S, \Theta, K \mid \lambda) = \sum_{i=1}^{q} \log f^{(i)}(S, \Theta, K \mid \lambda_i)$, where for each i and almost all (Θ, K) $\log f^{(i)}(S, \Theta, K \mid \lambda_i)$ has a unique global maximum as a function of λ_i. Note that $\lambda = \{\lambda_i\}$ and q is the number of parameter sets after separation. Define $Q_i(\lambda, \lambda'_i)$ by

$$Q_i(\lambda, \lambda'_i) = \int_{\Omega} f(S, \Theta, K \mid \lambda) \log f^{(i)}(S, \Theta, K \mid \lambda_i) d\mu(\Theta, K). \quad (15)$$

Then for λ fixed, $Q_i(\lambda, \lambda'_i)$ as a function of λ'_i has a unique global maximum $\bar{\lambda}_i$ that is a critical point of $Q_i(\lambda, \lambda'_i)$. The reestimation transformation $\mathscr{T}$ is thus defined as $\mathscr{T}: \lambda \to \bar{\lambda} = \{\bar{\lambda}_i\}$. We further define $\mathscr{T}_i: \lambda \to \tilde{\lambda}_i = \{\lambda_1, \lambda_2, \cdots, \bar{\lambda}_i, \cdots, \lambda_q\}$.

Theorem 3: Under the above assumptions, for all $\lambda \in \Lambda$, *and every* i, $f(S \mid \mathscr{T}_i(\lambda)) \geq f(S \mid \lambda)$ *with equality if and only if* λ_i *is a critical point of* $f(S \mid \lambda)$ *with respect to* λ_i *or, equivalently,* $\tilde{\lambda}_i$ *is a fixed point of* $\mathscr{T}_i$ *and furthermore,* $f(S \mid \mathscr{T}(\lambda)) \geq f(S \mid \lambda)$ *with equality if and only if* λ *is a critical point of* $f(S \mid \lambda)$ *or, equivalently, a fixed point of* $\mathscr{T}$.

Proof:

$$Q(\lambda, \tilde{\lambda}_i) = \sum_{\substack{j=1 \\ j \neq i}}^{q} Q_j(\lambda, \lambda_j) + Q_i(\lambda, \bar{\lambda}_i)$$

$$\geqslant \sum_{j=1}^{q} Q_j(\lambda, \lambda_j) = Q(\lambda, \lambda),$$

so Theorem 1 implies $f(S \mid \tilde{\lambda}_i) \geqslant f(S \mid \lambda)$. Since $Q_i(\lambda, \lambda_i')$ has a unique global maximum as a function of λ_i', the inequality $Q(\lambda, \tilde{\lambda}_i) \geqslant Q(\lambda, \lambda)$ is strict unless $\bar{\lambda}_i = \lambda_i$. Furthermore,

$$Q(\lambda, \bar{\lambda}) = \sum_{i=1}^{q} Q_i(\lambda, \bar{\lambda}_i) \geqslant \sum_{i=1}^{q} Q_i(\lambda, \lambda_i) = Q(\lambda, \lambda),$$

the second half of the theorem is thus true.

The separation of (14) is seen to be the key to the increased versatility of the reestimation algorithm in accommodating mixture-observation densities. Let $\mathbf{b}_{jk}$ be the parameter set defining the density $b_{jk}(s)$. Obviously, if $b_{jk}(s)$ is multivariate Gaussian, $\mathbf{b}_{jk} = (\eta_{jk}, \mathbf{R}_{jk})$, where η_{jk} is the mean vector and $\mathbf{R}_{jk}$ is the covariance matrix. We now write the auxiliary function in a separated form using the simplified expression of (13) without loss of generality.

$$Q(\lambda, \lambda') \triangleq \sum_{\Theta} \sum_{K} f(S, \Theta, K \mid \lambda) \log f(S, \Theta, K \mid \lambda')$$

$$= \sum_{\Theta} \sum_{K} f(S, \Theta, K \mid \lambda) \left\{ \log u'_{\theta_0} + \sum_{t=1}^{T} \log a_{\theta_{t-1}\theta_t} \right.$$

$$\left. + \sum_{t=1}^{T} \log b'_{\theta_t k_t}(s_t) + \sum_{t=1}^{T} \log c'_{\theta_t k_t} \right\}$$

$$= Q_u(\lambda, \mathbf{u}') + \sum_{i=1}^{N} Q_{a_i}[\lambda, \{a'_{ij}\}_{j=1}^{N}]$$

$$+ \sum_{j=1}^{N} \sum_{k=1}^{M} Q_b(\lambda, \mathbf{b}'_{jk}) + \sum_{j=1}^{N} Q_{c_j}[\lambda, \{c'_{jk}\}_{k=1}^{M}], \qquad (16)$$

where

$$Q_u(\lambda, \mathbf{u}') = \sum_{\Theta} \sum_{K} f(S, \Theta, K \mid \lambda) \log u'_{\theta_0}$$

$$= \sum_{i=1}^{N} \sum_{K} f(S, \theta_0 = i, K \mid \lambda) \log u'_i \qquad (17)$$

$$Q_{a_i}[\lambda, \{a'_{ij}\}_{j=1}^N] = \sum_{\Theta} \sum_{K} f(S, \Theta, K \mid \lambda) \sum_{t=1}^{T} \log a'_{\theta_{t-1}\theta_t} \delta(\theta_{t-1} - i)$$

$$= \sum_{j=1}^{N} \sum_{t=1}^{T} \sum_{K} f(S, \theta_{t-1} = i, \theta_t = j, K \mid \lambda) \log a'_{ij} \quad (18)$$

$$Q_b(\lambda, \mathbf{b}'_{jk}) = \sum_{\Theta} \sum_{K} f(S, \Theta, K \mid \lambda)$$

$$\cdot \sum_{t=1}^{T} \log b'_{\theta_t k_t}(s_t) \delta(\theta_t - j) \delta(k_t - k)$$

$$= \sum_{t=1}^{T} f(S, \theta_t = j, k_t = k \mid \lambda) \log b'_{jk}(s_t), \quad (19)$$

and

$$Q_{c_j}(\lambda, \{c'_{jk}\}_{k=1}^M) = \sum_{\Theta} \sum_{K} f(S, \Theta, K \mid \lambda) \sum_{t=1}^{T} \log c'_{\theta_t k_t} \delta(\theta_t - j)$$

$$= \sum_{k=1}^{M} \sum_{t=1}^{T} f(S, \theta_t = j, k_t = k \mid \lambda) \log c'_{jk}. \quad (20)$$

The above expression $\delta(\cdot)$ is the Kronecker delta function.

Individual maximization of Q_u, Q_{a_i} and Q_{c_i} for $i = 1, 2, \cdots, N$ subject to the constraints

$$1. \sum_{j=1}^{N} u_j = 1, \qquad u_j \geq 0$$

$$2. \sum_{j=1}^{N} a_{ij} = 1, \qquad a_{ij} \geq 0 \qquad \text{for all appropriate } i \text{ and } j$$

$$3. \sum_{j=1}^{M} c_{ij} = 1, \qquad c_{ij} \geq 0, \quad (21)$$

respectively, is well known.[13,14] These individual auxiliary functions have the same form $\sum_{j=1}^{N} w_j \log y_j$, which as a function of $\{y_j\}_{j=1}^N$, subject to the constraints $\sum_{j=1}^{N} y_j = 1$ and $y_j \geq 0$, attains a global maximum at the single point

$$y_j = \frac{w_j}{\sum_{i=1}^{N} w_i} \quad j = 1, 2, \cdots, N. \quad (22)$$

The result has been proved in many ways.[14]

When $b_{jk}(s)$ is strictly log concave in $\mathbf{b}_{jk}$ and $\lim_{|\mathbf{b}_{jk}| \to \infty} \log b_{jk}(s) = -\infty$, it is easily seen that for λ fixed $Q_b(\lambda, \mathbf{b}'_{jk})$ has a unique global maximum that is a critical point of $Q_b(\lambda, \mathbf{b}'_{jk})$. When $b_{jk}(s)$ is elliptically symmetric, the following theorem due to Liporace[7] is applicable.

Theorem 4: If (i) $b_{jk}(s)$ *has the representation of eq.* (7), *and* (ii) *there*

are among $s_1, s_2, \cdots, s_T$, $d + 1$ *observations, any* d *(the dimension of each observation vector) of which are linearly independent, for fixed* λ, $Q_b(\lambda, \mathbf{b}'_{jk})$ *has a unique global maximum as a function of* $\mathbf{b}'_{jk} = (\eta'_{jk}, \mathbf{R}'_{jk})$, *and this maximum is the one and only critical point of* $Q_b(\lambda, \mathbf{b}'_{jk})$.

Proof of this theorem is easily obtained by following the Appendix in Ref. 7.

The reestimation algorithm has thus been extended to accommodate the hidden Markov joint density (8) with mixture observation densities.

III. APPLICATIONS

We now explicitly derive the reestimation transformation. By applying eq. (22), we can easily calculate $\bar{\mathbf{u}}$, $\bar{\mathbf{A}}$, and $\{\bar{c}_{ik}\}_{k=1}^{M}$ for $i = 1, 2, \cdots, n$, the reestimates that for fixed λ maximize $Q_u(\lambda, \mathbf{u}')$, $Q_{a_i}(\lambda, \{a'_{ij}\}_{j=1}^{N})$ and $Q_{c_i}(\lambda, \{c'_{ik}\}_{k=1}^{M})$ for $i = 1, 2, \cdots, N$, as a function of $\mathbf{u}'$, $\{a'_{ij}\}_{j=1}^{N}$ and $\{c'_{ik}\}_{k=1}^{M}$, respectively.

1. Initial probability:

$$Q_u(\lambda, \mathbf{u}') \triangleq \sum_{j=1}^{N} \sum_{K} f(S, \theta_0 = i, K \mid \lambda) \log u'_j.$$

Hence, for $i = 1, 2, \cdots, N$,

$$\bar{u}_i = \sum_{K \in \Omega_S^T} f(S, \theta_0 = i, K \mid \lambda) \Big/ \sum_{K \in \Omega_S^T} f(S, K \mid \lambda)$$

$$= f(S, \theta_0 = i \mid \lambda)/f(S \mid \lambda). \tag{23}$$

2. Transition probability:
For every $i = 1, 2, \cdots, N$,

$$Q_{a_i}(\lambda, \{a'_{ij}\}_{j=1}^{N}) = \sum_{j=1}^{N} \sum_{t=1}^{T} \sum_{K} f(S, \theta_{t-1} = i, \theta_t = j, K \mid \lambda) \log a'_{ij}.$$

Therefore, for $i, j = 1, 2, \cdots, N$,

$$\bar{a}_{ij} = \sum_{t=1}^{T} \sum_{K} f(S, \theta_{t-1} = i, \theta_t = j, K \mid \lambda) \Big/ \cdot \sum_{t=1}^{T} \sum_{K} f(S, \theta_{t-1} = i, K \mid \lambda)$$

$$= \sum_{t=1}^{T} f(S, \theta_{t-1} = i, \theta_t = j \mid \lambda) \Big/ \sum_{t=1}^{T} f(S, \theta_{t-1} = i \mid \lambda). \tag{24}$$

3. Branch probability:
For every $i = 1, 2, \cdots, N$,

$$Q_{c_i}(\lambda, \{c'_{ik}\}_{k=1}^{M}) = \sum_{k=1}^{M} \sum_{t=1}^{T} f(S, \theta_t = i, k_t = k \mid \lambda) \log c'_{ij}.$$

Then obviously, for $i = 1, 2, \cdots, N$, $k = 1, 2, \cdots, M$,

$$\bar{c} = \sum_{t=1}^{T} f(S, \theta_t = i, k_t = k \mid \lambda) \Big/ \sum_{t=1}^{T} f(S, \theta_t = i \mid \lambda). \tag{25}$$

4. Branch density:
For every $i = 1, 2, \cdots, N$, and $k = 1, 2, \cdots, M$,

$$Q_b(\lambda, \mathbf{b}'_{ik}) = \sum_{t=1}^{T} f(S, \theta_t = i, k_t = k \mid \lambda) \log b'_{ik}(s_t).$$

Maximization of $Q_b(\lambda, \mathbf{b}'_{ij})$ with respect to $\mathbf{b}'_{ik}$ is well known for many familiar density functions. The solution to the maximization problem is, in general, obtained through differentiation; i.e., we find $\bar{\mathbf{b}}_{ik}$ that satisfies

$$\begin{aligned} &\nabla_{\mathbf{b}'_{ik}} Q_b(\lambda, \mathbf{b}'_{ik}) \Big|_{\mathbf{b}'_{ik} = \bar{\mathbf{b}}_{ik}} \\ &\quad = \sum_{t=1}^{T} f(S, \theta_t = i, k_t = k \mid \lambda) \frac{\nabla_{\mathbf{b}'_{ik}} b'_{ik}(s_t)}{b'_{ik}(s_t)} \Big|_{\mathbf{b}'_{ik} = \bar{\mathbf{b}}_{ik}} \\ &\quad = 0. \end{aligned} \tag{26}$$

For strictly log concave $b_{ik}(s)$, the solution can be easily found. For elliptically symmetric $b_{ik}(s)$,

$$b_{ik}(s) = |\mathbf{R}_{ik}|^{-1/2} h_{ik}(g_{ik}(s)),$$

where

$$g_{ik}(s) = (s - \eta_{ik})^* \mathbf{R}_{ik}^{-1} (s - \eta_{ik}),$$

with representation (7), Liporace's results apply.[7] In particular, the solution to (26), i.e., reestimates $\bar{\eta}_{ik}$ and $\bar{\mathbf{R}}_{ik}$, is given by

$$\bar{\eta}_{ik} = \frac{\sum_{t=1}^{T} f(S, \theta_t = i, k_t = k \mid \lambda) \cdot s_t}{\sum_{t=1}^{T} f(S, \theta_t = i, k_t = k \mid \lambda)} \tag{27}$$

$$\bar{\mathbf{R}}_{ik} = \frac{\sum_{t=1}^{T} f(S, \theta_t = i, k_t = k \mid \lambda) \cdot (s_t - \eta_{ik})(s_t - \eta_{ik})^*}{\sum_{t=1}^{T} f(S, \theta_t = i, k_t = k \mid \lambda)} \cdot \tag{28}$$

Note that if $b_{ik}(s)$ is multivariate Gaussian, the reestimates $\bar{\eta}_{ik}$ and $\bar{\mathbf{R}}_{ik}$ above are readily applicable. It is also easy to see that each $\bar{\mathbf{R}}_{ik}$ is positive definite. If s is any d-dimensional vector,

$$s^* \bar{\mathbf{R}}_{ik} s = \sum_{t=1}^{T} x_{ik}(t)[s^*(s_t - \eta_{ik})]^2 \geqslant 0, \tag{29}$$

where

$$x_{ik}(t) = f(S, \theta_t = i, k_t = k \mid \lambda) \Big/ \sum_{t=1}^{T} f(S, \theta_t = i, k_t = k \mid \lambda)$$

$$\geqslant 0.$$

The inequality (29) is strict provided for any η_{ik} the vectors $\{s_t - \eta_{ik}\}$ span the d-dimensional observation space, i.e., the observation process $S = (s_1, s_2, \cdots, s_T)$ satisfies the condition laid out in Theorem 4.

The above reestimates can be conveniently calculated with the forward-backward inductive procedure. Define "forward probabilities" $\alpha_0(i) = u_i$, $i = 1, 2, \cdots, N$, and

$$\begin{aligned} \alpha_t(i) &= f(s_1, s_2, \cdots, s_t, \theta_t = i \mid \lambda) \\ &= \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} f_i(s_t), \end{aligned} \tag{30}$$

for $i = 1, 2, \cdots, N$ and $t = 1, 2, \cdots, T$. Branch densities $f_i(s)$ are defined in (5). Similarly, define "backward probabilities"

$$\begin{aligned} \beta_t(i) &= f(s_{t+1}, s_{t+2}, \cdots, s_T \mid \theta_t = i, \lambda) \\ &= \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} f_j(s_{t+1}), \end{aligned} \tag{31}$$

and $\beta_T(i) = 1$, for $i = 1, 2, \cdots, N$ and $t = T - 1, T - 2, \cdots, 0$. Further define "branch probability" $\gamma_t(i, k)$

$$\begin{aligned} \gamma_t(i, k) &= f(s_1, s_2, \cdots, s_t, \theta_t = i, k_t = k \mid \lambda) \\ &= \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} c_{ik} b_{ik}(s_t), \end{aligned} \tag{32}$$

for $i = 1, 2, \cdots, N$, $k = 1, 2, \cdots, M$, and $t = 1, 2, \cdots, T$. Then,

$$f(S, \theta_t = i \mid \lambda) = \alpha_t(i) \beta_t(i), \tag{33}$$

which leads to

$$f(S \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i), \tag{34}$$

and in particular at $t = T$,

$$f(S \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{35}$$

Furthermore,

$$\begin{aligned} f(S, \theta_{t-1} = i, \theta_t = j \mid \lambda) &= \alpha_{t-1}(i) a_{ij} f_j(s_t) \beta_t(j) \\ &= \alpha_{t-1}(i) a_{ij} \left[\sum_{k=1}^{M} c_{jk} b_{jk}(s_t) \right] \beta_t(j), \end{aligned} \tag{36}$$

and

$$\begin{aligned} f(S, \theta_t = i, k_t = k \mid \lambda) &= \gamma_t(i, k) \beta_t(i) \\ &= \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} c_{ik} b_{ik}(s_t) \beta_t(i). \end{aligned} \tag{37}$$

As a result, the reestimates are expressed in terms of the forward and backward probabilities:

1. Initial Probability:

$$\begin{aligned} \bar{u}_i &= \alpha_0(i) \beta_0(i) \Big/ \sum_{j=1}^{N} \alpha_0(j) \beta_0(j) \\ &= \alpha_0(i) \beta_0(i) \Big/ \sum_{j=1}^{N} \alpha_T(j) \end{aligned} \tag{38}$$

2. Transition probability:

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T} \alpha_{t-1}(i) a_{ij} \left[\sum_{k=1}^{M} c_{jk} b_{jk}(s_t) \right] \beta_t(j)}{\sum_{t=1}^{T} \alpha_{t-1}(i) \beta_{t-1}(i)} \tag{39}$$

3. Branch probability:

$$\bar{c}_{ik} = \frac{\sum_{t=1}^{T} \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} c_{ik} b_{ik}(s_t) \beta_t(i)}{\sum_{t=1}^{T} \alpha_t(i) \beta_t(i)} \tag{40}$$

4. Branch density:

$$\bar{\eta}_{ik} = \frac{\sum_{t=1}^{T} \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} c_{ik} b_{ik}(s_t) \beta_t(i) \cdot s_t}{\sum_{t=1}^{T} \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} c_{ik} b_{ik}(s_t) \beta_t(i)} \tag{41}$$

and

$$\bar{\mathbf{R}}_{ik} = \frac{\sum_{t=1}^{T} \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} c_{ik} b_{ik}(s_t) \beta_t(i) \cdot (s_t - \eta_{ik})(s_t - \eta_{ik})^*}{\sum_{t=1}^{T} \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} c_{ik} b_{ik}(s_t) \beta_t(i)} \cdot \tag{42}$$

Note that the results of (41) and (42) apply to the case of mixture of elliptically symmetric densities with the representation (7). Mixtures of multivariate Gaussian densities, of course, fall into such a category. For other strictly log-concave densities, (26) applies.

Note that the above results can be easily applied to conventional parametric estimation of mixture distributions by setting the number of states, N, to unity.

IV. CONCLUSIONS

We have extended the reestimation algorithm to accommodate a broad class of mixtures of strictly log-concave or elliptically symmetric multivariate distributions. The algorithm is particularly useful in modeling nonstationary stochastic processes with multimodal non-symmetric probabilistic functions of Markov chains that could not be dealt with previously. Explicit reestimates in terms of the well-known forward-backward inductive probabilities are derived for computational ease. Due to the greatly expanded capability of the reestimation method, more accurate modeling of sophisticated signals and thus improvements in various applications such as speech recognition are expected.

REFERENCES

1. L. E. Baum et al. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains,"Ann. Math. Statist., *41* (1970), pp. 164–71.
2. 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), pp. 360–3.
3. J. K. Baker, "The DRAGON System—An Overview," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-23* (February 1975), pp. 24–9.
4. F. Jelinek, "Continuous Speech Recognition By Statistical Methods," Proc. IEEE, *64* (April 1976), pp. 532–56.
5. L. R. Rabiner, S. E. Levinson, and M. M. Sondhi. "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent Isolated Word Recognition," B.S.T.J. *62*, No. 4, Part 1 (April 1983), pp. 1075–106.

6. S. Kullback and R. A. Leibler, "On Information and Sufficiency," Ann. Math. Statist., 22 (March 1951), pp. 79–86.
7. L. R. Liporace, "Maximum Likelihood Estimation for Multivariate Observations of Markov Sources," IEEE Trans. Inform. Theory, *IT-28* (September 1982), pp. 729–34.
8. K. Fan, "Les Fonction Dèfinies—Positives et les Fonctions Complètement Monotones," Mèmonial des Sciences Math., *114,* 1950.
9. A. H. Gray and J. D. Markel, "Quantization and Bit Allocation in Speech Processing," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-24* (December 1976), pp. 459–73.
10. L. R. Rabiner, J. G. Wilpon, and J. G. Ackenhusen, "On the Effects of Varying Analysis Parameters on an LPC-Based Isolated Word Recognizer," B.S.T.J., *60* (July–August 1981), pp. 893–911.
11. S. E. Levinson, private communication.
12. J. L. Doob, *Stochastic Processes*, New York; Wiley, 1953.
13. L. E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes," Inequalities, *3* (1972), pp. 1–8.
14. 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," B.S.T.J., *62,* No. 4, Part 1 (April 1983), pp. 1035–74.

AUTHOR

Biing-Hwang Juang, B.Sc. (Electrical Engineering), 1973, National Taiwan University, Republic of China; M.Sc. and Ph.D. (Electrical and Computer Engineering), University of California, Santa Barbara, 1979 and 1981, respectively; Speech Communications Research Laboratory (SCRL), 1978; Signal Technology, Inc., 1979–1982; AT&T Bell Laboratories, 1982; AT&T Information Systems Laboratories, 1983; AT&T Bell Laboratories, 1983—. Before joining AT&T Bell Laboratories, Mr. Juang worked on vocal tract modeling at the Speech Communications Research Laboratory, and on speech coding and interference suppression at Signal Technology, Inc. Presently, he is a member of the Acoustics Research Department, where he is researching speech communications techniques and stochastic modeling of speech signals.