The Annals of Statistics
1994, Vol. 22, No. 4, 1884–1895

CONSISTENT AND ASYMPTOTICALLY NORMAL PARAMETER ESTIMATES FOR HIDDEN MARKOV MODELS

BY TOBIAS RYDÉN[1]

Lund Institute of Technology

Hidden Markov models are today widespread for modeling of various phenomena. It has recently been shown by Leroux that the maximum-likelihood estimate (MLE) of the parameters of a such a model is consistent, and local asymptotic normality has been proved by Bickel and Ritov. In this paper we propose a new class of estimates which are consistent, asymptotically normal and almost as good as the MLE.

1. Introduction. A hidden Markov model (HMM) is, loosely speaking, a sequence $\{Y_k\}_{k=1}^{\infty}$ of random variables obtained in the following way. First, a realization of a finite state Markov chain $\{X_k\}$ is created. This chain is sometimes called the *regime*. Then, conditioned on $\{X_k\}$, the Y-variables are independent of each other, and the distribution of Y_k depends on $\{X_k\}$ only through X_k. This definition will be made more formal below.

HMMs have during the last decade become widespread for modeling sequences of dependent random variables with applications in areas like speech processing [Rabiner (1989)], biochemistry [Fredkin and Rice (1992)] and biology [Leroux and Puterman (1992)]. Markov-modulated Poisson processes (MMPPs), a kind a doubly stochastic Poisson process used, for example, to model arrival processes in modern communication networks [Heffes and Lucantoni (1986)], are also closely related to HMMs. Sometimes the hidden Markov chain $\{X_k\}$ does indeed exist, so that the physical nature of the problem suggests the use of an HMM; in other cases HMMs just provide a good fit to the data.

Inference for HMMs was first considered by Baum and Petrie (1966), who treated the case when $\{Y_k\}$ takes values in the finite set $\{1, \ldots, p\}$. In Baum and Petrie (1966), results on consistency and asymptotic normality of the maximum-likelihood estimate (MLE) are given, and the conditions for consistency are weakened in Petrie (1969). In the latter paper the identifiability problem is also discussed; that is, under what conditions there are no other parameters that induce the same law for $\{Y_k\}$ as the true parameter does, with exception for permutations of states. For general HMMs, with Y_k, conditioned on X_k, having density $f(\cdot\,; \theta_{X_k})$, Lindgren (1978) constructed consistent and asymptotically normal estimates of the θ-parameters, but no results on the estimation of the transition probabilities were given. Later, Leroux (1992) proved consistency of

Received July 1993; revised January 1994.
[1]Supported by the National Board for Industrial and Technical Development Grant 88-02060P.
AMS 1991 *subject classifications*. Primary 62M09; secondary 62F12, 62E25.
Key words and phrases. Hidden Markov model, consistency, asymptotic normality, identifiability, regenerative process.

the MLE for general HMMs under mild conditions, and local asymptotic normality in the sense of Le Cam has been proved by Bickel and Ritov (1993). However, in order to apply the latter result to the MLE, $\sqrt{n}$-consistency of this estimate must be proved, which has not been done. Thus, asymptotic normality of the MLE is still an open question. Recursive likelihood estimation for HMMs has been studied by Holst and Lindgren (1991), and ML estimation for MMPPs has been treated by Rydén (1994).

In this paper we propose a new class of estimates which under fairly general conditions are consistent and asymptotically normal, and which are almost as good as the MLE. These estimates are obtained by splitting the observations into groups of fixed size, viewing these groups as independent and then maximizing the resulting likelihood. This approach was also used in Rydén (1993) for establishing similar results for MMPPs.

The paper is organized as follows. In Section 2 the notation is given and some regularity conditions are stated. In Section 3 the identifiability problem is discussed, and the new class of estimates is introduced in Section 4. Consistency and asymptotic normality are proved in Sections 5 and 6, respectively, and finally some numerical results are given in Section 7.

2. Preliminaries. We will parametrize the problem in a way that is essentially the same as in Leroux (1992). The parameter space is denoted by $\Phi \subseteq \mathbb{R}^s$, and the regime $\{X_k\}_{k=1}^{\infty}$ is a stationary Markov chain with state space $\{1, \ldots, r\}$ and transition probability matrix $\{\alpha_{ij}(\phi)\}$, where $\phi \in \Phi$ is any parameter. Let $\{f(\cdot\,; \theta); \theta \in \Theta\}$ be a family of densities on a Euclidian space $\mathcal{Y}$ with respect to a measure μ, parametrized by $\theta \in \Theta \subseteq \mathbb{R}^q$. Given $\{X_k\}$, $\{Y_k\}_{k=1}^{\infty}$ is a sequence of independent $\mathcal{Y}$-valued random variables, Y_k having density $f(\cdot\,; \theta_{X_k}(\phi))$, where θ_i, $i = 1, \ldots, r$, are functions $\Phi \to \Theta$.

The most common case is $\phi = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{rr}, \theta_1, \ldots, \theta_r)$ with $\alpha_{ij}(\cdot)$ and $\theta_i(\cdot)$ being the coordinate projections. We will refer to this case as the "usual parametrization" in the sequel.

The stationary distribution of the transition probability matrix $\{\alpha_{ij}(\phi)\}$ will be denoted by $\{\alpha_i(\phi)\}$. This distribution need not be unique, and therefore we will also need the enlarged parameter space

$$\Psi = \bigcup (\phi, \alpha),$$

where the union runs over all pairs (ϕ, α) such that $\phi \in \Phi$ and α is a stationary distribution for $\{\alpha_{ij}(\phi)\}$.

Having introduced this notation, the likelihood for a sequence of observations $y_1, \ldots, y_m$, or the joint density of $Y_1, \ldots, Y_m$, is

$$(1) \quad p_m(y_1, \ldots, y_m; \psi) = \sum_{x_1=1}^{r} \cdots \sum_{x_m=1}^{r} \alpha_{x_1} f(y_1; \theta_{x_1}(\phi)) \prod_{k=2}^{m} \alpha_{x_{k-1} x_k} f(y_k; \theta_{x_k}(\phi)),$$

where $\psi = (\phi, \alpha)$. On the subset Φ^e of Φ such that $\{\alpha_{ij}(\phi)\}$ has a unique stationary distribution, $p_m(y_1, \ldots, y_m; \cdot)$ can be viewed as a function of ϕ. For such ϕ we

will use both ϕ and ψ as function arguments in the sequel. The true parameter value will be denoted by ϕ^0.

The following conditions will be used throughout the paper. Note that condition C1 ensures that there is a neighborhood of ϕ^0 which is contained in Φ^e.

C1. The transition probability matrix $\{\alpha_{ij}(\phi^0)\}$ is irreducible.

C2. Φ is a closed set.

C3. For all i and j, $\alpha_{ij}(\cdot)$ and $\theta_i(\cdot)$ are continuous, $f(y;\cdot)$ is continuous for all y, and for all y, $f(y;\phi) \to 0$ as $|\phi| \to \infty$.

There is an integer $m > 0$ such that:

C4. For each $\psi \neq \psi^0$, $p_m(y_1,\ldots,y_m;\psi)$ and $p_m(y_1,\ldots,y_m;\psi^0)$ are not equal μ^m-a.e.

C5. $E_{\psi^0}|\log p_m(Y_1,\ldots,Y_m;\psi^0)| < \infty$.

C6. For each ψ there is a $\delta > 0$ such that $E_{\psi^0}[\sup_{|\psi'-\psi|\le\delta}(\log p_m(Y_1,\ldots,Y_m;\psi'))^+] < \infty$, and there is a $b > 0$ such that $E_{\psi^0}[\sup_{|\psi'|\ge b}(\log p_m(Y_1,\ldots,Y_m;\psi'))^+] < \infty$.

C7. ϕ^0 is an interior point of Φ.

C8. For all $y_1,\ldots,y_m$, all partial derivatives of $p_m(y_1,\ldots,y_m;\phi)$ with respect to ϕ of order three or less exist in a neighborhood of ϕ^0.

C9. For each i,

$$E_{\phi^0}\left[\left(\frac{\partial}{\partial\phi_i}\log p_m(Y_1,\ldots,Y_m;\phi^0)\right)^2\right] < \infty,$$

for each i and j,

$$E_{\phi^0}\left|\frac{\partial^2}{\partial\phi_i\,\partial\phi_j}\log p_m(Y_1,\ldots,Y_m;\phi^0)\right| < \infty,$$

and $E_{\phi^0}[\log p_m(Y_1,\ldots,Y_m;\phi)]$ may be differentiated twice under the integral sign with respect to ϕ at ϕ^0.

C10. For each i, j and k, there exists a function $M\colon \mathbb{R}^m \to \mathbb{R}$ and a neighborhood G of ϕ^0 such that $E_{\phi^0}M(Y_1,\ldots,Y_m) < \infty$ and

$$\sup_{\phi'\in G}\left|\frac{\partial^3}{\partial\phi_i\,\partial\phi_j\,\partial\phi_k}\log p_m(y_1,\ldots,y_m;\phi')\right| \le M(y_1,\ldots,y_m)$$

for all $y_1,\ldots,y_m$.

We will use a technique developed by Wald (1949) to prove consistency [this technique is also used by Leroux (1992)], and conditions C2–C6 are similar in spirit to those stated in Wald (1949), although the continuity assumptions can be weakened slightly. They are also very close to the conditions in Leroux (1992), but here partly formulated in terms of m-dimensional densities.

Condition C1 implies that $\{Y_k\}$ is ergodic; see Lemma 1 of Leroux (1992). Condition C4 says that the m-dimensional distribution of $\{Y_k\}$ uniquely identifies ψ^0 within Ψ; we will discuss this assumption further in the next section.

Conditions C7–C10 are of the kind that are normally used to prove asymptotic normality in a "Cramér fashion." Bickel and Ritov (1993) assume similar conditions to hold, but, in addition, they also need some moment conditions which can be omitted in the current approach.

3. Identifiability. In this section we will discuss the identifiability condition C4 in closer detail. Suppose that the following condition is satisfied.

C4′. The family of mixtures of at most r elements of $\{f(y;\ \theta);\ \theta \in \Theta\}$ is identifiable.

This condition means that if $\theta_i \in \Theta$ and $\theta_i' \in \Theta$ for $i = 1, \ldots, r$, and $(a_1, \ldots, a_r)$ and $(a_1', \ldots, a_r')$ are probability vectors, then (δ_θ is the point mass at θ)

$$\sum_{i=1}^{r} a_i f(y;\theta_i) = \sum_{i=1}^{r} a_i' f(y;\theta_i') \ \mu\text{-a.e.} \quad \Rightarrow \quad \sum_{i=1}^{r} a_i \delta_{\theta_i} = \sum_{i=1}^{r} a_i' \delta_{\theta_i'},$$

that is, we can identify the mixing distribution. This holds, for example, for the Poisson family, the negative exponential family and the normal family with fixed variance. So far we have just discussed identifiability of one-dimensional distributions, but it turns out that this property carries over to multidimensional ones [see Teicher (1967)]; that is, the family of mixtures of at most r elements of $\Pi_1^m f(y_i;\theta_i)$ (over Θ^m) is identifiable.

Now, assume that we have the usual parametrization and that condition C4′ is satisfied. Let r_1 be the number of distinct θ_i^0 and define the function γ: $\{1, \ldots, r\} \to \{1, \ldots, r_1\}$ by $\gamma(i) = \gamma(j)$ iff $\theta_i^0 = \theta_j^0$; that is, γ describes how the states of $\{X_k\}$ are "clustered." Then, by the result in Teicher (1967), it is obvious that C4 will be satisfied iff ψ^0 is uniquely determined by the m-dimensional distribution of $\{\gamma(X_k)\}$ over Ψ, the "product space" of $r \times r$ stochastic matrices and the corresponding stationary distributions. Of course, we can always permute the states of $\{X_k\}$ without changing the distribution of $\{Y_k\}$, but this unessential ambiguity can be taken care of, for example, by ordering the θ_i. If $r_1 = r$, that is, all θ_i^0 are distinct, then it is clear that for $m \geq 2$, condition C4 will be satisfied if C1 is. On the other hand, if $\gamma(i) = \gamma(j)$ for some $i \neq j$, then there is, in general, an infinite number of stochastic matrices that induce the same finite-dimensional, and hence also m-dimensional, distributions for $\{\gamma(X_k)\}$ as ψ^0 does. This is shown in Ito, Amari and Kobayashi (1992), where an explicit algorithm for constructing such other matrices is also given.

To summarize, if we have the usual parametrization and condition C4′ holds, the most interesting case is that when all θ_i^0 are distinct. It should also be noted that in any case the finite-dimensional distributions of $\{Y_k\}$ are uniquely determined by the $2r$-dimensional one. This is true since the corresponding property holds for $\{\gamma(X_k)\}$; see Gilbert (1959).

4. Maximum split data likelihood estimates. Suppose that we have observed a data sequence $y_1, \ldots, y_{mn}$. If the m-dimensional random variables $(Y_1, \ldots, Y_m)$, $(Y_{m+1}, \ldots, Y_{2m})$, and so on, were independent, then the likelihood

would be

$$\mathcal{L}^s(\psi; y_1, \ldots, y_{mn}) = \prod_{k=1}^{n} p_m\big(y_{m(k-1)+1}, \ldots, y_{mk}; \psi\big), \tag{2}$$

where p_m is the joint density of $(Y_1, \ldots, Y_m)$ [see (1)]. We call $\mathcal{L}^s$ a *split data likelihood*, and any global maximum point $\hat{\psi}^{(m)}$ of it is a *maximum split data likelihood estimate* (MSDLE). To stress the dependence on m, we sometimes write "the m-dimensional MSDLE." Although $\mathcal{L}^s$ is a "false" likelihood, it will nevertheless turn out that the MSDLE is strongly consistent, because of the identifiability condition C4, asymptotically normal and in many cases as good as the MLE.

The case $m = 1$ was treated by Lindgren (1978), but, as noted in Section 1, this m is, in general, too small for estimation of the complete parameter ϕ.

5. Consistency. In this section we prove strong consistency of the MSDLE. The basic technique of the proof is due to Wald (1949). We start with a lemma concerning the m-dimensional Kullback–Leibler information

$$K_m(\psi^0, \psi) = E_{\psi^0}\left[\log \frac{p_m(Y_1, \ldots, Y_m; \psi^0)}{p_m(Y_1, \ldots, Y_m; \psi)}\right].$$

LEMMA 1. *Assume that conditions* C4–C6 *hold. Then* $K_m(\psi^0, \psi) \geq 0$ *with equality iff* $\psi = \psi^0$.

PROOF. Let $H_m(\psi^0, \psi) = E_{\psi^0}[\log p_m(Y_1, \ldots, Y_m; \psi)]$. By conditions C5 and C6, $H_m(\psi^0, \psi^0)$ is finite and $H_m(\psi^0, \psi) < \infty$, so that $K_m(\psi^0, \psi)$ is well defined. By Jensen's inequality,

$$-K_m(\psi^0, \psi) = E_{\psi^0}\left[\log \frac{p_m(Y_1, \ldots, Y_m; \psi)}{p_m(Y_1, \ldots, Y_m; \psi^0)}\right] \leq \log 1 = 0,$$

with equality iff $p_m(y_1, \ldots, y_m; \psi) = p_m(y_1, \ldots, y_m; \psi^0)$ μ^m-a.e. However, by condition C4, this is true iff $\psi = \psi^0$. □

We now prove strong consistency.

THEOREM 1. *Assume that conditions* C1–C6 *hold and let* $\hat{\psi}^{(m)}(n)$ *be the* m*-dimensional MSDLE based on* mn *observations. Then* $\hat{\psi}^{(m)}(n) \to \psi^0$ P_{ψ^0}*-a.s. as* $n \to \infty$.

PROOF. By condition C3, $p_m(y_1, \ldots, y_m; \cdot)$ is continuous for all $y_1, \ldots, y_m$. As in Wald (1949), one readily shows

$$\lim_{\delta \downarrow 0} E_{\psi^0}\left[\sup_{|\psi' - \psi| \leq \delta} \log p_m(Y_1, \ldots, Y_m; \psi')\right] = E_{\psi^0}\big[\log p_m(Y_1, \ldots, Y_m; \psi)\big] \tag{3}$$

for any $\psi \in \Psi$, and

$$\lim_{b \to \infty} E_{\psi^0}\left[\sup_{|\psi| \geq b} \log p_m(Y_1, \ldots, Y_m; \psi)\right] = -\infty. \tag{4}$$

Now, let $\varepsilon > 0$ be arbitrary, let $S_\varepsilon = \{\psi \in \Psi; |\psi - \psi^0| < \varepsilon\}$ and let $C = \Psi \cap S_\varepsilon^c$. Choose $b_0 > 0$ such that

$$E_{\psi^0}\left[\sup_{|\psi| > b_0} \log p_m(Y_1, \ldots, Y_m; \psi)\right] \leq E_{\psi^0}\left[\log p_m(Y_1, \ldots, Y_m; \psi^0)\right] - 1,$$

[this can be done by (4)], and let $C_1 = C \cap \{\psi \in \Psi; |\psi| \leq b_0\}$. It follows from Lemma 1 and (3) that for each $\psi \in C_1$ there is an ε_ψ and a neighborhood G_ψ of ψ such that

$$E_{\psi^0}\left[\sup_{\psi' \in G_\psi} \log p_m(Y_1, \ldots, Y_m; \psi')\right] \leq E_{\psi^0}\left[\log p_m(Y_1, \ldots, Y_m; \psi^0)\right] - \varepsilon_\psi.$$

Note that conditions C2 and C3 imply that Ψ is closed (as a subset of $\mathbb{R}^{s+r}$), so that C_1 is compact, and thus there is a finite set $\{\psi_1, \ldots, \psi_d\} \subseteq \Psi$ such that $C_1 \subseteq \bigcup_1^d G_i$, where $G_i = G_{\psi_i}$. Define also $G_0 = \{\psi \in \Psi; |\psi| > b_0\}$. The ergodicity of $\{Y_k\}$ now yields

$$\begin{aligned}
&\sup_{\psi \in S_\varepsilon^c} \left(\log \mathcal{L}^s(Y_1, \ldots, Y_{mn}; \psi) - \log \mathcal{L}^s(Y_1, \ldots, Y_{mn}; \psi^0)\right) \\
&\quad = \max_{0 \leq i \leq d} \left(\sup_{\psi \in G_i} \log \mathcal{L}^s(Y_1, \ldots, Y_{mn}; \psi) - \log \mathcal{L}^s(Y_1, \ldots, Y_{mn}; \psi^0)\right) \to -\infty,
\end{aligned}$$

and since ε is arbitrary, $\widehat{\psi}^{(m)}(n)$ is strongly consistent. □

If condition C4 does not hold, then we can introduce an equivalence relation $\sim$ on Ψ by writing $\psi \sim \psi'$ if ψ and ψ' induce the same m-dimensional distribution for $\{Y_k\}$, and the proof of the theorem above can easily be modified to yield strong consistency in the quotient topology generated by $\sim$; that is, if G is an open set containing the equivalence class of ψ^0, then $\widehat{\psi}^{(m)}(n) \in G$ for n sufficiently large, P_{ψ^0}-a.s. The main theorem of Leroux (1992) is formulated in this way. With the last paragraph of Section 3 in mind, it is clear that if we have the usual parametrization, condition C4′ holds and $m \geq 2r$, then $\sim$, in fact, denotes equivalence of the finite-dimensional distributions.

6. Asymptotic normality. The hardest part in proving asymptotic normality of an MLE is often to obtain a central limit theorem for the score function, that is, the derivative of the log-likelihood. In many cases a martingale approach is successful. For the MSDLE, asymptotic normality is very easily verified, since $\mathcal{L}^s$ is a product of scalars, and hence $\log \mathcal{L}^s$ is a sum. We will use the regenerative properties of $\{Y_k\}$, and therefore we give some basic results for such processes.

A discrete-time random process $\{Z_k\}_{k=1}^{\infty}$ is said to be regenerative with independent cycles if there exists a (possibly delayed) renewal process $\{S_k\}_{k=0}^{\infty}$, $S_k = T_0 + \cdots + T_k$, such that:

(i) for each $n \geq 0$, $\{T_{n+1}, T_{n+2}, \ldots, \{Z_{S_n+k}\}_{k=0}^{\infty}\}$ is independent of $S_0, \ldots, S_n$ and its distribution does not depend on n, and

(ii) conditioned on $\{S_k\}$, $\{\{Z_j; S_{k-1} \leq j \leq S_k - 1\}\}_{k=1}^{\infty}$ is a sequence of independent random elements.

We will study sums of the form $\Sigma_{k=1}^{n} g(Z_k)$, where $g = (g^{(1)}, \ldots, g^{(d)})$ is an $\mathbb{R}^d$-valued function, and for this we introduce the random variable $U = \Sigma_{k=1}^{T_1 - 1} g(Z_k)$, the mean cycle length $\overline{T} = E[T_1]$ and the probability measure P^p defined by $P^p(A) = P(A \mid S_0 = 1)$, corresponding to the case when $\{S_k\}$ is a pure renewal process.

A straightforward modification of Theorem 5.3.1 of Asmussen (1987) proves the following law of large numbers.

THEOREM 2. *If $\overline{T} < \infty$ and $E^p|U^{(i)}| < \infty$ for all i, then*

$$\frac{1}{n}\sum_{k=1}^{n} g(Z_k) \xrightarrow{P} E^p[U]/\overline{T}.$$

A similar modification of Theorem 5.3.2 of Asmussen (1987) and an application of the Cramér–Wold device gives a central limit theorem.

THEOREM 3. *If $E^p T_1^2 < \infty$ and $E^p|U^{(i)}|^2 < \infty$ for all i, then*

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} \{g(Z_k) - E^p[U]/\overline{T}\} \xrightarrow{d} \mathcal{N}(0, \Sigma/\overline{T}),$$

where

$$\Sigma_{ij} = \mathrm{Cov}^p\left(U^{(i)} - \frac{E^p U^{(i)}}{\overline{T}} T_1,\ U^{(j)} - \frac{E^p U^{(j)}}{\overline{T}} T_1\right).$$

We now have the necessary tools for proving asymptotic normality of the

MSDLE. Let $i_0 \in \{1, \ldots, r\}$ be arbitrary and define

$$h_i(y_1, \ldots, y_m; \phi) = \frac{\partial}{\partial \phi_i} \log p_m(y_1, \ldots, y_m; \phi),$$

$$T = \min\{k > 1: X_{m(k-1)+1} = i_0\},$$

$$A_{ij}^{(m)} = E_{\phi^0}\left[\sum_{k=1}^{T-1} h_i\left(Y_{m(k-1)+1}, \ldots, Y_{mk}; \phi^0\right) \times h_j\left(Y_{m(k-1)+1}, \ldots, Y_{mk}; \phi^0\right) \middle| X_1 = i_0\right],$$

$$V_{ij}^{(m)} = E_{\phi^0}\left[\sum_{k=1}^{T-1} h_i\left(Y_{m(k-1)+1}, \ldots, Y_{mk}; \phi^0\right) \times \sum_{k=1}^{T-1} h_j\left(Y_{m(k-1)+1}, \ldots, Y_{mk}; \phi^0\right) \middle| X_1 = i_0\right].$$

THEOREM 4. *Assume that conditions* C1–C10 *hold and let* $\widehat{\psi}^{(m)}(n) = (\widehat{\phi}^{(m)}(n), \widehat{\alpha}^{(m)}(n))$ *be the m-dimensional MSDLE based on mn observations. If* $A^{(m)}$ *is nonsingular and m is a multiple of the period of* $\{\alpha_{ij}(\phi^0)\}$, *then*

$$\sqrt{n}\left(\widehat{\phi}^{(m)}(n) - \phi^0\right) \to \mathcal{N}\left(0, C^{(m)}\right),$$

where $C^{(m)} = [A^{(m)}]^{-1} V^{(m)} [A^{(m)}]^{-1} / \alpha_{i_0}(\phi^0)$.

PROOF. The proof is essentially the same as the proof of Theorem 6.4.1(ii) of Lehmann (1991). Since $\widehat{\phi}^{(m)}$ is strongly consistent, for n sufficiently large $\widehat{\phi}^{(m)}(n)$ is an interior point of Φ for which the stationary distribution of the corresponding transition probability matrix is unique. For such n a Taylor expansion of the gradient of

$$\ell(\phi) = \log \mathcal{L}^s(\phi) = \sum_{k=1}^{n} \log p_m\left(Y_{m(k-1)+1}, \ldots, Y_{mk}; \phi\right)$$

about ϕ^0 yields

$$0 = \ell_i'\left(\widehat{\phi}^{(m)}\right) = \ell_i'\left(\phi^0\right) + \sum_{j=1}^{s} \left(\widehat{\phi}_j^{(m)} - \phi_j^0\right) \ell_{ij}''\left(\phi^0\right) + \frac{1}{2} \sum_{j=1}^{s} \sum_{k=1}^{s} \left(\widehat{\phi}_j^{(m)} - \phi_j^0\right)\left(\widehat{\phi}_k^{(m)} - \phi_k^0\right) \ell_{ijk}'''\left(\overline{\phi}\right),$$

where $\overline{\phi}$ is a point on the line between ϕ^0 and $\widehat{\phi}^{(m)}$, and the dependence of $\widehat{\phi}^{(m)}$

on n has been suppressed. Following Lehmann (1991), we have to prove

$$A_{ijn} = -\frac{1}{n}\ell''_{ij}(\phi^0) - \frac{1}{2n}\sum_{k=1}^{s}(\hat{\phi}_k^{(m)} - \phi_k^0)\ell'''_{ijk}(\bar{\phi}) \xrightarrow{P_{\phi^0}} \alpha_{i_0}^0 A_{ij}^{(m)}, \tag{5}$$

$$\{W_{in}\}_{i=1}^{s} = \left\{\frac{1}{\sqrt{n}}\ell'_i(\phi^0)\right\} \to \mathcal{N}(0, \alpha_{i_0}^0 V^{(m)}) \quad P_{\phi^0}\text{-weakly}, \tag{6}$$

where $\alpha^0 = \alpha(\phi^0)$.

By the ergodicity of $\{Y_k\}$, the strong consistency of $\hat{\phi}^{(m)}$ and condition C10, the second term on the right-hand side of (5) tends to 0 P_{ϕ^0}-a.s., and it follows from condition C9 that

$$-\frac{1}{n}\ell''_{ij}(\phi^0) \to A'_{ij} = E_{\phi^0}\left[-\frac{\partial^2}{\partial\phi_i\,\partial\phi_j}\log p_m(Y_1,\ldots,Y_m;\phi^0)\right] \quad P_{\phi^0}\text{-a.s.}$$

and

$$A'_{ij} = E_{\phi^0}\left[h_i(Y_1,\ldots,Y_m;\phi^0)\,h_j(Y_1,\ldots,Y_m;\phi^0)\right]. \tag{7}$$

Now, note that $\{(Y_{m(k-1)+1},\ldots,Y_{mk})\}_{k=1}^{\infty}$ is a regenerative process with independent cycles, k being a regeneration point if $X_{m(k-1)+1} = i_0$. Fix i and j, let $g(y_1,\ldots,y_m) = h_i(y_1,\ldots,y_m;\phi^0)h_j(y_1,\ldots,y_m;\phi^0)$ and define U as above. Since $\{X_{m(k-1)+1}\}$ is a Markov chain with stationary state probabilities α^0, the mean cycle length is $\bar{T} = 1/\alpha_{i_0}^0 < \infty$, and hence the hidden Markov structure of $\{Y_k\}$ and condition C9 imply $E_{\phi^0}^p|U| < \infty$. By Theorem 2, the ergodicity of $\{Y_k\}$ and (7), it follows that $A'_{ij} = \alpha_{i_0}^0 E_{\phi^0}^p U$, proving (5).

It remains to prove (6). Redefine g by $g^{(i)} = h_i$ for $i = 1,\ldots,s$ and redefine U in the obvious way. Since $E_{\phi^0}^p T_1^2 < \infty$, we have $E_{\phi^0}^p|U^{(i)}|^2 < \infty$ for each i. Moreover, $E_{\phi^0}g^{(i)}(Y_1,\ldots,Y_m;\phi^0) = 0$, and, in view of the ergodicity of $\{Y_k\}$ and Theorem 2, we get $E_{\phi^0}^p U = 0$. Hence, by Theorem 3,

$$\frac{1}{\sqrt{n}}\ell'(\phi^0) = \frac{1}{\sqrt{n}}\sum_{k=1}^{n} g(Y_{m(k-1)+1},\ldots,Y_{mk};\phi^0) \to \mathcal{N}(0, \alpha_{i_0}^0 V), \quad P_{\phi^0}\text{-weakly},$$

which proves (6) and thus also the theorem. □

Expressing $V^{(m)}$ and $A^{(m)}$ in terms of regenerative theory is particularly useful if we want to obtain these matrices by simulation. It is simple to generate samples of the random variable U which are independent, and thus we may easily estimate the mean of U with desired accuracy by the sample mean. In practice, we do, of course, not know the true parameters, but we may then use the MSDLE instead.

We could also have used the mixing properties of $\{Y_k\}$ to derive the central limit result needed in the proof. By an argument given in Lindgren (1978), page

TABLE 1
True parameter values, MLEs and 30-dimensional MSDLEs with approximate 95% confidence intervals for cases A–D. The number of observed samples N was in all cases 5000

Case		a	b	θ_1	θ_2
A	True values	0.1	0.4	1	4
	MLE	0.0938	0.4043	1.007	3.909
	MSDLE & 95% C.I.	0.0943 ± 0.015	0.4069 ± 0.048	1.012 ± 0.045	3.919 ± 0.22
B	True values	0.1	0.1	1	4
	MLE	0.0870	0.0893	0.984	3.997
	MSDLE & 95% C.I.	0.0878 ± 0.014	0.0874 ± 0.015	0.988 ± 0.049	4.004 ± 0.094
C	True values	0.1	0.4	1	2
	MLE	0.0839	0.4287	1.012	2.079
	MSDLE & 95% C.I.	0.0791 ± 0.072	0.4626 ± 0.19	1.026 ± 0.095	2.128 ± 0.51
D	True values	0.1	0.1	1	2
	MLE	0.0777	0.0851	0.988	2.042
	MSDLE & 95% C.I.	0.0881 ± 0.031	0.0846 ± 0.031	0.964 ± 0.087	2.006 ± 0.11

87, $\{(Y_{m(k-1)+1}, \ldots, Y_{mk})\}$ is strongly mixing, and thus a central limit theorem like Theorem 18.5.3 of Ibragimov and Linnik (1971) can be applied.

For the theorem to hold, A must be nonsingular. If A is singular, then $\{h_i(\cdot\,; \phi^0)\}_{i=1}^s$ are linearly dependent, but in most cases $\{h_i(\cdot\,; \phi^0)\}$ are nonlinear functions such that this is impossible.

7. Numerical examples. With the usual parameterization and $f(y; \theta)$ being the Poisson family, the negative exponential family or the normal family with fixed variance, on readily verifies that conditions C1–C10 are satisfied for any $m \geq 2$ if θ_i^0 are distinct. Thus, so far we know that for any $m \geq 2$ the MSDLE is a consistent and asymptotically normal estimate of the parameters, but we do not know *which* m to choose. This question will be addressed now.

We will study four different two-component Poisson mixtures with transition probability matrix

$$\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

and Poisson parameters θ_1 and θ_2, respectively, listed in Table 1 and denoted cases A–D. For each case, one run of length $N = 5000$ samples was simulated, and all point estimates are computed from these four runs. Intuitively, cases A and B are the easiest ones for estimation, and cases C and D are the hardest ones. This is confirmed by the MLEs; see Table 1.

First, we study the MSDLE errors. In Figure 1 the errors in the m-dimensional MSDLE relative to the errors in the MLE are plotted. The curves indicate that for $m \geq 20$, say, the MSDLE is as good as the MLE. Of course, the figure shows results only for one sample run from each of two different HMMs, but we have also studied other examples which all gave similar curves.

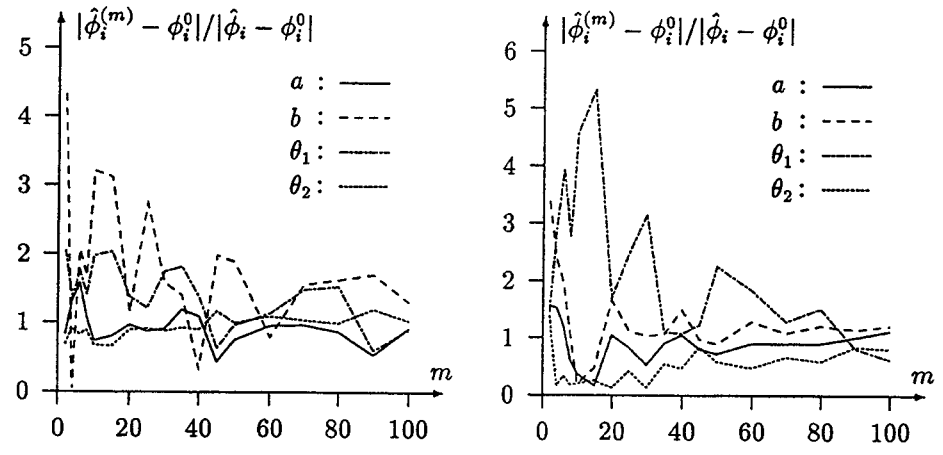

FIG. 1. *Error in the m-dimensional MSDLE relative to error in the MLE versus m for case A (left) and case D (right). The number of observed samples N was in all cases* 5000.

Next, we study the confidence intervals obtained from Theorem 4 above. If the total number of observed samples N is fixed, the number of groups $n = \lfloor N/m \rfloor$ will decrease as m grows, but, on the other hand, $C^{(m)}$ will also change. In Figure 2 half of the relative widths of the approximate 95% confidence intervals for the parameters are plotted versus m for cases A and D. The covariance matrices $C^{(m)}$ were obtained by simulation as outlined above, using 20,000 independent samples of U calculated for the true parameters. The curves indicate that it suffices to choose m larger than a threshold; making m even larger does not improve the confidence intervals much. This threshold depends on the specific example, though, but these curves together with other examples show that

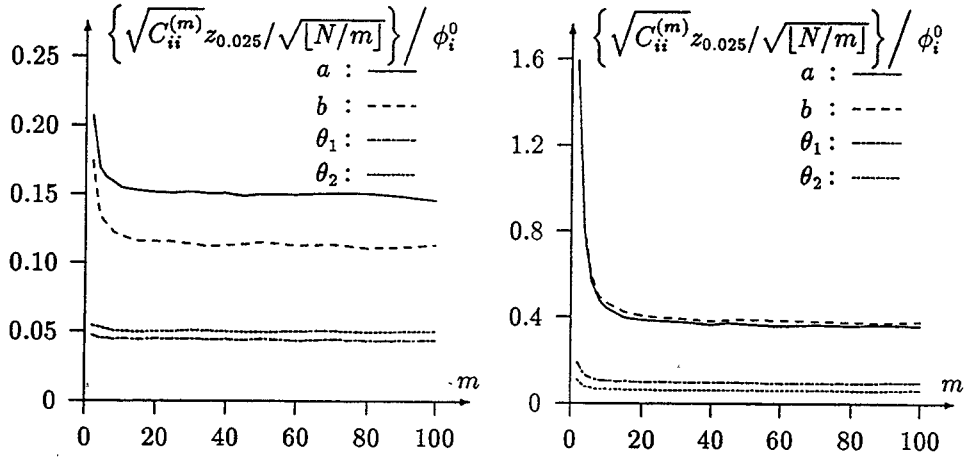

FIG. 2. *Half of the relative widths of the approximate* 95% *confidence intervals versus m for case A (left) and case D (right). The number of observed samples N was in all cases* 5000.

$m = 30$ is a safe choice. Not surprisingly, the curves also show that the transition probabilities are harder to estimate than the Poisson parameters.

Finally, in Table 1 we give MSDLEs and approximate 95% confidence intervals for all cases. Here, $m = 30$ and the MSDLEs were used in the simulations performed to obtain the covariance matrices $C^{(m)}$.

REFERENCES

ASMUSSEN, S. (1987). *Applied Probability and Queues*. Wiley, New York.

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

BICKEL, P. J. and RITOV, Y. (1993). Inference in hidden Markov models. I. Local asymptotic normality in the stationary case. Preprint.

FREDKIN, D. R. and RICE, J. A. (1992). Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. Roy. Soc. London Ser. B* **249** 125–132.

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

HEFFES, H. and LUCANTONI, D. (1986). A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communication* **4** 856–867.

HOLST, U. and LINDGREN, G. (1991). Recursive estimation in mixture models with Markov regime. *IEEE Trans. Inform. Theory* **37** 1683–1690.

IBRAGIMOV, I. A. and LINNIK, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.

ITO, H., AMARI, S.-I. and KOBAYASHI, K. (1992). Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Trans. Inform. Theory* **38** 324–333.

LEHMANN, E. L. (1991). *Theory of Point Estimation*. Wadsworth and Brooks/Cole, Belmont, CA.

LEROUX, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.* **40** 127–143.

LEROUX, B. G. and PUTERMAN, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48** 545–558.

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

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

RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77** 257–284.

RYDÉN, T. (1993). Consistent and asymptotically normal parameter estimates for Markov modulated Poisson processes. Technical Report 1993:5, Dept. Mathematical Statistics, Lund Institute of Technology.

RYDÉN, T. (1994). Parameter estimation for Markov modulated Poisson processes. *Stochastic Models.* **10** 795–829.

TEICHER, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Statist.* **38** 1300–1302.

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

DEPARTMENT OF MATHEMATICAL STATISTICS
LUND INSTITUTE OF TECHNOLOGY
BOX 118
S-221 00 LUND
SWEDEN