THIELE CENTRE
FOR APPLIED MATHEMATICS IN NATURAL SCIENCE

Asymptotics for Estimating Equations in Hidden Markov Models

Jørgen Vinsløv Hansen and Jens Ledet Jensen

Research Report

No. 07 I June 2008

ASYMPTOTICS FOR ESTIMATING EQUATIONS IN HIDDEN MARKOV MODELS

Jørgen Vinsløv Hansen and Jens Ledet Jensen

University of Aarhus

Abstract

Results on asymptotic normality for the maximum likelihood estimate in hidden Markov models are extended in two directions. The stationarity assumption is relaxed, which allows for a covariate process influencing the hidden Markov process. Furthermore a class of estimating equations is considered instead of the maximum likelihood estimate. The basic ingredients are mixing properties of the process and a general central limit theorem for weakly dependent variables. The results are illustrated with a cyclic model for the progesterone concentration in cowmilk.

Key words and phrases: Cyclic model, Estimating equation, Mixing properties, Progesterone concentration.

1 Introduction

Unless simulation based methods are used inference in hidden Markov models is based on the asymptotic normality of the parameter estimates. For the case of a finite state space for both the hidden variable x and the observed variable y, asymptotic normality for the maximum likelihood estimate was established in the pioneering paper of Baum and Petrie (1966). More than thirty years elapsed until this result was generalised to a general state space for the observed variable y by Bickel, Ritov and Rydén (1998), and still further generalized to a non-discrete state space for the hidden variable x by Jensen and Petersen (1999). In these papers stationarity is a crucial assumption. The log likelihood is a sum where the individual terms are the log densitites of y_i given the past $y_1, \dots, y_{i-1}$. These are replaced by the similar expressions conditioned instead on the infinite past $\dots, y_{-1}, y_0, \dots, y_{i-1}$. A martingale central limit theorem is then used to establish asymptotic normality of the score function. In this paper we use a different approach that allows us to consider nonhomogeneous processes and to consider alternatives to the maximum likelihood estimates. To illustrate the scope of the setup we briefly describe an example from evolutionary biology.

Example 1. Let $v(t) = (v_1(t), \dots, v_n(t))$ be a sequence of letters from the alphabet $\{A, G, C, T\}$ of nucleotides at time t. The sequence at time $t = 0$ is fixed and known. Time is discrete. The process is observed at time $t = T$, but not observed at the times

$t = 1, 2, \ldots, T-1$ in between. The sequence $v(t)$ evolves according to a Markov chain with transition probabilitites of the form

$$p(v(t+1)|v(t)) = \prod_{i=1}^{n} h(v_i(t+1)|v_{i-1}(t+1), v_i(t), v_{i+1}(t)),$$

for some transition probability h. This formalizes a time discretized version of a model where the probability of a change of a nucleotide $v_i(t)$ depends on the two neighbouring nucleotides. Let now $x_i = (v_i(1), v_i(2), \ldots, v_i(T))$ be the complete history for nucleotide i. It can be seen that the conditional distribution of x_i given $x_1, \ldots, x_{i-1}$ depends on (x_{i-2}, x_{i-1}) only. We thus have a second order hidden Markov model where the observed variable is $y_i = v_i(T)$. The underlying Markov structure is inhomogeneous due to the fixed initial sequence $v(0)$.

Asymptotic normality for a class of estimating equations, in the setting af evolutionary models for DNA, has been treated in Jensen (2005). In that paper both the state space of the hidden variable x and the observed variable y is finite. Here we extend the results in Jensen (2005) to a setup akin that of Jensen and Petersen (1999) with a general state space for the observed variable and a general state space for the hidden variable. Nonhomogeneity is introduced through a covariate. We base the asymptotic normality of the "score function" directly on the mixing properties of the process, using a central limit theorem extracted from Götze and Hipp (1983). Although the state space is general the conditions imposed effectively restricts the space to be compact.

In section 2 we describe the setup and results in detail and define the class of estimating equations that we consider. In section 3 we illustrate the results for a hidden cyclic model used to describe the progesterone concentration in cowmilk. The proofs of the results are split into three sections. In section 4 we study the mixing properties of the process and use these in section 5 to derive a central limit theorem for the "score function". Finally, in section 6 we derive the uniform convergence of the "observed information".

2 Assumptions and results

We consider an observed process $y_1, \ldots, y_n$ controlled by an unobserved Markov process $\{x_i\}$. Conditionally on the x-process the y_is are independent. Both the observed y_i and the unobserved x_i may be influenced by a covariate z_i, making the process inhomogeneous. The transition density of the Markov process is $p_\theta(x_i|x_{i-1}; z_i)$, where $p_\theta(\tilde{x}|x; z)$ is a density in $\tilde{x}$ with respect to a probability measure μ on the state space for the hidden variable. The conditional density of y_i is $p_\theta(y_i|x_i; z_i)$, where $p_\theta(y|x; z)$ is a density in y with respect to a measure ν. Both these densities are parametrized by the d-dimensional parameter θ. We split the assumptions into two parts, one part concerned with the process itself, Conditions 2 and 3 below, and another part concerned with the estimating function used, Condition 5 below.

Condition 2 ensures mixing of the underlying Markov chain. In order to allow for the possibility that in a single step the Markov chain can reach only a subset of the state space, we use the m_0–step transition probabilities in the condition. This transition density depends on several z_i's, but in order not to overburden the notation we write simply z instead. We can start by establishing exponential mixing of the m_0–step chain $\{x_{jm_0}\}$, and from this trivially obtain mixing of the original chain $\{x_j\}$. To avoid complicated notation we consider in the proofs the case with $m_0 = 1$. In the setting of a DNA sequence as in Jensen (2005) the two-step transition probabilities will suffice, whereas in the setting of a process with a cyclic nature as described in section 3 higher order transitions may be needed. Point (i) of Condition 3 limits the influence of the hidden variable on the observed variable. This condition is needed when studying the mixing properties of the hidden chain conditioned on the observed y–process. Point (ii) of Condition 3 limits the conditional score function based on (x_i, y_i) given x_{i-1}. The true value of the parameter is θ_0.

Condition 2. *There exists $\delta_0 > 0$, a positive integer m_0, and constants $0 < \tau < \sigma < \infty$, such that*

$$\tau \leq p_\theta(x_{m_0}|x_0; z) \leq \sigma \quad \textit{for all } (x_0, x_{m_0}, z) \textit{ and all } |\theta - \theta_0| \leq \delta_0.$$

To state the next condition we introduce some notation. Likelihood quantities for the chain (x_i, y_i) are denoted by ω as follows

$$\omega_i(\theta) = \log[p_\theta(x_i|x_{i-1}; z_i)p_\theta(y_i|x_i; z_i)] \quad \text{and} \quad \omega_i^r(\theta) = \frac{\partial}{\partial\theta_r}\omega_i(\theta).$$

With δ_0, τ, and σ from Condition 2 define

$$\xi(y) = \sup_{x_1, x_2, z, |\theta-\theta_0|\leq\delta_0} \frac{p_\theta(y|x_1; z)}{p_\theta(y|x_2; z)}, \qquad \rho(y) = 1 - \tau^2/(\sigma\xi(y)),$$

and

$$\beta_1 = \inf_{x,z} \int p_{\theta_0}(y|x; z)/\xi(y)\nu(dy).$$

Condition 3. *Let δ_0 be as in Condition 2.*

i) Assume that $\xi(y) < \infty$ for all y and that $\beta_1 > 0$.

ii) Assume that there exists a function $h_0(y)$ with

$$c_0 = \sup_{x,z} \int h_0(y)p_{\theta_0}(y|x; z)\nu(dy) < \infty,$$

such that for all $r = 1, \ldots, d$ and all i,

$$\sup_{x_{i-1}, x_i, z_i, |\theta-\theta_0|\leq\delta_0} |\omega_i^r(\theta)| \leq h_0(y_i).$$

The second part of the conditions relates to the estimating equation. Let $\psi(\theta, \bar{x}, y; z)$ be a function of the parameter θ, a triple $\bar{x}$ of consequtive states, an observed variable y and covariates z. Let $\psi_i(\theta) = \psi(\theta, \bar{x}_i, y_i; z)$, where $\bar{x}_i = (x_{i-1}, x_i, x_{i+1})$. We think of $\sum_{i=1}^n \psi_i(\theta) = 0$ as an estimating equation had both x and y being observed. Having observed y only we use the estimating equation

$$\sum_{i=1}^n E_\theta[\psi_i(\theta)|(1, n)] = 0, \tag{1}$$

where $E_\theta(\cdot|(1, n))$ is the conditional mean given $y_1, \ldots, y_n$. The coordinates of $\psi_i(\theta)$ are denoted by $\psi_i^r(\theta)$, $r = 1, \ldots, d$ and the derivatives of these are $\psi_i^{rs}(\theta) = \frac{\partial}{\partial\theta_s}\psi_i^r(\theta)$.

In Appendix I a recursive formula for evaluating the estimating function on the left hand side of (1) is given. To solve (1) one often uses an EM-type algorithm, that is, $\sum_{i=1}^n E_\theta[\psi_i(\tilde\theta)|(1,n)] = 0$ is solved with respect to $\tilde\theta$, and this defines a new value improving on the old value θ. This *EEE*-algorithm (Expectation–Estimating–Equation) has been considered in Heyde and Morton (1996), Rosen, Jiang and Tanner (2000) and Elashoff and Ryan (2004).

The "observed information" in this setting, that is, the derivative of the left hand side of (1), is given by

$$\begin{aligned} J_n(\theta) &= -\frac{\partial}{\partial\theta} E_\theta\Big[\sum_{i=1}^n \psi_i(\theta)|(1,n)\Big] \\ &= -E_\theta\Big[\sum_{i=1}^n \frac{\partial}{\partial\theta}\psi_i(\theta)|(1,n)\Big] - V_\theta\Big[\sum_{i=1}^n \psi_i(\theta), \sum_{i=1}^n \frac{\partial}{\partial\theta}\omega_i(\theta)|(1,n)\Big]. \end{aligned}$$

This formula corresponds to the formula in Louis (1982) for the maximum likelihood equation. A derivation can be found in Jensen (2005).

Before stating the second part of the conditions it is convenient to introduce a notation for a class of functions satisfying suitable conditions.

Definition 4. *Consider for each i a function $a_i(\theta)$ which also depends on $(\bar x_i, y_i, z)$. We say that these functions belong to class G_k if there exist $\delta_0 > 0$, a function $a_0(y)$ and a finite constant $c_0^k(a)$ such that*

$$|a_i(\theta)| \le a_0(y_i) \quad \textit{for all } (\bar x_i, z) \textit{ and all } |\theta - \theta_0| \le \delta_0,$$

and

$$\sup_{x,z} \int a_0(y)^k p_{\theta_0}(y|x;z)\nu(dy) \le c_0^k(a).$$

If, furthermore, there exist a function $a_1(y)$ and finite constants $c_1(a), c_1^m(a)$ such that

$$|a_i(\theta) - a_i(\theta_0)| \le |\theta - \theta_0| a_1(y_i) \quad \textit{for all } (\bar x_i, z) \textit{ and all } |\theta - \theta_0| \le \delta_0,$$

and

$$\sup_{x,z} \int a_1(y) p_{\theta_0}(y|x;z)\nu(dy) \le c_1(a), \quad \sup_{x,z} \int a_0(y)^m h_0(y) p_{\theta_0}(y|x;z)\nu(dy) \le c_1^m(a),$$

we say that the set of functions belong to class $G_{k,m}$.

Condition 5. *i) For all $r = 1, \dots, d$ the set of functions $\{\psi_i^r(\theta)\}$ belongs to class G_3 and $E_\theta \psi_i(\theta) = 0$.*

ii) For all $r, s = 1, \dots, d$ the set of functions $\{\psi_i^{rs}(\theta)\}$ belongs to class $G_{1,2}$.

iii) For all $r = 1, \dots, d$ the set of functions $\{\omega_i^r(\theta)\}$ belongs to class $G_{1,2}$.

We now formulate the results of this paper.

Theorem 6. *Assumme that Condition 2, Condition 3(i), and Condition 5(i) hold. Define $S_n = \sum_{i=1}^n E_{\theta_0}(\psi_i(\theta_0)|(1,n)/\sqrt{n}$ and assume that the covariates $\{z_i\}$ are such that the variance of S_n converges to a positive definite limit. Then a central limit theorem holds for the normalized sum S_n.*

Theorem 7. *Assume that Condition 2, 3, and 5 hold. Let $\delta_n \to 0$ for $n \to \infty$. Then*

$$E_{\theta_0}\Big\{\sup_{|\theta-\theta_0|\le\delta_n} \frac{1}{n}\big|J_n(\theta) - J_n(\theta_0)\big|\Big\} \to 0.$$

Corollary 8. *Assume that Condition 2, 3, and 5 hold. Assume that the covariates $\{z_i\}$ are such that the variance of $S_n = \sum_{i=1}^n E_{\theta_0}(\psi_i(\theta_0)|(1,n)/\sqrt{n}$ converges to a positive definite limit $V(\theta_0)$, and also $\frac{1}{n}J_n(\theta_0)$ converges to a positive definite limit $I(\theta_0)$. Then there exists a sequence $\hat\theta_n$ solving the estimating equation such that $\hat\theta \to \theta_0$ in probability and $\sqrt{n}(\hat\theta - \theta_0)$ has a limiting normal distribution with mean zero and variance $I(\theta_0)^{-1}V(\theta_0)I(\theta_0)^{-1}$.*

3 Example: cyclic model

In Hansen (2008) a cyclic hidden Markov model is described for the progesterone concentration in cowmilk. The observed process y_j is the measured progesterone concentration in the milk at each milking. The underlying dynamic is described by a stage i_j, a level v_j giving the mean of the observed process, a slope s_j which defines the increase in the level v_j, and a waiting time r_j until the next change of the stage. The stage describes a cyclic nature where $i = 1$ corresponds to a low stage, this is followed by an increasing stage $i = 2$, followed next by a high stage $i = 3$, and ending in a decreasing stage $i = 4$. Below, when $i = 4$ the sum $i + 1$ means the stage 1. The process is controlled by two transition probabilities, $p(r|i;\gamma)$ which is the probability of a new waiting time r at a point in time where the stage changes from $i - 1$ to i and which depends on a parameter γ, and $p(s|r, v, i)$ which is the probability of a new slope s at a point in time where the stage changes from $i - 1$ to i, the present level is v, and the new waiting time is r. Formally, the Markov structure for the hidden variable $x_j = (i_j, r_j, v_j, s_j)$ is given by

$$\left.\begin{aligned} i_{j+1} &= i_j \\ r_{j+1} &= r_j - 1 \\ v_{j+1} &= v_j + s_j \\ s_{j+1} &= s_j \end{aligned}\right\} \quad r_j > 1,$$
$$\left.\begin{aligned} i_{j+1} &= i_j + 1 \\ r_{j+1} &\sim p(\cdot|i_{j+1};\gamma) \\ v_{j+1} &= v_j + s_j \\ s_{j+1} &\sim p(\cdot|r_{j+1}, v_{j+1}, i_{j+1}) \end{aligned}\right\} \quad r_j = 1.$$

Conditionally on the hidden state the observed variable y_j is normally distributed with mean v_j and variance σ^2.

We consider γ and σ^2 to be cow specific parameters with γ allowing for variation in the mean cycle length from cow to cow, and with σ^2 allowing for varying degree of fit of the hidden model. Finally, we consider the case where s_j and v_j belong to compact sets and r_j belongs to a finite set (this is slightly different from the setup in Hansen, 2008).

The full likelihood, having observed both x_j and y_j, $j = 1, \dots, n$, and conditioning on x_0, leads to the likelihood equations

$$\sum_{j=1}^n \big[(y_j - v_j)^2 - \sigma^2\big] = 0,$$

$$\sum_{j=1}^n \Big[\frac{d}{d\gamma}\log p(r_j|i_j;\gamma)\Big]1(i_{j-1} \neq i_j) = 0.$$

We replace the first of these equations with one giving a more robust estimate of σ. Thus we use instead

$$\sum_{j=1}^{n}\big[|y_j - v_j| - \sigma\sqrt{\frac{2}{\pi}}\big] = 0.$$

In relation to our general setup we thus have $\theta = (\sigma, \gamma)$ and

$$\psi_j^1 = |y_j - v_j| - \sigma\sqrt{\frac{2}{\pi}}, \qquad \psi_j^2 = \big[\frac{d}{d\gamma}\log p(r_j|i_j;\gamma)\big]1(i_{j-1} \neq i_j).$$

The derivatives of these with respect to σ and γ are

$$\begin{aligned} \psi_j^{11} &= -\sqrt{\frac{2}{\pi}}, & \psi_j^{12} &= 0, \\ \psi_j^{21} &= 0, & \psi_j^{22} &= \big[\frac{d^2}{d\gamma^2}\log p(r_j|i_j;\gamma)\big]1(i_{j-1} \neq i_j). \end{aligned}$$

Furthermore, we have

$$\omega_j = \begin{cases} \log\big[\varphi(y_j; v_j, \sigma)\big] & i_j = i_{j-1}, \\ \log\big[p(r_j|i_j;\gamma)p(s_j|r_j, v_j, i_j)\varphi(y_j; v_j, \sigma)\big] & i_j \neq i_{j-1}, \end{cases}$$

where $\varphi(y; v, \sigma)$ is the density of a normal distribution with mean v and variance σ^2.

The derivatives of ω_j with respect to σ and γ are

$$\omega_j^1 = \frac{1}{\sigma^3}(y_j - v_j)^2 - \frac{1}{\sigma}, \quad \omega_j^2 = \frac{d}{d\gamma}\log p(r_j|i_j;\gamma)\big]1(i_{j-1} \neq i_j).$$

Condition 2 will hold under mild conditions on the transition densities due to the compactness of the state space. We do not discuss this further here. Since the state space is bounded we make the assumption that the first three derivatives of $p(\cdot|i;\gamma)$ are bounded. For condition 3 i) we find the bound $\xi(y) \leq \exp(b_0 + b_1|y|)$ for suitable constants b_0 and b_1, due to the finiteness of the level v. Then clearly also $\beta_1 > 0$. For condition 3 ii) we can take $h_0(y) = b_0 + b_1|y| + b_2y^2$ for suitable constants b_0, b_1, b_2. Similarly in condition 5 i) we can use a bound on the form $a_0(y) = b_0 + b_1|y|$. For condition 5 ii) $a_0(y)$ and $a_1(y)$ can be taken as constants. And, finally, for condition 5 iii) both $a_0(y)$ and $a_1(y)$ can be bounded by $b_0 + b_1|y| + b_2y^2$ for suitable constants b_0, b_1, b_2.

In conclusion we see that the standard asymptotic results hold for the estimates in a cyclic model as described here and considered (with slight modifications) in Hansen (2008).

4 Mixing

As a first step in the proof of the main results we study the mixing properties of the process. We use throughout Condition 2 with $m_0 = 1$. Our results hold for all $|\theta - \theta_0| \leq \delta_0$, and we skip θ in the notation below. First we state bounds on the transition densities for the hidden chain conditioned on the observed y–process. The lemma has been proved in Jensen and Petersen (1999).

Lemma 9. *Assume Condition 2. Conditioned on the y–process $\{x_n\}$ constitute a Markov chain with*

$$\frac{\tau^2}{\sigma\xi(y_s)} \leq p(x_s|x_{s-1}, x_{s+1}, y; z) \leq \frac{\sigma^2\xi(y_s)}{\tau}.$$

For the original Markov chain (not conditioned on y) we have trivially from Condition 2 that

$$\frac{\tau^2}{\sigma} \leq p(x_{s+1}|x_s, x_{s+2}; z) \leq \frac{\sigma^2}{\tau}. \tag{2}$$

For easy reference we state here a Lemma from Jensen and Petersen (1999) that will be used repeatedly.

Lemma 10. *Assume that ν_1 and ν_2 are dominated by μ and $\nu_1(\mathcal{X}) = \nu_2(\mathcal{X})$. Then for any real valued measurable function h on $\mathcal{X}$ we have*

$$\Big|\int_{\mathcal{X}} h d\nu_1 - \int_{\mathcal{X}} h d\nu_2\Big| \leq \{\sup_x h(x) - \inf_x h(x)\}\{\nu_1(S^+) - \nu_2(S^+)\},$$

where $S^+ = \{d\nu_1/d\mu - d\nu_2/d\mu > 0\}$.

To establish mixing results for both the original hidden Markov chain and for the chain conditioned on the y–process we consider a general Markov chain $\{x_s\}$ satisfying

$$\tau_s \leq p(x_s|x_{s-1}, x_{s+1}) \leq \sigma_s. \tag{3}$$

with $0 < \tau_s < \sigma_s < \infty$ for all s. We start with a result on one-sided and two-sided mixing. To make the notation more transparent we let u_r, for a lower case letter u, denote $x_r = u$, and let A_s, for a upper case letter A, denote $x_s \in A$.

Lemma 11. *Assume (3). Let $r < s < t$ and let $\rho_j = 1 - \tau_j$. Then*

$$\sup_u P(A_s|u_r) - \inf_v P(A_s|v_r) \leq \prod_{j=r+1}^{s} \rho_j,$$

and

$$\sup_{a,b} P(A_s|a_r, b_t) - \inf_{u,v} P(A_s|u_r, v_t) \leq \prod_{j=r+1}^{s} \rho_j + \prod_{j=s}^{t-1} \rho_j.$$

Proof. The proof of the one-sided case is given in Jensen and Petersen (1999) based on Doob (1953, page 198). In Jensen(2005) a similar proof for the two-sided case is indicated. We give here the details of this proof.

Let $r < s < t$. Define, for a fixed set A and a fixed state w, $D(r) = \max_u P(A_s|u_r, w_t)$, $d(r) = \min_u P(A_s|u_r, w_t)$, and $S_r = \{x : p(x_r = x|u_{r-1}, w_t) > p(x_r = x|v_{r-1}, w_t)\}$. Using Lemma 10 in the first inequality below we find

$$\begin{aligned}
&D(r-1) - d(r-1) \\
&\quad = \max_{u,v} [P(A_s|u_{r-1}, w_t) - P(A_s|v_{r-1}, w_t)] \\
&\quad = \max_{u,v} \int P(A_s|\alpha_r, w_t)\, [p(\alpha_r|u_{r-1}, w_t) - p(\alpha_r|v_{r-1}, w_t)]\, \mu(d\alpha) \\
&\quad \leq (D(r) - d(r)) \max_{u,v} [P(S_r|u_{r-1}, w_t) - P(S_r|v_{r-1}, w_t)] \\
&\quad \leq (D(r) - d(r)) \max_{u,v} [1 - P(S_r^c|u_{r-1}, w_t) - P(S_r|v_{r-1}, w_t)] \\
&\quad \leq (D(r) - d(r))\big(1 - \tau_r\big) \\
&\quad = (D(r) - d(r))\rho_r,
\end{aligned}$$

where we used the bound

$$p(x_r|u_{r-1}, w_t) = \int p(x_r|u_{r-1}, v_{r+1})p(v_{r+1}|u_{r-1}, w_t)\mu(dv) \geq \tau_r.$$

Iterating, we obtain

$$\max_{u,v} |P(A_s|u_r, w_t) - P(A_s|v_r, w_t)| \leq \prod_{j=r+1}^{s} \rho_j,$$

A similar argument gives

$$\max_{u,v} |P(A_s|w_r, u_t) - P(A_s|w_r, v_t)| \leq \prod_{j=s}^{t-1} \rho_j.$$

Combining the two latter bounds lead to

$$\begin{aligned}
&\max_{a,b,u,v} |P(A_s|a_r, b_t) - P(A_s|u_r, v_t)| \\
&\quad \leq |P(A_s|a_r, b_t) - P(A_s|u_r, b_t)| + |P(A_s|u_r, b_t) - P(A_s|u_r, v_t)| \\
&\quad \leq \prod_{j=r+1}^{s} \rho_j + \prod_{j=s}^{t-1} \rho_j. \qquad (4)
\end{aligned}$$

□

Lemma 12. *Assume Condition 2. Define $\rho = 1 - \tau^2/\sigma$. For the y-process we have mixing as in Lemma 11 with $\rho_j \equiv \rho$.*

Proof. For the original Markov chain $\{X_n\}$ we have the bounds in Lemma 11 with $\rho_j \equiv \rho$. Letting y_r^j denote $y_r = y^j$ and similarly with x_r^j, we find by using Lemma 10 twice

$$\begin{aligned}
&P(y_s \in A|y_r^1, y_t^1; z) - P(y_s \in A|y_r^2, y_t^2; z) \\
&\quad = \iint P(y_s \in A|x_s; z)p(x_s|x_r, x_t; z)\mu(dx_s) \\
&\qquad \times [p(d(x_r, x_t)|y_r^1, y_t^1; z) - p(d(x_r, x_t)|y_r^2, y_t^2; z)] \\
&\quad \leq \sup_{x_r^1, x_t^1, x_r^2, x_t^2} \Big[\int P(y_s \in A|x_s; z)p(x_s|x_r^1, x_t^1; z)\mu(dx_s) \\
&\qquad - \int P(y_s \in A|x_s; z)p(x_s|x_r^2, x_t^2; z)\mu(dx_s) \Big] \\
&\quad \leq \sup_{x_r^1, x_t^1, x_r^2, x_t^2, B} \big[p(x_s \in B|x_r^1, x_t^1; z) - p(x_s \in B|x_r^2, x_t^2; z) \big] \\
&\quad \leq \rho^{s-r} + \rho^{t-s}.
\end{aligned}$$

□

5 Central limit theorem

In this section we prove Theorem 6. First some notation. Mean values and probabilities are with respect to the true measure corresponding to $\theta = \theta_0$. We do not show θ_0 in the notation. The conditional mean given $(y_s, y_{s+1}, \ldots, y_t)$ is denoted by $E(\cdot|(s,t))$. If, furthermore, we condition on x_s and x_t we use the notation $E(\cdot|[s,t])$. The expression $\prod_{j=s(-u)}^{t} c_j$ is a short hand notation for the expression $\prod_{j=s}^{u-1} c_j + \prod_{j=u+1}^{t} c_j$.

From Götze and Hipp (1983), which deals with Edgeworth expansion, we can extract a central limit theorem suitable for our purpose. We have already seen in Lemma 12 that the observed process is exponentially fast mixing. If w_i is a sequence of random variables with uniformly bounded third absolute moment a central limit theorem holds for the normalized sum under two additional assumptions. The first condition is the standard assumption that the variance of the normalized sum converge. The second condition says that each w_i can for each m be approximated by a function of y_{i-m}^{i+m} introducing an error that is exponentially small in m. To handle this last requirement we have the following lemma.

Lemma 13. *Assume Condition 2 and 3. Let a_i be a function of $(\bar{x}_i, y_i, z)$. Assume that the set $\{a_i\}$ belongs to class G_1. Then*

$$E\left|E\left(a_i\middle|(1,n)\right) - E\left(a_i\middle|(i-l,i+l)\right)\right| \le 4c_0^1(a)(1-\tau^2\beta_1/\sigma)^{l-1},$$

where $i-l$ is replaced by 1 when $i-l<1$ and, similarly, $i+l$ is replaced by n when $i+l>n$.

Proof. For the case $i-l \ge 1$ and $i+l \le n$, one finds using Lemma 10 and Lemma 11 with $\rho_j = \rho(y_j) = 1 - \tau^2/(\sigma\xi(y_j))$ (see Lemma 9) that

$$\begin{aligned}
&\left|E\left(a_i\middle|(1,n)\right) - E\left(a_i\middle|(i-l,i+l)\right)\right| \\
&\quad = \left|\int E\left(a_i\middle|[i-l,i+l]\right)\left\{P\left(d(x_{i-l},x_{i+l})\middle|(1,n)\right)\right.\right. \\
&\qquad\qquad \left.\left. - P\left(d(x_{i-l},x_{i+l})\middle|(i-l,i+l)\right)\right\}\right| \\
&\quad \le 2a_0(y_i)\max_{A,a,b,u,v}|P(\bar{x}_i \in A|a_{i-l},b_{i+l},y;z) - P(\bar{x}_i \in A|u_{i-l},v_{i+l},y;z)| \\
&\quad \le 2a_0(y_i)\prod_{j=i-l+1(-i)}^{i+l-1}\rho(y_j). \qquad (5)
\end{aligned}$$

To bound the mean of this we condition on the x–process and use the conditional independence of the y's given the x's,

$$\begin{aligned}
&E\left|E\left(a_i\middle|(1,n)\right) - E\left(a_i\middle|(i-l,i+l)\right)\right| \\
&\quad \le 2E\Big\{E(a_0(y_i)|x_i)\prod_{j=i-l+1(-i)}^{i+l-1}E(\rho(y_j)|x_j)\Big\} \\
&\quad \le 2c_0^1(a)\prod_{j=i-l+1(-i)}^{i+l-1}\Big(1-\frac{\tau^2}{\sigma}\beta_1\Big) \\
&\quad = 4c_0^1(a)(1-\tau^2\beta_1/\sigma)^{l-1},
\end{aligned}$$

where we have used Assumption 2 and 5. The two cases $i-k<1$ and $i+k>n$ are treated similarly using one-sided mixing. □

Proof of Theorem 6. Since $\{\psi_i^r\}$ are assumed to be of class G_3 the third absolute moments are uniformly bounded. Furthermore, since $G_3 \subseteq G_1$ we can use Lemma 13 with a_i replaced by $\psi_i^r(\theta_0)$. The central limit theorem extracted from Götze and Hipp (1983) is then applicable. □

6 Uniform convergence of "observed information"

As a final step we prove here Theorem 7. In particular then we work under Condition 2.

To show uniform convergence of $\frac{1}{n}J_n(\theta)$ we need to bound the difference between conditional mean values evaluated under θ and under θ_0.

Lemma 14. *Let b^u be a funtion of $\bar{x}_u$ with $|b^u| \leq 1$. Let $s \leq u-2$ and let $t \geq u+2$. For $|\theta - \theta_0| \leq \delta_0$ we have*

$$|E_\theta(b^u|[s,t]) - E_{\theta_0}(b^u|[s,t])| \leq 2d|\theta - \theta_0| \sum_{i=s+1}^{t} h_0(y_i).$$

Proof. This lemma corresponds to Lemma 5 in Jensen (2005) with sums replaced by integrals. The representation of the conditional density of $\bar{x}_u$ given $[s,t]$ is in our case

$$\frac{\int \prod_{i=s+1}^{t} \omega_i(\theta) \prod_{i=s+1}^{u-2} \mu(dx_i) \prod_{i=u+2}^{t} \mu(dx_i)}{\int \prod_{i=s+1}^{t} \omega_i(\theta) \prod_{i=s+1}^{t} \mu(dx_i)},$$

with $\omega_i(\theta) = p_\theta(x_i|x_{i-1}; z_i)p_\theta(y_i|x_i; z_i)$. An interchange of differentiation and integration is possible since the derivative of the integrand is bounded. The details of the proof can be seen in Jensen (2005). □

Lemma 15. *Let b^u be a function of $\bar{x}_u$ with $|b^u| \leq 1$. For $|\theta - \theta_0| \leq \delta_0$ and any integer $l \geq 1$ we have*

$$|E_\theta(b^u|(1,n)) - E_{\theta_0}(b^u|(1,n))| \leq 2d|\theta - \theta_0| \sum_{i=u-l+1}^{u+l} h_0(y_i) + 4 \prod_{j=u-l+1(-u)}^{u+l-1} \rho(y_j).$$

Proof. We can replace $E_\theta(b^u|(1,n))$ by $E_\theta(b^u|[u-l,u+l])$ with an error less than

$$\sup_{x_{u-l},x_{u+l}} E_\theta(b^u|(u-l,u+l),x_{u-l},x_{u+l}) - \inf_{x_{u-l},x_{u+l}} E_\theta(b^u|(u-l,u+l),x_{u-l},x_{u+l}).$$

Combining Lemma 11 and Lemma 10 this gives the bound $2\prod_{j=u-l+1(-u)}^{u+l-1} \rho(y_j)$. We use this for both E_θ and for E_{θ_0}. Finally we use the bound from Lemma 14 for $E_\theta(b^u|[u-l,u+l])$ $- E_{\theta_0}(b^u|[u-l,u+l])$. □

Lemma 16. *Let the functions $a_i(\theta)$ belong to class $G_{1,1}$ and let $\delta_n \to 0$ for $n \to \infty$. Then*

$$\lim_{n\to\infty} E_{\theta_0} \sup_{|\theta-\theta_0|\leq\delta_n} \Big|\frac{1}{n}\sum_{i=1}^{n}\big\{E_\theta(a_i(\theta)|(1,n)) - E_{\theta_0}(a_i(\theta_0)|(1,n))\big\}\Big| = 0$$

Proof. We can replace $E_\theta(a_i(\theta)|(1,n))$ by $E_\theta(a_i(\theta_0)|(1,n))$ with an error bounded by $\delta_n a_1(y_i)$. Next, from Lemma 15, we can replace $E_\theta(a_i(\theta_0)|(1,n))$ with $E_{\theta_0}(a_i(\theta_0)|(1,n))$. Adding together the error terms we need to consider

$$E_{\theta_0}\left\{\frac{1}{n}\sum_{u=1}^{n}\Big[\delta_n a_1(y_u)+a_0(y_u)\Big(2d\delta_n\sum_{i=u-l+1}^{u+l}h_0(y_i)+4\prod_{j=u-l+1(-u)}^{u+l-1}\rho(y_j)\Big)\Big]\right\}.$$

Conditioning first on the hidden process this gives the bound

$$\delta_n c_1(a)+2d\delta_n\big[c_1^1(a)+2lc_0^1(a)c_0\big]+8c_0^1(a)(1-\tau^2\beta_1/\sigma)^{l-1}.$$

If we take $l=\delta_n^{-1/2}$ the last expression tends to zero for $n\to\infty$. □

Lemma 17. *Let the functions $a_i(\theta)$ and $b_j(\theta)$ belong to the class $G_{1,2}$. Then there exist constants q_1,q_2,q_3 such that for any integer $l\geq 1$*

$$\begin{aligned}E_{\theta_0}\sup_{|\theta-\theta_0|\leq\delta}&|V_\theta(a_u(\theta),b_v(\theta)|(1,n))-V_{\theta_0}(a_u(\theta_0),b_v(\theta_0)|(1,n))|\\&\leq d\delta\big[q_1+q_2(|v-u|+6l)\big]+q_3(1-\tau^2\beta_1/\sigma)^{l-1}.\end{aligned}$$

Proof. Let $u\leq v$. The difference between the covariances can be written as the sum of the two terms

$$E_\theta(a_u(\theta)b_v(\theta)|(1,n))-E_{\theta_0}(a_u(\theta_0)b_v(\theta_0)|(1,n))$$

and

$$\begin{aligned}&E_\theta(a_u(\theta)|(1,n))E_\theta(b_v(\theta)|(1,n))-E_{\theta_0}(a_u(\theta_0)|(1,n))E_{\theta_0}(b_v(\theta_0)|(1,n))\\&\quad=E_\theta(a_u(\theta)|(1,n))\{E_\theta(b_v(\theta)|(1,n))-E_{\theta_0}(b_v(\theta_0)|(1,n))\}\\&\qquad+\{E_\theta(a_u(\theta)|(1,n))-E_{\theta_0}(a_u(\theta_0)|(1,n))\}E_{\theta_0}(b_v(\theta_0)|(1,n)).\end{aligned}$$

For each of these terms we apply Lemma 15. For the first term this gives the bound

$$a_0(y_u)b_0(y_v)\Big\{2d\delta\sum_{i=u-l+1}^{v+l}h_0(y_i)+4\prod_{j=u-l+1(-(u:v))}^{v+l-1}\rho(y_j)\Big\}$$

for $|\theta-\theta_0|\leq\delta$. For the second term the bound becomes

$$\begin{aligned}a_0(y_u)b_0(y_v)\Big\{&2d\delta\sum_{i=v-l+1}^{v+l}h_0(y_i)+4\prod_{j=v-l+1(-v)}^{v+l-1}\rho(y_j)\\&+2d\delta\sum_{i=u-l+1}^{u+l}h_0(y_i)+4\prod_{j=u-l+1(-u)}^{u+l-1}\rho(y_j)\Big\}.\end{aligned}$$

We next bound the mean of the sum of these two terms by first bounding the conditional mean given the hidden process $\{x_i\}$. For the case $u\neq v$ we get the bound in the lemma with

$$q_1=4(c_1^1(a)c_0^1(b)+c_1^1(b)c_0^1(a)),\quad q_2=2c_0^1(a)c_0^1(b)c_0,\quad q_3=24c_0^1(a)c_0^1(b),$$

and for the case $u=v$ we get bound in the lemma with

$$q_1=6\sqrt{c_1^2(a)c_1^2(b)},\quad q_2=2\sqrt{c_0^2(a)c_0^2(b)}c_0,\quad q_3=24\sqrt{c_0^2(a)c_0^2(b)}.$$

We use here that $\int a_0(y)b_0(y)p_{\theta_0}(y|x;z)\nu(dy)$ is bounded by $\sqrt{c_0^2(a)c_0^2(b)}$ and, similarly, $\int a_0(y)b_0(y)h_0(y)p_{\theta_0}(y|x;z)\nu(dy)$ is bounded by $\sqrt{c_1^2(a)c_1^2(b)}$. □

Lemma 18. *Let the assumptions be as in Lemma 17. Let $\delta_n \to 0$ for $n \to \infty$. Then*

$$\lim_{n\to\infty} E_{\theta_0}\left\{\sup_{|\theta-\theta_0|\le\delta_n}\left|\frac{1}{n}\sum_{u,v=1}^{n}\big\{V_\theta(a_u(\theta),b_v(\theta)|(1,n)) - V_{\theta_0}(a_u(\theta_0),b_v(\theta_0)|(1,n))\big\}\right|\right\}$$
$$= 0$$

Proof. The mixing result in Lemma 11 for the hidden process conditioned on the observed process gives (for the case $v > u$)

$$|V_\theta(a_u(\theta),b_v(\theta)|(1,n))| \le 4a_0(y_u)b_0(y_v)\prod_{i=u+2}^{v-2}\rho(y_i),$$

see Ibragimov and Linnik (1971, Theorem 17.2.1). Taking the mean of this, by first evaluating the conditional mean given the hidden process, gives the bound

$$4c_0^1(a)c_0^1(b)(1-\tau^2\beta_1/\sigma)^{|v-u|-3}. \tag{6}$$

Consider now a fixed u and the sum over v of the difference between the two covariances. We split this sum into terms with $|u-v| > l$ and terms with $|u-v| \le l$. For the first set we use the bound in (6) for each covariance, and for the second set we use the bound from Lemma 17. This gives the bound

$$\frac{16c_0^1(a)c_0^1(b)}{\tau^2\beta_1/\sigma}(1-\tau^2\beta_1/\sigma)^{l-3} + d\delta_n\big[(2l+1)q_1 + q_2(l(l+1)+6l(2l+1))\big]$$
$$+ q_3(2l+1)(1-\tau^2\beta_1/\sigma)^{l-1}.$$

Taking $l = \delta_n^{-1/4}$ this bound tends to zero as $\delta_n^{1/2}$ and the lemma has been proved. □

Proof of Theorem 7. The theorem follows directly from Lemma 18. □

Appendix I: Recursions

Let us write the traditional recursive filter for the hidden Markov process in terms of the joint density $p(x_k, y_1^k)$ of the state x_k at time k and the observations $y_1^k = (y_1, y_2, \dots, y_k)$. We skip the covariates $\{z_i\}$ from the notation here. The recursion takes the form

$$p(x_{k+1}, y_1^{k+1}) = p(y_{k+1}|x_{k+1})\int p(x_k, y_1^k)p(x_{k+1}|x_k)\mu(dx_k).$$

We next state a similar recursion for the estimating function on the left hand side of (1). Define $a_k(x_k) = E\big(\sum_{i=1}^{k-1}\psi_i|x_k, y_1^k\big)$, where ψ_i is a function of y_i and $\bar{x}_i = (x_{i-1}, x_i, x_{i+1})$.

We then have

$$
\begin{aligned}
a_{k+1}(x_{k+1}) &= E\Big(\sum_{i=1}^{k-1}\psi_i + \psi_k | x_{k+1}, y_1^{k+1}\Big) \\
&= \int \Big\{ E\Big(\sum_{i=1}^{k-1}\psi_i | x_k, x_{k+1}, y_1^{k+1}\Big) + E(\psi_k | x_k, x_{k+1}, y_1^{k+1}) \Big\} \\
&\qquad \times p(x_k | x_{k+1}, y_1^{k+1}) \mu(dx_k) \\
&= \int \Big\{ a_k(x_k) + \int \psi_k p(x_{k-1} | x_k, x_{k+1}, y_1^{k+1}) \mu(dx_{k-1}) \Big\} \\
&\qquad \times \frac{p(x_k, y_1^k) p(x_{k+1}|x_k) p(y_{k+1}|x_{k+1})}{p(x_{k+1}, y_1^{k+1})} \mu(dx_k) \\
&= \int \Big\{ a_k(x_k) + \int \psi_k \frac{p(x_{k-1}, y_1^{k-1}) p(x_k|x_{k-1}) p(y_k|x_k)}{p(x_k, y_1^k)} \mu(dx_{k-1}) \Big\} \\
&\qquad \times \frac{p(x_k, y_1^k) p(x_{k+1}|x_k) p(y_{k+1}|x_{k+1})}{p(x_{k+1}, y_1^{k+1})} \mu(dx_k) \\
&= \int \Big\{ a_k(x_k) p(x_k, y_1^k) + \int \psi_k p(x_{k-1}, y_1^{k-1}) p(x_k|x_{k-1}) p(y_k|x_k) \mu(dx_{k-1}) \Big\} \\
&\qquad \times \frac{p(x_{k+1}|x_k) p(y_{k+1}|x_{k+1})}{p(x_{k+1}, y_1^{k+1})} \mu(dx_k).
\end{aligned}
$$

A similar calculation gives that the estimating function is

$$
\begin{aligned}
E\Big(\sum_{i=1}^{n}\psi_i | y_1^n\Big) &= \Big\{ \int p(x_n, y_1^n) \mu(dx_n) \Big\}^{-1} \int \Big\{ a_n(x_n) p(x_n, y_1^n) + \iint \psi_n \\
&\times p(x_{n-1}, y_1^{n-1}) p(x_n|x_{n-1}) p(y_n|x_n) p(x_{n+1}|x_n) \mu(dx_{n-1}) \mu(dx_{n+1}) \Big\} \mu(dx_n).
\end{aligned}
$$

For the special case where ψ_i depends on y_i and (x_{i-1}, x_i) only, we define instead $\tilde{a}_k(x_k) = E\big(\sum_{i=1}^k \psi_i | x_k, y_1^k\big)$. The recursion becomes

$$
\tilde{a}_{k+1}(x_{k+1}) = \int \{ \tilde{a}_k(x_k) + \psi_k \} \frac{p(x_k, y_1^k) p(x_{k+1}|x_k) p(y_{k+1}|x_{k+1})}{p(x_{k+1}, y_1^{k+1})} \mu(dx_k).
$$

References

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

Bickel, P.J., Ritov, Y. and Rydén, T. (1998). Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, **26**, 1614–1635.

Doob, J.L. (1953). *Stochastic Processes.* Wiley.

Elashoff, M. and Ryan, L. (2004). An em algorithm for estimating equations. *J. Comput. Graph. Statist.*, **13**, 485–465.

Götze, F. and Hipp, C. (1983). Asymptotic expansions for sums of weakly dependent random vectors. *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, **64**, 211–240.

Hansen, J.V., Jensen J.L., Friggens, N.C. and Højsgaard, S. (2008). A state space model exhibiting a cyclic structure with an application to progesterone concentration in cow milk.

Heyde, C. and Morton, R. (1996). Quasi-likilhood and generalizing the em algortihm. *J. Roy. Statist. Soc. B*, **58**, 317–327.

Ibragimov, I.A. and Linnik, Yu.V. (1971). *Independent and stationary sequences of random variables.* Wolters-Noordhoff Series of Monographs and Textbooks on Pure and Applied Mathematics.

Jensen, J.L. (2005). Context dependent DNA evolutionary models. Research Report No. 458, Department of Theoretical Statistics, University of Aarhus.

Jensen, J.L. and Petersen, N.V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Ann. Statist*, **27**, 514–535.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, **44**, 226–233.

Rosen, O., Jiang, W. and Tanner, M. (2000). Mixtures of marginal models. *Biometrika*, **87**, 391–404.

Department of Genetics and Biotechnology and Department of Mathematical Sciences,
University of Aarhus, DK-8000 Aarhus C, Denmark.

E-mail: `jvhansen@imf.au.dk`

Department of Mathematical Sciences, University of Aarhus, DK-8000 Aarhus C,
Denmark.

E-mail: `jlj@imf.au.dk`