J. Appl. Prob. Spec. Vol. **48A,** 295–306 (2011)

NEW FRONTIERS IN APPLIED PROBABILITY

A Festschrift for SØREN ASMUSSEN
Edited by P. GLYNN, T. MIKOSCH and T. ROLSKI

Part 6. Statistics

ASYMPTOTIC NORMALITY OF M-ESTIMATORS IN NONHOMOGENEOUS HIDDEN MARKOV MODELS

JENS LEDET JENSEN, *University of Aarhus*
Department of Mathematical Sciences, University of Aarhus, Ny Munkegade Building 1530, DK-8000 Aarhus C, Denmark. Email address: jlj@imf.au.dk

APPLIED PROBABILITY TRUST
AUGUST 2011

ASYMPTOTIC NORMALITY OF M-ESTIMATORS IN NONHOMOGENEOUS HIDDEN MARKOV MODELS

BY JENS LEDET JENSEN

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.

Keywords: Estimating equation; mixing properties

2010 Mathematics Subject Classification: Primary 62F12
Secondary 62M09

1. Introduction

In a hidden Markov model the observed variables $y_1, \dots, y_n$ are conditionally independent given the values of the hidden variables $x_1, \dots, x_n$, the latter constituting a Markov chain. In this paper we consider the asymptotic normality of a parameter estimate. Contrary to previous research, the Markov chain need not be homogeneous and we consider a class of M-estimators instead of the maximum likelihood estimator.

The class of estimators can be described as follows. Let $\theta \in \mathbb{R}^p$ be the parameter of the model. We start from an estimating function $T_n(\theta) \in \mathbb{R}^p$ based on complete observation, and calculate the conditional mean given the observed variables $y_1, \dots, y_n$, $S_n(\theta) = \mathrm{E}(T_n(\theta) \mid y_1, \dots, y_n)$, to obtain the estimating function of interest. The original function $T_n(\theta)$ is of the form $T_n(\theta) = \sum_{i=1}^n \psi_i(\theta)$, where $\psi_i(\theta) = \psi_i(\theta; \bar{x}_i, y_i)$ depends on the local data $(\bar{x}_i, y_i)$, with $\bar{x}_i = (x_{i-1}, x_i, x_{i+1})$. Thus, the estimating function based on the observables $(y_1, \dots, y_n)$ becomes

$$S_n(\theta) = \sum_{i=1}^n \mathrm{E}_\theta(\psi_i(\theta; \bar{x}_i, y_i) \mid y_1, \dots, y_n). \tag{1}$$

The index i on $\psi_i(\theta)$ allows for the modelling of an inhomogeneous process. For the maximum likelihood estimator, the estimating function becomes the score function and is obtained on taking ψ_i equal to the derivative of the logarithm of the product of the transition density times the emission density of y_i given x_i. The dependency in ψ_i on both x_{i-1} and x_{i+1} allows us to consider estimating equations based on pseudo-likelihood ideas, where we condition on the neighbouring values.

In [12] a situation is considered where the nonhomogeneity of the Markov chain is a natural part of the model. In that paper the evolution of a DNA string is considered. The data consist of two strings where the second has evolved from the first. It is natural to consider the process conditioned on the first string. The hidden variable is the complete evolutionary history for one site along the string, and, because of the conditioning on the initial string, the transition probabilities of the hidden variable are nonhomogeneous. Asymptotic results for the case of a discrete space for both the hidden and the observed variables are given in [12]. In this paper we derive asymptotic normality for more general state spaces.

To prove asymptotic normality of an estimator, we need a central limit theorem for the estimating function and a result on uniform convergence of the derivative of the estimating function. For the maximum likelihood estimator, Baum and Petrie [2] considered the case of discrete state spaces for both the observed and hidden variables, Bickel *et al.* [3] considered a general state space for the observed variable, and Jensen and Petersen [13] allowed for a more general state space for the hidden variable (corresponding roughly to a compact state space). Douc *et al.* [5] extended this to a framework where the observed process conditioned on the hidden variables is autoregressive. In all of these papers the central limit theorem for the score function is obtained by approximating the score function by a stationary martingale increment sequence. Also, the uniform convergence of the observed information is obtained by approximating the information by the average of an ergodic stationary process. The homogeneity of the process, and to some degree also the use of the score function for estimation, are essential to the approach of the abovementioned papers.

In this paper the central limit theorem for the estimating function is based on a general theorem of Götze and Hipp [6] for weakly dependent variables, where homogeneity is not an issue. The uniform convergence of the derivative of the estimating function is obtained in a more direct way, utilizing the mixing properties of the process. In Section 2 we state three assumptions and the main result together with an example illustrating the models under consideration. The first assumption is used in Section 3 for a study of the mixing properties. We use an idea of Douc *et al.* [5] and extend this into a 'two-sided' version, which is of relevance for establishing the central limit theorem for the estimating function. The central limit theorem is derived in Section 4, where we first write down a slight generalization of the result from Götze and Hipp [6]. The second assumption from Section 2 is needed for the central limit theorem and the third assumption comes into play when considering the convergence properties of the derivative of the estimating function in Sections 5 and 6. In the final section we state a general result that explains how the results of Sections 4–6 lead to the main result in Section 2.

The present paper is a rewriting of the report by Hansen and Jensen [7].

2. Notation and main results

The transition densities for the hidden process and the emission densities for the observed process given the hidden process depend on a parameter $\theta \in \mathbb{R}^p$. We do not show this dependency unless needed. The transition density for the Markov chain with respect to a probability measure μ is $p_j(x_j \mid x_{j-1};\theta)$, and the emission density with respect to a measure ν is $g_j(y_j \mid x_j;\theta)$. The dependency on j of these densities allows for the modelling of an inhomogeneous process. We do not make any assumptions on the state spaces for the hidden and the observed variables. The true parameter value is θ_0. We use the following notation for likelihood quantities:

$$\omega_i(\theta) = \log\{p_i(x_i \mid x_{i-1};\theta)g_i(y_i \mid x_i;\theta)\}, \qquad \omega_i^r(\theta) = \frac{\partial}{\partial\theta_r}\omega_i(\theta). \tag{2}$$

When conditioning on $x_r = u$, we simply write u_r. When conditioning on $(y_r, \dots, y_s)$, we write (r, s), and when conditioning on both $(y_r, \dots, y_s)$ and (x_r, x_s), we write $[r, s]$. The triple

https://doi.org/10.1239/jap/1318940472 Published online by Cambridge University Press

(x_{i-1}, x_i, x_{i+1}) is denoted by $\bar{x}_i$. More generally, we denote a consecutive set of variables $(x_r, \dots, x_s)$ by x_r^s.

Finally, we define the function classes C_k and $C_{k,m}$. Let $B(\delta)$ be a closed ball centred at θ_0 with radius δ. Consider a set of functions $a_i = a_i(\bar{x}_i, y_i; \theta)$, $i = 1, \dots, n$. We say that $\{a_i\}$ belongs to the class C_k if there exist functions $a_i^0(y_i)$, a constant $\delta_0 > 0$, and a constant K such that, for all i,

$$\sup_{\bar{x}_i, \theta \in B(\delta_0)} |a_i(\bar{x}_i, y_i; \theta)| \le a_i^0(y_i) \quad \text{and} \quad \mathrm{E}_{\theta_0}(a_i^0(y_i)^k) \le K.$$

Note that, for the case where a_i depends on y_i only, belonging to the class C_k simply means a bound on the kth moment. Furthermore, $\{a_i\}$ belongs to the class $C_{k,m}$ if the set belongs to C_k and there exist functions $\bar{a}_i(y_i)$ and $\delta_0 > 0$ such that, for $\theta \in B(\delta_0)$,

$$|a_i(\bar{x}_i, y_i; \theta) - a_i(\bar{x}_i, y_i; \theta_0)| \le |\theta - \theta_0| \bar{a}_i(y_i) \quad \text{for all } \bar{x}_i, \quad \text{and} \quad \mathrm{E}_{\theta_0}(\bar{a}_i^0(y_i)^m) \le K.$$

We next state the three sets of conditions we need. The first set allows us to study the mixing properties, the second set is used to establish a central limit theorem for the estimating function, and the third set is used to show uniform convergence of the derivative of the estimating function.

Assumption 1. (Mixing.) *There exist* $\delta_0 > 0$ *and* $0 < \sigma_-, \sigma_+ < \infty$ *such that, for* $\theta \in B(\delta_0)$,

$$\sigma_- \le p_j(x_j \mid x_{j-1}; \theta) \le \sigma_+ \quad \textit{for all } j, x_j, x_{j-1}.$$

Furthermore, for all j, y_j *and all* $\theta \in B(\delta_0)$, *we have* $0 < \int g_j(y_j \mid x_j; \theta)\mu(\mathrm{d}x_j) < \infty$.

Assumption 2. (Central limit theorem.) *Assume that the terms of the estimating function are unbiased,* $\mathrm{E}_{\theta_0}(\psi_i(\theta_0)) = 0$ *for all* i, *and that* $\{\psi_i\}$ *is of class* C_3. *Furthermore, there exist constants* $K_0 > 0$ *and* n_0 *such that, for* $n > n_0$,

$$a^\top \operatorname{var}_{\theta_0}(S_n) a \ge n K_0 |a|^2 \quad \textit{for all } a \in \mathbb{R}^p, \tag{3}$$

where $S_n = S_n(\theta_0)$ *is the estimating function from (1).*

Assumption 3. (Uniform convergence.) *Let* $F_n = \mathrm{E}_{\theta_0}(-\partial S_n(\theta_0)/\partial \theta^\top)$. *Assume that there exist* $c_0 > 0$ *and* n_0 *such that, for* $n > n_0$, *the eigenvalues of* $F_n^\top F_n$ *are bounded below by* c_0. *Furthermore, assume that* $\{\psi_i\}$ *and* $\{\omega_i^r\}$, $r = 1, \dots, d$, *are of class* C_4, *and the* $\{\partial \psi_i / \partial \theta_r\}$, $r = 1, \dots, d$, *are of class* $C_{3,1}$.

With the assumptions defined we are in a position to state the main theorem. The proof of the theorem is based on the general result in Section 7 that ties together the results of Sections 4–6.

Theorem 1. *Assume that Assumption 1, Assumption 2, and Assumption 3 hold. Let* $G_n = \operatorname{var}_{\theta_0}(S_n(\theta_0))$. *Then there exists a consistent sequence* $\hat{\theta}_n$, *solving the estimating equation* $S_n(\theta) = 0$, *such that* $\sqrt{n} G_n^{-1/2} F_n(\hat{\theta}_n - \theta_0)$ *has a limiting standard normal distribution under* P_{θ_0}.

Finally, we end this section with an example illustrating the setup.

Example 1. We consider a situation where the observed variables are counts. Thus, conditionally on the hidden variables, we assume that $y_i \mid x_i \sim \mathrm{Poisson}(e_i x_i)$, where the e_is are known covariates. Typically, the e_is reflect some sort of population size or size of sampling window. As a concrete example, we use the counts of clover leaves in 200 windows of size 5 cm $\times$ 5 cm along a line transect from Augustin *et al.* [1]. For this particular example, we have $e_i \equiv 1$. For the hidden variable, we choose the state space $\{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$, and the possible transitions are one step down with probability $\rho(1-\beta)$ and one step up with probability $(1-\rho)(1-\beta)$, except for the first state where this probability is $1-\alpha$. Thus, the model has the three parameters (α, β, ρ), which can be translated into a mean, a variance, and a correlation between two neighbouring observations. In Møller *et al.* [16] the same data are analyzed. In that paper a hidden variable model is considered, but the hidden variables do not constitute a Markov chain. They considered the use of a composite likelihood that in our setting would correspond to the estimating function $\psi_i = (\partial/\partial\theta) \log p(y_i, y_{i-1}; \theta)$ (the dependency on both y_i and y_{i-1} as opposed to ψ_i in (1) has no importance). Since this function does not depend on $\bar{x}_i$, we have $\mathrm{E}(\psi_i \mid (1, n)) = (\partial/\partial\theta) \log p(y_i, y_{i-1}; \theta)$ and the asymptotic analysis is in this specialized case much easier than the general treatment we give in this paper. For our model described above, the maximum likelihood estimates are $(0.045, 0.012, 0.552)$, whereas the estimates using the abovementioned composite likelihood is $(0.279, 0.000, 0.569)$. The maximum aposteriori estimates of the hidden process are almost the same for the two sets of parameter estimates, and resemble those given in [16].

3. Geometric decay of the mixing rate

In this section we use Assumption 1 to establish a bound on the transition density of the conditional Markov chain given the observed process y. In [3] and [13] this bound depends on y and is for the density with respect to μ. The dependency on y necessitates further assumptions on the density $g_j(y_j \mid x_j)$ for the main results of those two papers. Contrary to this, Douc *et al.* [5] established a bound independent of y by considering the transition density with respect to a measure dependent on y. The latter dependency, however, has no influence on the ensuing mixing rates. We follow here the approach of Douc *et al.* [5], except that we also use a two-sided version of the argument. To handle both the original Markov chain and the chain conditioned on the observed process y, in the formulation below we let $g_j(x_j)$ be either identically 1 or the function $g_j(y_j \mid x_j)$.

Lemma 1. *Consider an inhomogeneous Markov chain with joint density*

$$c \prod_{k=1}^{n} p_k(x_k \mid x_{k-1}) g_k(x_k)$$

with respect to $\mu^{\otimes n}$, where c is a normalizing constant. There exist probability measures μ_k such that the transition densities q_k with respect to these satisfy

$$\frac{\sigma_-}{\sigma_+} \le q_k(x_k \mid x_{k-1}) \le \frac{\sigma_+}{\sigma_-}.$$

Also, there exists a probability measure $\tilde{\mu}_k$ such that, when conditioning on both x_{k-1} and x_{k+1}, the conditional density satisfies

$$\left(\frac{\sigma_-}{\sigma_+}\right)^2 \le q_k(x_k \mid x_{k-1}, x_{k+1}) \le \left(\frac{\sigma_+}{\sigma_-}\right)^2.$$

Proof. We formulate the proof through the standard filtering equations for hidden Markov models. Define $a_n(x_n) = 1$, and recursively define

$$a_{k-1}(x_{k-1}) = \int a_k(x_k) p_k(x_k \mid x_{k-1}) g_k(x_k) \mu(\mathrm{d}x_k)$$

for $k = n-1, \dots, 1$. Note that these numbers are bounded from below and above

https://doi.org/10.1239/jap/1318940472 Published online by Cambridge University Press

according to Assumption 1. The transition density with respect to μ can then be written as $p_k(x_k \mid x_{k-1})g_k(x_k)a_k(x_k)/a_{k-1}(x_{k-1})$. Next, let μ_k be the probability measure with density $g_k(x_k)a_k(x_k)/\int g_k(z)a_k(z)\mu(\mathrm{d}z)$ with respect to μ. The transition density with respect to μ_k is $q_k(x_k \mid x_{k-1}) = p_k(x_k \mid x_{k-1})/\int p_k(z \mid x_{k-1})\mu_k(\mathrm{d}z)$, which clearly satisfies the bounds given in the lemma.

Conditioning on both x_{k-1} and x_{k+1}, the density with respect to μ is $\zeta_k(x_k)/\int \zeta_k(z)\mu(\mathrm{d}z)$ with $\zeta_k(x_k) = p_k(x_k \mid x_{k-1})p_{k+1}(x_{k+1} \mid x_k)g_k(x_k)$. Now define the probability measure $\tilde{\mu}_k$ through the density $g_k(x_k)/\int g_k(z)\mu(\mathrm{d}z)$ with respect to μ. Then the conditional density with respect to $\tilde{\mu}_k$ is $\tilde{\zeta}_k(x_k)/\int \tilde{\zeta}_k(z)\tilde{\mu}_k(\mathrm{d}z)$ with $\tilde{\zeta}_k(x_k) = p_k(x_k \mid x_{k-1})p_k(x_{k+1}|x_k)$. Clearly, this conditional density satisfies the bounds given in the lemma.

Corollary 1. *Consider the same inhomogeneous Markov chain as in Lemma 1. Let $r < s$, and let $\rho = 1 - \sigma_-/\sigma_+$. Then, for any subset A,*

$$\sup_u \mathrm{P}(x_s \in A \mid x_r = u) - \inf_v \mathrm{P}(x_s \in A \mid x_r = v) \leq \rho^{s-r}.$$

Let $r < s_1 \leq s_2 < t$, and let $\tilde{\rho} = 1 - (\sigma_-/\sigma_+)^2$. Then, for any subset B,

$$\sup_{a,b} \mathrm{P}(x_{s_1}^{s_2} \in B \mid x_r = a,\ x_t = b) - \inf_{u,v} \mathrm{P}(x_{s_1}^{s_2} \in B \mid x_r = u,\ x_t = v) \leq \tilde{\rho}^{s_1-r} + \tilde{\rho}^{t-s_2}.$$

Proof. The method of proof basically goes back to [4, p. 198] for the one-sided case. Details for the one-sided case are given in [13], and details for the case of a finite state space are given in [2]. Douc *et al.* [5] refered to [14] for the one-sided case. Here we give a proof for the two-sided case using similar ideas.

Let $k < s_1$. Define, for a fixed set B and a fixed state w, $D(k) = \sup_u \mathrm{P}(x_{s_1}^{s_2} \in B \mid u_k, w_t)$ and $d(k) = \inf_u \mathrm{P}(x_{s_1}^{s_2} \in B \mid u_k, w_t)$, and, for fixed u and v, define

$$S_k = \{x_k \colon q_k(x_k \mid u_{k-1}, w_t) > q_k(x_k \mid v_{k-1}, w_t)\},$$

where q_k is the density with respect to $\tilde{\mu}_k$ from Lemma 1 (remember that the notation u_r means conditioning on $x_r = u$). From Lemma 1 we have

$$q_k(x_k \mid u_{k-1}, w_t) = \int q_k(x_k \mid u_{k-1}, v_{k+1})q_{k+1}(v_{k+1} \mid u_{k-1}, w_t)\tilde{\mu}_{k+1}(\mathrm{d}v) \geq \left(\frac{\sigma_-}{\sigma_+}\right)^2.$$

We then find that

$$\begin{aligned}
&D(k-1) - d(k-1)\\
&\quad = \sup_{u,v}[\mathrm{P}(x_{s_1}^{s_2} \in B \mid u_{k-1}, w_t) - \mathrm{P}(x_{s_1}^{s_2} \in B \mid v_{k-1}, w_t)]\\
&\quad = \sup_{u,v} \int \mathrm{P}(x_{s_1}^{s_2} \in B \mid \alpha_k, w_t)[q_k(\alpha_k \mid u_{k-1}, w_t) - q_k(\alpha_k \mid v_{k-1}, w_t)]\tilde{\mu}_k(\mathrm{d}\alpha)\\
&\quad \leq (D(k) - d(k)) \sup_{u,v}[\mathrm{P}(S_k \mid u_{k-1}, w_t) - \mathrm{P}(S_k \mid v_{k-1}, w_t)]\\
&\quad \leq (D(k) - d(k)) \sup_{u,v}[1 - \mathrm{P}(S_k^{\mathrm{c}} \mid u_{k-1}, w_t) - \mathrm{P}(S_k \mid v_{k-1}, w_t)]\\
&\quad \leq (D(k) - d(k))\left(1 - \left(\frac{\sigma_-}{\sigma_+}\right)^2\right)\\
&\quad = (D(k) - d(k))\tilde{\rho}.
\end{aligned}$$

https://doi.org/10.1239/jap/1318940472 Published online by Cambridge University Press

Iterating, for $k = s_1, s_1 - 1, \dots, r+1$, we obtain

$$\sup_{u,v} |\mathrm{P}(x_{s_1}^{s_2} \in B \mid u_r, w_t) - \mathrm{P}(x_{s_1}^{s_2} \in B \mid v_r, w_t)| \leq \prod_{k=r+1}^{s_1} \tilde{\rho} = \tilde{\rho}^{s_1 - r}.$$

A similar argument shows that $\sup_{u,v} |\mathrm{P}(x_{s_1}^{s_2} \in B \mid w_r, u_t) - \mathrm{P}(x_{s_1}^{s_2} \in B \mid w_r, v_t)|$ is bounded by $\tilde{\rho}^{t-s_2}$. Combining the two latter bounds we obtain the result of the corollary.

The mixing statement of Corollary 1 immediately leads to a similar mixing statement for the observed process y.

Corollary 2. *Let $r < s < t$. For any values of y^1, y^2, $\tilde{y}^1$, and $\tilde{y}^2$, and any set B, we have*

$$\sup_{y^1,y^2} \mathrm{P}(y_s \in B \mid y_r = y^1,\ y_t = y^2) - \inf_{\tilde{y}^1,\tilde{y}^2} \mathrm{P}(y_s \in B \mid y_r = \tilde{y}^1,\ y_t = \tilde{y}^2) \leq \tilde{\rho}^{s-r} + \tilde{\rho}^{t-s}.$$

Proof. We use the results of Corollary 1 for the original Markov chain (where $\tilde{\mu}_k = \mu$). Using the structure of the process, we find that

$$\begin{aligned}
&\mathrm{P}(y_s \in B \mid y_r = y^1,\ y_t = y^2) - \mathrm{P}(y_s \in B \mid y_r = \tilde{y}^1,\ y_t = \tilde{y}^2) \\
&\quad = \int\!\!\int \mathrm{P}(y_s \in B \mid x_s) p(x_s \mid x_r, x_t) \mu(\mathrm{d}x_s) \\
&\qquad\qquad \times [\mathrm{P}(\mathrm{d}(x_r, x_t) \mid y_r = y^1,\ y_t = y^2) - \mathrm{P}(\mathrm{d}(x_r, x_t) \mid y_r = \tilde{y}^1,\ y_t = \tilde{y}^2)] \\
&\quad \leq \sup_{a,b,u,v} \biggl[\int \mathrm{P}(y_s \in B \mid x_s) p(x_s \mid a_r, b_t) \mu(\mathrm{d}x_s) \\
&\qquad\qquad - \int \mathrm{P}(y_s \in B \mid x_s) p(x_s \mid u_r, v_t) \mu(\mathrm{d}x_s) \biggr] \\
&\quad \leq \sup_{a,b,u,v,A} [\mathrm{P}(x_s \in A \mid a_r, b_t) - \mathrm{P}(x_s \in A \mid u_r, v_t)] \\
&\quad \leq \tilde{\rho}^{s-r} + \tilde{\rho}^{t-s}.
\end{aligned}$$

4. Central limit theorem

In [6] an Edgeworth expansion for a sum of weakly dependent random variables is derived. From this result we can extract a central limit theorem that suits our needs well. We state here a slightly generalized version of the result. This generalization is indicated in [11] and the proof is obtained by following the detailed proofs in [9] and [10]. The direct result from [6] corresponds to having $\gamma_1 = \gamma_2 = 0$ in (5) below and replacing $\mathrm{dist}(I_1, I_2)^{-\lambda}$ by $\rho^{\mathrm{dist}(I_1,I_2)}$ in that same formula. We first introduce some notation.

The central limit theorem is for the sum of random variables $Z_i \in \mathbb{R}^p$, $i \in \mathbb{Z}$. We make the assumption that there exist $\varepsilon > 0$ and $K_0 < \infty$ such that, for all i,

$$\mathrm{E}(Z_i) = 0 \quad \text{and} \quad \mathrm{E}\,|Z_i|^{2+\varepsilon} \leq K_0. \tag{4}$$

We consider a set of σ-algebras $\mathcal{D}_j$ indexed by $j \in \mathbb{Z}$ and satisfying the following strong mixing property. There exist constants γ_0, γ_1, γ_2, and λ such that, for any index sets I_1 and I_2, and any sets $A_i \in \sigma(\mathcal{D}_j : j \in I_i)$, we have

$$\begin{gathered}
|\mathrm{P}(A_1 \cap A_2) - \mathrm{P}(A_1)\,\mathrm{P}(A_2)| \leq \gamma_0 |I_1|^{\gamma_1} |I_2|^{\gamma_2} \,\mathrm{dist}(I_1, I_2)^{-\lambda} \\
\text{with} \quad \lambda > \gamma_1 + \gamma_2 + \max\left\{ \frac{2+\varepsilon}{\varepsilon}, 1 + \gamma_2, 2 \right\}.
\end{gathered} \tag{5}$$

https://doi.org/10.1239/jap/1318940472 Published online by Cambridge University Press

Here $|I_i|$ is the number of elements in I_i and $\operatorname{dist}(I_1, I_2)$ is the Euclidean distance between the two sets, $\operatorname{dist}(I_1, I_2) = \min\{|j_1 - j_2|, j_1 \in I_1, j_2 \in I_2\}$. For the case where $\operatorname{dist}(I_1, I_2)^{-\lambda}$ is replaced by $\rho^{\operatorname{dist}(I_1, I_2)}$ for some $\rho < 1$, the second part of condition (5) is not relevant. (In the case of a random field, that is, the index of Z_i is $i \in \mathbb{Z}^d$, the lower bound on λ must be multiplied by d.) We do not assume that the random variable Z_j is $\mathcal{D}_j$-measurable. Instead, we assume that, for any j and any $m \in \mathbb{N}$, there exists a random variable $Z_j(m)$ which is $\sigma(\mathcal{D}_k\colon \operatorname{dist}(k, j) \le m)$-measurable, and such that

$$\mathrm{E}\,|Z_j - Z_j(m)| \le K_1 m^{-\lambda} \tag{6}$$

for some constant K_1.

Finally, as in [6], we need to assume that the variance of the sum scales with the number of terms. (In [5] and [13] the corresponding condition appears for the main result on asymptotic normality of the maximum likelihood estimate.) Thus, with $S_n = \sum_{i=1}^n Z_i$ we assume that the variance scales as in (3).

Theorem 2. *Under assumptions (3), (4), (5), and (6), we find, as $n \to \infty$, that the eigenvalues of $(1/n)\operatorname{var}(S_n)$ are bounded and $\operatorname{var}(S_n)^{-1/2} S_n \xrightarrow{\mathrm{D}} N_p(0, I)$.*

We now use this theorem for the estimating function (1). For this, we need Assumption 2, which parallels Assumption (A7) of [5] and Assumption (A4) of [13].

Theorem 3. *Let $S_n = S_n(\theta_0)$. Under Assumption 1 and Assumption 2, we have the conclusions of Theorem 2.*

Proof. To use Theorem 2, we let the σ-algebra $\mathcal{D}_j$ be the one generated by y_j. From Corollary 2, it then follows that the mixing assumption (5) is fulfilled with $\gamma_1 = \gamma_2 = 0$ and with $\operatorname{dist}(I_1, I_2)^{-\lambda}$ replaced by $\tilde{\rho}^{\operatorname{dist}(I_1, I_2)}$.

Letting $Z_i = \mathrm{E}_{\theta_0}(\psi_i(\theta_0) \mid (1, n))$ we have $\mathrm{E}\,|Z_i|^3 \le \mathrm{E}((\psi_i^0)^3)$, which, by Assumption 2, is bounded. The only thing left to check is (6). In the formula below we suppress θ_0. For the cases $i - l \ge 1$ and $i + l \le n$, we obtain, from Corollary 1,

$$\begin{aligned}
&|\mathrm{E}(\psi_i \mid (1, n)) - \mathrm{E}(\psi_i \mid (i-l, i+l))| \\
&\quad = \left| \int \mathrm{E}(\psi_i \mid [i-l, i+l])\{\mathrm{P}(\mathrm{d}(x_{i-l}, x_{i+l}) \mid (1, n)) - \mathrm{P}(\mathrm{d}(x_{i-l}, x_{i+l}) \mid (i-l, i+l))\} \right| \\
&\quad \le 2\psi_i^0 \sup_{A,a,b,u,v} |\mathrm{P}(\bar{x}_i \in A \mid a_{i-l}, b_{i+l}, (1, n)) - \mathrm{P}(\bar{x}_i \in A \mid u_{i-l}, v_{i+l}, (1, n))| \\
&\quad \le 2\psi_i^0 \{2\tilde{\rho}^{l-1}\}.
\end{aligned} \tag{7}$$

Taking the mean value we see from Assumption 2 that this is bounded by $4q_3^{1/3}\tilde{\rho}^{l-1}$, where q_3 is an upper bound on the third moment of ψ_i^0. Thus, (6) is proved. The two cases $i - l < 1$ and $i + l > n$ are treated similarly using one-sided mixing.

5. Uniform convergence of the 'observed information'

Throughout this section, we work under Assumption 1.

By the observed information $J_n(\theta)$ we refer in our setting to minus the derivative of the estimating function $S_n(\theta)$ from (1). We can write the observed information as

$$J_n(\theta) = -\mathrm{E}_\theta\left(\sum_{i=1}^n \frac{\partial}{\partial\theta}\psi_i(\theta) \,\middle|\, (1, n) \right) - \operatorname{cov}_\theta\left(\sum_{i=1}^n \psi_i(\theta), \sum_{i=1}^n \frac{\partial}{\partial\theta}\omega_i(\theta) \,\middle|\, (1, n) \right). \tag{8}$$

https://doi.org/10.1239/jap/1318940472 Published online by Cambridge University Press

This formula corresponds to the formula in [15] for the maximum likelihood equation. A derivation can be found in [12].

To show uniform convergence of $(1/n)J_n(\theta)$, we need to bound the difference between conditional mean values evaluated under θ and under θ_0. For the next lemma, we define

$$l_s^t(\theta) = \sum_{i=s+1}^{t} \omega_i(\theta) \quad \text{and} \quad h_i(y_i) = \sup_{x_{i-1}, x_i, \theta \in B(\delta_0), r} |\omega_i^r(\theta)|,$$

where ω_i and ω_i^r are defined in (2).

Lemma 2. *Let b_r^s be a function of x_r^s with $|b_r^s| \leq 1$. For $\theta \in B(\delta_0)$ and any integer $l \geq 0$, we have*

$$|\mathrm{E}_\theta(b_r^s \mid (1,n)) - \mathrm{E}_{\theta_0}(b_r^s \mid (1,n))| \leq 2p|\theta - \theta_0| \sum_{i=r-l+1}^{s+l} h_i(y_i) + 8\tilde{\rho}^l.$$

Proof. We can replace $\mathrm{E}_\theta(b_r^s \mid (1,n))$ by $\mathrm{E}_\theta(b_r^s \mid [r-l, s+l])$ with an error of less than

$$\sup_{x_{r-l}, x_{s+l}} \mathrm{E}_\theta(b_r^s \mid (r-l, s+l), x_{r-l}, x_{s+l}) - \inf_{x_{r-l}, x_{s+l}} \mathrm{E}_\theta(b_r^s \mid (r-l, s+l), x_{r-l}, x_{s+l}).$$

Since $\sup b_r^s - \inf b_r^s \leq 2$, this expression is, from Corollary 1, bounded by $2 \cdot 2\tilde{\rho}^l$. We use this for both E_θ and E_{θ_0}.

We thus need to bound $\mathrm{E}_\theta(b_r^s \mid [r-l, s+l]) - \mathrm{E}_{\theta_0}(b_r^s \mid [r-l, s+l])$. For this, we show the more general statement that

$$|\mathrm{E}_\theta(b \mid [s,t]) - \mathrm{E}_{\theta_0}(b \mid [s,t])| \leq 2p|\theta - \theta_0| \sum_{i=s+1}^{t} h_i(y_i), \tag{9}$$

where $t > s+1$ and b is a function of x_s^t with $|b| \leq 1$. When the bound on the right-hand side is finite, the interchange of integration and differentiation below is valid. We write the conditional mean as

$$\mathrm{E}_\theta(b \mid [s,t]) = \frac{\int b \exp\{l_s^t(\theta)\} \mu(\mathrm{d}x_{s+1}^{t-1})}{\int \exp\{l_s^t(\theta)\} \mu(\mathrm{d}x_{s+1}^{t-1})}.$$

The derivative of the numerator with respect to θ_r is bounded by

$$\left| \int b \sum_{i=s+1}^{t} \omega_i^r(\theta) \exp\{l_s^t(\theta)\} \mu(\mathrm{d}x_{s+1}^{t-1}) \right| \leq \left(\sum_{i=s+1}^{t} h_i(y_i) \right) \int \exp\{l_s^t(\theta)\} \mu(\mathrm{d}x_{s+1}^{t-1}),$$

and this bound can also be used for the derivative of the denominator. Using this, the derivative of the conditional mean with respect to θ_r is bounded by $2\sum_{i=s+1}^{t} h_i(y_i)$. Finally, we write the difference of the conditional means at θ and θ_0 as the integral $\int_0^1 (\mathrm{d}/\mathrm{d}t)\, \mathrm{E}_{\theta_0 + t(\theta - \theta_0)}(b \mid [s,t])\mathrm{d}t$. This then gives (9).

Proposition 1. *Let the functions $\{a_i\}$ belong to the class $C_{2,1}$, and let the functions $\{h_i\}$ belong to the class C_2. For any sequence δ_n tending to 0 as $n \to \infty$, we have*

$$\lim_{n\to\infty} \mathrm{E}_{\theta_0}\left(\sup_{\theta \in B(\delta_n)} \left| \frac{1}{n} \sum_{i=1}^{n} \{\mathrm{E}_\theta(a_i(\theta) \mid (1,n)) - \mathrm{E}_{\theta_0}(a_i(\theta_0) \mid (1,n))\} \right| \right) = 0.$$

Proof. We can replace $\mathrm{E}_\theta(a_i(\theta) \mid (1,n))$ by $\mathrm{E}_\theta(a_i(\theta_0) \mid (1,n))$ with an error bounded by $\delta_n \bar{a}_i(y_i)$. Next, Lemma 2 gives an upper bound when replacing $\mathrm{E}_\theta(a_i(\theta_0) \mid (1,n))$ by $\mathrm{E}_{\theta_0}(a_i(\theta_0) \mid (1,n))$. Adding together the error terms we need to consider

$$\mathrm{E}_{\theta_0}\biggl(\frac{1}{n}\sum_{u=1}^{n}\biggl[\delta_n \bar{a}_u(y_u) + a_u^0(y_u)\biggl(2p\delta_n \sum_{i=u-l}^{u+1+l} h_i(y_i) + 8\tilde{\rho}^l\biggr)\biggr]\biggr).$$

From the moment assumptions, $\mathrm{E}_{\theta_0}\, \bar{a}_u(y_u) \le K$, $\mathrm{E}_{\theta_0}\, a_i^0(y_u) \le K$, and $\mathrm{E}_{\theta_0}\, a_u^0(y_u)h_i(y_i) \le \sqrt{KK} = K$ for some constant K. The bound then becomes

$$\delta_n K + 2p\delta_n(2l+2)K + 8\tilde{\rho}^l K.$$

If we take $l = \delta_n^{-1/2}$, this last expression tends to 0 for $n \to \infty$.

Lemma 3. *Let the functions $\{a_i\}$ and $\{b_i\}$ belong to the class $C_{2,1}$, and let the functions $\{h_i\}$ belong to the class C_3. Then there exist constants q_2 and q_3 such that, for any integer $l \ge 0$,*

$$\begin{aligned}\mathrm{E}_{\theta_0}\Bigl(\sup_{|\theta-\theta_0|\le\delta} &|\operatorname{cov}_\theta(a_u(\theta), b_v(\theta) \mid (1,n)) - \operatorname{cov}_{\theta_0}(a_u(\theta_0), b_v(\theta_0) \mid (1,n))|\Bigr)\\ &\le \delta\{2q_2 + 2pq_3[|v-u| + 3(2l+2)]\} + 24q_2\tilde{\rho}^l.\end{aligned}$$

Proof. The difference $\operatorname{cov}_\theta(a_u(\theta), b_v(\theta) \mid (1,n)) - \operatorname{cov}_\theta(a_u(\theta_0), b_v(\theta_0) \mid (1,n))$ is bounded by $\delta[\bar{a}_u b_v^0 + a_u^0 \bar{b}_v]$, and the mean of this is bounded by $2\delta q_2$, where q_2 is an upper bound on the second moments of the terms involved.

Next, let a^u and b^v be the respective functions evaluated at θ_0. The difference

$$\mathrm{E}_\theta(a^u b^v \mid (1,n)) - \mathrm{E}_{\theta_0}(a^u b^v \mid (1,n))$$

is, from Lemma 2, bounded by $a_u^0 b_u^0[2p\delta \sum_{i=u-l}^{v+1+l} h_i(y_i) + 8\tilde{\rho}^l]$ for any $l \ge 0$. Similarly, the difference

$$\mathrm{E}_\theta(a^u \mid (1,n))\, \mathrm{E}_\theta(b^v \mid (1,n)) - \mathrm{E}_{\theta_0}(a^u \mid (1,n))\, \mathrm{E}_{\theta_0}(b^v \mid (1,n))$$

is bounded by $a_u^0 b_u^0[2p\delta(\sum_{i=u-l}^{u+1+l} h_i(y_i) + \sum_{i=v-l}^{v+1+l} h_i(y_i) + 16\tilde{\rho}^l]$ for any $l \ge 0$. Combining the latter two bounds and taking the mean value, we obtain the bound $2p\delta q_3[|v-u| + 3(2l+2)] + 24q_2\tilde{\rho}^l$ for the difference of the covariance evaluated under θ and under θ_0. Here q_j is an upper bound on the jth moments of the terms involved.

Combining all the bounds, completes the proof.

Proposition 2. *Suppose that the assumptions in Lemma 3 hold. Let $\delta_n \to 0$ for $n \to \infty$. Then*

$$\begin{aligned}\lim_{n\to\infty} \mathrm{E}_{\theta_0}\biggl\{\sup_{|\theta-\theta_0|\le\delta_n}\biggl|\frac{1}{n}\sum_{u,v=1}^{n}\{\operatorname{cov}_\theta(a_u(\theta), b_v(\theta) \mid (1,n)) - \operatorname{cov}_{\theta_0}(a_u(\theta_0), b_v(\theta_0) \mid (1,n))\}\biggr|\biggr\}\\ = 0.\end{aligned}$$

Proof. The mixing result in Corollary 1 for the hidden process conditioned on the observed process gives

$$|\operatorname{cov}_\theta(a_u(\theta), b_v(\theta) \mid (1,n))| \le 4a_u^0 b_v^0 \rho^{|v-u|-3}; \tag{10}$$

see Theorem 17.2.1 of [8]. Taking the mean of this gives the bound $4q_2\rho^{|v-u|-3}$.

Now consider 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 above for each covariance, and, for the second set, we use the bound from Lemma 3. This gives the final bound

$$8q_2\frac{\rho^{l-2}}{1-\rho}+\delta_n\{2q_2(2l+1)+2pq_3[l(l+1)+3(2l+2)(2l+1)]\}+24q_2(2l+1)\tilde{\rho}^l.$$

Taking $l=\delta_n^{-1/4}$, this bound tends to 0 as $\delta_n^{1/2}$, completing the proof.

6. Nonrandom limit of the 'observed information'

Throughout this section, we work under Assumption 1. We show that the derivative $(1/n)J_n(\theta_0)$ of the estimating equation has a nonrandom limit, that is, we show that the limiting variance of the entries of this matrix is 0. We consider first the conditional mean value part of J_n in (8).

Lemma 4. *Let the functions $\{a_i\}$ belong to the class C_3. As $n\to\infty$, the variance of $(1/n)\sum_{u=1}^n \mathrm{E}_{\theta_0}(a_u \mid (1,n))$ is of order $O(1/n)$.*

Proof. From the argument used in (7) we have $|\mathrm{E}(a_u\mid(1,n))-\mathrm{E}(a_u\mid(u-l,u+l))|\le 4a_u^0\tilde{\rho}^{l-1}$. This gives

$$\begin{aligned}&\operatorname{cov}(\mathrm{E}(a_u\mid(1,n)),\mathrm{E}(a_v\mid(1,n)))\\&\qquad=\operatorname{cov}(\mathrm{E}(a_u\mid(u-l,u+l)),\mathrm{E}(a_v\mid(u-l,u+l)))+O(q_2\tilde{\rho}^l),\end{aligned}$$

where q_2 is an upper bound for the second moment of a_u^0. Using the mixing of the observed process and Theorem 17.2.2 of [8], we find that the latter covariance in the above expression is of order $O(q_3[\tilde{\rho}^{1/3}]^{\max\{0,|v-u|-2l\}})$. Taking $l=|v-u|/4$, we find that $\sum_{u,v=1}^n\operatorname{cov}(\mathrm{E}(a_u\mid(1,n)),\mathrm{E}(a_v\mid(1,n)))$ is of order n.

Lemma 5. *Let the functions $\{a_i\}$ and $\{b_i\}$ belong to the class C_4. As $n\to\infty$, the variance of $(1/n)\sum_{u,v=1}^n\operatorname{cov}_{\theta_0}(a_u,b_u\mid(1,n))$ tends to 0.*

Proof. The proof parallels that of Lemma 4, although the details are more complicated.

Let $\xi_u=\sum_{v=1}^n\operatorname{cov}(a_u,b_v\mid(1,n))$, and let ξ_u^l be the same expression with the sum being over $v=u-l$ to $v=u+l$. Using (10), we find that the difference $\xi_u-\xi_u^l$ is of order $a_u^0\rho^l\sum_{k=0}^\infty(b_{u+l+k}^0+b_{u-l-k}^0)$. This in turn implies that the difference $\operatorname{cov}(\xi_u,\xi_z)-\operatorname{cov}(\xi_u^l,\xi_z^l)$ is of order $q_4l\rho^l$, where q_4 is an upper bound on the fourth moments of a_u^0 and b_v^0.

Using the argument behind (7), we can show that the difference $\operatorname{cov}(a_u,b_v\mid(1,n))-\operatorname{cov}(a_u,b_v\mid(u-l,v+l))$ is of order $a_u^0b_v^0\tilde{\rho}^l$. Let $\tilde{\xi}_u^l$ be ξ_u^l, where each covariance term is replaced by $\operatorname{cov}(a_u,b_v\mid(u-l,v+l))$. Then the difference $\operatorname{cov}(\xi_u^l,\xi_z^l)-\operatorname{cov}(\tilde{\xi}_u^l,\tilde{\xi}_z^l)$ is of order $q_4l\tilde{\rho}^l$.

Using (10), we see that $\tilde{\xi}_u^l$ is bounded by $a_u^0\sum_{v=u-l}^{u+l}b_v^0\tilde{\rho}^{|v-u|}$. Using Hölders inequality, the third moment of $\tilde{\xi}_u^l$ can be bounded by a term of order q_4l^3. Finally, we use [8, Theorem 17.2.2] to bound $\operatorname{cov}(\tilde{\xi}_u^l,\tilde{\xi}_z^l)$ by a term of order $q_4l^3[\tilde{\rho}^{1/3}]^{\max\{0,|v-u|-4l\}}$. Combining all the above estimates, we find that $\sum_{z=1}^n\operatorname{cov}(\xi_u,\xi_v)=O(l^4+nl\tilde{\rho}^l)$. Taking $l=n^{1/8}$, we obtain the result of the lemma.

7. Asymptotics for estimators from estimating equations

We state here a general theorem that directly gives the result of Theorem 1 on combining the results of the previous sections. The proof is based on the method outlined in [17]. We consider a general situation with an estimating function $S_n(\theta)$ with minus the derivative given by $J_n(\theta)$. We define

$$\gamma(n, \delta) = \sup_{\theta \in B(\delta)} \left| \frac{1}{n}(J_n(\theta) - J_n(\theta_0)) \right|.$$

Theorem 4. *Below, probability statements are with respect to the true measure* P_{θ_0}*. Assume that*

(i) *there exist a constant* $c_0 > 0$ *and nonrandom matrices* F_n*, with eigenvalues of* $F_n^\top F_n$ *bounded below by* c_0*, such that* $J_n(\theta_0)/n - F_n \xrightarrow{\mathrm{P}} 0$*;*

(ii) *there exist constants* $0 < c_1 < c_2 < \infty$ *and nonrandom positive definite matrices* G_n*, with eigenvalues between* c_1 *and* c_2*, such that* $(1/\sqrt{n})G_n^{-1/2}S_n(\theta_0) \xrightarrow{\mathrm{D}} N_p(0, I)$*.*

Assume further that $\gamma(n, c/\sqrt{n}) \xrightarrow{\mathrm{P}} 0$ *for any* $c > 0$*. Then*

$$\lim_{c \to \infty} \liminf_{n \to \infty} \mathrm{P}_{\theta_0}\left(\textit{there exists } \hat{\theta}_n \in B\left(\frac{c}{\sqrt{n}} \right) \right) = 1,$$

and, for such an estimate, $\sqrt{n}G_n^{-1/2}F_n(\hat{\theta}_n - \theta_0) \xrightarrow{\mathrm{D}} N_p(0, I)$*. Under the stronger assumption that* $\gamma(n, \delta_n) \xrightarrow{\mathrm{P}} 0$ *for any sequence* $\delta_n \to 0$*, we have* $\sqrt{n}G_n^{-1/2}F_n(\hat{\theta}_n - \theta_0) \xrightarrow{\mathrm{D}} N_p(0, I)$ *for any consistent solution* $\hat{\theta}_n$ *to the estimating equation.*

Acknowledgement

I thank Jan Pedersen for fruitful discussions.

References

[1] Augustin, N. H., McNicol, J. and Marriott, C. A. (2006). Using the truncated auto-Poisson model for spatially correlated counts of vegetation. *J. Agricult. Biol. Environ. Statist.* **11,** 1–23.
[2] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **37,** 1554–1563.
[3] 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.
[4] Doob, J. L. (1953). *Stochastic Processes.* John Wiley, New York.
[5] Douc, R., Moulines, É. and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.* **32,** 2254–2304.
[6] Götze, F. and Hipp, C. (1983). Asymptotic expansions for sums of weakly dependent random vectors. *Z. Wahrscheinlichkeitsth.* **64,** 211–239.
[7] Hansen, J. V. and Jensen, J. L. (2008). Asymptotics for estimating equations in hidden Markov models. Thiele Res. Rep., No. 7, Department of Mathematical Sciences, University of Aarhus.
[8] Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables.* Wolters-Nordhoff, Groningen.
[9] Jensen, J. L. (1986). A note on the work of Götze and Hipp concerning asymptotic expansions for sums of weakly dependent random vectors. Memoir, No. 10, Department of Theoretical Statistics, University of Aarhus.
[10] Jensen, J. L. (1988). A note on the work of Götze and Hipp concerning asymptotic expansions for sums of weakly dependent random vectors. In *Proc. 4th Prague Symp. Asymptotic Statistics,* eds P. Mandel and M. Huskova, Charles University, Prague, pp. 295–303.
[11] Jensen, J. L. (1993). A note on asymptotic expansions for sums over a weakly dependent random field with application to the Poisson and Strauss processes. *Ann. Inst. Statist. Math.* **45,** 353–360.
[12] Jensen, J. L. (2005). Context dependent DNA evolutionary models. Res. Rep., No. 458, Department of Mathematical Sciences, University of Aarhus.

[13] JENSEN, J. L. AND PETERSEN, N. V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Ann. Statist.* **27,** 514–535.
[14] LINDVALL, T. (1992). *Lectures on the Coupling Method.* John Wiley, New York.
[15] LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. Ser. B* **44,** 226–233.
[16] MØLLER, J., MCCULLAGH, P. AND RUBAK, E. (2008). Statistical inference for a class of multivariate negative binomial distributions. Unpublished manuscript, Department of Mathematical Sciences, Aalborg University.
[17] SWEETING, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Ann. Statist.* **8,** 1375–1381.

JENS LEDET JENSEN, *University of Aarhus*

Department of Mathematical Sciences, University of Aarhus, Ny Munkegade Building 1530, DK-8000 Aarhus C, Denmark. Email address: jlj@imf.au.dk

https://doi.org/10.1239/jap/1318940472 Published online by Cambridge University Press