Hidden Markov model likelihoods and their derivatives behave like i.i.d. ones : Details. [*]

Peter J. Bickel[†] Ya'acov Ritov[‡] Tobias Rydén[§]

Abstract

We consider the log-likelihood function of hidden Markov models, its derivatives and expectations of these (such as different information functions). We give explicit expressions for these functions and bound them as the size of the chain increases. We apply our bounds to obtain partial second order asymptotics and some qualitative properties of a special model as well as to extend some results of Petrie's (1969).

Key words and phrases: Hidden Markov model, incomplete data, missing data, asymptotic normality.

AMS 1991 classification: 62M09

1 Introduction

In two previous papers, Bickel and Ritov (1996) and Bickel, Ritov and Rydén (1998), we developed the theory of inference for hidden Markov models (HMMs) when the state space of the chain is finite but the hiding mechanism is not finitely-valued. We recall the definition of these models. An HMM is a discrete-time stochastic process $\{(X_t, Y_t)\}$ such that (i) $\{X_t\}$ is a Markov chain, and (ii) given $\{X_t\}$, $\{Y_t\}$ is a sequence of conditionally independent random variables with the conditional distribution of Y_n depending on $\{X_t\}$ only through X_n. The distribution of $\{Y_t\}$ is assumed to depend smoothly on a Euclidean parameter θ. Equivalently an HMM can be thought of as $Y_n = h(X_n, \varepsilon_n)$, where $\{\varepsilon_t\}$ is an i.i.d. sequence independent of $\{X_t\}$, that is a stochastic function of X_n. The parameter mentioned before labels both the transition probability matrix of the chain, the function h and the distribution of $\{\varepsilon_t\}$, although in principle the latter can be taken as fixed,

[*] The research of the first two authors was partially supported by NSF grant DMS-9115577 and by US/Israel Bi-National Science Foundation grant 90-00031/2. The third author was supported by a Swedish Research Council grant.

[†] Department of Statistics, University of California, Evans Hall, Berkeley, CA 94720, USA

[‡] Department of Statistics, The Hebrew University, Jerusalem 91905, Israel

[§] Centre for Mathematical Sciences, Lund University, Box 118, S-221 00 Lund. Sweden

for example uniform on $(0,1)$. In this generality, HMMs include state space models, cf. Kalman (1977). We, in this paper as before, restrict ourselves to the case where the state space of the $\{X_t\}$ is finite. HMMs have during the last decade become wide spread for modeling sequences of weakly dependent random variables, with applications in areas like speech processing (Rabiner, 1989), neurophysiology (Freaking and Rice, 1992), and biology (Leroux and Puterman, 1992). See also the monograph by MacDonald and Zucchini (1997).

Inference for HMMs was first considered by Baum and Petrie, who treated the case when $\{Y_t\}$ takes values in a finite set. In Baum and Petrie (1966), results on consistency and asymptotic normality of the maximum-likelihood estimator (MLE) are given, and the conditions for consistency are weakened in Petrie (1969). Baum and Petrie and particularly Petrie also studied the structure of the map $\vartheta \to E_\vartheta \log p_\vartheta(Y_1|Y_0, Y_{-1}, \ldots)$, the conditional limiting entropy per observation.

The results of Bickel et al. (1998) depend critically on bounds on derivatives of the log-likelihood of the observations. Specifically, we showed that under Cramér-type conditions at ϑ_0,

$$\sup\left\{\frac{\partial^k}{\partial\vartheta^k}\log p_\vartheta(Y_1,\ldots,Y_n) \,:\, |\vartheta-\vartheta_0| \le \varepsilon\right\} \le M_k(Y_1,\ldots,Y_n), \tag{1}$$

where $E_{\vartheta_0}|M_k(Y_1,\ldots,Y_n)| \le C < \infty$ for $k = 1, 2$, as well as quasi-continuity of the map $\vartheta \mapsto (\partial^2/\partial\vartheta^2)\log p_\vartheta(Y_1,\ldots,Y_n)$ at ϑ_0. The same type of bounds were applied to state space models by Jensen and Petersen (1999).

Similar bounds were obtained and used in the case where Y is finitely supported by Baum and Petrie (1966) and Petrie (1969). They were actually able to show that if ϑ is the transition probability matrix of the Markov chain (and when Y is finitely supported this is the most general model) and $\Theta = \{\vartheta \,:\, \vartheta_{ij} \ge \delta > 0, \forall i, j\}$, then the map $\vartheta \mapsto E_{\vartheta_0}\log p_\vartheta(Y_1|Y_0, Y_{-1}, \ldots)$ has a convergent series expansion everywhere on Θ.

Our technical goals are threefold:

(i) To exhibit bounds on the derivatives of the form (1) and to show that, under some conditions, M_k grows no faster than $nC^k k!$. We shall derive these bounds by a unified argument relying on results of Saulis and Statulevičius (1991) (henceforth referred to as S&S).

(ii) To obtain bounds on

$$\sup\left\{\frac{\partial^k}{\partial\vartheta^k}E_\vartheta\left(\frac{\partial^m}{\partial\vartheta^m}\log p_\vartheta(Y_1|Y_0, Y_{-1}, \ldots)\right)\right\}.$$

We give conditions under which this expression is bounded by $C^{k+m}(k+m)!$.

(iii) To establish bounds on

$$\left| \frac{\partial^k}{\partial \vartheta^k} \Big(\log p_\vartheta(Y_1|Y_0, Y_{-1}, \ldots, Y_t) - \log p_\vartheta(Y_1|Y_0, Y_{-1}, \ldots, Y_s) \Big) \right|$$

of the form $C\rho^{|t-s|}$ for $\rho < 1$.

Using these results, we shall, under suitable conditions

(a) Show how to establish stochastic asymptotic expansions for the MLE in terms of derivatives of the loglikelihood at ϑ_0. We sketch how in conjunction with Edgeworth type expansions for sums of functions of Markov chains, these establish the validity of procedures such as debiasing the MLE and other second order methods as available in the i.i.d. case.

(b) Show that the following functions are analytic. The Fisher information

$$I(\vartheta) = -E_\vartheta \left\{ \frac{\partial^2}{\partial \vartheta^2} \log p_\vartheta(Y_1|Y_0, Y_{-1}, \ldots) \right\},$$

the Kullback-Leibler distance

$$K(\vartheta) = E_\vartheta \left\{ \log \frac{p_{\vartheta_0}}{p_\vartheta}(Y_1|Y_0, Y_{-1}, \ldots) \right\}$$

and the entropy

$$H(\vartheta) = -E_\vartheta \{ \log p_\vartheta(Y_1|Y_0, Y_{-1}, \ldots) \}.$$

(c) Study the behavior of these functions and their derivatives at points ϑ_0 under which the X's are i.i.d. (the transition probability matrix is degenerate). We show that at such points, in principle, these quantities can be computed explicitly.

(d) Show how to use these expansions qualitatively to guess properties of $I(\vartheta)$ which can be established in other ways.

2 Assumptions and main results

We observe $\mathcal{Y}$-valued random variables $Y_1, \ldots, Y_n$, where $\mathcal{Y}$ is a general space, distributed as follows. We let $\{X_t\}_{t=1}^n$ be a stationary Markov chain with state space $\{1, 2, \ldots, R\}$ and transition probability matrix $A_\vartheta = \{\alpha_\vartheta(\cdot, \cdot)\}$. Then, $Y_1, \ldots, Y_n$ are conditionally independent given $X_1, \ldots, X_n$ and the conditional distribution of Y_t depends on X_t only. We assume that these distributions have densities $g_\vartheta(y|x)$, where $y \in \mathcal{Y}$ and $x \in \{1, \ldots, R\}$, with respect to some common σ-finite measure on $\mathcal{Y}$. Thus, given $X_t = i$, Y_t has conditional density $g_\vartheta(\cdot|i)$. We assume that $\{X_t\}$ is ergodic so that the stationary distribution exists and is unique. We denote it by $\pi_\vartheta(\cdot)$. The parameter ϑ lies in $\vartheta \in \Theta \subseteq \mathbf{R}^d$, where Θ is open. All computations will be done under a particular value of ϑ, denoted by ϑ_0.

We write $\mathbf{Y}_s^t$ for $(Y_s, \ldots, Y_t)$ and define $\mathbf{X}_s^t$ similarly. We can and will embed $Y_1, \ldots, Y_n$ into $\{Y_t\}_{t=-\infty}^{\infty}$ related to $\{X_t\}_{t=-\infty}^{\infty}$ by the mechanism described above. When this is done, we write $\mathbf{Y}_{-\infty}^t$ for $(\ldots, Y_{t-1}, Y_t)$ and define $\mathbf{X}_{-\infty}^t$ similarly. Probabilities and expectations will be denoted by P and E, respectively, and the conditional expectation given some random variable $\mathbf{Y}$ by $E^{\mathbf{Y}}$. Likelihood ratios (with respect to ϑ_0) will be denoted by L and loglikelihood ratios by ℓ. Note that we use the same characters to denote different functions; the specific function will be clear from the argument. If the value of the parameter is ϑ_0, we often replace it by 0. Thus $L_\vartheta(\mathbf{Y}_1^n)$ is the likelihood ratio of $\mathbf{Y}_1^n$ calculated at ϑ, whereas $\ell_0(\mathbf{X}, \mathbf{Y})$ is the loglikelihood ratio of $\mathbf{X}$ and $\mathbf{Y}$ (being some general random variables) calculated at ϑ_0. To further shorten the notation, we let $h_t(\vartheta) = \ell_\vartheta(X_t, Y_t|\mathbf{X}_1^{t-1}, \mathbf{Y}_1^{t-1})$ for $t \geq 1$. Thus, by the very definition of an HMM, $h_t(\vartheta) = \log \alpha_\vartheta(X_{t-1}, X_t) + \log g_\vartheta(Y_t|X_t)$ for $t > 1$ and $h_1(\vartheta) = \log \pi_\vartheta(X_1) + \log g_\vartheta(Y_1|X_1)$. If $\mathbf{a} = (a_1, \ldots, a_d)$ is a d-dimensional vector with non-negative integer entries, then $D_{\mathbf{a}}$ denotes the corresponding partial derivative $\partial^{|\mathbf{a}|_+}/(\partial_{\vartheta_1}^{a_1} \cdots \partial_{\vartheta_d}^{a_d})$, where $|\mathbf{a}|_+ = \sum_1^d a_i$. We call such $\mathbf{a}$ a multi-index. Furthermore we define

$$C_k(y) = \sup_{\vartheta \in V_0} \max_{|\mathbf{a}|_+ = k} \max_{i,j} \{|D_{\mathbf{a}} \log \alpha_\vartheta(i,j)| + |D_{\mathbf{a}} \log g_\vartheta(y|i)| + |D_{\mathbf{a}} \log \pi_\vartheta(i)|\}, \qquad (2)$$

with V_0 being a neighborhood of ϑ_0 and the outer maximum being taken over all multi-indices $\mathbf{a}$ with $|\mathbf{a}|_+ = k$, and

$$B_k = \max \left\{ \prod_{t=1}^{r} E_0 \left(\prod_{i=p_{t-1}+1}^{p_t} \frac{C_{j_i}(Y_t)}{j_i!} \,\middle|\, X_t = x_t \right) \,:\, 1 \leq m \leq k,\, 1 \leq r \leq m, \right.$$
$$\left. j_1, \ldots, j_m \geq 1,\, \sum j_i = k,\, 0 = p_0 < p_1 < \ldots < p_r = m,\; x_1, \ldots, x_r \in \{1, \ldots, R\} \right\}.$$

The last quantity measures how big, in expectation, we can make a product of partial derivatives of the h_t by distributing a total of k derivatives and possible time indices over different components of the parameter and time, respectively. In particular, if $\overline{C}_k(y) = \max\{C_m(y) :\, 1 \leq m \leq k\}$, then

$$B_k \leq \max\{E_0(\overline{C}_k(Y_1)|X_1 = x)^m :\, x \in \{1, \ldots, R\},\, 1 \leq m \leq k\} \leq \overline{C}^k,$$

if all $\overline{C}_k \le \overline{C} < \infty$. On the other hand, if the Y_t are mutually independent, then

$$B_k \le \prod_{m=1}^{k} \Big(1 \vee \max\{E_0(C_m(Y_1) \mid X_1 = x) :\ x \in \{1, \dots, R\}\}\Big).$$

We now state the main assumptions being used in this paper.

(A1) The entries of transition probability matrix A_ϑ are bounded away from zero on V_0.

(A1∞) Condition (**A1**) holds for all ϑ_0.

(A2k) For all i, j, $\vartheta \mapsto \log \alpha_\vartheta(i,j)$ and $\vartheta \mapsto \log \pi_\vartheta(i)$ has k continuous derivatives in the neighborhood $\vartheta \in V_0$ of ϑ_0, and for all i and $y \in \mathcal{Y}$, $\vartheta \mapsto \log g_\vartheta(y|i)$ has k continuous derivatives in the same neighborhood.

(A2∞) All $\log \alpha_\vartheta(i,j)$ and $\log \pi_\vartheta(i)$ and all their derivatives are uniformly bounded.

(A3k) $B_k < \infty$.

(A3∞) All derivatives of $\log g_\vartheta(y|i)$ are uniformly bounded in y.

We now state our main results.

Theorem 2.1. *Assume (**A1**), (**A2**k) and (**A3**k) hold. Then for all multi-indices $\mathbf{a}$ with $|\mathbf{a}|_+ = k$,*

$$E_0 \left| \sup_{\vartheta \in V_0} D_{\mathbf{a}} \ell_\vartheta(\mathbf{Y}_1^n) \right| \le C_1 n B_k C_2^k k!$$

for some C_1 and C_2 that depend on the transition probability matrix A_ϑ, $\vartheta \in V_0$, only.

Theorem 2.2. *Let $\mathbf{a}$ and $\mathbf{b}$ be multi indices with $|\mathbf{a}|_+ = k$ and $|\mathbf{b}|_+ = m$, respectively. Then the following assertions hold true.*

- *(i) Under (**A1**), (**A2**k) and (**A3**k), $E_0|D_{\mathbf{a}}\ell_0(Y_1|\mathbf{Y}_{-n}^0)| \le C_1 B_k C_2^k k!$ where C_1 and C_2 depend on the transition probability matrix A_0 only.*
- *(ii) Under (**A1**), (**A2**∞) and (**A3**∞), if $-n \le -t \le 0$ then $|D_{\mathbf{a}} P_0(X_1 = x|\mathbf{Y}_{-n}^0) - D_{\mathbf{a}} P_0(X_1 = x|\mathbf{Y}_{-t}^0)| \le C_3 \rho^t$, $\rho < 1$ where C_3 depends on the uniform bound on all derivatives and A_0.*
- *(iii) Under (**A1**), (**A2**k), and (**A3**k), $E_0|D_{\mathbf{a}}\ell_0(Y_1|\mathbf{Y}_{-n}^0) D_{\mathbf{b}} L_0(\mathbf{Y}_{-n}^1)| \le C_3 k! m!$.*

Remarks.

1. The bounds in (i) and (iii) are the same as in the case $\{Y_i\}$ i.i.d.
2. The bound of (ii) expresses a strong mixing property. We could prove versions of (ii) and (iii) under (**A1**), (**A2**k), and (**A3**k), but the results are technically complicated and we leave them to the reader.

Theorem 2.3. *Under (**A1**), (**A2**∞) and (**A3**∞) the following assertions hold true.*

(i) $D_{\mathbf{a}}\ell_0(Y_1|\mathbf{Y}_{-n}^0)$ *converges* P_0*-a.s. to* $D_{\mathbf{a}}\ell_0(Y_1|\mathbf{Y}_{-\infty}^0)$.

(ii) $E_0(D_{\mathbf{a}}\ell_0(Y_1|\mathbf{Y}_{-n}^0)D_{\mathbf{b}}L_0(\mathbf{Y}_{-n}^1))$ *converges to an appropriate limit as* $n \to \infty$.

(iii) $n^{-1/2}(D_{\mathbf{a}}\ell_0(\mathbf{Y}_1^n) - E_0D_{\mathbf{a}}\ell(\mathbf{Y}_1^n))$ *converges weakly to a* $N(0, Var_0(D_{\mathbf{a}}\ell_0(Y_1|\mathbf{Y}_{-\infty}^0)))$ *distribution under* P_0.

3 Proofs of main results

Throughout the remainder of the paper we shall assume that the parameter space Θ is one-dimensional, that is $d = 1$. This causes no loss of generality, but simplifies the notations as we do not need to work with mixed partial derivatives. Derivatives with respect to ϑ of order k will be denoted with superindex k, for example $L_\vartheta^{(k)}$.

In many of the proofs, cumulants play a major role. Let $Z_1, \ldots, Z_k$ be k random variables. We denote their cumulant by

$$\Gamma(Z_1, \ldots, Z_k) = \frac{1}{\iota^k}\frac{\partial}{\partial u_1}\cdots\frac{\partial}{\partial u_k}\log\left(Ee^{\iota(u_1Z_1+\cdots+u_kZ_k)}\right)\Big|_{u_1=\ldots=u_k=0}, \tag{3}$$

where $\iota = \sqrt{-1}$. The cumulant is a multilinear function (that is, it is linear in any of the random variables if all other variables are kept fixed) and, in particular, if $Z_1 = \ldots = Z_k$ then $\Gamma(Z_1, \ldots, Z_k)$ is the standard kth cumulant of Z_1. Finally, if $Z = (Z_1, \ldots, Z_k)$ we write $\Gamma(Z) = \Gamma(Z_1, \ldots, Z_k)$ and $\Gamma^W(Z)$ for the cumulant of the conditional distribution of Z given W.

For the proof of Theorem 2.1 we proceed as follows.

(i) We write $\ell_0^{(k)}(\mathbf{Y}_1^n)$ as a linear combination of conditional cumulants of $\sum_{t=1}^n h_t^{(j)}(\vartheta_0)$, $1 \le j \le k$, by means of a general formula valid for any latent variable model, see (4).

(ii) By a generalization of results by S&S we give a bound on the individual conditional cumulants of the form (1) to yield the result.

For the proof of Theorem 2.2(i), we use the same formula expressing $\ell_0^{(k)}(Y_1|\mathbf{Y}_{-n}^0)$ as $\ell_0^{(k)}(\mathbf{Y}_{-n}^1) - \ell_0^{(k)}(\mathbf{Y}_{-n}^0)$ and analyzing it in the same way as in Theorem 2.1. For part (ii) we use a further decomposition into so-called centered moments; see below and part (ii) of Theorem 2.2. We also relate $L_0^{(m)}$ to the $h^{(i)}$, $1 \le i \le m$, by another general formula.

For the general formula, we need additional notation. Let $\mathbf{Z}_+$ be the set of positive integers. We let

$$\mathcal{J} = \{(J_1, \ldots, J_k) : k > 0, J_j \in \mathbf{Z}_+,\ j = 1, \ldots, k\}\,.$$

For $J \in \mathcal{J}$, $|J|$ denotes the dimension of the vector J and $|J|_+ = \sum J_j$. We define the following subsets of $\mathcal{J}$: $\mathcal{J}(k) = \{J \in \mathcal{J} : |J| = k\}$ and $\mathcal{J}^+(k) = \{J \in \mathcal{J} : |J|_+ = k\}$. Another useful set of integers is

$$\mathcal{J}_m^n(k) = \{(I_1, \ldots, I_k) : I_i \in \mathbf{Z}_+, m \le I_i \le n, i = 1, \ldots, k\}\,.$$

For any integer vector I as above, $\min I = \min I_i$, $\max I = \max I_i$ and $\Delta(I) = \max I - \min I$. Furthermore, if $a_1, a_2, \ldots$ is any sequence and $|I| = k$, then $a_I = (a_{I_1}, \ldots, a_{I_k})$. An operation between two sequences is done term wise, so that $a_I/b_I = (a_{I_1}/b_{I_1}, \ldots, a_{I_k}/b_{I_k})$. In general any operation is meant to be term by term, so $J! = (J_1!, J_2!, \ldots)$, $\prod a_I = \prod_{i=1}^k a_{I_i}$ and a very typical expression in this paper is

$$\frac{f_I^{(J)}}{J!} = \left(\frac{f_{I_1}^{(J_1)}}{J_1!}, \ldots, \frac{f_{I_k}^{(J_k)}}{J_k!} \right).$$

We now let $\mathbf{X}$ and $\mathbf{Y}$ be any random vectors such that only $\mathbf{Y}$ is observed with $\mathbf{X}$ being missing at random.

Proposition 3.1. *Suppose the loglikelihood ratio of the full data model $\ell_\vartheta(\mathbf{X}, \mathbf{Y})$ is k times differentiable. Then the loglikelihood ratio of the observable model is k times differentiable and*

$$\ell_0^{(k)}(\mathbf{Y}) = \sum_{J \in \mathcal{J}^+(k)} \frac{k!}{|J|!} \Gamma_0^{\mathbf{Y}} \left(\frac{\ell_0^{(J)}(\mathbf{X}, \mathbf{Y})}{J!} \right), \tag{4}$$

$$L_0^{(k)}(\mathbf{Y}) = \sum_{J \in \mathcal{J}^+(k)} \frac{k!}{|J|!} \prod \frac{\ell_0^{(J)}(\mathbf{Y})}{J!}. \tag{5}$$

The first of these formulae may be viewed as a generalization of results of Louis (1982) and Meilijson (1989), relating the score function and observed information of the observable vector $\mathbf{Y}$ to those of the full model. The above theorem provides results also for higher order derivatives. Both of the statements of the theorem are closely related to the "exlog relations" in Barndorff-Nielsen and Cox (1989, pp. 140). For the proof we need the following form of Faa di Bruno's formula, whose proof is in the Appendix.

Lemma 3.1. *(i) For any functions $f : \mathbf{R} \to \mathbf{R}$ and $h : \mathbf{R} \to \mathbf{R}$ with $h(0) = 0$ and f and h being k times differentiable,*

$$\frac{\partial^k}{\partial \vartheta^k} f(h(\vartheta))\bigg|_{\vartheta=0} = \frac{\partial^k}{\partial \vartheta^k} f\left(\sum_{i=0}^{k} \vartheta^i h^{(i)}(0)/i!\right)\bigg|_{\vartheta=0}.$$

(ii) If $f : \mathbf{R}^k \to \mathbf{R}$, then

$$\frac{\partial^k}{\partial \vartheta^k} f(\vartheta, \vartheta^2/2, \ldots, \vartheta^k/k!)\bigg|_{\vartheta=0} = \sum_{J \in \mathcal{J}^+(k)} \frac{k!}{|J|! \prod J!} \frac{\partial^{|J|}}{\partial_{u_{J_1}} \cdots \partial_{u_{J_{|J|}}}} f(u_1, \ldots, u_k)\bigg|_{u_1=\ldots=u_k=0}.$$

Proof of Proposition 3.1. We start with the representation

$$\ell_\vartheta(\mathbf{Y}) = \log E_0^{\mathbf{Y}} e^{\ell_\vartheta(\mathbf{X},\mathbf{Y})}.$$

Note that for random variables whose joint moment generating function exists in a vicinity of 0, the joint characteristic function in the definition (3) of the cumulant can be replaced by the joint moment generating function, and the factor ι^k in the denominator then also disappears. Hence

$$\frac{\partial^{|J|}}{\partial_{J_1} \cdots \partial_{J_{|J|}}} \log E_0^{\mathbf{Y}} e^{\sum u_i W_i}\bigg|_{u_i=\ldots=u_{|J|}=0} = \Gamma^{\mathbf{Y}}(W_{J_1}, \ldots, W_{J_{|J|}}) = \Gamma^{\mathbf{Y}}(W_J).$$

Now apply the first part of Lemma 3.1 to obtain

$$\ell_0^{(k)}(\mathbf{Y}) = \frac{\partial^k}{\partial \vartheta^k} \log E_0^{\mathbf{Y}} \exp\left\{\sum_{i=1}^{k} \vartheta^i \ell_0^{(i)}(\mathbf{X},\mathbf{Y})/i!\right\}\bigg|_{\vartheta=0}.$$

Apply the second part of the lemma to this expression with

$$f(u_1, \ldots, u_k) = \log E_0^{\mathbf{Y}} \exp\left\{\sum_{i=1}^{k} u_i \ell_0^{(i)}(\mathbf{X},\mathbf{Y})\right\}$$

to see that

$$\ell_0^{(k)}(\mathbf{Y}) = \sum_{J \in \mathcal{J}^+(k)} \frac{k!}{|J|! \prod J!} \Gamma^{\mathbf{Y}}(\ell_0^{(J)}(\mathbf{X},\mathbf{Y})).$$

The product $\prod J!$ can be taken inside Γ because of the multilinearity of the cumulant function and the proof of the first part of the lemma is complete.

Similarly,

$$L_0^{(k)}(\mathbf{Y}) = \left.\frac{\partial^k}{\partial \vartheta^k} e^{\ell_\vartheta(\mathbf{Y})}\right|_{\vartheta=0} = \left.\frac{\partial^k}{\partial \vartheta^k} \exp\left\{\sum_{j=1}^{k} \vartheta^j \ell_0^{(j)}(\mathbf{Y})/j!\right\}\right|_{\vartheta=0} = \sum_{J \in \mathcal{J}^+(k)} \frac{k!}{|J|!} \prod \frac{\ell_0^{(J)}(\mathbf{Y})}{J!}$$

by Lemma 3.1 with $f(u_1, \ldots, u_k) = \exp\{\sum_{j=1}^k u_j \ell_0^{(j)}(\mathbf{Y})\}$. □

The next lemma requires introduction of so-called centered moments and notation of mixing. For any random variables $Z_1, Z_2, \ldots$, let $\chi'(Z_1) = Z_1$ and $\chi(Z_1) = EZ_1$, and define recursively

$$\begin{aligned} \chi'(Z_1, \ldots, Z_k) &= Z_1(\chi'(Z_2, \ldots, Z_k) - \chi(Z_2, \ldots, Z_k)) \\ \chi(Z_1, \ldots, Z_k) &= E\chi'(Z_1, \ldots, Z_k). \end{aligned}$$

χ is called the centered moment function (S&S, p. 12). For example,

$$\chi(Z_1, Z_2, Z_3) = E(Z_1 Z_2 Z_3) - E(Z_1)E(Z_2 Z_3) - E(Z_1 Z_2)E(Z_3) + E(Z_1)E(Z_2)E(Z_3). \quad (6)$$

Similar to the notation for cumulants, if $Z = (Z_1, \ldots, Z_k)$ then $\chi(Z) = \chi(Z_1, \ldots, Z_k)$ and $\chi^W(Z)$ is the centered moment of the conditional distribution of Z given W.

Let $Z_t = g_t(T_t)$ for some measurable functions g_t, where $\{T_t\}_{t=-\infty}^{\infty}$ is Markovian and obeys the following mixing condition in terms of constants φ_t, $-\infty < t < \infty$. If $\underline{\mathcal{F}}_m$ is the σ-field generated by Z_t, $-\infty < t \le m$, and $\overline{\mathcal{F}}_n$ is the σ-field generated by Z_t, $n \le t < \infty$, then for all $m < n$,

$$\sup\{|P(B \mid A) - P(B)| : A \in \underline{\mathcal{F}}_m, B \in \overline{\mathcal{F}}_n, P(A) > 0\} \le \prod_{t=m+1}^{n} \varphi_t. \quad (7)$$

Lemma 3.2. *With $\{Z_t\}$ as above, assume $|Z_t| \le C_t$ a.s., $1 \le t \le n$, and let $1 \le t_1 \le t_2 \le \ldots \le t_k \le n$. Then*

$$\chi(Z_{t_1}, \ldots, Z_{t_k}) \le 2^{k-1} \prod_{j=1}^{k} C_{t_j} \prod_{j=t_1+1}^{t_k} \varphi_j.$$

PROOF. This is essentially Theorem 4.4 of S&S. The only difference is that we allow different bounds C_t on the Z_t; the validity of this extension follows easily as multiplicative constants can be moved in and out of centered moments. □

We now express the cumulants of sums in terms of centered moments, a generalization of a formula of S&S. Let $W_1, W_2, \ldots$ be random vectors, i.e. $W_i = (W_{i,1}, W_{i,2}, \ldots)$ etc. If $J \in \mathcal{J}$ is a set of indices, $W_{i,J}$ denotes the vector with elements $W_{i,j}$, $j \in J$.

Lemma 3.3. *(i) The multivariate cumulant* $\Gamma(\cdot)$ *can be expanded and bounded as*

$$\left|\Gamma\left(\sum_{i=1}^{n} W_{i,J}\right)\right| = \left|\sum_{I \in \mathcal{J}_1^n(|J|)} \Gamma(W_{I,J})\right|$$
$$\leq \sum_{i=1}^{n} \sum_{\substack{I \in \mathcal{J}_1^n(|J|) \\ \min I = i}} \sum_{\nu=1}^{|J|} \sum_{\uplus K_q = \{1,\ldots,|J|\}} M_\nu(K_1, \ldots, K_\nu) \prod_{q=1}^{\nu} |\chi(W_{I(K_q),J(K_q)})|,$$

where the inner sum is over all partitions $K_1, \ldots, K_\nu$ *of the set* $\{1, \ldots, |J|\}$, $I(K_q) = (I_{K_{q,1}}, I_{K_{q,2}}, \ldots)$ *and* $J(K_q)$ *is defined similarly. The* M_ν *are non-negative combinatorial constants satisfying, in particular, that* $M_\nu(\cdot) > 0$ *implies* $\sum_{q=1}^{\nu} \Delta(I(K_q)) \geq \Delta(I)$.

(ii) For all i *and* $0 \leq \rho < 1$,

$$\left|\sum_{\substack{I \in \mathcal{J}_1^n(|J|) \\ \min I = i}} \sum_{\nu=1}^{|J|} \sum_{\uplus K_q = \{1,\ldots,|J|\}} M_\nu(K_1, \ldots, K_\nu) \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))}\right| \leq |J|! \left(\frac{4}{1-\rho}\right)^{|J|-1}.$$

To clarify the notation once more we remark that $\Gamma(W_{I,J}) = \Gamma(W_{i_1,j_1}, \ldots, W_{i_\ell,j_\ell})$, where $\ell = |I| = |J|$.

PROOF. The multilinearity of the cumulant function is one of its basic properties. The bound in (i) comes from S&S Lemma 1.1, where also the property of the M_ν's is found. For part (ii), note that the proof of S&S Lemma 4.6 starts with their (4.55), which is equivalent to the expression in part (i) of the lemma. Then, in S&S's notation, we use $C_0 = C_2 = u = 1$, $f(s,t) = \rho^{|t-s|}$ and the bound $\rho^{\Delta(I_p)}$ on $\chi(W_{I_p,J_p})$; the result now follows from (4.60) in S&S. □

We remark that we tacitly assume that for any cumulant $\Gamma(W_{I,J})$, the vectors I (and J) are rearranged so that the elements of I become sorted in non-decreasing order before the cumulant is expanded into centered moments as in the above lemma. Since cumulants are invariant with respect to permutations of the random variables involved, such a rearrangement does not change the value of the cumulant, but it is necessary as we want to apply results like Lemmas 3.2 and 3.4 which do require sorted time indices.

We shall now examine the mixing condition (7) for HMMs and identify the φ's in this particular case. Define ρ by

$$1-\rho = \inf_{\vartheta \in V_0} \left(\min_{i,j} \alpha_0(i,j) \wedge \min_{i,j} \alpha_0^*(i,j) \right)$$

with $\alpha^*(i,j) = \pi_0(j)/\pi_0(i) \times \alpha_0(j,i)$; note that $\alpha_0^*(i,j)$ are the transition probabilities of the time-reversed Markov chain. Under (**A1**), $\rho < 1$. Generally, if A_ϑ is ergodic there is an m, $m \le R$, such that all m-step transition probabilities are positive. Assuming $m = 1$, it holds that if H_t is a set defined in terms of X_u and Y_u, $t \le u \le n$ only, then for $s < t$,

$$\max_i P_\vartheta(H_t \mid \mathbf{Y}_1^n, X_s = i) - \min_i P_\vartheta(H_t \mid \mathbf{Y}_1^n, X_s = i) \le \rho^{t-s}. \tag{8}$$

This result is proved in Douc, Moulines, and Rydén (2001, Corollary 1). A simple conditioning argument then yields

$$\max_i |P_\vartheta(H_t \mid \mathbf{Y}_1^n, X_s = i) - P_\vartheta(H_t \mid \mathbf{Y}_1^n)| \le \rho^{t-s} \quad \text{for all } \vartheta \in V_0, \tag{9}$$

which is our particular version of (7). Note that we work with the conditional Markov chain $\mathbf{X}|\mathbf{Y}$ (it is straightforward to verify that the conditional process is still Markov, although non-homogeneous), because in view of Proposition 3.1 we want to examine conditional cumulants. When $\mathbf{Y}_1^n$ is fixed we can identify φ_t in (7) with ρ in (9). Following Bickel et al. (1998, Lemma 5), one can also prove that for $1 \le s \le t-1$,

$$\sup_{A \subseteq \{1,\ldots,R\}} |P_\vartheta(X_s \in A \mid \mathbf{Y}_1^t) - P_\vartheta(X_s \in A \mid \mathbf{Y}_1^{t-1})| \le \rho^{t-1-s} \quad \text{for all } \vartheta \in V_0. \tag{10}$$

In the case $m > 1$, an inequality similar to (8) still holds true but with the bound on the conditional mixing now depending on the Y_t. This causes an additional degree of difficulty in our subsequent arguments and we do not treat this case.

The next result now follows from the above and Lemma 3.2.

Lemma 3.4. *Let $K_q = (K_{q,1}, \ldots, K_{q,\ell})$ be an element of a partition as in Lemma 3.3. Then*

$$|\chi_\vartheta^{\mathbf{Y}_1^n}(h_{I(K_q)}^{(J(K_q))}(\vartheta))| \le 2^{\ell-1} \prod_{j=1}^{\ell} C_{J_{K_{q,j}}}(Y_{I_{K_{q,j}}}) \rho^{\Delta(I(K_q))} \quad \textit{for all } \vartheta \in V_0,$$

where $I(K_q) = (I_{K_{q,1}}, \ldots, I_{K_{q,\ell}})$ etc. and $C_k(y)$ is defined in (2).

Proof of Theorem 2.1. First note that we may replace ϑ_0 by any ϑ in Proposition 3.1 without changing the definition of h_t, as only derivatives of this function appear in the proposition. Hence

$$|\ell_\vartheta^{(k)}(\mathbf{Y}_1^n)| \le k! \sum_{J\in\mathcal{J}^+(k)} \frac{1}{|J|!} \left| \Gamma_\vartheta^{\mathbf{Y}_1^n} \left(\frac{\sum_{i=1}^n h_i^{(J)}(\vartheta)}{J!} \right) \right| \quad \text{for all } \vartheta \in V_0 .$$

Expand $\Gamma_\vartheta^{\mathbf{Y}_1^n}(\sum_{i=1}^n h_i^{(J)}(\vartheta))$ as in Lemma 3.3(i). We can employ Lemma 3.4 to obtain the bound

$$\begin{aligned} \left| \prod_{q=1}^{\nu} \chi_\vartheta^{\mathbf{Y}_1^n}(h_{I(K_q)}^{(J(K_q))}(\vartheta)) \right| &\le \prod_{q=1}^{\nu} \left\{ 2^{|K_q|-1} \prod_{j=1}^{|K_q|} C_{J_{K_{q,j}}}(Y_{I_{K_{q,j}}}) \rho^{\Delta(I(K_q))} \right\} \\ &= 2^{|J|-\nu} \prod_{j=1}^{|J|} C_{J_j}(Y_{I_j}) \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))} \quad \text{for all } \vartheta \in V_0. \qquad (11) \end{aligned}$$

Taking the expectation of this bound and letting $i_1' < i_2' < \ldots < i_r'$ denote the distinct points of the vector I, the structure of an HMM yields

$$\begin{aligned} &E_0 \left\{ \sup_{\vartheta\in V_0} \left| \prod_{q=1}^{\nu} \chi^{\mathbf{Y}_1^n}(h_{I(K_q)}^{(J(K_q))}(\vartheta)) \right| \right\} \\ &\quad \le 2^{|J|-\nu} E_0 \left\{ \prod_{j=1}^{|J|} C_{J_j}(Y_{I_j}) \right\} \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))} \\ &\quad \le 2^{|J|-\nu} E_0 \left\{ E_0 \left(\prod_{\ell=1}^{r} \prod_{j:\, I_j = i_\ell'} C_{J_j}(Y_{i_\ell'}) \,\middle|\, \mathbf{X}_1^n \right) \right\} \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))} \\ &\quad = 2^{|J|-\nu} E_0 \left\{ \prod_{\ell=1}^{r} E_0 \left(\prod_{j:\, I_j = i_\ell'} C_{J_j}(Y_{i_\ell'}) \,\middle|\, X_{i_\ell'} \right) \right\} \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))} \\ &\quad \le 2^{|J|-1} \prod_{\ell=1}^{r} \max_x E_0 \left(\prod_{j:\, I_j = i_\ell'} C_{J_j}(Y_{i_\ell'}) \,\middle|\, X_{i_\ell'} = x \right) \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))}. \qquad (12) \end{aligned}$$

Multiplying by $1/\prod J!$ and using Lemma 3.3(ii) we obtain

$$E_0 \left\{ \sup_{\vartheta\in V_0} \left| \Gamma_\vartheta^{\mathbf{Y}_1^n} \left(\frac{\sum_{i=1}^n h_i^{(J)}(\vartheta)}{J!} \right) \right| \right\} \le \sum_{i=1}^{n} B_k |J|! \left(\frac{8}{1-\rho} \right)^{|J|-1} = n B_k |J|! \left(\frac{8}{1-\rho} \right)^{|J|-1} .$$

Hence

$$E_0\left|\sup_{\vartheta\in V_0}\ell_\vartheta^{(k)}(\mathbf{Y}_1^n)\right|\le nk!B_k\sum_{J\in\mathcal{J}^+(k)}\left(\frac{8}{1-\rho}\right)^{|J|-1}\le nk!B_k\frac{8}{1-\rho}\left(1+\frac{8}{1-\rho}\right)^{k-1},$$

where the last inequality is from Lemma 4.1(ii) in the Appendix. □

The next lemma is needed for the proof of Theorem 2.2.

Lemma 3.5. *Let $a'\le a\le b\le b'$ and suppose that W measurable w.r.t. the sigma-field generated by $\mathbf{Y}_a^b$. Then*

$$E_0\left\{WL_0^{(m)}(\mathbf{Y}_{a'}^{b'})\right\}=E_0\left\{WL_0^{(m)}(\mathbf{Y}_a^b)\right\},\quad m=0,1,2,\ldots$$

The proof is given in the Appendix.

Proof of Theorem 2.2. In this proof we again use the notation of Lemma 3.3. and drop the argument ϑ_0 of the function h as this parameter stays fixed throughout the proof. Moreover, since this lemma and other ones are formulated in terms of positive time indices we shift the indices of the statement of the theorem and set out to prove

$$E_0|\ell_0^{(k)}(Y_n|\mathbf{Y}_1^{n-1})|\le C_1B_kC_2^kk!.$$

By Proposition 3.1 and multilinearity of the cumulant function,

$$\begin{aligned}\ell_0^{(k)}(Y_n|\mathbf{Y}_1^{n-1})&=\ell_0^{(k)}(\mathbf{Y}_1^n)-\ell_0^{(k)}(\mathbf{Y}_1^{n-1})\\&=\sum_{J\in\mathcal{J}^+(k)}\frac{k!}{|J|!}\left\{\Gamma_0^{\mathbf{Y}_1^n}\left(\frac{\sum_{i=1}^{n-1}h_i^{(J)}+h_n^{(J)}}{J!}\right)-\Gamma_0^{\mathbf{Y}_1^{n-1}}\left(\frac{\sum_{i=1}^{n-1}h_i^{(J)}}{J!}\right)\right\}\\&=\sum_{J\in\mathcal{J}^+(k)}\frac{k!}{|J|!}\left\{\Gamma_0^{\mathbf{Y}_1^n}\left(\frac{\sum_{i=1}^{n-1}h_i^{(J)}}{J!}\right)-\Gamma_0^{\mathbf{Y}_1^{n-1}}\left(\frac{\sum_{i=1}^{n-1}h_i^{(J)}}{J!}\right)\right.\\&\qquad\left.+\sum_{\substack{J'\uplus J''=J\\J'\ne J}}\Gamma_0^{\mathbf{Y}_1^n}\left(\frac{\sum_{i=1}^{n-1}h_i^{(J')}}{J'!},\frac{h_n^{(J'')}}{J''!}\right)\right\},\end{aligned}\tag{13}$$

where the last sum is over all partitions (J',J'') of the set J except $(J',J'')=(J,\emptyset)$. This partition is excluded since it is the first sum of the right hand side and will be compared to the second one. Clearly, there are two types of cumulants here. The first two ones are similar and their difference will be shown to remain bounded in expectation as $n\to\infty$.

The last sum involves cumulants that contain at least one h_n, and this is sufficient to keep them bounded as $n \to \infty$.

We start by considering

$$\gamma = \left| \prod_{q=1}^{\nu} \chi_0^{\mathbf{Y}_1^n}(h_{I(K_q)}^{(J(K_q))}) - \prod_{q=1}^{\nu} \chi_0^{\mathbf{Y}_1^{n-1}}(h_{I(K_q)}^{(J(K_q))}) \right|,$$

where we assume that

$$\sum_{q=1}^{\nu} \Delta(I(K_q)) \geq \Delta(I).$$

This difference can be bounded in two ways. First, each term of the difference can be bounded separately. Arguing as for (11), we obtain

$$\begin{aligned} \gamma &\leq \prod_{q=1}^{\nu} \left| \chi_0^{\mathbf{Y}_1^n}(h_{I(K_q)}^{(J(K_q))}) \right| + \prod_{q=1}^{\nu} \left| \chi_0^{\mathbf{Y}_1^{n-1}}(h_{I(K_q)}^{(J(K_q))}) \right| \\ &\leq 2 \times 2^{|J|-1} \prod_{j=1}^{|J|} C_{J_j}(Y_{I_j}) \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))}. \end{aligned} \tag{14}$$

Secondly, we can write

$$\gamma = \prod_{j=1}^{|J|} C_{J_j}(Y_{I_j}) \times \left| \prod_{q=1}^{\nu} \chi_0^{\mathbf{Y}_1^n}(W_{I(K_q),J(K_q)}) - \prod_{q=1}^{\nu} \chi_0^{\mathbf{Y}_1^{n-1}}(W_{I(K_q),J(K_q)}) \right|, \tag{15}$$

where $W_{ij} = h_i^{(j)}/C_j(Y_i)$. Consider the scheme

$$\prod_{i=1}^{m} B_i - \prod_{i=1}^{m} A_i = \sum_{j=1}^{m} \left(\prod_{i=1}^{j-1} B_i \right) (B_j - A_j) \left(\prod_{i=j+1}^{m} A_i \right). \tag{16}$$

We expand $\chi(W_{I(K_q),J(K_q)})$ into a sum of $2^{|K_q|-1}$ products of expected values of products of W_{ij}'s, cf. (6), with each random factor being bounded by one. Pick one of the terms in this sum. This term is thus a product of no more than $|K_q|$ factors (with each factor being a conditional expectation of a product of W_{ij}'s). We call these factors A_i and B_i, respectively, when the expectation is conditional on $\mathbf{Y}_1^n$ and $\mathbf{Y}_1^{n-1}$, respectively. Using (10) we find that for each factor, the difference between its conditional expectations under $\mathbf{Y}_1^n$ and $\mathbf{Y}_1^{n-1}$, respectively, is bounded by $\rho^{n-1-\max I(K_q)} \leq \rho^{n-1-\max I}$, that is $|A_i - B_i|$

is bounded by this expression. Employing (16) the product, we can bound it by we arrive at the bound $|K_q|\rho^{n-1-\max I}$. Using this bound for each term, we find

$$|\chi_0^{\mathbf{Y}_1^n}(W_{I(K_q),J(K_q)}) - \chi_0^{\mathbf{Y}_1^{n-1}}(W_{I(K_q),J(K_q)})| \le 2^{|K_q|-1}|K_q|\,\rho^{n-1-\max I},$$

where $2^{|K_q|-1}$ is the total number of terms in the sum and $|K_q|$ upper bounds the number of factors in each term. In addition, just as Lemma 3.4 follows from Lemma 3.2 we obtain $|\chi_0^{\mathbf{Y}_1^n}(W_{I(K_q),J(K_q)})| \le 2^{|K_q|-1}\rho^{\Delta(I(K_q))}$ and similarly for $\mathbf{Y}_1^{n-1}$. Hence, by applying (16) to (15),

$$\begin{aligned}
\gamma &\le \prod_{j=1}^{|J|} C_{J_j}(Y_{I_j}) \sum_{p=1}^{\nu} \left(\prod_{\substack{q=1\\ q\ne p}}^{\nu} 2^{|K_q|-1}\rho^{\Delta(I(K_q))} \right) 2^{|K_p|-1}|K_p|\rho^{n-1-\max I} \\
&\le \prod_{j=1}^{|J|} C_{J_j}(Y_{I_j})\,|J|\,2^{|J|-1}\rho^{n-1-\max I} \qquad (17)
\end{aligned}$$

We can combine these two bounds, (14) and (17), by taking a geometric mean;

$$\gamma \le 2^{|J|-1}\prod_{j=1}^{|J|} C_{J_j}(Y_{I_j}) 2^{1/2}|J|^{1/2}\rho^{\sum_{q=1}^{\nu}\Delta(I(K_q))/2+(n-1-\max I)/2}. \qquad (18)$$

As in the proof of Theorem 2.1, it follows that

$$\begin{aligned}
E_0 &\left| \Gamma_0^{\mathbf{Y}_1^n}\left(\frac{\sum_{i=1}^{n-1} h_i^{(J)}}{J!}\right) - \Gamma_0^{\mathbf{Y}_1^{n-1}}\left(\frac{\sum_{i=1}^{n-1} h_i^{(J)}}{J!}\right) \right| \\
&\qquad \le 2^{1/2}|J|^{1/2}B_k \sum_{i=1}^{n-1}\rho^{(n-1-i)/2}|J|!\left(\frac{8}{1-\rho^{1/2}}\right)^{|J|-1} \\
&\qquad \le 2^{1/2}(1-\rho^{1/2})^{-1}B_k|J|^{1/2}|J|!\left(\frac{8}{1-\rho^{1/2}}\right)^{|J|-1}. \qquad (19)
\end{aligned}$$

We now proceed to bounding the second type of cumulants appearing in (13). Let (J', J'') be a partition of some $J \in \mathcal{J}^+(k)$ with $J' \ne J$. We can expand the cumulant similarly to Lemma 3.3(i) to obtain

$$\begin{aligned}
&\left| \Gamma_0^{\mathbf{Y}_1^n}\left(\sum_{i=1}^{n-1} h_i^{(J')}, h_n^{(J'')}\right) \right| \\
&\qquad \le \sum_{I\in\mathcal{J}_1^{n-1}(|J'|)\times\{n\}^{|J''|}} \sum_{\nu=1}^{|J|} \sum_{\uplus K_q=\{1,\dots,|J|\}} M_\nu(K_1,\dots,K_\nu)\prod_{q=1}^{\nu}|\chi_0^{\mathbf{Y}_1^n}(h_{I(K_q)}^{(J',J'')(K_q)})|.
\end{aligned}$$

Taking expectations, applying the bound (12) and multiplying by $1/(\prod J'! \prod J''!)$ yields

$$\begin{aligned} E_0 &\left| \Gamma_0^{\mathbf{Y}_1^n} \left(\frac{\sum_{i=1}^{n-1} h_i^{(J')}}{J'!}, \frac{h_n^{(J'')}}{J''!} \right) \right| \\ &\qquad \leq 2^{|J|-1} B_k \sum_{I \in \mathcal{J}_1^{n-1}(|J'|) \times \{n\}^{|J''|}} \sum_{\nu=1}^{|J|} \sum_{\uplus K_q = \{1,\ldots,|J|\}} M_\nu(K_1, \ldots, K_\nu) \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))}. \end{aligned}$$

Fix a partition (J', J'') of J such that $J'' \neq J$ and an $I \in \mathcal{J}_1^{n-1}(|J'|) \times \{n\}^{|J''|}$. We need to look closer at the combinatorial constants $M_\nu(K_1, \ldots, K_\nu)$. If the partition $(K_1, \ldots, K_\nu)$ is such that there is no K_q with $I(K_q)$ containing an element less than n as well as an element n, then $M_\nu(K_1, \ldots, K_\nu) = 0$. This is because in the graph for $M_\nu(K_1, \ldots, K_\nu)$ (see S&S, pp. 80) there can be no edge over the vertex corresponding to the first occurrence of n in I. Hence, we may disregard partitions of this kind.

Now consider a partition $(K_1, \ldots, K_\nu)$ with at least one $I(K_q)$ containing an element less than n and an element n, and let $\max' I$ denote the second largest element of the vector I, not counting multiple n's. We can form a new vector I' from I by replacing all elements of I being equal to n by $\max' I + 1$. Then $\sum_{q=1}^{\nu} \Delta(I'(K_q)) \leq \sum_{q=1}^{\nu} \Delta(I(K_q)) - (n - \max' I - 1)$. This vector I' is not a member of $\mathcal{J}_1^{n-1}(|J'|) \times \{n\}^{|J''|}$ (unless $\max' I = n-1$), but does belong to $\mathcal{J}_1^{\max' I}(|J'|) \times \{\max' I + 1\}^{|J''|}$; indeed, there is a one-to-one correspondence between vectors I and I' with these characteristics. Therefore

$$\begin{aligned} &\sum_{\substack{J' \uplus J'' = J \\ J' \neq J}} \sum_{I \in \mathcal{J}_1^{n-1}(|J'|) \times \{n\}^{|J''|}} \sum_{\nu=1}^{|J|} \sum_{\uplus K_q = \{1,\ldots,|J|\}} M_\nu(K_1, \ldots, K_\nu) \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))} \\ &\leq \sum_{i=1}^{n-1} \rho^{n-i-1} \sum_{\substack{J' \uplus J'' = J \\ J' \neq J}} \sum_{I \in \mathcal{J}_1^{i}(|J'|) \times \{i+1\}^{|J''|}} \sum_{\nu=1}^{|J|} \sum_{\uplus K_q = \{1,\ldots,|J|\}} M_\nu(K_1, \ldots, K_\nu) \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))}. \end{aligned}$$

For a fixed dimension $|J''|$ the summation above is done over $I \in \mathcal{J}_1^i(|J'|) \times \{i+1\}^{|J''|}$, a subset of $\mathcal{J}_1^n(|J|)$ characterized as vectors having exactly $|J''|$ elements of maximal size $i+1$, all being located at the end of the vector. In $\mathcal{J}_1^n(|J|)$ there are more elements having exactly $|J''|$ elements of maximal size $i+1$, disregarding their location. This number is in exact correspondence with the number of partitions (J', J'') of J with $|J''|$ as prescribed; their common value is the combinatorial constant $C(|J|, |J''|)$. Hence the above expression is bounded by

$$
\begin{aligned}
&\sum_{i=1}^{n-1} \rho^{n-i-1} \sum_{\substack{I \in \mathcal{J}_1^n(|J|) \\ \max I = i+1}} \sum_{\nu=1}^{|J|} \sum_{\uplus K_q = \{1,\ldots,|J|\}} M_\nu(K_1,\ldots,K_\nu) \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))} \\
&\qquad\qquad \le \sum_{i=1}^{n-1} \rho^{n-i-1} |J|! \left(\frac{4}{1-\rho}\right)^{|J|-1} \le \frac{1}{1-\rho} |J|! \left(\frac{4}{1-\rho}\right)^{|J|-1},
\end{aligned}
$$

where the second last inequality is Lemma 3.3(ii); by symmetry, the bound is still valid when $\min I = i$ is replaced by $\max I = i$. We note that this bound does in fact also take the partition $(J', J'') = (\emptyset, J)$, which was not considered above, into account; it corresponds to $I = \{n\}^{|J|}$. Thus

$$
E_0 \left| \Gamma_0^{\mathbf{Y}_1^n} \left(\frac{\sum_{i=1}^{n-1} h_i^{(J')}}{J'!}, \frac{h_n^{(J'')}}{J''!} \right) \right| \le \frac{1}{1-\rho} B_k |J|! \left(\frac{8}{1-\rho}\right)^{|J|-1}. \tag{20}
$$

Adding (19) and (20) as in (13) we find, using Lemma 4.1(ii) in the Appendix, that

$$
\begin{aligned}
E_0 |\ell_0^{(k)}(Y_n | \mathbf{Y}_1^{n-1})| &\le C' B_k k! \sum_{J \in \mathcal{J}^+(k)} |J|^{1/2} \left(\frac{8}{1-\rho^{1/2}}\right)^{|J|-1} \\
&\le C' B_k k! \sum_{J \in \mathcal{J}^+(k)} C^{|J|-1} \\
&\le C' B_k (1+C)^{k-1} k!
\end{aligned}
$$

for some $C > 8/(1-\rho^{1/2})$ and $C' = 2^{1/2}/(1-\rho^{1/2}) + 1/(1-\rho)$.

Here is the proof of part (ii). Since $D_k P_0(X_1 = x \mid \mathbf{Y}_{-j}^0)$, $j = t, \ldots, n$, can be expressed as polynomials in $P_0(X_1 = x \mid \mathbf{Y}_{-j}^0)$ and $D_r \log P_0(X_1 = x \mid \mathbf{Y}_{-j}^0)$, $1 \le r \le k$, it is enough to establish bounds for these quantities. The claim for $P_0(X_1 = x \mid \mathbf{Y}_{-j}^0)$ is essentially part 3 of Lemma 5 of Bickel et al. (1998). In general, note that, since

$$
\begin{aligned}
P_\vartheta(X_1 = x \mid \mathbf{Y}_{-j}^0) &= E_0 \left\{ I(X_1 = x) \frac{L_\vartheta(\mathbf{X}_{-j}^1, \mathbf{Y}_{-j}^0)}{L_\vartheta(\mathbf{Y}_{-j}^0)} \,\middle|\, \mathbf{Y}_{-j}^0 \right\} \\
&= P_0(X_1 = x | \mathbf{Y}_{-j}^0) E_0 \left\{ \frac{L_\vartheta(\mathbf{X}_{-j}^1, \mathbf{Y}_{-j}^0)}{L_\vartheta(\mathbf{Y}_{-j}^0)} \,\middle|\, \mathbf{Y}_{-j}^0, X_1 = x \right\},
\end{aligned}
$$

it holds that

$$
D_r \log P_0(X_1 = x \mid \mathbf{Y}_{-j}^1) = \sum_{J \in \mathcal{J}^+(r)} \frac{r!}{|J|! \prod J!} \Gamma\left(\ell_0^{(J)}(\mathbf{X}_{-j}^0, \mathbf{Y}_{-j}^0) - \ell_0^{(J)}(\mathbf{Y}_{-j}^0) \,\middle|\, \mathbf{Y}_{-j}^0, X_1 = x \right).
$$

One can argue as for part (i) of the theorem with the critical step being a bound on

$$\widetilde{\gamma} = \left| \prod_{q=1}^{\nu} \chi_0 \left(W_{I(K_q),J(K_q)} \,\middle|\, \mathbf{Y}^0_{-j}, X_1 = x \right) - \prod_{q=1}^{\nu} \chi_0 \left(W_{I(K_q),J(K_q)} \,\middle|\, \mathbf{Y}^0_{-j}, X_1 = x \right) \right|.$$

The argument follows the route in going from (15) to (18),

We now prove part (iii) of Theorem 2.2. We start by establishing a bound on derivative of the likelihood. Note that by Proposition 3.1, Theorem 2.1 and Lemma 4.1,

$$|L_0^{(k)}(\mathbf{Y}_1^n)| \leq C_1 n B_k k! \sum_{J \in \mathcal{J}^+(k)} \frac{1}{|J|!} \prod C_2^{|J|} \leq C_1 n k! B_k C_2^k 2^k \tag{21}$$

for some constants C_2, C_2, and B_k.

Let

$$\Lambda_i^j = \prod_{q=1}^{\nu} \chi_0^{\mathbf{Y}_i^j} (h_{I(K_q)}^{(J(K_q))}).$$

Then

$$\begin{aligned} E_0\{\gamma L_0^{(m)}(\mathbf{Y}_{-n}^1)\} &\leq E_0\{|\Lambda^1_{\min I} - \Lambda^0_{\min I}| L_0^{(m)}(\mathbf{Y}_{-n}^1)\} \\ &+ \sum_{i=0}^{n+\min I} E_0\{|\Lambda^1_{\min I - i - 1} - \Lambda^0_{\min I - i - 1} - \Lambda^1_{\min I - i} + \Lambda^0_{\min I - i}| L_0^{(m)}(\mathbf{Y}_{-n}^1)\} \\ &= \gamma_1 + \gamma_2, \end{aligned}$$

say. We bound now each of the terms. First

$$\gamma_1 \leq C_1^m C_2^{|J|} (|\min I| \vee m)^m \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))/2 + |\min I|/2}$$

by Lemma 3.5, (21), and (18). But $(|\min I| \vee m)^m \rho^{\max I/4} < C_3^m m!$ for some $C_3 > 0$, whence

$$\gamma_1 \leq C_1^m C_2^{|J|} m! \rho^{|\min I|/4 + \sum_{q=1}^{\nu} \Delta(I(K_q))}.$$

Similarly, by considering the appropriate differences in the expression for γ_2 depending on whether $i > \max(I)$ or vice-versa,

$$\begin{aligned} \gamma_2 &\leq \sum_{i=1}^{\infty} C_1^m C_2^{|J|} ((|\min I| + i) \vee m)^m \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))/2 + (i \bigvee |\min I|)/2} \\ &\leq \sum_{i=1}^{\infty} C_3^m C_2^{|J|} m! \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))/2 + (i + |\min I|)/4} \\ &\leq C_4^m C_2^{|J|} m! \rho^{\sum_{q=1}^{\nu} \Delta(I(K_q))/3 + |\min I|/12}. \end{aligned}$$

Having the bound on $E_0\{\gamma L^{(m)}\} = \gamma_1 + \gamma_2$, similar to the bound (18) on γ (except for the factor $C^m m!$), we continue as in the first part of the proof to prove the theorem.

This completes the proof of Theorem 2.2.

□

Proof of Theorem 2.3.

We proceed under the given assumptions. The first part of Theorem 2.3 follows readily from part (ii) of Theorem 2.2, since

$$D_{\mathbf{a}}\ell_0(Y_1|\mathbf{Y}_{-n}^0) = D_{\mathbf{a}} \log\Big(\sum_x P(X_1 = x|\mathbf{Y}_{-n}^0) g_\vartheta(Y_1|x)\Big)\Big|_{\vartheta=0}. \tag{22}$$

For the second part note that

$$D_{\mathbf{a}}\ell_0(\mathbf{Y}_1^n) = \sum_{i=1}^n D_{\mathbf{a}}\ell_0(Y_i|\mathbf{Y}_1^{i-1}) = \sum_{i=(\log n)^2+1}^n D_{\mathbf{a}}\ell_0(Y_i|\mathbf{Y}_{i-(\log n)^2}^{i-1}) + o_p(n^{1/2}), \tag{23}$$

by (22) and part (ii) of Theorem 2.2.

Now, under our assumptions the variables in the sum in (23) are uniformly bounded and geometrically mixing since the ith one is a function of $U_{i,n} = (Y_{i-(\log n)^2}, \ldots, Y_i)$, and the $\{U_{i,n}\}$ are uniformly in n geometrically φ-mixing. Then asymptotic normality, with natural centering by means and scaling by standard deviations, follows by the obvious extension to triangular arrays of the classical theorem of Ibragimov, see Doukhan (1994, p. 47) for instance. That the means and variances converge to the limit postulated is again an exercise in applying part (ii) of Theorem 2.2. □

4 Applications

4.1 A start at higher order asymptotics

It is well known, see for example Barndorff-Nielsen and Cox (1989) that in the i.i.d. case it is possible under suitable smoothness and moment conditions to 'debias' the MLE $\hat{\vartheta}$ to first order, that is to construct $\widehat{b}(\cdot)$ such that

$$E_\vartheta(\widehat{\vartheta} + n^{-1}\widehat{b}(\widehat{\vartheta})) = \vartheta + O(n^{-3/2})$$

and $\widehat{b} \to b$ in probability, uniformly in ϑ, for a fixed continuous b. With further conditions, $O(n^{-3/2})$ can be turned into $O(n^{-2})$.

Other second order asymptotics results of interest are Pfanzagl's second order optimality of functions of the MLE within classes of estimates with the same bias function (see for example Bickel, Götze and van Zwet, 1985), the validity of Bartlett's correction to the likelihood ratio test (see for example Bickel and Ghosh, 1990), the second order validity of bootstrap t-tests (see for example Hall, 1988) etc. The basic ingredients of debiasing are:

(a) A stochastic expansion for the MLE in terms of polynomials of the derivatives of $\ell_\vartheta(\mathbf{Y}_1^n)$.

(b) Probability bounds on probabilities of intermediate and large deviations of the derivatives $\ell_\vartheta(\mathbf{Y}_1^n)$.

For the other types of results one further needs,

(c) Edgeworth expansions for the joint distribution of the first few derivatives of $\ell_\vartheta(\mathbf{Y}_1^n)$ at ϑ_0.

As we shall see, our bounds give (a) directly. We conjecture that (b) can be established using results for sums of functions of Markov Variables as in S&S. Results of type (c) under simple assumptions, although plausible appear difficult to attain.

Here is the argument for (a) under (**A1**), (**A2**∞), and (**A3**∞) and real ϑ. Write $\widehat{\vartheta}$ for the MLE. Then, by a Taylor expansion,

$$-n^{-1/2}D\ell_0(\mathbf{Y}_1^n) = n^{1/2}(\widehat{\vartheta}-\vartheta_0)\, n^{-1}D_2\ell_0(\mathbf{Y}_1^n) + \frac{1}{2}n^{-1/2}n(\widehat{\vartheta}-\vartheta_0)^2 n^{-1}D_3\ell_{\vartheta^*}(\mathbf{Y}_1^n),$$

where ϑ^* lies between $\widehat{\vartheta}$ and ϑ_0. Suppose for simplicity that all entries of A_0 are positive and that the derivatives of $\log g_\vartheta$ are uniformly bounded in a neighborhood of ϑ_0. Then, by Theorem 2.1, under ϑ_0,

$$\begin{aligned} n^{1/2}(\widehat{\vartheta}-\vartheta_0) \;=\; & -\frac{n^{-1/2}D\ell_0(\mathbf{Y}_1^n)}{n^{-1}D_2\ell_0(\mathbf{Y}_1^n)} \\ & -\frac{1}{2}n^{-1/2}\left(\frac{n^{-1/2}D\ell_0(\mathbf{Y}_1^n)}{n^{-1}D_2\ell_0(\mathbf{Y}_1^n)}\right)^2 \frac{n^{-1}D_3\ell_0(\mathbf{Y}_1^n)}{n^{-1}D_2\ell_0(\mathbf{Y}_1^n)} + O_p(n^{-1}). \end{aligned}$$

But

$$n^{-1}D_2\ell_0(\mathbf{Y}_1^n) = -I(\vartheta_0) + O_p(n^{-1/2})$$

by Theorem 2.3 (which can be viewed as a refinement of Lemma 2 of Bickel et al., 1998). Here

$$I(\vartheta_0) = E_{\vartheta_0}(D_2 \log p_{\vartheta_0}(Y_1|\mathbf{Y}_{-\infty}^0))^2.$$

Finally we get

$$\begin{aligned} n^{1/2}(\widehat{\vartheta}-\vartheta_0) &= n^{-1/2}D\ell_0(\mathbf{Y}_1^n)I(\vartheta_0)^{-1} \\ &\quad - n^{-1/2}D\ell_0(\mathbf{Y}_1^n)I(\vartheta_0)^{-2}(n^{-1}D_2\ell_0(\mathbf{Y}_1^n)+I(\vartheta_0)) \\ &\quad - \frac{1}{2}n^{-5/2}(D\ell_0(\mathbf{Y}_1^n))^2 I(\vartheta_0)^{-3}D_3\ell_0(\mathbf{Y}_1^n)+O_p(n^{-1}), \end{aligned}$$

the desired stochastic expansion.

4.2 Asymptotic expansions

The following results generalizes Theorem 3.18 of Petrie (1969). For simplicity we assume that the parameter is real.

Theorem 4.1. *Assume that (**A1**∞)–(**A3**∞) hold. Then $I(\vartheta)$ and $K(\vartheta)$ are analytic functions. For instance,*

$$K(\vartheta)=\sum_{j=1}^{\infty} D_jK(\vartheta_0)\frac{(\vartheta-\vartheta_0)^j}{j!}$$

in some neighborhood of every $\vartheta_0\in\Theta$.

The series converges absolutely in some neighborhood of ϑ_0, since for every ϑ_0, $|D_jK(\vartheta_0)|\le C(\vartheta_0)^j j!$.

Moreover,

$$\begin{aligned} D_1K(\vartheta_0) &= E_0 D\ell_0(Y_1|\mathbf{Y}_{-\infty}^0)=0, \\ D_2K(\vartheta_0) &= -I(\vartheta_0), \\ D_{j+2}K(\vartheta) &= \lim_{n\to\infty} n^{-1}I_{n2}^{(j)}(\vartheta_0), \end{aligned}$$

where

$$I_{nd}^{(j)}(\vartheta_0)=n^{-1}\frac{\partial^j}{\partial\vartheta^j}E_0[\ell_\vartheta^{(d)}(\mathbf{Y}_1^n)L_\vartheta(\mathbf{Y}_1^n)]\bigg|_{\vartheta_0}.$$

The limit $I_{\infty d}^{(j)}(\vartheta_0)$ exists under our assumptions and can be represented by

$$\sum_{k=0}^{j}\binom{j}{k}\sum_{J\in\mathcal{J}^+(k+d)}\frac{(k+d)!}{|J!|}\sum_{I\in\mathcal{J}_1^\infty(|J|),\,\min(I)=1}E_0\left\{\Gamma_0^{\mathbf{Y}_1^\infty}\left(\frac{L_I^{(J)}}{J!}\right)L_0^{(j-k)}(\mathbf{Y}_1^{\max(I)})\right\},(24)$$

where $L_I=L_0(Y_I)$.

Proof. Note that $I(\vartheta) = -D_2K(\vartheta)$, so that it is enough to establish the claim for $K(\vartheta)$.

Since

$$I_{nd} = \sum_{i=1}^{n} E_0\{\ell_\vartheta^{(d)}(Y_i|\mathbf{Y}_1^{i-1})L_\vartheta(\mathbf{Y}_1^n)\},$$

we obtain

$$\left|n^{-1}I_{nd}^{(j)}(\vartheta_0)\right| = \frac{1}{n}\left|\sum_{i=1}^{n}\sum_{k=0}^{j}\binom{j}{k} E_0\{\ell_0^{(k+d)}(Y_i|\mathbf{Y}_1^{i-1})L_0^{(j-k)}(\mathbf{Y}_1^n)\}\right|$$

and the bound follows from Theorem 2.2. The limit is clearly given by the similarly bounded derivative of the expression

$$\lim_{n\to\infty} E_0\{\ell_0^{(d)}(Y_1|\mathbf{Y}_{-n}^0)L_0(\mathbf{Y}_{-n}^1)\}.$$

We need the limit in this expression since $L_0(\mathbf{Y}_{-\infty}^1)$ is not defined.

The other representation follows by expanding the derivatives as in Proposition 3.1, expanding the cumulant function as in Lemma 3.3 and using this lemma and Lemma 3.5 to argue that the limit exists.

$$\begin{aligned}
&\lim_{n\to\infty} n^{-1}I_{nd}^{(j)}(\vartheta_0) \\
&\quad = \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{j}\binom{j}{k}\sum_{J\in\mathcal{J}^+(k+d)}\frac{(k+d)!}{|J|!}\sum_{\substack{I\in\mathcal{J}_1^\infty(|J|)\\ \min(I)=1}} E_0\left\{\Gamma_0^{\mathbf{Y}_1^\infty}\left(\frac{L_I^{(J)}}{J!}\right)L_0^{(j-k)}(\mathbf{Y}_1^{\max(I)})\right\} \\
&\quad = \sum_{k=0}^{j}\binom{j}{k}\sum_{J\in\mathcal{J}^+(k+d)}\frac{(k+d)!}{|J|!}\sum_{\substack{I\in\mathcal{J}_1^\infty(|J|)\\ \min(I)=1}} E_0\left\{\Gamma_0^{\mathbf{Y}_1^\infty}\left(\frac{L_I^{(J)}}{J!}\right)L_0^{(j-k)}(\mathbf{Y}_1^{\max(I)})\right\}.
\end{aligned}$$

□

Corollary 4.1. *If under ϑ_0 the Y_i are i.i.d., then the sum in (24) becomes in principle computable as*

$$E_0\Gamma_0^{\mathbf{Y}_1^\infty}\left(\frac{L_I^{(J)}}{J!}\right)L_0^{(j-k)}(\mathbf{Y}_1^{\max(I)}) = 0$$

unless $I_1 = 1$, $I_1 - I_{i-1} = 0$ or 1, $i = 1, \ldots, |J|$.

Proof. Unless the conditions above are satisfied, $\Gamma_0^{\mathbf{Y}_1^\infty}(L_I^{(J)}/J!) = 0$, since the indices involved could be split into two blocks of the form $\{i_1 \leq \ldots \leq i_k\}$ and $\{i_{k+1} \leq \ldots \leq i_{\max(I)}\}$ with $i_{k+1} - i_k > 1$.

Since all variables in $L_I^{(J)}$ are at most 2-dependent, the conditional cumulant would have to vanish because the variables involved could be split into independent blocks. $\square$

We give some explicit computations for a special case below. We note that unfortunately the number of non-zero terms in $\Gamma_0^{\mathbf{Y}_1^\infty}(L_I^{(J)}/J!)$ grows exponentially as a function of $|J|$.

Corollary 4.2. *If $E_0|\log p_0(Y_1)| < \infty$ and the conditions of Theorem 2.2 hold then $H(\vartheta)$ is analytic.*

Proof. $H(\vartheta) = K(\vartheta) - E_\vartheta \log p_0(Y_1|\mathbf{Y}_{-\infty}^0)$ in this case. $\square$

4.3 Example: Information under independence and a two-state Markov chain with Gaussian observations

We consider now a reversible two state Markov chain with normal observations. Let $X_i \in \{-1, 1\}$, $P(X_{i+1} \neq X_i \mid X_i) = p$ and $Y_i = X_i + \varepsilon_i$ where $\ldots, \varepsilon_0, \varepsilon_1, \ldots$ are i.i.d. $N(0, \sigma^2)$, σ^2 known, random variables independent of the X process. We identify the parameter p with the ϑ of the general discussion and take $\vartheta_0 = 1/2$.

One can derive the information from Proposition 3.1. Alternatively, one can compute directly, cf. Louis (1982) and Meilijson (1989):

$$\begin{aligned}
\ell_\vartheta(\mathbf{Y}_1^n) &= E_0(\ell_\vartheta(\mathbf{X}_1^n, \mathbf{Y}_1^n) \mid \mathbf{Y}_1^n) - E_0(\ell_\vartheta(\mathbf{X}_1^n|\mathbf{Y}_1^n) \mid \mathbf{Y}_1^n) \\
\ell_0^{(2)}(\mathbf{Y}_1^n) &= E_0(\ell_\vartheta^{(2)}(\mathbf{X}_1^n, \mathbf{Y}_1^n) \mid \mathbf{Y}_1^n) - E_0(\ell_\vartheta^{(2)}(\mathbf{X}_1^n|\mathbf{Y}_1^n) \mid \mathbf{Y}_1^n)\Big|_{\vartheta_0} \\
&= E_0(\ell_0^{(2)}(\mathbf{X}_1^n, \mathbf{Y}_1^n) \mid \mathbf{Y}_1^n) + \mathrm{var}_0(\ell_0^{(1)}(\mathbf{X}_1^n|\mathbf{Y}_1^n) \mid \mathbf{Y}_1^n) \\
E_0\ell_\vartheta^{(2)}(\mathbf{Y}_1^n) &= E_0\ell_0^{(2)}(\mathbf{X}_1^n, \mathbf{Y}_1^n) + E_0\,\mathrm{var}_0(\ell_0^{(1)}(\mathbf{X}_1^n|\mathbf{Y}_1^n) \mid \mathbf{Y}_1^n)
\end{aligned}$$

However, if $X_1, X_2, \ldots$ are i.i.d. under ϑ_0, then we can simplify this expression. Write $\ell(\mathbf{X}_1^n|\mathbf{Y}_1^n) = \sum h_{\vartheta,t} = \sum h_{\vartheta,t}(X_{t-1}, X_t, Y_t)$, and note that, under independence, $\dot{h}_{\vartheta,i}$ and $\dot{h}_{\vartheta,j}$, $|j-i| > 1$ are independent given $\mathbf{Y}_1^n$. Moreover,

$$E_0\,\mathrm{cov}(\dot{h}_{\vartheta_0,2}, \dot{h}_{\vartheta_0,3}|\mathbf{Y}_1^n) = E_0E_0(\dot{h}_{\vartheta_0,2}\dot{h}_{\vartheta_0,3}|Y_2, Y_2, Y_3) - E_0(E_0(\dot{h}_{\vartheta_0,2}|Y_1, Y_2)E(\dot{h}_{\vartheta_0,3}|Y_2, Y_3)) = 0.$$

Hence

$$I(\vartheta_0) = -E_0E_0(\ddot{h}_{\vartheta_0,2}|Y_1, Y_2) - E_0\,\mathrm{var}(\dot{h}_{\vartheta_0,2}|Y_1, Y_2).$$

We specialize to our model where $h_{p,2} = (1 - X_1X_2)\log(p)/2 + (1 + X_1X_2)\log(1 - p)/2 + c_0$ for some c_0. Here

$$I(1/2) = 4(1 - E_0\,\mathrm{var}(X_1X_2|Y_1, Y_2)) = 4E_0(E_0(X_1|Y_1))^4, \quad I'(1/2) = 0.$$

It is reasonable to conjecture the following result.

Theorem 4.2. *Consider the two state symmetric Markov chain with normal observations as above. Then $I(p)$ has the following properties:*

(i) Symmetry: $I(p) = I(1-p)$;

(ii) Unimodality with minimum at 1/2;

(iii) Unboundness: $0 < \liminf_{p\downarrow 0} pI(p) \le \limsup_{p\downarrow 0} pI(p) \le 1$.

The proof of this result cannot depend on the expansion. Here is an argument.

Proof. The first two properties will be proved by showing that for any p^* between p and $1-p$ there is a Markov kernel that does not depend on the unknown parameter p, and transforms the observations $Y_1, Y_2, \dots$ to another sequence of variables $Y_1^*, Y_2^*, \dots$, such that the latter follows the same model as the original observations but with parameter p^*. This shows that $I(p^*) \le I(p)$, for any such p, and in particular $I(p) = I(1-p)$. Let $S_1, S_2, \dots$ be i.i.d. Bernoulli random variables with mean α, independent of the Y process and define

$$Y_i^* = (-1)^{\sum_{j=1}^i S_j} Y_i = (-1)^{\sum_{j=1}^i S_j}(X_i + \varepsilon_i) = X_i^* + \varepsilon_i^*.$$

Now, ε_i^* are still i.i.d. Gaussian, and $X_1^*, X_2^*, \dots$ is still Markovian, with values in $\{-1, 1\}$, but with probability of switching given by $p^* = (1-\alpha)p + \alpha(1-p)$.

We now prove the third property. We will argue that for any p_0 there is an estimator of p, valid for values of the parameter in a small neighborhood of p_0, whose asymptotic variance converges to 0 as $p_0 \to 0$. Since the information at p_0 is larger than the inverse of the variance of any regular estimator, $\lim_{p_0\to 0} I(p_0) = \infty$.

Here are the details.

We consider a net of models indexed by $p_0 \in (0,1)$. The parameter space of the p_0 model is $(p_0 - p_0^2, p_0 + p_0^2)$. We consider the limit as $p_0 \to 0$. Let $m = m(p_0)$ be such that $p_0 m \to 1/2$. For a given p_0:

It is easy to see, for example by induction, that

$$P(X_i = 1 \mid X_1 = 1) = \frac{1}{2} + \frac{1}{2}(1-2p)^{i-1}.$$

$$\mu_m(p) = \frac{1}{m} E\left(\sum_{i=1}^m X_i \,\middle|\, X_1 = 1 \right)$$

$$
\begin{aligned}
&= \frac{1}{m}\sum_{i=1}^{m}(1-2p)^{i-1} \\
&= \frac{1-(1-2p)^m}{2pm} \\
&\to 1-e^{-1},
\end{aligned}
$$

Note that

$$
\begin{aligned}
v_m(p) &= E\left(\frac{1}{m}\sum_{i=1}^{m}X_i\right)^2 \\
&= \frac{1}{m}+\frac{2}{m^2}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}E(X_iX_j) \\
&= \frac{1}{m}+\frac{2}{m^2}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(1-2p)^{j-i} \\
&= \frac{1}{m}+\frac{2(1-2p)}{m^2}\sum_{i=1}^{m}\frac{1-(1-2p)^{m-i}}{2p} \\
&= \frac{1}{pm}(1-p)+\frac{1-(1-2p)^m}{2p^2m^2} \\
&\to 4-2e^{-1},
\end{aligned}
$$

Let $1 < T_1 < \dots T_M \leq m$ be the times of switches, that is $X_{T_{i-1}} = \dots = X_{T_i-1} \neq X_{T_i}$ (with random M). Then the process $T_1/m, T_2/m, \dots, T_M/m$ converges weakly to a Poisson process on $(0,1)$, and hence the limiting distribution of $(m^{-1}\sum_{i=1}^{m}X_i)^2$ is non-degenerate. Since the limit distribution of $(m^{-1}\sum_{i=1}^{m}\varepsilon_i)^2$ is degenerate, the limit distribution of $(m^{-1}\sum_{i=1}^{m}Y_i)^2$ is non-degenerate and equals the limiting distribution of $(m^{-1}\sum_{i=1}^{m}X_i)^2$.

Consider now the statistics

$$
\widehat{v}_m = \frac{m}{n}\sum_{j=0}^{[n/m]-1}\left(\frac{1}{m}\sum_{i=1}^{m}Y_{jm+i}\right)^2 .
$$

Then

$$
\lim_{p_0\to 0}\lim_{n\to\infty}(n/m)\operatorname{var}(\widehat{v}_m - v_m(p)) = c_v < \infty.
$$

Suppose it is known that $p \in (p_0 - p_0^2, p_0 + p_0^2)$ and let the estimator $\widehat{p}$ be defined by

$v_m(\widehat{p}) = \widehat{v}_m$. Since

$$\frac{dv_m(p)}{dp} = -\frac{1}{mp^2} + \frac{m(1-2p)^{m-1}}{p^2m^2} - \frac{1-(1-2p)^m}{p^3m^2},$$

it follows that

$$p_0\frac{dv_m(p)}{dp} \to -6 + 6e^{-1},$$

where the limit is taken as above. Hence

$$\liminf_{p_0\to 0} p_0 I(p_0) \geq \liminf_{p_0\to 0} \frac{p_0}{n\operatorname{var}(\hat{p})} = 72c_v^{-1}(1-e^{-1})^{-1} > 0.$$

On the other hand the information cannot be larger than the information when the X process is observed directly. The latter is p_0^{-1} in the limit. □

Acknowledgement

We thank the anonymous referee for an exeptionally careful reading of the first version of the manuscript, leading to a substantially improved paper.

Appendix

Proof of Lemma 3.1. The first part is clear since the kth derivative depends on h only through its k first derivatives at 0, and hence it will not be changed if h is replaced by any other function with the same k first derivatives.

For the second part, we first observe that

$$\left.\frac{\partial^k}{\partial\vartheta^k} f(\vartheta, \vartheta^2/2, \ldots, \vartheta^k/k!)\right|_{\vartheta=0} = \sum_{J\in\mathcal{J}} C(J,k)\left.\frac{\partial^{|J|}}{\partial u_{J_1}\cdots\partial u_{J_{|J|}}} f(u_1,\ldots,u_k)\right|_{u_1=\cdots=u_k=0},$$

where the constants $C(J,k)$ do not depend on f, and, without loss of generality, $C(J,k) = C(J',k)$ if J' is a permutation of J. We will find these constants by considering a convenient family of functions f. Let

$$f(u_1,\ldots,u_k) = \prod_{j=1}^k \frac{u_j^{m_j}}{m_j!}.$$

Then

$$\frac{\partial^{|J|}}{\partial u_{J_1} \cdots \partial u_{J_{|J|}}} f(u_1, \dots, u_k) \bigg|_{u_1 = \cdots = u_k = 0} = \begin{cases} 1 \text{ if } J \in A(m_1, \dots, m_k), \\ 0 \text{ otherwise,} \end{cases}$$

where $A(m_1, \dots, m_k) = \{J \in \mathcal{J}^+(\sum m_i) : \sum_j \mathbf{1}(J_j = i) = m_i, \, i = 1, \dots, k\}$. Hence

$$\begin{aligned} \frac{\partial^k}{\partial \vartheta^k} f(\vartheta, \vartheta^2/2, \dots, \vartheta^k/k!) \bigg|_{\vartheta=0} &= C(J^*, k) \, |A(m_1, \dots, m_k)| \\ &= C(J^*, k) \frac{(\sum_{i=1}^k m_i)!}{\prod_{i=1}^k m_i!}, \end{aligned} \tag{25}$$

where J^* is any member of $A(m_1, \dots, m_k)$. On the other hand, we can compute directly that

$$f(\vartheta, \vartheta^2/2, \dots, \vartheta^k/k!) = \frac{\vartheta^{\sum_{i=1}^k i m_i}}{\prod_{i=1}^k m_i! \prod_{i=1}^k (i!)^{m_i}}$$

and

$$\frac{\partial^k}{\partial \vartheta^k} f(\vartheta, \vartheta^2/2, \dots, \vartheta^k/k!) \bigg|_{\vartheta=0} = \begin{cases} \frac{(\sum_{i=1}^k i m_i)!}{\prod_{i=1}^k m_i! \prod_{i=1}^k (i!)^{m_i}} & \text{if } \sum_{i=1}^k i m_i = k, \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

Comparing (25) to (26) and noting that $|J^*| = \sum_{i=1}^k m_i$ and $|J^*|_+ = \sum_{i=1}^k i m_i$ we obtain that

$$C(J, k) = \begin{cases} \frac{k!}{|J|! \prod J!} & \text{if } |J|_+ = k, \\ 0 & \text{otherwise,} \end{cases}$$

and the proof is complete. □

Proof of Lemma 3.5. The lemma follows since for any two random variables U_1 and U_2 with joint density $h_{U_1 U_2}(\cdot, \cdot)$ with respect to some measure $\mu_1 \times \mu_2$:

$$h^{(m)}_{U_1 U_2}(u_1, u_2) = \textstyle\sum_{j=0}^m \binom{m}{j} h^{(j)}_{U_1}(u_1) h^{(m-j)}_{U_2|U_1}(u_2|u_1).$$

Hence, for any function $f(\cdot)$:

$$\begin{aligned} E\left\{ f(U_1) \frac{h^{(m)}_{U_1 U_2}(U_1, U_2)}{h_{U_1 U_2}(U_1, U_2)} \right\} &= \sum_{j=0}^m \binom{m}{j} \int f(u_1) h^{(j)}_{U_1}(u_1) h^{(m-j)}_{U_2|U_1}(u_2|u_1) d\mu_1(u_1) d\mu_2(u_2) \\ &= \int f(u_1) h^{(m)}_{U_1}(u_1) d\mu_1(u_1) \\ &= E\left\{ f(U_1) \frac{h^{(m)}_{U_1}(U_1)}{h_{U_1}(U_1)} \right\}, \end{aligned}$$

since

$$\int h^{(m-j)}_{U_2|U_1}(u_2|u_1)d\mu_2(u_2) = \begin{cases} 1 \; j = m \\ 0 \; j \neq m. \end{cases}$$

Now, let U_1 represents $\mathbf{Y}_a^b$, U_2 be all the Y outside this range.

□

Lemma 4.1. *(i) For any real-valued sequence* $a_1, \dots, a_n$*,*

$$\left(\sum_{i=1}^n a_i \right)^k = \sum_{I \in \mathcal{J}_1^n(k)} \prod a_I = \sum_{J \in \mathcal{J}(k)} \prod a_J,$$

with $a_i = 0$ *for* $i > n$*.*

(ii) For any $c \geq 0$ *and* $k > 0$*,*

$$\sum_{J \in \mathcal{J}^+(k)} c^{|J|} = c(c+1)^{k-1},$$

$$\sum_{J \in \mathcal{J}^+(k)} \frac{c^{|J|}}{|J|!} \leq \frac{(c \vee k)^k}{k!} 2^{k-1}.$$

Proof. Part (i) follows by expanding the expression on the left hand side.

For part (ii), let $a_k = \sum_{J \in \mathcal{J}^+(k)} c^{|J|}$. Then for $|x|$ small enough,

$$\begin{aligned} \sum_{k=1}^\infty a_k x^k &= \sum_{k=1}^\infty \sum_{J \in \mathcal{J}^+(k)} c^{|J|} x^{|J|_+} \\ &= \sum_{i=1}^\infty c^i \sum_{J \in \mathcal{J}(i)} x^{|J|_+} \\ &= \sum_{i=1}^\infty c^i \sum_{J \in \mathcal{J}(i)} \prod x^J \\ &= \sum_{i=1}^\infty c^i \left(\sum_{j=1}^\infty x^j \right)^i \\ &= \frac{c}{c+1} \sum_{k=1}^\infty (c+1)^k x^k. \end{aligned}$$

The first claim follows.

To prove the second claim note that if $c \geq k$ then $c^{|J|}/|J|! \leq c^k/k!$ (since $|J| \leq k$) and $\sum_{J \in \mathcal{J}^+(k)} 1 = 2^{k-1}$ (e.g. by the first part). On the other hand if $c \leq k$ then $c^{|J|}/|J|! \leq k^k/k!$. □

References

Barndorff-Nielsen, O. and Cox, D.R. (1989). *Asymptotic Techniques for Use in Statistics.* Chapman and Hall, London.

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., Götze, F. and van Zwet, W.R. (1985). A simple analysis of third-order efficiency of estimates. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II*, 749–768. Wadsworth, Belmont, CA.

Bickel, P.J. and Ghosh, J.K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction — A Bayesian argument. *Ann. Statist.* **18**, 1070–1090.

Bickel, P.J. and Ritov, Y. (1996). Inference in hidden Markov models I: Local asymptotic normality in the stationary case. *Bernoulli* **2**, 199–228.

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.

Douc, R., Moulines, E. and Rydén, T. (2001). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. Preprint.

Doukhan, P. (1994) *Mixing. Properties and Examples.* Springer Lecture Notes in Statistics **85**. Springer-Verlag, New York.

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

Hall, P. (1988). Rate of convergence in bootstrap approximations. *Ann. Probab.* **16**, 1665–1684.

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

Kalman, R.E. (1977). A new approach to linear filtering and prediction problems. In *Linear Least-Squares Estimation*, 254–264. Dowden, Hutchinson & Ross, Stroudsburg, PA.

Leroux, B.G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stoch. Proc. 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.

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

MacDonald, I.L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series.* Chapman & Hall, London.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *J. Roy. Statist. Soc. B* **51**, 127–138.

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. *Proc. IEEE* **77**, 257–284.

Saulis, L. and Statulevičius, V.A. (1991). *Limit Theorems for Large Deviations.* Kluwer, Dordrecht.