ASYMPTOTIC NORMALITY OF THE MAXIMUM LIKELIHOOD ESTIMATOR IN STATE SPACE MODELS

JENS LEDET JENSEN[1] NIELS VÆVER PETERSEN[2]

University of Aarhus[1,2] *and MaPhySto*[1*]

Abstract

State space models is a very general class of time series models capable of modeling dependent observations in a natural and interpretable way. Inference in such models have been studied by Bickel *et al.*, who consider hidden Markov models, which are a special kind of state space models, and prove that the maximum likelihood estimator is asymptotically normal under mild regularity conditions. In this paper we generalize the results of Bickel *et al.* to state space models, where the latent process is a continuous state Markov chain satisfying regularity conditions, which are fulfilled if the latent process takes values in a compact space.

AMS 1991 subject classification. Primary 62F12; secondary 62M09.
Keywords and phrases. State space models, asymptotic normality, maximum likelihood estimation.

1. Introduction. A state space model is a discrete time model for dependent observations $\{Y_k\}$, where the dependence is modelled via an unobserved Markov process $\{X_k\}$ such that, conditionally on $\{X_k\}$ the Y_k's are independent, and the distribution of Y_k depends on X_k only. The unobserved process $\{X_k\}$ is often referred to as the *latent* process. The state space framework encompasses the classical ARMA models, but, more interestingly, non-linear and non-Gaussian models can be formulated in this framework as well.

We will consider inference in state space models by the likelihood method. The likelihood function can not always be calculated explicitly in these models, however, for linear state space models with Gaussian errors the likelihood function can be calculated

*Centre for Mathematical Physics and Stochastics, funded by a grant from the Danish National Research Foundation.

by the Kalman filter. There is an extensive literature on Kalman filtering, see for instance West & Harrison (1989) who gives a comprehensive treatment of linear state space models with many examples.

For non-linear state space models and for state space models with non-Gaussian errors the likelihood function can rarely be calculated explicitly. Instead different approximations to the likelihood function have been proposed. Kitagawa & Gersch (1996) discusses an approximation to the likelihood function based on numerical integration techniques, an approach which is also studied in Frühwirth-Schnatter (1994). However, with these techniques the likelihood function can only be approximated to a certain degree of accuracy. Alternatively the likelihood function can be approximated to any degree of accuracy by simulation techniques. This approach is investigated by Durbin & Koopman (1997), Shephard & Pitt (1997) and references therein.

Inference in state space models has mainly been studied in the case of hidden Markov models where the latent process takes values in a finite set. Leroux (1992) proved consistency of the maximum likelihood estimator and Bickel, Ritov & Rydén (1998) proved asymptotic normality. The purpose of this paper is to extend the results of Bickel et al. to cover more general state space models where the latent process is a continuous Markov process. We show that the distributional inequality in Lemma 4 in Bickel et al. is valid in our setup also, under regularity conditions which can be fulfilled if the state space of the latent process is a compact set. The inequality states a mixing property of the latent process, given the observed process, and is the main key to proving asymptotic normality. Having established this mixing result we follow Bickel et al. in their proof of the central limit theorem for the score function and in the proof of the uniform law of large numbers for the observed information.

In Section 2 we state the model and the assumptions we will work under. In Section 3 we state our main results, the central limit theorem for the score function, the uniform law of large numbers for the observed information, and, finally, asymptotic normality of the maximum likelihood estimate. In Section 4 we prove the central limit theorem and in Section 5 we prove the law of large numbers.

2. Notation and assumptions. Let $\{X_k\}$ denote a stationary homogenous Markov chain on the measurable space $(\mathcal{X}, \mathcal{A}, \mu)$. Here $\mathcal{X}$ may be continuous or discrete. A typical setting fulfilling our assumptions below, is where $\mathcal{X}$ is a compact set. Let $\alpha_\theta(x,z)$ denote the transition densities with respect to μ which are parametrized by a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$. Let $\{Y_k\}$ be a sequence of stochastic variables on the measurable space $(\mathcal{Y}, \mathcal{B}, \nu)$ such that given $\{X_k\}$ the Y_k's are independent, and the distribution of Y_i depends through $\{X_k\}$ on X_i only and has density $g_\theta(y_i|x_i)$ wrt. ν. The model can thus be formulated as

$$\begin{aligned} Y_k \mid X_k &\sim g_\theta(y_k \mid x_k), \\ X_k \mid X_{k-1} &\sim \alpha_\theta(x_{k-1}, x_k). \end{aligned}$$

We will let π_θ denote the density wrt. μ of the stationary distribution of X.

We observe values $Y_1, Y_2, \ldots, Y_n$ of the process $\{Y_k\}$ while $\{X_k\}$ remains unobserved, and we wish to estimate θ by the maximum likelihood method. We will let $l_n(\theta)$ denote the log likelihood function based on $Y_1, \ldots, Y_n$. In Section 4 an expression for this function is derived. For the moment we only give the expression for the simultaneous density of $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ wrt. $\mu^n \times \nu^n$,

$$p_\theta(x_1, \ldots, x_n, y_1, \ldots, y_n) = \pi_\theta(x_1) g_\theta(y_1 \mid x_1) \prod_{k=2}^{n} \{\alpha_\theta(x_{k-1}, x_k) g_\theta(y_k | x_k)\}. \tag{1}$$

Above, as everywhere else in this paper, we use the sloppy, but hopefully clear notation $p_\theta(z)$ for the density of a stochastic vector Z with respect to a measure given by the context.

We will let Dg_θ denote the gradient of g_θ wrt. θ and $D^2 g_\theta$ will denote the Hessian, and we will let $\tau_\theta(x) = D \log \pi_\theta(x)$, $\lambda_\theta(x,z) = D \log \alpha_\theta(x,z)$ and $\gamma_\theta(y|x) = D \log g_\theta(y|x)$. The true parameter will be denoted θ_0 and a notation like τ_0 is short for τ_{θ_0}. Throughout the paper X_1^n will denote the vector $(X_1, \ldots, X_n)$ and $\mathbf{c}$ will denote an unspecified finite constant. In the assumptions below we will let $\|\cdot\|$ denote the max-norm of a $d \times d$ maxtrix, $\|A\| = \max_{i,j} |A_{ij}|$.

We will assume that there exists a $\delta > 0$ such that with $B_0 = \{\theta \in \Theta \mid |\theta - \theta_0| < \delta\}$ the following conditions hold.

(A1) There exists a $\sigma > 0$ and an $M < \infty$ such that $\sigma \le \alpha_\theta(x,z) \le M$ for all $x, z \in \mathcal{X}$ and all $\theta \in B_0$.

(A2) For all $x, z \in \mathcal{X}$ the maps $\theta \mapsto \alpha_\theta(x,z)$ and $\theta \mapsto \pi_\theta(x)$ are twice continuously differentiable on B_0. Likewise, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ the map $\theta \mapsto g_\theta(y|x)$ is twice continuously differentiable on B_0.

(A3) Define $\rho(y) = \sup_{\theta \in B_0} \sup_{x,z \in \mathcal{X}} g_\theta(y|z)/g_\theta(y|x)$, then

$$\inf_{x \in \mathcal{X}} \int_{\mathcal{Y}} g_0(y|x)/\rho(y)\, \nu(dy) > 0.$$

(A4) (i) $\sup_{\theta \in B_0} \sup_{x,z \in \mathcal{X}} |\lambda_\theta(x,z)| < \infty$ and $\sup_{\theta \in B_0} \sup_{x \in \mathcal{X}} |\tau_\theta(x)| < \infty$.

(ii) $\sup_{\theta \in B_0} \sup_{x,z \in \mathcal{X}} \|D\lambda_\theta(x,z)\| < \infty$ and $\sup_{\theta \in B_0} \sup_{x \in \mathcal{X}} \|D\tau_\theta(x)\| < \infty$.

(iii) Define $\gamma^*(Y_1) = \sup_{\theta \in B_0} \sup_{x \in \mathcal{X}} |\gamma_\theta(Y_1|x)|$ then $\gamma^*(Y_1) \in \mathbb{L}^2(P_0)$ and $\sup_{\theta \in B_0} \sup_{x \in \mathcal{X}} \|D\gamma_\theta(Y_1|x)\| \in \mathbb{L}^1(P_0)$.

(A5) (i) For ν-almost all $y \in \mathcal{Y}$ there exists a function $h_y : \mathcal{X} \to \mathbb{R}_+$ in $\mathbb{L}^1(\mu)$ such that $|g_\theta(y|x)| \le h_y(x)$ for all $\theta \in B_0$.

(ii) For μ-almost all $x \in \mathcal{X}$ there exist functions $h_x^1 : \mathcal{Y} \to \mathbb{R}_+$ and $h_x^2 : \mathcal{Y} \to \mathbb{R}_+$ in $\mathbb{L}^1(\nu)$ such that $|Dg_\theta(y|x)| \le h_x^1(y)$ and $\|D^2 g_\theta(y|x)\| \le h_x^2(y)$ for all $\theta \in B_0$.

(A6) $\theta_0 \in \operatorname{int}(\Theta)$.

Remark 2.1 Note that if $\sup_{x,z\in\mathcal{X}} |\lambda_\theta(x,z)| < \infty$ for a $\theta \in B_0$ and $\sup_{x\in\mathcal{X}} |\tau_\theta(x)| < \infty$ for a $\theta \in B_0$ then A4(ii) implies A4(i). Likewise in A5(ii) the local dominated ν-integrability assumption of $y \mapsto \|D^2 g_\theta(y|x)\|$ for μ-almost all x implies a similar property of $y \mapsto |Dg_\theta(y|x)|$, provided that $|Dg_\theta(y|x)| \in \mathbb{L}^1(P_0)$ for a $\theta \in B_0$.

Remark 2.2 By assumptions A5(i), A1 and A4 the function $x_1^n \mapsto D^i p_\theta(x_1^n, y_1^n)$ is locally dominated μ^n-integrable around θ_0 for ν^n-almost all y_1^n, any $n \in \mathbb{N}$ and $i = 1, 2$. This is seen by noting that by (1) $Dp_\theta(x_1^n, y_1^n)$ consists of a sum of terms like, for instance,

$$\pi_\theta(x_1) g_\theta(y_1|x_1) \prod_{k=2,k\neq j}^{n-1} \{\alpha_\theta(x_{k-1}, x_k) g_\theta(y_k|x_k)\} \alpha_\theta(x_{j-1}, x_j) Dg_\theta(y_j|x_j).$$

By A1 and A5(i) this term is absolutely dominated by

$$M^n \prod_{k=1}^n h_{y_k}(x_k) |D \log g_\theta(y_j|x_j)| \leq M^n \sup_{\theta\in B_0} \sup_{x\in\mathcal{X}} |\gamma_\theta(y_j|x)| \prod_{k=1}^n h_{y_k}(x_k),$$

which for almost all fixed y_1^n is a μ^n-integrable function, by assumption A4(iii). The domination of the second derivative is similar. The local integrability assumption is needed to "interchange integration and differentiation" in some expressions below.

Remark 2.3 The process Y is ergodic under A1. To see this, we observe that Y_k can be described as $Y_k = g(X_k, U_k; \theta_0)$, where $U_1, U_2, \ldots$ are uniformly distributed $U(0,1)$, and independent of X. Now $\{(X_k, U_k)\}$ is a stationary Markov chain, with transition density $p(x_1, u_1 \mid x_0, u_0) = \alpha_0(x_0, x_1)$ and hence ergodic by A1. Thus Y is also ergodic.

Remark 2.4 The assumptions A1, A3 and A4 are restrictive and are not fulfilled in a general state space model. A typical example where A1 to A5 are fulfilled is the following. Suppose $\mathcal{X}$ is a compact set in $\mathbb{R}^q$ and μ is the Lebesque measure. If the transition density $\alpha_\theta(x,z)$ and the stationary density $\pi_\theta(x)$ are positive and satisfies A2 and if $\alpha_\theta(x,z)$, $\pi_\theta(x)$ and their first and second derivatives are continuous functions of (θ, x, z) and (θ, x), respectively, then A1, A4(ii) and A4(i) are satisfied. Suppose, furthermore, that $g_\theta(y|x)$ is an exponential family density,

$$g_\theta(y|x) = e^{\phi(x,\theta)\cdot t(y) - \kappa(\phi(x,\theta))}.$$

Here κ denotes the cumulant transform of $t(Y)$ defined on the full parameter space $\Lambda \subseteq \mathbb{R}^k$, and $\phi(x,\theta) \in \Lambda_0$ where Λ_0 is a subset of $\mathrm{int}(\Lambda)$. Suppose that $\phi(x,\theta)$ is twice differentiable wrt. θ and that ϕ and its derivatives are continuous functions of (x,θ), then $\phi(x,\theta) : \mathcal{X} \times \bar{B}_0 \to \Lambda_0$ takes values in a compact set. By continuity of κ we have

$$\begin{aligned} g_\theta(y|x)/g_\theta(y|z) &= \exp\left[\{\phi(x,\theta) - \phi(z,\theta)\} \cdot t(y) - \{\kappa(\phi(x,\theta)) - \kappa(\phi(z,\theta))\}\right] \\ &\leq \mathbf{c}_1 \exp(\mathbf{c}_2 |t(y)|). \end{aligned}$$

Then

$$\begin{aligned}\inf_{x\in\mathcal{X}}\int_{\mathcal{Y}} g_0(y|x)/\rho(y)\,\nu(dy) &\geq \mathbf{c}_1^{-1}\inf_{x\in\mathcal{X}}\int_{\mathcal{Y}} e^{\phi(x,\theta_0)\cdot t(y)-\kappa(\phi(x,\theta_0))}e^{-\mathbf{c}_2|t(y)|}\,\nu(dy)\\ &\geq \mathbf{c}_3\inf_{x\in\mathcal{X}}\int_{\mathcal{Y}} e^{-|\phi(x,\theta_0)|\,|t(y)|}e^{-\mathbf{c}_2|t(y)|}\,\nu(dy)\\ &\geq \mathbf{c}_3\int_{\mathcal{Y}} e^{-\mathbf{c}_4|t(y)|}\,\nu(dy)>0,\end{aligned}$$

hence A3 is fulfilled. As for A4(iii) we have

$$D\log g_\theta(y|x)=D\{\phi(x,\theta)\cdot t(y)-\kappa(\phi(x,\theta))\}=\frac{\partial\phi(x,\theta)}{\partial\theta^T}\{t(y)-\tau(\phi(x,\theta))\},$$

where $\tau(\phi)=\frac{\partial\kappa(\phi)}{\partial\phi}$ is the mean of $t(Y)$ under P_ϕ and $\frac{\partial\phi(x,\theta)}{\partial\theta^T}$ denotes the $d\times k$ matrix of partial derivatives of ϕ wrt. θ. Then because of compactness of $\mathcal{X}\times\bar{B}_0$ we get

$$\sup_{\theta\in B_0}\sup_{x\in\mathcal{X}}|D\log g_\theta(y|x)|\leq \mathbf{c}_1+\mathbf{c}_2|t(y)|,$$

and hence

$$E_0(\sup_{\theta\in B_0}\sup_{x\in\mathcal{X}}|D\log g_\theta(y|x)|^2)\leq 2\mathbf{c}_1^2+2\mathbf{c}_2^2E_0(|t(Y)|^2)<\infty.$$

The second derivative $D^2\log g_\theta(y|x)$ can be dominated in the same way, and hence A4 follows. Assumption A5(i) follows again by compactness of the parameter space, and finally A5(ii) follows by the continuity of ϕ.

3. Main results. Our main results are stated in this Section. These are a central limit theorem for the score function and a uniform law of large numbers for the observed information. As a consequence of these we find that with a probability that tends to one as n tends to infinity, there exists a (local) maximum point of the likelihood function, which is consistent in probability and asymptotically normal. If especially the maximum likelihood estimator exists and is consistent, it is asymptotically normal.

Let $l_n(\theta)$ denote the log likelihood function based on observations $Y_1,\ldots,Y_n$. Below, $\mathcal{I}_0$ will denote a Fisher information matrix given by

$$\mathcal{I}_0=E_0(\eta\eta^T),\quad \text{where } \eta=\lim_{n\to\infty}D\log p_0(Y_1\,|\,Y_{-n}^0).$$

This will be formally defined in Section 4, but as the following theorems show it can be thought of as the asymptotic covariance matrix of the score function or the limit of the normed observed information as the number of observations tends to infinity.

THEOREM 3.1 *As n tends to infinity $n^{-1/2}Dl_n(\theta_0)\to N(0,\mathcal{I}_0)$, P_0-weakly.*

This Theorem is proved in Section 4.

Theorem 3.2 *Let $\{\theta_n^*\}$ denote any stochastic sequence in Θ such that $\theta_n^* \to \theta_0$ in P_0-probability as $n \to \infty$. Then $n^{-1}D^2 l_n(\theta_n^*) \to -\mathcal{I}_0$ in P_0-probability as $n \to \infty$.*

This Theorem is proved in Section 5. Having established these two results the following result is proved in Jensen (1986) (see also Sweeting (1980).)

Theorem 3.3 *Assume that $\mathcal{I}_0$ is positive definite. With a P_0-probability that tends to 1 as n tends to infinity, there exists a sequence of local maximum points of the likelihood function $\{\hat{\theta}_n\}$ such that $\hat{\theta}_n \to \theta_0$ in P_0 probability, and*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, \mathcal{I}_0^{-1}), P_0 - \textit{weakly}.$$

If especially the maximum likelihood estimator exists and is consistent in P_0-probability, then this estimator has the same limit distribution.

The proof of the second part of the theorem follows by a Taylor expansion of the likelihood function around θ_0, as the proof of Theorem 1 in Bickel et al. The proof of the first part relies on the assumption that $\mathcal{I}_0$ is positive definite, thus in the limit the likelihood function has negative curvature and hence a local maximum at θ_0.

4. A central limit theorem for the score function. In this Section we prove the central limit theorem for the score function stated in Theorem 3.1. Bickel et al. proved the same result in case where the state space of the latent process is finite. Here we start with some Lemmas which will replace Lemma 4 and 5 in Bickel et al. For notational reasons we will assume that d is equal to one in the rest of this paper. If derivatives are replaced by gradients and second derivatives by Hessians all results are valid for general d.

Lemma 4.1 *Let $J \subseteq \mathbb{Z}$ be an integer set and let $\theta \in B_0$. Conditionally on $Y_J = \{Y_j \,|\, j \in J\}$, X constitute an inhomogeneous Markov chain with $p_\theta(X_k|X_{k-1}, Y_J) \geq \omega_k$, where*

$$\omega_k = \begin{cases} \sigma^2/(M\rho(Y_k)) & \text{if } k \in J \\ \sigma^2/M & \text{if } k \notin J. \end{cases}$$

The inequality is also true for the reversed chain $\{X_{-k}\}_{k\in\mathbb{Z}}$.

Proof The Markov property is proved by considering $n < k < m$, assuming for simplicity that $n \leq j \leq m$ for all $j \in J$. Then

$$\begin{aligned} p_\theta(X_n^{k-1}, X_{k+1}^m \,|\, X_k, Y_J) &= \pi_\theta(X_n) \prod_{i=n}^{m-1} \alpha_\theta(X_i, X_{i+1}) \prod_{j\in J} g_\theta(Y_j|X_j)/p_\theta(X_k, Y_J) \\ &= h_1(X_n^k, Y_J) h_2(X_k^m, Y_J), \end{aligned}$$

where h_1 and h_2 are functions of (X_n^k, Y_J) and (X_k^m, Y_J), respectively. It follows that X_n^{k-1} and X_{k+1}^m are conditionally independent given (X_k, Y_J).

Suppose $k \in J$. Conditionally on X_{k-1} and X_{k+1}, X_k and $Y_{J\setminus\{k\}}$ are independent by definition of the state space model. Hence the conditional density of X_k given (X_{k-1}, X_{k+1}, Y_J) is

$$\begin{aligned} p_\theta(X_k|X_{k-1}, X_{k+1}, Y_J) &= \frac{\alpha_\theta(X_{k-1}, X_k)\alpha_\theta(X_k, X_{k+1})g_\theta(Y_k|X_k)}{\int_{\mathcal{X}} \alpha_\theta(X_{k-1}, x)\alpha_\theta(x, X_{k+1})g_\theta(Y_k|x)\,\mu(dx)} \\ &\geq \sigma^2 \left(M \int_{\mathcal{X}} \alpha_\theta(X_{k-1}, x) \frac{g_\theta(Y_k|x)}{g_\theta(Y_k|X_k)}\,\mu(dx)\right)^{-1} \\ &\geq \sigma^2/(M\rho(Y_k)). \end{aligned} \tag{2}$$

Integrating the conditional density wrt. $p_\theta(X_{k+1}|X_{k-1}, Y_J)$ gives the stated result. When $k \notin J$ the term $g_\theta(Y_k|X_k)$ vanishes.

The proof of the statement for the reversed chain follows by integrating (2) wrt. $p_\theta(X_{k-1}\,|\,X_{k+1}, Y_J)$ instead. $\square$

We state the following Lemma for future reference, leaving the proof to the reader.

LEMMA 4.2 *Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space and let $h\colon \mathcal{X} \to \mathbb{R}$ be a measurable function on $\mathcal{X}$. Let ν_1 and ν_2 be two measures dominated by μ with $\nu_1(\mathcal{X}) = \nu_2(\mathcal{X})$. Then*

$$|\int_{\mathcal{X}} h\,d\nu_1 - \int_{\mathcal{X}} h\,d\nu_2| \leq \{\sup_{x\in\mathcal{X}} h(x) - \inf_{x\in\mathcal{X}} h(x)\}\{\nu_1(S^+) - \nu_2(S^+)\},$$

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

In the next Lemma we will let $\{X_k\}_{k\in\mathbb{Z}}$ denote any inhomogenous Markov chain, that is, $\{X_k\}$ is not necessarily the latent process in the model.

LEMMA 4.3 *Let $\{X_k\}_{k\in\mathbb{Z}}$ be a Markov chain with state space $(\mathcal{X}, \mathcal{A}, \mu)$. Assume*

$$\frac{dP_{X_k|X_{k-1}}}{d\mu}(x|z) = p_k(z, x) \geq \delta_k,$$

for all $x, z \in \mathcal{X}$ and $k \in \mathbb{Z}$, where $P_{X_k|X_{k-1}}$ denotes the conditional distribution of X_k given X_{k-1}. Then for any $A \in \mathcal{A}$,

$$\sup_{\xi\in\mathcal{X}} P(X_n \in A|X_0 = \xi) - \inf_{\eta\in\mathcal{X}} P(X_n \in A|X_0 = \eta) \leq \prod_{k=1}^{n}(1 - \delta_k\mu(\mathcal{X})).$$

PROOF Let $S_k^+ = \{x \in \mathcal{X} \mid p_k(\xi, x) - p_k(\eta, x) > 0\}$ for fixed ξ and η in $\mathcal{X}$, and let $S_k^- = (S_k^+)^c$. Define $M_A^{(k)} = \sup_{\xi\in\mathcal{X}} P(X_n \in A|X_k = \xi)$ and $m_A^{(k)} = \inf_{\eta\in\mathcal{X}} P(X_n \in$

$A|X_k = \eta)$. Then

$$\begin{aligned}
&M_A^{(k-1)} - m_A^{(k-1)} \\
&= \sup_{\xi,\eta} \left(P(X_n \in A | X_{k-1} = \xi) - P(X_n \in A | X_{k-1} = \eta) \right) \\
&= \sup_{\xi,\eta} \int_{\mathcal{X}} P(X_n \in A | X_k = z) \{ p_k(\xi, z) - p_k(\eta, z) \} \, \mu(dz) \\
&\leq \sup_{\xi,\eta} \{ P(X_k \in S_k^+ | X_{k-1} = \xi) - P(X_k \in S_k^+ | X_{k-1} = \eta) \} (M_A^{(k)} - m_A^{(k)}) \\
&= \sup_{\xi,\eta} \{ 1 - P(X_k \in S_k^- | X_{k-1} = \xi) - P(X_k \in S_k^+ | X_{k-1} = \eta) \} (M_A^{(k)} - m_A^{(k)}) \\
&\leq \{ 1 - \delta_k \mu(\mathcal{X}) \} (M_A^{(k)} - m_A^{(k)}),
\end{aligned}$$

where the first inequality follows from Lemma 4.2. The result now follows by induction with $k = n, n-1, \ldots, 1$. (Proof based on Doob (1953, p. 198)) $\square$

We are now ready to prove a result corresponding to Lemma 4 in Bickel et al. Let $\omega(y) = \mu(\mathcal{X})\sigma^2/(M\rho(y))$.

LEMMA 4.4 *Let $k < l$ and let $J \subseteq \mathbb{Z}$ such that $\{k, k+1, ..., l-1\} \subseteq J$. Let $Y_J = \{Y_j \mid j \in J\}$ then for any $\theta \in B_0$,*

$$\sup_{A \in \mathcal{A}} \sup_{\xi,\eta \in \mathcal{X}} |P_\theta(X_k \in A \mid Y_J, X_l = \xi) - P_\theta(X_k \in A \mid Y_J, X_l = \eta)| \leq \prod_{i=k}^{l-1} (1 - \omega(Y_i)).$$

Likewise, if $l < k$ and $\{l+1, l+2, \ldots, k\} \subseteq J$ then

$$\sup_{A \in \mathcal{A}} \sup_{\xi,\eta \in \mathcal{X}} |P_\theta(X_k \in A \mid Y_J, X_l = \xi) - P_\theta(X_k \in A \mid Y_J, X_l = \eta)| \leq \prod_{i=l+1}^{k} (1 - \omega(Y_i)).$$

PROOF Consider the case $k < l$. Applying Lemma 4.1 on the reversed chain $\{X_{-k}\}_{k \in \mathbb{Z}}$ we get

$$p_\theta(X_i \mid X_{i+1}, Y_J) \geq \sigma^2/(M\rho(Y_i)) = \omega(Y_i)/\mu(\mathcal{X}) \quad \text{for } i = k, \ldots, l-1.$$

Using Lemma 4.3 with $\delta_i = \omega(Y_i)/\mu(\mathcal{X})$ we get the stated result. The proof is similar when $l < k$, applying Lemma 4.1 on the original chain $\{X_k\}_{k \in \mathbb{Z}}$. $\square$

LEMMA 4.5 *Let $-m \leq -n \leq k \leq 0$. For any θ in B_0 and any $A, B \in \mathcal{A}$ we have*

$$|P_\theta(X_k \in A \,|\, Y_{-n}^1) - P_\theta(X_k \in A \,|\, Y_{-n}^0)| \leq \prod_{i=k}^{0}(1-\omega(Y_i)),$$

$$|P_\theta(X_k \in A, X_{k+1} \in B \,|\, Y_{-n}^1) - P_\theta(X_k \in A, X_{k+1} \in B \,|\, Y_{-n}^0)| \leq \prod_{i=k+1}^{0}(1-\omega(Y_i)),$$

$$|P_\theta(X_k \in A \,|\, Y_{-n}^1) - P_\theta(X_k \in A \,|\, Y_{-m}^1)| \leq \prod_{i=-n}^{k}(1-\omega(Y_i)),$$

$$|P_\theta(X_k \in A, X_{k+1} \in B \,|\, Y_{-n}^1) - P_\theta(X_k \in A, X_{k+1} \in B \,|\, Y_{-m}^1)| \leq \prod_{i=-n}^{k}(1-\omega(Y_i)).$$

The first and second expression hold P_θ-almost surely if n is replaced by ∞. The third and fourth hold P_θ-almost surely if m is replaced by ∞ and for both we can replace Y_{-n}^1 and Y_{-m}^1 by Y_{-n}^0 and Y_{-m}^0, respectively.

PROOF The first expression can be evaluated as

$$\begin{aligned}
&|P_\theta(X_k \in A \,|\, Y_{-n}^1) - P_\theta(X_k \in A \,|\, Y_{-n}^0)| \\
&= |\int_{\mathcal{X}} P_\theta(X_k \in A \,|\, Y_{-n}^0, x_1)\{p_\theta(x_1 \,|\, Y_{-n}^1) - p_\theta(x_1 \,|\, Y_{-n}^0)\}\,\mu(dx_1)| \\
&\leq \sup_{\xi \in \mathcal{X}} P_\theta(X_k \in A \,|\, Y_{-n}^0, X_1 = \xi) - \inf_{\eta \in \mathcal{X}} P_\theta(X_k \in A \,|\, Y_{-n}^0, X_1 = \eta) \leq \prod_{i=k}^{0}(1-\omega(Y_i)),
\end{aligned}$$

where the inequalities follows from Lemma 4.2 and 4.4, respectively. As for the second expression,

$$\begin{aligned}
&|P_\theta(X_k \in A, X_{k+1} \in B \,|\, Y_{-n}^1) - P_\theta(X_k \in A, X_{k+1} \in B \,|\, Y_{-n}^0)| \\
&= |\int_B P_\theta(X_k \in A | x_{k+1}, Y_{-n}^k)\{P_\theta(x_{k+1} \,|\, Y_{-n}^1) - P_\theta(x_{k+1} \,|\, Y_{-n}^0)\}\,\mu(dx_{k+1})| \\
&\leq |P_\theta(X_{k+1} \in S^+ \,|\, Y_{-n}^1) - P_\theta(X_{k+1} \in S^+ \,|\, Y_{-n}^0)| \leq \prod_{i=k+1}^{0}(1-\omega(Y_i)).
\end{aligned}$$

Here S^+ is a set chosen as in Lemma 4.2 and the second inequality follows from above.

By a martingale convergence theorem by Levy (Hoffmann-Jørgensen 1994, p. 505) we get, for instance, that $P_\theta(X_k \in A \,|\, Y_{-n}^1) \to P_\theta(X_k \in A \,|\, Y_{-\infty}^1)$ P_θ-almost surely as $n \to \infty$. This result shows that we can replace n by ∞ in the inequalities above.

The third expression is proved as the first by conditioning on $X_{-n-1} = x_{-n-1}$ in the integral, and the fourth expression follows from the third by an argument similar to the one used to deduce the second from the first. The arguments are identical when replacing Y_{-n}^1 and Y_{-m}^1 with Y_{-n}^0 and Y_{-m}^0, and the extension to the case $m = \infty$ follows from the martingale convergence argument above. □

Lemma 4.5 corresponds to Lemma 5 in Bickel et al. Having established this result the rest of the proof of the CLT for the score function follows the line of these authors closely. However, we will repeat some of the arguments here since there are some differences due to our latent process being continuous.

We will for notational reasons denote our observations $Y_{-n}, \ldots, Y_1$. The score function $Dl(\theta)$ is then given by

$$Dl(\theta) = \sum_{k=-n}^{1} D \log p_\theta(Y_k \,|\, Y_{-n}^{k-1}),$$

where $p_\theta(Y_k \,|\, Y_{-n}^{k-1})$ denotes the conditional density of Y_k given Y_{-n}^{k-1} given by

$$D \log p_\theta(Y_k \,|\, Y_{-n}^{k-1}) = D \log p_\theta(Y_{-n}^{k}) - D \log p_\theta(Y_{-n}^{k-1}).$$

Using assumption A5(i) to interchange integration and differentiation below we find that for any $j = k-1, k$,

$$D \log p_\theta(Y_{-n}^{j}) = E_\theta(D \log p_\theta(Y_{-n}^{j}, X_{-n}^{k}) \,|\, Y_{-n}^{j}).$$

Hence $D \log p_\theta(Y_k \,|\, Y_{-n}^{k-1})$ is given by

$$\begin{aligned} &D \log p_\theta(Y_k \,|\, Y_{-n}^{k-1}) \\ &\qquad = E_\theta(D \log p_\theta(Y_{-n}^{k}, X_{-n}^{k}) \,|\, Y_{-n}^{k}) - E_\theta(D \log p_\theta(Y_{-n}^{k-1}, X_{-n}^{k}) \,|\, Y_{-n}^{k-1}). \end{aligned} \tag{3}$$

Using the expression for $p_\theta(Y_{-n}^{k}, X_{-n}^{k})$ in (1) we find

$$\begin{aligned} &D \log p_\theta(Y_k \,|\, Y_{-n}^{k-1}) = \\ &\sum_{i=-n}^{k-1} \Big\{ E_\theta(\lambda_\theta(X_i, X_{i+1}) + \gamma_\theta(Y_i|X_i)|Y_{-n}^{k}) - E_\theta(\lambda_\theta(X_i, X_{i+1}) + \gamma_\theta(Y_i|X_i)|Y_{-n}^{k-1}) \Big\} \\ &\qquad + E_\theta(\tau_\theta(X_{-n}) \,|\, Y_{-n}^{k}) - E_\theta(\tau_\theta(X_{-n}) \,|\, Y_{-n}^{k-1}) + E_\theta(\gamma_\theta(Y_k|X_k) \,|\, Y_{-n}^{k}). \end{aligned} \tag{4}$$

Now, let

$$\begin{aligned} \eta_1 = \sum_{i=-\infty}^{0} &\Big\{ E_0(\lambda_0(X_i, X_{i+1}) + \gamma_0(Y_i|X_i)|Y_{-\infty}^{1}) - \\ &E_0(\lambda_0(X_i, X_{i+1}) + \gamma_0(Y_i|X_i)|Y_{-\infty}^{0}) \Big\} + E_0(\gamma_0(Y_1|X_1) \,|\, Y_{-\infty}^{1}). \end{aligned} \tag{5}$$

The infinite sum is absolutely convergent in $\mathbb{L}^2(P_0)$, as will be shown in Lemma 4.6, so η_1 is a well defined variable in $\mathbb{L}^2(P_0)$. Let

$$\mathcal{I}_0 = E_0(\eta_1^2).$$

Letting $||\cdot||_2$ denote the $\mathbb{L}^2(P_0)$-norm we have:

LEMMA 4.6 *There exists a* $\beta \in [0,1)$ *and a constant* $\mathbf{c}$ *such that*

$$||D \log p_0(Y_1 \mid Y_{-n}^0) - \eta_1||_2 \le \mathbf{c}\beta^n,$$

for all n.

PROOF Let

$$Z_k = \lambda_0(X_k, X_{k+1}) + \gamma_0(Y_k|X_k).$$

By splitting the sums in (4) and (5) we can dominate $||D \log p_0(Y_1 \mid Y_{-n}^0) - \eta_1||_2$ by the sum of the following terms:

$$||E_0(\tau_0(X_{-n}) \mid Y_{-n}^1) - E_0(\tau_0(X_{-n}) \mid Y_{-n}^0)||_2, \tag{6}$$

$$||E_0(\gamma_0(Y_1|X_1) \mid Y_{-n}^1) - E_0(\gamma_0(Y_1|X_1) \mid Y_{-\infty}^1)||_2, \tag{7}$$

$$\sum_{k=-[n/2]}^{0} ||E_0(Z_k \mid Y_{-n}^j) - E_0(Z_k \mid Y_{-\infty}^j)||_2, \quad j = 0, 1, \tag{8}$$

$$\sum_{k=-n}^{-[n/2]-1} ||E_0(Z_k \mid Y_{-n}^1) - E_0(Z_k \mid Y_{-n}^0)||_2, \tag{9}$$

$$\sum_{k=-\infty}^{-[n/2]} ||E_0(Z_k \mid Y_{-\infty}^1) - E_0(Z_k \mid Y_{-\infty}^0)||_2, \tag{10}$$

where $[\cdot]$ denotes the integer part. We will show that each of the terms (6)–(10) can be dominated by $\mathbf{c}\beta^n$, where $0 \le \beta < 1$, which proves the Lemma. Furthermore, the domination of (10) shows that the sum in (5) is absolutely convergent as stated earlier.

We will show the domination of (9) and leave the remaining terms to the reader. We will first consider the part of Z_k given by $\gamma_0(Y_k|X_k)$ in (9). By applying Lemma 4.2 and 4.5 we have the following inequality:

$$\begin{aligned}
&\left|E_0(\gamma_0(Y_k|X_k) \mid Y_{-n}^1) - E_0(\gamma_0(Y_k|X_k) \mid Y_{-n}^0)\right| \\
&= \left|\int_{\mathcal{X}} \gamma_0(Y_k|x_k)\{p_0(x_k \mid Y_{-n}^1) - p_0(x_k \mid Y_{-n}^0)\}\, \mu(dx_k)\right| \\
&\le 2 \sup_{x \in \mathcal{X}} |\gamma_0(Y_k|x)| \prod_{i=k+1}^{0} (1 - \omega(Y_i)).
\end{aligned}$$

Hence the $\mathbb{L}^2$-norm can be dominated as

$$\begin{aligned}
&||E_0(\gamma_0(Y_k|X_k) \mid Y_{-n}^1) - E_0(\gamma_0(Y_k|X_k) \mid Y_{-n}^0)||_2^2 \\
&\le 4E_0\left(E_0\left(\sup_{x \in \mathcal{X}} \gamma_0(Y_k|x)^2 \prod_{i=k+1}^{0} (1 - \omega(Y_i))^2 \middle| X_k^0\right)\right) \\
&= 4E_0\left(E_0(\sup_{x \in \mathcal{X}} \gamma_0(Y_k|x)^2|X_k) \prod_{i=k+1}^{0} E_0((1 - \omega(Y_i))^2|X_i)\right) \\
&\le \mathbf{c}\beta^{-k},
\end{aligned} \tag{11}$$

where the equality follows by definition of the state space model and where β is given by

$$\begin{aligned}\beta &= \sup_{x\in\mathcal{X}} \int_{\mathcal{Y}} \left(1 - \frac{\mu(\mathcal{X})\sigma^2}{M\rho(y)}\right)^2 g_0(y|x)\,\nu(dy)\\ &\le \sup_{x\in\mathcal{X}} \int_{\mathcal{Y}} \left(1 - \frac{\mu(\mathcal{X})\sigma^2}{M\rho(y)}\right) g_0(y|x)\,\nu(dy)\\ &= 1 - \frac{\mu(\mathcal{X})\sigma^2}{M} \inf_{x\in\mathcal{X}} \int_{\mathcal{Y}} g_0(y|x)/\rho(y)\,\nu(dy) < 1,\end{aligned}$$

by assumption A3. The constant in (11) is finite by assumption A4. For a sum of $\mathbb{L}^2$-norms we get,

$$\sum_{k=-n}^{-[n/2]-1} ||E_0(\gamma_0(Y_k|X_k)\,|\,Y_{-n}^1) - E_0(\gamma_0(Y_k|X_k)\,|\,Y_{-n}^0)||_2 \le \mathbf{c} \sum_{k=-n}^{-[n/2]-1} \beta^{-k/2} \le \mathbf{c}\beta^{[n/2]/2}.$$

The part of (9) involving $\lambda_0(X_k, X_{k+1})$ can be dominated in a similar way, using A4 and the second inequality in Lemma 4.4. Hence we have proved the claimed domination of (9). $\square$

Lemma 4.6 is the final brick needed to prove Theorem 3.1; it tells us that in the limit the score function function is equivalent to a sum of terms like η_1. These constitute a stationary martingale increment sequence, and hence by a martingal central limit theorem we obtain the stated limit distribution of the score function. The proof is identical to the proof of Lemma 1 in Bickel et al. (p. 1626)

5. A law of large numbers for the observed information. In this Section we will show Theorem 3.2. As in the previous Section we will start with some Lemmas providing inequalities for conditional probabilities. Lemmas 5.1 and 5.3 are multivariate versions of Lemma 4.5.

LEMMA 5.1 *Let $-m \le -n \le k \le l \le 0$, and let $\theta \in B_0$. Then for all $C \in \sigma\{(X_t, Y_t) : t \le l\}$ we have*

$$|P_\theta(C\,|\,Y_{-n}^1) - P_\theta(C\,|\,Y_{-n}^0)| \le \prod_{i=l}^{0}(1-\omega(Y_i)).$$

Likewise for all $C \in \sigma\{(X_t, Y_t) : t \ge k\}$ and for $j = 0, 1$ we have,

$$|P_\theta(C\,|\,Y_{-n}^j) - P_\theta(C\,|\,Y_{-m}^j)| \le \prod_{i=-n}^{k}(1-\omega(Y_i)).$$

PROOF Let $C \in \sigma\{(X_t, Y_t), t \le l\}$. Then

$$\begin{aligned}
&|P_\theta(C \,|\, Y_{-n}^1) - P_\theta(C \,|\, Y_{-n}^0)| \\
&= |\int_{\mathcal{X}} P_\theta(C \,|\, x_l, Y_{-n}^l)\{p_\theta(x_l \,|\, Y_{-n}^1) - p_\theta(x_l \,|\, Y_{-n}^0)\}\, \mu(dx_l)| \\
&\le P_\theta(X_l \in S^+ \,|\, Y_{-n}^1) - P_\theta(X_l \in S^+ \,|\, Y_{-n}^0) \le \prod_{i=l}^{0}(1-\omega(Y_i)),
\end{aligned}$$

where $S^+ = \{x_l \in \mathcal{X} \mid p_\theta(x_l \,|\, Y_{-n}^1) - p_\theta(x_l \,|\, Y_{-n}^0) > 0\}$ is chosen as in Lemma 4.2, and the last inequality follows from 4.5. The second inequality is derived by a similar argument, by conditioning on X_k instead of X_l. □

In the next Lemma $\{X_k\}$ denotes any inhomogenous Markov chain, as in Lemma 4.3.

LEMMA 5.2 *Let the setup be as in Lemma 4.3. Let $n \in \mathbb{Z}$ and let Q be the measure on $\mathcal{A} \otimes \mathcal{A}$ defined by,*

$$Q(A \times B) = P(X_0 \in A)P(X_n \in B),$$

for $A, B \in \mathcal{A}$. Then for all $C \in \mathcal{A} \otimes \mathcal{A}$,

$$|P((X_0, X_n) \in C) - Q(C)| \le \prod_{k=1}^{n}(1-\delta_k \mu(\mathcal{X})).$$

PROOF Let $C_{x_0} = \{x_n \in \mathcal{X} \mid (x_0, x_n) \in C\}$, then

$$\begin{aligned}
&|P((X_0, X_n) \in C) - Q(C)| \\
&\quad = |\int_{\mathcal{X}} \{P(X_n \in C_{x_0} \,|\, X_0 = x_0) - P(X_n \in C_{x_0})\}\, P_{X_0}(dx_0)| \le \prod_{k=1}^{n}(1-\delta_k \mu(\mathcal{X})).
\end{aligned}$$

Here the last inequality follows from Lemma 4.3 since

$$\begin{aligned}
&|P(X_n \in A \,|\, X_0 = \xi) - P(X_n \in A)| \\
&= |\int_{\mathcal{X}} \{P(X_n \in A \,|\, X_0 = \xi) - P(X_n \in A \,|\, X_0 = \eta)\}\, P_{X_0}(d\eta) \\
&= \sup_{\xi \in \mathcal{X}} P(X_n \in A \,|\, X_0 = \xi) - \inf_{\eta \in \mathcal{X}} P(X_n \in A \,|\, X_0 = \eta) \le \prod_{k=1}^{n}(1-\delta_k \mu(\mathcal{X})).
\end{aligned}$$

□

LEMMA 5.3 *Let $-m \le -n \le k \le l \le 0$. Let $Q_{\theta,-n}^j$ be the measure on $\mathcal{A} \otimes \mathcal{A}$ defined by*

$$Q_{\theta,-n}^j(A \times B) = P_\theta(X_k \in A \,|\, Y_{-n}^j)P_\theta(X_l \in B \,|\, Y_{-n}^j)$$

for $j = 0, 1$ *and* $A, B \in \mathcal{A}$. *Then for all* $\theta \in B_0$, *for* $C \in \mathcal{A} \otimes \mathcal{A}$ *and for* $j = 0, 1$,

$$|P_\theta((X_k, X_l) \in C \mid Y^j_{-n}) - Q^j_{\theta,-n}(C)| \leq \prod_{i=k}^{l-1}(1 - \omega(Y_i)),$$

$$|Q^1_{\theta,-n}(C) - Q^0_{\theta,-n}(C)| \leq 2\prod_{i=l}^{0}(1 - \omega(Y_i)),$$

$$|Q^j_{\theta,-n}(C) - Q^j_{\theta,-m}(C)| \leq 2\prod_{i=-n}^{k}(1 - \omega(Y_i)).$$

Proof The first inequality follows from Lemma 5.2 and 4.1. To prove the second expression we will let $C_y = \{x \in \mathcal{X} \mid (x, y) \in C\}$, $C'_x = \{y \in \mathcal{X} \mid (x, y) \in C\}$ and proceed as follows,

$$\begin{aligned}
&|Q^1_{\theta,-n}(C) - Q^0_{\theta,-n}(C)| \\
&= |\int_C \{p_\theta(x_k \mid Y^1_{-n}) p_\theta(x_l \mid Y^1_{-n}) - p_\theta(x_k \mid Y^0_{-n}) p_\theta(x_l \mid Y^0_{-n})\}\, \mu(dx_k)\mu(dx_l)| \\
&\leq |\int_C \{p_\theta(x_k \mid Y^1_{-n}) - p_\theta(x_k \mid Y^0_{-n})\}\, p_\theta(x_l \mid Y^1_{-n}) \mu(dx_k)\mu(dx_l)| \\
&\quad + |\int_C \{p_\theta(x_l \mid Y^1_{-n}) - p_\theta(x_l \mid Y^0_{-n})\}\, p_\theta(x_k \mid Y^0_{-n})\, \mu(dx_l)\mu(dx_k)| \\
&\leq \int_{\mathcal{X}} |P_\theta(X_k \in C_{x_l} \mid Y^1_{-n}) - P_\theta(X_k \in C_{x_l} \mid Y^0_{-n})|\, p_\theta(x_l|Y^1_{-n})\, \mu(dx_l) + \\
&\quad + \int_{\mathcal{X}} |P_\theta(X_l \in C'_{x_k} \mid Y^1_{-n}) - P_\theta(X_l \in C'_{x_k} \mid Y^0_{-n})|\, p_\theta(x_k|Y^0_{-n})\, \mu(dx_k) \\
&\leq \prod_{i=k}^{0}(1 - \omega(Y_i)) + \prod_{i=l}^{0}(1 - \omega(Y_i)) \leq 2\prod_{i=l}^{0}(1 - \omega(Y_i)),
\end{aligned}$$

where the third inequality is given by Lemma 4.5.

The third expression is proved as the second. □

Having established these inequalities we are ready to prove the law of large numbers for the observed information. Using A5(i) to interchange integration and differentiation we find

$$\begin{aligned}
&D^2 \log p_\theta(Y_1 \mid Y^0_{-n}) = D^2 \log p_\theta(Y^1_{-n}) - D^2 \log p_\theta(Y^0_{-n}) \\
&= E_\theta(D^2 \log p_\theta(X^1_{-n}, Y^1_{-n}) \mid Y^1_{-n}) - E_\theta(D^2 \log p_\theta(X^1_{-n}, Y^0_{-n}) \mid Y^0_{-n}) \\
&\quad + \mathrm{var}_\theta(D \log p_\theta(X^1_{-n}, Y^1_{-n}) \mid Y^1_{-n}) - \mathrm{var}_\theta(D \log p_\theta(X^1_{-n}, Y^0_{-n}) \mid Y^0_{-n})
\end{aligned} \tag{12}$$

Define for notational reasons

$$Z_{\theta,k} = \lambda_\theta(X_k, X_{k+1}) + \gamma_\theta(Y_k|X_k) \text{ and } \dot{Z}_{\theta,k} = D\lambda_\theta(X_k, X_{k+1}) + D\gamma_\theta(Y_k|X_k).$$

Inserting expression (1) in (12) we get,

$$
\begin{aligned}
&D^2 \log p_\theta(Y_1 \mid Y_{-n}^0) \qquad (13)\\
&= E_\theta(D\tau_\theta(X_{-n}) \mid Y_{-n}^1) - E_\theta(D\tau_\theta(X_{-n}) \mid Y_{-n}^0) + E_\theta(D\gamma_\theta(Y_1|X_1) \mid Y_{-n}^1)\\
&\quad + \sum_{k=-n}^{0} \{E_\theta(\dot{Z}_{\theta,k} \mid Y_{-n}^1) - E_\theta(\dot{Z}_{\theta,k} \mid Y_{-n}^0)\}\\
&\quad + \mathrm{var}_\theta(\gamma_\theta(Y_1|X_1) \mid Y_{-n}^1) + \mathrm{var}_\theta(\tau_\theta(X_{-n}) \mid Y_{-n}^1) - \mathrm{var}_\theta(\tau_\theta(X_{-n}) \mid Y_{-n}^0)\\
&\quad + \sum_{k=-n}^{0} \sum_{l=-n}^{0} \Big\{\mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-n}^1) - \mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-n}^0)\Big\}\\
&\quad + 2\sum_{k=-n}^{0} \Big\{\mathrm{cov}_\theta(\tau_\theta(X_{-n}), Z_{\theta,k} \mid Y_{-n}^1) - \mathrm{cov}_\theta(\tau_\theta(X_{-n}), Z_{\theta,k} \mid Y_{-n}^0)\Big\}\\
&\quad + 2\sum_{k=-n}^{0} \mathrm{cov}_\theta(\gamma_\theta(Y_1|X_1), Z_{\theta,k} \mid Y_{-n}^1) + 2\mathrm{cov}_\theta(\gamma_\theta(Y_1|X_1), \tau_\theta(X_{-n}) \mid Y_{-n}^1).
\end{aligned}
$$

We then have the following convergence result.

Lemma 5.4 *As $m, n \to \infty$,*

$$\left|\left|\sup_{\theta \in B_0} |D^2 \log p_\theta(Y_1 \mid Y_{-n}^0) - D^2 \log p_\theta(Y_1 \mid Y_{-m}^0)|\right|\right|_1 \to 0,$$

where $||\cdot||_1$ denotes the $\mathbb{L}^1(P_0)$-norm.

This Lemma states that $\{D^2 \log p_\theta(Y_1 \mid Y_{-n}^0)\}$ is a uniform Cauchy sequence in $\mathbb{L}^1(P_0)$. This is important because it proves the existence of a limit in $\mathbb{L}^1(P_0)$ of $D^2 \log p_\theta(Y_1 \mid Y_{-n}^1)$ as $n \to \infty$ for any $\theta \in B_0$, and not only $\theta = \theta_0$. In the proof we will need the following Lemma.

Lemma 5.5 *Let $-m \le -n \le k \le l \le 0$ and let $Z_{\theta,k}$ be defined as above. Then there exists a $\beta \in [0,1)$ such that the following inequalities hold for $j = 0, 1$,*

$$\left|\left|\sup_{\theta \in B_0} |\mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-n}^1) - \mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-n}^0)|\right|\right|_1 \le \mathbf{c}\beta^{-l}, \tag{14}$$

$$\left|\left|\sup_{\theta \in B_0} |\mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-n}^j) - \mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-m}^j)|\right|\right|_1 \le \mathbf{c}\beta^{k+n}, \tag{15}$$

$$\left|\left|\sup_{\theta \in B_0} |\mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-n}^j)|\right|\right|_1 \le \mathbf{c}\beta^{l-k}. \tag{16}$$

Above $Z_{\theta,i}$ may be replaced by $\tau_\theta(X_i)$ or $\gamma_\theta(Y_i \mid X_i)$ for $i = k, l$.

Proof Recall that $Z_{\theta,k} = \lambda_\theta(X_k, X_{k+1}) + \gamma_\theta(Y_k \mid X_k)$. Thus the covariance of $Z_{\theta,k}$ and $Z_{\theta,l}$ splits into the sum of four covariance terms involving λ_θ and γ_θ. We will show that $\mathrm{cov}_\theta(\gamma_\theta(Y_k \mid X_k), \gamma_\theta(Y_l \mid X_l) \mid Y)$ satisfies the claimed inequalities. The three remaining terms are similar.

To show the first inequality we will consider the expression

$$\begin{aligned}
&\sup_{\theta\in B_0} |E_\theta\{\gamma_\theta(Y_k \mid X_k)\gamma_\theta(Y_l \mid X_l) \mid Y_{-n}^1\} - E_\theta\{\gamma_\theta(Y_k \mid X_k)\gamma_\theta(Y_l \mid X_l) \mid Y_{-n}^0\}| \\
&= \sup_{\theta\in B_0} |\int_{\mathcal{X}^2} \gamma_\theta(Y_k \mid x_k)\gamma_\theta(Y_l \mid x_l)\{p_\theta(x_k, x_l \mid Y_{-n}^1) - p_\theta(x_k, x_l \mid Y_{-n}^0)\}\,\mu(dx_k)\mu(dx_l)| \\
&\le 2\gamma^*(Y_k)\gamma^*(Y_l) \sup_{\theta\in B_0} |P_\theta((X_k, X_l) \in S^+ \mid Y_{-n}^1) - P_\theta((X_k, X_l) \in S^+ \mid Y_{-n}^0)| \\
&\le 2\gamma^*(Y_k)\gamma^*(Y_l) \prod_{i=l}^{0}(1 - \omega(Y_i)).
\end{aligned}$$

Here $\gamma^*(Y_k) = \sup_{\theta\in B_0}\sup_{x\in\mathcal{X}} |\gamma_\theta(Y_k|x)|$ as defined in assumption A4, and the inequalities follows from Lemma 4.2 and 5.1, respectively. The $\mathbb{L}^1(P_0)$-norm of such a term is thus less than

$$\begin{aligned}
&2E_0\left(\gamma^*(Y_k)\gamma^*(Y_l)\prod_{i=l}^{0}(1 - \omega(Y_i))\right) \\
&= 2E_0\left(E_0\left(\gamma^*(Y_k)\gamma^*(Y_l)\prod_{i=l}^{0}(1 - \omega(Y_i))\middle| X_k^0\right)\right) \\
&\le 2E_0\left(E_0(\gamma^*(Y_k)|X_k)E_0(\gamma^*(Y_l)|X_l)\prod_{i=l+1}^{0} E_0(1 - \omega(Y_i)|X_i)\right) \\
&\le 2E_0(\gamma^*(Y_1)^2)\beta^{-l} = \mathbf{c}\beta^{-l},
\end{aligned}$$

where the first inequality follows by definition of the state space model, and where β is given by

$$\beta = \sup_{x\in\mathcal{X}} E_0(1 - \omega(Y_i)|X_i = x) = \sup_{x\in\mathcal{X}} \int_{\mathcal{Y}} \left(1 - \frac{\mu(\mathcal{X})\sigma^2}{M\rho(y)}\right) g_0(y|x)\,\nu(dy) < 1,$$

by assumption A3. Assumption A4 assures that the constant $\mathbf{c}$ above is finite.

The expression

$$\begin{aligned}
\Bigg\|&\sup_{\theta\in B_0} |E_\theta\{\gamma_\theta(Y_k \mid X_k) \mid Y_{-n}^1\}E_\theta\{\gamma_\theta(Y_l \mid X_l) \mid Y_{-n}^1\} - \\
&\qquad E_\theta\{\gamma_\theta(Y_k \mid X_k) \mid Y_{-n}^0\}E_\theta\{\gamma_\theta(Y_l \mid X_l) \mid Y_{-n}^0\}|\Bigg\|_1
\end{aligned}$$

can be dominated by the same technique, using the second expression in Lemma 5.3. Hence (14) is proved.

The second inequality (15) is proved as (14) using the second expression in Lemma 5.1 and the third expression in Lemma 5.3, respectively. As for (16) we have

$$
\begin{aligned}
&\sup_{\theta\in B_0} |\mathrm{cov}_\theta(\gamma_\theta(Y_k \mid X_k), \gamma_\theta(Y_l \mid X_l) \mid Y_{-n}^j)| \\
&= \sup_{\theta\in B_0} \Big| \int_{\mathcal{X}^2} \gamma_\theta(Y_k \mid x_k)\gamma_\theta(Y_l \mid x_l) \\
&\qquad\qquad \{p_\theta(x_k, x_l \mid Y_{-n}^j) - p_\theta(x_k \mid Y_{-n}^j)p_\theta(x_l \mid Y_{-n}^j)\}\,\mu(dx_k)\mu(dx_l)\Big| \\
&\le 2\gamma^*(Y_k)\gamma^*(Y_l)\prod_{i=k}^{l-1}(1-\omega(Y_i)),
\end{aligned}
$$

by the first expression in Lemma 5.3. The claimed domination of the $\mathbb{L}^1(P_0)$-norm of this term is proved as above. □

PROOF (Lemma 5.4.) Considering the expression for $D^2 \log p_\theta(Y_1 \mid Y_{-n}^0)$ in (13) we will show that the term

$$
\begin{aligned}
&\Bigg\| \sup_{\theta\in B_0} \Bigg| \sum_{k=-m}^{0}\sum_{l=-m}^{0} \Big\{\mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-m}^1) - \mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-m}^0)\Big\} \\
&\qquad - \sum_{k=-n}^{0}\sum_{l=-n}^{0} \Big\{\mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-n}^1) - \mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-n}^0)\Big\} \Bigg| \Bigg\|_1 \to 0 \quad (17)
\end{aligned}
$$

as $n, m \to \infty$. The remaining terms in (13) can be treated with similar arguments.

Suppose $m > n$. By symmetry of k and l in the sum in (17) it suffices to consider the sum over the region where $k \le l$. This region can be further divided into 5 subregions,

$$
\begin{aligned}
D_1 &= \{(k,l) \in \mathbb{Z}^2 \mid -[n/2] \le k \le 0, k \le l \le 0\}, \\
D_2 &= \{(k,l) \in \mathbb{Z}^2 \mid -n \le k \le -[n/2], [k/2] \le l \le 0\}, \\
D_3 &= \{(k,l) \in \mathbb{Z}^2 \mid -m \le k \le -n, [k/2] \le l \le 0\}, \\
D_4 &= \{(k,l) \in \mathbb{Z}^2 \mid -n \le k \le -[n/2], k \le l \le [k/2]\}, \\
D_5 &= \{(k,l) \in \mathbb{Z}^2 \mid -m \le k \le -n, k \le l \le [k/2]\}.
\end{aligned}
$$

We will show that the sum over each of these regions tends to zero in $\mathbb{L}^1(P_0)$ as $n, m \to \infty$, hence proving (17). Using (15) we find that

$$
\begin{aligned}
&\sum_{(k,l)\in D_1} \Bigg\| \sup_{\theta\in B_0} |\{\mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-m}^1) - \mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-m}^0)\} \\
&\qquad - \{\mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-n}^1) - \mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \mid Y_{-n}^0)\}| \Bigg\|_1 \le \mathbf{c} \sum_{k=-[n/2]}^{0}\sum_{l=k}^{0} \beta^{k+n}
\end{aligned}
$$

Using (16) we find that the corresponding sums over D_2 and D_3 are less than

$$
\mathbf{c} \sum_{k=-n}^{-[n/2]} \sum_{l=[k/2]}^{0} \beta^{l-k} \qquad \text{and} \qquad \mathbf{c} \sum_{k=-m}^{-n} \sum_{l=[k/2]}^{0} \beta^{l-k},
$$

respectively, and by (14) the sums over D_4 and D_5 are dominated by

$$\mathbf{c}\sum_{k=-n}^{-[n/2]}\sum_{l=k}^{[k/2]}\beta^{-l} \qquad \text{and} \qquad \mathbf{c}\sum_{k=-m}^{-n}\sum_{l=k}^{[k/2]}\beta^{-l},$$

respectively. Since $0 \le \beta < 1$ these sums all tend to zero as $n, m \to \infty$ and the proof is complete. $\square$

Lemma 5.6 *The map $\theta \mapsto D^2 \log p_\theta(Y_1|Y_{-n}^0)$ from B_0 to $\mathbb{L}^1(P_0)$ is continuous.*

Proof Let $\{\theta_m\} \subseteq B_0$ be a sequence such that $\theta_m \to \theta$ as $m \to \infty$. We will show that $E_0\{|D^2 \log p_{\theta_m}(Y_1 \,|\, Y_{-n}^0) - D^2 \log p_\theta(Y_1 \,|\, Y_{-n}^0)|\} \to 0$, as $m \to \infty$. Considering the expression in (13) we must show that terms like, for instance,

$$E_0\Big[|E_{\theta_m}\{\gamma_{\theta_m}(Y_k|X_k)\gamma_{\theta_m}(Y_l|X_l) \,|\, Y_{-n}^1\} - E_\theta\{\gamma_\theta(Y_k|X_k)\gamma_\theta(Y_l|X_l) \,|\, Y_{-n}^1\}|\Big]$$

tend to zero as $m \to \infty$. The integrand can be evaluated as

$$\begin{aligned}
&|E_{\theta_m}\{\gamma_{\theta_m}(Y_k|X_k)\gamma_{\theta_m}(Y_l|X_l) \,|\, Y_{-n}^1\} - E_\theta\{\gamma_\theta(Y_k|X_k)\gamma_\theta(Y_l|X_l) \,|\, Y_{-n}^1\}| \\
&\le |\int_{\mathcal{X}^2} \gamma_{\theta_m}(Y_k|x_k)\gamma_{\theta_m}(Y_l|x_l)\{p_{\theta_m}(x_k, x_l \,|\, Y_{-n}^1) - p_\theta(x_k, x_l \,|\, Y_{-n}^1)\}\, \mu(dx_k)\mu(dx_l)| \\
&+ |\int_{\mathcal{X}^2} \{\gamma_{\theta_m}(Y_k|x_k)\gamma_{\theta_m}(Y_l|x_l) - \gamma_\theta(Y_k|x_k)\gamma_\theta(Y_l|x_l)\}p_\theta(x_k, x_l \,|\, Y_{-n}^1)\, \mu(dx_k)\mu(dx_l)|.
\end{aligned}$$

The first term is less than

$$\gamma^*(Y_k)\gamma^*(Y_l)\int_{\mathcal{X}^2} |p_{\theta_m}(x_k, x_l \,|\, Y_{-n}^1) - p_\theta(x_k, x_l \,|\, Y_{-n}^1)|\, \mu(dx_k)\mu(dx_l) \tag{18}$$

$$= \frac{\gamma^*(Y_k)\gamma^*(Y_l)}{p_{\theta_m}(Y_{-n}^1)}\int_{\mathcal{X}^2} |p_{\theta_m}(x_k, x_l, Y_{-n}^1) - p_\theta(x_k, x_l \,|\, Y_{-n}^1)p_{\theta_m}(Y_{-n}^1)|\, \mu(dx_k)\mu(dx_l). \tag{19}$$

The integral tends to zero as $m \to \infty$ as can be seen by considering the simultaneous density

$$p_{\theta_m}(x_k, x_l, Y_{-n}^1) = \int_{\mathcal{X}^n} \pi_{\theta_m}(x_{-n})g_{\theta_m}(Y_{-n}|x_{-n}) \prod_{i=-n+1}^{1} \{\alpha_{\theta_m}(x_{i-1}, x_i)g_{\theta_m}(Y_i|x_i)\} \prod_{\substack{i=-n \\ i\ne k,l}}^{1} \mu(dx_i). \tag{20}$$

Since the integrand here is continuous and can be dominated by

$$M^{n+2}\prod_{i=-n}^{1} h_{Y_i}(x_i) \in \mathbb{L}^1(\mu^n), \tag{21}$$

by assumption A1 and A5, we have from Lebesgue's dominated convergence theorem that

$$p_{\theta_m}(x_k, x_l, Y_{-n}^1) \to p_\theta(x_k, x_l, Y_{-n}^1) \quad \text{as } m \to \infty.$$

Likewise $p_{\theta_m}(Y_{-n}^1) \to p_\theta(Y_{-n}^1)$ as $m \to \infty$, and hence the integrand in (19) tends to zero. By (21) the integrand can be dominated in $\mathbb{L}^1(\mu^2)$, and therefore (19) tends to zero.

Since the expression in (18) is less than

$$\gamma^*(Y_k)\gamma^*(Y_l)\int_{\mathcal{X}^2}\{p_{\theta_m}(x_k,x_l|Y_{-n}^1)+p_\theta(x_k,x_l|Y_{-n}^1)\}\,\mu(dx_k)\mu(dx_l) \\ = 2\gamma^*(Y_k)\gamma^*(Y_l), \quad (22)$$

it is dominated in $\mathbb{L}^1(P_0)$ and hence tends to zero in $\mathbb{L}^1(P_0)$ as $m \to \infty$.

The second term can be dominated similarly and tends to zero P_0-almost surely, and therefore also in $\mathbb{L}^1(P_0)$, by the continuity of γ_θ. □

Lemma 5.4 and 5.6 show that $\{D^2 \log p_\theta(Y_1|Y_{-n}^0)\}_{n\in\mathbb{N}}$ is a uniform Cauchy sequence of continuous functions in $\mathbb{L}^1(P_0)$, which proves Lemma 10 of Bickel et al. The final Lemma states a usual property of the Fisher information. With this result, the remaining part of the proof of Theorem 3.2 is now identical to the proof of Lemma 2 in Bickel et al. (p. 1633)

LEMMA 5.7 *For any n,*

$$E_0\{D^2 \log p_0(Y_1|Y_{-n}^0)\} = -E_0\{[D\log p_0(Y_1|Y_{-n}^0)]^2\}.$$

PROOF By (3) and (12) we have

$$\begin{aligned}
&(D\log p_0(Y_1\,|\,Y_{-n}^0))^2 + D^2\log p_0(Y_1\,|\,Y_{-n}^0)\\
&= 2\Big(E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\,|\,Y_{-n}^0)^2\\
&\qquad - E_0\Big(D\log p_0(X_{-n}^1,Y_{-n}^1)\,|\,Y_{-n}^1\Big)E_0\Big(D\log p_0(X_{-n}^1,Y_{-n}^0)\,|\,Y_{-n}^0\Big)\Big) \qquad (23)\\
&+ E_0((D\log p_0(X_{-n}^1,Y_{-n}^1))^2\,|\,Y_{-n}^1) - E_0((D\log p_0(X_{-n}^1,Y_{-n}^0))^2\,|\,Y_{-n}^0)\\
&+ E_0(D^2\log p_0(X_{-n}^1,Y_{-n}^1)\,|\,Y_{-n}^1) - E_0(D^2\log p_0(X_{-n}^1,Y_{-n}^0)\,|\,Y_{-n}^0).
\end{aligned}$$

The first term has zero mean. This follows by noting from (1) that

$$D\log p_0(X_{-n}^1,Y_{-n}^1) = D\log p_0(X_{-n}^1,Y_{-n}^0) + \gamma_0(Y_1|X_1), \qquad (24)$$

thus

$$
\begin{aligned}
&E_0\Big[E_0\{D\log p_0(X_{-n}^1,Y_{-n}^1)\,|\,Y_{-n}^1\}E_0\{D\log p_0(X_{-n}^1,Y_{-n}^0)\,|\,Y_{-n}^0\}\Big]\\
&= E_0\Big\{D\log p_0(X_{-n}^1,Y_{-n}^1)E_0\{D\log p_0(X_{-n}^1,Y_{-n}^0)\,|\,Y_{-n}^0\}\Big\}\\
&= E_0\Big\{D\log p_0(X_{-n}^1,Y_{-n}^0)E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\,|\,Y_{-n}^0)\Big\}\\
&\qquad + E_0\Big\{\gamma_0(Y_1|X_1)E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\,|\,Y_{-n}^0)\Big\}\\
&= E_0\Big\{D\log p_0(X_{-n}^1,Y_{-n}^0)E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\,|\,Y_{-n}^0)\Big\}\\
&\qquad + E_0\Big\{E_0\Big[\gamma_0(Y_1|X_1)|X_1\Big]E_0\Big[E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\,|\,Y_{-n}^0)\,|\,X_1\Big]\Big\}\\
&= E_0\Big\{D\log p_0(X_{-n}^1,Y_{-n}^0)E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\,|\,Y_{-n}^0)\Big\},
\end{aligned}
$$

where the third equality follows from the conditional independence of Y_{-n}^0 and Y_1 given X_1, and the last equality from the fact that $E_0(\gamma_0(Y_1|X_1)\,|\,X_1) = 0$ by A5(ii). The mean of the first term in (23) is then

$$
\begin{aligned}
2E_0\Big\{&E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\,|\,Y_{-n}^0)\\
&\Big[E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\,|\,Y_{-n}^0) - D\log p_0(X_{-n}^1,Y_{-n}^0)\Big]\Big\} = 0.
\end{aligned}
$$

By (24) the mean of the sum of the two last terms is given by

$$
\begin{aligned}
E_0\Big\{D^2\log g_0(Y_1\,|\,X_1)\Big\} &+ E_0\Big\{(D\log g_0(Y_1\,|\,X_1))^2\Big\}\\
&+ 2E_0\Big\{\gamma_0(Y_1\,|\,X_1)D\log p_0(X_{-n}^1,Y_{-n}^0)\Big\}.
\end{aligned}
$$

The last term is zero, which is seen by conditioning on X_1 and using the argument from above. The sum of the two first terms is zero by assumption A5(ii), which completes the proof. □

References

Bickel, P. J., Ritov, Y. & Rydén, T. (1998), 'Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models', *Ann. Statist.* **26**(4), 1614–1635.

Doob, J. L. (1953), *Stochastic Processes*, John Wiley & Sons, Inc.

Durbin, J. & Koopman, S. J. (1997), 'Monte Carlo maximum likelihood estimation for non-Gaussian state space models', *Biometrika* **84**(3), 669–684.

Frühwirth-Schnatter, S. (1994), 'Applied state space modelling of non-Gaussian time series using integration-based Kalman filtering', *Stat. Comp.* **4**, 259–269.

Hoffmann-Jørgensen, J. (1994), *Probability with a View toward Statistics*, Vol. 1, Chapman & Hall.

Jensen, J. L. (1986), Nogle asymptotiske resultater. University of Aarhus (In Danish).

Kitagawa, G. & Gersch, W. (1996), *Smoothness Priors Analysis of Time Series*, Springer-Verlag, New York.

Leroux, B. G. (1992), 'Maximum-likelihood estimation for hidden Markov models', *Stoch. Proc. Appl.* **40**, 127–143.

Shephard, N. & Pitt, M. K. (1997), 'Likelihood analysis of non-Gaussian measurement time series', *Biometrika* **84**(3), 653–667.

Sweeting, T. (1980), 'Uniform asymptotic normality of the maximum likelihood estimator', *Ann. Statist.* **8**(6), 1375–1381.

West, M. & Harrison, J. (1989), *Bayesian Forecasting and Dynamic Models*, Springer-Verlag, New York.

DEPT. OF THEORETICAL STATISTICS
UNIVERSITY OF AARHUS
NY MUNKEGADE
DK-8000 AARHUS C, DENMARK

E-MAIL: jlj@imf.au.dk
vaever@imf.au.dk