Asymptotic Analysis of Model Selection Criteria for General Hidden Markov Models

Shouto Yonekura[1,3], Alexandros Beskos[1,3], Sumeetpal S.Singh[2,3]

March 31, 2020

1 Department of Statistical Science, University College London, UK.
2 Department of Engineering, University of Cambridge, UK.
3 The Alan Turing Institute for Data Science, UK.

Abstract

The paper obtains analytical results for the asymptotic properties of Model Selection Criteria – widely used in practice – for a general family of hidden Markov models (HMMs), thereby substantially extending the related theory beyond typical 'i.i.d.-like' model structures and filling in an important gap in the relevant literature. In particular, we look at the Bayesian and Akaike Information Criteria (BIC and AIC) and the model evidence. In the setting of nested classes of models, we prove that BIC and the evidence are strongly consistent for HMMs (under regularity conditions), whereas AIC is not weakly consistent. Numerical experiments support our theoretical results.

Keywords Hidden Markov models, Akaike information criteria, Bayesian information criteria, Model evidence.

1   Introduction

Owning to their rich structure, hidden Markov models (HMMs) are routinely used in such diverse disciplines as finance (Mamon and Elliott, 2007), speech recognition (Gales and Young, 2008), epidemiology (Green and Richardson, 2002), biology (Yoon, 2009), signal processing (Crouse et al., 1998). We refer to Del Moral (2004, 2013) for a comprehensive study of the theory of HMMs and of accompanying Monte Carlo methods for their calibration to observations. Model Selection has been one of the most well studied topics in Statistics. BIC (Schwarz, 1978) or AIC (Akaike, 1974) – as well as their generalisations (Konishi and Kitagawa, 1996) – and the evidence, are used in a wide range of contexts, including time series analysis (Shibata, 1976), regression (Hurvich and Tsai, 1989), bias correction (Hurvich and Tsai, 1990), composite likelihoods (Varin

and Vidoni, 2005). For a comprehensive treatment of the subject of Model Selection, see e.g. Claeskens and Hjort (2008).

There has been relatively limited research on Model Selection for general classes of HMMs used in practice. A fundamental property of a Model Selection Criterion is that of *consistency* (to be defined analytically later on in the paper). Csiszár and Shields (2000) prove strong consistency of BIC for discrete-time, finite-alphabet Markov chains. In the HMM context, Gassiat and Boucheron (2003) consider discrete-time, finite-alphabet HMMs and provide asymptotic and finite-sample analysis of code-based and penalised maximum likelihood estimators (MLEs) using tools from Information Theory and Stein's Lemma. With regards to the Bayesian approach to Model Selection, this typically involves the marginal likelihood of the data (or evidence) (Jeffreys, 1998; Kass and Raftery, 1995). Shao et al. (2017) show numerically that the evidence can be consistent for HMMs, however this has yet to be proven analytically.

The work in this paper makes a number of contributions, relevant for HMMs on general state spaces – thus of wide practical significance and such that cover an important gap in the theory of HMMs established in the existing literature. We remark that our analysis assumes smoothness conditions of involved functions w.r.t. the parameter of interest, thus is intrinsically not relevant for interesting problems of discrete nature, an example being the identification of the number of states of the underlying Markov chain. Our main results can be summarised as follows:

(i) We establish sharp asymptotic results (in the sense of obtaining $\limsup_n$ for the quantity of interest) for the log-likelihood function for HMMs evaluated at the MLE, in an a.s. sense. A lot of the initial developments are borrowed from Douc et al. (2014) (see also citations therein for more works on asymptotic properties of the MLE for HMMs). Moving from the study of the MLE to that of Model Selection Criteria is non-trivial, involving for instance use of the Law of Iterated Logarithm (LIL) for – carefully developed – martingales (Stout, 1970).

(ii) We show that BIC and the evidence are strongly consistent in the context of nested HMMs, whereas AIC is not consistent. To the best of our knowledge, this is the first time that such statements are proven in the literature for general HMMs. For AIC, we show that, w.p. 1, this criterion will occasionally choose the wrong model even under an infinite amount of information.

The rest of the paper is organised as follows. In Section 2, we briefly review some asymptotic results for the log-likelihood function and the MLE without assuming model correctness. An important departure from the i.i.d. setting is that the log-likelihood function itself does not make up a stationary time-series process even if the data are assumed to be derived from one. Section 3 begins with some asymptotic results for the MLE and the log-likelihood under model correctness. Later on, we move beyond the established literature and, by calling upon LIL for martingales, we establish a number of fundamental

asymptotic results, relevant for Model Selection Criteria. In Section 4, we study the derivation of BIC (and its connection with the evidence) and AIC for general HMMs. In particular, an explicit result binding BIC and evidence will later on be used to show that the two criteria share similar consistency properties. Section 5 contains our main results. We prove strong consistency of BIC and the evidence and non-consistency of AIC for a class of nested HMMs. Section 6 reviews (for completeness) an algorithm borrowed from the literature, based on Sequential Monte Carlo, for approximating AIC and BIC. We use this algorithm in Section 7 to present some numerical results that agree with our theory. We conclude in Section 8.

2 Asymptotics under No-Model-Correctness

We briefly summarise some asymptotic results for general HMMs needed in later sections. The development follows closely (Douc et al., 2014, Ch. 13). An HMM is a bivariate process $\{x_k, z_k\}_{k \geq 0}$ such that state component $\{x_k\}_{k \geq 0}$ is an unobservable Markov chain with initial law $x_0 \sim \eta$ and transition kernel $Q_\theta(\cdot|x)$, with values in the measurable space $(\mathsf{X}, \mathcal{X})$. We have adopted a parametric setting with $\theta \in \Theta \subseteq \mathbb{R}^d$, for some $d \geq 1$. Conditionally on $\{x_k\}_{k \geq 0}$, the distribution of the observation process instance z_k depends only on $x_k = x$, independently over $k \geq 0$, and is given by the kernel $G_\theta(\cdot|x)$ defined on $(\mathsf{Y}, \mathcal{Y})$. We assume that X, Y are Polish spaces and $\mathcal{X}$, $\mathcal{Y}$ the corresponding Borel σ-algebras. The notation $\{y_k\}_{k \geq 0}$ is reserved for the true *data generating process*, which may or may not belong in the parametric family of HMMs we specified above – meant to be distinguished from $\{z_k\}_{k \geq 0}$ which is the process driven by the model dynamics. In particular, in this section we work under no-model correctness, i.e. we do not have to assume the existence of a correct parameter value for the prescribed model that delivers the distribution of the data generating process.

Throughout the article, we assume that the following hold.

Assumption 1. *The data generating process* $\{y_k\}_{k \geq 0}$ *is strongly stationary and ergodic.*

Assumption 2.   *(i) $Q_\theta(\cdot|x)$ is absolutely continuous w.r.t. a probability measure μ on $(\mathsf{X}, \mathcal{X})$ with density $q_\theta(x'|x)$ – μ is fixed for all $(x, \theta) \in \mathsf{X} \times \Theta$.*

  (ii) $G_\theta(\cdot|x)$ is absolutely continuous w.r.t. a measure ν on $(\mathsf{Y}, \mathcal{Y})$ with density $g_\theta(y|x)$ – ν is fixed for all $(x, \theta) \in \mathsf{X} \times \Theta$.

  (iii) The initial distribution $\eta = \eta(dx_0)$ has density, denoted $\eta(x_0)$ – with some abuse of notation –, w.r.t. μ.

  (iv) The parameter space Θ is a compact subset of $\mathbb{R}^d$; w.p. 1, $p_\theta(y_{0:n-1}) > 0$ for all $\theta \in \Theta$, for all $n \geq 1$, where $p_\theta(\cdot)$ denotes here the density of the distribution of the observations under the model (for given θ and size n).

Without loss of generality, we have assumed that $\eta(dx_0)$ does not depend on θ. Probability statements – as in Assumption 2(iv) – and expectations throughout

the paper are to be understood w.r.t. the law of the data generating process $\{y_k\}$. Henceforth we make use of the notation $a_{i:j} = (a_i, \ldots, a_j)$, for integers $i \leq j$, for a given sequence $\{a_k\}$. We need the following conditions.

Assumption 3. *There exist σ^-, $\sigma^+ \in (0, \infty)$ so that*

$$\sigma^- \leq q_\theta(x'|x) \leq \sigma^+$$

for any x, $x' \in \mathsf{X}$ and any $\theta \in \Theta$.

This is the strong mixing condition typically used in this context (e.g. Del Moral (2004, 2013)), providing a Dobrushin coefficient of $1 - \sigma_-/\sigma_+$ for the hidden Markov chain; it is critical for most of the results reviewed or developed in the sequel. Assumption 3 implies, for instance, that for any $x \in \mathsf{X}$, $A \in \mathcal{X}$, $Q_\theta(A|x) \geq \sigma^- \mu(A)$, that is, for any $\theta \in \Theta$, X is a 1-small set for the process $\{x_k\}_{k \geq 0}$. The chain has the unique invariant measure π_θ^X and is uniformly ergodic, so for any $x \in \mathsf{X}$, $n \geq 0$, $\left\| Q_\theta^n(\cdot|x) - \pi_\theta^X \right\|_{TV} \leq (1 - \sigma^-/\sigma^+)^n$ – with $\|\cdot\|_{TV}$ denoting total variation norm.

We calculate the likelihood and log-likelihood functions,

$$p_\theta(y_{0:n-1}) = \int \eta(dx_0) g_\theta(y_0|x_0) \prod_{k=1}^{n-1} \left\{ q_\theta(x_k|x_{k-1}) g_\theta(y_k|x_k) \right\} \mu^{\otimes n}(dx_{1:n-1}); \quad (1)$$

$$\ell_\theta(y_{0:n-1}) = \log p_\theta(y_{0:n-1}) = \sum_{k=0}^{n-1} \log p_\theta(y_k|y_{0:k-1}). \quad (2)$$

Though $\{y_k\}_{k \geq 0}$ is stationary and ergodic, terms $\{\log p_\theta(y_k|y_{0:k-1})\}_{k \geq 0}$ do not form a stationary process (in general). To obtain stationary and ergodic log-likelihood terms, following Douc et al. (2004); Cappé et al. (2005); Douc et al. (2014), we work with the standard extension of the stationary y-process onto the whole of the integers, and write $\{y_k\}_{k=-\infty}^\infty$. One can then define the variable $\log p_\theta(y_k|y_{-\infty:k-1})$ as the a.s. limit of the Cauchy sequence (uniformly in θ) $\log p_\theta(y_k|y_{-t:k-1})$ – found as in (1) for initial law $x_{-t} \sim \eta$ – as $t \to \infty$; see (Douc et al., 2014, Ch. 13) for more details. We can now define the modified, stationary version of the log-likelihood

$$\ell_\theta^s(y_{-\infty:n-1}) := \sum_{k=0}^{n-1} \log p_\theta(y_k|y_{-\infty:k-1}). \quad (3)$$

Assumption 4. *We have that $b^+ := \sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} g_\theta(y|x) < \infty$ and*

$$\mathbb{E}\,|\log b^-(y_0)| < \infty,$$

where $b^-(y) := \inf_{\theta \in \Theta} \int_\mathsf{X} g_\theta(y|x) \mu(dx) > 0$.

The finite-moment part implies that $\mathbb{E}\,|\log p_\theta(y_0|y_{-\infty:-1})| < \infty$, thus Birkhoff's ergodic theorem can be applied for averages deduced from (3).

Proposition 1. *Under Assumptions 1-4,*

$$\sup_{\theta \in \Theta} |\tfrac{1}{n}\ell_\theta(y_{0:n-1}) - \tfrac{1}{n}\ell_\theta^s(y_{-\infty:n-1})| \le \tfrac{C}{n},$$

for a constant $C > 0$.

Proof. This is Proposition 13.5 of Douc et al. (2014); the upper bound C/n is implied from the proof of that proposition. □

We consider the maximum likelihood estimator (MLE) defined as the set

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \ell_\theta(y_{0:n-1}). \tag{4}$$

Assumption 5. *For all $(x, x') \in \mathsf{X} \times \mathsf{X}$ and $y \in \mathsf{Y}$, the mappings $\theta \mapsto q_\theta(x'|x)$ and $\theta \mapsto g_\theta(y|x)$ are continuous.*

Such requirements imply continuity of the log-likelihood $\theta \mapsto \tfrac{1}{n}\ell_\theta(y_{0:n-1})$ and its limit $\theta \mapsto \mathbb{E}\left[\log p_\theta(y_0|y_{-\infty:-1})\right]$, which – together with other conditions – provide convergence of the MLE to the maximiser of the limiting function. For sets $A, B \subseteq \Theta$, we define $d(A, B) := \inf_{a \in A, b \in B} |a - b|$.

Proposition 2. *Under Assumptions 1-5 we have the following.*

(i) *Let $\ell(\theta) := \mathbb{E}\left[\log p_\theta(y_0|y_{-\infty:-1})\right]$. The function $\theta \mapsto \ell(\theta)$ is continuous, and we have*

$$\lim_{n \to \infty} \sup_{\theta \in \Theta} |\tfrac{1}{n}\ell_\theta(y_{0:n-1}) - \ell(\theta)| = 0, \quad w.p.\ 1.$$

(ii) *We have $\lim_{n \to \infty} d(\hat{\theta}_n, \theta_*) = 0$, w.p. 1, where*

$$\theta_* := \arg\max_{\theta \in \Theta} \ell(\theta)$$

is the set of global maxima of $\ell(\theta)$.

Proof. This is Theorem 13.7 of Douc et al. (2014). The proof of (i) is based on working with the stationary version of the log-likelihood in (3), permitted due to Proposition 2, and using Birkhoff's ergodic theorem. □

Recall that θ_* need not be thought of as the correct parameter value here, as no assumption of the class of HMMs containing the correct data-generating model is made in this section. To avoid identifiability issues, we make the following assumption on the HMM model.

Assumption 6. *θ_* is a singleton.*

This implies immediately the following.

Corollary 1. *The set of maxima $\hat{\theta}_n$ is a singleton for all large enough n, w.p. 1, and $\lim_{n \to \infty} \hat{\theta}_n = \theta_*$, w.p. 1.*

3 Asymptotics under Model-Correctness

To examine the asymptotic behaviour of Information Criteria like AIC or BIC one has to investigate the behaviour of the log-likelihood evaluated at the MLE, $\ell_{\hat{\theta}_n}(y_{0:n-1})$, for increasing data size n. Following closely Douc et al. (2014), we first pose the following assumption, with $\theta_* \in \Theta$ as determined in Proposition 2 and Assumption 6. Here and in the sequel, all gradients and Hessians – represented by ∇ and $\nabla\nabla^\top$ respectively, adopting an 'applied mathematics' notation – are w.r.t. the model parameter(s) θ.

Assumption 7. *θ_* is in the interior of Θ, and there exists $\epsilon > 0$ and an open neighbourhood $\mathcal{B}_\epsilon(\theta_*) := \{\theta \in \Theta : |\theta - \theta_*| < \epsilon\}$ of θ_* such that the following hold.*

(i) For any $(x, x') \in \mathsf{X} \times \mathsf{X}$ and $y \in \mathsf{Y}$, $\theta \mapsto q_\theta(x'|x)$ and $\theta \mapsto g_\theta(y|x)$ are twice continuously differentiable on $\mathcal{B}_\epsilon(\theta_*)$.

(ii) $\sup_{\theta \in \mathcal{B}_\epsilon(\theta_*)} \sup_{x,x' \in \mathsf{X}^2} \left\{ \left|\nabla \log q_\theta(x'|x)\right| + \left|\nabla\nabla^\top \log q_\theta(x'|x)\right| \right\} < \infty$.

(iii) For some $\delta > 0$,

$$\mathbb{E}\left[\sup_{\theta \in \mathcal{B}_\epsilon(\theta_*)} \sup_{x \in \mathsf{X}} \left\{ \left|\nabla \log g_\theta(y_0|x)\right|^{2+\delta} + \left|\nabla\nabla^\top \log g_\theta(y_0|x)\right| \right\} \right] < \infty.$$

$|\cdot|$ denotes the Euclidean norm for vector input or one of the standard equivalent matrix norms for matrix input. Assumption 7 implies that, for any fixed n the log-likelihood function is twice continuously differentiable in $\mathcal{B}_\epsilon(\theta_*)$ (standard use of bounded convergence theorem from Assumption 7(i)). Also, the score function has finite $(2 + \delta)$–moment and the Hessian finite first moment, for any $\theta \in \mathcal{B}_\epsilon(\theta_*)$; the proof of these statements requires use of Fisher's identity (used later on) together with parts (ii), (iii) of Assumption 7 involving the gradient for the score function, and Louis' identity (see e.g. Poyiadjis et al. (2011) for background on Fisher's, Louis' identities) for the Hessian together with the stated conditions for the matrices of second derivatives. We avoid further details.

We start off with a standard Taylor expansion,

$$\ell_{\hat{\theta}_n}(y_{0:n-1}) = \ell_{\theta_*}(y_{0:n-1}) + \frac{\nabla^\top \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}} \sqrt{n}(\hat{\theta}_n - \theta_*)$$
$$+ \tfrac{1}{2}\sqrt{n}(\hat{\theta}_n - \theta_*)^\top \left[\int_0^1 \frac{\nabla\nabla^\top \ell_{s\hat{\theta}_n + (1-s)\theta_*}(y_{0:n-1})}{n} ds \right] \sqrt{n}(\hat{\theta}_n - \theta_*), \quad (5)$$

together with a corresponding one for the score function,

$$0 \equiv \frac{\nabla \ell_{\hat{\theta}_n}(y_{0:n-1})}{\sqrt{n}}$$
$$= \frac{\nabla \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}} + \left[\int_0^1 \frac{\nabla\nabla^\top \ell_{s\hat{\theta}_n + (1-s)\theta_*}(y_{0:n-1}) ds}{n} \right] \sqrt{n}(\hat{\theta}_n - \theta_*). \quad (6)$$

We will look at the asymptotic properties of the score function terms and the integral involving the Hessian, i.e. of,

$$\nabla \ell_{\theta_*}(y_{0:n-1})/\sqrt{n}, \quad \int_0^1 \frac{\nabla \nabla^\top \ell_{s\hat{\theta}_n + (1-s)\theta_*}(y_{0:n-1})ds}{n}, \tag{7}$$

starting from the former.

We will sometimes work under the assumption of model-correctness and we shall be clear when that is the case.

Assumption 8. *The dynamics of the data generating process $\{y_k\}_{k\geq 0}$ correspond to those of the HMM with initial distribution $x_0 \sim \eta(\cdot) \equiv \pi_{\theta_*}^X(\cdot)$, transition kernel $Q_{\theta_*}(\cdot|x)$ and observation kernel $G_{\theta_*}(\cdot|x)$.*

For results that do not refer to Assumption 8, θ_* still makes sense as per its definition in Proposition 2. Using Jensen's inequality, and for θ_* corresponding to the true parameter, one can easily check that $\ell(\theta) \leq \ell(\theta_*)$, so indeed the true parameter coincides with θ_* given in Proposition 2.

Following (Douc et al., 2014, Ch. 13) we obtain:

1. Re-write the score function evaluated at $\theta = \theta_*$ as

$$\nabla \ell_{\theta_*}(y_{0:n-1}) = \sum_{i=0}^{n-1} \left[\nabla \ell_{\theta_*}(y_{0:i}) - \nabla \ell_{\theta_*}(y_{0:i-1}) \right], \tag{8}$$

   under the convention that $\nabla \ell_{\theta_*}(y_{0:-1}) \equiv 0$. The above differences will be shown to converge – for increasing data size n, in an appropriate sense – to stationary (and ergodic) martingale increments.

2. Using Fisher's identity, one has, for $y_{0:k} \in \mathsf{Y}^{k+1}$, $k \geq 0$,

$$\nabla \ell_{\theta_*}(y_{0:k}) = \int_{\mathsf{X}^{k+1}} \nabla \log p_{\theta_*}(x_{0:k}, y_{0:k}) \, p_{\theta_*}(x_{0:k}|y_{0:k}) \mu^{\otimes(k+1)}(dx_{0:k})$$

$$= \sum_{j=0}^k \int_{\mathsf{X}^2} d_{\theta_*}(x_{j-1}, x_j, y_j) p_{\theta_*}(x_{j-1:j}|y_{0:k}) \mu^{\otimes 2}(dx_{j-1:j}),$$

   where we have defined

$$d_{\theta_*}(x_{j-1}, x_j, y_j) := \nabla \log \left[q_{\theta_*}(x_j|x_{j-1}) \, g_{\theta_*}(y_j|x_j) \right], \quad j \geq 0,$$

   with the conventions

$$d_{\theta_*}(x_{-1}, x_0, y_0) \equiv d_{\theta_*}(x_0, y_0) \equiv \nabla \log \left[\eta(x_0) \, g_{\theta_*}(y_0|x_0) \right]$$

   and the one

$$\int_{\mathsf{X}^2} d_{\theta_*}(x_{-1}, x_0, y_0) p_{\theta_*}(x_{-1:0}|y_{0:k}) \mu^{\otimes 2}(dx_{-1:0})$$

$$\equiv \int_{\mathsf{X}} d_{\theta_*}(x_0, y_0) p_{\theta_*}(x_0|y_{0:k}) \mu(dx_0).$$

7

Thus, we have, for $i \geq 0$,

$$h_{\theta_*}(y_{0:i}) := \nabla\ell_{\theta_*}(y_{0:i}) - \nabla\ell_{\theta_*}(y_{0:i-1}) \tag{9}$$

$$= \int_{\mathsf{X}^2} d_{\theta_*}(x_{i-1}, x_i, y_i) p_{\theta_*}(x_{i-1:i}|y_{0:i})\mu^{\otimes 2}(dx_{i-1:i})$$

$$+ \sum_{j=0}^{i-1}\Big[\int_{\mathsf{X}^2} d_{\theta_*}(x_{j-1}, x_j, y_j) p_{\theta_*}(x_{j-1:j}|y_{0:i})\mu^{\otimes 2}(dx_{j-1:j})$$

$$- \int_{\mathsf{X}^2} d_{\theta_*}(x_{j-1}, x_j, y_j) p_{\theta_*}(x_{j-1:j}|y_{0:i-1})\mu^{\otimes 2}(dx_{j-1:j})\Big].$$

3. To obtain stationary increments for increasing n, Douc et al. (2014) work with (for $i \geq 0$)

$$h_{\theta_*}(y_{-\infty:i}) := \int_{\mathsf{X}^2} d_{\theta_*}(x_{i-1}, x_i, y_i) p_{\theta_*}(x_{i-1:i}|y_{-\infty:i})\mu^{\otimes 2}(dx_{i-1:i})$$

$$+ \sum_{j=-\infty}^{i-1}\Big[\int_{\mathsf{X}^2} d_{\theta_*}(x_{j-1}, x_j, y_j) p_{\theta_*}(x_{j-1:j}|y_{-\infty:i})\mu^{\otimes 2}(dx_{j-1:j})$$

$$- \int_{\mathsf{X}^2} d_{\theta_*}(x_{j-1}, x_j, y_j) p_{\theta_*}(x_{j-1:j}|y_{-\infty:i-1})\mu^{\otimes 2}(dx_{j-1:j})\Big].$$

Following (Douc et al., 2014, Proposition 13.20), integrals involving infinitely long data sequences of the form

$$\int_{\mathsf{X}^2} d_{\theta_*}(x_{j-1}, x_j, y_j) p_{\theta_*}(x_{j-1:j}|y_{-\infty:i})\mu^{\otimes 2}(dx_{j-1:j}), \quad j \leq i, \quad i \geq 0,$$

appearing above are defined as a.s. or L_2-limits of the random variables $\int_{\mathsf{X}^2} d_{\theta_*}(x_{j-1}, x_j, y_j) p_{\theta_*}(x_{j-1:j}|y_{-m:i})\mu^{\otimes 2}(dx_{j-1:j})$, with $m \to \infty$, given Assumptions 1-7. A small modification of the derivations in (Douc et al., 2014, Ch.13) (they look at second moments) gives that, under Assumptions 1-7, and for constant $\delta > 0$ as defined in Assumption 7(iii), for $i \geq 0$,

$$\big\|h_{\theta_*}(y_{0:i}) - h_{\theta_*}(y_{-\infty:i})\big\|_{2+\delta}$$

$$\leq 12\,\mathbb{E}^{1/(2+\delta)}\Big[\sup_{x,x'\in\mathsf{X}}|d_{\theta_*}(x, x', y_0)|^{2+\delta}\Big]\frac{\rho^{i/2-1}}{1-\rho}, \tag{10}$$

where $\rho = 1 - \sigma^-/\sigma^+$. (The expectation in the upper bound is finite due to Assumption 7(ii),(iii).) Here and below, $\|\cdot\|_a$, $a \geq 1$, denotes the L_a–norm of the variable under consideration. From triangular inequality we have,

$$\Big\|n^{-1/2}\sum_{i=0}^{n-1}\big\{h_{\theta_*}(y_{0:i}) - h_{\theta_*}(y_{-\infty:i})\big\}\Big\|_{2+\delta}$$

$$\leq n^{-1/2}\sum_{i=0}^{n-1}\|h_{\theta_*}(y_{0:i}) - h_{\theta_*}(y_{-\infty:i})\|_{2+\delta}.$$

Thus, recalling equation (8) and definition (9), the bound (10) implies

$$\frac{\nabla \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}} = n^{-1/2} \sum_{i=0}^{n-1} h_{\theta_*}(y_{-\infty:i}) + \mathcal{O}_{L_{2+\delta}}(n^{-1/2}). \tag{11}$$

For $a \geq 1$ and a sequence of positive reals $\{b_k\}$, $\mathcal{O}_{L_a}(b_n)$ denotes a sequence of random variables with L_a-norm being $\mathcal{O}(b_n)$.

4. At this point we are required to make explicit use of the model-correctness Assumption 8. We have

$$\mathbb{E}\left[h_{\theta_*}(y_{-\infty:i})|y_{-\infty:i-1}\right] = \mathbb{E}\left[\mathbb{E}\left[d_{\theta_*}(x_{i-1}, x_i, y_i)|y_{-\infty:i}\right]\Big| y_{-\infty:i-1}\right]$$
$$+ \sum_{j=-\infty}^{i-1} \mathbb{E}\left[\left\{\mathbb{E}\left[d_{\theta_*}(x_{j-1}, x_j, y_j)|y_{-\infty:i}\right]\right.\right.$$
$$\left.\left. - \mathbb{E}\left[d_{\theta_*}(x_{j-1}, x_j, y_j)|y_{-\infty:i-1}\right]\right\}\Big| y_{-\infty:i-1}\right]$$

Each term in the sum is trivially 0. For the first term, we have,

$$\mathbb{E}\left[d_{\theta_*}(x_{i-1}, x_i, y_i)|y_{-\infty:i-1}\right] = \mathbb{E}\left[\mathbb{E}\left[d_{\theta_*}(x_{i-1}, x_i, y_i)\,|\,x_{i-1}, y_{-\infty:i-1}\right]\right]$$
$$\equiv 0.$$

Notice that we have indeed used the model correctness assumption to obtain the latter result. So, terms $h_{\theta_*}(y_{-\infty:i})$ make up a strongly stationary, ergodic (they inherit the properties of the data generating process) martingale increment sequence – of finite second moment – under the filtration generated by the data. Using a CLT (Hall and Heyde, 1980) and the LIL of Stout (1970) for such sequences allows for control over the martingales

$$M_{n,j} := \sum_{i=0}^{n-1} (h_{\theta_*}(y_{-\infty:i}))_j \ , \quad 1 \leq j \leq d.$$

Subscript j indicates the j-th component of the d-dimensional vectors. In particular, we have the CLT ('$\Rightarrow$' denotes weak convergence, and $N_d(a, B)$ the d-dimensional Gaussian law with mean a and covariance matrix B)

$$\frac{M_n}{\sqrt{n}} \Rightarrow N_d(0, \mathcal{J}_{\theta_*}), \tag{12}$$

where we have defined,

$$\mathcal{J}_{\theta_*} = \mathbb{E}\left[h_{\theta_*}(y_{-\infty:0})h_{\theta_*}(y_{-\infty:0})^\top\right]. \tag{13}$$

Also, we have the LIL (Stout, 1970),

$$\limsup_n \frac{|M_{n,j}|}{\sqrt{2n\log\log n}} = \mathbb{E}^{1/2}\left[(h_{\theta_*}(y_{-\infty:0}))_j^2\right], \quad 1 \leq j \leq d, \quad w.p.\,1. \tag{14}$$

9

We now turn to the second term in (7).

Proposition 3. *Under Assumptions 1-7, we have that, w.p. 1,*

$$\lim_{\delta \to 0} \lim_{n \to \infty} \sup_{\theta \in \mathcal{B}_\delta(\theta_*)} \left| (-\nabla \nabla^\top \ell_\theta(y_{0:n-1})/n) - \mathcal{J}_{\theta_*} \right| = 0.$$

Proof. This is Theorem 13.24 of Douc et al. (2014). $\square$

Proposition 4. *Under Assumptions 1-7 we have that, w.p. 1,*

$$J_{\theta_*}(y_{0:n-1}) := -\int_0^1 \frac{\nabla \nabla^\top \ell_{s\hat{\theta}_n + (1-s)\theta_*}(y_{0:n-1})ds}{n} \longrightarrow \mathcal{J}_{\theta_*}.$$

Proof. This is implied immediately from Proposition 3 (recall that $\hat{\theta}_n \to \theta_*$). $\square$

Notice that this result does not require the assumption of model correctness. We do make the following assumption on the HMM model under consideration.

Assumption 9. *The matrix $\mathcal{J}_{\theta_*} \in \mathbb{R}^{d \times d}$ is non-singular.*

We summarise the results in this part with a proposition and theorem.

Proposition 5. *(i) Under Assumptions 1-7, 9 we have, for all large enough n,*

$$\ell_{\hat{\theta}_n}(y_{0:n-1}) = \ell_{\theta_*}(y_{0:n-1}) + \frac{1}{2} \frac{\nabla^\top \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}} J_{\theta_*}^{-1}(y_{0:n-1}) \frac{\nabla \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}},$$

where $J_{\theta_*}(y_{0:n-1}) \to \mathcal{J}_{\theta_*}$, w.p. 1, for the non-singular matrix $\mathcal{J}_{\theta_*}$ defined in (13).

(ii) Under Assumptions 1-7, 9 we have, for all large enough n,

$$\frac{\nabla \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}} = n^{-1/2} \sum_{i=0}^{n-1} h_{\theta_*}(y_{-\infty:i}) + n^{-1/2} R_n$$

where $\|h_{\theta_*}(y_{-\infty:i})\|_{2+\delta} + \|R_n\|_{2+\delta} \leq C$, for $\delta > 0$ as determined in Assumption 7(iii) and a constant $C > 0$.

(iii) Under Assumptions 1-9, w.p. 1, as $n \to \infty$, $n^{-1/2} R_n \to 0$, and we have the CLT

$$\frac{\nabla \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}} \Rightarrow N_d(0, \mathcal{J}_{\theta_*}).$$

Proof. The equation in part (i) is a combination of equations (5), (6), assuming that n is big enough to permit inversion of the involved matrix. (ii) is simply a rewriting of earlier calculations. The CLT in (iii) is trivial. $\square$

Theorem 1. *Under Assumptions 1-9, we have the LIL,*

$$\limsup_n \frac{|(\nabla \ell_{\theta_*}(y_{0:n-1}))_j|}{\sqrt{2n \log \log n}} = \mathbb{E}^{1/2} \left[(h_{\theta_*}(y_{-\infty:0}))_j^2 \right], \quad 1 \leq j \leq d.$$

Proof. From Proposition 5(ii), using Markov inequality, we have for $1 \leq j \leq d$ and any $\epsilon > 0$,

$$\mathbb{P}\left[\,|R_{n,j}| \geq \epsilon\,\sqrt{n}\,\right] = \mathbb{P}\left[\,|R_{n,j}|^{2+\delta} \geq \epsilon^{2+\delta}\,n^{1+\delta/2}\,\right] \leq \frac{\mathbb{E}\,|R_{n,j}|^{2+\delta}}{\epsilon^{2+\delta}\,n^{1+\delta/2}}.$$

Thus,

$$\sum_{n=0}^{\infty} \mathbb{P}\left[\,|R_{n,j}| \geq \epsilon\,\sqrt{n}\,\right] < \infty,$$

and the Borel-Cantelli lemma gives that

$$\mathbb{P}\left[\,|R_{n,j}| \geq \epsilon\,\sqrt{n},\ \text{infinitely often in } n\,\right] = 0.$$

Equivalently, w.p. 1, $|R_{n,j}|/\sqrt{n} \to 0$. The proof is completed via the martingale LIL in (14). $\qquad\square$

4 Model Selection Criteria for HMMs

We provide a brief illustration for the derivation of AIC and BIC, with focus on HMMs. Results obtained that explicitly connect BIC and the evidence will allow for deriving consistency properties for the evidence directly after studying the BIC criterion later in the paper.

4.1 BIC and Evidence for HMMs

We consider the derivation of BIC for a general HMM. BIC is used by Schwarz (1978) and can be obtained by applying a Laplace approximation at the calculation of the marginal likelihood (or evidence) of the model under consideration. Consideration of the sequence of log-likelihood functions over the data size n (see e.g. Kass et al. (1990) for the concept of 'Laplace-regular' models) provide analytical, sufficient conditions for controlling the difference between the evidence and BIC. We briefly review the Taylor expansions underlying the derivation of BIC and provide the regularity conditions that control its difference from the evidence in the context of HMMs. Compared with Kass et al. (1990), weaker conditions are required here, as BIC derives from an $\mathcal{O}(n^{-1})$ approximation, in an a.s. sense, of the evidence (rather than $\mathcal{O}(n^{-2})$ expansions looked at in the Laplace-regular framework).

Let $\pi(\theta)d\theta$ be a prior for parameter θ – for simplicity we assume that $d\theta$ is the Lebesgue measure on $\mathbb{R}^d$. The evidence is given by

$$p(y_{0:n-1}) = \int_{\Theta} \pi(\theta) \exp\left\{\ell_\theta(y_{0:n-1})\right\}d\theta. \tag{15}$$

We define

$$\mathsf{J}(y_{0:n-1}) := -\frac{\nabla\nabla^{\top}\ell_{\hat{\theta}_n}(y_{0:n-1})}{n}.$$

We will be explicit on regularity conditions in the statement of the proposition that follows. Following similar steps as in Kass et al. (1990), we apply a fourth-order Taylor expansion around the MLE $\hat{\theta}_n$ that gives – for $u := \sqrt{n}(\theta - \hat{\theta}_n)$,

$$\ell_\theta(y_{0:n-1}) = \ell_{\hat{\theta}_n}(y_{0:n-1}) - \tfrac{1}{2}\, u^\top \mathsf{J}(y_{0:n-1})\, u$$
$$+ \tfrac{1}{6}\, n^{-1/2} \sum_{i,j,k=1}^{d} u_i u_j u_k\, \frac{\partial_{\theta_i}\partial_{\theta_j}\partial_{\theta_k}\ell_{\hat{\theta}_n}(y_{0:n-1})}{n} + R_{1,n}, \quad (16)$$

for residual term $R_{1,n}$ (in the integral form expansion) involving fourth-order derivatives of $\theta \mapsto \ell_\theta(y_{0:n-1})/n$ evaluated at

$$\xi = a\hat{\theta}_n + (1-a)\theta,$$

for some $a \in [0,1]$, fourth order polynomials of u, and a factor of n^{-1}, see e.g. Ch.14 of Lang (2012) for details on such expansions. Notice we have used $\nabla\ell_{\hat{\theta}_n}(y_{0:n-1}) = 0$. For the prior density we have

$$\pi(\theta) = \pi(\hat{\theta}_n) + n^{-1/2}\, \nabla^\top\pi(\hat{\theta}_n)\, u + R_{2,n},$$

for the integral residual term $R_{2,n}$ with second-order derivatives of $\pi(\theta)$, second-order polynomial of u and a factor of n^{-1}. Using a second order expansion for $x \mapsto e^x$, only for the terms beyond the quadratic in u in (16), we get

$$\frac{p(y_{0:n-1})}{p_{\hat{\theta}_n}(y_{0:n-1})} = \int_\Theta e^{-\frac{1}{2}\, u^\top \mathsf{J}(y_{0:n-1})\, u} \times$$
$$\{\pi(\hat{\theta}_n) + n^{-1/2}\, m(u, y_{0:n-1}) + R_n\}\, d\theta, \quad (17)$$

where we have separated the term (later on removed as having zero mean under a Gaussian integrator)

$$m(u, y_{0:n-1}) = \tfrac{1}{6}\, n^{-1/2} \sum_{i,j,k=1}^{d} u_i u_j u_k\, \frac{\partial_{\theta_i}\partial_{\theta_j}\partial_{\theta_k}\ell_{\hat{\theta}_n}(y_{0:n-1})}{n} + \nabla^\top\pi(\hat{\theta}_n)\, u;$$

the residual term R_n can be deduced from the calculations.

Remark 1. *The Laplace-regular setting of Kass et al. (1990) provides concrete conditions for the above derivations to be valid and for controlling the deduced residual terms. Apart from the standard assumptions on the existence of derivatives and a bound on the fourth order derivatives of $\ell_\theta(y_{0:n-1})$ close to θ_* – the latter being defined in Proposition 2 as the limit of $\hat{\theta}_n$ – the following are also required:*

(i) For any $\delta > 0$, w.p. 1,

$$\limsup_n \sup_{\theta \in \Theta - \mathcal{B}_\delta(\theta_*)} \left\{\tfrac{1}{n}\left(\ell_\theta(y_{0:n-1}) - \ell_{\theta_*}(y_{0:n-1})\right)\right\} < 0;$$

12

(ii) For some $\epsilon > 0$, $\mathcal{B}_\epsilon(\theta^*) \subseteq \Theta$, and w.p. 1,

$$\limsup_n \sup_{\theta \in \mathcal{B}_\epsilon(\theta_*)} \left\{ \det\left(-\nabla\nabla^\top \ell_\theta(y_{0:n-1})/n \right) \right\} > 0.$$

Note that (i) is implied by Proposition 2 and identifiability Assumption 6. Also, Proposition 3 and Assumption 9 imply (ii).

Here, $\det(\cdot)$ denotes the determinant of a square matrix. Following the above remark, the Laplace-regular setting of Kass et al. (1990) translates into the following assumption and proposition.

Assumption 10. *(i) W.p. 1, $\theta \mapsto q_\theta(x'|x)$ and $\theta \mapsto g_\theta(y|x)$ are four-times continuously differentiable for $x, x' \in \mathsf{X}$, $y \in \mathsf{Y}$; the prior $\theta \mapsto \pi(\theta)$ is two-times continuously differentiable.*

(ii) For some $\epsilon > 0$, $\mathcal{B}_\epsilon(\theta_*) \subseteq \Theta$ and w.p. 1, for all $0 \leq j_1 \leq \cdots \leq j_k \leq d$, $k \leq 4$

$$\limsup_n \sup_{\theta \in \mathcal{B}_\epsilon(\theta_*)} \left\{ \tfrac{1}{n} \left| \partial_{\theta_{j_1}} \cdots \partial_{\theta_{j_k}} \ell_\theta(y_{0:n-1}) \right| \right\} < \infty.$$

Proposition 6. *Under Assumptions 1-7, 9-10, we have that, w.p. 1,*

$$\frac{p(y_{0:n-1})}{p_{\hat\theta_n}(y_{0:n-1})} = (2\pi)^{d/2} \, n^{-d/2} \left\{ \det(\mathsf{J}(y_{0:n-1})) \right\}^{-1/2} \pi(\hat\theta_n) \, (1 + \mathcal{O}(n^{-1})).$$

Proof. Under the assumptions, Theorem 2.1 of Tadic and Doucet (2018) ensures that the the log-likelihood $\theta \mapsto \ell_\theta(y_{0:n-1})$ is four-times continuously differentiable. Then, recall from Proposition 2 that $\theta \mapsto \ell_\theta(y_{0:n-1})/n$ converges uniformly to the continuous function $\theta \mapsto \ell(\theta)$ defined therein, which implies that $\hat\theta_n \to \theta_*$, with θ_* the unique maximiser of $\ell(\cdot)$ (under Assumption 6) – all these statements hold w.p. 1. We choose sufficiently small $\delta > 0$ (in Remark 1(i)), then $\epsilon = \epsilon_1$ and $\epsilon = \epsilon_2$ in Assumption 10(ii) and Remark 1(ii) respectively, and $\gamma > 0$ such that for large enough n, $\mathcal{B}_\delta(\theta_*) \subseteq \mathcal{B}_\gamma(\hat\theta_n) \subseteq \mathcal{B}_{\min\{\epsilon_1, \epsilon_2\}}(\theta_*)$. We have that

$$\begin{aligned}
\frac{p(y_{0:n-1})}{p_{\hat\theta_n}(y_{0:n-1})} &= \int_{\Theta - \mathcal{B}_\gamma(\hat\theta_n)} \pi(\theta) \, e^{n \times \frac{1}{n}\{\ell_\theta(y_{0:n-1}) - \ell_{\hat\theta_n}(y_{0:n-1})\}} d\theta \\
&\quad + \int_{\mathcal{B}_\gamma(\hat\theta_n)} \pi(\theta) \, e^{\ell_\theta(y_{0:n-1}) - \ell_{\hat\theta_n}(y_{0:n-1})} d\theta \\
&\leq e^{-cn} + \int_{\mathcal{B}_\gamma(\hat\theta_n)} \pi(\theta) \, e^{\ell_\theta(y_{0:n-1}) - \ell_{\hat\theta_n}(y_{0:n-1})} d\theta,
\end{aligned}$$

for some $c > 0$, where we used Remark 1(i) to obtain the inequality. It remains to treat the integral on $\mathcal{B}_\gamma(\hat\theta_n)$. Applying the Taylor expansions as described in

the main text and continuing from (17) – with the domain of integration now being $\mathcal{B}_\gamma(\hat{\theta}_n)$ – will give,

$$
\mathcal{I}_n := \int_{\mathcal{B}_\gamma(\hat{\theta}_n)} \pi(\theta)\, e^{\ell_\theta(y_{0:n-1}) - \ell_{\hat{\theta}_n}(y_{0:n-1})} d\theta
$$

$$
= \int_{\mathcal{B}_\gamma(\hat{\theta}_n)} e^{-\frac{1}{2} u^\top \mathsf{J}(y_{0:n-1})\, u} \{\pi(\hat{\theta}_n) + n^{-1/2}\, m(u, y_{0:n-1}) + R_n\}\, du. \tag{18}
$$

A careful, but otherwise straightforward, consideration of the structure of the residual R_n gives that, under Remark 1(ii) and Assumption 10(ii),

$$
\frac{1}{(2\pi)^{d/2} \{\det(\mathsf{J}(y_{0:n-1}))\}^{-1/2}} \int_{\mathcal{B}_\gamma(\hat{\theta}_n)} e^{-\frac{1}{2} u^\top \mathsf{J}(y_{0:n-1})\, u} |R_n|\, d\theta = \mathcal{O}(n^{-1}).
$$

Thus, continuing from (18), the change of variables $u = \sqrt{n}(\theta - \hat{\theta}_n)$ implies that, for $f(\cdot; \Omega)$ denoting the pdf of a centred d-dimensional Gaussian distribution with precision matrix Ω,

$$
\mathcal{I}_n = (2\pi)^{d/2}\, n^{-d/2} \{\det(\mathsf{J}(y_{0:n-1})\}^{-1/2}
$$

$$
\times \int_{\mathcal{B}_{\gamma\sqrt{n}}(0)} f(u; \mathsf{J}(y_{0:n-1}))\{\pi(\hat{\theta}_n) + n^{-1/2}\, m(u, y_{0:n-1})\}\, du
$$

$$
\times (1 + \mathcal{O}(n^{-1}))
$$

The final result follows from the fact, that using Assumption 10(i), the integral appearing above is $\mathcal{O}(e^{-c'n})$ apart from the same integral over the whole $\mathbb{R}^d$, for some constant $c' > 0$. □

Proposition 6 implies that, w.p. 1,

$$
\log p(y_{0:n-1}) = \ell_{\hat{\theta}_n}(y_{0:n-1}) - \tfrac{d}{2} \log n + \mathcal{O}(1) + \mathcal{O}(n^{-1}).
$$

Ignoring the terms which are $\mathcal{O}(1)$ w.r.t. n, we obtain that

$$
2 \log p(y_{0:n-1}) \approx 2\ell_{\hat{\theta}_n}(y_{0:n-1}) - d \log n.
$$

Thus, working with the Laplace approximation to the evidence, one can derive the standard formulation of the BIC,

$$
\mathrm{BIC} = -2\ell_{\hat{\theta}_n}(y_{0:n-1}) + d \log n. \tag{19}
$$

Remark 2. *The above results provide an interesting conceptual reassurance. Admitting the evidence as the core principle under which model comparison is carried out, if amongst a family of parametric HMM models, w.p. 1 one has the largest evidence for any big enough n, then BIC is guaranteed to eventually select that model as the optimal one.*

Remark 3. *There is considerable work in the literature regarding consistency properties of the evidence (or Bayes Factor) for classes of models beyond the i.i.d. setting, see e.g. Chatterjee et al. (2018) and the references therein. In our approach, we have brought together results in the literature to deliver assumptions that – whilst being fairly general – were produced with HMMs in mind (and the connection between AIC and the evidence) and are relatively straightforward to be verified, indeed, for HMMs. Alternative approaches typically provide higher level conditions (see e.g. above reference) in an attempt to preserve generality.*

4.2 AIC for HMMs

AIC is developed in Akaike (1974) with its derivation discussed for i.i.d. data and Gaussian models of ARMA type. Following more recent expositions (see e.g. Claeskens and Hjort (2008)), AIC is based on the use of the Kullback-Leibler (KL) divergence for quantifying the distance between the true data-generating distribution and the probability model; an effort to reduce the bias of a 'naive' estimator of the KL divergence leads to the formula for AIC. The case that one does not assume that the parametric model contains the true data distribution corresponds to a generalised version of AIC often called the Takeuchi Information Criterion (TIC), first proposed in Takeuchi (1976). The above ideas are easy to be demonstrated in simple settings (e.g. Claeskens and Hjort (2008) consider i.i.d. and linear regression models).

The framework connecting KL with AIC, in the context of HMMs, can be developed as follows. Let $v(dz_{0:n-1})$ denote the true data-generating distribution, $n \geq 1$. A model is suggested in the form of a family of distributions $\{p_\theta(dz_{0:n-1}); \theta \in \Theta\}$. We assume that $v(dz_{0:n-1})$, $p_\theta(dz_{0:n-1})$ admit densities $v(z_{0:n-1})$, $p_\theta(z_{0:n-1})$ w.r.t. $\nu^{\otimes n}$, $n \geq 1$. We work with the KL distance,

$$
\begin{aligned}
\mathrm{KL}_n(\theta) : = \tfrac{1}{n} \int v(dz_{0:n-1}) \log \frac{v(z_{0:n-1})}{p_\theta(z_{0:n-1})} \\
= \tfrac{1}{n} \int v(dz_{0:n-1}) \log v(z_{0:n-1}) - \tfrac{1}{n} \int v(dz_{0:n-1}) \log p_\theta(z_{0:n-1}). \quad (20)
\end{aligned}
$$

Therefore, minimising the above discrepancy is equivalent to maximising

$$
\mathcal{R}_n(\theta) := \tfrac{1}{n} \int v(dz_{0:n-1}) \log p_\theta(z_{0:n-1}).
$$

Following standard ideas from cases models (e.g. i.i.d. models), one is interested in the quantity $\mathcal{R}_n(\hat{\theta}_n)$, but, in practice, has access only to the naive estimator $\frac{1}{n}\ell_{\hat{\theta}_n}(y_{0:n-1})$, the latter tending to have positive bias versus $\mathcal{R}_n(\hat{\theta}_n)$ due to the use of both the data and the data-induced MLE in its expression. AIC is then derived by finding the larger order term (of size $\mathcal{O}(1/n)$) in the discrepancy of the expectation and appropriately adjusting the naive estimator.

Assumption 11. *(i) There exists a constant $C > 0$, such that w.p. 1,*

$$
\sup_{n \geq 1} \sup_{\theta \in \Theta} \left\{ \tfrac{1}{n} \left| \nabla \nabla^\top \ell_\theta(y_{0:n-1}) \right| \right\} < C.
$$

15

(ii) *There is some $n_0 \geq 1$ such that w.p. 1, matrix $J_{\theta_*}^{-1}(y_{0:n-1})$ – defined in Proposition 4 – is well-posed for all $n \geq n_0$, and there is a constant $C' > 0$, such that w.p. 1,*

$$\sup_{n \geq n_0} |J_{\theta_*}^{-1}(y_{0:n-1})| < C'.$$

These are high-level assumptions – especially Assumption 11(ii) – and a more analytical study is required for them to be of immediate practical use (or weakening them); but such a study would considerably deviate from the main purposes of this work. Our contribution is contained in the following proposition.

Proposition 7. *Under Assumptions 1-9, 11, we have that*

$$\mathbb{E}\left[\tfrac{1}{n}\ell_{\hat{\theta}_n}(y_{0:n-1}) - \mathcal{R}_n(\hat{\theta}_n)\right] = \tfrac{d}{n} + o(n^{-1}).$$

Proof. Use of a second-order Taylor expansion gives,

$$\tfrac{1}{n}\ell_{\hat{\theta}_n}(y_{0:n-1}) - \mathcal{R}_n(\hat{\theta}_n)$$

$$= \tfrac{1}{n}\ell_{\theta_*}(y_{0:n-1}) - \tfrac{1}{n}\int \ell_{\theta_*}(z_{0:n-1})v(dz_{0:n-1})$$

$$+ \tfrac{1}{n}\frac{\nabla^\top \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}}\sqrt{n}(\hat{\theta}_n - \theta_*) - \tfrac{1}{n}\left\{\int v(dz_{0:n-1})\nabla^\top \ell_{\theta_*}(z_{0:n-1})\right\}(\hat{\theta}_n - \theta_*)$$

$$+ \tfrac{1}{2n}\sqrt{n}(\hat{\theta}_n - \theta_*)^\top \left\{\int \mathcal{E}_{\theta_*}(y_{0:n-1}, z_{0:n-1})v(dz_{0:n-1})\right\}\sqrt{n}(\hat{\theta}_n - \theta_*). \qquad (21)$$

where we have set

$$\mathcal{E}_{\theta_*}(y_{0:n-1}, z_{0:n-1}) := \int_0^1 \frac{\nabla\nabla^\top \ell_{s\hat{\theta}_n+(1-s)\theta_*}(y_{0:n-1}) - \nabla\nabla^\top \ell_{s\hat{\theta}_n+(1-s)\theta_*}(z_{0:n-1})}{n}ds.$$

Taking expectations in (21), notice that: i) the expectation of the first difference on the right-hand-side is trivially 0; ii) the integral appearing in the second difference is identically zero, since we are working under the correct model Assumption 8. It remains to consider the expectation of the terms,

$$\zeta_n := \tfrac{1}{n}\frac{\nabla^\top \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}}\sqrt{n}(\hat{\theta}_n - \theta_*);$$

$$\zeta_n' := \tfrac{1}{2n}\sqrt{n}(\hat{\theta}_n - \theta_*)^\top \left\{\int \mathcal{E}_{\theta_*}(y_{0:n-1}, z_{0:n-1})v(dz_{0:n-1})\right\}\sqrt{n}(\hat{\theta}_n - \theta_*). \qquad (22)$$

The first term rewrites as, using (6),

$$\zeta_n = \tfrac{1}{n} \times \frac{\nabla^\top \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}}J_{\theta_*}^{-1}(y_{0:n-1})\frac{\nabla\ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}}$$

Thus, Proposition 5 gives that,

$$n\zeta_n \Rightarrow Z^\top \mathcal{J}_{\theta_*}^{-1}Z; \quad Z \sim N(0, \mathcal{J}_{\theta_*}).$$

For weak convergence to imply convergence in expectation, we require uniform integrability. Assumption 11(ii) takes care of the difficult term $J_{\theta_*}^{-1}(y_{0:n-1})$.

Then, Proposition 5(iii) and the Marcinkiewicz–Zygmund inequality applied for martingales (Ibragimov and Sharakhmetov, 1999), give that

$$\sup_n \|\frac{\nabla \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}}\|_2 < \infty.$$

Thus, from Cauchy–Schwarz, we have

$$\sup \|n\zeta_n\|_2 < \infty,$$

which implies uniform integrability for $\{n\zeta_n\}_n$. So, we have shown that,

$$\mathbb{E}\left[\, n\zeta_n \,\right] \to \mathbb{E}\left[\, Z^\top \mathcal{J}_{\theta_*}^{-1} Z \,\right] \equiv d. \tag{23}$$

We proceed to term ζ_n' in (22). Using again (6) and setting

$$A_{\theta_*}(y_{0:n-1}) := \nabla^\top \ell_{\theta_*}(y_{0:n-1})/\sqrt{n} \cdot J_{\theta_*}^{-1}(y_{0:n-1}),$$

we have that,

$$2n\zeta_n' = A_{\theta_*}(y_{0:n-1})\Big\{ \int \mathcal{E}_{\theta_*}(y_{0:n-1}, z_{0:n-1}) v(dz_{0:n-1}) \Big\} A_{\theta_*}^\top(y_{0:n-1}).$$

Clearly, we can write,

$$\mathbb{E}\left[\, 2n\zeta_n' \,\right]$$
$$= \int \big\{ A_{\theta_*}(y_{0:n-1}) \mathcal{E}_{\theta_*}(y_{0:n-1}, z_{0:n-1}) A_{\theta_*}^\top(y_{0:n-1}) \big\}(v \otimes v)(dy_{0:n-1}, dz_{0:n-1}).$$

From Proposition 3 we obtain that $(v \otimes v)(dy_{0:n-1}, dz_{0:n-1})$-a.s., we have that $\lim_n \mathcal{E}_{\theta_*}(y_{0:n-1}, z_{0:n-1}) \to \mathcal{J}_{\theta_*} - \mathcal{J}_{\theta_*} = 0$. This implies the weak convergence of $A_{\theta_*}(y_{0:n-1}) \mathcal{E}_{\theta_*}(y_{0:n-1}, z_{0:n-1}) A_{\theta_*}^\top(y_{0:n-1}) \Rightarrow 0$. Assumption 11, and arguments similar to the ones used for ζ_n, imply uniform integrability for $\{n\zeta_n'\}$. We thus have $\mathbb{E}\left[\, n\zeta_n' \,\right] \to 0$.

This latter result together with (23) complete the proof. $\qquad\square$

Proposition 7 provides the underlying principle for use of the standard AIC,

$$\mathrm{AIC} := -\, 2\ell_{\hat{\theta}_n}(y_{0:n-1}) + 2d. \tag{24}$$

5 BIC, Evidence, AIC Consistency Properties

We will now use the results in Sections 2-4 to examine the asymptotic properties of BIC, the evidence and AIC in the context of HMMs. We define the notions of strong and weak consistency in model selection in a nested setting as follows.

Definition 1 (Consistency of Model Selection Criterion). *Assume a sequence of nested parametric models*

$$\mathcal{M}_1 \subset \cdots \subset \mathcal{M}_k \subset \mathcal{M}_{k+1} \subset \cdots \subset \mathcal{M}_p,$$

17

for some fixed $p \geq 1$, specified via a sequence of corresponding parameter spaces $\Theta_1 \subseteq \mathbb{R}^{d_1}$, and $\Theta_{k+1} = \Theta_k \times \Delta\Theta_k$, $\Delta\Theta_k \subseteq \mathbb{R}^{d_{k+1}-d_k}$, $k \geq 1$, with $d_k < d_l$ for $k < l$. Let $\mathcal{M}_{k^*}$, for some $k^* \geq 1$, be the smallest model containing the correct one – the latter determined by the true parameter value $\theta_*^{k^*}(=: \theta_*) \in \Theta_{k^*}$.

Let $\mathcal{M}_{\hat{k}_n}$, for index $\hat{k}_n \geq 1$ based on data $\{y_0, \ldots, y_{n-1}\} \in \mathsf{Y}^n$, $n \geq 1$, be the model selected via optimising a Model Selection Criterion. If it holds that $\lim_{n\to\infty} \hat{k}_n = k^*$, w.p. 1, then the Model Selection Criterion is called strongly consistent. If it holds that $\lim_{n\to\infty} \hat{k}_n = k^*$, in probability, then the Model Selection Criterion is called weakly consistent.

We henceforth assume that for each $1 \leq k \leq p$, $\mathcal{M}_k$ corresponds to a parametric HMM as defined in Section 2. The particular model under consideration will be implied by the corresponding parameter appearing in an expression or the superscript k used in relevant functions; i.e., a quantity involving θ^k will refer to model $\mathcal{M}_k$. E.g., $\theta_*^k \in \Theta^k$ is the a.s. limit of the MLE, $\hat{\theta}_n^k$, for model $\mathcal{M}_k$, and such a limit has been shown to exist under Assumptions 1-6 for model $\mathcal{M}_k$.

Assumption 12. *Assumptions 2-6 hold for each parametric model $\mathcal{M}_k$, for index $1 \leq k < k^*$; Assumptions 2-9 hold for each parametric model $\mathcal{M}_k$, for index $k^* \leq k \leq p$.*

Remark 4. *For a model $\mathcal{M}_k$ that contains $\mathcal{M}_{k^*}$ ($k > k^*$), for all of Assumptions 2-9 to hold, it is necessary that the parameterisation of the larger model $\mathcal{M}_k$ is such that non-identifiability issues are avoided. In a trivial example, for $\mathcal{M}_{k^*}$ corresponding to i.i.d. data from $N(\theta_1, 1)$, a larger model of the form $N(\theta_1, \exp(\theta_2))$ would satisfy Assumptions 2 and 9 (the main ones the relate to the shape, in the limit, of the log-likelihood and, consequently identifiability) – one can check this – whereas model $N(\theta_1 + \theta_2, 1)$ would not. In practice, for a given application with nested models, one can most times easily deduce whether identifiability issues are taken care of, thus Assumptions 2-9 correspond to reasonable requirements over the larger models. In general, only 'atypical' parameterisations can produce non-identifyibility issues – thus, also abnormal behavior of the log-likelihood function – for the case of the larger model.*

Proposition 8. *Let $\lambda_n := \ell_{\hat{\theta}_n^k}^k(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k^*}}^{k^*}(y_{0:n-1})$, for a $k \neq k^*$. Under Assumptions 2-9, 12 we have the following.*

   (i) If $\mathcal{M}_k \subset \mathcal{M}_{k^*}$, then $\lim_{n\to\infty} n^{-1}\lambda_n = \ell^k(\theta_*^k) - \ell^{k^*}(\theta_*) < 0$, w.p. 1.

   (ii) If $\mathcal{M}_k \supset \mathcal{M}_{k^*}$, then $\lambda_n \geq 0$ and $\lambda_n = \mathcal{O}(\log \log n)$, w.p. 1.

Proof. (i) From Proposition 2 we have, w.p. 1,

$$n^{-1}\big(\ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k^*}}^{k^*}(y_{0:n-1})\big) \to \ell^k(\theta_*^k) - \ell^{k^*}(\theta_*)$$
$$\equiv \mathbb{E}\left[\log p_{\theta_*^k}(y_0|y_{-\infty:-1})\right] - \mathbb{E}\left[\log p_{\theta_*}(y_0|y_{-\infty:-1})\right].$$

Using Jensen's inequality and simple calculations, one obtains that

$$\mathbb{E}\left[\log p_{\theta_*^k}(y_0|y_{-\infty:-1})\right] - \mathbb{E}\left[\log p_{\theta_*}(y_0|y_{-\infty:-1})\right]$$

$$\leq \log \int \frac{p_{\theta_*^k}(y_0|y_{-\infty:-1})}{p_{\theta_*}(y_0|y_{-\infty:-1})} p_{\theta_*}(y_{-\infty:0}) dy_{-\infty:0} \equiv \log 1.$$

For strict inequality, Assumptions 6 and 12 imply that mapping $\theta \mapsto \ell^{k^*}(\theta)$ has the unique maximum $\theta_* \in \Theta_{k^*}$. Thus, we cannot have $\ell^k(\theta_*^k) = \ell^{k^*}(\theta_*)$, as this would give (from the nested model structure) $\ell^{k^*}(\theta_*) = \ell^{k^*}(\theta_k, \theta_0)$ for some $\theta_0 \in \prod_{l=k+1}^{k^*} \Delta\Theta_l$, with $(\theta_k, \theta_0)^\top \neq \theta_*$ (otherwise the definition of correct model class would be violated).

(ii) Having $\lambda_n \geq 0$ is a consequence of the log-likelihood for model $\mathcal{M}_k$ being maximised over a larger parameter domain than $\mathcal{M}_{k^*}$. Then, notice that the limiting matrix $\mathcal{J}_{\theta_*}$ in Proposition 4 (for the notation used therein) is positive-definite: it is non-negative-definite following its definition in (13); then, non-singularity Assumption 9 provides the positive-definiteness. From Proposition 5(i), the difference in the definition of λ_n equals the difference of two quadratic forms, as the constants in the expression for the log-likelihood provided by Proposition 5(i) are equal for models $\mathcal{M}_{k^*}$, $\mathcal{M}_k$ and cancel out. As $\lambda_n \geq 0$, and both quadratic forms are non-negative, it suffices to consider the one for model $\mathcal{M}_k$. The a.s. convergence of the positive-definite matrix in the quadratic form implies a.s. convergence of its eigenvalues and eigenvectors. Thus, using Theorem 1, overall one has that $\lambda_n = \mathcal{O}(\sum_{i=1}^d \log\log n) = \mathcal{O}(\log\log n)$. $\square$

5.1 Asymptotic Properties of BIC and Evidence

BIC is known to be strongly consistent in i.i.d. settings and some particular non-i.i.d. ones (Claeskens and Hjort, 2008). In the context of HMMs, Gassiat and Boucheron (2003) show strong consistency of BIC for observations that take a finite set of values. The key tool to obtain strong consistency of BIC in a general HMM is LIL we obtained in Section 3. Nishii (1988) also uses LIL for the i.i.d. setting to prove strong (and weak) consistency of BIC.

Recall that k^* denotes the index of the correct model.

Proposition 9. *(i) Let $\hat{k}_n$ be the index of the selected model obtained via minimizing BIC as defined in (19). Then, under Assumptions 1-9, 12, we have that $\hat{k}_n \to k^*$, w.p. 1.*

(ii) If $\hat{\mathsf{k}}_n$ denotes the index obtained via maximising the evidence in (15), then Assumptions 1-12 imply that $\hat{\mathsf{k}}_n \to k^*$, w.p. 1.

Proof. (i) We make use of Proposition 8. Indeed, in the case that $\mathcal{M}_k \subset \mathcal{M}_{k^*}$, Proposition 8(i) gives

$$\text{BIC}_n(\mathcal{M}_k) - \text{BIC}_n(\mathcal{M}_{k^*})$$

$$= n\left\{\tfrac{1}{n}\ell_{\hat{\theta}_n^{k^*}}(y_{0:n-1}) - \tfrac{1}{n}\ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \tfrac{(d_k - d_{k^*})\log n}{n}\right\},$$

$$\to +\infty \quad w.p.\, 1,$$

therefore $\liminf_n \hat{k}_n \geq k^*$, w.p. 1. In the case $\mathcal{M}_k \supset \mathcal{M}_{k^*}$, we obtain, from Proposition 8(ii),

$$
\begin{aligned}
\mathrm{BIC}_n(\mathcal{M}_k) &- \mathrm{BIC}_n(\mathcal{M}_{k^*}) \\
&= (d_k - d_{k^*}) \log n - \{\ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k^*}}(y_{0:n-1})\} \\
&= (d_k - d_{k^*}) \log n - \mathcal{O}(\log \log n).
\end{aligned}
$$

Thus, w.p. 1, for all large enough n, $\mathrm{BIC}_n(\mathcal{M}_k) - \mathrm{BIC}_n(\mathcal{M}_{k^*}) > c_k > 0$, for some constant c_k. Therefore, we have $\limsup_n \hat{k}_n \leq k^*$, w.p. 1.

(ii) Given part (i), this now follows directly from Proposition 6.

$\square$

Therefore, BIC is strongly consistent for a general class of HMMs in the nested model setting we are considering here – with a model assumed to be a correctly specified one.

5.2 Asymptotic Properties of AIC

We can be quite explicit about the behaviour of AIC. Consider the case $k > k^*$. Making use of Proposition 5 we have

$$
\begin{aligned}
\ell_{\hat{\theta}_n^k}(y_{0:n-1}) &- \ell_{\hat{\theta}_n^{k^*}}(y_{0:n-1}) \\
&= \frac{1}{2} \frac{\nabla^\top \ell_{\theta_*^k}(y_{0:n-1})}{\sqrt{n}} \mathcal{J}_{\theta_*^k}^{-1} \frac{\nabla \ell_{\theta_*^k}(y_{0:n-1})}{\sqrt{n}} \\
&\quad - \frac{1}{2} \frac{\nabla^\top \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}} \mathcal{J}_{\theta_*}^{-1} \frac{\nabla \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}} + \epsilon_n,
\end{aligned}
\tag{25}
$$

where $\epsilon_n = o(\log \log n)$, w.p. 1. Due to working with nested models, we have (immediately from the definition of $\mathcal{J}_{\theta_*^k}$, $\mathcal{J}_{\theta_*}$)

$$
\mathcal{J} := \mathcal{J}_{\theta_*^k} = \begin{pmatrix} \mathcal{J}_{11} & \mathcal{J}_{12} \\ \mathcal{J}_{21} & \mathcal{J}_{22} \end{pmatrix} \in \mathbb{R}^{d_k \times d_k}.
$$

where $\mathcal{J}_{11} \equiv \mathcal{J}_{\theta_*}$, and $\mathcal{J}_{12}$, $\mathcal{J}_{21} = \mathcal{J}_{12}^\top$ as deduced from $\mathcal{J}_{\theta_*^k}$. Similarly, the quantity $\nabla \ell_{\theta_*}(y_{0:n-1})$ forms the upper d_{k^*}-dimensional part of $\nabla \ell_{\theta_*^k}(y_{0:n-1})$. We will make use of the matrix equations implied by

$$
\mathcal{J}\mathcal{J}^{-1} = I_{d_k} \iff
$$
$$
\begin{pmatrix} \mathcal{J}_{11} & \mathcal{J}_{12} \\ \mathcal{J}_{21} & \mathcal{J}_{22} \end{pmatrix} \begin{pmatrix} (\mathcal{J}^{-1})_{11} & (\mathcal{J}^{-1})_{12} \\ (\mathcal{J}^{-1})_{21} & (\mathcal{J}^{-1})_{22} \end{pmatrix} = \begin{pmatrix} I_{d_{k^*}} & 0_{d_{k^*} \times (d_k - d_{k^*})} \\ 0_{(d_k - d_{k^*}) \times d_{k^*}} & I_{(d_k - d_{k^*})} \end{pmatrix}.
$$

Given the above nesting considerations, some cumbersome but otherwise straightforward calculations give

$$
\begin{aligned}
\frac{\nabla^\top \ell_{\theta_*^k}(y_{0:n-1})}{\sqrt{n}} &\mathcal{J}_{\theta_*^k}^{-1} \frac{\nabla \ell_{\theta_*^k}(y_{0:n-1})}{\sqrt{n}} - \frac{\nabla^\top \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}} \mathcal{J}_{\theta_*}^{-1} \frac{\nabla \ell_{\theta_*}(y_{0:n-1})}{\sqrt{n}} \\
&\equiv \frac{\{M \nabla \ell_{\theta_*^k}(y_{0:n-1})\}^\top}{\sqrt{n}} D \frac{M \nabla \ell_{\theta_*^k}(y_{0:n-1})}{\sqrt{n}},
\end{aligned}
\tag{26}
$$

where we have set

$$M := \left((\mathcal{J}^{-1})_{21} \ \ \{(\mathcal{J}^{-1})_{22}\}^{-1} \right) \in \mathbb{R}^{(d_k - d_{k^*}) \times d_k};$$

$$D := \{(\mathcal{J}^{-1})_{22}\}^{-1} \in \mathbb{R}^{(d_k - d_{k^*}) \times (d_k - d_{k^*})}.$$

Consider the standard decomposition of the symmetric positive-definite D,

$$D = P \Lambda P^\top,$$

for orthonormal $P \in \mathbb{R}^{(d_k - d_k^*) \times (d_k - d_k^*)}$ and diagonal $\Lambda \in \mathbb{R}^{(d_k - d_k^*) \times (d_k - d_k^*)}$.

Assumption 13. *Define the martingale increments in $\mathbb{R}^{d_k - d_{k^*}}$, $k > k^*$.*

$$\tilde{h}_{\theta_*^k}(y_{-\infty:0}) := (\sqrt{\Lambda} P^\top M) \, h_{\theta_*^k}(y_{-\infty:0}).$$

We have that $\mathbb{E} \left| \tilde{h}_{\theta_*^k}(y_{-\infty:0}) \right|^2 > 0$.

Proposition 10. *Under Assumptions 1-9, 12-13, we have that, for $k > k^*$,*

$$\mathbb{P} \left[\, \mathrm{AIC}_n(\mathcal{M}_k) < \mathrm{AIC}_n(\mathcal{M}_{k^*}), \ \text{infinitely often in } n \geq 1 \right] = 1.$$

Proof. Continuing from (25), (26), the use of LIL for martingale increments will give that, w.p. 1,

$$\limsup_n \frac{\sqrt{2}\{\ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k^*}}(y_{0:n-1})\}}{\log \log n}$$
$$\geq \sup_{1 \leq j \leq d_k - d_{k^*}} \mathbb{E} \left[(\tilde{h}_{\theta_*^k}(y_{-\infty:0}))_j^2 \right] > 0.$$

As the difference $\ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k^*}}(y_{0:n-1})$ is of size $\Theta(\log \log n) - o(\log \log n)$ infinitely often, the result follows immediately. (The notation $\Theta(a_n)$ for a positive sequence $\{a_n\}$ means that the sequence of interest is lower and upper bounded by ca_n and $c'a_n$ respectively for constants $0 < c < c'$.) $\square$

Thus, AIC is not a consistent Model Selection Criterion – in contrast with BIC. Still, it is well known that AIC has desirable properties, e.g. with regards to prediction error (in many cases the model chosen by AIC attains the minimum maximum error in terms of prediction among models being considered), or its minimax optimality. Barron et al. (1999) is a comprehensive article of this topic and shows that minimax optimality of AIC holds in many cases, including the i.i.d., some non-linear models and for density estimation. For works on the efficiency of AIC terms of prediction see Shibata (1980, 1981); Shao (1997). AIC is equivalent to LOO cross-validation (Stone, 1977) for i.i.d.-type model structures. We refer to Ding et al. (2017, 2018) for a comprehensive review of AIC and BIC. Note that Yang (2005) shows that consistency of model selection and minimax optimality do not necessarily hold simultaneously. Our main focus in this work is asymptotic behaviour of criteria from a model selection viewpoint, so we will not further examine the prediction perspective.

5.3 A General Result

Following e.g. Sin and White (1996), one can generalise some of the above results for arbitrary penalty functions. Assume that we consider Information Criterion (IC) of the form

$$\text{IC}_n(\mathcal{M}_k) = -\ell_{\hat{\theta}_n^k}(y_{0:n-1}) + pen_n(k), \tag{27}$$

for a penalty function $pen_n(k) \in \mathbb{R}$, (strictly) increasing in $k \geq 1$.

Proposition 11. *(i) If the following hold, for $k' > k \geq 1$,*

$$\lim_{n\to\infty} \frac{pen_n(k') - pen_n(k)}{n} = 0, \quad \lim_{n\to\infty} \frac{pen_n(k') - pen_n(k)}{\log\log n} = +\infty,$$

then, under Assumptions 1-9, 12, the information criterion IC_n in (27) is strongly consistent.

(ii) If the following hold, for $k' > k \geq 1$,

$$\lim_{n\to\infty} \frac{pen_n(k') - pen_n(k)}{n} = 0, \quad \lim_{n\to\infty} pen_n(k') - pen_n(k) = +\infty,$$

then, under Assumptions 1-9, 12, the Information Criterion IC_n in (27) is weakly consistent.

Proof. (i) It is an immediate generalisation of the proof of Proposition 9.

(ii) First, let us consider the case when $\mathcal{M}_k \subset \mathcal{M}_{k^*}$. Then, for any $\epsilon > 0$ we have

$$\begin{aligned}
\mathbb{P}\left[\,\text{IC}_n(\mathcal{M}_k) - \text{IC}_n(\mathcal{M}_{k^*}) > \epsilon\,\right] \\
= \mathbb{P}\left[\,\ell_{\hat{\theta}_n^{k^*}}(y_{0:n-1}) - \ell_{\hat{\theta}_n^k}(y_{0:n-1}) + (pen_n(k) - pen_n(k^*)) > \epsilon\,\right] \\
\to 1.
\end{aligned}$$

The limit follows from Proposition 8(i), as the random variable on the left side of the inequality above diverges to $+\infty$ w.p. 1, and a.s. convergence implies convergence in probability. This result implies directly $\lim_{n\to\infty} \mathbb{P}\left[\,\hat{k}_n \geq k^*\,\right] = 1$, where $\hat{k}_n$ denotes the model index minimising $\text{IC}_n(\mathcal{M}_k)$, $1 \leq k \leq p$.

Next, we consider the case where $\mathcal{M}_{k^*} \subset \mathcal{M}_k$. We have that

$$\begin{aligned}
\mathbb{P}\left[\,\text{IC}_n(\mathcal{M}_k) \leq \text{IC}_n(\mathcal{M}_{k^*})\,\right] \\
= \mathbb{P}\left[\,\ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k^*}}(y_{0:n-1}) \geq (pen_n(k) - pen_n(k^*))\,\right]. \tag{28}
\end{aligned}$$

Recall the expression for $\ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k^*}}(y_{0:n-1})$ implied by Proposition 5(i). The martingale CLT in Proposition 5(iii) gives that

$$\frac{\nabla\ell_{\theta_*^k}(y_{0:n-1})}{\sqrt{n}} \Rightarrow N_{d_k}\left(0, \mathcal{J}_{\theta_*^k}\right),$$

with $\mathcal{J}_{\theta_*^k}$ defined in the obvious manner. Let $Z \sim N_{d_k}(0, \mathcal{J}_{\theta_*^k})$; the continuous mapping theorem gives that

$$\ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k*}}(y_{0:n-1}) \Rightarrow$$
$$= \tfrac{1}{2} Z^\top \mathcal{J}_{\theta_*^k}^{-1} Z - \tfrac{1}{2} Z_{1:d_{k*}}^\top \mathcal{J}_{\theta_*}^{-1} Z_{1:d_{k*}} =: Z_0$$

Continuing from (28), since $|Z_0| < \infty$, w.p. 1, for any $\epsilon > 0$ we can have some n_0 so that for all $n_1 \geq n_0$, $\mathbb{P}\left[\, Z_0 \geq pen_{n_1}(k) - pen_{n_1}(k^*) \,\right] < \epsilon$. Thus, for all n large enough, we have that

$$\mathbb{P}\left[\, \ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k*}}(y_{0:n-1}) \geq (pen_n(k) - pen_n(k^*)) \,\right]$$
$$\leq \mathbb{P}\left[\, \ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k*}}(y_{0:n-1}) \geq (pen_{n_1}(k) - pen_{n_1}(k^*)) \,\right]$$
$$\rightarrow \mathbb{P}\left[\, Z_0 \geq pen_{n_1}(k) - pen_{n_1}(k^*) \,\right] < \epsilon.$$

Thus, we conclude that $\mathbb{P}\left[\, \mathrm{IC}_n(\mathcal{M}_k) \leq \mathrm{IC}_n(\mathcal{M}_{k^*}) \,\right] \rightarrow 0$, so that we have obtained $\lim_{n \to \infty} \mathbb{P}\left[\, \hat{k}_n \leq k^* \,\right] = 1$.

Overall, we have shown that $\lim_{n \to \infty} \mathbb{P}\left[\, \hat{k}_n = k^* \,\right] = 1$. $\qquad\square$

The above results highlight that the penalty function should grow to infinity with the sample size (at certain rate) to deliver strongly or weakly consistent IC.

6 Particle Approximation of AIC and BIC

BIC and AIC can be used for model selection for HMMs but are typically impossible to calculate analytically due to intractability of the likelihood function for general HMMs. Thus, an approximation technique is required. We adopt the computational approach of Poyiadjis et al. (2011) which – for completeness – we briefly review in this section. It involves a particle filtering algorithm coupled with a recursive construction for an integral approximation.

The description follows closely Poyiadjis et al. (2011). The marginal Fisher identity gives,

$$\nabla \ell_\theta(y_{0:n}) = \int_{\mathsf{X}} \nabla \log p_\theta(x_n, y_{0:n}) p_\theta(x_n | y_{0:n}) \mu(dx_n).$$

At step n, let $(x_n^{(i)}, w_n^{(i)})_{i=1}^N$ be a particle approximation of $p_\theta(dx_n | y_{0:n})$, with standardised weights, i.e. $\sum_i w_n^{(i)} = 1$, obtained via some particle filtering algorithm, so that,

$$\nabla \ell_\theta(y_{0:n}) \approx \sum_{i=1}^N w_n^{(i)} \nabla \log p_\theta(x_n^{(i)}, y_{0:n}). \tag{29}$$

We explore the unknown quantity $\nabla \log p_\theta(x_n, y_{0:n})$. We have,

$$p_\theta(x_n, y_{0:n})$$
$$= p_\theta(y_{0:n-1}) g_\theta(y_n | x_n) \int_{\mathsf{X}} q_\theta(x_n | x_{n-1}) p_\theta(x_{n-1} | y_{0:n-1}) \mu(dx_{n-1}). \tag{30}$$

23

This implies that,

$$\nabla p_\theta(x_n, y_{0:n}) = p_\theta(y_{0:n-1})g_\theta(y_n|x_n)\int_{\mathsf{X}}q_\theta(x_n|x_{n-1})p_\theta(x_{n-1}|y_{0:n-1})\times$$

$$\left\{\nabla\log g_\theta(y_n|x_n) + \nabla\log q_\theta(x_n|x_{n-1}) + \nabla\log p_\theta(x_{n-1}, y_{0:n-1})\right\}\mu(dx_{n-1}). \quad (31)$$

At step $n-1$, let $(x_{n-1}^{(j)}, w_{n-1}^{(j)})_{j=1}^N$ be a particle approximation of the filtering distribution $p_\theta(dx_{n-1}|y_{0:n-1})$ and $(\alpha_{n-1}^{(j)})_{j=1}^N$ a sequence of approximations to $(\nabla\log p_\theta(x_{n-1}^{(j)}, y_{0:n-1}))_{j=1}^N$. Equations (30), (31) suggest the following recursive approximation of $\nabla\log p_\theta(x_n^{(i)}, y_{0:n})$, for $1\leq i\leq N$,

$$\alpha_n^{(i)} = \frac{\sum_{j=1}^N w_{n-1}^{(j)}q_\theta(x_n^{(i)}|x_{n-1}^{(j)})}{\sum_{j=1}^N w_{n-1}^{(j)}q_\theta(x_n^{(i)}|x_{n-1}^{(j)})}\times$$

$$\left\{\nabla\log g_\theta(y_n|x_n^{(i)}) + \nabla\log q_\theta(x_n^{(i)}|x_{n-1}^{(j)}) + \alpha_{n-1}^{(j)}\right\}, \quad (32)$$

Thus, from (29), one obtains an estimate of the score function at step n, as

$$\nabla\ell_\theta(y_{0:n}) \approx \sum_{i=1}^N w_n^{(i)}\alpha_n^{(i)}. \quad (33)$$

The calculation in (32), and the adjoining particle filtering algorithm, can be applied recursively to provide the approximation of the score function in (33) for $n = 0, 1, \ldots$. Note that the computational cost of this algorithm is $\mathcal{O}(N^2)$, but is robust for increasing n as it is based on the approximation of the filtering distributions rather than the smoothing ones, see results and comments on this point in Poyiadjis et al. (2011).

Moreover, Poyiadjis et al. (2011) use the score function estimation methodology to propose an *online* gradient ascent algorithm for obtaining an MLE-type parameter estimate. In more detail, the method is based on the recursion

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\nabla\log p_\theta(y_n|y_{0:n-1})\big|_{\theta=\theta_n}, \quad (34)$$

where $\{\gamma_k\}_{k\geq 1}$ is a positive decreasing sequence with

$$\sum_{k=1}^\infty \gamma_k = \infty, \quad \sum_{k=1}^\infty \gamma_k^2 < \infty.$$

To deduce an online algorithm – following ideas in Le Gland and Mevel (1997) – intermediate quantities involved in the recursions in (29)-(32) are calculated at different, consecutive parameter values. See Poyiadjis et al. (2011) for more details, and Le Gland and Mevel (1997); Tadic and Doucet (2018) for analytical studies on the convergence properties of the algorithm. In particular, under strict conditions the algorithm is shown to converge to the maximiser θ_* of the limiting function of $\theta \mapsto \ell_n(\theta)/n$, as $n \to \infty$.

Remark 5. *In our setting, we want to use the numerical studies to illustrate the theoretical results obtained for AIC and BIC, so we will use the outcome of the online recursion as proxy for the MLE. Then, the AIC and BIC will be approximated by running a particle filter for the chosen MLE value to obtain an approximation of the log-likelihood of the data at this parameter value.*

7   Empirical Study

We consider the following stochastic volatility model (labeled as $\mathcal{SV}$)

$$\mathcal{SV}: \quad \left\{ \begin{array}{l} X_t = \phi X_{t-1} + W_t, \\ Y_t = \exp(X_t/2)V_t, \quad t \geq 1, \end{array} \right.$$

and the one with jumps (labeled as $\mathcal{SVJ}$),

$$\mathcal{SVJ}: \quad \left\{ \begin{array}{l} X_t = \phi X_{t-1} + W_t \\ Y_t = \exp(X_t/2)V_t + q_t J_t, \quad t \geq 1, \end{array} \right.$$

where $W_t \sim N(0, \sigma_X^2)$, $V_t \sim N(0,1)$, $J_t \sim N(0, \sigma_J^2)$ and $q_t \sim Bernoulli(p)$, all variables assumed independent over the time index $t \geq 1$. In both cases $X_0 = 0$. The extended model $\mathcal{SVJ}$ can be used to capture instantaneous big jumps in the relative changes of the values of the underlying asset; the choice of models has been motivated by their use in the literature, see e.g. Pitt et al. (2014). Figure 1 shows two sets of $n = 10^4$ simulated observations, one from $\mathcal{SV}$ and one from $\mathcal{SVJ}$, under the corresponding true parameter values

$$(\phi, \sigma_X) = (0.9, \sqrt{0.3}); \quad (\phi, \sigma_X, \sigma_J, p) = (0.9, \sqrt{0.3}, \sqrt{0.6}, 0.6).$$

These simulated data will be used in the experiments that follow. Scenario 1 (resp. Scenario 2) corresponds to the case when the true model is $\mathcal{SVJ}$ (resp. $\mathcal{SV}$). We will compare the two models, using AIC and BIC, in both Scenarios. The estimated parameter values for $\mathcal{SV}$ and $\mathcal{SVJ}$ – and the estimates for AIC and BIC using a particle filter – are obtained via the method of Poyiadjis et al. (2011), reviewed in Section 6. Note that, as we have established in this work, BIC is expected to be consistent for both Scenarios, whereas AIC only for the first Scenario.

We set $\gamma_n = n^{-2/3}$ for all numerical experiments in the sequel. Figure 2 shows estimated parameter values for $\mathcal{SV}$, $\mathcal{SVJ}$, for both simulated datasets, sequentially in the data size, using the online version of the method of Section 6, with $N = 200$ particles. (We tried also larger number of particles, with similar results.) To further investigate the stability of the algorithm in Section 6 we summarize in Figures 3, 4 estimates of AIC and BIC for the two models from $R = 200$ replications of the same algorithm, for different data sizes. Figure 3 corresponds to Scenario 1 and Figure 4 to Scenario 2. The results obtained seem to indicate that the numerical algorithm used for approximating AIC and BIC is fairly robust in all cases. Also, it appears that in the challenging Scenario 2,
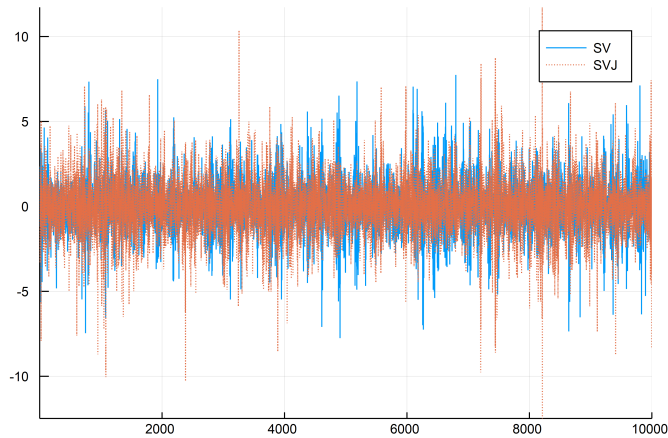
Figure 1: The $n = 10^4$ simulated observations from models $\mathcal{SV}$, $\mathcal{SVJ}$, with parameter values $(\phi, \sigma_X) = (0.9, \sqrt{0.3})$, $(\phi, \sigma_X, \sigma_J, p) = (0.9, \sqrt{0.3}, \sqrt{0.6}, 0.6)$, respectively, to be used in the experiments.
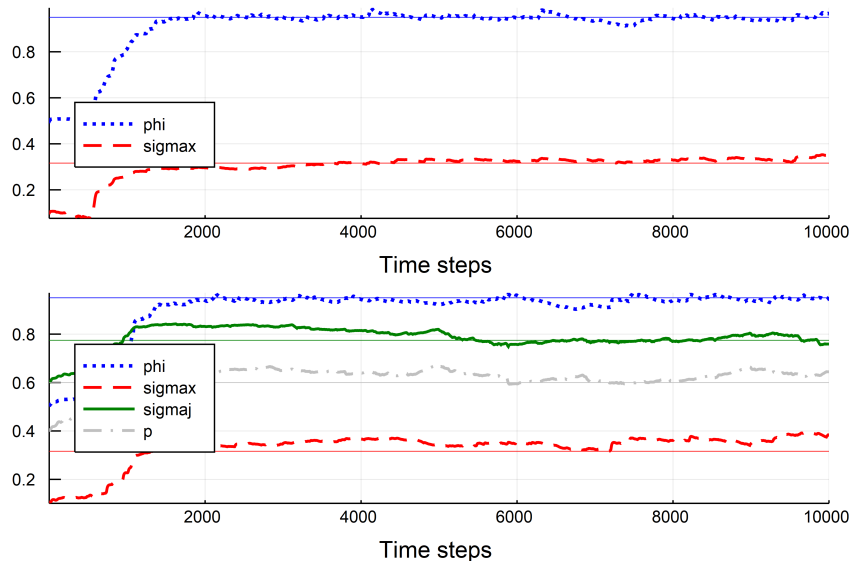

Figure 2: Estimated parameters for the $\mathcal{SV}$ (top panel) and $\mathcal{SVJ}$ (bottom panel) models as obtained – sequentially in time – via the data simulated from the $\mathcal{SV}$ (top panel) and $\mathcal{SVJ}$ (bottom panel) models respectively and the algorithm reviewed in Section 6 with $N = 200$ particles. The horizontal lines indicate the true parameter values in each case.
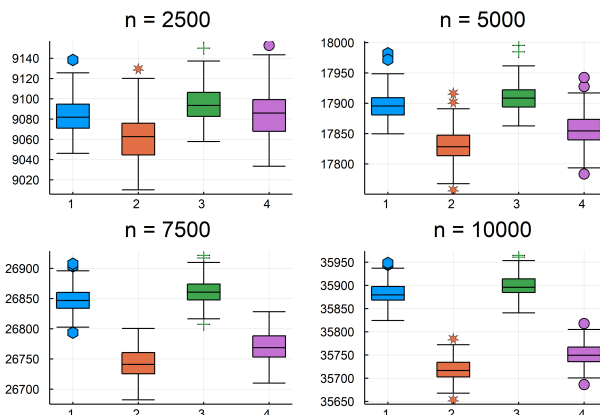
Figure 3: Boxplots for Scenario 1 ($\mathcal{SVJ}$ model is true) from $R = 200$ estimates (R denotes the replications of the numerical algorithm) of an Information Criterion (IC) and various observation sizes. Blue: AIC($\mathcal{SV}$), Orange: AIC($\mathcal{SVJ}$), Green: BIC($\mathcal{SV}$), Purple: BIC($\mathcal{SVJ}$).
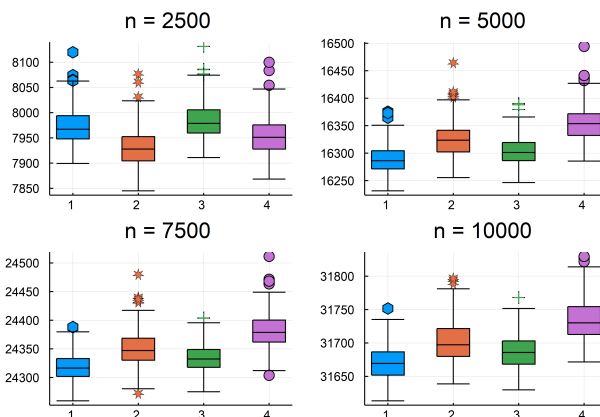

Figure 4: Boxplots for Scenario 2 ($\mathcal{SV}$ model is true) from $R = 200$ estimates of an IC and various observation sizes. Blue: AIC($\mathcal{SV}$), Orange: AIC($\mathcal{SVJ}$), Green: BIC($\mathcal{SV}$), Purple: BIC($\mathcal{SVJ}$).

even with $n = 10^4$ observations, the boxplots do not seem to provide any decisive evidence in favor of true model $\mathcal{SV}$.

Table 1 shows results from the same $R = 200$ replications for the estimation of AIC and BIC for each of the two models and two simulated datasets. In agreement with our theory, BIC appears more robust (than AIC) at choosing the correct model for the dataset simulated from $\mathcal{SV}$. Figure 5 plots differences

27

in AIC and BIC in Scenario 2, sequentially in the data size – more accurately, a proxy of the differences, see Remark 5. To be precise, the blue line denotes the 'path' of $\text{AIC}(\mathcal{SV}) - \text{AIC}(\mathcal{SVJ})$, and the red line denotes the one of $\text{BIC}(\mathcal{SV}) - \text{BIC}(\mathcal{SVJ})$. Since model $\mathcal{SV}$ is true in this case, the difference should be lower than zero for large enough n if the used IC were consistent. As one can see, the difference in BIC is always negative after a large enough sample size n. In contrast, and in agreement with our theory, the difference in AIC never has such property. For instance, sometime after $n = 10^4$, the difference increased and exceeded the zero line. This is a clear empirical manifestation of Proposition 10; so, whereas in the previous plots the deficiency of AIC was difficult to showcase when looking at *fixed* data sizes, such deficiency became clear when we look at the evolution of AIC as a function of data size.

n	$2,500$	$5,000$	$7,500$	$10,000$
$\text{AIC}(\mathcal{SV})$	$32/200$	$4/200$	$0/200$	$0/200$
$\text{AIC}(\mathcal{SVJ})$	$168/200$	$196/200$	$200/200$	$200/200$
$\text{BIC}(\mathcal{SV})$	$51/200$	$5/200$	$0/200$	$0/200$
$\text{BIC}(\mathcal{SVJ})$	$149/200$	$195/200$	$200/200$	$200/200$

n	$2,500$	$5,000$	$7,500$	$10,000$
$\text{AIC}(\mathcal{SV})$	$154/200$	$161/200$	$153/200$	$158/200$
$\text{AIC}(\mathcal{SVJ})$	$46/200$	$39/200$	$47/200$	$42/200$
$\text{BIC}(\mathcal{SV})$	$179/200$	$173/200$	$184/200$	$192/200$
$\text{BIC}(\mathcal{SVJ})$	$21/200$	$27/200$	$16/200$	$8/200$

Table 1: Results after $R = 200$ replications of the approximation algorithm with $N = 200$ particles. The top (resp. bottom) table shows results for the data simulated from $\mathcal{SVJ}$ (resp. $\mathcal{SV}$). The 1st (resp. 2nd) row in each table shows the fraction of the replications where the estimated AIC is smaller for the $\mathcal{SV}$ model (resp. $\mathcal{SVJ}$ model) for different data sizes; rows 3 and 4 show similar results for BIC.

8   Conclusions and Remarks

We have investigated the asymptotic behaviour of BIC, the evidence and AIC for nested HMMs, and have derived new results concerning their consistency properties. Our work shows that BIC – and the evidence – are strongly consistent for a general class of HMMs. In contrast, for a similarly posed Model Selection problem, AIC is not even weakly consistent. Our study focuses on asymptotics for increasing data size, so we do not investigate finite sample-size results for BIC, evidence and AIC, such as optimality properties. It is well-known that AIC is minimax-rate optimal but BIC is not in many cases, see e.g. Barron et al. (1999). We conjecture this might also be the case for general HMMs.

The technique of constructing stationary, ergodic processes by introducing a backward infinite extension of the observations – see (3) in Section 2 – has been
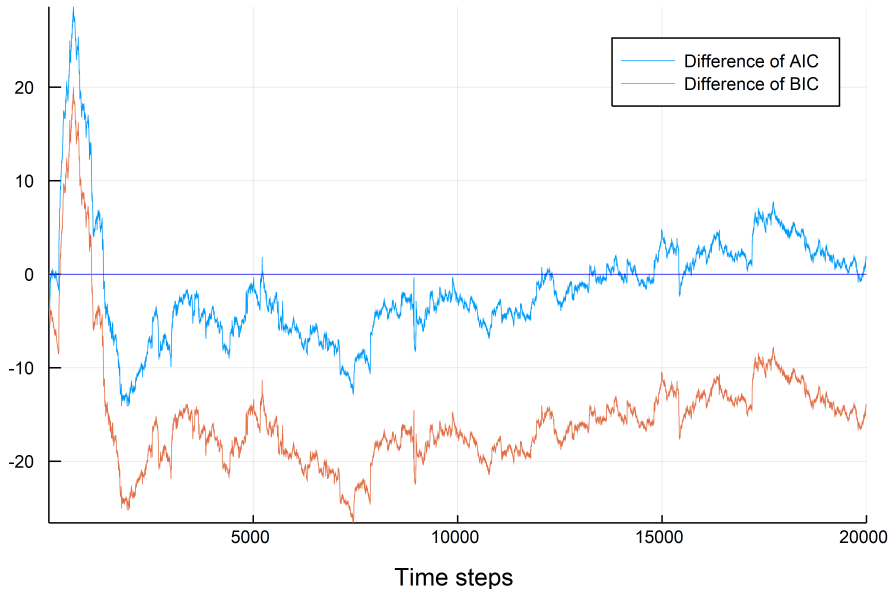
Figure 5: The path of differences in AIC and BIC in Scenario 2 ($\mathcal{SV}$ is the true model). That is, the blue line show the approximated value of $\text{AIC}(\mathcal{SV}) - \text{AIC}(\mathcal{SVJ})$ as a function of data size, and the red line the corresponding function for $\text{BIC}(\mathcal{SV}) - \text{BIC}(\mathcal{SVJ})$.

used in many other studies, even beyond HMMs. E.g. Douc et al. (2020) use this approach to study posterior consistency for a class of partially observed Markov models; Lehéricy (2018) use it to investigate non-asymptotic behaviour of MLE for (finite state space) HMMs; Le Corff and Fort (2013) apply the technique within an online EM setting for HMMs; Diel et al. (2020) consider more general classes of latent variable models.

We note that asymptotic results about the MLE for HMMs have recently been obtained under weaker conditions. See, e.g., Douc et al. (2011, 2016) for developments that go beyond compact spaces. Here, we have worked with strict assumptions on the state space (see – indicatively – Assumption 3, Section 2), so that we obtain an important first set of illustrative results for Model Selection Criteria, avoiding at the same time an overload of technicalities. Future investigations are expected to further weaken the conditions we have used here.

There are challenges when trying to move beyond the Model-Correctness setting. As we have described in the first parts of the paper, Douc and Moulines (2012) show that the MLE converges a.s. even for misspecified models under mild assumptions. However, a CLT for the MLE in the context of general state-space misspecified HMMs has yet to been proven. To the best of our knowledge, only Pouzo et al. (2016) obtain such a result for a finite state space X. Thus,

extending our results to non-nested settings or/and ones where one does not assume correctness of a model, is a non-trivial undertaking that requires extensive further research. Also, we note that AIC is asymptotically prediction efficient in some misspecified models whilst BIC is not. The above discussion suggests that investigating asymptotic behaviour of model selection criteria under No-Model-Correctness for general HMM models is an important open problem that requires further research.

One can use alternative numerical algorithms instead of the one we have used here, and describe in Section 6 – see e.g. the approach in Olsson and Alenlöv (2017). Note that the numerical algorithm used in the paper is mostly a tool for illustrating our theoretical findings, which is the main focus of our work. The numerical study shown in the paper already delivers the points stemming from the theory, so we have refrained from describing/implementing further methods to avoid diverting attention from our main findings.

From a practitioner point of view, our results and numerical study indicate that AIC can wrongly select the more complex model due to ineffective penalty term. Critically, this can be difficult to assess using standard experiments. Our study has shown that one needs to investigate the evolution of AIC against data size to clearly highlight its deficiency in the context of a numerical study. We stress here that in the numerical experiment we have knowingly used models for which several of the stated Assumptions will not hold (maybe most notably, the strong mixing Assumption 3). The aim is to illustrate at least numerically, that while our assumptions are standard in the literature, they serve for simplifying the path to otherwise too technical derivations and provide results that are expected to hold in much more general settings.

Acknowledgments

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models. Springer Series in Statistics*. Springer.

Chatterjee, D., Maitra, T., and Bhattacharya, S. (2018). A short note on almost sure convergence of Bayes factors in the general set-up. *The American Statistician*, pages 1–4.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*, volume 330. Cambridge University Press, Cambridge.

Crouse, M., Nowak, R., and Baraniuk, R. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing*, 46(4):886–902.

Csiszár, I. and Shields, P. (2000). The consistency of the BIC Markov order estimator. *The Annals of Statistics*, 28(6):1601–1619.

Del Moral, P. (2004). *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer.

Del Moral, P. (2013). *Mean Field Simulation for Monte Carlo Integration*. Chapman and Hall/CRC.

Diel, R., Le Corff, S., and Lerasle, M. (2020). Learning the distribution of latent variables in paired comparison models with round-robin scheduling. *arXiv preprint arXiv:1707.0136*.

Ding, J., Tarokh, V., and Yang, Y. (2017). Bridging AIC and BIC: a new criterion for autoregression. *IEEE Transactions on Information Theory*, 64(6):4024–4043.

Ding, J., Tarokh, V., and Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6):16–34.

Douc, R. and Moulines, E. (2012). Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models. *The Annals of Statistics*, 40(5):2697–2732.

Douc, R., Moulines, E., Olsson, J., and Van Handel, R. (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 39(1):474–513.

Douc, R., Moulines, E., and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of Statistics*, 32(5):2254–2304.

Douc, R., Moulines, E., and Stoffer, D. (2014). *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. CRC Press.

Douc, R., Olsson, J., and Roueff, F. (2020). Posterior consistency for partially observed Markov models. *Stochastic Processes and their Applications*, 130(2):733–759.

Douc, R., Roueff, F., and Sim, T. (2016). The maximizing set of the asymptotic normalized log-likelihood for partially observed Markov chains. *The Annals of Applied Probability*, 26(4):2357–2383.

Gales, M. and Young, S. (2008). The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304.

Gassiat, E. and Boucheron, S. (2003). Optimal error exponents in hidden Markov models order estimation. *IEEE Transactions on Information Theory*, 49(4):964–980.

Green, P. and Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97(460):1055–1070.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and its Application*. Academic Press.

Hurvich, C. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

Hurvich, C. and Tsai, C.-L. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44(3):214–217.

Ibragimov, R. and Sharakhmetov, S. (1999). Analogues of Khintchine, Marcinkiewicz–Zygmund and Rosenthal inequalities for symmetric statistics. *Scandinavian Journal of Statistics*, 26(4):621–633.

Jeffreys, H. (1998). *The Theory of Probability*. OUP Oxford.

Kass, R. and Raftery, A. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kass, R. E., Tierney, L., and Kadane, J. B. (1990). The validity of posterior expansions based on Laplace's method. In *In Bayesian and Likelihood Methods in Statistics and Econometrics, edited by S. Geisser, J. S. Hodges, S. J. Press and A. Zellner*, pages 473–488. North-Holland Amsterdam.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, 83(4):875–890.

Lang, S. (2012). *Real and functional analysis*, volume 142. Springer Science & Business Media.

Le Corff, S. and Fort, G. (2013). Online expectation maximization based algorithms for inference in hidden Markov models. *Electronic Journal of Statistics*, 7:763–792.

Le Gland, F. and Mevel, L. (1997). Asymptotic behaviour of the MLE in hidden Markov models. In *Proceedings of the 4th European Control Conference, Bruxelles 1997*.

Lehéricy, L. (2018). Nonasymptotic control of the MLE for misspecified non-parametric hidden Markov models. *arXiv preprint arXiv:1807.03997*.

Mamon, R. and Elliott, R. (2007). *Hidden Markov Models in Finance*, volume 460. Springer.

Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis*, 27(2):392–403.

Olsson, J. and Alenlöv, J. W. (2017). Particle-based online estimation of tangent filters with application to parameter estimation in nonlinear state-space models. *Annals of the Institute of Statistical Mathematics*, pages 1–32.

Pitt, M., Malik, S., and Doucet, A. (2014). Simulated likelihood inference for stochastic volatility models using continuous particle filtering. *Annals of the Institute of Statistical Mathematics*, 66(3):527–552.

Pouzo, D., Psaradakis, Z., and Sola, M. (2016). Maximum likelihood estimation in possibly misspecified dynamic models with time inhomogeneous Markov regimes. *arxiv preprint arXiv:1612.04932*.

Poyiadjis, G., Doucet, A., and Singh, S. (2011). Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65–80.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, pages 221–242.

Shao, S., Jacob, P., Ding, J., and Tarokh, V. (2017). Bayesian model comparison with the Hyvärinen score: Computation and consistency. *arXiv preprint arXiv:1711.00136*.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 63(1):117–126.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The annals of statistics*, pages 147–164.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, 68(1):45–54.

Sin, C.-Y. and White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics*, 71(1-2):207–225.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47.

Stout, W. (1970). The Hartman-Wintner law of the iterated logarithm for martingales. *The Annals of Mathematical Statistics*, 41(6):2158–2160.

Tadic, V. Z. and Doucet, A. (2018). Asymptotic properties of recursive maximum likelihood estimation in non-linear state-space models. *arXiv preprint arXiv:1806.09571*.

Takeuchi, K. (1976). Distribution of information statistics and validity criteria of models. *Mathematical Science*, 153:12–18.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

Yoon, B.-J. (2009). Hidden Markov models and their applications in biological sequence analysis. *Current Genomics*, 10(6):402–415.