Supplement for *Bayesian model comparison with the Hyvärinen score: computation and consistency*

Stephane Shao*, Pierre E. Jacob*, Jie Ding†, Vahid Tarokh‡

Contents

S1 Implementation of sequential Monte Carlo methods **2**

S2 Numerical illustration of robustness with Normal models **3**

S3 Numerical illustration of consistency with ARMA models **3**

S4 Numerical experiments: posterior density plots **5**
S4.1 Diffusion models for population dynamics of red kangaroos . . . 5
S4.2 Lévy-driven stochastic volatility models . . . 6

S5 H-score for discrete observations **7**

S6 Identities for the H-score **9**
S6.1 Proof of (4) . . . 9
S6.2 Proof of (11) and (12) . . . 11

S7 Consistency of the H-score **12**
S7.1 Proofs of Theorem 1 and 2 . . . 12
S7.1.1 Models for i.i.d. data . . . 13
S7.1.2 State-space models . . . 14
S7.2 Sufficient conditions and intermediary results . . . 15
S7.2.1 Assumption A3: Concentration of the posterior distribution . . . 15
S7.2.2 Assumption A4: Convergence of specific posterior moments . . . 16
S7.2.3 Assumption A9: Forgetting property of the H-score . . . 17
S7.2.4 Assumption A11: H-score of conditional density given the infinite past . . . 18
S7.3 Proofs of intermediary results . . . 19
S7.3.1 Proof of (s12) . . . 19
S7.3.2 Proof of (s13) . . . 20
S7.3.3 Proof of Lemma 1 . . . 20
S7.3.4 Proof of Lemma 2 . . . 21
S7.3.5 Proof of Lemma 3 . . . 22
S7.4 Heuristic proof for well-specified nested i.i.d. models . . . 26
S7.5 Numerical illustration with nested i.i.d. Normal models . . . 30

*Department of Statistics, Harvard University. Emails: stephaneshao@g.harvard.edu, pjacob@fas.harvard.edu.
†School of Statistics, University of Minnesota. Email: jie.ding959@duke.edu.
‡Department of Electrical and Computer Engineering, Duke University. Email: vahid.tarokh@duke.edu

S1 Implementation of sequential Monte Carlo methods

At each step $t \in \{0, ..., T\}$, an SMC sampler produces a set of N_θ particles $\theta_t^{(1:N_\theta)} = (\theta_t^{(1)}, ..., \theta_t^{(N_\theta)})$ with associated normalized weights $W_t^{(1:N_\theta)} = (W_t^{(1)}, ..., W_t^{(N_\theta)})$, targeting the posterior distribution $p(d\theta|y_{1:t})$ (Chopin, 2002; Del Moral, Doucet and Jasra, 2006). These particles are initialized as i.i.d. draws $\theta_0^{(1:N_\theta)}$ from a proposal distribution $q(d\theta)$. When the prior $p(d\theta)$ is proper and can be sampled from, one can chose $q(d\theta) = p(d\theta)$. Otherwise, $q(d\theta)$ should be chosen as an approximation of the first posterior distribution $p(d\theta|y_{1:\tau})$ that is proper. Going from an approximate sample of the posterior $p(d\theta|y_{1:t})$ to the next posterior $p(d\theta|y_{1:t+1})$ is achieved by successively targeting the intermediate bridging distributions with densities $p_{\gamma_{t,j}}(\theta) \propto p(\theta|y_{1:t})p(y_{t+1}|y_{1:t}, \theta)^{\gamma_{t,j}}$, where the $\gamma_{t,j}$'s are well-chosen *temperatures* satisfying $0 = \gamma_{t,0} < \gamma_{t,1} < ... < \gamma_{t,J_t} = 1$, with $J_t \in \mathbb{N}^*$. Let N denote the minimum effective sample size, fixed at some desired value (e.g. $N = N_\theta/2$ in all our numerical experiments). Given the current temperature $\gamma_{t,j} < 1$ and current particles $\theta_{t,j}^{(1:N_\theta)}$ with weights $W_{t,j}^{(1:N_\theta)}$ targeting $p_{\gamma_{t,j}}(\theta)$, the next temperature is determined adaptively and the particles are moved as follows:

1. For each $m \in \{1, ..., N_\theta\}$, compute $G_t^{(m)} = p(y_{t+1}|y_{1:t}, \theta_{t,j}^{(m)})$.
2. Find the largest $\gamma_{t,j+1} \in (\gamma_{t,j}, 1]$ such that $\text{ESS}\left(\gamma_{t,j}\,, \gamma_{t,j+1}\,, G_t^{(1:N_\theta)}, \theta_{t,j}^{(1:N_\theta)}\right) \geq N$, where

$$\text{ESS}\left(\gamma_{t,j}\,, \gamma_{t,j+1}\,, G_t^{(1:N_\theta)}, \theta_{t,j}^{(1:N_\theta)}\right) = \frac{\left(\sum_{m=1}^{N_\theta} W_{t,j}^{(m)} \left(G_t^{(m)}\right)^{\gamma_{t,j+1}-\gamma_{t,j}}\right)^2}{\sum_{m=1}^{N_\theta} \left(W_{t,j}^{(m)} \left(G_t^{(m)}\right)^{\gamma_{t,j+1}-\gamma_{t,j}}\right)^2}\,.$$

3. For each $m \in \{1, ..., N_\theta\}$, compute $w_{t,j+1}^{(m)} = W_{t,j}^{(m)}(G_t^{(m)})^{\gamma_{t,j+1}-\gamma_{t,j}}$ and set $W_{t,j+1}^{(m)} = w_{t,j+1}^{(m)}/\sum_{i=1}^{N_\theta} w_{t,j+1}^{(i)}$.
4. If $\gamma_{t,j+1} = 1$ and $t < T$, then set $\gamma_{t+1,0} = 0$, $\theta_{t+1}^{(m)} = \theta_{t+1,0}^{(m)} = \theta_{t,j}^{(m)}$, and $W_{t+1}^{(m)} = W_{t+1,0}^{(m)} = W_{t,j+1}^{(m)}$, for each $m \in \{1, ..., N_\theta\}$. If $\gamma_{t,j+1} < 1$, then resample the particles according to the weights $W_{t,j+1}^{(1:N_\theta)}$, reset the weights to $W_{t,j+1}^{(m)} = 1/N_\theta$ for each $m \in \{1, ..., N_\theta\}$, and move the particles by sampling independently $\theta_{t,j+1}^{(m)} \sim K(\theta_{t,j}^{(m)}, d\theta)$ for each $m \in \{1, ..., N_\theta\}$, where K is a Markov kernel that leaves $p_{\gamma_{t,j+1}}$ invariant. More generally, these moves can be performed as many times as desired in order to improve the rejuvenation rate of the particles, albeit at the expense of additional computations.

Starting from $\text{ESS}(\gamma_{0,0}\,, \gamma_{0,0}\,, G_0^{(1:N_\theta)}, \theta_{0,j}^{(1:N_\theta)}) = N_\theta$, we have $\text{ESS}(\gamma_{t,j}\,, \gamma_{t,j}\,, G_t^{(1:N_\theta)}, \theta_{t,j}^{(1:N_\theta)}) \geq N$ at any given time, by construction. Besides, $\gamma \mapsto \text{ESS}(\gamma_{t,j}\,, \gamma\,, G_t^{(1:N_\theta)}, \theta_{t,j}^{(1:N_\theta)})$ is a continuous function of γ. These two facts guarantee the existence and uniqueness of the $\gamma_{t,j+1}$ defined in step 2. Other diagnostics than the effective sample size could be used to monitor the degeneracy of the weights. In step 4, we choose to perform the resampling using SSP resampling (Gerber, Chopin and Whiteley, 2017). Regarding the Markov kernel K, we use a Metropolis-Hastings kernel, with a mixture of Normals as a proposal. This mixture of Normals is fitted to the latest set of weighted particles, using five components by default throughout our numerical experiments.

For general state-space models, the SMC^2 algorithm attaches one particle filter made of N_x particles $x_t^{(1:N_x)}$ to each of the particles $\theta_t^{(m)}$. These particle filters produce estimators $\hat{p}(y_{t+1}|y_{1:t}, \theta_{t,j}^{(m)})$, which are used in place of the intractable incremental likelihoods $p(y_{t+1}|y_{1:t}, \theta_{t,j}^{(m)})$ in step 1. Our implementation of SMC^2 starts with a

small initial number N_x, and adaptively doubles it whenever the acceptance rate of the moves in step 4 are below some desired threshold. Increasing N_x is achieved by using conditional sequential Monte Carlo steps, which sample new filters made of a larger number of particles, as detailed in Section 3.6.2 of Chopin, Jacob and Papaspiliopoulos (2013). We rely on the bootstrap particle filter for simplicity, but more efficient filters — such as the auxiliary particle filter (Pitt and Shephard, 1999) — could be used within SMC^2, as illustrated in Golightly and Kypraios (2017). Other relevant considerations are discussed in Chopin, Ridgway, Gerber and Papaspiliopoulos (2015); Duan and Fulop (2015).

S2 Numerical illustration of robustness with Normal models

This section complements Section 2.3 of the main paper. We consider the two Normal models

$$M_1: \quad Y_1, ..., Y_T \,|\, \theta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\theta_1, 1\right), \quad \theta_1 \sim \mathcal{N}\left(0, \sigma_0^2\right),$$
$$M_2: \quad Y_1, ..., Y_T \,|\, \theta_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \theta_2\right), \quad \theta_2 \sim \text{Inv-}\chi^2\left(\nu_0, s_0^2\right).$$

The positive hyperparameters are chosen as $\nu_0 = 0.1$ and $s_0^2 = 1$. We compare M_1 and M_2 using observations generated as $Y_1, ..., Y_T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(1, 1)$, for different $\log(\sigma_0) \in \{0, 150, 350\}$, i.e. for increasingly vague priors on μ. In this setting, M_1 is well-specified whereas M_2 is misspecified.

Under M_1, we have $Y_t \,|\, Y_{1:t-1} \sim \mathcal{N}\left(\mu_{t-1}, \sigma_{t-1}^2 + 1\right)$ for all $t \in \{0, ..., T\}$ by conjugacy, with $\sigma_t^2 = (t + \sigma_0^{-2})^{-1}$ and $\mu_t = \sigma_t^2 \sum_{i=1}^t Y_i$ for all $t \in \{1, ..., T\}$. Under M_2, we have $Y_t \,|\, Y_{1:t-1} \sim t_{\nu_{t-1}}\left(0, s_{t-1}^2\right)$ for all $t \in \{0, ..., T\}$ by conjugacy, with $\nu_t = \nu_0 + t$ and $s_t = (\nu_0 s_0^2 + \sum_{i=1}^t Y_i^2)/(\nu_0 + t)$ for all $t \in \{1, ..., T\}$, where $t_\nu\left(0, s^2\right)$ denotes a (centered) scaled Student's t-distribution with density $x \to \Gamma(\nu/2)^{-1}(\nu\pi s^2)^{-1/2}\Gamma((\nu+1)/2)(1 + x^2/(\nu s^2))^{-(\nu+1)/2}$ for $x \in \mathbb{R}$. Given $Y_{1:T}$, these conjugacy results allow us to compute the log-Bayes factor and H-factor analytically.

We generate 100 independent samples, each consisting of $T = 1000$ i.i.d. draws from $\mathcal{N}(1, 1)$. For each sample $Y_{1:T}$ and every $\log(\sigma_0) \in \{0, 150, 350\}$, we compute the log-Bayes factor and H-score of M_1 against M_2. The results are shown in Figure 1. For any fixed value of σ_0, both the H-factor and log-Bayes factor are consistent, in the sense of asymptotically choosing the correct model M_1, as $T \to +\infty$. However, for any fixed T, no matter how large, there always exists a large enough σ_0 such that the log-Bayes factor chooses the wrong model M_2 with arbitrarily high probability. This is because $\log p_{M_1}(Y_{1:T})$ behaves equivalently to $-\log(\sigma_0)$ as $\sigma_0 \to +\infty$. This sensitivity of the log-Bayes factor to the vagueness of priors is arguably undesirable, as it can lead to choosing a misspecified model over a well-specified one. This artifact is even more unsettling when considering that, with $T = 1000$ observations, all three prior specifications $\log(\sigma_0) \in \{0, 150, 350\}$ essentially lead to the same posterior on μ, i.e. the same model fit in some sense. By contrast, the H-factor is virtually unchanged when increasing σ_0 beyond a certain value.

S3 Numerical illustration of consistency with ARMA models

Define the stationarity triangle $\mathbb{S} = \{(\phi_1, \phi_2) \in \mathbb{R}^2 : |\phi_2| < 1, \phi_2 - \phi_1 < 1, \phi_2 + \phi_1 < 1\}$. Let Unif($\mathbb{S}$) denote the bivariate uniform distribution on the set $\mathbb{S}$ and let $(\varepsilon_t)_{t \in \mathbb{N}}$ denote a sequence of i.i.d. standard Normal variables. We consider the following time series models, corresponding respectively to AR(1), AR(2), and MA(1) models.

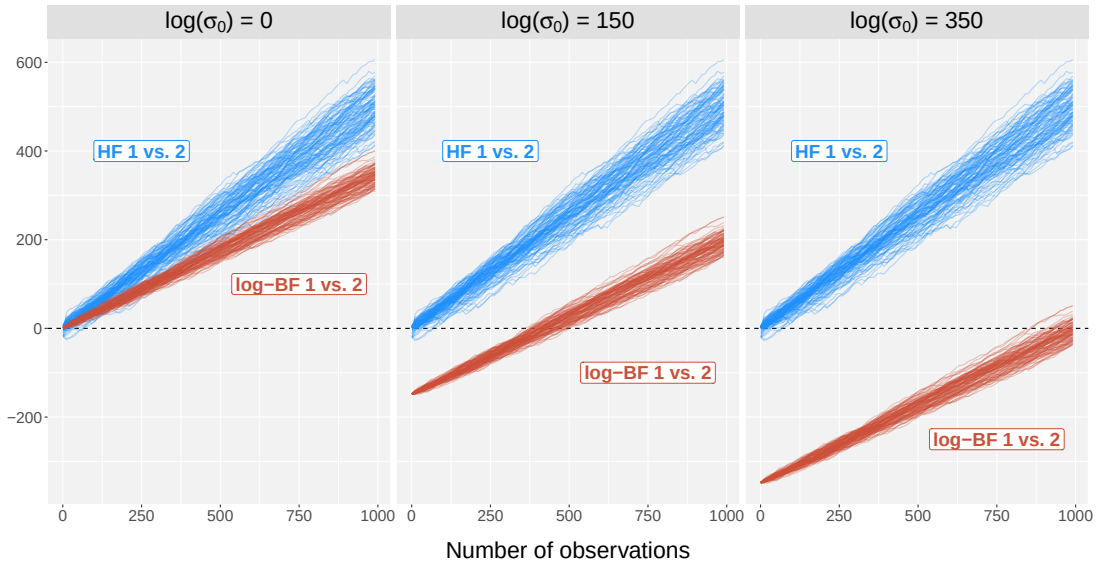


Figure 1. *Exact log-Bayes factors (log-BF) and H-factors (HF) of M_1 against M_2, computed for 100 independent samples (thin solid lines) generated as i.i.d. $\mathcal{N}(1,1)$, under three increasingly vague priors on μ:* $\log(\sigma_0) = 0$ *(left panel),* $\log(\sigma_0) = 150$ *(middle panel),* $\log(\sigma_0) = 350$ *(right panel). See Section S2.*

M_1: $Y_1 \,|\, \phi, \sigma^2 \sim \mathcal{N}\left(0, \sigma^2/(1-\phi^2)\right)$; $Y_t = \phi Y_{t-1} + \sigma\varepsilon_t$ *for all* $t \geq 2$;
with independent priors $\phi \sim \text{Unif}(-1,1)$ *and* $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, s_0^2)$.

M_2: $Y_1, Y_2 \,|\, \phi_1, \phi_2, \sigma^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{(1-\phi_2)\sigma^2/(1+\phi_2)}{(1-\phi_2-\phi_1)(1-\phi_2+\phi_1)}\right)$; $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \sigma\varepsilon_t$ *for all* $t \geq 3$;
with independent priors $(\phi_1, \phi_2) \sim \text{Unif}(\mathbb{S})$ *and* $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, s_0^2)$.

M_3: $Y_t = \sigma\left(\varepsilon_t + \theta\varepsilon_{t-1}\right)$ *for all* $t \geq 1$;
with independent priors $\theta \sim \text{Unif}(-1,1)$ *and* $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, s_0^2)$.

The positive hyperparameters are set to $\nu_0 = 1$ and $s_0^2 = 1$. First, we consider a non-nested setting by comparing M_1 and M_3 under the following two data-generating processes:

(1) AR(1) with $Y_1 \sim \mathcal{N}(0,1)$ and $Y_t = 0.6\, Y_{t-1} + 0.8\, \varepsilon_t$, i.e. M_1 is well-specified while M_3 is not,

(2) MA(1) with $Y_t = \varepsilon_t + 0.5\, \varepsilon_{t-1}$, i.e. M_3 is well-specified while M_1 is not.

ARMA models can be regarded as particular cases of linear Gaussian state-space models, whose likelihood can be computed using Kalman filters. Thus, H-scores of ARMA models can be estimated by directly using SMC in conjunction with Kalman filters, instead of more sophisticated SMC^2 algorithms. For each data-generating process, we generate $T = 1000$ observations and estimate the H-score of M_1 and M_3 via SMC with $N_\theta = 1024$ particles. The estimated H-factors and log-Bayes factors of M_1 against M_3 are shown in Figure 2. We see that the H-factor asymptotically chooses the correct model.

We now consider a nested setting by comparing M_1 and M_2 under the following two data-generating processes:

(3) AR(1) with $Y_1 \sim \mathcal{N}(0,1)$ and $Y_t = 0.6\, Y_{t-1} + 0.8\, \varepsilon_t$, i.e. both M_1 and M_2 are well-specified,

(4) AR(2) with $Y_1, Y_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $Y_t = 0.25\, Y_{t-1} + 0.5\, Y_{t-2} + 0.75\, \varepsilon_t$, i.e. M_2 is well-specified but M_1 is not.

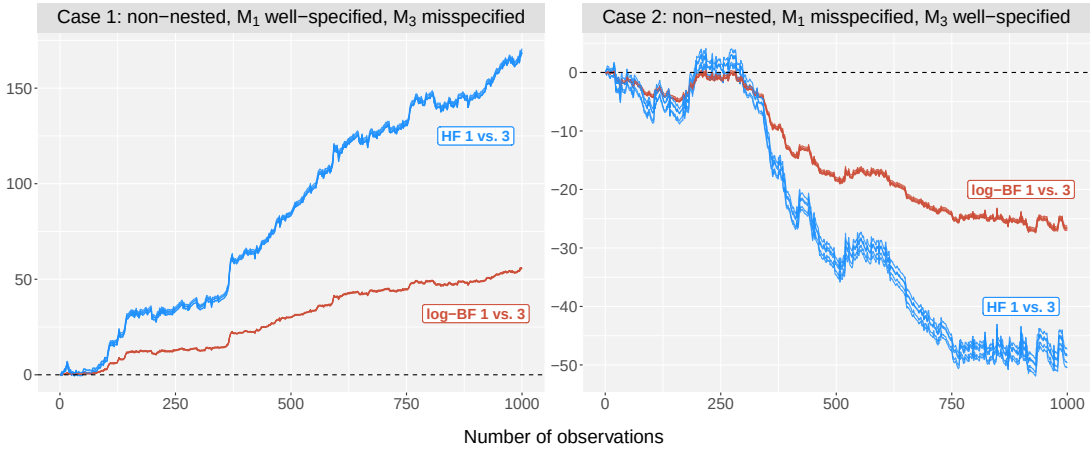


Figure 2. *Estimated log-Bayes factors (log-BF) and H-factors (HF) of M_1 against M_3, computed for 5 replications (thin solid lines), under two data-generating processes: AR(1) (Case 1) and MA(1) (Case 2). See Section S3.*

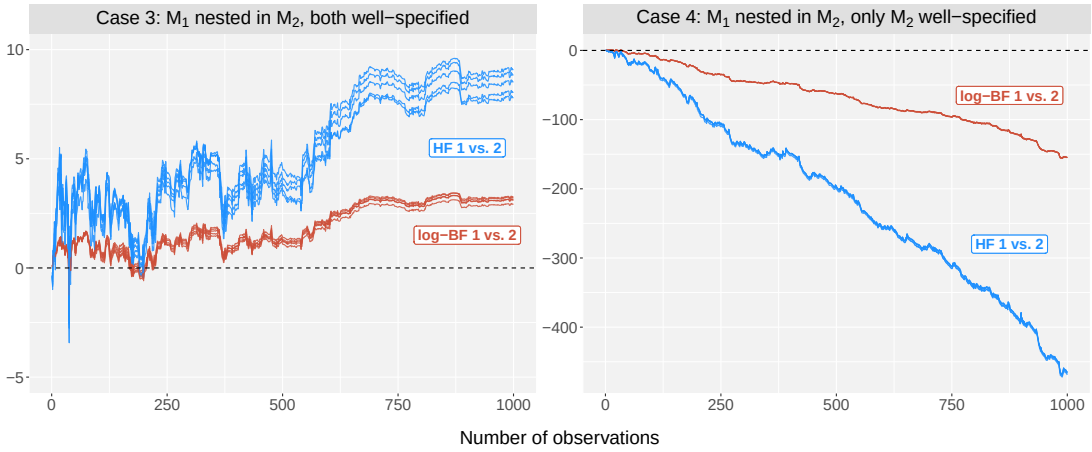


Figure 3. *Estimated log-Bayes factors (log-BF) and H-factors (HF) of M_1 against M_2, computed for 5 replications (thin solid lines), under two data-generating processes: AR(1) (Case 3) and AR(2) (Case 4). See Section S3.*

The data-generating processes are initialized at their respective stationary distributions. For each case, we generate $T = 1000$ observations and estimate the H-score of M_1 and M_2 via SMC with $N_\theta = 1024$ particles. The respective H-factors and log-Bayes factors of M_1 against M_2 are shown in Figure 3. Case 3 suggests that, when dealing with nested well-specified models, the H-factor asymptotically favors the model of smallest dimension.

S4 Numerical experiments: posterior density plots

S4.1 Diffusion models for population dynamics of red kangaroos

This section complements the numerical example presented in Section 4.2 of the main paper. For each population model M_1, M_2, and M_3, the respective posteriors of the parameters are estimated via SMC^2 across 5 replications. The marginal posterior densities are shown in Figures 4, 5, and 6. These estimated posterior densities should be contrasted with the vague independent priors σ, τ, $b \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 10)$, and $r \sim \text{Unif}(-10, 10)$. The plots suggest that posterior concentration may be a reasonable assumption, even when the conditions of Section S7.2 are not met.

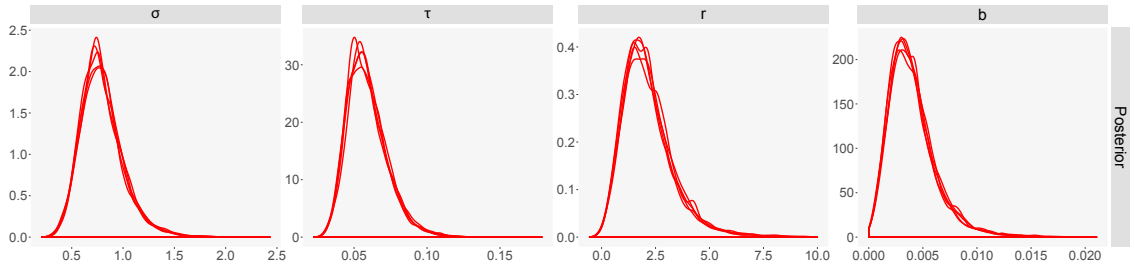


Figure 4. *Estimated marginal posterior densities of* (σ, τ, r, b) *under model* M_1*, given 41 observations, with independent priors* σ, τ, $b \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 10)$ *and* $r \sim \text{Unif}(-10, 10)$*, plotted for 5 replications (solid lines). See Section S4.1.*

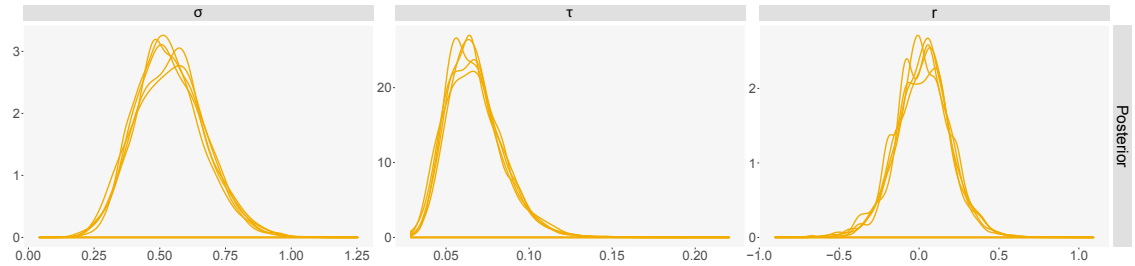


Figure 5. *Estimated marginal posterior densities of* (σ, τ, r) *under model* M_2*, given 41 observations, with independent priors* σ, $\tau \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 10)$ *and* $r \sim \text{Unif}(-10, 10)$*, plotted for 5 replications (solid lines). See Section S4.1.*

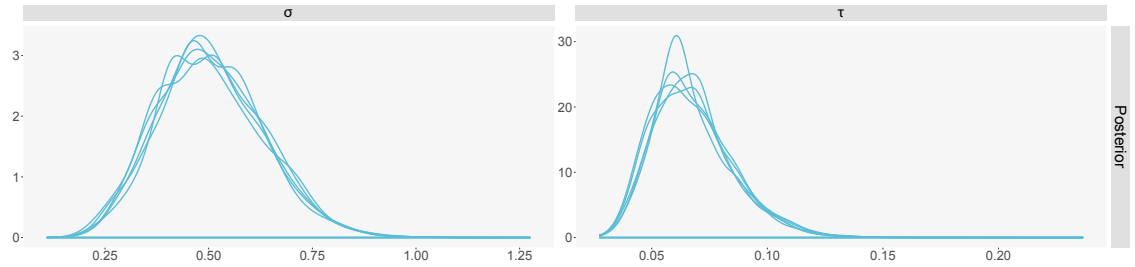


Figure 6. *Estimated marginal posterior densities of* (σ, τ) *under model* M_3*, given 41 observations, with independent priors* σ, $\tau \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 10)$*, plotted for 5 replications (solid lines). See Section S4.1.*

S4.2 Lévy-driven stochastic volatility models

This section complements the numerical example presented in Section 3.3 of the main paper. For each Lévy-driven stochastic volatility model M_1 and M_2, the respective posterior densities of the parameters are estimated via SMC^2 across 5 replications. The estimated marginal posterior densities are shown in Figures 7 and 8, along with the corresponding marginal prior densities. For comparability, the respective marginal prior densities are plotted over the same support as their corresponding marginal posterior densities, albeit with different scales on the y-axis for better readability. Similarly to the previous example, posterior concentration seems to be a reasonable assumption. The exception is on λ_2 under model M_2, whose posterior after 1000 observations resembles the prior. This can be explained by the posterior of w concentrating near 1 as the data are generated from M_1, thus making the second factor irrelevant in model M_2. The parameter λ_2 associated with the second factor is then not identified.

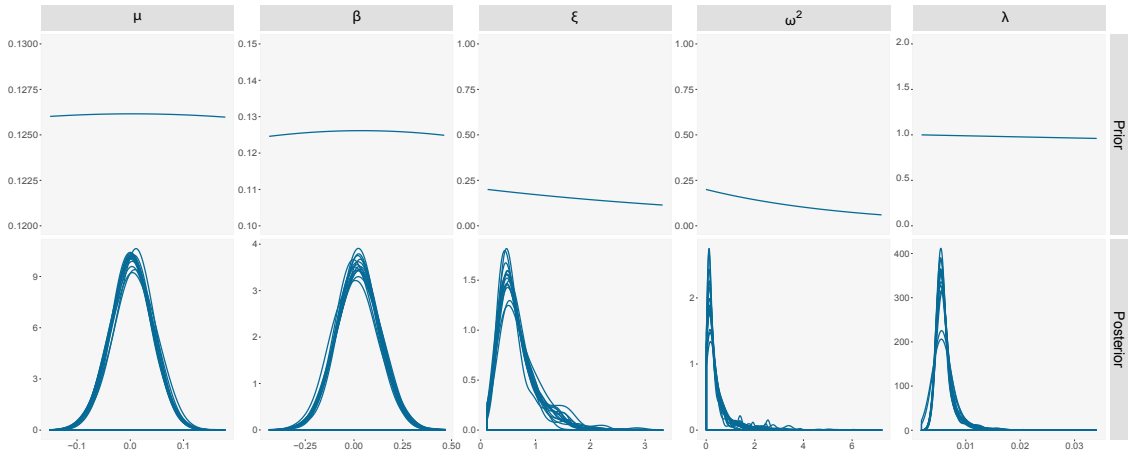


Figure 7. *Top panels: marginal prior densities* μ, $\beta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 10)$; ξ, $\omega^2 \overset{\text{i.i.d.}}{\sim} \text{Exp}(1/5)$; $\lambda \sim \text{Exp}(1)$, *plotted over the support of the posterior. Bottom panels: estimated marginal posterior densities under model* M_1, *given 1000 observations, plotted for 5 replications (solid lines). See Section S4.2.*

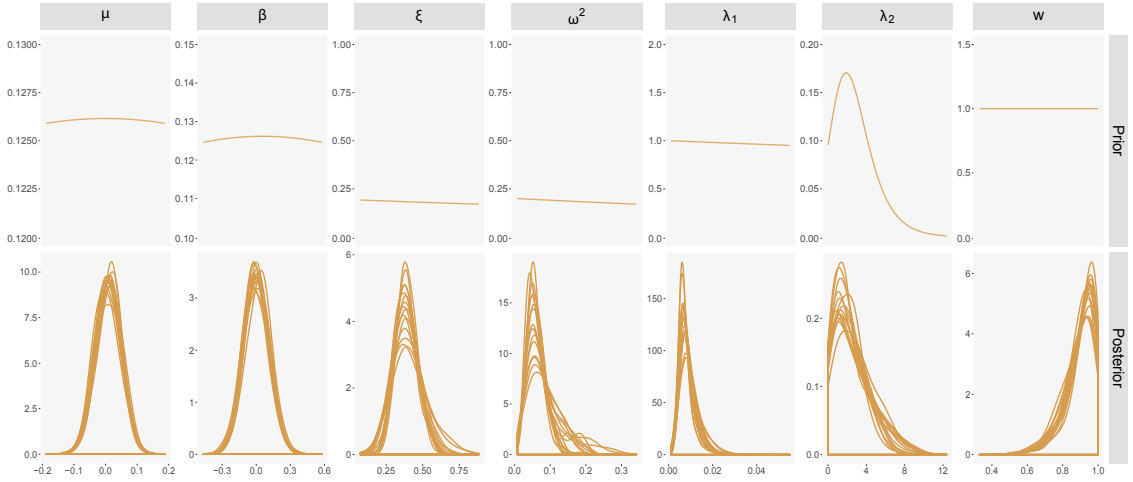


Figure 8. *Top panels: marginal prior densities* μ, $\beta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 10)$; ξ, $\omega^2 \overset{\text{i.i.d.}}{\sim} \text{Exp}(1/5)$; $\lambda_1 \sim \text{Exp}(1)$; $\lambda_2 - \lambda_1 \sim \text{Exp}(1/2)$; $\text{w} \sim \text{Unif}(0, 1)$, *plotted over the support of the posterior. Bottom panels: estimated marginal posterior densities under model* M_2, *given 1000 observations, plotted for 5 replications (solid lines). See Section S4.2.*

S5 H-score for discrete observations

0-homogeneous score functions for discrete observations are proper if and only if they are super-gradients of 1-homogeneous concave entropy functions (McCarthy, 1956; Hendrickson and Buehler, 1971). It follows that we can construct a proper 0-homogeneous scoring rule in terms of a collection of homogeneous functions over the cliques of an undirected graph on the space $\mathbb{Y} = [\![a_1, b_1]\!] \times ... \times [\![a_{d_y}, b_{d_y}]\!]$ (Dawid, Lauritzen and Parry, 2012). More precisely, let $\mathcal{G}$ denote an undirected graph with a set of nodes equal to $\mathbb{Y}$ and a set of edges defined as $\{(y_1, y_2) \in \mathbb{Y}^2 : y_1 - y_2 \in \{-2e_k, -e_k, e_k, 2e_k\} \text{ for some } k \in [\![1, d_y]\!]\}$. Here e_k denotes the canonical vector of $\mathbb{Z}^{d_y}$ that has all coordinates equal to 0 except for its k-th coordinate that equals 1. The cliques (maximal complete subsets) of this graph are of the form $\{y - e_k, y, y + e_k\}$. Define the function $H : (0, \infty)^3 \to \mathbb{R}$ as $H(p_1, p_2, p_3) = -(p_3 - p_1)^2/p_2$. This function is 1-homogeneous and concave. Indeed, for any $\lambda > 0$, we have $H(\lambda p_1, \lambda p_2, \lambda p_3) = \lambda H(p_1, p_2, p_3)$.

Besides, the Hessian of H at any $(p_1, p_2, p_3) \in (0, \infty)^3$ is given by

$$\begin{pmatrix} -\frac{2(p_3-p_1)^2}{p_2^3} & \frac{2(p_3-p_1)}{p_2^2} & -\frac{2(p_3-p_1)}{p_2^2} \\ \frac{2(p_3-p_1)}{p_2^2} & -\frac{2}{p_2} & \frac{2}{p_2} \\ -\frac{2(p_3-p_1)}{p_2^2} & \frac{2}{p_2} & -\frac{2}{p_2} \end{pmatrix}.$$

For all $(p_1, p_2, p_3) \in (0, \infty)^3$, the determinants of the extracted matrices

$$\left(-\frac{2(p_3-p_1)^2}{p_2^3}\right), \left(-\frac{2}{p_2}\right), \begin{pmatrix} -\frac{2(p_3-p_1)^2}{p_2^3} & \frac{2(p_3-p_1)}{p_2^2} \\ \frac{2(p_3-p_1)}{p_2^2} & -\frac{2}{p_2} \end{pmatrix}, \begin{pmatrix} -\frac{2}{p_2} & \frac{2}{p_2} \\ \frac{2}{p_2} & -\frac{2}{p_2} \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} -\frac{2(p_3-p_1)^2}{p_2^3} & -\frac{2(p_3-p_1)}{p_2^2} \\ -\frac{2(p_3-p_1)}{p_2^2} & -\frac{2}{p_2} \end{pmatrix}$$

are respectively negative, negative, 0, 0, and 0. The determinant of the Hessian is also equal to 0. In other words, all the principal minors of the negative Hessian are non-negative. By Sylvester's criterion (Horn and Johnson, 1985), this implies that the negative Hessian of H at (p_1, p_2, p_3) is positive semi-definite, for all $(p_1, p_2, p_3) \in (0, \infty)^3$, which proves that the function H is concave.

Following the construction from Section 3.3 of Dawid et al. (2012), we can define, for all probability mass functions p on $\mathbb{Y}$, the concave entropy function

$$\mathcal{E}_{\mathcal{H}^D}(p) = -\sum_{k=1}^{d_y} \sum_{\substack{y \in \mathbb{Y} \text{ s.t.} \\ a_k < y_{(k)} < b_k}} p(y) \left(\frac{p(y+e_k) - p(y-e_k)}{2\,p(y)}\right)^2, \tag{s1}$$

whose associated score function is given by

$$\mathcal{H}^D(y, p) = \sum_{k=1}^{d_y} \mathcal{H}_k^D(y, p),$$

where

$$\mathcal{H}_k^D(y,p) = \begin{cases} \frac{p(y+2\,e_k)-p(y)}{2\,p(y+e_k)} & \text{if } y_{(k)} = a_k, \\ \frac{p(y+2\,e_k)-p(y)}{2\,p(y+e_k)} + \left(\frac{p(y+e_k)-p(y-e_k)}{2\,p(y)}\right)^2 & \text{if } y_{(k)} = a_k + 1, \\ \frac{p(y+2\,e_k)-p(y)}{2\,p(y+e_k)} - \frac{p(y)-p(y-2\,e_k)}{2\,p(y-e_k)} + \left(\frac{p(y+e_k)-p(y-e_k)}{2\,p(y)}\right)^2 & \text{if } a_k + 1 < y_{(k)} < b_k - 1, \\ -\frac{p(y)-p(y-2\,e_k)}{2\,p(y-e_k)} + \left(\frac{p(y+e_k)-p(y-e_k)}{2\,p(y)}\right)^2 & \text{if } y_{(k)} = b_k - 1, \\ -\frac{p(y)-p(y-2\,e_k)}{2\,p(y-e_k)} & \text{if } y_{(k)} = b_k. \end{cases}$$

The concavity of the entropy function guarantees that $\mathcal{H}^D$ is a proper scoring rule. The entropy in (s1) can be interpreted as a discrete analog of the entropy function of the H-score for continuous observations, which is given by $-\int_{\mathbb{Y}} \|\nabla_y \log p(y)\|^2 \, p(y) dy$ under mild regularity assumptions (Hyvärinen, 2005; Dawid and Musio, 2015).

The alternative definition using forward differences, given by

$$\begin{cases} 2\left(\frac{p(y+e_k)-p(y)}{p(y)}\right) + \left(\frac{p(y+e_k)-p(y)}{p(y)}\right)^2 & \text{if } y_{(k)} = a_k, \\ 2\left(\frac{p(y+e_k)-p(y)}{p(y)} - \frac{p(y)-p(y-e_k)}{p(y-e_k)}\right) + \left(\frac{p(y+e_k)-p(y)}{p(y)}\right)^2 & \text{if } a_k < y_{(k)} < b_k, \\ -2\left(\frac{p(y)-p(y-e_k)}{p(y-e_k)}\right) & \text{if } y_{(k)} = b_k, \end{cases}$$

is a particular case of the pair scoring rule from Example 4.1 in Dawid et al. (2012), where we choose the concave function G to be $u \mapsto -(u-1)^2$.

S6 Identities for the H-score

In this section, we fix a model M and drop the dependence on the model in the notation. Equations (4), (11), and (12) result from algebraic manipulations, under assumptions guaranteeing the existence of all the relevant derivatives and integrals, as well as enabling differentiation under the integral sign. Such assumptions can be stated as follows.

Assumption A1. *For all $t \in \mathbb{N}^*$, the following conditions hold:*

(a) For all $y_{1:t} \in \mathbb{Y}^t$, the function $\theta \mapsto p(y_t|y_{1:t-1},\theta)\, p(\theta|y_{1:t-1})$ is integrable on $\mathbb{T}$.

(b) For all $(y_{1:t-1},\theta) \in \mathbb{Y}^{t-1} \times \mathbb{T}$, the function $y_t \mapsto p(y_t|y_{1:t-1},\theta)$ is twice differentiable on $\mathbb{Y}$.

(c) For all $k \in \{1,...,d_y\}$, there exist integrable functions $h_{1,k,t}$ and $h_{2,k,t}$ such that, for all $(y_{1:t},\theta) \in \mathbb{Y}^t \times \mathbb{T}$,

$$\left|\frac{\partial p(y_t|y_{1:t-1},\theta)}{\partial y_{t_{(k)}}}\, p(\theta|y_{1:t-1})\right| \leq h_{1,k,t}(\theta) \qquad \text{and} \qquad \left|\frac{\partial^2 p(y_t|y_{1:t-1},\theta)}{\partial y_{t_{(k)}}^2}\, p(\theta|y_{1:t-1})\right| \leq h_{2,k,t}(\theta).$$

Assumption A2. *For all $t \in \mathbb{N}^*$, the following conditions hold:*

(a) For all $(y_{1:t},\theta) \in \mathbb{Y}^t \times \mathbb{T}$, the function $x_t \mapsto p(x_t|y_{1:t-1},\theta)\, g_\theta(y_t|x_t)$ is integrable on $\mathbb{X}$.

(b) For all $(\theta, x_t) \in \mathbb{T} \times \mathbb{X}$, the function $y_t \mapsto g_\theta(y_t|x_t)$ is twice differentiable on $\mathbb{Y}$.

(c) For all $k \in \{1,...,d_y\}$, there exist integrable functions $h_{3,k,t}$ and $h_{4,k,t}$ such that, for all $(y_{1:t},\theta,x_t) \in \mathbb{Y}^t \times \mathbb{T} \times \mathbb{X}$,

$$\left|\frac{\partial g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}}\, p(x_t|y_{1:t-1},\theta)\right| \leq h_{3,k,t}(x_t) \qquad \text{and} \qquad \left|\frac{\partial^2 g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}^2}\, p(x_t|y_{1:t-1},\theta)\right| \leq h_{4,k,t}(x_t).$$

S6.1 Proof of (4)

Consider some generic prior $p(\theta)$ and likelihood $p(y|\theta)$. Assume that $\theta \mapsto p(y|\theta)p(\theta)$ is integrable for every $y \in \mathbb{Y}$, $y \mapsto p(y|\theta)$ is twice differentiable on $\mathbb{Y}$ for every $\theta \in \mathbb{T}$, and, for all $k \in \{1,...,d_y\}$, both $\theta \mapsto \left|\frac{\partial\, p(y|\theta)}{\partial y_{(k)}}\, p(\theta)\right|$ and $\theta \mapsto \left|\frac{\partial^2\, p(y|\theta)}{\partial {y_{(k)}}^2}\, p(\theta)\right|$ are dominated by integrable functions on $\mathbb{T}$. Let $p(y) = \int_{\mathbb{T}} p(y|\theta)p(\theta)d\theta$. The previous

assumptions allow us to partially differentiate $y \mapsto p(y)$ twice under the integral sign with respect to each coordinate. Recall from (1) the definition of the H-score,

$$\mathcal{H}(y,p) = \sum_{k=1}^{d_y} \left(2\,\frac{\partial^2 \log p(y)}{\partial y_{(k)}^2} + \left(\frac{\partial \log p(y)}{\partial y_{(k)}} \right)^2 \right).$$

For all $k \in \{1, ..., d_y\}$, partial differentiation under the integral sign yields, on the one hand,

$$\frac{\partial \log p(y)}{\partial y_{(k)}} = \frac{1}{p(y)} \int \left(\frac{\partial\, p(y|\theta)}{\partial y_{(k)}} \right) p(\theta) d\theta = \int \left(\frac{\partial \log p(y|\theta)}{\partial y_{(k)}} \right) p(\theta|y) d\theta = \mathbb{E}\left[\frac{\partial \log p(y|\Theta)}{\partial y_{(k)}} \middle| y \right].$$

On the other hand, partially differentiating twice under the integral sign yields

$$\frac{\partial^2 \log p(y)}{\partial {y_{(k)}}^2} = -\left(\frac{\partial \log p(y)}{\partial y_{(k)}} \right)^2 + \frac{1}{p(y)} \frac{\partial^2 p(y)}{\partial {y_{(k)}}^2} = -\left(\frac{\partial \log p(y)}{\partial y_{(k)}} \right)^2 + \frac{1}{p(y)} \int \left(\frac{\partial^2 p(y|\theta)}{\partial {y_{(k)}}^2} \right) p(\theta) d\theta.$$

Regarding the integrand in the last term, we have

$$\frac{\partial^2 p(y|\theta)}{\partial {y_{(k)}}^2} = p(y|\theta) \left[\frac{\partial^2 \log p(y|\theta)}{\partial {y_{(k)}}^2} + \left(\frac{\partial \log p(y|\theta)}{\partial y_{(k)}} \right)^2 \right].$$

This leads to

$$\begin{aligned} \frac{\partial^2 \log p(y)}{\partial {y_{(k)}}^2} &= -\left(\frac{\partial \log p(y)}{\partial y_{(k)}} \right)^2 + \int p(\theta|y) \left[\frac{\partial^2 \log p(y|\theta)}{\partial {y_{(k)}}^2} + \left(\frac{\partial \log p(y|\theta)}{\partial y_{(k)}} \right)^2 \right] d\theta \\ &= -\left(\frac{\partial \log p(y)}{\partial y_{(k)}} \right)^2 + \mathbb{E}\left[\frac{\partial^2 \log p(y|\Theta)}{\partial {y_{(k)}}^2} + \left(\frac{\partial \log p(y|\Theta)}{\partial y_{(k)}} \right)^2 \middle| y \right]. \end{aligned}$$

By putting everything together we finally get

$$\mathcal{H}(y,p) = \sum_{k=1}^{d_y} \left(\mathbb{E}\left[2\,\frac{\partial^2 \log p(y|\Theta)}{\partial {y_{(k)}}^2} + 2 \left(\frac{\partial \log p(y|\Theta)}{\partial y_{(k)}} \right)^2 \middle| y \right] - \left(\mathbb{E}\left[\frac{\partial \log p(y|\Theta)}{\partial y_{(k)}} \middle| y \right] \right)^2 \right). \tag{s2}$$

For a given model M with parameter $\theta \in \mathbb{T}$, we have

$$p(y_t|y_{1:t-1}) = \int_{\mathbb{T}} p(y_t|y_{1:t-1}, \theta)\, p(\theta|y_{1:t-1}).\, d\theta \tag{s3}$$

Therefore, under Assumption A1, we can apply (s2) to (s3) for each term of the sum in (3) to get

$$\mathcal{H}_T(M) = \sum_{t=1}^{T} \sum_{k=1}^{d_y} \left(2\,\mathbb{E}\left[\frac{\partial^2 \log p(y_t|y_{1:t-1}, \Theta)}{\partial y_{t_{(k)}}^2} + \left(\frac{\partial \log p(y_t|y_{1:t-1}, \Theta)}{\partial y_{t_{(k)}}} \right)^2 \middle| y_{1:t} \right] - \left(\mathbb{E}\left[\frac{\partial \log p(y_t|y_{1:t-1}, \Theta)}{\partial y_{t_{(k)}}} \middle| y_{1:t} \right] \right)^2 \right),$$

which proves (4). □

S6.2 Proof of (11) and (12)

Under Assumption A2, we can partially differentiate under the integral sign, so that for all $k \in \{1, ..., d_y\}$, we have

$$\begin{aligned}
\frac{\partial \log p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}} &= \frac{1}{p(y_t|y_{1:t-1}, \theta)} \int p(x_t|y_{1:t-1}, \theta) \left(\frac{\partial g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}} \right) dx_t \\
&= \frac{1}{p(y_t|y_{1:t-1}, \theta)} \int p(x_t|y_{1:t-1}, \theta)\, g_\theta(y_t|x_t) \left(\frac{\partial \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}} \right) dx_t \\
&= \int \left(\frac{\partial \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}} \right) p(x_t|y_{1:t}, \theta)\, dx_t \,,
\end{aligned}$$

where the last equality comes from the fact that

$$\frac{p(x_t|y_{1:t-1}, \theta)\, g_\theta(y_t|x_t)}{p(y_t|y_{1:t-1}, \theta)} = \frac{p(x_t, y_t|y_{1:t-1}, \theta)}{p(y_t|y_{1:t-1}, \theta)} = p(x_t|y_{1:t}, \theta). \tag{s4}$$

This proves (11).

Regarding (12), we proceed similarly and have, for all $k \in \{1, ..., d_y\}$,

$$\frac{\partial^2 \log p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}^2} = -\left(\frac{\partial \log p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}} \right)^2 + \frac{1}{p(y_t|y_{1:t-1}, \theta)} \frac{\partial^2 p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}^2} \,. \tag{s5}$$

The second term can be rewritten as

$$\frac{1}{p(y_t|y_{1:t-1}, \theta)} \frac{\partial^2 p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}^2} = \frac{1}{p(y_t|y_{1:t-1}, \theta)} \int p(x_t|y_{1:t-1}, \theta) \left(\frac{\partial^2 g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}^2} \right) dx_t \,, \tag{s6}$$

where the integrand can be written as

$$\frac{\partial^2 g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}^2} = g_\theta(y_t|x_t) \left(\frac{\partial^2 \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}^2} + \left(\frac{\partial \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}} \right)^2 \right). \tag{s7}$$

By plugging (s7) into (s6) and using again (s4), we get

$$\frac{1}{p(y_t|y_{1:t-1}, \theta)} \frac{\partial^2 p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}^2} = \int p(x_t|y_{1:t}, \theta) \left(\frac{\partial^2 \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}^2} + \left(\frac{\partial \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}} \right)^2 \right) dx_t \,.$$

By plugging this back into (s5), we finally get

$$\frac{\partial^2 \log p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}^2} = -\left(\frac{\partial \log p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}} \right)^2 + \int \left(\frac{\partial^2 \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}^2} + \left(\frac{\partial \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}} \right)^2 \right) p(x_t|y_{1:t}, \theta)\, dx_t \,,$$

which proves (12). □

S7 Consistency of the H-score

Without much loss of generality, we prove the results in the case of continuous univariate observations ($d_y = 1$). Thanks to (3), the proofs can be generalized to multivariate observations by working on each dimension separately. Unless stated otherwise, we assume that A1 and A2 hold, so that we may use (13) and (5).

Section S7.1 should be read as a proof of concept: we prove Theorem 1 and 2 by using intermediary results as high-level assumptions (A3 to A12). This allows us to highlight the key steps of the proofs. In Section S7.2, we present explicit conditions (C1 to C6) that are sufficient for these assumptions to hold. Some of these conditions are strong, which enables intuitive proofs; our simulation studies suggest that the consistency of the H-score is likely to hold under weaker conditions. Detailed proofs of all the intermediary results are provided in Section S7.3. Proofs under weaker conditions or discrete observations are left for future work.

S7.1 Proofs of Theorem 1 and 2

The first ingredient is the $\mathbb{P}_\star$-almost sure concentration of the posterior distribution $p(d\theta|Y_{1:t})$ around some limit value $\theta^\star \in \mathbb{T}$, as the number of observations increases (Assumption A3).

Assumption A3. *$\mathbb{P}_\star$-a.s., there exists $\theta^\star \in \mathbb{T}$ such that, if $\Theta_t \sim p(d\theta|Y_{1:t})$ for all $t \in \mathbb{N}^*$, then $\Theta_t \xrightarrow[t\to+\infty]{\mathcal{D}} \theta^\star$.*

Posterior concentration in i.i.d. settings can be formally enforced by explicit regularity conditions (e.g. Condition C1 in Section S7.2.1). In the case of state-space models with dependent observations, we treat posterior concentration as a working assumption. From now on, we assume that Assumption A3 holds, so that we can unambiguously refer to the limit point $\theta^\star$ around which the posterior distribution concentrates.

In addition to concentration of the posterior distribution, we also want the posterior moments of specific test functions to converge, $\mathbb{P}_\star$-almost surely. In particular, as the posterior distribution concentrates to a point mass, we want the posterior expectations and variances appearing in (5) to respectively converge to a finite limit and to 0, as the number of observations increases (Assumption A4).

Assumption A4. *The following limits hold:*

(a) $\mathbb{E}\left[\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \Theta)\right) \,\middle|\, Y_{1:t}\right] - \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)\right) \xrightarrow[t\to+\infty]{\mathbb{P}_\star-a.s.} 0.$

(b) $\mathrm{Var}\left(\dfrac{\partial \log p(Y_t|Y_{1:t-1}, \Theta)}{\partial y_t} \,\middle|\, Y_{1:t}\right) \xrightarrow[t\to+\infty]{\mathbb{P}_\star-a.s.} 0.$

By Stolz-Cesàro's theorem, the $\mathbb{P}_\star$-a.s. convergence of the posterior moments in Assumption A4 implies the $\mathbb{P}_\star$-a.s. convergence of their Cesàro means. This leads to the convergence of the prequential quantities, so that

$$\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \Theta)\right) \,\middle|\, Y_{1:t}\right]\right) - \left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)\right)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star-a.s.} 0\,, \tag{s8}$$

and

$$\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p(Y_t|Y_{1:t-1}, \Theta)}{\partial y_t} \,\middle|\, Y_{1:t}\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star-a.s.} 0\,. \tag{s9}$$

At this stage, the proof starts to differ depending on which setting we consider.

S7.1.1 Models for i.i.d. data

For i.i.d. models, we have $\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)\right) = \mathcal{H}\left(Y_t, p(dy|\theta^\star)\right)$ for all $t \in \mathbb{N}^*$. If the Y_t's are generated as i.i.d. from $p_\star$ (Assumption A5), then the integrability of $\mathcal{H}\left(Y, p(dy|\theta^\star)\right)$ with respect to $Y \sim p_\star$ (Assumption A6) enables the application of the law of large numbers to the quantity $T^{-1}\sum_{t=1}^{T}\mathcal{H}\left(Y_t, p(dy|\theta^\star)\right)$.

Assumption A5. *The observations $(Y_t)_{t\in\mathbb{N}^*}$ are i.i.d. draws from $p_\star$.*

Assumption A6. *The H-score of $p(dy|\theta^\star)$ is integrable: $\mathbb{E}_\star\left[\left|\mathcal{H}\left(Y, p(dy|\theta^\star)\right)\right|\right] < +\infty$.*

Under Assumptions A5 and A6, the law of large numbers reduces (s8) to

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathcal{H}\left(Y_t, p(dy|\Theta)\right)\,\middle|\, Y_{1:t}\right] \quad \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} \quad \mathbb{E}_\star\left[\mathcal{H}\left(Y, p(dy|\theta^\star)\right)\right], \tag{s10}$$

where the expectation is taken with respect to $Y \sim p_\star$. If M_1 and M_2 are both i.i.d. models satisfying A1, A3, A4, and A6, then combining (5), (s9) and (s10) leads to

$$\frac{1}{T}\mathcal{H}_T(M_j) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_j(dy|\theta_j^\star)\right)\right],$$

for each $j \in \{1, 2\}$. Taking the difference of the respective scores yields

$$\frac{1}{T}\left(\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_2(dy|\theta_2^\star)\right)\right] - \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_1(dy|\theta_1^\star)\right)\right], \tag{s11}$$

In order to interpret the consistency of the H-score in terms of an appropriate divergence, we impose further regularity assumptions on the models and the data-generating process itself (Assumption A7). Assumption A7(a) allows us to define the H-score of $p_\star$, assumed to be integrable by A7(b). Assumption A7(c) ensures the strict propriety of the H-score.

Assumption A7. *The data-generating process and the model satisfy the following conditions:*

(a) $y \mapsto p_\star(y)$ *is twice differentiable.*

(b) $\mathbb{E}_\star\left[\left|\mathcal{H}\left(Y, p_\star(dy)\right)\right|\right] < +\infty$.

(c) $\dfrac{\partial \log p(y|\theta^\star)}{\partial y} p_\star(y) \xrightarrow[|y|\to+\infty]{} 0$.

Under Assumptions A3, A6, and A7, we can define the divergence $D_{\mathcal{H}}(p_\star, M_j)$ as in (7). By adding and subtracting $\mathbb{E}_\star\left[\mathcal{H}\left(Y, p_\star(dy)\right)\right]$ in (s11), we get

$$\frac{1}{T}\left(\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} D_{\mathcal{H}}(p_\star, M_2) - D_{\mathcal{H}}(p_\star, M_1).$$

Under Assumption A7(c), integration by parts (Hyvärinen, 2005; Dawid and Musio, 2015) leads to

$$D_{\mathcal{H}}(p_\star, M_j) = \int\left(\frac{\partial \log p_\star(y)}{\partial y} - \frac{\partial \log p_j(y|\theta_j^\star)}{\partial y}\right)^2 p_\star(y)dy.$$

Therefore, we have $D_{\mathcal{H}}(p_\star, M_j) \geq 0$.

If $D_{\mathcal{H}}(p_\star, M_j) = 0$, then $\partial \log p_\star(y)/\partial y = \partial \log p_j(y|\theta_j^\star)/\partial y$ for all $y \in \mathbb{Y}$. Hence, $\log p_\star(y) = \log p_j(y|\theta_j^\star) + \log(c)$ for all $y \in \mathbb{Y}$ and some constant $c > 0$. This leads to $p_\star(y) = c\, p_j(y|\theta_j^\star)$ for all $y \in \mathbb{Y}$. Since probability densities integrate to 1, we necessarily have $c = 1$, i.e. $p_\star(y) = p_j(y|\theta_j^\star)$ for all $y \in \mathbb{Y}$. This concludes the proof of Theorem 1.

S7.1.2 State-space models

In the case of state-space models and dependent observations, more subtle arguments are needed since we can no longer apply the standard law of large numbers to the term $T^{-1}\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}\theta^\star)\right)$ in (s8). Instead, we approximate this term by a stationary analog, to which ergodic theorems will apply.

To this end, we assume the process $(Y_t)_{t\in\mathbb{N}^*}$ is strongly stationary and ergodic (Assumption A8). Under strong stationarity, we can artificially extend the index set to negative integers and consider the two-sided process $(Y_t)_{t\in\mathbb{Z}}$. We also need the dependence of the H-score on the initial distribution of the latent Markov chain to vanish quickly enough. This will be referred to as the *forgetting property* of the H-score (Assumption A9).

Assumption A8. *The process $(Y_t)_{t\in\mathbb{N}^*}$ is strongly stationary and ergodic.*

Assumption A9. *There exist $\rho \in (0,1)$ and $\gamma > 0$ such that, for all $t \in \mathbb{N}^*$, all $m \in \mathbb{N}$, and all $y_{-m:t} \in \mathbb{Y}^{m+t+1}$,*

$$|\mathcal{H}(y_t, p(dy_t|y_{-m+1:t-1}, \theta^\star)) - \mathcal{H}(y_t, p(dy_t|y_{-m:t-1}, \theta^\star))| \leq \gamma\, \rho^{t+m-1}.$$

Under Assumptions A8 and A9, we can prove that, $\mathbb{P}_\star$-a.s., for all $t \in \mathbb{N}^*$, $(\mathcal{H}(Y_t, p(dy_t|Y_{-m+1:t-1}, \theta^\star)))_{m\in\mathbb{N}}$ is a real-valued Cauchy sequence, and thus converges to a $\mathbb{P}_\star$-a.s. limit denoted by $\mathcal{H}(Y_t, p(dy_t|Y_{-\infty:t-1}, \theta^\star))$. In other words, $\mathbb{P}_\star$-almost surely, and for all $t \in \mathbb{N}^*$, we have

$$\mathcal{H}(Y_t, p(dy_t|Y_{-m+1:t-1}, \theta^\star)) \quad \xrightarrow[m\to+\infty]{} \quad \mathcal{H}(Y_t, p(dy_t|Y_{-\infty:t-1}, \theta^\star))\,. \tag{s12}$$

Using (s12) and the forgetting property in A9, we can prove that

$$\left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)\right)\right) - \left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{-\infty:t-1}, \theta^\star)\right)\right) \quad \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} \quad 0\,. \tag{s13}$$

The proofs of (s12) and (s13) are provided in Section S7.3. Equation (s13) implies that, $\mathbb{P}_\star$-almost surely, the term $T^{-1}\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}\theta^\star)\right)$ in (s8) can be asymptotically approximated by the stationary quantity $T^{-1}\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{-\infty:t-1}, \theta^\star)\right)$, to which ergodic theorems can be applied under adequate integrability conditions (Assumption A10).

Assumption A10. *The limit in (s12) is integrable:* $\mathbb{E}_\star\left[\left|\mathcal{H}\left(Y_1, p(dy_1|Y_{-\infty:0}, \theta^\star)\right)\right|\right] < +\infty$.

Under Assumption A10, using Birkhoff's ergodic theorem after combining (s8) and (s13) yields

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \Theta)\right) \,\middle|\, Y_{1:t}\right] \quad \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} \quad \mathbb{E}_\star\left[\mathcal{H}\left(Y_1, p(dy_1|Y_{-\infty:0}, \theta^\star)\right)\right]. \tag{s14}$$

Under Assumption A8, and assuming both models M_1 and M_2 satisfy A1, A3, A4, A9, and A10, we can piece together (5), (s9), and (s14). This leads to

$$\frac{1}{T}\left(\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} \mathbb{E}_\star\left[\mathcal{H}\left(Y_1, p_2(dy_1|Y_{-\infty:0}, \theta_2^\star)\right)\right] - \mathbb{E}_\star\left[\mathcal{H}\left(Y_1, p_1(dy_1|Y_{-\infty:0}, \theta_1^\star)\right)\right]. \tag{s15}$$

In order to interpret this consistency result in terms of a divergence, we need $\mathcal{H}\left(Y_1, p_j(dy_1|Y_{-\infty:0}, \theta_j^\star)\right)$ to correspond to the actual H-score at Y_1 of a twice differentiable probability density function $y_1 \mapsto p_j(y_1|Y_{-\infty:0}, \theta_j^\star)$, conditional on $Y_{-\infty:0}$, $\mathbb{P}_\star$-almost surely (Assumption A11).

Assumption A11. *$\mathbb{P}_\star$-almost surely, we can define the conditional density $p(y_1|Y_{-\infty:0}, \theta^\star)$ of Y_1 given $Y_{-\infty:0}$, and the limit $\mathcal{H}(Y_1, p(dy_1|Y_{-\infty:0}, \theta^\star))$ in (s12) corresponds to the actual H-score at Y_1 of $p(dy_1|Y_{-\infty:0}, \theta^\star)$.*

In order to define the divergence $D_{\mathcal{H}}$, we make further regularity assumptions on the models and the data-generating process (Assumption A12). Similarly to Assumption A7 in the i.i.d. setting, A12(a) allows us to define the H-score of $p_\star(dy_1|Y_{-\infty:0})$, conditional on $Y_{-\infty:0}$, $\mathbb{P}_\star$-almost surely, while A12(b) enforces its integrability. Assumption A12(c) ensures the strict propriety of the H-score, conditional on $Y_{-\infty:0}$, $\mathbb{P}_\star$-almost surely.

Assumption A12. *The data-generating process and the model satisfy the following conditions:*

(a) The conditional density $y_1 \mapsto p_\star(y_1|Y_{-\infty:0})$ of Y_1 given $Y_{-\infty:0}$ is well-defined and twice differentiable.

(b) $\mathbb{E}_\star\left[\left|\mathcal{H}\left(Y_1, p_\star(dy_1|Y_{-\infty:0})\right)\right|\right] < +\infty$.

(c) $\dfrac{\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1} p_\star(y_1|Y_{-\infty:0}) \xrightarrow[|y_1|\to+\infty]{\mathbb{P}_\star - a.s.} 0$.

Under Assumptions A3, A10, and A12, we can define the divergence $D_{\mathcal{H}}(p_\star, M_j)$ as in (14). By adding and subtracting $\mathbb{E}_\star\left[\mathcal{H}\left(Y_1, p_\star(dy_1|Y_{-\infty:0})\right)\right]$ in (s15), we get

$$\frac{1}{T}\left(\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star - a.s.} D_{\mathcal{H}}(p_\star, M_2) - D_{\mathcal{H}}(p_\star, M_1).$$

The tower property of conditional expectations, combined with an integration by parts using A12(c), leads to

$$D_{\mathcal{H}}(p_\star, M_j) = \mathbb{E}_\star\left[\int \left(\frac{\partial \log p_\star(y_1|Y_{-\infty:0})}{\partial y_1} - \frac{\partial \log p_j(y_1|Y_{-\infty:0}, \theta_j^\star)}{\partial y_1}\right)^2 p_\star(y_1|Y_{-\infty:0}) dy_1\right].$$

Thus, we have $D_{\mathcal{H}}(p_\star, M_j) \geq 0$. If $D_{\mathcal{H}}(p_\star, M_j) = 0$, the same reasoning as in the proof of (6) shows that, $\mathbb{P}_\star$-almost surely, we have $p_j(y_1|Y_{-\infty:0}, \theta_j^\star) = p_\star(y_1|Y_{-\infty:0})$ for all $y_1 \in \mathbb{Y}$. This concludes the proof of Theorem 2.

S7.2 Sufficient conditions and intermediary results

S7.2.1 Assumption A3: Concentration of the posterior distribution

For a generic i.i.d. model $\{p(dy|\theta) : \theta \in \mathbb{T}\}$ with prior distribution $p(d\theta)$, if the observations are assumed to be i.i.d. from $p_\star$, then Theorem 1.3.4. in Ghosh and Ramamoorthi (2003) shows that the following set of regularity conditions (Condition C1) ensures the concentration of the posterior. In other words, for i.i.d. models and data, Condition C1 and Assumption A5 guarantee Assumption A3.

Condition C1. *The model satisfies the following conditions:*

(a) $\mathbb{T}$ *is a compact metric space, and* $p(\theta) > 0$ *for all* $\theta \in \mathbb{T}$.

(b) $y \mapsto p(y|\theta)$ *is measurable for all* $\theta \in \mathbb{T}$, *and* $\theta \mapsto p(y|\theta)$ *is continuous for all* $y \in \mathbb{Y}$.

(c) $\int_{\mathbb{Y}} \sup_{\theta \in \mathbb{T}} |\log p(y|\theta)| \, p_\star(y) dy < +\infty$.

Condition C1 can be relaxed to allow for semi-continuity and non-compact parameter spaces, as discussed in Remark 1.3.5 of Ghosh and Ramamoorthi (2003) and its references (e.g. Wald, 1949; Le Cam, 1953; Kiefer and Wolfowitz, 1956; Huber, 1967; Perlman et al., 1972).

Posterior concentration for general state-space models with dependent data is less standard, especially when allowing for misspecification. Some concentration results have been proved in specific cases (e.g. Lijoi, Prünster and Walker, 2007; De Gunst and Shcherbakova, 2008; Shalizi, 2009; Gassiat, Rousseau et al., 2014; Douc, Olsson and Roueff, 2016, and references therein). However, as far as we know, a formal proof of posterior concentration with explicit conditions on possibly misspecified state-space models has yet to be derived.

S7.2.2 Assumption A4: Convergence of specific posterior moments

Concentration of the posterior distribution does not guarantee convergence of any posterior moments. The latter can be ensured by further imposing equicontinuity (Condition C2) and uniform integrability (Condition C3).

Condition C2. $\mathbb{P}_\star$*-almost surely, the following statements hold:*

(a) $\{\theta \mapsto \mathcal{H}(Y_t, p(dy_t|Y_{1:t-1}, \theta)) : t \in \mathbb{N}^*\}$ *is equicontinuous at* $\theta^\star$.

(b) $\left\{\theta \mapsto \dfrac{\partial \log p(Y_t|Y_{1:t-1}, \theta)}{\partial y_t} : t \in \mathbb{N}^*\right\}$ *is equicontinuous at* $\theta^\star$.

Equicontinuity at $\theta^\star$ of a family of functions $\{\theta \longmapsto h_t(\theta) : t \in \mathbb{N}^*\}$ means that all the functions in the family share a common (i.e. not depending on t) modulus of continuity at $\theta^\star$. Equicontinuity can be enforced by the stronger but more explicit condition that there exists a neighborhood $\mathcal{U}_{\theta^\star}$ of $\theta^\star$, on which the functions $\theta \longmapsto h_t(\theta)$ are differentiable for all $t \in \mathbb{N}^*$, and such that $\sup_{(t,\theta) \in \mathbb{N}^* \times \mathcal{U}_{\theta^\star}} \|\nabla_\theta h_t(\theta)\| = L < +\infty$. Indeed, by the mean value theorem, such uniform boundedness of the gradients ensures that the functions $\theta \longmapsto h_t(\theta)$ are L-Lipschitz on $\mathcal{U}_{\theta^\star}$ for all $t \in \mathbb{N}^*$, where L does not depend on t. Then, for any arbitrary $\varepsilon > 0$, we can find $\delta_\varepsilon > 0$ not depending on t (e.g. $\delta_\varepsilon = \varepsilon / L$ if $L > 0$, or else any $\delta_\varepsilon > 0$ if $L = 0$) such that, for all $\theta \in \mathcal{U}_{\theta^\star}$, $\|\theta - \theta^\star\| < \delta_\varepsilon$ implies $\sup_{t \in \mathbb{N}^*} \|h_t(\theta) - h_t(\theta^\star)\| < \varepsilon$, which proves the equicontinuity at $\theta^\star$ of the family $\{\theta \longmapsto h_t(\theta) : t \in \mathbb{N}^*\}$.

Condition C3. $\mathbb{P}_\star$*-almost surely, if* $\Theta_t \sim p(d\theta|Y_{1:t})$ *for all* $t \in \mathbb{N}^*$, *then the following statements hold:*

(a) $\{\mathcal{H}(Y_t, p(dy_t|Y_{1:t-1}, \Theta_t)) : t \in \mathbb{N}^*\}$ *is uniformly integrable given* $(Y_t)_{t \in \mathbb{N}^*}$.

(b) $\left\{\left(\dfrac{\partial \log p(Y_t|Y_{1:t-1}, \Theta_t)}{\partial y_t}\right)^2 : t \in \mathbb{N}^*\right\}$ *is uniformly integrable given* $(Y_t)_{t \in \mathbb{N}^*}$.

Uniform integrability of a family of random variables $\{H_t : t \in \mathbb{N}^*\}$ can be enforced by the stronger but more explicit condition of L_α-boundedness: if there exists $\alpha > 1$ such that $\sup_{t \in \mathbb{N}^*} \mathbb{E}[|H_t|^\alpha] < +\infty$, then $\{H_t : t \in \mathbb{N}^*\}$ is uniformly integrable (e.g. see Theorem 25.12 and its corollary in Billingsley, 1995).

Convergence of the relevant posterior moments in Assumption A4 can be obtained as a consequence of Assumption A3 combined with Conditions C2 and C3. This is summarized by the following lemma.

Lemma 1. *Assume A3, C2, and C3. Then Assumption A4 holds and we have:*

(a) $\mathbb{E}\left[\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \Theta)\right) \middle| Y_{1:t}\right] - \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)\right) \xrightarrow[t\to+\infty]{\mathbb{P}_\star - a.s.} 0.$

(b) $\mathrm{Var}\left(\frac{\partial \log p(Y_t|Y_{1:t-1}, \Theta)}{\partial y_t} \middle| Y_{1:t}\right) \xrightarrow[t\to+\infty]{\mathbb{P}_\star - a.s.} 0.$

The proof of Lemma 1 is provided in Section S7.3.3.

S7.2.3 Assumption A9: Forgetting property of the H-score

For state-space models, the forgetting property of the H-score can be obtained as a consequence of the forgetting property of the latent Markov chain stated in (s16) (following from Condition C4) and some appropriate boundedness conditions on the first two derivatives of the observation log-density (Condition C5). Condition C4 corresponds to a simplified version of Assumption A13.1 in Douc, Moulines and Stoffer (2014).

Condition C4. *The model satisfies the following conditions:*

(a) There exists a dominating probability measure η *on* $\mathbb{X}$ *such that the transition kernel* $f_{\theta^\star}(dx_{t+1}|x_t)$ *has density* $\nu_{\theta^\star}(x_{t+1}|x_t) = (df_{\theta^\star}(\cdot|x_t)/d\eta)(x_{t+1})$ *with respect to* η.

(b) There exist positive constants σ^- *and* σ^+ *such that, for all* $(x_t, x_{t+1}) \in \mathbb{X} \times \mathbb{X}$, *the transition density* $\nu_{\theta^\star}(x_{t+1}|x_t)$ *satisfies* $0 < \sigma^- < \nu_{\theta^\star}(x_{t+1}|x_t) < \sigma^+ < +\infty$.

(c) For all $y_t \in \mathbb{Y}$, *the integral* $\int_{\mathbb{X}} g_{\theta^\star}(y_t, x_t)\, \eta(dx_t)$ *is bounded away from* 0 *and* $+\infty$.

Under strong stationarity of the process $(Y_t)_{t\in\mathbb{N}^*}$, Lemma 13.2 in Douc et al. (2014) guarantees that for all $t \in \mathbb{N}^*$, all $m \in \mathbb{N}$, and all realizations $y_{-m:t} \in \mathbb{Y}^{m+t+1}$, the filtering distributions of the latent states satisfy

$$d_{TV}\left(p(dx_t|y_{-m+1:t}, \theta^\star), p(dx_t|y_{-m:t}, \theta^\star)\right) \leq \rho^{t+m-1}, \tag{s16}$$

where d_{TV} stands for the total variation distance and $\rho = 1 - (\sigma^-/\sigma^+) \in (0, 1)$. Condition C4 would typically require the latent space $\mathbb{X}$ to be finite or compact, and ensures that the transition kernel is geometrically ergodic. Such a condition can generally be weakened to allow for non-finite and non-compact spaces (e.g. Douc, Moulines, Olsson, Van Handel et al., 2011; Douc, Moulines et al., 2012).

When (11) and (12) hold, H-scores for a fixed $\theta^\star$ can be written in terms of expectations with respect to the corresponding filtering distributions. Differences of H-scores can then be related to total variation distance between filtering distributions by assuming the integrands in (11) and (12) are bounded (Condition C5).

Condition C5. *The model satisfy the following domination conditions:*

(a) $b = \sup_{\substack{x\in\mathbb{X}\\ y\in\mathbb{Y}}} \left| \frac{\partial^2 \log g_{\theta^\star}(y|x)}{\partial y^2} + \left(\frac{\partial \log g_{\theta^\star}(y|x)}{\partial y}\right)^2 \right| < +\infty.$

(b) $c = \sup_{\substack{x\in\mathbb{X}\\ y\in\mathbb{Y}}} \left|\frac{\partial \log g_{\theta^\star}(y|x)}{\partial y}\right| < +\infty$.

Condition C5 could be enforced by the stronger conditions that $\mathbb{X}$ and $\mathbb{Y}$ are compact, and the first two derivatives of $y \mapsto \log g_{\theta^\star}(y|x)$ are continuous with respect to (x, y). Compactness conditions may look quite restrictive, since most well-known continuous distributions have non compact supports. In reality, for all practical purposes, we could always envision a sufficiently large compact space in which all our numerical values would lie. As stated earlier, Conditions C4 and C5 should only be regarded as mere sufficient conditions that allow for more straightforward proofs. The models of our simulation studies do not satisfy these conditions, and yet we observe the consistency of H-scores. This indicates that consistency is likely to hold under weaker conditions.

Under Assumptions A2, A3, and A8, Conditions C4 and C5 combined with (11)-(12) guarantee that the forgetting property of the H-score in Assumption A9 holds, as stated by the following lemma.

Lemma 2. *Assume A2, A3, A8, C4, and C5. Then, for all $t \in \mathbb{N}^*$, all $m \in \mathbb{N}$, and all $y_{-m:t} \in \mathbb{Y}^{m+t+1}$,*

$$|\mathcal{H}(y_t, p(dy_t|y_{-m+1:t-1}, \theta^\star)) - \mathcal{H}(y_t, p(dy_t|y_{-m:t-1}, \theta^\star))| \leq 2(b + c^2)\, \rho^{t+m-1}, \tag{s16}$$

$$\sup_{m\in\mathbb{N}} |\mathcal{H}(y_t, p(dy_t|y_{-m+1:t-1}, \theta^\star))| \leq 2\, b + c^2, \tag{s17}$$

where $\rho = 1 - \frac{\sigma^-}{\sigma^+} \in (0, 1)$.

Equation (s16) in Lemma 2 enforces Assumption A9 with $\gamma = 2(b+c^2)$, while (s17) directly buys us Assumptions A6 and A10. The proof of Lemma 2 is provided in Section S7.3.

S7.2.4 Assumption A11: H-score of conditional density given the infinite past

Ensuring that we may define $y_1 \mapsto p(y_1|Y_{-\infty:0}, \theta^\star)$ as an actual probability density function can be done under further domination and integrability conditions on the observation density (Condition C6).

Condition C6. *Let $\nu_{\theta^\star}$ be the probability measure from Condition C4. The following statements hold:*

(a) $\sup_{\substack{x\in\mathbb{X}\\ y\in\mathbb{Y}}} g_{\theta^\star}(y|x) < +\infty$.

(b) $\mathbb{E}_\star\left[\left|\log\left(\int_{\mathbb{X}} g_{\theta^\star}(Y_1|x)\nu_{\theta^\star}(dx)\right)\right|\right] < +\infty$.

Condition C6 corresponds to a simplified statement of Assumption A13.3 in Douc et al. (2014). Under Assumption A8 with Conditions C4 and C6, Lemma 13.12 and Proposition 13.5 from Douc et al. (2014) show that $y_1 \mapsto \log p(y_1|Y_{-\infty:0}, \theta^\star) = \lim_{m\to+\infty} \log p(y_1|Y_{-m+1:0}, \theta^\star)$ exists and defines an actual log-density, $\mathbb{P}_\star$-almost surely. The $\mathbb{P}_\star$-almost sure twice differentiability of $y_1 \mapsto \log p(y_1|Y_{-\infty:0}, \theta^\star)$ follows from the uniform convergence of the first two derivatives of $y_1 \mapsto \log p(y_1|Y_{-m+1:0}, \theta^\star)$ as $m \to +\infty$ (e.g. Theorem 7.17 from Rudin, 1964), which can be proved using (11)-(12) and the domination conditions from C5. In other words, under Assumptions A2 and A8, Conditions C4 to C6 ensure that Assumption A11 holds. This is stated by the following lemma.

Lemma 3. *Assume A2, A8, C4, C5, and C6. Then, $\mathbb{P}_\star$-almost surely, there exists a continuous probability density function $x_1 \mapsto p(x_1|Y_{-\infty:0},\theta^\star) = \lim_{m\to+\infty} p(x_1|Y_{-m+1:0},\theta^\star)$ with respect to $\nu_{\theta^\star}$. Define the function*

$$y_1 \mapsto p(y_1|Y_{-\infty:0},\theta^\star) = \int g_{\theta^\star}(y_1|x_1)p(x_1|Y_{-\infty:0},\theta^\star)\nu_{\theta^\star}(dx_1).$$

Then, $\mathbb{P}_\star$-almost surely, $p(Y_1|Y_{-\infty:0},\theta^\star) = p(y_1|Y_{-\infty:0},\theta^\star)|_{y_1=Y_1}$, and $y_1 \mapsto p(y_1|Y_{-\infty:0},\theta^\star)$ is the conditional density with respect to the Lebesgue measure of Y_1 given the σ-algebra generated by $(Y_{-m})_{m\in\mathbb{N}}$ under $\mathbb{P}_\star$. Moreover, $\mathbb{P}_\star$-a.s., $y_1 \mapsto \log p(y_1|Y_{-\infty:0},\theta^\star) = \lim_{m\to+\infty} \log p(y_1|Y_{-m+1:0},\theta^\star)$ exists and is twice differentiable on $\mathbb{Y}$, with

$$\frac{\partial \log p(y_1|Y_{-\infty:0},\theta^\star)}{\partial y_1} = \lim_{m\to+\infty} \frac{\partial \log p(y_1|Y_{-m+1:0},\theta^\star)}{\partial y_1},$$
$$\frac{\partial^2 \log p(y_1|Y_{-\infty:0},\theta^\star)}{\partial y_1^2} = \lim_{m\to+\infty} \frac{\partial^2 \log p(y_1|Y_{-m+1:0},\theta^\star)}{\partial y_1^2},$$

and

$$\mathcal{H}(y_1, p(dy_1|Y_{-\infty:0},\theta^\star)) = 2\,\frac{\partial^2 \log p(y_1|Y_{-\infty:0},\theta^\star)}{\partial y_1^2} + \left(\frac{\partial^2 \log p(y_1|Y_{-\infty:0},\theta^\star)}{\partial y_1^2}\right)^2,$$

for all $y_1 \in \mathbb{Y}$.

The proof of Lemma 3 is provided in Section S7.3.

S7.3 Proofs of intermediary results

S7.3.1 Proof of (s12)

Fix some arbitrary $\varepsilon > 0$ and $t \in \mathbb{N}^*$. Since $\rho \in (0,1)$, we have $\rho^N \to 0$ as $N \to +\infty$, so there exists some $N \in \mathbb{N}$ large enough such that $\gamma\,\rho^{t+N}(1-\rho)^{-1} < \varepsilon$. Using Assumption A9, we get, $\mathbb{P}_\star$-almost surely, for any $n > m > N$,

$$\begin{aligned}
|\mathcal{H}(Y_t, p(dy_t|Y_{-m+1:t-1},\theta^\star)) - \mathcal{H}(Y_t, p(dy_t|Y_{-n+1:t-1},\theta^\star))| &\leq \sum_{k=m}^{n-1} |\mathcal{H}(Y_t, p(dy_t|Y_{-k+1:t-1},\theta^\star)) - \mathcal{H}(Y_t, p(dy_t|Y_{-k:t-1},\theta^\star))| \\
&\leq \gamma\,\rho^{t-1} \sum_{k=m}^{n-1} \rho^k \\
&\leq \gamma\,\rho^{t-1} \sum_{k=N+1}^{+\infty} \rho^k \\
&\leq \varepsilon.
\end{aligned}$$

Therefore $(\mathcal{H}(Y_t, p(dy_t|Y_{-m+1:t-1},\theta^\star))_{m\in\mathbb{N}}$ is a Cauchy sequence for every $t \in \mathbb{N}^*$, $\mathbb{P}_\star$-almost surely. Since $\mathbb{R}$ is complete, this sequence converges $\mathbb{P}_\star$-almost surely to a limit, denoted by

$$\mathcal{H}(Y_t, p(dy_t|Y_{-m+1:t-1},\theta^\star)) \xrightarrow[m\to+\infty]{\mathbb{P}_\star-a.s.} \mathcal{H}(Y_t, p(dy_t|Y_{-\infty:t-1},\theta^\star)).$$

□

S7.3.2 Proof of (s13)

We have, $\mathbb{P}_\star$-almost surely, for every $T \in \mathbb{N}^*$,

$$\begin{aligned}
&\left| \frac{1}{T} \sum_{t=1}^{T} \left(\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)\right) - \mathcal{H}\left(Y_t, p(dy_t|Y_{-\infty:t-1}, \theta^\star)\right) \right) \right| \\
\leq\ & \frac{1}{T} \sum_{t=1}^{T} \sum_{m=0}^{+\infty} \left| \mathcal{H}\left(Y_t, p(dy_t|Y_{-m+1:t-1}, \theta^\star)\right) - \mathcal{H}\left(Y_t, p(dy_t|Y_{-m:t-1}, \theta^\star)\right) \right| \\
\leq\ & \frac{\gamma}{T} \sum_{t=1}^{T} \sum_{m=0}^{+\infty} \rho^{t+m-1},
\end{aligned}$$

where $\rho \in (0,1)$ and $\gamma > 0$ are given by Assumption A9. Properties of geometric series lead to

$$\left| \left(\frac{1}{T} \sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)\right) \right) - \left(\frac{1}{T} \sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{-\infty:t-1}, \theta^\star)\right) \right) \right| \leq \frac{\gamma}{T} \sum_{t=1}^{+\infty} \rho^{t-1} \sum_{m=0}^{+\infty} \rho^m \leq \frac{\gamma}{T(1-\rho)^2}.$$

The upper bound goes to 0 as $T \to +\infty$, therefore

$$\left(\frac{1}{T} \sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1}, \theta^\star)\right) \right) - \left(\frac{1}{T} \sum_{t=1}^{T} \mathcal{H}\left(Y_t, p(dy_t|Y_{-\infty:t-1}, \theta^\star)\right) \right) \xrightarrow[T \to +\infty]{\mathbb{P}_\star - a.s.} 0.$$

□

S7.3.3 Proof of Lemma 1

Any finite intersection of almost sure events is an almost sure event, thus we can find a common event A such that $\mathbb{P}_\star(A) = 1$, and on which all the assumptions and conditions hold simultaneously. Fix some arbitrary $\omega \in A$. For all $t \in \mathbb{N}^*$, define $y_t = Y_t(\omega)$ and let $\Theta_t \sim p(d\theta|y_{1:t})$. By Assumption A3, we have $\Theta_t \xrightarrow[t \to +\infty]{\mathcal{D}} \theta^\star$. The space $\mathbb{T}$ is a metric space and the support of the limit distribution δ_{θ^*} is the singleton $\{\theta^*\}$, which is separable, so by Skorokhod's representation theorem (e.g. see Theorem 6.7 in Billingsley, 1968), we can construct random variables $(\Theta'_t)_{t \in \mathbb{N}^*}$ on some instrumental probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\Theta'_t \sim \Theta_t$ for all $t \in \mathbb{N}^*$ and $\Theta'_t \xrightarrow[t \to +\infty]{\mathbb{P}-a.s.} \theta^\star$, where $\mathbb{P}$ captures the randomness of $(\Theta'_t)_{t \in \mathbb{N}^*}$ conditional on the realizations $(y_t)_{t \in \mathbb{N}^*}$. We have, $\mathbb{P}$-almost surely, for any arbitrary $\varepsilon > 0$ and the corresponding $\delta_\varepsilon > 0$ given by the equicontinuity stated in Condition C2(a), the existence of some $t_0 \in \mathbb{N}^*$ such that, for every $t > t_0$, we have $d(\Theta'_t, \theta^\star) < \delta_\varepsilon$ and

$$\left| \mathcal{H}(y_t, p(dy_t|y_{1:t-1}, \Theta'_t)) - \mathcal{H}(y_t, p(dy_t|y_{1:t-1}, \theta^\star)) \right| \leq \varepsilon.$$

Therefore, we have

$$\mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \Theta'_t)\right) - \mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \theta^\star)\right) \xrightarrow[t \to +\infty]{\mathbb{P}-a.s.} 0. \tag{s18}$$

Similarly, using C2(b), we get

$$\frac{\partial \log p(y_t|y_{1:t-1},\Theta'_t)}{\partial y_t} - \frac{\partial \log p(y_t|y_{1:t-1},\theta^\star)}{\partial y_t} \xrightarrow[t\to+\infty]{\mathbb{P}-a.s.} 0. \tag{s19}$$

The family $\{\mathcal{H}\left(y_t, p(dy_t|y_{1:t-1},\Theta'_t)\right) - \mathcal{H}\left(y_t, p(dy_t|y_{1:t-1},\theta^\star)\right)\}_{t\in\mathbb{N}^*}$ is uniformly integrable by Condition C3(a) and the fact that $\Theta'_t \sim \Theta_t \sim p(d\theta|y_{1:t})$ for all $t \in \mathbb{N}^*$, so that the convergence from (s18) implies the convergence of the first moments (e.g. see Theorem 25.12 in Billingsley, 1995). In other words, we get

$$\mathbb{E}\left[\mathcal{H}\left(y_t, p(dy_t|y_{1:t-1},\Theta'_t)\right) \mid y_{1:t}\right] - \mathcal{H}\left(y_t, p(dy_t|y_{1:t-1},\theta^\star)\right) \xrightarrow[t\to+\infty]{} 0.$$

By construction, we have $\Theta'_t \sim \Theta_t$, thus

$$\mathbb{E}\left[\mathcal{H}\left(y_t, p(dy_t|y_{1:t-1},\Theta_t)\right) \mid y_{1:t}\right] - \mathcal{H}\left(y_t, p(dy_t|y_{1:t-1},\theta^\star)\right) \xrightarrow[t\to+\infty]{} 0.$$

Since this holds for all $\omega \in A$ and $\mathbb{P}_\star(A) = 1$, we conclude that

$$\mathbb{E}\left[\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1},\Theta)\right) \mid Y_{1:t}\right] - \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1},\theta^\star)\right) \xrightarrow[t\to+\infty]{\mathbb{P}_\star-a.s.} 0\,,$$

where the expectation is taken with respect to the posterior distribution of Θ given $Y_{1:t}$, which proves A4(a).

Similarly, the family $\left\{\left(\partial \log p(y_t|y_{1:t-1},\Theta'_t)/\partial y_t - \partial \log p(y_t|y_{1:t-1},\theta^\star)/\partial y_t\right)^2\right\}_{t\in\mathbb{N}^*}$ is uniformly integrable by Condition C3(b) and the fact that $\Theta'_t \sim \Theta_t \sim p(d\theta|y_{1:t})$ for all $t \in \mathbb{N}^*$, so that the convergence from (s19) implies the convergence of the first two moments, and a fortiori the convergence of the variance. Thus,

$$\mathrm{Var}\left(\frac{\partial \log p(y_t|y_{1:t-1},\Theta'_t)}{\partial y_t} - \frac{\partial \log p(y_t|y_{1:t-1},\theta^\star)}{\partial y_t} \,\middle|\, y_{1:t}\right) \xrightarrow[t\to+\infty]{} 0.$$

By construction, we have $\Theta'_t \sim \Theta_t$. Besides, $\partial \log p(y_t|y_{1:t-1},\theta^\star)/\partial y_t$ is constant given $y_{1:t}$. Therefore,

$$\mathrm{Var}\left(\frac{\partial \log p(y_t|y_{1:t-1},\Theta_t)}{\partial y_t} \,\middle|\, y_{1:t}\right) \xrightarrow[t\to+\infty]{} 0.$$

Since this holds for all $\omega \in A$ and $\mathbb{P}_\star(A) = 1$, we conclude that

$$\mathrm{Var}\left(\frac{\partial \log p(Y_t|Y_{1:t-1},\Theta)}{\partial y_t} \,\middle|\, Y_{1:t}\right) \xrightarrow[t\to+\infty]{\mathbb{P}_\star-a.s.} 0\,,$$

where the variance is taken with respect to the posterior distribution of Θ given $Y_{1:t}$, which proves A4(b). □

S7.3.4 Proof of Lemma 2

By (11)-(12) under Assumption A2, the H-score $\mathcal{H}(y_t, p(dy_t|y_{-m+1:t-1},\theta^\star))$ is equal to

$$2\int\left[\frac{\partial^2 \log g_{\theta^\star}(y_t|x_t)}{\partial y_t^2} + \left(\frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t}\right)^2\right] p(dx_t|y_{-m+1:t},\theta^\star) + \left(\int \frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t} p(dx_t|y_{-m+1:t},\theta^\star)\right)^2. \tag{s20}$$

Under Condition C5, the triangular inequality and the fact that probability densities integrate to 1 lead to

$$
\begin{aligned}
&|\mathcal{H}(y_t, p(dy_t|y_{-m+1:t-1}, \theta^\star)) - \mathcal{H}(y_t, p(dy_t|y_{-m:t-1}, \theta^\star))| \\
&\leq 2 \left| \int \left[\frac{\partial^2 \log g_{\theta^\star}(y_t|x_t)}{\partial y_t^2} + \left(\frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t} \right)^2 \right] (p(dx_t|y_{-m+1:t}, \theta^\star) - p(dx_t|y_{-m:t}, \theta^\star)) \right| \\
&\qquad + \left| \left(\int \frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t} p(dx_t|y_{-m+1:t}, \theta^\star) \right)^2 - \left(\int \frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t} p(dx_t|y_{-m+1:t}, \theta^\star) \right)^2 \right| \\
&\leq 2\, b \left| \int (p(dx_t|y_{-m+1:t}, \theta^\star) - p(dx_t|y_{-m:t}, \theta^\star)) \right| \\
&\qquad + c^2 \left| \int \left(p(dx_t|y_{-m+1:t}, \theta^\star) - p(dx_t|y_{-m:t}, \theta^\star) \right) \right| \left| \int \left(p(dx_t|y_{-m+1:t}, \theta^\star) + p(dx_t|y_{-m:t}, \theta^\star) \right) \right| \\
&\leq 2\, b\, d_{TV} \left(p(dx_t|y_{-m+1:t}, \theta^\star), p(dx_t|y_{-m:t}, \theta^\star) \right) + 2\, c^2\, d_{TV} \left(p(dx_t|y_{-m+1:t}, \theta^\star), p(dx_t|y_{-m:t}, \theta^\star) \right) \\
&\leq 2 \left(b + c^2 \right) d_{TV} \left(p(dx_t|y_{-m+1:t}, \theta^\star), p(dx_t|y_{-m:t}, \theta^\star) \right) \\
&\leq 2 \left(b + c^2 \right) \rho^{t+m-1}, \qquad \text{(s21)}
\end{aligned}
$$

where the last inequality comes from (s16) under Condition C4. This proves (s16). From (s20) and Condition C5, the triangular inequality and the fact that probability densities integrate to 1 yield (s17). □

S7.3.5 Proof of Lemma 3

We closely follow the proof of Lemma 13.12 in Douc et al. (2014). We have

$$
p(x_1|Y_{-m+1:0}, \theta^\star) = \int \nu_{\theta^\star}(x_1|x_0)\, p(dx_0|Y_{-m+1:0}, \theta^\star), \qquad \text{(s22)}
$$

for all $x_1 \in \mathbb{X}$ and all $m \in \mathbb{N}^*$, $\mathbb{P}_\star$-almost surely. By Condition C4 and (s16), we get

$$
|p(x_1|Y_{-m+1:0}, \theta^\star) - p(x_1|Y_{-m:0}, \theta^\star)| \leq \sigma^+\, d_{TV} \left(p(dx_0|Y_{-m+1:0}, \theta^\star), p(dx_0|Y_{-m:0}, \theta^\star) \right) \leq \sigma^+ \rho^{m-1},
$$

for all $x_1 \in \mathbb{X}$ and all $m \in \mathbb{N}^*$, $\mathbb{P}_\star$-almost surely. The upper bound does not depend on x_1, hence

$$
\sup_{x_1 \in \mathbb{X}} |p(x_1|Y_{-m+1:0}, \theta^\star) - p(x_1|Y_{-m:0}, \theta^\star)| \leq \sigma^+ \rho^{m-1},
$$

for all $m \in \mathbb{N}^*$, $\mathbb{P}_\star$-almost surely. The geometric series $\sum_m \rho^m$ converges as $m \to +\infty$, since $\rho \in (0, 1)$, thus

$$
\sum_{m=1}^{+\infty} \sup_{x_1 \in \mathbb{X}} |p(x_1|Y_{-m+1:0}, \theta^\star) - p(x_1|Y_{-m:0}, \theta^\star)| < +\infty,
$$

$\mathbb{P}_\star$-almost surely. In other words, we have

$$
\mathbb{P}_\star \left(\sum_{m=1}^{+\infty} \sup_{x_1 \in \mathbb{X}} |p(x_1|Y_{-m+1:0}, \theta^\star) - p(x_1|Y_{-m:0}, \theta^\star)| < +\infty \right) = 1. \qquad \text{(s23)}
$$

For any $\varepsilon > 0$, the convergence of the series in (s23) guarantees that, $\mathbb{P}_\star$-almost surely, there exists some $N \in \mathbb{N}^*$, such that $\sum_{m=N}^{+\infty} \sup_{x_1 \in \mathbb{X}} |p(x_1|Y_{-m+1:0}, \theta^\star) - p(x_1|Y_{-m:0}, \theta^\star)| < \varepsilon$. Then, for all $r > s > N$,

$$\begin{aligned}
\sup_{x_1 \in \mathbb{X}} |p(x_1|Y_{-s:0}, \theta^\star) - p(x_1|Y_{-r:0}, \theta^\star)| &= \sup_{x_1 \in \mathbb{X}} \left| \sum_{m=s+1}^{r} p(x_1|Y_{-m+1:0}, \theta^\star) - p(x_1|Y_{-m:0}, \theta^\star) \right| \\
&\leq \sum_{m=s+1}^{r} \sup_{x_1 \in \mathbb{X}} |p(x_1|Y_{-m+1:0}, \theta^\star) - p(x_1|Y_{-m:0}, \theta^\star)| \\
&\leq \sum_{m=N}^{+\infty} \sup_{x_1 \in \mathbb{X}} |p(x_1|Y_{-m+1:0}, \theta^\star) - p(x_1|Y_{-m:0}, \theta^\star)| \\
&\leq \varepsilon.
\end{aligned}$$

This implies that, $\mathbb{P}_\star$-almost surely, the sequence of non-negative continuous functions $(x_1 \mapsto p(x_1|Y_{-m:0}, \theta^\star))_{m \in \mathbb{N}}$ converges uniformly to a limit function $x_1 \mapsto p(x_1|Y_{-\infty:0}, \theta^\star) = \lim_{m \to +\infty} p(x_1|Y_{-m:0}, \theta^\star)$, which is itself necessarily non-negative and continuous, as a uniform limit of such functions. We can now check that $x_1 \mapsto p(x_1|Y_{-\infty:0}, \theta^\star)$ is indeed a probability density function.

On the one hand, applying Fatou's Lemma to the non-negative functions $(x_1 \mapsto p(x_1|Y_{-m:0}, \theta^\star))_{m \in \mathbb{N}}$ yields

$$\int p(x_1|Y_{-\infty:0}, \theta^\star) \eta(dx_1) = \int \liminf_{m \to +\infty} p(x_1|Y_{-m:0}, \theta^\star) \eta(dx_1) \leq \liminf_{m \to +\infty} \int p(x_1|Y_{-m:0}, \theta^\star) \eta(dx_1) = 1,$$

where η is the dominating measure introduced in Condition C4(a).

On the other hand, (s22) and Condition C4 imply that $0 \leq p(x_1|Y_{-m:0}, \theta^\star) \leq \sigma^+$. Applying Fatou's Lemma to the non-negative functions $(x_1 \mapsto \sigma^+ - p(x_1|Y_{-m:0}, \theta^\star))_{m \in \mathbb{N}}$ yields

$$1 = \limsup_{m \to +\infty} \int p(x_1|Y_{-m:0}, \theta^\star) \eta(dx_1) \leq \int \limsup_{m \to +\infty} p(x_1|Y_{-m:0}, \theta^\star) \eta(dx_1) = \int p(x_1|Y_{-\infty:0}, \theta^\star) \eta(dx_1).$$

These two inequalities hold $\mathbb{P}_\star$-almost surely, and prove that, $\mathbb{P}_\star$-almost surely, $x_1 \mapsto p(x_1|Y_{-\infty:0}, \theta^\star)$ is a probability density with respect to η.

Furthermore, for all $y_1 \in \mathbb{Y}$, all $x_1 \in \mathbb{X}$, and all $m \in \mathbb{N}^*$, we have, $\mathbb{P}_\star$-almost surely,

$$p(y_1|Y_{-m+1:0}, \theta^\star) = \int g_{\theta^\star}(y_1|x_1) \nu_{\theta^\star}(x_1|x_0) \, p(dx_0|Y_{-m+1:0}, \theta^\star) dx_1.$$

By using again (s16), we get

$$\begin{aligned}
|p(y_1|Y_{-m+1:0}, \theta^\star) - p(y_1|Y_{-m:0}, \theta^\star)| &\leq \sigma^+ \sup_{\substack{x \in \mathbb{X} \\ y \in \mathbb{Y}}} g_{\theta^\star}(y|x) \;\; d_{TV}\left(p(dx_0|Y_{-m+1:0}, \theta^\star), p(dx_0|Y_{-m:0}, \theta^\star) \right) \\
&\leq \sigma^+ \sup_{\substack{x \in \mathbb{X} \\ y \in \mathbb{Y}}} g_{\theta^\star}(y|x) \;\; \rho^{m-1},
\end{aligned}$$

for all $y_1 \in \mathbb{Y}$ and all $m \in \mathbb{N}^*$, $\mathbb{P}_\star$-almost surely. The supremum is finite thanks to Condition C6. Using a similar

reasoning as in the first part of the proof, we get

$$\mathbb{P}_\star\left(\sum_{m=1}^{+\infty}\sup_{y_1\in\mathbb{Y}}|p(y_1|Y_{-m+1:0},\theta^\star)-p(y_1|Y_{-m:0},\theta^\star)|<+\infty\right)=1, \tag{s24}$$

so that, $\mathbb{P}_\star$-almost surely, the sequence of functions $(y_1\mapsto p(y_1|Y_{-m:0},\theta^\star))_{m\in\mathbb{N}}$ converges uniformly to a limit function $y_1\mapsto p(y_1|Y_{-\infty:0},\theta^\star)$, and $p(Y_1|Y_{-\infty:0},\theta^\star)=p(y_1|Y_{-\infty:0},\theta^\star)|_{y_1=Y_1}$.

Consider an event $K\subseteq\mathbb{Y}$ such that $\lambda(K)<+\infty$, where λ denotes the Lebesgue measure. On the one hand, martingale convergence theorems (e.g. Corollary B.13 in Douc et al., 2014) guarantee that, $\mathbb{P}_\star$-almost surely,

$$\mathbb{E}\left[\mathbb{1}_K(Y_1)|Y_{-\infty:0},\theta^\star\right]=\lim_{m\to+\infty}\mathbb{E}\left[\mathbb{1}_K(Y_1)|Y_{-m:0},\theta^\star\right]. \tag{s25}$$

On the other hand, the uniform convergence of the functions $(y_1\mapsto p(y_1|Y_{-m:0},\theta^\star))_{m\in\mathbb{N}}$ and the finiteness of $\lambda(K)$ allow us to interchange the order of limits and integration. This implies that, $\mathbb{P}_\star$-almost surely, we have

$$\begin{aligned}\lim_{m\to+\infty}\mathbb{E}\left[\mathbb{1}_K(Y_1)|Y_{-m:0},\theta^\star\right]&=\lim_{m\to+\infty}\int\mathbb{1}_K(y_1)p(y_1|Y_{-m:0},\theta^\star)\lambda(dy_1)\\&=\int\mathbb{1}_K(y_1)\lim_{m\to+\infty}p(y_1|Y_{-m:0},\theta^\star)\lambda(dy_1)\\&=\int\mathbb{1}_K(y_1)p(y_1|Y_{-\infty:0},\theta^\star)\lambda(dy_1).\end{aligned} \tag{s26}$$

Combining (s25) and (s26) leads to

$$\mathbb{E}\left[\mathbb{1}_K(Y_1)|Y_{-\infty:0},\theta^\star\right]=\int\mathbb{1}_K(y_1)p(y_1|Y_{-\infty:0},\theta^\star)\lambda(dy_1),$$

for any event $K\subseteq\mathbb{Y}$ with $\lambda(K)<+\infty$, $\mathbb{P}_\star$-almost surely. This proves that, $\mathbb{P}_\star$-almost surely, $y_1\mapsto p(y_1|Y_{-\infty:0},\theta^\star)$ is the conditional density of Y_1 given $Y_{-\infty:0}$. Finally, we get $\log p(y_1|Y_{-\infty:0},\theta^\star)=\lim_{m\to+\infty}\log p(y_1|Y_{-m+1:0},\theta^\star)$ for all $y_1\in\mathbb{Y}$, $\mathbb{P}_\star$-almost surely, by applying Proposition 13.5 from Douc et al. (2014).

Under Assumption A2, the function $y_1\mapsto\log p(y_1|Y_{-m+1:0},\theta^\star)$ is $\mathbb{P}_\star$-almost surely twice differentiable for all $m\in\mathbb{N}$. $\mathbb{P}_\star$-almost surely, for all $y_1\in\mathbb{Y}$, the first derivative is

$$\frac{\partial\log p(y_1|Y_{-m+1:0},\theta^\star)}{\partial y_1}=\int\left(\frac{\partial\log g_{\theta^\star}(y_1|x_1)}{\partial y_1}\right)p(x_1|Y_{-m+1:0},\theta^\star)\,dx_1\,,$$

and the second derivative satisfies

$$\frac{\partial^2\log p(y_1|Y_{-m+1:0},\theta^\star)}{\partial y_1^2}=-\left(\frac{\partial\log p(y_1|Y_{-m+1:0},\theta^\star)}{\partial y_1}\right)^2+\int\left[\frac{\partial^2\log g_{\theta^\star}(y_1|x_1)}{\partial y_1^2}+\left(\frac{\partial\log g_{\theta^\star}(y_1|x_1)}{\partial y_1}\right)^2\right]p(x_1|Y_{-m+1:0},\theta^\star)\,dx_1.$$

We will prove the $\mathbb{P}_\star$-almost sure twice differentiability of $y_1\mapsto\log p(y_1|Y_{-\infty:0},\theta^\star)$ by proving that the sequences of derivatives $(y_1\mapsto\partial\log p(y_1|Y_{-m+1:0},\theta^\star)/\partial y_1)_{m\in\mathbb{N}}$ and $(y_1\mapsto\partial^2\log p(y_1|Y_{-m+1:0},\theta^\star)/\partial y_1^2)_{m\in\mathbb{N}}$ converge uniformly to well-defined limit functions, $\mathbb{P}_\star$-almost surely. Such uniform convergences imply the twice differentiability of the limit of $(y_1\mapsto\log p(y_1|Y_{-m+1:0},\theta^\star))_{m\in\mathbb{N}}$ by virtue of Theorem 7.17 from Rudin (1964).

From Condition C5 and (s16), we have, $\mathbb{P}_\star$-almost surely, for all $m \in \mathbb{N}$ and all $y_1 \in \mathbb{Y}$,

$$\left| \frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} - \frac{\partial \log p(y_1|Y_{-m:0}, \theta^\star)}{\partial y_1} \right| \leq c \, d_{TV} \left(p(dx_1|Y_{-m+1:0}, \theta^\star), p(dx_1|Y_{-m:0}, \theta^\star) \right) \leq c \, \rho^m .$$

As the upper bound does not depend on $y_1 \in \mathbb{Y}$, we have, $\mathbb{P}_\star$-almost surely, for all $m \in \mathbb{N}$,

$$\sup_{y_1 \in \mathbb{Y}} \left| \frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} - \frac{\partial \log p(y_1|Y_{-m:0}, \theta^\star)}{\partial y_1} \right| \leq c \, \rho^m ,$$

where $\rho \in (0, 1)$. By using the triangle inequality, we have, $\mathbb{P}_\star$-almost surely,

$$\begin{aligned} \sup_{y_1 \in \mathbb{Y}} \left| \sum_{k=m}^{+\infty} \left(\frac{\partial \log p(y_1|Y_{-k+1:0}, \theta^\star)}{\partial y_1} - \frac{\partial \log p(y_1|Y_{-k:0}, \theta^\star)}{\partial y_1} \right) \right| &\leq \sum_{k=m}^{+\infty} \sup_{y_1 \in \mathbb{Y}} \left| \frac{\partial \log p(y_1|Y_{-k+1:0}, \theta^\star)}{\partial y_1} - \frac{\partial \log p(y_1|Y_{-k:0}, \theta^\star)}{\partial y_1} \right| \\ &\leq c \sum_{k=m}^{+\infty} \rho^k \\ &\leq c \, \frac{\rho^m}{1-\rho} . \end{aligned}$$

Using telescopic sums, and $\rho^m \to 0$ when $m \to +\infty$ since $\rho \in (0, 1)$, we get

$$\sup_{y_1 \in \mathbb{Y}} \left| \frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} - \frac{\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1} \right| \xrightarrow[m \to +\infty]{\mathbb{P}_\star - a.s.} 0 ,$$

where

$$\frac{\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1} = \lim_{\substack{m \to +\infty \\ \mathbb{P}_\star\text{-a.s.}}} \frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} .$$

In other words, $\mathbb{P}_\star$-almost surely, the sequence of derivatives $(y_1 \mapsto \partial \log p(y_1|Y_{-m+1:0}, \theta^\star)/\partial y_1)_{m \in \mathbb{N}}$ converges uniformly to the function $y_1 \mapsto \partial \log p(y_1|Y_{-\infty:0}, \theta^\star)/\partial y_1$. Besides, we have proved earlier that the sequence of functions $(y_1 \mapsto \log p(y_1|Y_{-m+1:0}, \theta^\star))_{m \in \mathbb{N}}$ converges pointwise to the limit function $y_1 \mapsto \log p(y_1|Y_{-\infty:0}, \theta^\star)$. By using Theorem 7.17 from Rudin (1964), the limit function $y_1 \mapsto \log p(y_1|Y_{-\infty:0}, \theta^\star)$ is $\mathbb{P}_\star$-almost surely differentiable and its derivative is given $\mathbb{P}_\star$-almost surely by

$$\frac{\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1} = \lim_{m \to +\infty} \frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} .$$

Regarding the second derivative, we can follow the approach used to derive (s21) in the proof of Lemma 2, so that, $\mathbb{P}_\star$-almost surely, for all $m \in \mathbb{N}$ and all $y_1 \in \mathbb{Y}$, we have

$$\left| \left(\frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} \right)^2 - \left(\frac{\partial \log p(y_1|Y_{-m:0}, \theta^\star)}{\partial y_1} \right)^2 \right| \leq 2 \, c^2 \, \rho^m ,$$

By using again the triangle inequality, telescopic sums, and the fact that $\rho^m \to 0$ when $m \to +\infty$, we get

$$\sup_{y_1 \in \mathbb{Y}} \left| \sum_{k=m}^{+\infty} \left(\left(\frac{\partial \log p(y_1|Y_{-k+1:0}, \theta^\star)}{\partial y_1} \right)^2 - \left(\frac{\partial \log p(y_1|Y_{-k:0}, \theta^\star)}{\partial y_1} \right)^2 \right) \right| \xrightarrow[m \to +\infty]{\mathbb{P}_\star - a.s.} 0 ,$$

which implies that, $\mathbb{P}_\star$-almost surely, the sequence of functions $\left(y_1 \mapsto \left(\partial \log p(y_1|Y_{-k+1:0}, \theta^\star)/\partial y_1\right)^2 \right)_{m\in\mathbb{N}}$ converges uniformly to some limit function

$$y_1 \mapsto \lim_{m\to+\infty} \left(\frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} \right)^2 . \tag{s27}$$

By following again the derivation of (s21) in the proof of Lemma 2, we get, $\mathbb{P}_\star$-a.s., for all $m \in \mathbb{N}$ and all $y_1 \in \mathbb{Y}$,

$$\left| \int \left(\frac{\partial^2 \log g_{\theta^\star}(y_1|x_1)}{\partial y_1^2} + \left(\frac{\partial \log g_{\theta^\star}(y_1|x_1)}{\partial y_1} \right)^2 \right) \left(p(dx_1|Y_{-m+1:0}, \theta^\star) - p(dx_1|Y_{-m:0}, \theta^\star) \right) \right| \leq b\, \rho^m .$$

As previously, the triangle inequality, telescopic sums, and $\rho \in (0,1)$ imply that, $\mathbb{P}_\star$-almost surely, the sequence

$$\left(y_1 \mapsto \int \left(\frac{\partial^2 \log g_{\theta^\star}(y_1|x_1)}{\partial y_1^2} + \left(\frac{\partial \log g_{\theta^\star}(y_1|x_1)}{\partial y_1} \right)^2 \right) p(dx_1|Y_{-m+1:0}, \theta^\star) \right)_{m\in\mathbb{N}}$$

converges uniformly to some limit function

$$y_1 \mapsto \lim_{m\to+\infty} \int \left(\frac{\partial^2 \log g_{\theta^\star}(y_1|x_1)}{\partial y_1^2} + \left(\frac{\partial \log g_{\theta^\star}(y_1|x_1)}{\partial y_1} \right)^2 \right) p(dx_1|Y_{-m+1:0}, \theta^\star). \tag{s28}$$

Since a sum of two uniformly convergent sequences of functions is still uniformly convergent, with the limit function being the sum of the two limit functions, the previous results imply that the sequence of second derivatives $(y_1 \mapsto \partial^2 \log p(y_1|Y_{-m+1:0}, \theta^\star)/\partial y_1^2)_{m\in\mathbb{N}}$ converges uniformly to the function $y_1 \mapsto \partial^2 \log p(y_1|Y_{-\infty:0}, \theta^\star)/\partial y_1^2$ defined as the sum of the limit functions in (s27) and (s28), $\mathbb{P}_\star$-almost surely. By using again Theorem 7.17 from Rudin (1964), the function $y_1 \mapsto \log p(y_1|Y_{-\infty:0}, \theta^\star)$ is twice differentiable with second derivative equal to $y_1 \mapsto \partial^2 \log p(y_1|Y_{-\infty:0}, \theta^\star)/\partial y_1^2$, $\mathbb{P}_\star$-almost surely.

By (s12) and the previous results, we get, $\mathbb{P}_\star$-almost surely, for all $y_1 \in \mathbb{Y}$,

$$\begin{aligned}
\mathcal{H}(y_1, p(dy_1|Y_{-\infty:0}, \theta^\star)) &= \lim_{m\to+\infty} \mathcal{H}(y_1, p(dy_1|Y_{-m+1:0}, \theta^\star)) \\
&= \lim_{m\to+\infty} \left(2\, \frac{\partial^2 \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1^2} + \left(\frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} \right)^2 \right) \\
&= 2 \lim_{m\to+\infty} \left(\frac{\partial^2 \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1^2} \right) + \left(\lim_{m\to+\infty} \frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} \right)^2 \\
&= 2\, \frac{\partial^2 \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1^2} + \left(\frac{\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)}{\partial y_1} \right)^2 .
\end{aligned}$$

□

S7.4 Heuristic proof for well-specified nested i.i.d. models

In this section, we go back to the i.i.d. setting of Section S7.1.1, but we now assume that model M_1 is nested in model M_2, in the sense of Eq. (9) in Berger and Pericchi (1996). In other words, with $k_1, k_2 \in \mathbb{N}$ denoting the dimensions of the parameter spaces $\mathbb{T}_1$ and $\mathbb{T}_2$ with $k_2 > k_1 > 0$, we have $\mathbb{T}_2 = \{(\theta_1, \eta) \in \Xi_1 \times \Xi_2\} \subseteq \mathbb{R}^{k_1} \times \mathbb{R}^{k_2 - k_1}$

and $\mathbb{T}_1 \subseteq \Xi_1$, and there exists $\eta_1^\star \in \Xi_2$ such that for all $y \in \mathbb{Y}$ and all $\theta_1 \in \mathbb{T}_1$, we have $p_1(y|\theta_1) = p_2(y|\theta_1, \eta_1^\star)$. We also assume that both M_1 and M_2 are well-specified, so that there exists $\theta_1^\star \in \mathbb{T}_1$ such that, for all $y \in \mathbb{Y}$, we have $p_\star(y) = p_1(y|\theta_1^\star) = p_2(y|\theta_2^\star)$, where $\theta_2^\star = (\theta_1^\star, \eta_1^\star)$. For simplicity, we assume $\theta_1^\star$ is in the interior of $\mathbb{T}_1$. In this setting, the H-score is said to be consistent if it asymptotically chooses the model of smaller dimension, i.e. M_1. It is enough to show that $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) \to +\infty$ as $T \to +\infty$, in $\mathbb{P}_\star$-probability. The rest of this section is meant as a mere proof of concept with some heuristic arguments for this consistency to hold. We define

$$\Lambda_T(M_2, M_1) = \sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p_2(dy|\Theta_2)\right) \,\middle|\, Y_{1:t}\right] - \sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p_1(dy|\Theta_1)\right) \,\middle|\, Y_{1:t}\right],$$
$$\Delta_T(M_2, M_1) = \sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_2(Y_t|\Theta_2)}{\partial y_t}\,\middle|\, Y_{1:t}\right) - \sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_1(Y_t|\Theta_1)}{\partial y_t}\,\middle|\, Y_{1:t}\right),$$

so that we can write, using (5),

$$\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) = \Lambda_T(M_2, M_1) + \Delta_T(M_2, M_1). \tag{s29}$$

In the non-nested misspecified setting, the first term $\Lambda_T(M_2, M_1)$ would typically dominate and drive the difference $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)$, which would then behave asymptotically as $T\left(D_{\mathcal{H}}(p_\star, M_2) - D_{\mathcal{H}}(p_\star, M_1)\right)$. However, in the nested well-specified setting, we have $D_{\mathcal{H}}(p_\star, M_2) - D_{\mathcal{H}}(p_\star, M_1) = 0$, so that we would expect the second term $\Delta_T(M_2, M_1)$ to take over, and act as a penalty reflecting the difference in dimensions. This penalty term will dictate the asymptotic behavior of $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)$, provided that the first term $\Lambda_T(M_2, M_1)$ does not grow too fast (Assumption A13).

This can be informally motivated by the following arguments. Under appropriate regularity conditions, for each $j \in \{1, 2\}$, the posterior of Θ_j given $Y_{1:T}$ under model M_j concentrates in $\mathbb{P}_\star$-probability around the maximum likelihood estimator (MLE) defined as $\hat{\theta}_{j,T}^{\mathrm{MLE}} = \arg\min_{\theta \in \mathbb{T}_j} -\sum_{t=1}^{T} \log p_j(Y_t|\theta)$ (e.g. see Theorem 1.3.4 in Ghosh and Ramamoorthi, 2003). Under conditions ensuring the consistency of M-estimators (e.g. see Theorem 5.7 in Van der Vaart, 1998), the minimum H-score estimator (mHE) defined as $\hat{\theta}_{j,T}^{\mathrm{mHE}} = \arg\min_{\theta \in \mathbb{T}_j} \sum_{t=1}^{T} \mathcal{H}\left(Y_t, p_j(dy|\theta)\right)$ gets asymptotically close to the MLE, in the sense that $\lim_{T\to+\infty}(\hat{\theta}_{j,T}^{\mathrm{MLE}} - \hat{\theta}_{j,T}^{\mathrm{mHE}}) = (\theta_j^\star - \theta_j^\star) = 0$ in $\mathbb{P}_\star$-probability, where the limits of the MLE and the mHE coincide by well-specification and identifiability of the models, and strict propriety of the H-score. We can then write

$$\Lambda_T(M_2, M_1) = \sum_{t=1}^{T} \mathcal{H}\left(Y_t, p_2(dy|\hat{\theta}_{2,T}^{\mathrm{mHE}})\right) - \sum_{t=1}^{T} \mathcal{H}\left(Y_t, p_1(dy|\hat{\theta}_{1,T}^{\mathrm{mHE}})\right) + R_{1,T}^{\mathrm{MLE}} + R_{2,T}^{\mathrm{MLE}} + R_{1,T}^{\mathrm{mHE}} + R_{2,T}^{\mathrm{mHE}},$$

where the remainder terms for each $j \in \{1, 2\}$ are defined as

$$R_{j,T}^{\mathrm{MLE}} = \sum_{t=1}^{T} \left(\mathbb{E}\left[\mathcal{H}\left(Y_t, p_j(dy|\Theta_j)\right) \,\middle|\, Y_{1:t}\right] - \mathcal{H}\left(Y_t, p_j(dy|\hat{\theta}_{j,T}^{\mathrm{MLE}})\right)\right),$$
$$R_{j,T}^{\mathrm{mHE}} = \sum_{t=1}^{T} \left(\mathcal{H}\left(Y_t, p_j(dy|\hat{\theta}_{j,T}^{\mathrm{MLE}})\right) - \mathcal{H}\left(Y_t, p_j(dy|\hat{\theta}_{j,T}^{\mathrm{mHE}})\right)\right).$$

The term $R_{j,T}^{\mathrm{MLE}}$ is controlled by how fast the posterior of Θ_j concentrates around $\hat{\theta}_{j,T}^{\mathrm{MLE}}$ and how well the posterior expectation of $\mathcal{H}(Y_t, p_j(dy|\Theta_j)$ can be approximated by its expectation with respect to a Dirac mass at that MLE, which would typically require uniform integrability conditions for the posterior moments to converge. The term $R_{j,T}^{\mathrm{mHE}}$ is controlled by how fast $\hat{\theta}_{j,T}^{\mathrm{mHE}}$ approaches $\hat{\theta}_{j,T}^{\mathrm{MLE}}$ and how smooth the function $(y, \theta_j) \mapsto \mathcal{H}(y, p_j(dy|\theta_j))$ is. Using the Landau notation (i.e. small and big O), we assume that $R_{1,T}^{\mathrm{MLE}} + R_{2,T}^{\mathrm{MLE}} + R_{1,T}^{\mathrm{mHE}} + R_{2,T}^{\mathrm{mHE}} = o(\log T)$ as $T \to +\infty$, in $\mathbb{P}_\star$-probability, for ease of exposition. Finally, we can reasonably assume that

$$\sum_{t=1}^{T} \mathcal{H}\left(Y_t, p_2(dy|\hat{\theta}_{2,T}^{\mathrm{mHE}})\right) - \sum_{t=1}^{T} \mathcal{H}\left(Y_t, p_1(dy|\hat{\theta}_{1,T}^{\mathrm{mHE}})\right) = O(1) \tag{s30}$$

as $T \to +\infty$, in $\mathbb{P}_\star$-probability. Indeed, the difference between H-score minima is analogous to the difference between log-likelihood maxima $\sum_{t=1}^{T} \log p_1(Y_t|\hat{\theta}_{1,T}^{\mathrm{MLE}}) - \sum_{t=1}^{T} \log p_2(Y_t|\hat{\theta}_{2,T}^{\mathrm{MLE}})$ appearing when trying to prove the consistency of the log-Bayes factor (e.g. see Chib and Kuffner, 2016). Under suitable conditions, the difference between log-likelihood maxima converges in distribution to a scaled $\chi^2_{k_2-k_1}$ distribution (e.g. see Vuong, 1989). This is known as Wilks's theorem in the likelihood ratio test literature (Wilks, 1938), and its proof essentially relies on a Taylor expansion combined with the asymptotic Normality of the MLE. Under differentiability assumptions on the functions $\theta_j \mapsto \sum_{t=1}^{T} \mathcal{H}(Y_t, p_j(dy|\theta_j))$ and asymptotic Normality of the M-estimators $\hat{\theta}_{j,T}^{\mathrm{mHE}}$ (e.g. see Theorem 5.23 in Van der Vaart, 1998), we may assume that a similar distributional result holds for the difference of H-score minima, leading to (s30). All the previous heuristic arguments motivate Assumption A13 stated below.

Assumption A13. *$\Lambda_T(M_2, M_1) = o(\log T)$ as $T \to +\infty$, in $\mathbb{P}_\star$-probability.*

Looking now at the penalty term $\Delta_T(M_2, M_1)$, we can define, for each $j \in \{1, 2\}$, the function

$$\begin{aligned} G_j \ : \ \mathbb{Y} \times \mathbb{T}_j &\longrightarrow \mathbb{R} \\ (y_t, \theta_j) &\longmapsto \frac{\partial \log p_j(y_t|\theta_j)}{\partial y_t}. \end{aligned}$$

Under conditions enabling the Bernstein-von Mises theorem to hold (e.g. see Theorem 1.4.2 in Ghosh and Ramamoorthi, 2003), as $t \to +\infty$, the posterior distribution of $\sqrt{t}(\Theta_{j,t} - \hat{\theta}_{j,t}^{\mathrm{MLE}})$ would approach a Normal distribution with mean 0 and covariance matrix $V_j^\star = I_j(\theta_j^\star)^{-1}$, where $I_j(\theta_j^\star) = \mathbb{E}_\star[-\nabla^2_{\theta_j} \log p_j(Y|\theta_j^\star)]$ corresponds to the Fisher information matrix at $\theta_j^\star$ when M_j is well-specified. Provided that the posterior second moments converge and G_j is differentiable with respect to θ_j such that $\nabla_{\theta_j} G_j(y_t, \theta_j^\star) \neq 0$ for all $y_t \in \mathbb{Y}$, applying the Delta method to the previous statement motivates Assumption A14.

Assumption A14. $\mathrm{Var}\left(\frac{\partial \log p_j(Y_t|\Theta_j)}{\partial y_t} \middle| Y_{1:t}\right) = \nabla_{\theta_j} G_j(Y_t, \theta_j^\star)^\top \frac{V_j^\star}{t} \nabla_{\theta_j} G_j(Y_t, \theta_j^\star) + o\left(\frac{1}{t}\right)$ *as $t \to +\infty$, $\mathbb{P}_\star$-almost surely, for each $j \in \{1, 2\}$.*

Using $\sum_{t=1}^{T}(1/t) = O(\log T)$ as $T \to +\infty$, and Assumption A14, we have, $\mathbb{P}_\star$-almost surely,

$$\Delta_T(M_2, M_1) = \sum_{t=1}^{T} \frac{1}{t} K_t + o(\log T), \tag{s31}$$

where $K_t = \left(\nabla_{\theta_2} G_2(Y_t, \theta_2^\star)^\top V_2^\star \nabla_{\theta_2} G_2(Y_t, \theta_2^\star) - \nabla_{\theta_1} G_1(Y_t, \theta_1^\star)^\top V_1^\star \nabla_{\theta_1} G_1(Y_t, \theta_1^\star)\right)$. The K_t's are i.i.d., so that if we

assume $\mathbb{E}_\star[K_t^2] < +\infty$ and define $\sigma^2 = \mathrm{Var}_\star(K_t)$, then we have

$$\mathrm{Var}_\star\left(\sum_{t=1}^{T}\frac{1}{t}K_t\right) = \sigma^2\sum_{t=1}^{T}\frac{1}{t^2} \xrightarrow[T\to+\infty]{} \sigma^2\frac{\pi^2}{6} < +\infty. \tag{s32}$$

On the other hand, if we define $\delta_{21} = \mathbb{E}_\star[K_t]$, then we have

$$\mathbb{E}_\star\left(\sum_{t=1}^{T}\frac{1}{t}K_t\right) = \delta_{21}\sum_{t=1}^{T}\frac{1}{t} = \delta_{21}\log T + o(\log T). \tag{s33}$$

Using Chebyshev's inequality, combining (s31), (s32), and (s33) leads to

$$\Delta_T(M_2, M_1) = \delta_{21}\log T + o(\log T),$$

in $\mathbb{P}_\star$-probability. Piecing everything together, we finally get

$$\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) = \delta_{21}\log T + o(\log T), \tag{s34}$$

in $\mathbb{P}_\star$-probability, where

$$\delta_{21} = \mathbb{E}_\star\left[\nabla_{\theta_2}G_2(Y,\theta_2^\star)^\top V_2^\star\nabla_{\theta_2}G_2(Y,\theta_2^\star) - \nabla_{\theta_1}G_1(Y,\theta_1^\star)^\top V_1^\star\nabla_{\theta_1}G_1(Y,\theta_1^\star)\right]. \tag{s35}$$

This implies that as $T\to+\infty$, the sign of $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)$ will be determined by the sign of δ_{21}. If $\theta_2 = (\theta_1, \eta)$ consists of orthogonal parameters, in the sense that $\mathbb{E}_\star[\nabla_\eta\nabla_{\theta_1}\log p_2(Y|\theta_1^\star,\eta_1^\star)] = 0$ so that the Fisher information matrix $I_2(\theta_2^\star)$ is block-diagonal, then $V_2^\star$ has the following block-diagonal structure

$$V_2^\star = \left(\begin{array}{c|c} V_1^\star & 0 \\ \hline 0 & V_{22}^\star \end{array}\right), \tag{s36}$$

where $V_{22}^\star = \mathbb{E}_\star[-\nabla_\eta^2\log p_2(Y|\theta_2^\star)]^{-1}$ is a symmetric positive definite matrix of size $(k_2 - k_1)$. Using (s36) and the fact that $\nabla_{\theta_2}G_2(y_t,\theta_2^\star)^\top = (\nabla_{\theta_1}G_1(y_t,\theta_1^\star)^\top, \nabla_\eta G_2(y_t,\theta_2^\star)^\top)$, we get $\delta_{21} = \mathbb{E}_\star[\nabla_\eta G_2(Y,\theta_2^\star)^\top V_{22}^\star\nabla_\eta G_2(Y,\theta_2^\star)] > 0$, where the positivity comes from the positive definiteness of $V_{22}^\star$. In other words, (s35) becomes

$$\delta_{21} = \mathbb{E}_\star\left[\left(\nabla_\eta\frac{\partial\log p_2(Y|\theta_2^\star)}{\partial y}\right)^\top \mathbb{E}_\star[-\nabla_\eta^2\log p_2(Y|\theta_2^\star)]^{-1}\left(\nabla_\eta\frac{\partial\log p_2(Y|\theta_2^\star)}{\partial y}\right)\right] > 0. \tag{s37}$$

This implies $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) \xrightarrow[T\to+\infty]{} +\infty$, so that we asymptotically choose the smaller model M_1, as desired.

S7.5 Numerical illustration with nested i.i.d. Normal models

We consider the following nested Normal models

$$M_1: \quad Y_1, ..., Y_T \,|\, \mu_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu_1, 1\right), \quad \mu_1 \sim \mathcal{N}\left(0, \sigma_0^2\right),$$
$$M_2: \quad Y_1, ..., Y_T \,|\, \mu_2, \sigma_2^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu_2, \sigma_2^2\right), \quad \mu_2 \,|\, \sigma_2^2 \sim \mathcal{N}\left(0, \sigma_2^2\right), \quad \sigma_2^2 \sim \text{Inv-}\chi^2\left(\nu_0, s_0^2\right).$$

The positive hyperparameters are chosen as $\sigma_0^2 = 10$, $\nu_0 = 0.1$, and $s_0^2 = 1$. We compare M_1 and M_2, using data generated as $Y_1, ..., Y_T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_\star, \sigma_\star^2)$, in the following settings: (1) $(\mu_\star, \sigma_\star^2) = (0, 5)$, i.e. M_2 is well-specified while M_1 is not; (2) $(\mu_\star, \sigma_\star^2) = (0, 1)$, i.e. both M_1 and M_2 are well-specified. For each case, we generate $T = 10^5$ observations and perform 5 runs of SMC with $N_\theta = 1024$ particles to estimate the log-Bayes factors and H-factors of M_1 against M_2. Each run averages the factors over 100 different orderings of the data, sampled uniformly from all the possible permutations. The results are shown in Figure 9.

In case 1, both factors correctly select the well-specified model M_2, as expected. In case 2, M_1 is nested in M_2 with respective dimensions $k_1 = 1$ and $k_2 = 2$. Besides, the Fisher information under the Normal model M_2 is diagonal. Using the same notation as in (s37), we get $\theta_2^\star = (0, 1)$ and $\delta_{21} = 2$, so that our postulated result in (s34) becomes $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) = 2 \log T + o(\log T)$. Regarding the log-Bayes factor, standard approximation via the BIC yields $\log p_2(Y_{1:T}) - \log p_1(Y_{1:T}) = \left(\frac{k_2 - k_1}{2}\right) \log T + o(\log T) = (1/2) \log T + o(\log T)$.

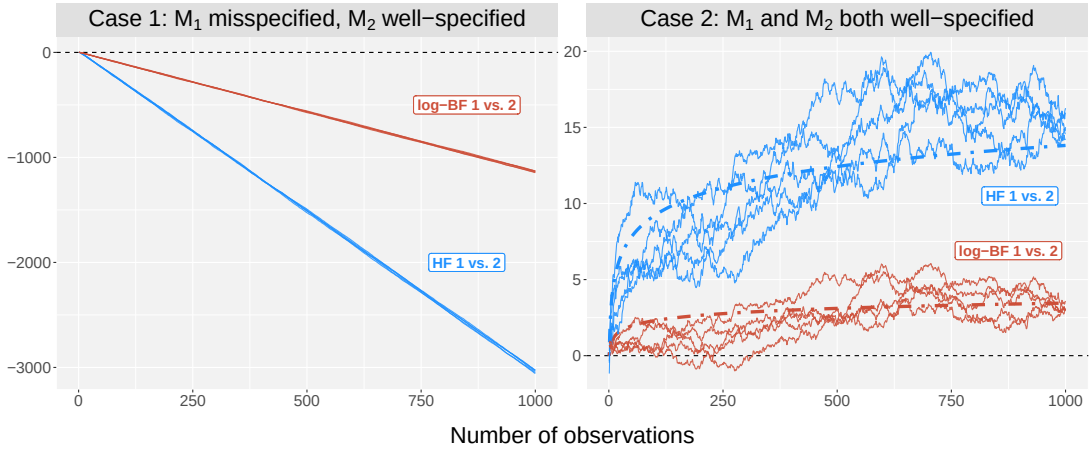


Figure 9. *Estimated log-Bayes factors (log-BF) and H-factors (HF) of M_1 against M_2, computed for 5 replications (thin solid lines), under two i.i.d. data-generating processes: $\mathcal{N}(0, 5)$ (Case 1) and $\mathcal{N}(0, 1)$ (Case 2). Each run averages the factors over 100 random orderings of the data. The variability within each factor is due to these random orderings and Monte Carlo error. In case 1 (left panel), this error is negligible relative to the magnitude of the factors, which makes the 5 replications hard to distinguish. In case 2 (right panel), the dot-dashed lines indicate the theoretical logarithmic growth of each factor: they correspond respectively to $t \mapsto (1/2) \log(t)$ (log-BF) and $t \mapsto 2 \log(t)$ (HF). See Section S7.5.*

References

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91 (433), 109–122.

Billingsley, P. (1968). *Convergence of probability measures*. Wiley Series in Probability and Statistics.

Billingsley, P. (1995). *Probability and measure*. Wiley Series in Probability and Mathematical Statistics.

Chib, S. and Kuffner, T. A. (2016). Bayes factor consistency. *Preprint*, arXiv:1607.00292.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89 (3), 539–552.

Chopin, N., Jacob, P. E. and Papaspiliopoulos, O. (2013). SMC^2: an efficient algorithm for sequential analysis of state-space models. *Journal of the Royal Statistical Society: Series B*, 75 (3), 397–426.

Chopin, N., Ridgway, J., Gerber, M. and Papaspiliopoulos, O. (2015). Towards automatic calibration of the number of state particles within the SMC^2 algorithm. *Preprint*, arXiv:1506.00570.

Dawid, A. P., Lauritzen, S. and Parry, M. (2012). Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40 (1), 593–608.

Dawid, A. P. and Musio, M. (2015). Bayesian model selection based on proper scoring rules. *Bayesian Analysis*, 10 (2), 479–499.

De Gunst, M. and Shcherbakova, O. (2008). Asymptotic behavior of Bayes estimators for hidden Markov models with application to ion channels. *Mathematical Methods of Statistics*, 17 (4), 342–356.

Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68 (3), 411—436.

Douc, R., Moulines, E., Olsson, J., Van Handel, R. et al. (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 Stoffer, D. (2014). *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. 1st edn. Chapman and Hall/CRC.

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

Douc, R., Olsson, J. and Roueff, F. (2016). Posterior consistency for partially observed Markov models. *Preprint*, arXiv:1608.06851.

Duan, J.-C. and Fulop, A. (2015). Density-tempered marginalized sequential Monte Carlo samplers. *Journal of Business & Economic Statistics*, 33 (2), 192–202.

Gassiat, E., Rousseau, J. et al. (2014). About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli*, 20 (4), 2039–2075.

Gerber, M., Chopin, N. and Whiteley, N. (2017). Negative association, ordering and convergence of resampling methods. *Preprint*, arXiv:1707.01845.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. Springer Science & Business Media.

Golightly, A. and Kypraios, T. (2017). Efficient SMC^2 schemes for stochastic kinetic models. *Preprint*, arXiv:1704.02791.

Hendrickson, A. D. and Buehler, R. J. (1971). Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42 (6), 1916–1921.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1 (1), 221–233.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6, 695–709.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pp. 887–906.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related results. *University of California publications in Statistics*, 1, 277–330.

Lijoi, A., Prünster, I. and Walker, S. G. (2007). Bayesian consistency for stationary models. *Econometric Theory*, 23 (4), 749–759.

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America*, 42 (9), 654–655.

Perlman, M. D. et al. (1972). On the strong consistency of approximate maximum likelihood estimators. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Theory of Statistics.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94 (446), 590–599.

Rudin, W. (1964). *Principles of mathematical analysis*. McGraw–Hill.

Shalizi, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, 3, 1039–1074.

Van der Vaart, A. W. (1998). *Asymptotic statistics*. Vol. 3. Cambridge University Press.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pp. 307–333.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20 (4), 595–601.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9 (1), 60–62.