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

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

Contents

*Department of Statistics, Harvard University. Emails: stephaneshao@g.harvard.edu, pjacob@fas.harvard.edu.
†School of Engineering and Applied Sciences, Harvard University. Emails: jieding@fas.harvard.edu, vahid@seas.harvard.edu.

S1 Identities for the H-score

In this section, we fix a candidate model M and drop the dependence on the model in the notation.

S1.1 Proof of Eq. (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 Eq. (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

$$\mathcal{H}(y, p) = \sum_{k=1}^{d_y} \left(\mathbb{E}\left[2\, \frac{\partial^2 \log p(y|\Theta)}{\partial {y_{(k)}}^2} + 2 \left(\frac{\partial \log p(y|\Theta)}{\partial y_{(k)}} \right)^2 \middle| y \right] - \left(\mathbb{E}\left[\frac{\partial \log p(y|\Theta)}{\partial y_{(k)}} \middle| y \right] \right)^2 \right). \tag{s1}$$

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

$$p(y_t|y_{1:t-1}) = \int_{\mathbb{T}} p(y_t|y_{1:t-1}, \theta)\, p(\theta|y_{1:t-1}).\, d\theta \tag{s2}$$

Therefore, under Assumption A1, we can apply (s1) to (s2) for each term of the sum in Eq. (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 Eq. (4).

□

S1.2 Proof of Proposition 1

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

$$\frac{p(x_t|y_{1:t-1}, \theta)\, g_\theta(y_t|x_t)}{p(y_t|y_{1:t-1}, \theta)} = \frac{p(x_t, y_t|y_{1:t-1}, \theta)}{p(y_t|y_{1:t-1}, \theta)} = p(x_t|y_{1:t}, \theta). \tag{s3}$$

This proves Eq. (5).

Regarding Eq. (6), we proceed similarly and have, for all $k \in \{1, ..., d_y\}$,

$$\frac{\partial^2 \log p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}^2} = -\left(\frac{\partial \log p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}} \right)^2 + \frac{1}{p(y_t|y_{1:t-1}, \theta)} \frac{\partial^2\, p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}^2} \,. \tag{s4}$$

The second term can be rewritten as

$$\frac{1}{p(y_t|y_{1:t-1}, \theta)} \frac{\partial^2\, p(y_t|y_{1:t-1}, \theta)}{\partial y_{t_{(k)}}^2} = \frac{1}{p(y_t|y_{1:t-1}, \theta)} \int p(x_t|y_{1:t-1}, \theta) \left(\frac{\partial^2\, g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}^2} \right) dx_t \,, \tag{s5}$$

where the integrand can be written as

$$\frac{\partial^2\, g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}^2} = g_\theta(y_t|x_t) \left(\frac{\partial^2 \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}^2} + \left(\frac{\partial \log g_\theta(y_t|x_t)}{\partial y_{t_{(k)}}} \right)^2 \right). \tag{s6}$$

By plugging (s6) into (s5) and using again Eq. (s3), 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 (s4), 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 Eq. (6).

□

S2 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 $\{(y_1, y_2) \in \mathbb{Y}^2 : y_1 - y_2 \in \{-2e_k, -e_k, e_k, 2e_k\}$ for some $k \in [\![1, d_y]\!]\}$. 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

$$\mathcal{E}_{\mathcal{H}^D}(p) = -\sum_{k=1}^{d_y} \sum_{\substack{y \in \mathbb{Y} \text{ s.t.} \\ a_k < y_{(k)} < b_k}} p(y) \left(\frac{p(y+e_k) - p(y-e_k)}{2\,p(y)}\right)^2, \tag{s7}$$

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 Eq. (s7) 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 $u \mapsto -(u-1)^2$.

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.

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 1. 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.

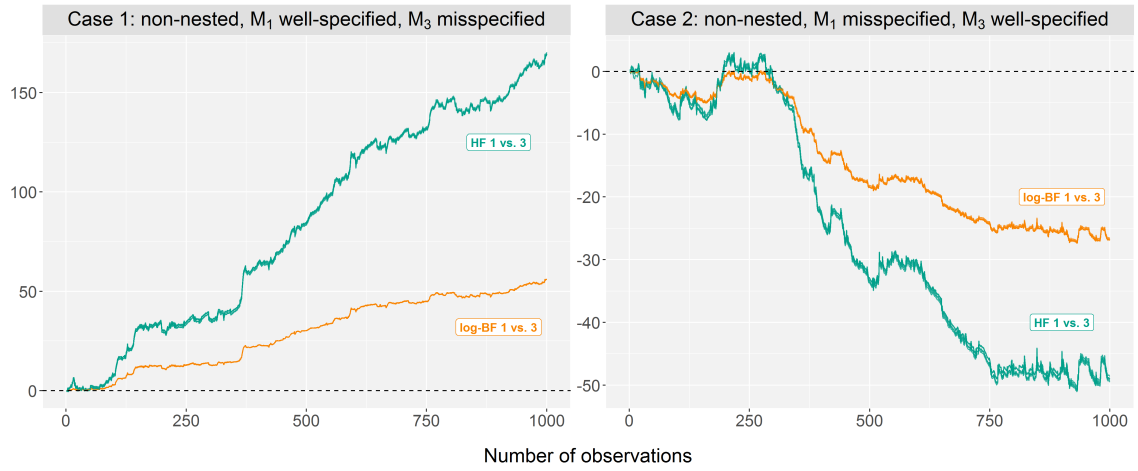

Figure 1. *Estimated log-Bayes factors (log-BF, orange) and H-factors (HF, green) of M_1 against M_3, computed for 5 replications (thin solid lines), under two data-generating processes: AR(1) (Case 1) and MA(1) (Case 2). See Section S3.*

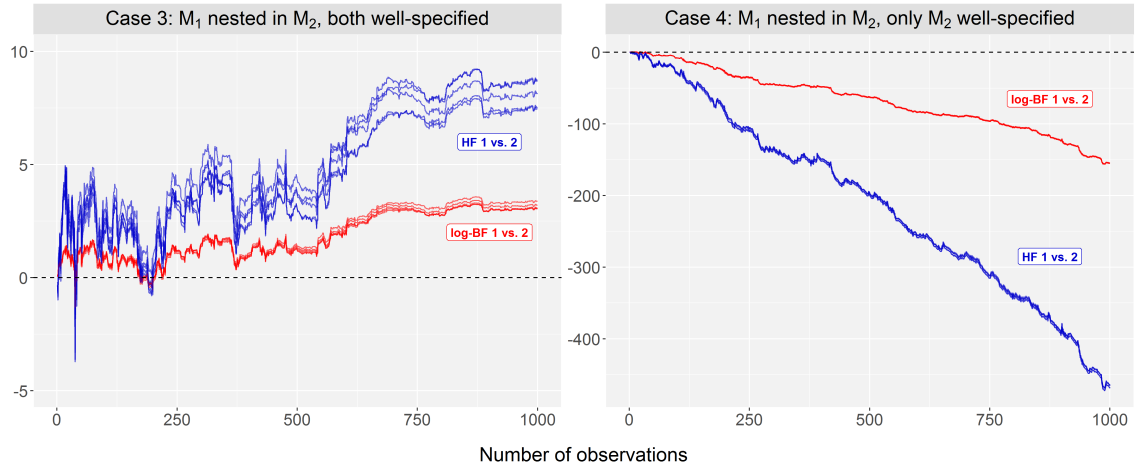

Figure 2. *Estimated log-Bayes factors (log-BF, red) and H-factors (HF, blue) of M_1 against M_2, computed for 5 replications (thin solid lines), under two data-generating processes: AR(1) (Case 3) and AR(2) (Case 4). See Section S3.*

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 2. Case 3 suggests that, when dealing with nested well-specified models, the H-factor asymptotically favors the model of smallest dimension.

S4 Applications: posterior density plots

S4.1 Diffusion models for population dynamics of red kangaroos

This section complements the numerical example presented in Section 6.1 of the main paper. For each population model M_1, M_2, and M_3, the respective posterior densities of the parameters are estimated via SMC^2 across 5 replications. The marginal posterior densities are shown in Figures 3, 4, and 5. 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 concentration of the posterior distributions may be a reasonable assumption, even when the strong conditions of Appendix C.2 are not met.

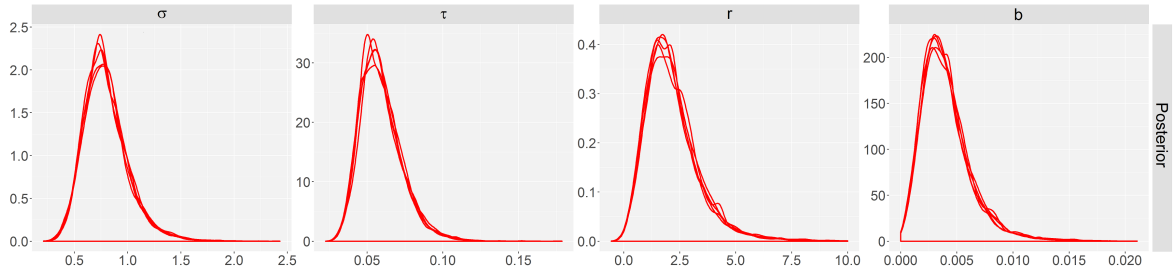

Figure 3. *Estimated marginal posterior densities of* (σ, τ, r, b) *under model* M_1*, given 41 observations, with independent priors* σ, τ, $b \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 10)$ *and* $r \sim \text{Unif}(-10, 10)$*, plotted for 5 replications (solid lines). See Section S4.1.*

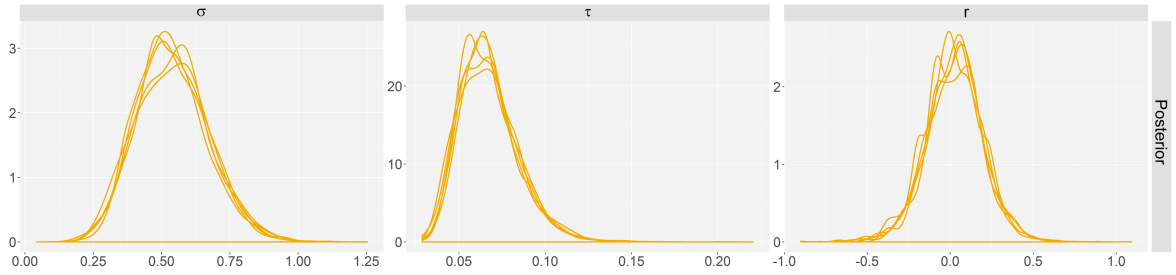

Figure 4. *Estimated marginal posterior densities of* (σ, τ, r) *under model* M_2*, given 41 observations, with independent priors* σ, $\tau \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 10)$ *and* $r \sim \text{Unif}(-10, 10)$*, plotted for 5 replications (solid lines). See Section S4.1.*

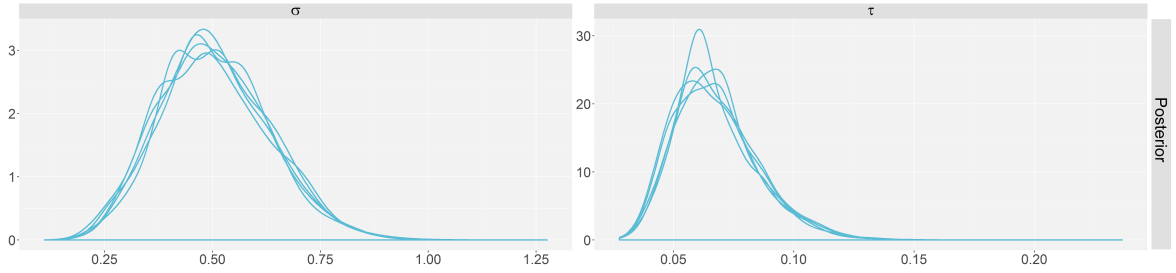

Figure 5. *Estimated marginal posterior densities of* (σ, τ) *under model* M_3*, given 41 observations, with independent priors* σ, $\tau \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 10)$*, plotted for 5 replications (solid lines). See Section S4.1.*

S4.2 Lévy-driven stochastic volatility models

This section complements the numerical example presented in Section 6.2 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 6 and 7, 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.

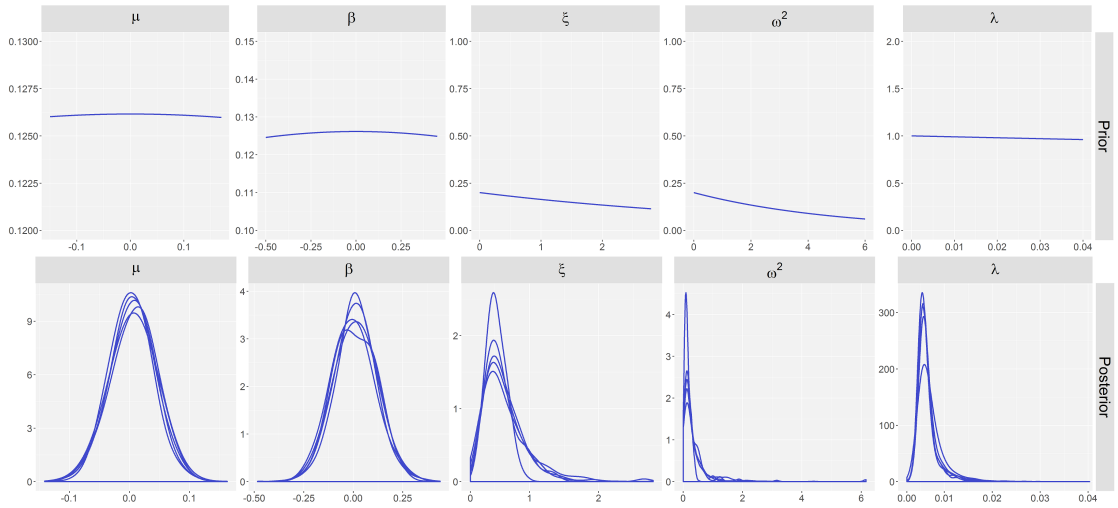

Figure 6. *Top panels: marginal prior densities* $\mu, \beta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 10)$; $\xi, \omega^2 \overset{\text{i.i.d.}}{\sim} \text{Exp}(1/5)$; $\lambda \sim \text{Exp}(1)$, *plotted over the support of the posterior. Bottom panels: estimated marginal posterior densities under model* M_1, *given 1000 observations, plotted for 5 replications (solid lines). See Section S4.2.*

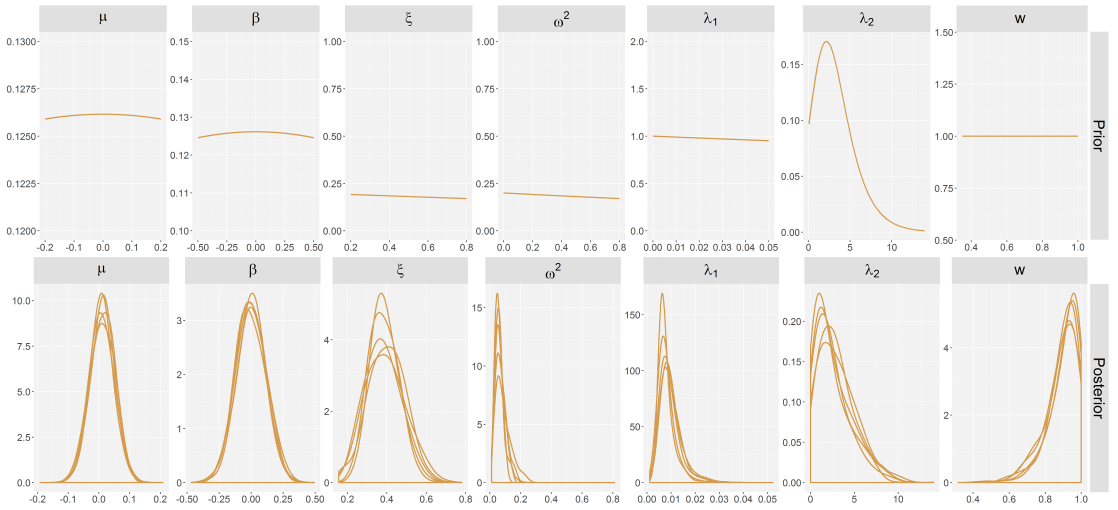

Figure 7. *Top panels: marginal prior densities* $\mu, \beta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 10)$; $\xi, \omega^2 \overset{\text{i.i.d.}}{\sim} \text{Exp}(1/5)$; $\lambda_1 \sim \text{Exp}(1)$; $\lambda_2 - \lambda_1 \sim \text{Exp}(1/2)$; $\text{w} \sim \text{Unif}(0, 1)$, *plotted over 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 Consistency of the H-score

S5.1 Proof of Eq. (C.1.4)

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

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

□

S5.2 Proof of Eq. (C.1.5)

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

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

□

S5.3 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

$$\Theta_t \xrightarrow[t\to+\infty]{\mathcal{D}} \theta^\star.$$

$\mathbb{T}$ 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 $(y_t)_{t\in\mathbb{N}^*}$. $\mathbb{P}$-almost surely, we have, 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

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

Therefore, we have

$$\mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \Theta'_t)\right) - \mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \theta^\star)\right) \quad \xrightarrow[t\to+\infty]{\mathbb{P}-a.s.} \quad 0. \tag{s8}$$

Similarly, using C2(b), we get

$$\frac{\partial \log p(y_t|y_{1:t-1}, \Theta'_t)}{\partial y_t} - \frac{\partial \log p(y_t|y_{1:t-1}, \theta^\star)}{\partial y_t} \quad \xrightarrow[t\to+\infty]{\mathbb{P}-a.s.} \quad 0. \tag{s9}$$

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 (s8) implies the convergence of the first moments (e.g. see Theorem 25.12 in Billingsley, 1995). In other words, we get

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

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

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

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

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

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 (s9) implies the convergence of the first two moments, and a fortiori the convergence of the variance. Thus,

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

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,

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

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

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

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

□

S5.4 Proof of Lemma 2

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

$$2\int\left[\frac{\partial^2 \log g_{\theta^\star}(y_t|x_t)}{\partial y_t^2} + \left(\frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t}\right)^2\right] p(dx_t|y_{-m+1:t}, \theta^\star) + \left(\int \frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t} p(dx_t|y_{-m+1:t}, \theta^\star)\right)^2. \tag{s10}$$

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

$$\begin{aligned}
&|\mathcal{H}(y_t, p(dy_t|y_{-m+1:t-1}, \theta^\star)) - \mathcal{H}(y_t, p(dy_t|y_{-m:t-1}, \theta^\star))| \\
&\leq 2\left|\int\left[\frac{\partial^2 \log g_{\theta^\star}(y_t|x_t)}{\partial y_t^2} + \left(\frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t}\right)^2\right](p(dx_t|y_{-m+1:t}, \theta^\star) - p(dx_t|y_{-m:t}, \theta^\star))\right| \\
&\qquad + \left|\left(\int \frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t} p(dx_t|y_{-m+1:t}, \theta^\star)\right)^2 - \left(\int \frac{\partial \log g_{\theta^\star}(y_t|x_t)}{\partial y_t} p(dx_t|y_{-m+1:t}, \theta^\star)\right)^2\right| \\
&\leq 2\,b\left|\int (p(dx_t|y_{-m+1:t}, \theta^\star) - p(dx_t|y_{-m:t}, \theta^\star))\right| \\
&\qquad + c^2\left|\int \left(p(dx_t|y_{-m+1:t}, \theta^\star) - p(dx_t|y_{-m:t}, \theta^\star)\right)\right|\left|\int \left(p(dx_t|y_{-m+1:t}, \theta^\star) + p(dx_t|y_{-m:t}, \theta^\star)\right)\right| \\
&\leq 2\,b\,d_{TV}\left(p(dx_t|y_{-m+1:t}, \theta^\star), p(dx_t|y_{-m:t}, \theta^\star)\right) + 2\,c^2\,d_{TV}\left(p(dx_t|y_{-m+1:t}, \theta^\star), p(dx_t|y_{-m:t}, \theta^\star)\right) \\
&\leq 2\,(b + c^2)\,d_{TV}\left(p(dx_t|y_{-m+1:t}, \theta^\star), p(dx_t|y_{-m:t}, \theta^\star)\right) \\
&\leq 2\,(b + c^2)\,\rho^{t+m-1},
\end{aligned} \tag{s11}$$

where the last inequality comes from Eq. (C.2.1) under Condition C4. This proves Eq. (C.2.2). From Eq. (s10) and Condition C5, the triangular inequality and the fact that probability densities integrate to 1 yield Eq. (C.2.3).

□

S5.5 Proof of Lemma 3

We closely follow the proof of Lemma 13.12 in Douc, Moulines and Stoffer (2014). We have

$$p(x_1|Y_{-m+1:0}, \theta^\star) = \int \nu_{\theta^\star}(x_1|x_0)\, p(dx_0|Y_{-m+1:0}, \theta^\star), \tag{s12}$$

for all $x_1 \in \mathbb{X}$ and all $m \in \mathbb{N}^*$, $\mathbb{P}_\star$-almost surely. By Condition C4 and Eq. (C.2.1), 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 doesn't 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

$$\mathbb{P}_\star\left(\sum_{m=1}^{+\infty} \sup_{x_1 \in \mathbb{X}} |p(x_1|Y_{-m+1:0}, \theta^\star) - p(x_1|Y_{-m:0}, \theta^\star)| < +\infty\right) = 1. \tag{s13}$$

For any $\varepsilon > 0$, the convergence of the series in (s13) 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, Eq. (s12) 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

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

These two inequalities hold $\mathbb{P}_\star$-almost surely, and lead to

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

which proves 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 Eq. (C.2.1), 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

$$\mathbb{P}_\star \left(\sum_{m=1}^{+\infty} \sup_{y_1 \in \mathbb{Y}} |p(y_1|Y_{-m+1:0}, \theta^\star) - p(y_1|Y_{-m:0}, \theta^\star)| < +\infty \right) = 1, \tag{s14}$$

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,

$$\mathbb{E}\left[\mathbb{1}_K(Y_1)|Y_{-\infty:0}, \theta^\star\right] = \lim_{m\to+\infty} \mathbb{E}\left[\mathbb{1}_K(Y_1)|Y_{-m:0}, \theta^\star\right]. \tag{s15}$$

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

$$\begin{aligned}
\lim_{m\to+\infty} \mathbb{E}\left[\mathbb{1}_K(Y_1)|Y_{-m:0}, \theta^\star\right] &= \lim_{m\to+\infty} \int \mathbb{1}_K(y_1) p(y_1|Y_{-m:0}, \theta^\star)\lambda(dy_1) \\
&= \int \mathbb{1}_K(y_1) \lim_{m\to+\infty} p(y_1|Y_{-m:0}, \theta^\star)\lambda(dy_1) \\
&= \int \mathbb{1}_K(y_1) p(y_1|Y_{-\infty:0}, \theta^\star)\lambda(dy_1). \qquad \text{(s16)}
\end{aligned}$$

Combining Eqs. (s15) and (s16) 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 Eq. (C.2.1), 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,

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

$$\leq c \sum_{k=m}^{+\infty} \rho^k$$
$$\leq c \, \frac{\rho^m}{1-\rho}.$$

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 Eq. (s11) 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

$$y_1 \mapsto \lim_{m \to +\infty} \left(\frac{\partial \log p(y_1|Y_{-m+1:0}, \theta^\star)}{\partial y_1} \right)^2 . \tag{s17}$$

By following again the derivation of Eq. (s11) 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

$$y_1 \mapsto \lim_{m \to +\infty} \int \left(\frac{\partial^2 \log g_{\theta^\star}(y_1|x_1)}{\partial y_1^2} + \left(\frac{\partial \log g_{\theta^\star}(y_1|x_1)}{\partial y_1} \right)^2 \right) p(dx_1|Y_{-m+1:0}, \theta^\star). \tag{s18}$$

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 (s17) and (s18), $\mathbb{P}_\star$-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 Eq. (C.1.4) and the previous results, we get, $\mathbb{P}_\star$-almost surely, for all $y_1 \in \mathbb{Y}$,

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

□

References

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.

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.

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

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.

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

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.

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