JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2018.1518237

Taylor & Francis
Taylor & Francis Group

Bayesian Model Comparison with the Hyvärinen Score: Computation and Consistency

Stephane Shao[a], Pierre E. Jacob[a], Jie Ding[b], and Vahid Tarokh[c]

[a]Department of Statistics, Harvard University, Cambridge, MA; [b]School of Statistics, University of Minnesota; [c]Department of Electrical and Computer Engineering, Duke University

ABSTRACT

The Bayes factor is a widely used criterion in model comparison and its logarithm is a difference of out-of-sample predictive scores under the logarithmic scoring rule. However, when some of the candidate models involve vague priors on their parameters, the log-Bayes factor features an arbitrary additive constant that hinders its interpretation. As an alternative, we consider model comparison using the Hyvärinen score. We propose a method to consistently estimate this score for parametric models, using sequential Monte Carlo methods. We show that this score can be estimated for models with tractable likelihoods as well as nonlinear non-Gaussian state-space models with intractable likelihoods. We prove the asymptotic consistency of this new model selection criterion under strong regularity assumptions in the case of nonnested models, and we provide qualitative insights for the nested case. We also use existing characterizations of proper scoring rules on discrete spaces to extend the Hyvärinen score to discrete observations. Our numerical illustrations include Lévy-driven stochastic volatility models and diffusion models for population dynamics. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received November 2017
Revised August 2018

KEYWORDS

Bayes factor; Noninformative prior; Model selection; SMC; State-space model

1. Introduction

1.1. Bayesian Model Comparison

Bayesian model comparison is challenging in situations where the candidate models involve either vague or improper prior distributions on some of their parameters. The Bayes factor (Jeffreys 1939) between two models—defined as the ratio of their marginal likelihoods—is a widely used approach to model comparison. If one of the candidate models includes the data-generating process, that model is termed well-specified or correct, and the Bayes factor can be interpreted as a ratio of odds, which updates the relative probabilities of the models being correct. In the misspecified or M-open setting (Bernardo and Smith 2000), the marginal log-likelihood can be interpreted as a measure of out-of-sample predictive performance assessed with the logarithmic scoring rule (e.g., Kass and Raftery 1995; Key, Pericchi, and Smith 1999; Bernardo and Smith 2000). Scoring rules are loss functions for the task of predicting an observation y with a probability distribution p, and the logarithmic scoring rule quantifies predictive performance with $-\log p(y)$. Under regularity conditions, the Bayes factor leads to consistent model selection as the number of observations goes to infinity (e.g., Dawid 2011; Lee and MacEachern 2011; Walker 2013; Chib and Kuffner 2016).

However, if any of the models involves either vague or improper prior distributions on their parameters, the Bayes factor can take arbitrary values and becomes unreliable for any fixed sample size. This is problematic as vague priors are extensively used in practice, for instance when uniform distributions are specified on intervals of plausible values (e.g., Knape and de Valpine 2012, sec. 4.2). Improper priors also arise from theoretical considerations, for instance as Jeffreys priors (e.g., Robert 2007, chap. 3). Our article takes the use of such priors by practitioners as a starting point, and addresses the question of model comparison in this context where one cannot rely on the Bayes factor. This limitation of the Bayes factor, sometimes referred to as Bartlett's paradox (Bartlett 1957; Kass and Raftery 1995), is a long-lasting challenge in Bayesian model comparison (Chapter 7 of Robert 2007), as it seems to suggest that prior specification should take into account the potential use (or misuse) of Bayes factors. Many approaches have been proposed to tackle this issue, either by modifying the Bayes factor (e.g., O'Hagan 1995; Berger and Pericchi 1996; Berger, Pericchi, and Varshavsky 1998; Berger and Pericchi 2001) or bypassing it altogether (e.g., Kamary et al. 2014, and references therein). In this article, we investigate an alternative criterion, that is, (1) principled for any sample size, thanks to an interpretation in terms of predictive performance and scoring rules, (2) enjoys asymptotic consistency properties, and (3) is robust to the arbitrary vagueness of prior distributions.

Since the Bayes factor is associated with predictive performance under the logarithmic scoring rule, natural alternatives arise by considering other scoring rules (Dawid and Musio 2015; Dawid et al. 2016). We consider the *Hyvärinen score* (Hyvärinen 2005), which is *proper*, *local*, and *homogeneous* (Dawid and Lauritzen 2005; Parry, Dawid, and Lauritzen 2012; Ehm and Gneiting 2012). Given T observations $y_{1:T} = (y_1, \dots, y_T) \in \mathbb{Y}^T$ and a finite set $\mathcal{M}$ of candidate models, each inducing a joint marginal density of $(Y_1, \dots, Y_T)$ denoted by p_M for $M \in \mathcal{M}$, we can regard the log-Bayes factor as a compari-

CONTACT Stephane Shao ✉ *stephaneshao@alumni.harvard.edu* Science Center 904, Harvard University, 1 Oxford street, Cambridge, MA 02138.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

son of predictive sequential (or *prequential*, Dawid 1984) log-score $-\log p_M(y_{1:T}) = \sum_{t=1}^{T} -\log p_M(y_t|y_{1:t-1})$, where by convention $p_M(y_1|y_{1:0})$ denotes the prior predictive distribution of Y_1 under model M. By contrast, for any d_y-dimensional observation $y \in \mathbb{R}^{d_y}$ and twice differentiable density p on $\mathbb{R}^{d_y}$, the Hyvärinen score is defined as

$$\mathcal{H}(y,p) = 2\Delta_y \log p(y) + \|\nabla_y \log p(y)\|^2, \qquad (1)$$

where ∇_y and Δ_y, respectively, denote the gradient and Laplacian operators with respect to the variable y. We would then select the model with the smallest prequential Hyvärinen score, defined as

$$\mathcal{H}_T(M) = \sum_{t=1}^{T} \mathcal{H}\left(y_t, p_M(dy_t|y_{1:t-1})\right). \qquad (2)$$

Homogeneity is the key property of the Hyvärinen score which is not shared by the logarithmic scoring rule. It ensures that the score does not depend on normalizing constants of candidate densities, hence offering robustness to vague priors and allowing for improper priors. For example, if M denotes the toy model $Y_1, \dots, Y_T \mid \mu \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$ with prior $\mu \sim \mathcal{N}(0, \sigma_0^2)$ and known hyperparameter $\sigma_0 > 0$, then $Y_t \mid Y_{1:t-1} \sim \mathcal{N}\left(\mu_{t-1}, \sigma_{t-1}^2 + 1\right)$ for all $t \in \{0, \dots, T\}$ by conjugacy, where $\sigma_t^2 = (t + \sigma_0^{-2})^{-1}$ and $\mu_t = \sigma_t^2 \sum_{i=1}^{t} Y_i$ for all $t \in \{1, \dots, T\}$. The log-score $-\log p_M(y_{1:T})$ becomes equivalent to $\log \sigma_0$ when $\sigma_0 \to +\infty$, and thus diverges to $+\infty$ as σ_0 increases. In other words, one could obtain Bayes factors that prefer virtually any other model over this one, by simply increasing σ_0 thus making the prior on μ arbitrarily vague, for any fixed number of observations T. On the other hand, the prequential Hyvärinen score, computed from Equations (1) and (2) using conjugacy, converges to a finite limit as $\sigma_0 \to +\infty$, so that increasing σ_0 can only influence the prequential Hyvärinen score to a limited extent. Throughout the article, the notion of robustness to arbitrary vagueness of priors is to be understood in that sense. Such a robustness is desirable when models are misspecified or when the specification of vague priors is dictated by practical considerations rather than a genuine reflection of one's prior knowledge, as is sometimes the case for parameters of complex state-space models (see, e.g., Section 4.2). The limit of $\mathcal{H}_T(M)$ as $\sigma_0 \to +\infty$ also unambiguously defines the value of the score for a flat prior $p(\mu) \propto 1$.

Without conjugacy, the calculation of the Hyvärinen score involves typically intractable integrals with respect to the sequence of partial posteriors. In this paper, we show how to use sequential Monte Carlo (SMC) methods to consistently estimate prequential Hyvärinen scores, thereby enabling their use in Bayesian model comparison for general parametric models. More specifically, we show that this estimation can be achieved for models with tractable likelihoods via SMC samplers (Chopin 2002; Del Moral, Doucet, and Jasra 2006; Zhou, Johansen, and Aston 2016). Furthermore, the case of generic state-space models can be covered by using SMC^2 (Fulop and Li 2013; Chopin, Jacob, and Papaspiliopoulos 2013) under the mild requirement that we can simulate the latent state process and evaluate the measurement density (Bretó, He, Ionides, and King 2009; Andrieu, Doucet, and Holenstein 2010), plus some integrability conditions. Our second contribution is to prove that, under regularity conditions allowing for misspecified settings, the prequential Hyvärinen score is consistent for model selection. Finally, motivated by an application to count-valued data in a population dynamics context, we propose a modified score for discrete observations that builds on recent complete characterizations of proper scoring rules on discrete spaces (McCarthy 1956; Hendrickson and Buehler 1971; Dawid, Lauritzen, and Parry 2012; Dawid, Musio, and Columbu 2017).

This article is organized as follows. In Section 2, we consider parametric models with tractable likelihoods. We present how the prequential Hyvärinen score can be estimated via SMC samplers, and show that it leads to consistent model selection, under regularity assumptions. In Section 3, we generalize the approach to nonlinear non-Gaussian state-space models, using SMC^2, and we present a simulation study with Lévy-driven stochastic volatility models. In Section 4, we extend the proposed criterion to discrete observations and compare diffusion models for population dynamics. Possible limitations and directions for future research are outlined in Section 5. Proofs, implementation details, and additional simulations are provided in the supplementary material. The R code producing the figures is available at *github.com/pierrejacob/bayeshscore*.

1.2. *Terminology and Notation*

We will abbreviate the prequential Hyvärinen score to *H-score*. Given two models M_1 and M_2, the difference of their H-scores $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)$ will be termed the *H-factor of* M_1 *against* M_2. We define $\mathbb{N}^* = \mathbb{N}\backslash\{0\}$ and use the colon notation for tuples of objects, for example,$y_{1:t} = (y_1, \dots, y_t)$ for all $t \in \mathbb{N}^*$, with the convention $y_{1:0} = \emptyset$. Unless specified otherwise, $\|\cdot\|$ denotes the Euclidean norm. Each observation $y = (y_{(1)}, \dots, y_{(d_y)})^\top$ is a vector of dimension $d_y \in \mathbb{N}^*$ and takes values in $\mathbb{Y} \subseteq \mathbb{R}^{d_y}$. Aside from Section 4, the observations are assumed to be continuous variables. Continuous probability distributions are assumed to admit densities with respect to the Lebesgue measure. We let $\mathbb{P}_\star$ (resp. $\mathbb{E}_\star$) denote the probability (resp. expectation) induced by the data-generating mechanism of the stochastic process $(Y_t)_{t \in \mathbb{N}^*}$. We use the abbreviation $\mathbb{P}_\star$-a.s. for $\mathbb{P}_\star$-*almost surely*. Assuming its existence, we let $p_\star$ denote the probability density or mass function associated with $\mathbb{P}_\star$. When dealing concurrently with several models from a set $\mathcal{M} = \{M_j : j = 1, \dots, k\}$, we use the subscript $j \in \{1, \dots, k\}$ to condition on a particular model. Each candidate model M_j is parameterized by a parameter θ_j in a space $\mathbb{T}_j \subseteq \mathbb{R}^{d_{\theta_j}}$ of dimension $d_{\theta_j} \in \mathbb{N}^*$. Explicit dependence on models is dropped from the notation whenever possible. For a differentiable function f on $\mathbb{Y}$, we use $\partial f(y_t)/\partial y_{t(k)}$ or $\partial f(y)/\partial y_{(k)}\big|_{y=y_t}$ to denote the kth partial derivative of f evaluated at $y_t \in \mathbb{Y}$. Hereafter, Gamma(α, β) distributions with shape $\alpha > 0$ and rate $\beta > 0$ have density $x \mapsto \beta^\alpha \Gamma(\alpha)^{-1} x^{\alpha-1} e^{-\beta x}$ for $x > 0$; a scaled inverse chi square distribution with degrees of freedom $\nu > 0$ and scale $s > 0$, denoted by Inv-$\chi^2(\nu, s^2)$, corresponds to the distribution of the inverse of a Gamma$(\nu/2, s^2\nu/2)$ variable, and has density $x \mapsto (\nu/2)^{\nu/2}\Gamma(\nu/2)^{-1} s^\nu x^{-(\nu/2+1)} e^{-\nu s^2/(2x)}$ for $x > 0$; NB(m, v), with $v > m > 0$, denotes a negative binomial distribution parameterized by its mean and variance, that is, with probability mass function $k \mapsto \binom{k+r-1}{k}(1-p)^r p^k$ for $k \in \mathbb{N}$, where $p = (v-m)/m$ and $r = m^2/(v-m)$.

2. H-score for Models with Tractable Likelihoods

We first describe how the H-score can be estimated with SMC samplers, before turning to asymptotic properties and numerical investigations. The H-score of a model M, defined in Equation (2), can be rewritten as

$$\mathcal{H}_T(M) = \sum_{t=1}^{T}\sum_{k=1}^{d_y}\left(2\,\frac{\partial^2 \log p(y_t|y_{1:t-1})}{\partial {y_t}_{(k)}^2} + \left(\frac{\partial \log p(y_t|y_{1:t-1})}{\partial {y_t}_{(k)}}\right)^2\right). \tag{3}$$

The marginal predictive densities appearing in Equation (3) correspond to integrals with respect to posterior distributions, as $p(y_t|y_{1:t-1}) = \int p(y_t|\theta, y_{1:t-1})\, p(\theta|y_{1:t-1})\, \mathrm{d}\theta$.

2.1. Computation of the H-score Using SMC

As noted in Dawid and Musio (2015), an interchange of differentiation and integration under appropriate regularity conditions (see Section S6 of the supplementary material) shows that $\mathcal{H}_T(M)$ is equal to

$$\sum_{t=1}^{T}\sum_{k=1}^{d_y}\left(2\,\mathbb{E}_t\left[\frac{\partial^2 \log p(y_t|y_{1:t-1},\Theta)}{\partial {y_t}_{(k)}^2} + \left(\frac{\partial \log p(y_t|y_{1:t-1},\Theta)}{\partial {y_t}_{(k)}}\right)^2\right] - \left(\mathbb{E}_t\left[\frac{\partial \log p(y_t|y_{1:t-1},\Theta)}{\partial {y_t}_{(k)}}\right]\right)^2\right), \tag{4}$$

where the conditional expectations $\mathbb{E}_t$ are taken with respect to the posterior distributions $\Theta \sim p(d\theta|y_{1:t})$. The terms of the sum in Equation (4) might not be well-defined when improper posterior distributions arise from improper priors. If τ denotes the first index such that the posterior $p(d\theta|y_{1:\tau})$ is proper, then we would redefine the H-score as $\sum_{t=\tau}^{T} \mathcal{H}\left(y_t, p(dy_t|y_{1:t-1})\right)$. This issue is not specific to the H-score, and for simplicity of exposition, we will thereafter assume that posterior distributions are proper after assimilating one observation.

In general, expectations with respect to $p(d\theta|y_{1:t})$ for all successive $t \geq 1$ can be consistently estimated using sequential or annealed importance sampling (Neal 2001) and SMC samplers (Chopin 2002; Del Moral, Doucet, and Jasra 2006). An SMC sampler starts by sampling a set of N_θ particles $\theta^{(1:N_\theta)} = (\theta^{(1)}, \ldots, \theta^{(N_\theta)})$ independently from an initial distribution $q(d\theta)$. The algorithm then assigns weights, resamples, and moves these particles in order to approximate $p(d\theta|y_{1:t})$ for each $t \geq 1$. We can move samples from a posterior distribution to the next by successively targeting intermediate distributions whose densities are proportional to $p(\theta|y_{1:t-1})p(y_t|y_{1:t-1},\theta)^{\gamma_{t,j}}$, where $0 = \gamma_{t,0} < \gamma_{t,1} < \ldots < \gamma_{t,J_t} = 1$ with $J_t \in \mathbb{N}^*$. The temperatures $\gamma_{t,j}$ can be determined adaptively to maintain a chosen level of nondegeneracy in the importance weights of the particles, for example,by forcing the effective sample size to stay above a desired threshold or by imposing a minimum number of unique particles. The resampling steps can be performed in various ways (see Douc and Cappé 2005; Murray, Lee, and Jacob 2016; Gerber, Chopin, and Whiteley 2017), and the move steps with any Markov chain Monte Carlo method. In the numerical experiments below, resampling is done with the Srinivasan Sampling Process (SSP, Gerber, Chopin, and Whiteley 2017), and move steps are independent Metropolis–Hastings steps with proposals obtained as mixtures of Normal distributions fitted on the current weighted particles. The initial distribution $q(d\theta)$ can be taken as the uniform distribution on a set (e.g., Fearnhead and Taylor 2013), as the prior distribution $p(d\theta)$ when it is proper, or more generally as an approximation of the first proper posterior distribution.

Sequential estimation of the H-score can thus be achieved at a cost comparable to that of estimating the log-evidence. Indeed, both can be obtained from the same SMC runs. However, numerical experiments suggest that the estimator of the H-score tends to have a larger relative variance than the estimator of the log-evidence, for a given number of particles. This can be explained informally as follows. For the evidence, the Monte Carlo approaches approximate expectations of the form $\mathbb{E}[p(y_t|y_{1:t-1},\Theta)]$ with respect to the posterior $p(d\theta|y_{1:t-1})$. On the other hand, the H-score involves expectations such as $\mathbb{E}[\nabla_y \log p(y_t|y_{1:t-1},\Theta)]$ with respect to $p(d\theta|y_{1:t})$. When t is large, the distributions $p(d\theta|y_{1:t-1})$ and $p(d\theta|y_{1:t})$ are similar, whereas the integrands $\theta \mapsto p(y_t|y_{1:t-1},\theta)$ and $\theta \mapsto \nabla_y \log p(y_t|y_{1:t-1},\theta)$ are different. In some generality, the first type of integrands will be easier to integrate than the second one, for example, when the former is bounded in θ while the latter is polynomial in θ, as in Normal location models (see Section 2.3).

2.2. Consistency of the H-score for iid Settings

Irrespective of model misspecification, the H-score can be justified for finite samples since it results from assessing predictions with a scoring rule that satisfies desirable properties such as propriety, locality, and homogeneity (Parry, Dawid, and Lauritzen 2012; Ehm and Gneiting 2012). Moreover, under regularity conditions, we can show that the H-score also satisfies sensible asymptotic properties: as the number of observations grows, choosing the model with the smallest H-score eventually leads to selecting the model closest to the data-generating process in a certain sense, as made precise below. Some general perspective on consistency of prequential scores can be found in Dawid and Musio (2015).

Here we consider iid models and assume that $(Y_t)_{t\in\mathbb{N}^*}$ is a sequence of iid observations drawn from $p_\star$. State-space models and more general data-generating processes will be covered in Section 3.2. For simplicity, we focus on continuous univariate ($d_y = 1$) observations. Our results will only be meaningful for models that are either nonnested, or nested with at most one model being well-specified. The case of well-specified nested models is discussed at the end of this section, with more details in Section S7.4 of the supplementary material. Our consistency result relies on the expression

$$\mathcal{H}_T(M) = \left(\sum_{t=1}^{T} \mathbb{E}_t\left[\mathcal{H}\left(y_t, p(dy_t|y_{1:t-1}, \Theta)\right)\right]\right) + \left(\sum_{t=1}^{T} \mathbb{V}_t\left[\frac{\partial \log p(y_t|y_{1:t-1}, \Theta)}{\partial y_t}\right]\right), \tag{5}$$

which follows directly from rearranging the terms in Equation (4), where $\mathbb{E}_t$ and $\mathbb{V}_t$, respectively, denote conditional expectations and variances with respect to $\Theta \sim p(d\theta|y_{1:t})$. The key insight is that, in non-nested settings, as the number of observations grows and the posterior distribution $p(d\theta|y_{1:T})$ concentrates to a point mass, the sum of the conditional expectations in Equation (5) will eventually dominate and drive the behavior of the H-score, while the sum of the conditional variances acts as a penalty term that becomes negligible. This penalty term only becomes crucial when comparing well-specified nested models, as discussed at the end of this section.

The result below considers model selection consistency for two iid models M_1 and M_2, each describing the data, respectively, as $Y_1, \dots, Y_T \mid \theta_j \overset{\text{iid}}{\sim} p_j(dy|\theta_j)$, with parameter $\theta_j \in \mathbb{T}_j$ and prior density $p_j(\theta_j)$, for $j \in \{1,2\}$.

Theorem 1. Assume $(Y_t)_{t\in\mathbb{N}^*}$ is a sequence of iid draws from $p_\star$. Assume M_1 and M_2 both satisfy the following conditions, where models are omitted from the notation and probabilistic statements are $\mathbb{P}_\star$-almost sure:

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

(b) For all $t \in \mathbb{N}^*$ and $\theta \in \mathbb{T}$, $y_t \mapsto p(y_t|\theta)$ is twice differentiable on $\mathbb{Y}$.

(c) For all $t \in \mathbb{N}^*$, there exist integrable functions $h_{1,t}$ and $h_{2,t}$ such that, for all $(y_{1:t}, \theta) \in \mathbb{Y}^t \times \mathbb{T}$, $\left|p(\theta|y_{1:t-1})\,\partial p(y_t|\theta)/\partial y_t\right| \leq h_{1,t}(\theta)$ and $\left|p(\theta|y_{1:t-1})\,\partial^2 p(y_t|\theta)/\partial {y_t}^2\right| \leq h_{2,t}(\theta)$.

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

(e) There exist a constant $L > 0$ and a neighborhood $\mathcal{U}_{\theta^\star}$ of $\theta^\star$ such that, for all $t \in \mathbb{N}^*$, $\theta \mapsto \mathcal{H}\left(Y_t, p(dy_t|\theta)\right)$ and $\theta \mapsto \partial \log p(Y_t|\theta)/\partial y_t$ are L-Lipschitz functions.

(f) There exist $\alpha_1 > 1$ and $\alpha_2 > 1$ such that $\sup_{t\in\mathbb{N}^*} \mathbb{E}\big[\,|\mathcal{H}\big(Y_t, p(dy_t|\Theta_t)\big)|^{\alpha_1} \mid Y_{1:t}\big] < +\infty$ and $\sup_{t\in\mathbb{N}^*} \mathbb{E}\big[\big(\partial \log p(Y_t|\Theta_t)/\partial y_t\big)^{2\alpha_2} \mid Y_{1:t}\big] < +\infty$, where the conditional expectations are with respect to the posterior distribution $\Theta_t \sim p(d\theta|Y_{1:t})$.

(g) $\mathbb{E}_\star\left[\left|\mathcal{H}\left(Y, p(dy|\theta^\star)\right)\right|\right] < +\infty$ and $p_\star(y)\,\partial \log p(y|\theta^\star)/\partial y \xrightarrow[|y|\to+\infty]{} 0$.

We also assume that the data-generating density $p_\star$ is such that $y \mapsto p_\star(y)$ is twice differentiable and $\mathbb{E}_\star[\left|\mathcal{H}\left(Y, p_\star(dy)\right)\right|] < +\infty$. If all the conditions are met, then we have

$$\frac{1}{T}\left(\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star-\text{a.s.}} D_{\mathcal{H}}(p_\star, M_2) - D_{\mathcal{H}}(p_\star, M_1)\,, \tag{6}$$

where, for each $j \in \{1,2\}$, the quantity

$$D_{\mathcal{H}}(p_\star, M_j) = \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_j(dy|\theta_j^\star)\right)\right] - \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_\star(dy)\right)\right] \tag{7}$$

satisfies $D_{\mathcal{H}}(p_\star, M_j) \geq 0$, with $D_{\mathcal{H}}(p_\star, M_j) = 0$ if and only if $p_j(y|\theta_j^\star) = p_\star(y)$ for all $y \in \mathbb{Y}$.

The assumptions listed in Theorem 1 are strong, which allows for more intuitive proofs. Our numerical experiments suggest that Equation (6) can hold when these conditions are not met (see, e.g., Section 2.3). Conditions (a) to (c) ensure the validity of Equation (5); (d) assumes the concentration of the posterior to a point mass; (e) to (f) ensure suitable convergence of posterior moments; and (g) ensures the strict propriety of the H-score and its definiteness for $p_\star$. Further discussion on these conditions and detailed proofs are provided in Section S7 of the supplementary material.

Theorem 1 provides insights into the asymptotic behavior of the H-score. Using integration by parts, we have

$$D_{\mathcal{H}}(p_\star, M_j) = \int \left(\frac{\partial \log p_\star(y)}{\partial y} - \frac{\partial \log p_j(y|\theta_j^\star)}{\partial y}\right)^2 p_\star(y)dy, \tag{8}$$

so that $D_{\mathcal{H}}(p_\star, M_j)$ can be interpreted as a divergence between the data-generating distribution $p_\star$ and model M_j. As long as $\mathbb{E}_\star\left[\mathcal{H}\left(Y, p_1(dy|\theta_1^\star)\right)\right] \neq \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_2(dy|\theta_2^\star)\right)\right]$, the H-score asymptotically chooses the model closest to the data-generating distribution $p_\star$ with respect to the divergence $D_{\mathcal{H}}$. In particular, if M_1 is well-specified and M_2 is misspecified, then $D_{\mathcal{H}}(p_\star, M_1) = 0 < D_{\mathcal{H}}(p_\star, M_2)$, which leads to $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) > 0$ for all sufficiently large T, $\mathbb{P}_\star$-almost surely. In other words, the H-score eventually chooses a well-specified model M_1 over a misspecified model M_2.

The divergence $D_{\mathcal{H}}(p_\star, M_j)$ appearing in (8) is sometimes referred to as the *relative Fisher information divergence* between $p_\star$ and $p_j(dy|\theta_j^\star)$ (e.g.,Walker 2016; Holmes and Walker 2017). It should be contrasted to the divergence associated with the log-score: under similar assumptions, one can prove (e.g., Dawid 2011) that

$$\frac{1}{T}\left(\left(-\log p_2(Y_{1:T})\right) - \left(-\log p_1(Y_{1:T})\right)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star-\text{a.s.}} \text{KL}(p_\star, M_2) - \text{KL}(p_\star, M_1),$$

where $\text{KL}(p_\star, M_j) = \mathbb{E}_\star\big[-\log p_j(Y|\theta_j^\star)\big] - \mathbb{E}_\star\big[-\log p_\star(Y)\big]$ denotes the Kullback-Leibler divergence between $p_\star$ and $p_j(dy|\theta_j^\star)$. In other words, the log-score $-\log p_j(Y_{1:T})$ asymptotically favors the model that is the closest to $p_\star$ with respect to the Kullback-Leibler divergence $\text{KL}(p_\star, M_j)$, whereas the H-score $\mathcal{H}_T(M_j)$ asymptotically favors the model that is the closest to $p_\star$ with respect to the divergence $D_{\mathcal{H}}(p_\star, M_j)$.

When only one of the candidate models is well-specified, the log-Bayes factor and the H-factor both agree on consistently selecting it. When both M_1 and M_2 are misspecified, each criterion selects a model according to its associated divergence. Despite being related (e.g., Bobkov, Gozlan, Roberto, and Samson 2014, and references therein), the geometries induced by these divergences differ, leading the log-Bayes factor and the H-factor to select possibly different models (see case 3 in Section 2.3). In the presence of informative priors, deciding which score to use in such misspecified settings is then a matter of preferences and further practical considerations; in this article we focus on the case of vague priors for which Bayes factors are not recommendable, as discussed earlier.

If $\mathbb{E}_\star\left[\mathcal{H}\left(Y, p_1(dy|\theta_1^\star)\right)\right] = \mathbb{E}_\star\left[\mathcal{H}\left(Y, p_2(dy|\theta_2^\star)\right)\right]$, the limit in Equation (6) becomes 0 and calls for a more careful look at the higher order penalty term formed by the conditional variances in Equation (5). Such a refinement is needed if M_1 is nested in M_2, in the sense of Eq. (9) in Berger and Pericchi (1996), and both models are well-specified. In other words, 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 $\mathbb{T}_1 \subseteq \Xi_1$ for some $k_1, k_2 \in \mathbb{N}$ with $k_2 > k_1 > 0$, and there exists $\eta_1^\star \in \Xi_2$ such that $p_1(y|\theta_1) = p_2(y|\theta_1, \eta_1^\star)$ for all $(y, \theta_1) \in \mathbb{Y} \times \mathbb{T}_1$. There also exists $\theta_1^\star \in \mathbb{T}_1$ such that $p_\star(y) = p_1(y|\theta_1^\star) = p_2(y|\theta_2^\star)$ for all $y \in \mathbb{Y}$, where $\theta_2^\star = (\theta_1^\star, \eta_1^\star)$. The particular case of nested Normal linear models is discussed in Sections 8 and 9 of Dawid and Musio (2015). Under regularity conditions, and if the parameters are orthogonal such that $\mathbb{E}_\star[\nabla_\eta \nabla_{\theta_1} \log p_2(Y|\theta_1^\star, \eta_1^\star)] = 0$, we conjecture that

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

as $T \to \infty$, in $\mathbb{P}_\star$-probability, where the difference δ_{21} in model dimensions appears as

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

This would imply that $\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1) \to +\infty$ as $T \to +\infty$, in $\mathbb{P}_\star$-probability, so that the H-score asymptotically chooses the model M_1 of smaller dimension, similarly to the log-Bayes factor for which $\log p_1(Y_{1:T}) - \log p_2(Y_{1:T}) = (1/2)(k_2 - k_1) \log T + o(\log T)$ under suitable assumptions (e.g., Moreno, Girón, and Casella 2010; Rousseau and Taeryon 2012; Chib and Kuffner 2016). Heuristic justification and numerical illustration of this postulate are provided in Sections S7.4 and S7.5 of the supplementary material. We leave more formal studies of the H-score in nested well-specified settings for future research.

As an aside, we need to contrast the prequential approach described in Equation (2) with a batch approach, where one would assess the predictive performance of model M at once via $\mathcal{H}_T^{\text{batch}}(M) = \mathcal{H}\left(y_{1:T}, p_M(dy_{1:T})\right)$. This batch approach would allow approximations using standard Markov chain Monte Carlo methods. However, the batch approach is generally not consistent for model selection (Dawid and Musio 2015, sec. 8.1). Therefore, the prequential framework not only has a natural interpretation that relates to sequential probability forecasts (Dawid 1984), but is also necessary for consistency. This leads to the task of approximating all the successive predictive distributions $p(dy_t|y_{1:t-1})$, as described in Section 2.1. This distinction does not arise for the log-score, for which we always have $-\log p(y_{1:T}) = -\sum_{t=1}^T \log p(y_t|y_{1:t-1})$. One consequence of the sequential approach is that different orderings of the observations lead to different sequences of predictive distributions, hence yielding different values of the H-score. This might be undesirable in settings where the observations are not naturally ordered (e.g., iid or spatial data). For large samples, this issue is mitigated by the convergence of rescaled H-scores to limits that do not depend on the ordering of the observations (see Theorem 1). For small samples, one could average the H-score over different permutations of the data, or use a random ordering of the data within each SMC run (see Section 2.3), at the cost of extra computations.

2.3. Numerical Illustration with Normal Models

Inspired by Section 3.2. of O'Hagan (1995), we consider the two Normal models

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

The positive hyperparameters are chosen as $\sigma_0^2 = 10$, $\nu_0 = 0.1$, and $s_0^2 = 1$. We compare M_1 and M_2, using data generated as $Y_1, \dots, Y_T \overset{\text{iid}}{\sim} \mathcal{N}(\mu_\star, \sigma_\star^2)$, in the following four settings: (1) $(\mu_\star, \sigma_\star^2) = (1, 1)$, that is, M_1 is well-specified while M_2 is not; (2) $(\mu_\star, \sigma_\star^2) = (0, 5)$, that is,$M_2$ is well-specified while M_1 is not; (3) $(\mu_\star, \sigma_\star^2) = (4, 3)$, that is,both M_1 and M_2 are misspecified; (4) $(\mu_\star, \sigma_\star^2) = (0, 1)$, that is,both M_1 and M_2 are well-specified.

Conjugacy allows all the posterior distributions, scores, and divergences to be computed in closed form. The posteriors under M_1 and M_2 concentrate, respectively, around $\theta_1^\star = \mu_\star$ and $\theta_2^\star = \sigma_\star^2 + \mu_\star^2$. We compute $D_\mathcal{H}$ and the Kullback–Leibler divergence for Normal densities analytically (Dawid and Musio 2015, sec. 6.1) and get the theoretical limits

$$D_\mathcal{H}(p_\star, M_2) - D_\mathcal{H}(p_\star, M_1) = \frac{\mu_\star^2}{\sigma_\star^2\left(\mu_\star^2 + \sigma_\star^2\right)} - \frac{\left(\sigma_\star^2 - 1\right)^2}{\sigma_\star^2}, \tag{9}$$

$$\text{KL}(p_\star, M_2) - \text{KL}(p_\star, M_1) = \frac{1}{2}\log\left(\frac{\mu_\star^2 + \sigma_\star^2}{\sigma_\star^2}\right) - \frac{\left(\sigma_\star^2 - 1\right) - \log\left(\sigma_\star^2\right)}{2}, \tag{10}$$

which depend on the values of $|\mu_\star|$ and $\sigma_\star^2$. For each of the four cases, we generate $T = 1000$ observations and perform five 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 uses a different ordering of the data, sampled uniformly from all the possible permutations. The results are shown in Figure 1. H-factors and log-Bayes factors are overlaid on the same plots in order to track their evolution jointly, but their values should not be directly compared. As expected in cases 1 and 2, the H-factor selects the well-specified model and diverges to infinity at a linear rate, with respective slopes matching the theoretical limits 0.5 and −3.2 from (9). Similar behavior is obtained for the log-Bayes factor, which correctly diverges to infinity at the same linear rate, with theoretical slopes given by (10). In case 3, both models are misspecified, and Equations (9) and (10) with $(\mu_\star, \sigma_\star^2) = (4, 3)$ yield $D_\mathcal{H}(p_\star, M_2) - D_\mathcal{H}(p_\star, M_1) \approx -1.05 < 0$ and $\text{KL}(p_\star, M_2) - \text{KL}(p_\star, M_1) \approx$

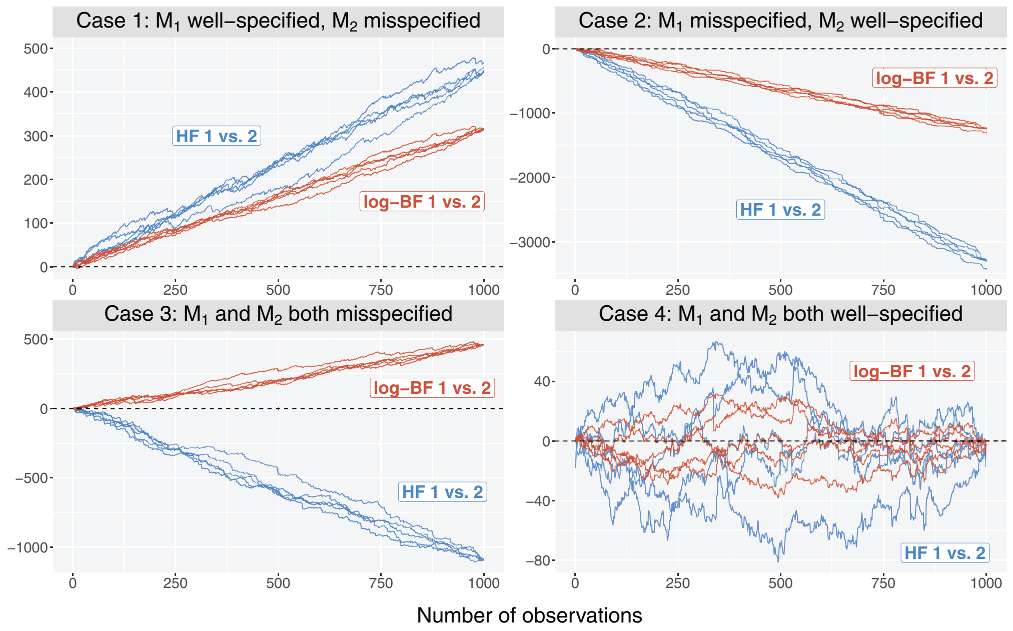

Figure 1. Estimated log-Bayes factors (log-BF) and H-factors (HF) of M_1 against M_2, computed for 5 replications (thin solid lines), under four iid data-generating processes: $\mathcal{N}(1, 1)$ (Case 1), $\mathcal{N}(0, 5)$ (Case 2), $\mathcal{N}(4, 3)$ (Case 3), and $\mathcal{N}(0, 1)$ (Case 4). In each plot, the observations are fixed but randomly ordered, so that the variability within each factor is due to Monte Carlo error and random permutation of the data. See Section 2.3.

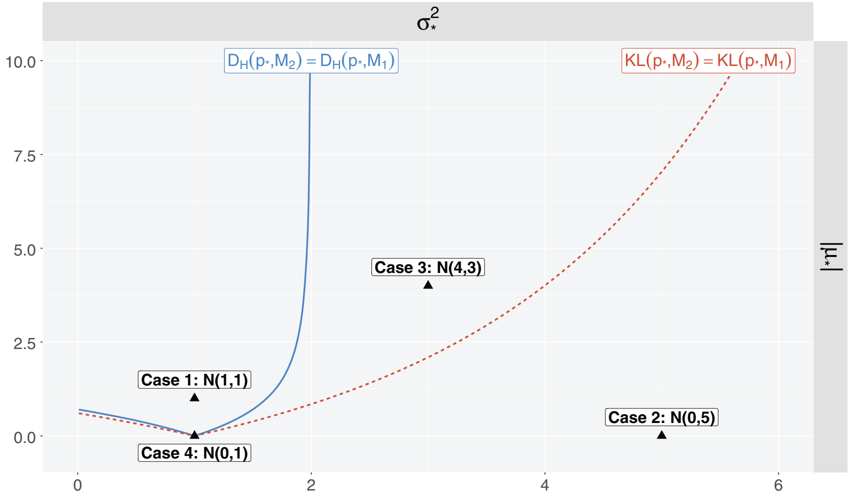

Figure 2. Phase plane of $d(p_\star, M_2) - d(p_\star, M_1)$ as a function of $(|\mu_\star|, \sigma_\star^2)$, where $d \in \{D_{\mathcal{H}}, KL\}$. The four cases from Section 2.3 are indicated as triangles. The lines (solid for $D_{\mathcal{H}}$, dashed for KL) are the sets of $(|\mu_\star|, \sigma_\star^2)$ such that $d(p_\star, M_2) = d(p_\star, M_1)$. The regions above (resp.below) the lines satisfy $d(p_\star, M_2) > d(p_\star, M_1)$ (resp.<), that is, M_1 (resp.M_2) is closer to $p_\star$.

$0.47 > 0$. This leads the Bayes factor and the H-factor to favor different misspecified models. In fact, when both M_1 and M_2 are misspecified, there are infinitely many combinations of $(|\mu_\star|, \sigma_\star^2) \in \mathbb{R}_+^2$ for which $D_{\mathcal{H}}(p_\star, M_2) < D_{\mathcal{H}}(p_\star, M_1)$ whereas $\mathrm{KL}(p_\star, M_2) > \mathrm{KL}(p_\star, M_1)$. Indeed, if we define the boundary $\mathcal{B}_{\mathcal{H}}(\sigma_\star^2) = |\sigma_\star^2 - 1|(2 - \sigma_\star^2)^{-1/2}$ for $\sigma_\star^2 \in (0, 2)$ and $\mathcal{B}_{\mathcal{H}}(\sigma_\star^2) = +\infty$ for $\sigma_\star^2 \geq 2$, then $D_{\mathcal{H}}(p_\star, M_2) = D_{\mathcal{H}}(p_\star, M_1)$ (resp.> and <) for $|\mu_\star| = \mathcal{B}_{\mathcal{H}}(\sigma_\star^2)$ (resp.> and <). By contrast, $\mathrm{KL}(p_\star, M_2) = \mathrm{KL}(p_\star, M_1)$ if and only if $|\mu_\star| = \mathcal{B}_{\mathrm{KL}}(\sigma_\star^2)$, where $\mathcal{B}_{\mathrm{KL}}(\sigma_\star^2) = (\exp(\sigma_\star^2 - 1) - \sigma_\star^2)^{1/2}$ for all $\sigma_\star^2 > 0$. Thus, whenever $\mathcal{B}_{\mathrm{KL}}(\sigma_\star^2) < |\mu_\star| < \mathcal{B}_{\mathcal{H}}(\sigma_\star^2)$, the divergences $D_{\mathcal{H}}$ and KL disagree on which model is closer to $p_\star$. This is illustrated in Figure 2. When both divergences are sensible, deciding which one to use would require further considerations (see, e.g., Jewson et al. 2018). As explained in Section 1, the log-Bayes factor might be inappropriate in the presence of vague priors. Looking back at case 1 for example, since $\log p_{M_1}(y_{1:T}) \to -\infty$ when $\sigma_0 \to +\infty$, one could always specify a σ_0 large enough such that the log-Bayes factor would wrongly pick M_2. On the other hand, the choice of M_1 by the H-factor remains unchanged when σ_0 increases. This robustness is further illustrated in Section S2 of the supplementary material.

Finally, in case 4, the theoretical slopes are exactly 0, while the models are of equal dimensions, hence no model prevails.

3. H-score for State-Space Models

The H-score raises additional computational challenges in the case of state-space models. State-space models, also known as hidden Markov models, are a flexible and widely used class of time series models (Cappé, Moulines, and Rydén 2005; Douc, Moulines, and Stoffer 2014), which describe the observations $(Y_t)_{t\in\mathbb{N}^*}$ as conditionally independent given a latent Markov chain $(X_t)_{t\in\mathbb{N}^*}$ living in $\mathbb{X} \subseteq \mathbb{R}^{d_x}$. A state-space model with parameter $\theta \in \mathbb{T} \subseteq \mathbb{R}^{d_\theta}$ specifies an initial distribution $\mu_\theta(dx_1)$ of the first state X_1, a Markov kernel $f_\theta(dx_{t+1}|x_t)$ for the transition of the latent process, a measurement distribution $g_\theta(dy_t|x_t)$, and a prior distribution $p(d\theta)$ on the parameter.

3.1. Computation of the H-score Using SMC^2

The conditional predictive distributions $p(y_t|y_{1:t-1},\theta)$ appearing in Equation (4) correspond to integrals over the latent states, that is, $p(y_t|y_{1:t-1},\theta) = \int p(x_t|y_{1:t-1},\theta)\, g_\theta(y_t|x_t)\, dx_t$, which are in general intractable. Interchanging differentiation and integration under suitable regularity conditions yields the following results, which are similar to Fisher's and Louis' identities (Cappé, Moulines, and Rydén 2005, prop. 10.1.6), except that differentiation here is with respect to the observation instead of the parameter. We obtain for all $\theta \in \mathbb{T}$, all observed $y_{1:T} \in \mathbb{Y}^T$, all $k \in \{1,\ldots,d_y\}$, and all $t \in \{1,\ldots,T\}$,

$$\frac{\partial \log p(y_t|y_{1:t-1},\theta)}{\partial y_{t(k)}} = \mathbb{E}_t\left[\left.\frac{\partial \log g_\theta(y_t|X_t)}{\partial y_{t(k)}}\right|\theta\right], \tag{11}$$

$$\frac{\partial^2 \log p(y_t|y_{1:t-1},\theta)}{\partial y_{t(k)}^2} + \left(\frac{\partial \log p(y_t|y_{1:t-1},\theta)}{\partial y_{t(k)}}\right)^2 = \mathbb{E}_t\left[\left.\frac{\partial^2 \log g_\theta(y_t|X_t)}{\partial y_{t(k)}^2} + \left(\frac{\partial \log g_\theta(y_t|X_t)}{\partial y_{t(k)}}\right)^2\right|\theta\right], \tag{12}$$

where the conditional expectations $\mathbb{E}_t$ are with respect to $X_t \sim p(dx_t|y_{1:t},\theta)$. Proofs of Equations (11) and (12) under regularity assumptions are presented in the supplementary material. Applying Equations (11) and (12) to each term in Equation (4) and using the tower property of conditional expectations yields

$$\mathcal{H}_T(M) = \sum_{t=1}^{T}\sum_{k=1}^{d_y}\left(2\,\mathbb{E}_t\left[\frac{\partial^2 \log g_\Theta(y_t|X_t)}{\partial y_{t(k)}^2} + \left(\frac{\partial \log g_\Theta(y_t|X_t)}{\partial y_{t(k)}}\right)^2\right] - \left(\mathbb{E}_t\left[\frac{\partial \log g_\Theta(y_t|X_t)}{\partial y_{t(k)}}\right]\right)^2\right), \tag{13}$$

where the expectations $\mathbb{E}_t$ are with respect to the joint posterior distributions of (Θ, X_t) given the observations $y_{1:t}$, whose densities are given by $p(\theta, x_t|y_{1:t}) = p(\theta|y_{1:t})p(x_t|y_{1:t},\theta)$.

For many state-space models, the log-derivatives of the measurement density $g_\theta(y|x)$ can be evaluated at any point $(\theta, y, x) \in \mathbb{T} \times \mathbb{Y} \times \mathbb{X}$. Assuming that we can simulate the transition kernel of the latent process, we can use SMC^2 (Fulop and Li 2013; Chopin, Jacob, and Papaspiliopoulos 2013) to consistently estimate all the conditional expectations appearing in Equation (13). At each time $t \in \{1,\ldots,T\}$, SMC^2 produces a set of weighted particles targeting the joint density $p(\theta, x_t|y_{1:t})$, which can be used to update the H-score.

3.2. Consistency of the H-score for State-Space Models

We revisit the asymptotic consistency results of the H-score in the case of state-space models. The observations are no longer assumed to be iid and we consider two candidate models, M_1 and M_2. An additional difficulty in proving consistency of the H-score with dependent observations lies in the approximation of $\mathcal{H}_T(M_j)$ by a stationary analog, to which ergodic theorems will apply. As in the iid setting, we only give results for univariate continuous observations.

Theorem 2. Assume $(Y_t)_{t\in\mathbb{N}^*}$ is ergodic and strongly stationary, so that we can artificially extend its set of indices to negative integers and consider the two-sided process $(Y_t)_{t\in\mathbb{Z}}$. Assume M_1 and M_2 both satisfy the following conditions, where models are omitted from the notation and probabilistic statements are $\mathbb{P}_\star$-almost sure:

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

(b) For all $t \in \mathbb{N}^*$ and $\theta \in \mathbb{T}$, $y_t \mapsto p(y_t|\theta)$ is twice differentiable on $\mathbb{Y}$.

(c) For all $t \in \mathbb{N}^*$, there exist integrable functions $h_{1,t}$ and $h_{2,t}$ such that, for all $(y_{1:t},\theta) \in \mathbb{Y}^t\times\mathbb{T}$, $\left|p(\theta|y_{1:t-1})\,\partial p(y_t|\theta)/\partial y_t\right| \leq h_{1,t}(\theta)$ and $\left|p(\theta|y_{1:t-1})\,\partial^2 p(y_t|\theta)/\partial {y_t}^2\right| \leq h_{2,t}(\theta)$.

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

(e) For all $t \in \mathbb{N}^*$ and $(\theta, x_t) \in \mathbb{T} \times \mathbb{X}$, $y_t \mapsto g_\theta(y_t|x_t)$ is twice differentiable on $\mathbb{Y}$.

(f) There exist integrable functions $h_{3,t}$ and $h_{4,t}$ such that, for all $(y_{1:t},\theta,x_t) \in \mathbb{Y}^t \times \mathbb{T} \times \mathbb{X}$, $\left|p(x_t|y_{1:t-1},\theta)\partial g_\theta(y_t|x_t)/\partial y_t\right| \leq h_{3,t}(x_t)$ and $\left|p(x_t|y_{1:t-1},\theta)\partial^2 g_\theta(y_t|x_t)/\partial y_t^2\right| \leq h_{4,t}(x_t)$.

(g) For all $t \in \mathbb{N}^*$, 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 $\Theta_t \xrightarrow[t\to+\infty]{\mathcal{D}} \theta^\star$.

(h) There exist a constant $L > 0$ and a neighborhood $\mathcal{U}_{\theta^\star}$ of $\theta^\star$ such that, for all $t \in \mathbb{N}^*$, $\theta \mapsto \mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1},\theta)\right)$ and $\theta \mapsto \partial \log p(Y_t|Y_{1:t-1},\theta)/\partial y_t$ are L-Lipschitz functions.

(i) There exist $\alpha_1 > 1$ and $\alpha_2 > 1$ such that $\sup_{t\in\mathbb{N}^*} \mathbb{E}\left[\,|\mathcal{H}\left(Y_t, p(dy_t|Y_{1:t-1},\Theta_t)\right)|^{\alpha_1} \mid Y_{1:t}\right] < +\infty$ and $\sup_{t\in\mathbb{N}^*} \mathbb{E}\left[\left(\partial \log p(Y_t|Y_{1:t-1},\Theta_t)/\partial y_t\right)^{2\alpha_2} \mid Y_{1:t}\right] < +\infty$, where the conditional expectations are with respect to the posterior distribution $\Theta_t \sim p(d\theta|Y_{1:t})$.

(j) 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.

(k) 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 $0 < \sigma^- < \nu_{\theta^\star}(x_{t+1}|x_t) < \sigma^+ < +\infty$.

(l) 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 $+\infty$.

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

(n) $\sup_{\substack{x\in\mathbb{X}\\ y\in\mathbb{Y}}} g_{\theta^\star}(y|x) < +\infty$ and $\mathbb{E}_\star \left[\left|\log\left(\int_{\mathbb{X}} g_{\theta^\star}(Y_1|x)\nu_{\theta^\star}(dx)\right)\right|\right] < +\infty.$

(o) The conditional density $y_1 \mapsto p_\star(y_1|Y_{-\infty:0})$ of Y_1 given $(Y_t)_{t\le 0}$ is well-defined and twice differentiable, and $\mathbb{E}_\star\left[\left|\mathcal{H}\left(Y_1, p_\star(dy_1|Y_{-\infty:0})\right)\right|\right] < +\infty.$

If these conditions are met, we may define, for each $j \in \{1,2\}$, the quantity

$$D_{\mathcal{H}}(p_\star, M_j) = \mathbb{E}_\star\left[\mathcal{H}\left(Y_1, p_j(dy_1|Y_{-\infty:0}, \theta_j^\star)\right)\right] - \mathbb{E}_\star\left[\mathcal{H}\left(Y_1, p_\star(dy_1|Y_{-\infty:0})\right)\right], \tag{14}$$

where $p_j(y_1|Y_{-\infty:0}, \theta_j^\star)$ is the provably well-defined conditional density of Y_1 given $(Y_t)_{t\le 0}$ under M_j and $\theta_j^\star$. Under these conditions, we have

$$\frac{1}{T}\left(\mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star-a.s.} D_{\mathcal{H}}(p_\star, M_2) - D_{\mathcal{H}}(p_\star, M_1)\,. \tag{15}$$

If $p_\star(y_1|Y_{-\infty:0})\,\partial \log p(y_1|Y_{-\infty:0}, \theta^\star)/\partial y_1 \xrightarrow[|y_1|\to+\infty]{\mathbb{P}_\star-a.s.} 0$, then we have $D_{\mathcal{H}}(p_\star, M_j) \ge 0$, with $D_{\mathcal{H}}(p_\star, M_j) = 0$ if and only if $p_j(y_1|Y_{-\infty:0}, \theta_j^\star) = p_\star(y_1|Y_{-\infty:0})$, $\mathbb{P}_\star$-almost surely.

Conditions (a) to (c) ensure the validity of Equation (5); (d) to (f) ensure the validity of (11) and (12); (g) assumes the concentration of the posterior to a point mass; (h) to (i) yield suitable convergence of posterior moments; (j) to (l) ensure the *forgetting propriety* of the latent Markov chain and the H-score; (m) to (n) relate to the well-definiteness of the conditional density $p_j(y_1|Y_{-\infty:0}, \theta_j^\star)$; finally, (o) and the last boundary condition ensure that the H-score is strictly proper and well-defined for $p_\star$. Further discussion on these conditions and detailed proofs is provided in Section S7 of the supplementary material.

For state-space models, posterior concentration results have been derived in specific cases (e.g., Lijoi, Prünster, and Walker 2007; De Gunst and Shcherbakova 2008; Shalizi 2009; Gassiat and Rousseau 2014; Douc, Moulines, and Stoffer 2014; Douc, Olsson, and Roueff 2016, and references therein). However, to the best of our knowledge, general results on posterior concentration for misspecified state-space models have yet to be established. As a consequence, our proof of Theorem 2 uses posterior concentration as a working assumption. Our numerical examples suggest that concentration of posterior distributions can be observed in practice, even for complex state-space models (see posterior density plots in Section S4 of the supplementary material). Further research on Bayesian asymptotics in state-space models might provide more theoretical understanding of such phenomena.

3.3. *Illustration with Lévy-Driven Stochastic Volatility Models*

In this simulation study we illustrate the consistency of the H-score in nonlinear, non-Gaussian state-space models with continuous observations. A simpler example with linear Gaussian state-space and ARMA models can be found in Section S3 of the supplementary material. Here we consider Lévy-driven stochastic volatility models (Barndorff-Nielsen and Shephard 2001, 2002). These models feature intractable transition kernels that can only be simulated, and describe the joint evolution of the log-returns Y_t and the instantaneous volatility V_t of a financial asset. The former is modeled as a continuous time process driven by a Brownian motion, while the latter is modeled as a Lévy process. Given a triplet of parameters (λ, ξ, ω), we can generate random variables $(V_t, Z_t)_{t\ge 1}$ recursively as

$$\left.\begin{array}{l} k \sim \text{Poisson}\left(\lambda\xi^2/\omega^2\right); \quad C_{1:k} \overset{\text{iid}}{\sim} \text{Unif}(t-1, t);\\ E_{1:k} \overset{\text{iid}}{\sim} \text{Exp}\left(\xi/\omega^2\right); \quad Z_0 \sim \text{Gamma}\left(\xi^2/\omega^2,\ \xi/\omega^2\right);\\ Z_t = e^{-\lambda} Z_{t-1} + \sum_{j=1}^k e^{-\lambda(t-C_j)} E_j;\\ V_t = \frac{1}{\lambda}\left(Z_{t-1} - Z_t + \sum_{j=1}^k E_j\right). \end{array}\right\}. \tag{16}$$

The first model (M_1) describes the volatility as driven by a *single factor*, expressed in terms of a finite rate Poisson process.

M_1:(V_t, Z_t) *from Equation* (16) *given* $(\lambda,\ \xi,\ \omega)$; $X_t = (V_t, Z_t)$; $Y_t\,|\,X_t \sim \mathcal{N}\left(\mu + \beta V_t, V_t\right)$; *with independent priors* $\lambda \sim \text{Exp}(1)$; $\xi,\ \omega^2 \overset{\text{iid}}{\sim} \text{Exp}\left(1/5\right)$; $\mu,\ \beta \overset{\text{iid}}{\sim} \mathcal{N}(0, 10)$.

The second model (M_2) introduces an additional independent component to drive the behavior of the volatility, leading to the *multifactor* model below.

M_2:$(V_{i,t}, Z_{i,t})$ *from Equation* (16) *independently for* $i \in \{1,2\}$ *given* $(\lambda_i, \xi\mathrm{w}_i, \omega\mathrm{w}_i)$, *with* $(\mathrm{w}_1, \mathrm{w}_2) = (\mathrm{w}, 1-\mathrm{w})$; $X_t = (V_{1,t}, V_{2,t}, Z_{1,t}, Z_{2,t})$; $Y_t\,|\,X_t \sim \mathcal{N}\left(\mu + \beta V_t, V_t\right)$ *where* $V_t = V_{1,t} + V_{2,t}$; *with independent priors* $\lambda_1 \sim \text{Exp}(1)$; $\lambda_2 - \lambda_1 \sim \text{Exp}\left(1/2\right)$; $\mathrm{w} \sim \text{Unif}(0,1)$; $\xi,\ \omega^2 \overset{\text{iid}}{\sim} \text{Exp}\left(1/5\right)$; $\mu,\ \beta \overset{\text{iid}}{\sim} \mathcal{N}(0, 10)$.

For model M_1, we can prove that there exist values of $\theta = (\lambda, \xi, \omega, \mu, \beta)$ such that $\mathbb{E}[\,|\partial \log g_\theta(y_1|X_1)/\partial y_1|\,] = +\infty$, which prevents the use of Equations (11) and (12) to estimate the H-score of model M_1. When Equations (11) and (12) do not hold, we can directly estimate the partial derivatives of $\tilde{y}_t \mapsto p(\tilde{y}_t|y_{1:t-1}, \theta)$ at the observed y_t, by using approximate draws from the conditional predictive distribution $p(dy_t|y_{1:t-1}, \theta)$. Approximate draws from $p(dy_t|y_{1:t-1}, \theta)$ can be obtained from a run of SMC^2, as long as one can sample from the measurement distribution $g_\theta(dy_t|x_t)$. For a chosen bandwidth $h > 0$ (e.g., Hardle, Marron, and Wand, 1990; Tsybakov, 2009, sec. 1.11) and a twice continuously differentiable kernel K integrating to 1, for example, a standard Gaussian kernel $K(u) = (2\pi)^{-1/2}\exp(-u^2/2)$, we can use n draws $\tilde{y}_t^{(1)}, \ldots, \tilde{y}_t^{(n)}$ from $p(dy_t|y_{1:t-1}, \theta)$ to consistently estimate $p(y_t|y_{1:t-1}, \theta)$ by the kernel density

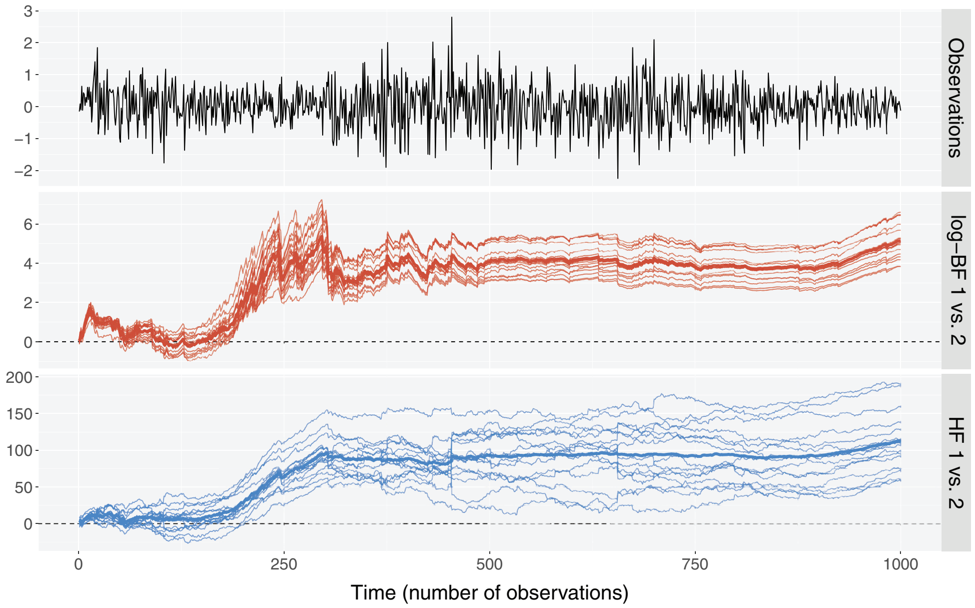

Figure 3. Top panel: log-returns simulated from model M_1 with parameters $\lambda = 0.01$, $\xi = 0.5$, $\omega^2 = 0.0625$, $\mu = 0$, and $\beta = 0$. Middle and bottom panels: estimated log-Bayes factor (log-BF) and H-factor (HF) of M_1 against M_2, computed for 15 replications (thin solid lines), along with the average scores across replications (thick solid lines). In each plot, the variability within each factor is due to Monte Carlo error. See Section 3.3.

estimator $\widehat{p}(y_t|y_{1:t-1},\theta) = (nh)^{-1}\sum_{i=1}^{n} K((y_t - \tilde{y}_t^{(i)})/h)$. This kernel density estimator is twice differentiable with respect to y_t, hence we can respectively use $\partial\,\widehat{p}(y_t|y_{1:t-1},\theta)/\partial y_{t(k)}$ and $\partial^2\,\widehat{p}(y_t|y_{1:t-1},\theta)/\partial y_{t(k)}^2$ as consistent estimators of the partial derivatives $\partial p(y_t|y_{1:t-1},\theta)/\partial y_{t(k)}$ and $\partial^2 p(y_t|y_{1:t-1},\theta)/\partial y_{t(k)}^2$, as $n \to +\infty$ and $h \to 0$ at an appropriate rate (e.g., Bhattacharya 1967).

We simulate $T = 1000$ observations from a single-factor Lévy-driven stochastic volatility model with parameters $\lambda = 0.01$, $\xi = 0.5$, $\omega^2 = 0.0625$, $\mu = 0$, and $\beta = 0$, following the simulations of Barndorff-Nielsen and Shephard (2002). The H-factor of M_1 against M_2 is computed for 15 replications of SMC^2, using $N_\theta = 1024$ particles in θ, and an adaptive number of particles in x starting at $N_x = 128$. The kernel density estimation is performed with a Gaussian kernel, using $n = 1024$ predictive draws and $h = 0.1$. The estimated log-Bayes factor and H-factor of M_1 against M_2 are plotted in Figure 3. Here the models are nested and well-specified, but their dimensions differ. We see that both criteria correctly select the smaller model M_1. As mentioned in Section 2.1, the estimated H-factor tends to have a larger relative variance than the estimated log-Bayes factor, especially in the presence of extreme observations (e.g., at times 454 and 656), and might thus call for a larger number of particles.

4. H-score for Discrete Observations

Motivated by an application in population dynamics (Section 4.2), we propose an extension of the H-score to discrete observations. We assume that each observation $y = (y_{(1)}, \dots, y_{(d_y)})^\top$ takes finite values (i.e., $\|y\| < +\infty$) in some discrete space $\mathbb{Y} = [\![a_1, b_1]\!] \times \cdots \times [\![a_{d_y}, b_{d_y}]\!]$, where $[\![a_k, b_k]\!] = [a_k, b_k] \cap \mathbb{Z}$ and $a_k, b_k \in \mathbb{Z} \cup \{-\infty, +\infty\}$, with $a_k < b_k$ for all $k \in \{1, \dots, d_y\}$. For ease of exposition, assume for now that $b_k - a_k \geq 3$ for all $k \in \{1, \dots, d_y\}$.

4.1. Extension of the H-score to Discrete Observations

Let e_k denote the canonical vector of $\mathbb{Z}^{d_y}$ that has all coordinates equal to 0 except for its kth coordinate that equals 1. For all $y \in \mathbb{Y}$, all nonnegative functions p on $\mathbb{Y}$, and all $k \in \{1, \dots, d_y\}$, we define $\partial_k p(y) = (p(y + e_k) - p(y - e_k))/2$ and $\partial_k \log p(y) = \partial_k p(y)/p(y)$. We define the score

$$\mathcal{H}^D(y, p) = \sum_{k=1}^{d_y} \mathcal{H}_k^D(y, p)\,, \tag{17}$$

where $\mathcal{H}_k^D(y, p) = 2\,\partial_k\left(\partial_k \log p(y)\right) + \left(\partial_k \log p(y)\right)^2$ if $a_k + 2 \leq y_{(k)} \leq b_k - 2$. At the boundaries, we define $\mathcal{H}_k^D(y, p)$, respectively, as $\partial_k \log p(y + e_k)$, $\partial_k \log p(y + e_k) + \left(\partial_k \log p(y)\right)^2$, $-\partial_k \log p(y - e_k) + \left(\partial_k \log p(y)\right)^2$, and $-\partial_k \log p(y - e_k)$ for $y \in \{a_k, a_k + 1, b_k - 1, b_k\}$.

The expression of $\mathcal{H}_k^D$ can be regarded as a discrete analog of the H-score where the partial derivatives are replaced by central finite differences. Proper scores for discrete observations can be entirely characterized as super-gradients of concave entropy functions (McCarthy 1956; Hendrickson and Buehler 1971; Dawid, Lauritzen, and Parry 2012). Using this characterization, we can prove that $\mathcal{H}^D$ is a proper scoring rule.

If $b_k = a_k + 1$ (e.g., for binary data) or $b_k = a_k + 2$, we could still define $\mathcal{H}_k^D$ by ignoring the cases $y_{(k)} = a_k + 1$, or $y_{(k)} = b_k - 1$, or both. Alternatively, we could use forward differences. All these definitions lead to scores that meet the requirements of being insensitive to prior vagueness, while being proper and local. Deciding which one to use is then a matter of further considerations, left for future research. The construction of $\mathcal{H}^D$

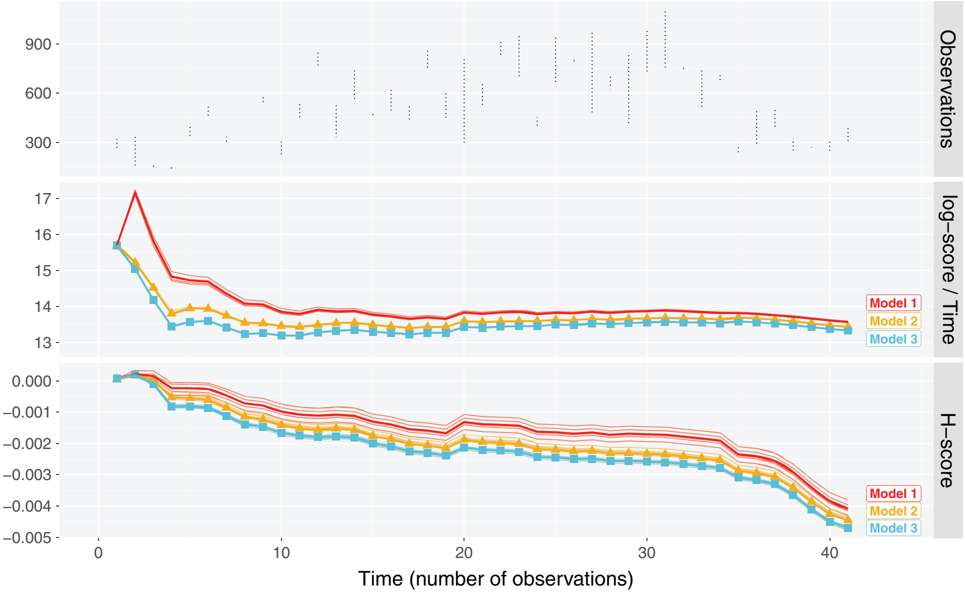

Figure 4. Top panel: double transect counts of red kangaroos. Middle and bottom panels: estimated log-scores and H-scores of M_1 (circles), M_2 (triangles), and M_3 (squares), for 5 replications (thin solid lines), along with the average scores across replications (thick lines with shapes). The log-scores are rescaled by the number of observations for better readability. The variability within each model is due to Monte Carlo error. See Section 4.2.

and the proof of its propriety are detailed in Section S5 of the supplement.

4.2. Diffusion Models for Population Dynamics of Red Kangaroos

We illustrate the H-score for discrete observations by comparing three nonlinear non-Gaussian state-space models, describing the dynamics of a population of red kangaroos (*Macropus rufus*) in New South Wales, Australia. These models were compared in Knape and de Valpine (2012) using Bayes factors, although the authors acknowledged the undesirable sensitivity of their results to their choice of prior distributions. The data (Caughley, Shepherd, and Short 1987) is a time series of 41 bi-variate observations $(Y_{1,t}, Y_{2,t})$, formed by double transect counts of red kangaroos, measured between 1973 and 1984 (see Figure 4). The small number of observations calls for a criterion that is principled for finite samples, contrarily to, for example, the Bayesian information criterion. The models are nested and will be referred to as M_1, M_2, and M_3, by decreasing order of complexity. The largest model (M_1) is a logistic diffusion model. Simpler versions include an exponential growth model (M_2) and a random-walk model (M_3). In these models, a latent population size (X_t) follows a stochastic differential equation (see further motivation in Dennis and Costantino 1988; Knape and de Valpine 2012). Each model is specified below, where $(W_t)_{t\geq 0}$ denotes a standard Brownian motion.

M_1: $X_1 \sim \mathrm{LN}(0,5)$; $\quad dX_t/X_t = (\sigma^2/2 + r - bX_t)\,dt + \sigma\,dW_t$; $Y_{1,t}, Y_{2,t} \mid X_t, \tau \overset{\text{iid}}{\sim} \mathrm{NB}(X_t, X_t + \tau X_t^2)$; *with independent priors;* $\sigma, \tau, b \overset{\text{iid}}{\sim} \mathrm{Unif}(0, 10)$, $r \sim \mathrm{Unif}(-10, 10)$.

M_2: *same as* M_1 *with* $b = 0$; *with independent priors* $\sigma, \tau \overset{\text{iid}}{\sim} \mathrm{Unif}(0, 10)$, $r \sim \mathrm{Unif}(-10, 10)$.

M_3: *same as* M_1 *with* $b = 0$ *and* $r = 0$; *with independent priors* $\sigma, \tau \overset{\text{iid}}{\sim} \mathrm{Unif}(0, 10)$.

We perform 5 runs of SMC^2 to estimate the log-score and H-score of each model, with an adaptive number N_x of latent particles. We use $N_\theta = 16,384$ particles in θ, and $N_x = 32$ initial particles in x. For model M_1, we simulate the latent process using the Euler-Maruyama method with discretization step $\Delta_t = 0.001$. The estimated log-scores and H-scores are shown in Figure 4. For better readability, the log-score is rescaled by the number of observations. Using the H-scores would lead to selecting model M_3, similarly to Knape and de Valpine (2012) who used log-scores. Their conclusion was mitigated by the sensitivity of the evidence to the choice of vague priors: for instance, changing the prior on r in model M_2 to $\mathrm{Unif}(-100, 100)$ effectively divides the evidence of M_2 by a factor 10. On the other hand, we have found the impact of that change of prior on the H-score to be indistinguishable from the Monte Carlo variation across runs.

5. Discussion

The H-factor constitutes a competitive alternative to the Bayes factor. It is justified non-asymptotically since it relies on assessing predictive performances using a proper local scoring rule, and it is robust to the arbitrary vagueness of prior distributions. It can be applied to a large variety of models — including nonlinear non-Gaussian state-space models—and it can be estimated sequentially with SMC or SMC^2, at a cost comparable to that of the Bayes factor. Using our R implementation, one SMC or SMC^2 replication took about a few minutes for each iid Normal models with 1000 observations, about an hour for each kangaroo population model with 41 observations, and about 5 h for each stochastic volatility model with 1000 observations. In all cases, the Monte Carlo error can be arbitrarily reduced by increasing the number of particles N_θ (Section 3 in Chopin, Jacob, and Papaspiliopoulos 2013). However, the H-score puts additional smoothness restrictions on the models, for example, the twice differentiability of their predictive distributions with respect to the observations (see Dawid and Musio 2015, and its rejoinder). Thus, there are models for which the Bayes factor is applicable but not the H-factor. We have also discussed in Section 2.3 a case where the two criteria disagree, even asymptotically, contrarily to , for example, partial and intrinsic Bayes factors (Santis and Spezzaferri 1999) that asymptotically agree with the Bayes factor.

The sequential form of the score is problematic when observations are not naturally ordered, leading to different values of the H-score for different orderings. This issue is mitigated by the following facts: if the sample is large enough, any ordering of the data would yield similar H-scores. For smaller samples, one could average the H-score over random permutations of the data. In that case, quantifying and controlling the extra variability induced by these permutations would deserve investigation.

For continuous observations and nonnested parametric models satisfying strong regularity assumptions, we have proved that the H-score leads to consistent model selection. The asymptotic behavior of the H-factor is determined by how close the candidate models are from the data-generating process, where closeness is quantified by the relative Fisher information divergence associated with the H-score, in contrast to the Kullback–Leibler divergence associated with the Bayes factor. Our proofs rely on strong assumptions, but the numerical experiments indicate that the results might hold in more generality. Results for discrete observations and nested well-specified models would be interesting topics of future research. It would be interesting to study frequentist properties of the proposed model choice procedure, for example, by deriving confidence intervals for the difference in expected H-scores. One could for instance complement the results of Theorems 1 and 2 with central limit theorems, which would enable further connections between Bayesian model selection criteria and likelihood ratio tests as described, for example, in Vuong (1989).

To deal with vague or improper priors, other alternatives to the log-evidence include Bayesian cross-validation criteria, for example, $\sum_{t=1}^{T} \log p(y_t|y_{-t})$, where $y_{-t} = \{y_s : 1 \leq s \leq T \text{ and } s \neq t\}$. Such criteria would be applicable under weaker smoothness assumptions on the predictive densities, while still being robust to arbitrary vagueness of prior distributions. Efficient computation of these criteria is challenging, and can be envisioned for iid models using MCMC (Alqallaf and Gustafson 2001) or SMC methods (Bornn, Doucet, and Gottardo 2010); the case of state-space models would be more challenging, due to standard difficulties arising when splitting time series. Another approach suggested in Kamary et al. (2014) is to cast model selection as a mixture estimation problem, which also raises questions in the case of time series.

Supplementary Materials

The supplementary material provides some guidance on the implementation of SMC methods to estimate H-scores, additional numerical experiments, details on the extension of the Hyvärinen score to discrete observations, and formal proofs of the consistency of H-scores in non-nested settings, along with heuristic arguments and illustrations for nested models.

Acknowledgments

Stephane Shao thanks the participants of Greek Stochastics iota (July 2017) for their helpful feedback.

Funding

Pierre E. Jacob gratefully acknowledges support by the National Science Foundation through grant DMS-1712872 and by the Harvard Data Science Initiative. This work was also supported in part by DARPA Grant No. W911NF1810134.

References

Alqallaf, F., and Gustafson, P. (2001), "On Cross-Validation of Bayesian Models," *Canadian Journal of Statistics*, 29, 333–340. [11]

Andrieu, C., Doucet, A., and Holenstein, R. (2010), "Particle Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Series B, 72, 269–342. [2]

Barndorff-Nielsen, O. E., and Shephard, N. (2001), "Non-Gaussian Ornstein–Uhlenbeck-based Models and Some of Their Uses in Financial Economics," *Journal of the Royal Statistical Society*, Series B, 63, 167âŁ"-241. [8]

Barndorff-Nielsen, O. E., and Shephard, N. (2002), "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society*, Series B, 64, 253–280. [8,9]

Bartlett, M. S. (1957), "A Comment on D. V. Lindley's Statistical Paradox," *Biometrika*, 44, 533–534. [1]

Berger, J. O., and Pericchi, L. R. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122. [1,5]

——— (2001), "Objective Bayesian Methods for Model Selection: Introduction and Comparison," *Model Selection, IMS Lecture Notes – Monograph Series*, 68, 135–207. [1]

Berger, J. O., Pericchi, L. R., and Varshavsky, J. (1998), "Bayes Factors and Marginal Distributions in Invariant Situations," *Sankhya A: The Indian Journal of Statistics*, 60, 307–321. [1]

Bernardo, J. M., and Smith, A. F. M. (2000), *Bayesian Theory*, New York: Wiley. [1]

Bhattacharya, P. (1967). "Estimation of a Probability Density Function and Its Derivatives," *Sankhyā: The Indian Journal of Statistics*, Series A, pp. 373–382. [9]

Bobkov, S. G., Gozlan, N., Roberto, C., and Samson, P.-M. (2014), "Bounds on the Deficit in the Logarithmic Sobolev Inequality," *Journal of Functional Analysis*, 267, 4110–4138. [4]

Bornn, L., Doucet, A., and Gottardo, R. (2010), "An Efficient Computational Approach for Prior Sensitivity Analysis and Cross-validation," *Canadian Journal of Statistics*, 38, 47–64. [11]

Bretó, C., He, D., Ionides, E. L., and King, A. A. (2009), "Time Series Analysis Via Mechanistic Models," *The Annals of Applied Statistics*, 3, 319–348. [2]

Cappé, O., Moulines, E., and Rydén, T. (2005), *Inference in Hidden Markov Models*, New York: Springer Series in Statistics. [7]

Caughley, G., Shepherd, N. and Short, J. (1987), *Kangaroos, Their Ecology and Management in the Sheep Rangelands of Australia*, Cambridge, UK: Cambridge University Press. [10]

Chib, S., and Kuffner, T. A. (2016), "Bayes Factor Consistency," Preprint, arXiv:1607.00292. [1,5]

Chopin, N. (2002), "A Sequential Particle Filter Method for Static Models," *Biometrika*, 89, 539–552. [2,3]

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, 397–426. [2,7,11]

Dawid, A. P. (2011), "Posterior Model Probabilities," *Handbook of the Philosophy of Science*, 7, 607–630. [1,4]

Dawid, A. P., and Lauritzen, S. L. (2005), "The Geometry of Decision Theory," In *Proceedings of the Second International Symposium on Information Geometry and Its Applications*, pp. 22–28. [1]

Dawid, A. P., Lauritzen, S., and Parry, M. (2012), "Proper Local Scoring Rules on Discrete Sample Spaces," *The Annals of Statistics*, 40, 593–608. [2,9]

Dawid, A. P., and Musio, M. (2015). "Bayesian Model Selection Based on Proper Scoring Rules," *Bayesian Analysis*, 10, 479–499. [1,3,5,11]

Dawid, A. P., Musio, M., and Columbu, S. (2017), "A Note on Bayesian Model Selection for Discrete Data Using Proper Scoring Rules," *Statistics and Probability Letters*, 129, 101–106. [2]

Dawid, A. P., Musio, M., and Ventura, L. (2016), "Minimum Scoring Rule Inference," *Scandinavian Journal of Statistics*, 43, 123–138. [1]

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, 342–356. [8]

Del Moral, P., Doucet, A., and Jasra, A. (2006), "Sequential Monte Carlo Samplers," *Journal of the Royal Statistical Society*, Series B, 68, 411–436. [2,3]

Dennis, B., and Costantino, R. (1988), "Analysis of Steady-state Populations with the Gamma-abundance Model: Application to Tribolium," *Ecology*, 69, 1200–1213. [10]

Douc, R., and Cappé, O. (2005), "Comparison of Resampling Schemes for Particle Filtering," *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pp. 64–69. [3]

Douc, R., Moulines, E., and Stoffer, D. (2014). *Nonlinear Time Series: Theory, Methods and Applications with R Examples* (1st edn.), Boca Raton, FL: Chapman and Hall/CRC. [7,8]

Douc, R., Olsson, J., and Roueff, F. (2016). "Posterior Consistency for Partially Observed Markov Models," Preprint, arXiv:1608.06851. [8]

Ehm, W., and Gneiting, T. (2012), "Local Proper Scoring Rules of Order Two," *The Annals of Statistics*, 40, 609–637. [1,3]

Fearnhead, P., and Taylor, B. M. (2013), "An Adaptive Sequential Monte Carlo Sampler," *Bayesian Analysis*, 8, 411–438. [3]

Fulop, A., and Li, J. (2013), " Efficient Learning Via Simulation: A Marginalized Resample-move Approach," *Journal of Econometrics*, 176, 146–161. [2,7]

Gassiat, E., and Rousseau, J. (2014), "About the Posterior Distribution in Hidden Markov Models with Unknown Number of States," *Bernoulli*, 20, 2039–2075. [8]

Gerber, M., Chopin, N. and Whiteley, N. (2017), "Negative Association, Ordering and Convergence of Resampling Methods," Preprint, arXiv:1707.01845. [3]

Hardle, W., Marron, J., and Wand, M. (1990), "Bandwidth Choice for Density Derivatives," *Journal of the Royal Statistical Society*, Series B, 52, 223–232. [8]

Hendrickson, A. D., and Buehler, R. J. (1971), "Proper Scores for Probability Forecasters," *The Annals of Mathematical Statistics*, 42, 1916–1921. [2,9]

Holmes, C. and Walker, S. (2017), "Assigning a Value to a Power Likelihood in a General Bayesian Model," *Biometrika*, 104, 497–503. [4]

Hyvärinen (1984), "The Prequential Approach," *Journal of the Royal Statistical Society*, Series A, 147, 278–292. [2,5]

Hyvärinen, A. (2005), "Estimation of Non-normalized Statistical Models by Score Matching," *Journal of Machine Learning Research*, 6, 695–709. [1]

Jeffreys, H. (1939), *Theory of Probability*, Oxford, UK: Oxford University Press. [1]

Jewson, J., Smith, J. Q., and Holmes, C. (2018), "Principled Bayesian Minimum Divergence Inference," Preprint, arXiv:1802.09411. [6]

Kamary, K., Mengersen, K., Robert, C. P., and Rousseau, J. (2014), "Testing Hypotheses Via a Mixture Estimation Model," Preprint, arXiv:1412.2044v2. [1,11]

Kass, R. and Raftery, A. (1995). "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [1]

Key, J. T., Pericchi, L. R., and Smith, A. F. (1999), "Bayesian Model Choice: What and Why," *Bayesian Statistics*, 6, 343–370. [1]

Knape, J., and de Valpine, P. (2012), "Fitting Complex Population Models by Combining Particle Filters with Markov Chain Monte Carlo," *Ecology*, 93, 256–263. [1,10]

Lee, J. and MacEachern, S. N. (2011), "Consistency of Bayes Estimators Without the Assumption that the Model Is Correct," *Journal of Statistical Planning and Inference*, 141, 748–757. [1]

Lijoi, A., Prünster, I. and Walker, S. G. (2007), "Bayesian Consistency for Stationary Models," *Econometric Theory*, 23, 749–759. [8]

McCarthy, J. (1956), "Measures of the Value of Information," *Proceedings of the National Academy of Sciences of the United States of America*, 42, 654–655. [2,9]

Moreno, E., Girón, F. J., and Casella, G. (2010), "Consistency of Objective Bayes Factors as the Dimension Grows," *The Annals of Statistics*, 38, 1937–1952. [5]

Murray, L. M., Lee, A., and Jacob, P. E. (2016), "Parallel Resampling in the Particle Filter," *Journal of Computational and Graphical Statistics*, 25, 789–805. [3]

Neal, R. M. (2001). "Annealed Importance Sampling," *Statistics and Computing*, 11, 125–139. [3]

O'Hagan, A. (1995). Fractional Bayes Factor for Model Comparison," *Journal of the Royal Statistical Society*, Series B, 57, 99–138. [1,5]

Parry, M., Dawid, A. P., and Lauritzen, S. (2012). "Proper Local Scoring Rules," *The Annals of Statistics*, 40, 561–592. [1,3]

Robert, C. (2007). *The Bayesian Choice: From Decision-theoretic Foundations to Computational Implementation*, New York: Springer Science & Business Media. [1]

Rousseau, J., and Taeryon, C. (2012), "Bayes Factor Consistency in Regression Problems." *HAL, archives ouvertes*, hal-00767469. [5]

Santis, F. and Spezzaferri, F. (1999). "Methods for Default and Robust Bayesian Model Comparison: The Fractional Bayes Factor Approach," *International Statistical Review*, 67, 267–286. [11]

Shalizi, C. R. (2009), "Dynamics of Bayesian Updating with Dependent Data and Misspecified Models," *Electronic Journal of Statistics*, 3, 1039–1074. [8]

Tsybakov, A. B. (2009), *Introduction to Nonparametric Estimation*, New York: Springer Series in Statistics. [8]

Vuong, Q. H. (1989), "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses," *Econometrica: Journal of the Econometric Society*, 57, 307–333. [11]

Walker, S. G. (2013), "Bayesian Inference with Misspecified Models," *Journal of Statistical Planning and Inference*, 143, 1621–1633. [1]

——— (2016), "Bayesian Information in an Experiment and the Fisher Information Distance," *Statistics & Probability Letters*, 112, 5–9. [4]

Zhou, Y., Johansen, A. M., and Aston, J. A. D. (2016), "Towards Automatic Model Comparison: An Adaptive Sequential Monte Carlo Approach," *Journal of Computational and Graphical Statistics*, 25, 701–726. [2]