Bayesian model comparison with the Hyvärinen score

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

Harvard University*, Duke University†

What this research is about

A new model selection criterion

Our goal is to perform **Bayesian model comparison when candidate models use vague or improper prior distributions** on their parameters. We present **a new criterion, based on the Hyvärinen score** advocated by Dawid and Musio (2015), and we provide a way to **estimate it using sequential Monte Carlo (SMC)** methods.

Why it is relevant

Limitations of the Bayes factor

Consider the toy model, with known $\sigma_0^2 > 0$,

$$Y_1, ..., Y_T \,|\, \mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$$
$$\mu \sim \mathcal{N}(0, \sigma_0^2)$$

The evidence (marginal likelihood) of this model given T observations $y_1, ..., y_T \in \mathbb{R}$ satisfies

$$p(y_{1:T}) \underset{\sigma_0^2 \to +\infty}{\sim} \frac{1}{\sigma_0} \left[\frac{1}{\sqrt{2\pi}} \prod_{t=2}^{T} \varphi\left(y_t \,\middle|\, \frac{\sum_{i=1}^{t-1} y_i}{t-1}, \frac{t}{t-1} \right) \right]$$

with $\varphi(\cdot|\mu, \sigma^2)$ denoting the density of a $\mathcal{N}(\mu, \sigma^2)$.

For any fixed sample size T, increasing σ_0 effectively multiplies the evidence by $1/\sigma_0$. Therefore, the conclusion of a model selection procedure using **the Bayes factor can be arbitrarily influenced by the choice of prior** (Figure 1). In the limit when the prior distribution becomes improper, the log-evidence is no longer well-defined.

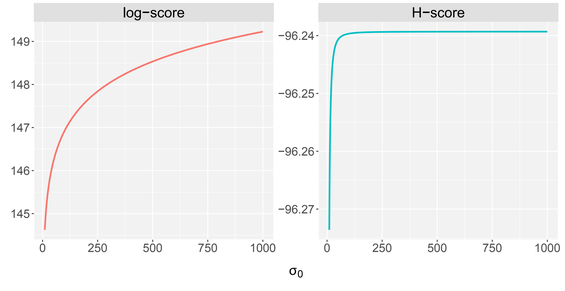

Fig. 1: Log-score and H-score of the Normal model $Y_1, ..., Y_T \,|\, \mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ with prior $\mu \sim \mathcal{N}(0, \sigma_0^2)$, for increasing values of σ_0, based on $T = 100$ observations simulated as i.i.d. $\mathcal{N}(0, 1)$. Notice the arbitrary shift in the log-score when σ_0 increases, in contrast to the relative robustness of the H-score when the prior becomes more vague.

What we propose

Prequential Hyvärinen score and its estimation via SMC methods

The Bayes factor selects the model with the smallest log-score $-\log p(y_{1:T}) = \Sigma_{t=1}^{T} - \log p(y_t|y_{1:t-1})$. Instead, we select the model with the smallest **prequential Hyvärinen score** (H-score), defined as

$$\mathcal{H}_T\left(y_{1:T}, p(dy_{1:T})\right) = \sum_{t=1}^{T} \mathcal{H}(y_t, p(dy_t|y_{1:t-1})),$$

with $p(y_1|y_{1:0}) = p(y_1)$, where for all $y = (y_{(1)}, ..., y_{(d_y)}) \in \mathbb{Y} \subseteq \mathbb{R}^{d_y}$ and all twice differentiable density p,

$$\mathcal{H}(y, p(dy)) = 2\,\Delta \log p(y) + \|\nabla \log p(y)\|^2.$$

This can be extended to discrete observations via finite differences. For unordered data, we may consider averaging $\mathcal{H}_T$ over different permutations of the data.

Estimation of the H-score

Under mild regularity assumptions, the prequential Hyvärinen score $\mathcal{H}_T\left(y_{1:T}, p\right)$ of a model is equal to

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

- For static and time series models with tractable likelihoods, this can be estimated with **SMC** samplers.
- For state-space models, further work leads to an expression involving only posterior expectations of the derivatives of the (typically available) observation log-density. These can be estimated with **SMC^2** even without analytical knowledge of the transition kernel, as long as one can simulate from it.

How it performs in practice

Numerical experiments

We illustrate our method by comparing three population models from Knape & de Valpine (2012) (Figure 3). We observe a time series $(Y_{1,t}, Y_{2,t})$ of double counts of kangaroos in New South Wales, Australia (Figure 2). The three state-space models introduce a latent population size (X_t) that follows an SDE.

Model 1 : $X_1 \sim \text{LN}(0, 5)$; $dX_t/X_t = (\sigma^2/2 + r - bX_t)\,dt + \sigma dW_t$; $Y_{1,t}, Y_{2,t} \overset{\text{i.i.d.}}{\sim} \text{NB}(X_t,\ X_t + \tau X_t^2)$

with independent priors $b,\ \sigma,\ \tau \sim \text{Unif}(0, 10)$ and $r \sim \text{Unif}(-10, 10)$

Model 2 and **Model 3** are nested versions of model 1, satisfying $b = 0$ and $b = r = 0$ respectively.

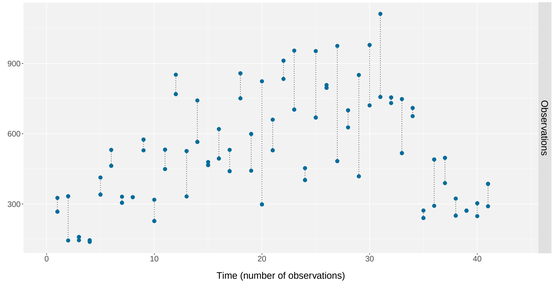

Fig. 2: Double total transect counts of red kangaroos performed on consecutive days at 41 occasions, spaced at irregular time intervals ranging from two to six months, from 1973 to 1984.

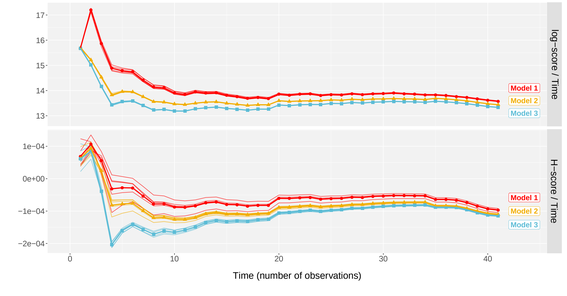

Fig. 3: Estimated scores, rescaled by the number of observations and computed for 5 replications (thin lines). The average scores across replications (thick lines) are plotted for better readability.

Why it works

Theoretical results

As a proper and local loss function, the H-score is **justified non-asymptotically** by decision theory.

Under conditions, **the H-score is also consistent**: the H-factor $\mathcal{H}_T^{2|1} = \mathcal{H}_T(M_2) - \mathcal{H}_T(M_1)$ satisfies

$$\frac{1}{T}\mathcal{H}_T^{2|1} \underset{T \to +\infty}{\overset{\mathbb{P}_\star - a.s.}{\longrightarrow}} D_{\mathcal{H}}(p_\star, M_2) - D_{\mathcal{H}}(p_\star, M_1),$$

where $D_{\mathcal{H}}(p_\star, M_j)$ is a divergence between the data generating process $p_\star$ and model M_j, which contrasts with the KL-divergence tied to the log-score.

This consistency is illustrated numerically with two i.i.d. models $M_1 = \{\mathcal{N}(\theta_1, 1) \,|\, \theta_1 \sim \mathcal{N}(0, \sigma_0^2)\}$ and $M_2 = \{\mathcal{N}(0, \theta_2) \,|\, \theta_2 \sim \text{Inv-}\chi^2(\nu_0, s_0^2)\}$ (Figure 4).

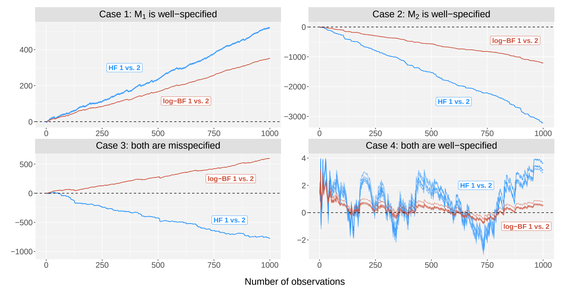

Fig. 4: Estimated log-Bayes factors (log-BF, red) and H-factors (HF, blue), computed for 5 replications (thin solid lines), under four different i.i.d. data generating processes: $\mathcal{N}(1, 1)$ (Case 1), $\mathcal{N}(0, 5)$ (Case 2), $\mathcal{N}(2, 3)$ (Case 3), $\mathcal{N}(0, 1)$ (Case 4).

Discussion

Why it is appealing and when it might fail

- Robust to arbitrary vagueness of priors and well-defined even for improper priors
- Can be estimated sequentially at the cost of SMC / SMC^2 under mild assumptions
- Regularity assumptions not always satisfied

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

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

[3] S. Shao, P. E. Jacob., J. Ding, V. Tarokh. *Bayesian model comparison with the Hyvärinen score.* Preprint, arXiv:1711.00136, 2017.