Bayesian model selection involving vague and improper priors

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

Harvard University

What this research is about

A new model selection criterion

Our goal is to perform **Bayesian model selection and comparison, in settings where candidate models make use of vague or improper prior distributions** on their parameters. We present **a new model selection 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 $\sigma_0^2 > 0$,

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

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

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

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

Increasing σ_0 effectively multiplies the evidence by $1/\sigma_0$. Therefore, **making the prior more vague can make the evidence arbitrarily small, thus misleading the conclusion of a model selection procedure based on the Bayes factor** (Figure 1). In the extreme case where the prior is improper, the log-evidence is not even well-defined.

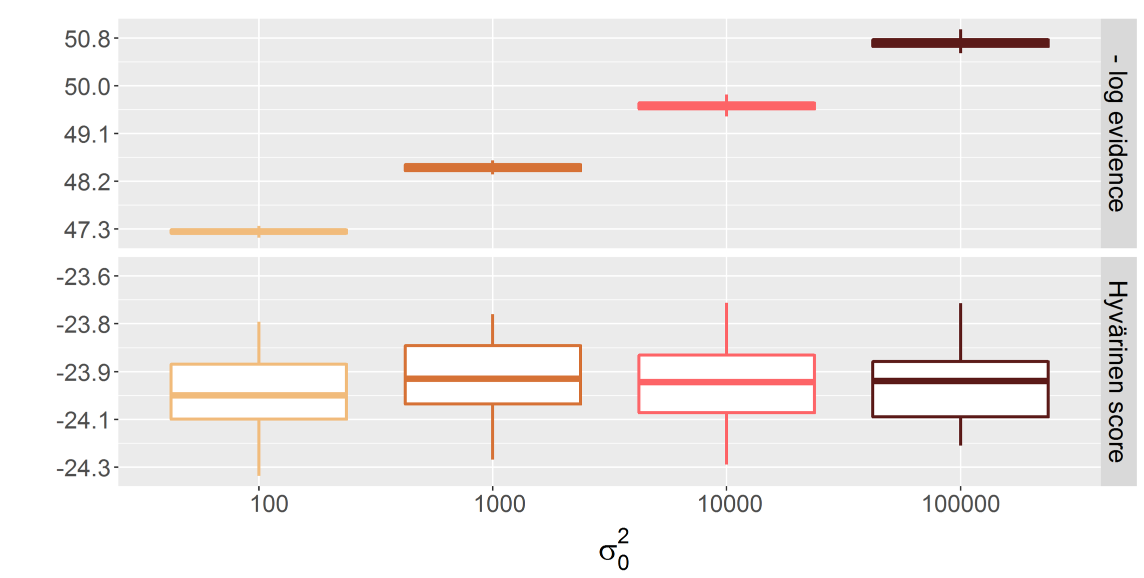

Fig. 1: Negative log-evidence and prequential Hyvärinen score of the Normal model $Y_1, \dots, Y_T \,|\, \mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ with prior $\mu \sim \mathcal{N}(0, \sigma_0^2)$, for increasing values of σ_0^2, based on $T = 30$ observations simulated as i.i.d. $\mathcal{N}(0,1)$ (box-plot over 50 replications). Notice how the log-evidence systematically shifts when the vagueness of the prior varies, as opposed to the prequential Hyvärinen score that remains relatively unchanged.

What we propose

Prequential Hyvärinen score and its estimation via SMC methods

The Bayes factor selects the model with the smallest *prequential* (predictive sequential) log score $\sum_{t=1}^{T} -\log p(y_t|y_{1:t-1})$. Instead, as suggested by Dawid and Musio (2015), **we select the model with the smallest prequential Hyvärinen score**, defined as

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

where, for all $y_t = (y_{t(1)}, \dots, y_{t(d_y)}) \in \mathbb{Y} \subseteq \mathbb{R}^{d_y}$ and assuming sufficient smoothness of the predictive densities,

$$\mathcal{H}(y_t, p(dy_t|y_{1:t-1})) = \sum_{k=1}^{d_y}\left[2\frac{\partial^2 \log p(y_t|y_{1:t-1})}{\partial y_{t(k)}^2} + \left(\frac{\partial \log p(y_t|y_{1:t-1})}{\partial y_{t(k)}}\right)^2\right]$$

Extension to discrete observations is possible but the details will be omitted.

Estimation of the prequential Hyvärinen score

Under mild regularity assumptions, the prequential Hyvärinen score $\mathcal{H}_T(y_{1:T}, p)$ 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 and 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). These 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.

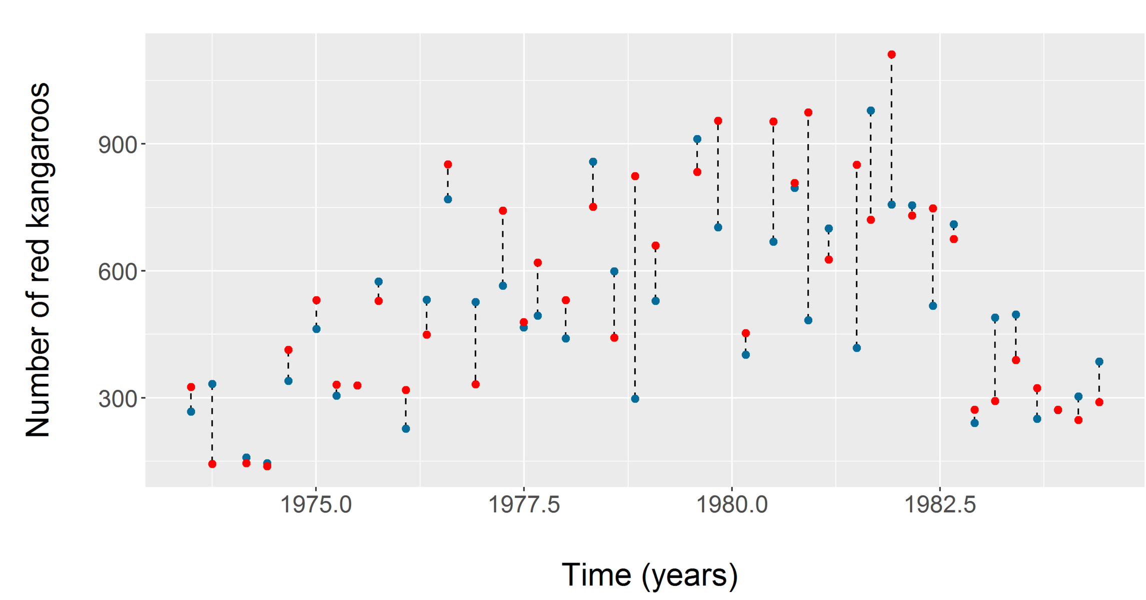

Fig. 2: Double total transect counts of red kangaroos performed on consecutive days at 41 occasions, spaced at irregular time intervals ranging from two to six months, from 1973 to 1984.

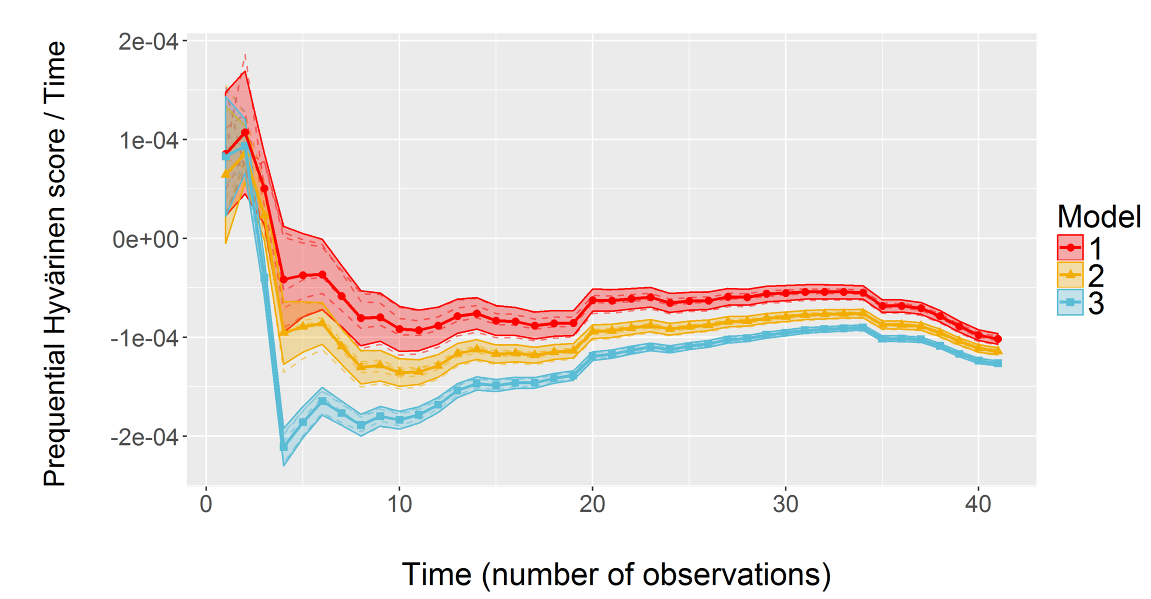

Fig. 3: Estimated prequential Hyvärinen scores of each model. Averages (solid) across 5 replications (dashed) and approximate t-based 95% confidence bands (filled) are plotted for readability.

Why it works

Theoretical results

Under mild regularity assumptions, the Hyvärinen score is a *proper* loss function, which provides a **non-asymptotic decision theoretic justification**.

Under technical conditions, **the prequential Hyvärinen score is consistent**: letting $p_\star$ denote the data generating process (DGP), the standardized *Hyvärinen factor* between models M_1 and M_2 (defined as M_2's score minus M_1's score) converges $p_\star$ - a.s. (when $T \to +\infty$) to $\Delta(p_\star, M_2) - \Delta(p_\star, M_1)$ where Δ is a measure of discrepancy.

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).

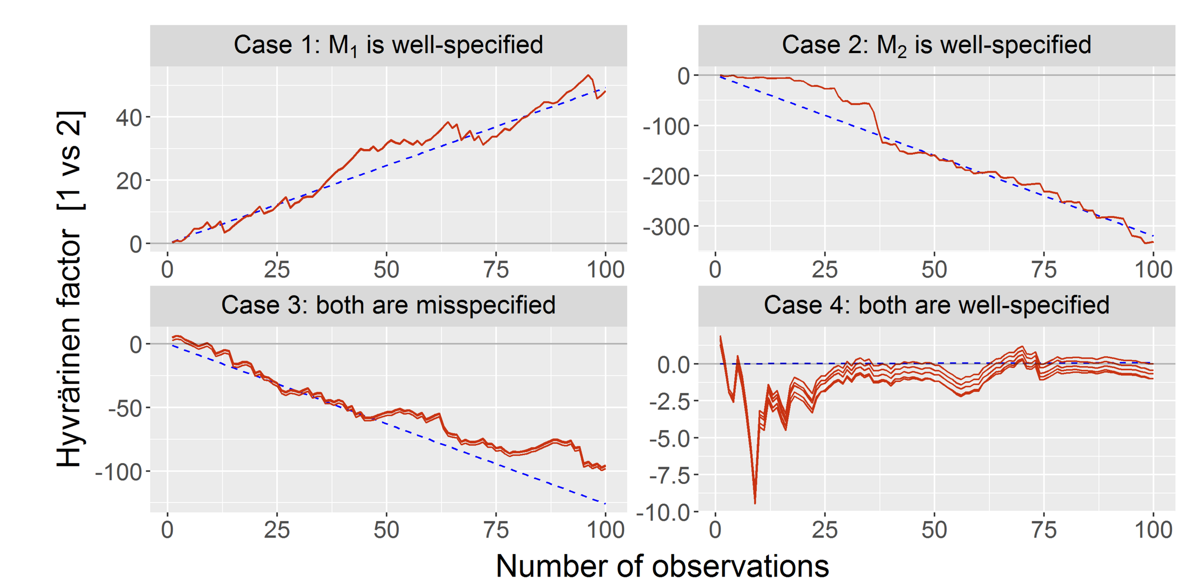

Fig. 4: Hyvärinen factor (solid red), selecting M_1 when positive and M_2 when negative, along with its theoretical slope (dashed blue), under four different i.i.d. DGP's, in this order: $\mathcal{N}(1,1)$ (M_1 is well-specified), $\mathcal{N}(0,5)$ (M_2 is well-specified), $\mathcal{N}(2,3)$ (both are misspecified), $\mathcal{N}(0,1)$ (both are well-specified).

Discussion

Why it is appealing and when it might fail

- Robust to arbitrary vagueness of priors, and well-defined even with improper priors
- Can be estimated sequentially at the cost of SMC / SMC^2 under mild assumptions
- Regularity assumptions not always satisfied, leading to inconsistency or large variance

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

[2] N. Chopin, P. E. Jacob, and 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.