Consistency of H-factors for model selection

Stephane Shao[*]

September 12, 2018

in collaboration with Pierre E. Jacob[*], Jie Ding[†] *and* Vahid Tarokh[‡]

[*]Department of Statistics, Harvard University
[†]School of Statistics, University of Minnesota
[‡]Department of Electrical and Computer Engineering, Duke University

Table of contents

Table of contents

Table of contents

Why use H-factors ?

Goal: compare and select Bayesian models

- We want to select a model from a set $\{M_1, ..., M_q\}$, given observations $Y_{1:T} = (Y_1, ..., Y_T) \in (\mathbb{R}^{d_y})^T$ from a data generating process $p_\star$.
- Each model M_j is a collection of distributions $p_{\theta_j}(dy_{1:T})$ parametrized by $\theta_j \in \mathbb{T}_j \subseteq \mathbb{R}^{d_j}$ with a prior distribution $p(d\theta_j)$.
- We are interested in settings where the priors may be vague, and the models may be misspecified ($p_\star \notin M_j$).

Goal: compare and select Bayesian models

- We want to select a model from a set $\{M_1, ..., M_q\}$, given observations $Y_{1:T} = (Y_1, ..., Y_T) \in (\mathbb{R}^{d_y})^T$ from a data generating process $p_\star$.
- Each model M_j is a collection of distributions $p_{\theta_j}(dy_{1:T})$ parametrized by $\theta_j \in \mathbb{T}_j \subseteq \mathbb{R}^{d_j}$ with a prior distribution $p(d\theta_j)$.
- We are interested in settings where the priors may be vague, and the models may be misspecified ($p_\star \notin M_j$).

Goal: compare and select Bayesian models

- We want to select a model from a set $\{M_1, ..., M_q\}$, given observations $Y_{1:T} = (Y_1, ..., Y_T) \in (\mathbb{R}^{d_y})^T$ from a data generating process $p_\star$.
- Each model M_j is a collection of distributions $p_{\theta_j}(dy_{1:T})$ parametrized by $\theta_j \in \mathbb{T}_j \subseteq \mathbb{R}^{d_j}$ with a prior distribution $p(d\theta_j)$.
- We are interested in settings where the priors may be vague, and the models may be misspecified ($p_\star \notin M_j$).

What kind of models ? i.i.d. , state-space models, ...

- **Example 1**. Normal i.i.d. models [*O'Hagan, 1995*]

Model 1	Model 2
$Y_{1:T} \mid \theta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_1, 1)$	$Y_{1:T} \mid \theta_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \theta_2)$
$\theta_1 \sim \mathcal{N}(0, \sigma_0^2)$	$\theta_2 \sim \text{Inv-}\chi^2(\nu_0, s_0^2)$

What kind of models ? i.i.d., state-space models, ...

- **Example 2**. Lévy-driven stochastic volatility models for log-returns of financial assets [*Barndorff-Nielsen & Shephard, 2001*]

Given parameters (λ, ξ, ω), generate random variables $(V_t, Z_t)_{t\geq 1}$ recursively as

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

Model 1

Given parameters $(\lambda,\xi,\omega,\mu,\beta)$:

$$(V_t, Z_t) \sim (⚙)$$

$$X_t=(V_t, Z_t)$$

$$Y_t \mid X_t \sim \mathcal{N}(\mu+\beta V_t, V_t)$$

with independent priors:

$$\lambda\sim\text{Exp}(1); \quad \xi, \omega^2\sim\text{Exp}(1/5); \quad \mu, \beta\sim\mathcal{N}(0,10)$$

Model 2

Given parameters $(\lambda_1,\lambda_2,\text{w}_1,\text{w}_2,\xi,\omega,\mu,\beta)$:

$$(V_{1,t}, Z_{1,t}) \sim (⚙) \text{ with } (\lambda_1,\ \xi\text{w}_1,\ \omega\text{w}_1)$$

$$(V_{2,t}, Z_{2,t}) \sim (⚙) \text{ with } (\lambda_2,\ \xi\text{w}_2,\ \omega\text{w}_2)$$

$$X_t=(V_{1,t}, V_{2,t}, Z_{1,t}, Z_{2,t})$$

$$V_t=V_{1,t}+V_{2,t}$$

$$Y_t \mid X_t \sim \mathcal{N}(\mu+\beta V_t, V_t)$$

with independent priors:

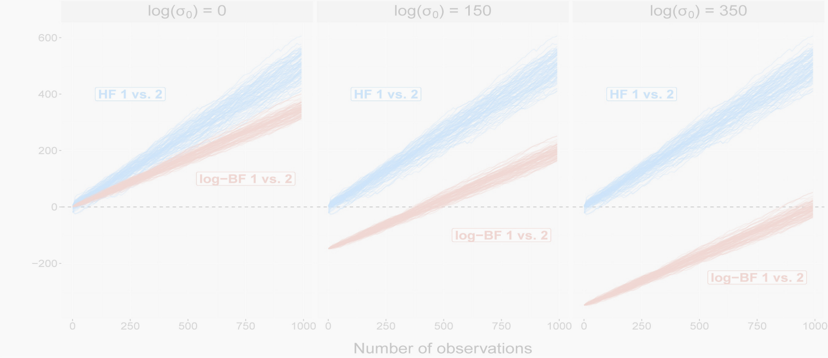
$$\lambda_1\sim\text{Exp}(1); \quad \lambda_2-\lambda_1\sim\text{Exp}(1/2); \quad \xi, \omega^2\sim\text{Exp}(1/5)$$

$$1-\text{w}_2=\text{w}_1\sim\text{Unif}(0,1); \quad \mu, \beta\sim\mathcal{N}(0,10)$$

Limitations of Bayes factors: sensitivity to vague priors

- Making the prior more vague effectively multiplies the evidence $p_j(y_{1:T})$ of a model M_j by an arbitrarily small constant, for any fixed sample size.
- Example 1. Observations generated as $Y_{1:1000} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(1,1)$.

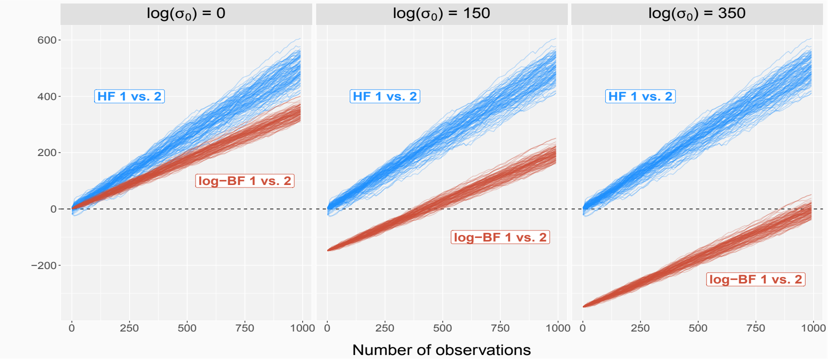
$$M_1 = \{Y_{1:T} \mid \theta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_1, 1)\ ;\ \theta_1 \sim \mathcal{N}(0, \sigma_0^2)\}$$
$$M_2 = \{Y_{1:T} \mid \theta_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \theta_2)\ ;\ \theta_2 \sim \text{Inv-}\chi^2(0.1, 1)\}$$



Limitations of Bayes factors: sensitivity to vague priors

- Making the prior more vague effectively multiplies the evidence $p_j(y_{1:T})$ of a model M_j by an arbitrarily small constant, for any fixed sample size.
- **Example 1**. Observations generated as $Y_{1:1000} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(1,1)$.

$$M_1 = \{Y_{1:T} \,|\, \theta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_1, 1) \; ; \; \theta_1 \sim \mathcal{N}(0, \sigma_0^2)\}$$
$$M_2 = \{Y_{1:T} \,|\, \theta_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \theta_2) \; ; \; \theta_2 \sim \text{Inv-}\chi^2(0.1, 1)\}$$



Decision theoretic solution: change of scoring rule

- Bayes factors select models maximizing $\log p(y_{1:T}) = \sum_{t=1}^{T} \log p(y_t \mid y_{1:t-1})$.

- This corresponds to minimizing the prequential score [*Dawid, 1984*]

$$\sum_{t=1}^{T} \mathcal{S}\left(y_t,\, p(dy_t \mid y_{1:t-1})\right)$$

  with the choice of scoring rule $\mathcal{S}(y, p) = -\log p(y)$ called the log-score.

- Each scoring rule has an associated divergence function

$$\mathrm{D}_{\mathcal{S}}(p, q) = \mathbb{E}_{Y\sim p}\left[\mathcal{S}(Y, q) - \mathcal{S}(Y, p)\right]$$

  $\mathcal{S}$ is (strictly) proper if $q \mapsto \mathrm{D}_{\mathcal{S}}(p, q)$ is (uniquely) minimized at $q = p$.

- The log-score is strictly proper and tied to the Kullback-Leibler divergence

$$\mathrm{KL}(p, q) = \int \left[\log p(y) - \log q(y)\right] p(y)\, dy$$

Decision theoretic solution: change of scoring rule

- Bayes factors select models maximizing $\log p(y_{1:T}) = \sum_{t=1}^{T} \log p(y_t \mid y_{1:t-1})$.
- This corresponds to minimizing the prequential score [*Dawid, 1984*]

$$\sum_{t=1}^{T} \mathcal{S}\left(y_t,\, p(dy_t \mid y_{1:t-1})\right)$$

with the choice of scoring rule $\mathcal{S}(y,\, p) = -\log p(y)$ called the log-score.

- Each scoring rule has an associated divergence function

$$\mathrm{D}_{\mathcal{S}}(p, q) = \mathbb{E}_{Y\sim p}\left[\mathcal{S}(Y, q) - \mathcal{S}(Y, p)\right]$$

$\mathcal{S}$ is (strictly) proper if $q \mapsto \mathrm{D}_{\mathcal{S}}(p, q)$ is (uniquely) minimized at $q = p$.

- The log-score is strictly proper and tied to the Kullback-Leibler divergence

$$\mathrm{KL}(p, q) = \int \left[\log p(y) - \log q(y)\right] p(y)\, dy$$

Decision theoretic solution: change of scoring rule

- Bayes factors select models maximizing $\log p(y_{1:T}) = \sum_{t=1}^{T} \log p(y_t \mid y_{1:t-1})$.
- This corresponds to minimizing the prequential score [*Dawid, 1984*]

$$\sum_{t=1}^{T} \mathcal{S}\left(y_t,\, p(dy_t \mid y_{1:t-1})\right)$$

with the choice of scoring rule $\mathcal{S}(y,\, p) = -\log p(y)$ called the log-score.

- Each scoring rule has an associated divergence function

$$\mathrm{D}_{\mathcal{S}}(p, q) = \mathbb{E}_{Y\sim p}\left[\mathcal{S}(Y, q) - \mathcal{S}(Y, p)\right]$$

$\mathcal{S}$ is (strictly) proper if $q \mapsto \mathrm{D}_{\mathcal{S}}(p, q)$ is (uniquely) minimized at $q = p$.

- The log-score is strictly proper and tied to the Kullback-Leibler divergence

$$\mathrm{KL}(p, q) = \int \left[\log p(y) - \log q(y)\right] p(y)\, dy$$

Decision theoretic solution: change of scoring rule

- Bayes factors select models maximizing $\log p(y_{1:T}) = \sum_{t=1}^{T} \log p(y_t \mid y_{1:t-1})$.
- This corresponds to minimizing the prequential score [*Dawid, 1984*]

$$\sum_{t=1}^{T} \mathcal{S}\left(y_t,\, p(dy_t \mid y_{1:t-1})\right)$$

  with the choice of scoring rule $\mathcal{S}(y,\, p) = -\log p(y)$ called the log-score.
- Each scoring rule has an associated divergence function

$$\mathrm{D}_{\mathcal{S}}(p, q) = \mathbb{E}_{Y\sim p}\left[\mathcal{S}(Y, q) - \mathcal{S}(Y, p)\right]$$

  $\mathcal{S}$ is (strictly) proper if $q \mapsto \mathrm{D}_{\mathcal{S}}(p, q)$ is (uniquely) minimized at $q = p$.
- The log-score is strictly proper and tied to the Kullback-Leibler divergence

$$\mathrm{KL}(p, q) = \int \left[\log p(y) - \log q(y)\right] p(y)\, dy$$

Change of scoring rule: the Hyvärinen score

- Instead of the KL-divergence, Dawid & Musio [2015] propose to use

$$\mathrm{D}_{\mathcal{H}}(p,q) = \int \left\| \nabla \log p(y) - \nabla \log q(y) \right\|^2 p(y)\, dy$$

  sometimes called the relative Fisher information divergence.

- It induces a scoring rule known as the Hyvärinen score [*Hyvärinen, 2005*]

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

  where the Laplacian Δ and the gradient ∇ are with respect to y. This score is strictly proper, local, homogeneous [*Parry, Dawid & Lauritzen, 2012*].

Select M_j minimizing the prequential Hyvärinen score (H-score)

$$\mathcal{H}_T(M_j) = \sum_{t=1}^{T} \mathcal{H}\left(y_t,\, p_j(dy_t \mid y_{1:t-1})\right)$$

which can be consistently estimated using SMC [*Chopin, 2002; Del Moral, Doucet & Jasra, 2006*] or SMC^2 [*Chopin, Jacob & Papaspiliopoulos, 2013*].

Change of scoring rule: the Hyvärinen score

- Instead of the KL-divergence, Dawid & Musio [2015] propose to use

$$D_{\mathcal{H}}(p, q) = \int \left\| \nabla \log p(y) - \nabla \log q(y) \right\|^2 p(y)\, dy$$

sometimes called the relative Fisher information divergence.

- It induces a scoring rule known as the Hyvärinen score [*Hyvärinen, 2005*]

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

where the Laplacian Δ and the gradient ∇ are with respect to y. This score is strictly proper, local, homogeneous [*Parry, Dawid & Lauritzen, 2012*].

Select M_j minimizing the prequential Hyvärinen score (H-score)

$$\mathcal{H}_T(M_j) = \sum_{t=1}^{T} \mathcal{H}\left(y_t,\, p_j(dy_t \mid y_{1:t-1})\right)$$

which can be consistently estimated using SMC [*Chopin, 2002; Del Moral, Doucet & Jasra, 2006*] or SMC^2 [*Chopin, Jacob & Papaspiliopoulos, 2013*].

Change of scoring rule: the Hyvärinen score

- Instead of the KL-divergence, Dawid & Musio [2015] propose to use

$$D_{\mathcal{H}}(p, q) = \int \left\| \nabla \log p(y) - \nabla \log q(y) \right\|^2 p(y)\, dy$$

sometimes called the relative Fisher information divergence.

- It induces a scoring rule known as the Hyvärinen score [*Hyvärinen, 2005*]

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

where the Laplacian Δ and the gradient ∇ are with respect to y. This score is strictly proper, local, homogeneous [*Parry, Dawid & Lauritzen, 2012*].

Select M_j minimizing the prequential Hyvärinen score (H-score)

$$\mathcal{H}_T(M_j) = \sum_{t=1}^{T} \mathcal{H}\left(y_t,\ p_j(dy_t \mid y_{1:t-1}) \right)$$

which can be consistently estimated using SMC [*Chopin, 2002; Del Moral, Doucet & Jasra, 2006*] or SMC^2 [*Chopin, Jacob & Papaspiliopoulos, 2013*].

Consistency of the H-score

Elements of proof in the univariate i.i.d. case $(d_y = 1)$

- Consider a generic model M : $Y_{1:T} \,|\, \theta \overset{\text{i.i.d.}}{\sim} p_\theta \;\; ; \;\; \theta \sim p(d\theta)$.

- By differentiating under the integral sign, the H-score $\mathcal{H}_T(M)$ equals

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p_\Theta\right) \Big| \, Y_{1:t}\right] + \sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_\Theta(Y_t)}{\partial y} \Big| \, Y_{1:t}\right)$$

where the expectations and variances are with respect to $\Theta \sim p(d\theta | y_{1:t})$.

Elements of proof in the univariate i.i.d. case $(d_y = 1)$

- Consider a generic model M : $Y_{1:T} \,|\, \theta \overset{\text{i.i.d.}}{\sim} p_\theta$; $\theta \sim p(d\theta)$.

- By differentiating under the integral sign, the H-score $\mathcal{H}_T(M)$ equals

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p_\Theta\right) \middle| Y_{1:t}\right] \; + \; \sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_\Theta(Y_t)}{\partial y} \middle| Y_{1:t}\right)$$

where the expectations and variances are with respect to $\Theta \sim p(d\theta | y_{1:t})$.

Elements of proof in the univariate i.i.d. case $(d_y = 1)$

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p_\Theta\right) \middle| Y_{1:t}\right] \quad + \quad \sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_\Theta(Y_t)}{\partial y} \middle| Y_{1:t}\right)$$

Concentration of the posterior

+ uniform integrability

+ equicontinuity

$$\mathbb{E}\left[\mathcal{H}(Y_t, p_\Theta) \middle| Y_{1:t}\right] \underset{t\to+\infty}{\approx} \mathcal{H}(Y_t, p_{\theta^\star}) \qquad \mathrm{Var}\left(\frac{\partial \log p_\Theta(y_t)}{\partial y} \middle| Y_{1:t}\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star\ a.s.} 0$$

Césaro's theorem

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}(Y_t, p_\Theta) \middle| Y_{1:t}\right] \underset{T\to+\infty}{\approx} \frac{1}{T}\sum_{t=1}^{T} \mathcal{H}(Y_t, p_{\theta^\star}) \qquad \frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_\Theta(Y_t)}{\partial y} \middle| Y_{1:t}\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star\ a.s.} 0$$

Law of large numbers

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{H}(Y_t, p_{\theta^\star}) \xrightarrow[T\to+\infty]{\mathbb{P}_\star\ a.s.} \mathbb{E}_\star\left[\mathcal{H}(Y, p_{\theta^\star})\right]$$

Elements of proof in the univariate i.i.d. case $(d_y = 1)$

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p_\Theta\right) \middle| Y_{1:t}\right] \qquad + \qquad \sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_\Theta(Y_t)}{\partial y} \middle| Y_{1:t}\right)$$

Concentration of the posterior

+ uniform integrability

+ equicontinuity

$$\mathbb{E}\left[\mathcal{H}(Y_t, p_\Theta) \middle| Y_{1:t}\right] \underset{t\to+\infty}{\approx} \mathcal{H}(Y_t, p_{\theta^\star}) \qquad \mathrm{Var}\left(\frac{\partial \log p_\Theta(y_t)}{\partial y} \middle| Y_{1:t}\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star\ a.s.} 0$$

Césaro's theorem

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}(Y_t, p_\Theta) \middle| Y_{1:t}\right] \underset{T\to+\infty}{\approx} \frac{1}{T}\sum_{t=1}^{T} \mathcal{H}(Y_t, p_{\theta^\star}) \qquad \frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_\Theta(Y_t)}{\partial y} \middle| Y_{1:t}\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star\ a.s.} 0$$

Law of large numbers

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{H}(Y_t, p_{\theta^\star}) \xrightarrow[T\to+\infty]{\mathbb{P}_\star\ a.s.} \mathbb{E}_\star\left[\mathcal{H}(Y, p_{\theta^\star})\right]$$

Elements of proof in the univariate i.i.d. case $(d_y = 1)$

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p_\Theta\right) \middle| Y_{1:t}\right] \quad + \quad \sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_\Theta(Y_t)}{\partial y} \middle| Y_{1:t}\right)$$

Concentration of the posterior

+ uniform integrability

+ equicontinuity

$$\mathbb{E}\left[\mathcal{H}(Y_t, p_\Theta) \middle| Y_{1:t}\right] \underset{t\to+\infty}{\approx} \mathcal{H}(Y_t, p_{\theta^\star}) \qquad \mathrm{Var}\left(\frac{\partial \log p_\Theta(y_t)}{\partial y} \middle| Y_{1:t}\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star\ a.s.} 0$$

Césaro's theorem

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}(Y_t, p_\Theta) \middle| Y_{1:t}\right] \underset{T\to+\infty}{\approx} \frac{1}{T}\sum_{t=1}^{T} \mathcal{H}(Y_t, p_{\theta^\star}) \qquad \frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_\Theta(Y_t)}{\partial y} \middle| Y_{1:t}\right) \xrightarrow[T\to+\infty]{\mathbb{P}_\star\ a.s.} 0$$

Law of large numbers

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{H}(Y_t, p_{\theta^\star}) \xrightarrow[T\to+\infty]{\mathbb{P}_\star\ a.s.} \mathbb{E}_\star\left[\mathcal{H}(Y, p_{\theta^\star})\right]$$

Elements of proof in the univariate i.i.d. case $(d_y = 1)$

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}\left(Y_t, p_\Theta\right) \middle| Y_{1:t}\right] \qquad + \qquad \sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_\Theta(Y_t)}{\partial y} \middle| Y_{1:t}\right)$$

Concentration of the posterior

+ uniform integrability

+ equicontinuity

$$\mathbb{E}\left[\mathcal{H}(Y_t, p_\Theta) \middle| Y_{1:t}\right] \underset{t \to +\infty}{\approx} \mathcal{H}(Y_t, p_{\theta^\star}) \qquad \mathrm{Var}\left(\frac{\partial \log p_\Theta(y_t)}{\partial y} \middle| Y_{1:t}\right) \xrightarrow[T \to +\infty]{\mathbb{P}_\star\ a.s.} 0$$

Césaro's theorem

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{H}(Y_t, p_\Theta) \middle| Y_{1:t}\right] \underset{T \to +\infty}{\approx} \frac{1}{T}\sum_{t=1}^{T} \mathcal{H}(Y_t, p_{\theta^\star}) \qquad \frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}\left(\frac{\partial \log p_\Theta(Y_t)}{\partial y} \middle| Y_{1:t}\right) \xrightarrow[T \to +\infty]{\mathbb{P}_\star\ a.s.} 0$$

Law of large numbers

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{H}(Y_t, p_{\theta^\star}) \xrightarrow[T \to +\infty]{\mathbb{P}_\star\ a.s.} \mathbb{E}_\star\left[\mathcal{H}(Y, p_{\theta^\star})\right]$$

Consistency result in the univariate i.i.d. case

- Under regularity conditions, the H-factor of M_1 vs. M_2 satisfies

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

where $\mathrm{D}_{\mathcal{H}}(p_\star, M_j) := \mathbb{E}_\star[\mathcal{H}(Y, p_{\theta_j^\star})] - \mathbb{E}_\star[\mathcal{H}(Y, p_\star)]$.

- In contrast, the log-Bayes factor of M_1 vs. M_2 satisfies

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

where $\mathrm{KL}(p_\star, M_j) := \mathbb{E}_\star[-\log p_{\theta_j^\star}(Y)] - \mathbb{E}_\star[-\log p_\star(Y)]$.

- This extends to state-space models and dependent data, with additional technicalities (e.g. forgetting properties and ergodic theorems).

- The limit is meaningless if $p_\star$ belongs to both models (e.g. nested well-specified setting): we need higher order Bayesian asymptotics , i.e. Bernstein-von-Mises-type results ... which are non-trivial for state-space models ⚠

Consistency result in the univariate i.i.d. case

- Under regularity conditions, the H-factor of M_1 vs. M_2 satisfies

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

  where $\mathrm{D}_{\mathcal{H}}(p_\star, M_j) := \mathbb{E}_\star[\mathcal{H}(Y, p_{\theta_j^\star})] - \mathbb{E}_\star[\mathcal{H}(Y, p_\star)]$.

- In contrast, the log-Bayes factor of M_1 vs. M_2 satisfies

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

  where $\mathrm{KL}(p_\star, M_j) := \mathbb{E}_\star[-\log p_{\theta_j^\star}(Y)] - \mathbb{E}_\star[-\log p_\star(Y)]$.

- This extends to state-space models and dependent data, with additional technicalities (e.g. forgetting properties and ergodic theorems).

- The limit is meaningless if $p_\star$ belongs to both models (e.g. nested well-specified setting): we need higher order Bayesian asymptotics , i.e. Bernstein-von-Mises-type results ... which are non-trivial for state-space models ⚠

Consistency result in the univariate i.i.d. case

- Under regularity conditions, the H-factor of M_1 vs. M_2 satisfies

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

where $\mathrm{D}_{\mathcal{H}}(p_\star, M_j) := \mathbb{E}_\star[\mathcal{H}(Y, p_{\theta_j^\star})] - \mathbb{E}_\star[\mathcal{H}(Y, p_\star)]$.

- In contrast, the log-Bayes factor of M_1 vs. M_2 satisfies

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

where $\mathrm{KL}(p_\star, M_j) := \mathbb{E}_\star[-\log p_{\theta_j^\star}(Y)] - \mathbb{E}_\star[-\log p_\star(Y)]$.

- This extends to state-space models and dependent data, with additional technicalities (e.g. forgetting properties and ergodic theorems).

- The limit is meaningless if $p_\star$ belongs to both models (e.g. nested well-specified setting): we need higher order Bayesian asymptotics , i.e. Bernstein-von-Mises-type results ... which are non-trivial for state-space models ⚠

Consistency result in the univariate i.i.d. case

- Under regularity conditions, the H-factor of M_1 vs. M_2 satisfies

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

where $\mathrm{D}_{\mathcal{H}}(p_\star, M_j) := \mathbb{E}_\star[\mathcal{H}(Y, p_{\theta_j^\star})] - \mathbb{E}_\star[\mathcal{H}(Y, p_\star)]$.

- In contrast, the log-Bayes factor of M_1 vs. M_2 satisfies

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

where $\mathrm{KL}(p_\star, M_j) := \mathbb{E}_\star[-\log p_{\theta_j^\star}(Y)] - \mathbb{E}_\star[-\log p_\star(Y)]$.

- This extends to state-space models and dependent data, with additional technicalities (e.g. forgetting properties and ergodic theorems).

- The limit is meaningless if $p_\star$ belongs to both models (e.g. nested well-specified setting): we need higher order Bayesian asymptotics , i.e. Bernstein-von-Mises-type results … which are non-trivial for state-space models ⚠

Illustration of consistency for i.i.d. observations

- **Example 1**. Given simulated $Y_1, \ldots, Y_{1000} \sim \mathcal{N}(\mu_\star, \sigma_\star^2)$, we compare

$$M_1: \quad Y_1, \ldots, Y_T \mid \theta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\theta_1, 1\right), \quad \theta_1 \sim \mathcal{N}\left(0, 10\right)$$
$$M_2: \quad Y_1, \ldots, Y_T \mid \theta_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \theta_2\right), \quad \theta_2 \sim \text{Inv-}\chi^2\left(0.1, 1\right)$$

in the following four cases $(\mu_\star, \sigma_\star^2) = (1,1), (0,5), (4,3), (0,1)$.

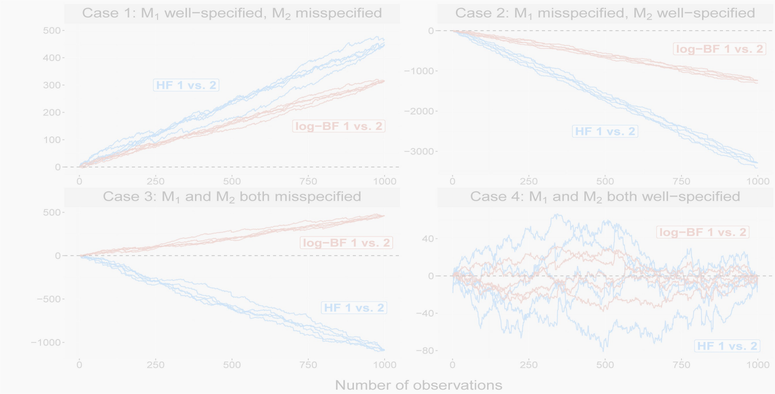
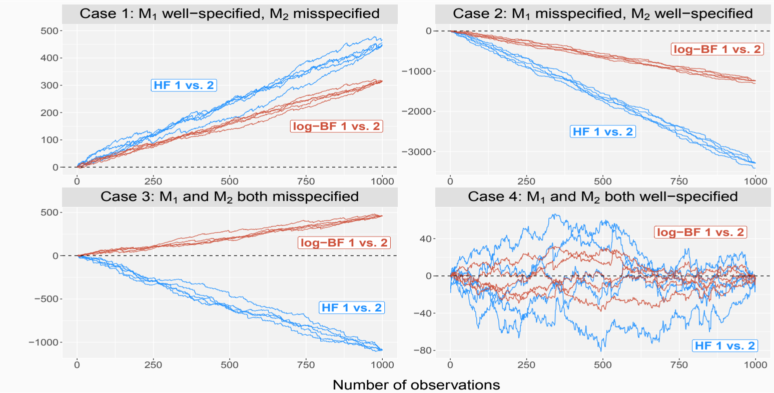


Illustration of consistency for i.i.d. observations

- **Example 1**. Given simulated $Y_1, \ldots, Y_{1000} \sim \mathcal{N}(\mu_\star, \sigma_\star^2)$, we compare

$$M_1: \quad Y_1, \ldots, Y_T \,|\, \theta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_1, 1), \quad \theta_1 \sim \mathcal{N}(0, 10)$$
$$M_2: \quad Y_1, \ldots, Y_T \,|\, \theta_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \theta_2), \quad \theta_2 \sim \text{Inv-}\chi^2(0.1, 1)$$

in the following four cases $(\mu_\star, \sigma_\star^2) = (1,1), (0,5), (4,3), (0,1)$.



Summary

The H-score has the advantage of being ...

- Robust to vagueness of priors and allows for improper priors
- Justified non-asymptotically and also generally consistent
- Applicable to a wide range of parametric models via SMC methods

... albeit at the cost of more regularity on the candidate densities and more expensive computation in practice.

Avenues for future research

- Confidence intervals using unbiased MCMC [Jacob, O'Leary, Atchadé, 2018]
- Posterior consistency + asymptotic Normality for state-space models

- More details in Shao, Jacob, Ding & Tarokh (2018)
- R package available at: github.com/pierrejacob/bayeshscore

Questions ?

References (1/2)

O. E. Barndorff-Nielsen and N. Shephard.
Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics.
Journal of the Royal Statistical Society: Series B, 63(2):167––241, 2001.

N. Chopin.
A sequential particle filter method for static models.
89:539–552, 2002.

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, 75 (3):397–426, 2013.

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

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

References (2/2)

J. Knape and P. D. Valpine.
Fitting complex population models by combining particle filters with markov chain monte carlo.
Ecology, 93 (2):256–263, 2012.

A. O'Hagan.
Fractional bayes factor for model comparison.
Journal of the Royal Statistical Society, 57 (1):99–138, 1995.

M. Parry, A. P. Dawid, and S. Lauritzen.
Proper local scoring rules.
The Annals of Statistics, 40 (1):561–592, 2012.

S. Shao, P. E. Jacob, J. Ding, and V. Tarokh.
Bayesian model comparison with the Hyvärinen score: computation and consistency.
Journal of the American Statistical Association. DOI: 10.1080/01621459.2018.1518237, 2018.

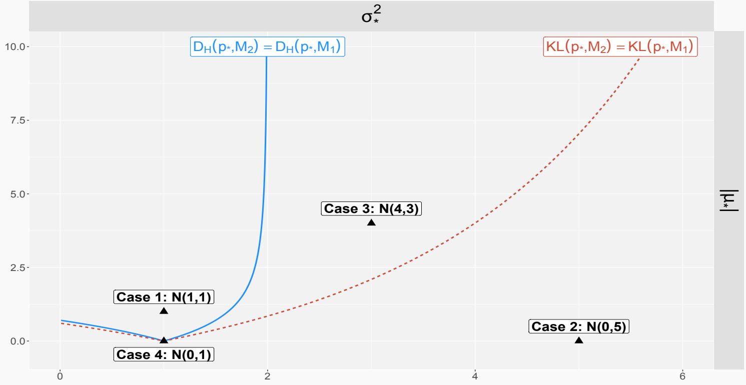
Illustration of consistency for i.i.d. observations (continued)

- **Example 1.** Given simulated $Y_1, \dots, Y_{1000} \sim \mathcal{N}(\mu_\star, \sigma_\star^2)$, we compare

$$M_1: \quad Y_1, \dots, Y_T \mid \theta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_1, 1), \quad \theta_1 \sim \mathcal{N}(0, 10)$$

$$M_2: \quad Y_1, \dots, Y_T \mid \theta_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \theta_2), \quad \theta_2 \sim \text{Inv-}\chi^2(0.1, 1)$$

in the following four cases $(\mu_\star, \sigma_\star^2) = (1,1), (0,5), (4,3), (0,1)$.



Nested models in the univariate i.i.d. case

- **Example 1**. Given simulated $Y_1, \ldots, Y_{1000} \sim \mathcal{N}(\mu_\star, \sigma_\star^2)$, we compare

$$M_1: \quad Y_1, \ldots, Y_T \mid \theta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_1, 1), \quad \mu_1 \sim \mathcal{N}(0, 10)$$

$$M_2: \quad Y_1, \ldots, Y_T \mid \mu_2, \sigma_2^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2), \quad (\mu_2, \sigma_2^2) \sim \mathcal{N}\text{-Inv-}\chi^2(0, 1, 0.1, 1)$$

in the following two cases $(\mu_\star, \sigma_\star^2) = (0, 5), (0, 1)$.