Bayesian model comparison with the Hyvärinen score

Stephane Shao*
November 21, 2017

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

*Department of Statistics, Harvard University
†School of Engineering and Applied Sciences, Harvard University

Table of contents

Table of contents

Table of contents

Table of contents

Table of contents

Table of contents

Setting and examples

Goal: compare models in a Bayesian framework

- Given T observations $Y_{1:T} = (Y_1, ..., Y_T) \in (\mathbb{R}^{d_y})^T$.
- We want to compare candidate models from a finite set $\{M_1, ..., M_q\}$.
- Each model M_j is a collection of distributions $p_j(dy_{1:T}|\theta_j)$ parametrized by $\theta_j \in \mathbb{T}_j$ with a prior distribution $p_j(d\theta_j)$ on the parameter.
- We are interested in settings where the prior may be vague or improper.

Goal: compare models in a Bayesian framework

- Given T observations $Y_{1:T} = (Y_1, ..., Y_T) \in (\mathbb{R}^{d_y})^T$.
- We want to compare candidate models from a finite set $\{M_1, ..., M_q\}$.
- Each model M_j is a collection of distributions $p_j(dy_{1:T}|\theta_j)$ parametrized by $\theta_j \in \mathbb{T}_j$ with a prior distribution $p_j(d\theta_j)$ on the parameter.
- We are interested in settings where the prior may be vague or improper.

Goal: compare models in a Bayesian framework

- Given T observations $Y_{1:T} = (Y_1, ..., Y_T) \in (\mathbb{R}^{d_y})^T$.
- We want to compare candidate models from a finite set $\{M_1, ..., M_q\}$.
- **Each model M_j is a collection of distributions $p_j(dy_{1:T}|\theta_j)$ parametrized by $\theta_j \in \mathbb{T}_j$ with a prior distribution $p_j(d\theta_j)$ on the parameter.**
- We are interested in settings where the prior may be vague or improper.

Goal: compare models in a Bayesian framework

- Given T observations $Y_{1:T} = (Y_1, \ldots, Y_T) \in (\mathbb{R}^{d_y})^T$.
- We want to compare candidate models from a finite set $\{M_1, \ldots, M_q\}$.
- Each model M_j is a collection of distributions $p_j(dy_{1:T}|\theta_j)$ parametrized by $\theta_j \in \mathbb{T}_j$ with a prior distribution $p_j(d\theta_j)$ on the parameter.
- We are interested in settings where the prior may be vague or improper.

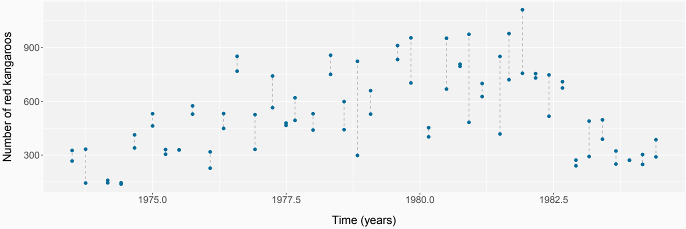
- **Example 1**. Normal i.i.d. models (*O'Hagan, 1995*)

Model 1	Model 2
$Y_1, \ldots, Y_T \mid \theta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_1, 1)$	$Y_1, \ldots, Y_T \mid \theta_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \theta_2)$
$\theta_1 \sim \mathcal{N}(0, \sigma_0^2)$	$\theta_2 \sim \text{Inv-}\chi^2(\nu_0, s_0^2)$

with known hyperparameters $\sigma_0 > 0$, $\nu_0 > 0$, and $s_0 > 0$.

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

- **Example 2.** Population dynamics of red kangaroos (*Knape & de Valpine, 2012*)



Model 1	Model 2	Model 3
Given parameters (b,r,σ,τ):	Given parameters (r,σ,τ):	Given parameters (σ,τ):
$X_1 \sim \mathrm{LN}(0,5)$	$X_1 \sim \mathrm{LN}(0,5)$	$X_1 \sim \mathrm{LN}(0,5)$
$dX_t/X_t = (r - bX_t + \frac{\sigma^2}{2})\,dt + \sigma dW_t$	$dX_t/X_t = (r + \frac{\sigma^2}{2})\,dt + \sigma dW_t$	$dX_t/X_t = (\frac{\sigma^2}{2})\,dt + \sigma dW_t$
$Y_{1,t}\,,\,Y_{2,t} \mid X_t \overset{\text{i.i.d.}}{\sim} \mathrm{NB}(X_t, X_t + \tau X_t^2)$	$Y_{1,t}\,,\,Y_{2,t} \mid X_t \overset{\text{i.i.d.}}{\sim} \mathrm{NB}(X_t, X_t + \tau X_t^2)$	$Y_{1,t}\,,\,Y_{2,t} \mid X_t \overset{\text{i.i.d.}}{\sim} \mathrm{NB}(X_t, X_t + \tau X_t^2)$
with independent priors:	with independent priors:	with independent priors:
$b, \sigma, \tau \sim \mathrm{Unif}(0,10)$	$\sigma, \tau \sim \mathrm{Unif}(0,10)$	$\sigma, \tau \sim \mathrm{Unif}(0,10)$
$r \sim \mathrm{Unif}(-10,10)$	$r \sim \mathrm{Unif}(-10,10)$	

Why not use Bayes factors ?

Limitations of Bayes factors: sensitivity to vague priors

- Bayes factors favor the model M_j with the largest *evidence*

$$p_j(y_{1:T}) = \int p_j(y_{1:T}|\theta_j)\, p_j(\theta_j)\, d\theta_j$$

(a.k.a. the *marginal likelihood* of M_j)

Sensitivity to the choice of prior

- The evidence of any model can be made arbitrarily small by making the prior arbitrarily vague, regardless of the sample size.
- Bayes factors do not allow for improper priors.

Yet, vague or improper priors often stem from reasonable approaches
(genuine non-informativeness, Jeffreys prior, …)

Limitations of Bayes factors: sensitivity to vague priors

- Bayes factors favor the model M_j with the largest *evidence*

$$p_j(y_{1:T}) = \int p_j(y_{1:T}|\theta_j)\, p_j(\theta_j)\, d\theta_j$$

(a.k.a. the *marginal likelihood* of M_j)

Sensitivity to the choice of prior

- The evidence of any model can be made arbitrarily small by making the prior arbitrarily vague, regardless of the sample size.
- Bayes factors do not allow for improper priors.

Yet, vague or improper priors often stem from reasonable approaches
(genuine non-informativeness, Jeffreys prior, …)

Limitations of Bayes factors: sensitivity to vague priors

- Bayes factors favor the model M_j with the largest *evidence*

$$p_j(y_{1:T}) = \int p_j(y_{1:T}|\theta_j)\, p_j(\theta_j)\, d\theta_j$$

(a.k.a. the *marginal likelihood* of M_j)

Sensitivity to the choice of prior

- The evidence of any model can be made arbitrarily small by making the prior arbitrarily vague, regardless of the sample size.
- Bayes factors do not allow for improper priors.

Yet, vague or improper priors often stem from reasonable approaches

(genuine non-informativeness, Jeffreys prior, …)

A new criterion: the Hyvärinen score

Bayes factors relate to a particular choice of scoring rule

- Bayes factors favor the model minimizing $-\log p(y_{1:T})$ or equivalently

$$\sum_{t=1}^{T} -\log p(y_t|y_{1:t-1})$$

- This is a particular case of a more general decision rule that favors the model with the smallest prequential score

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

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

Key idea: use the Hyvärinen score (H-score) instead (*Dawid & Musio, 2015*)

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

with $\mathcal{H}(y, p) = 2\,\Delta \log p(y) + \|\nabla \log p(y)\|^2$, where the Laplacian Δ and the gradient ∇ are taken with respect to the observation y

Bayes factors relate to a particular choice of scoring rule

- Bayes factors favor the model minimizing $-\log p(y_{1:T})$ or equivalently

$$\sum_{t=1}^{T} -\log p(y_t|y_{1:t-1})$$

- This is a particular case of a more general decision rule that favors the model with the smallest prequential score

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

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

Key idea: use the Hyvärinen score (H-score) instead (*Dawid & Musio, 2015*)

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

with $\mathcal{H}(y,\, p) = 2\,\Delta \log p(y) + \|\nabla \log p(y)\|^2$, where the Laplacian Δ and the gradient ∇ are taken with respect to the observation y

Bayes factors relate to a particular choice of scoring rule

- Bayes factors favor the model minimizing $-\log p(y_{1:T})$ or equivalently

$$\sum_{t=1}^{T} -\log p(y_t|y_{1:t-1})$$

- This is a particular case of a more general decision rule that favors the model with the smallest prequential score

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

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

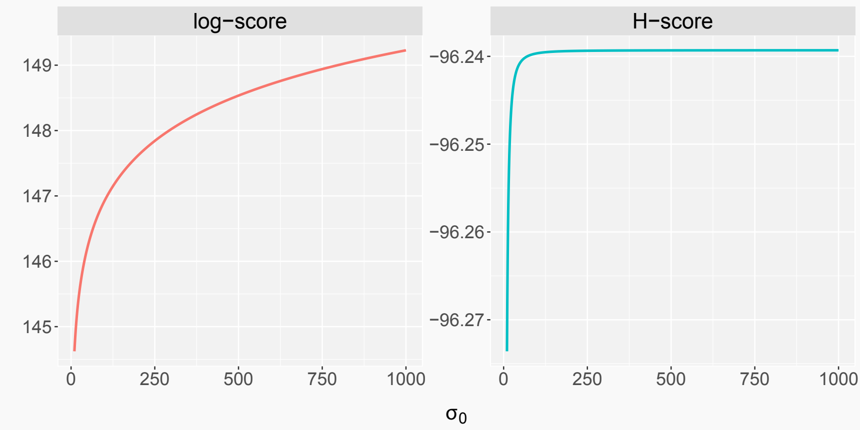
Key idea: use the Hyvärinen score (H-score) instead (*Dawid & Musio, 2015*)

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

with $\mathcal{H}(y, p) = 2\,\Delta \log p(y) + \|\nabla \log p(y)\|^2$, where the Laplacian Δ and the gradient ∇ are taken with respect to the observation y

H-scores are robust to prior vagueness

- **Example**. Given i.i.d. realizations $y_1, \ldots, y_{100}$ from a $\mathcal{N}(0,1)$, consider the model describing $Y_1, \ldots, Y_{100} \mid \mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ with prior $\mu \sim \mathcal{N}(0, \sigma_0^2)$.



Using H-scores is principled and justified non-asymptotically

- For univariate observations, the H-score can be written explicitly as

$$\sum_{t=1}^{T}\left(2\,\frac{\partial^2 \log p(y_t|y_{1:t-1})}{\partial {y_t}^2}+\left(\frac{\partial \log p(y_t|y_{1:t-1})}{\partial y_t}\right)^2\right)$$

- Similarly to the log-score, the H-score is
  - Proper: the expected loss $\mathbb{E}_\star\left[\mathcal{H}(Y, p)\right]$ under $Y \sim p_\star$ is minimized at $p = p_\star$
  - m-local: $\mathcal{H}(y, p)$ is a function of y and $p(y), p'(y), \ldots, p^{(m)}(y) \quad$ (with $m = 2$)
- The H-score has the additional advantage of being 0-homogeneous
  - $\mathcal{H}(y, p)$ is unchanged when multiplying $p(y), p'(y), \ldots, p^{(m)}(y)$ by any $\lambda > 0$
- The H-score is the simplest scoring rule satisfying propriety, m-locality, and 0-homogeneity (*Parry, Dawid & Lauritzen, 2012*)
- It can be extended to discrete observations via finite differences, while preserving all the above properties (*Dawid, Parry & Lauritzen, 2012*)
- **But it involves derivatives of typically intractable predictive densities $p(y_t|y_{1:t-1})$ at every $t \in \{1, \ldots, T\}$. This calls for sequential estimation.**

Using H-scores is principled and justified non-asymptotically

- For univariate observations, the H-score can be written explicitly as

$$\sum_{t=1}^{T}\left(2\,\frac{\partial^2 \log p(y_t|y_{1:t-1})}{\partial {y_t}^2}+\left(\frac{\partial \log p(y_t|y_{1:t-1})}{\partial y_t}\right)^2\right)$$

- **Similarly to the log-score, the H-score is**
  - Proper: the expected loss $\mathbb{E}_\star\left[\mathcal{H}(Y, p)\right]$ under $Y \sim p_\star$ is minimized at $p = p_\star$
  - m-local: $\mathcal{H}(y, p)$ is a function of y and $p(y), p'(y), \ldots, p^{(m)}(y)$ (with $m = 2$)
- The H-score has the additional advantage of being 0-homogeneous
  - $\mathcal{H}(y, p)$ is unchanged when multiplying $p(y), p'(y), \ldots, p^{(m)}(y)$ by any $\lambda > 0$
- The H-score is the simplest scoring rule satisfying propriety, m-locality, and 0-homogeneity (*Parry, Dawid & Lauritzen, 2012*)
- It can be extended to discrete observations via finite differences, while preserving all the above properties (*Dawid, Parry & Lauritzen, 2012*)
- **But it involves derivatives of typically intractable predictive densities $p(y_t|y_{1:t-1})$ at every $t \in \{1, \ldots, T\}$. This calls for sequential estimation.**

Using H-scores is principled and justified non-asymptotically

- For univariate observations, the H-score can be written explicitly as

$$\sum_{t=1}^{T}\left(2\,\frac{\partial^2 \log p(y_t|y_{1:t-1})}{\partial {y_t}^2}+\left(\frac{\partial \log p(y_t|y_{1:t-1})}{\partial y_t}\right)^2\right)$$

- Similarly to the log-score, the H-score is
  - **Proper**: the expected loss $\mathbb{E}_\star\left[\mathcal{H}(Y,\,p)\right]$ under $Y \sim p_\star$ is minimized at $p = p_\star$
  - *m*-**local**: $\mathcal{H}(y,\,p)$ is a function of y and $p(y), p'(y), \ldots, p^{(m)}(y)$ (with $m = 2$)
- The H-score has the additional advantage of being 0-homogeneous
  - $\mathcal{H}(y,\,p)$ is unchanged when multiplying $p(y), p'(y), \ldots, p^{(m)}(y)$ by any $\lambda > 0$
- The H-score is the simplest scoring rule satisfying propriety, m-locality, and 0-homogeneity (*Parry, Dawid & Lauritzen, 2012*)
- It can be extended to discrete observations via finite differences, while preserving all the above properties (*Dawid, Parry & Lauritzen, 2012*)
- **But it involves derivatives of typically intractable predictive densities $p(y_t|y_{1:t-1})$ at every $t \in \{1, \ldots, T\}$. This calls for sequential estimation.**

Using H-scores is principled and justified non-asymptotically

- For univariate observations, the H-score can be written explicitly as

$$\sum_{t=1}^{T} \left(2 \, \frac{\partial^2 \log p(y_t|y_{1:t-1})}{\partial {y_t}^2} + \left(\frac{\partial \log p(y_t|y_{1:t-1})}{\partial y_t} \right)^2 \right)$$

- Similarly to the log-score, the H-score is
  - **Proper**: the expected loss $\mathbb{E}_\star \left[\mathcal{H}(Y,\, p) \right]$ under $Y \sim p_\star$ is minimized at $p = p_\star$
  - **m-local**: $\mathcal{H}(y,\, p)$ is a function of y and $p(y), p'(y), \dots, p^{(m)}(y)$ (with $m = 2$)
- **The H-score has the additional advantage of being 0-homogeneous**
  - $\mathcal{H}(y,\, p)$ is unchanged when multiplying $p(y),\, p'(y),\, \dots,\, p^{(m)}(y)$ by any $\lambda > 0$
- The H-score is the simplest scoring rule satisfying propriety, m-locality, and 0-homogeneity (*Parry, Dawid & Lauritzen, 2012*)
- It can be extended to discrete observations via finite differences, while preserving all the above properties (*Dawid, Parry & Lauritzen, 2012*)
- **But it involves derivatives of typically intractable predictive densities $p(y_t|y_{1:t-1})$ at every $t \in \{1, \dots, T\}$. This calls for sequential estimation.**

Using H-scores is principled and justified non-asymptotically

- For univariate observations, the H-score can be written explicitly as

$$\sum_{t=1}^{T} \left(2 \, \frac{\partial^2 \log p(y_t|y_{1:t-1})}{\partial {y_t}^2} + \left(\frac{\partial \log p(y_t|y_{1:t-1})}{\partial y_t} \right)^2 \right)$$

- Similarly to the log-score, the H-score is
  - **Proper**: the expected loss $\mathbb{E}_\star \left[\mathcal{H}(Y, p) \right]$ under $Y \sim p_\star$ is minimized at $p = p_\star$
  - **m-local**: $\mathcal{H}(y, p)$ is a function of y and $p(y), p'(y), \ldots, p^{(m)}(y)$ (with $m = 2$)
- The H-score has the additional advantage of being **0-homogeneous**
  - $\mathcal{H}(y, p)$ is unchanged when multiplying $p(y)$, $p'(y)$, ..., $p^{(m)}(y)$ by any $\lambda > 0$
- **The H-score is the simplest scoring rule satisfying propriety, m-locality, and 0-homogeneity** (*Parry, Dawid & Lauritzen, 2012*)
- It can be extended to discrete observations via finite differences, while preserving all the above properties (*Dawid, Parry & Lauritzen, 2012*)
- **But it involves derivatives of typically intractable predictive densities $p(y_t|y_{1:t-1})$ at every $t \in \{1, \ldots, T\}$. This calls for sequential estimation.**

Using H-scores is principled and justified non-asymptotically

- For univariate observations, the H-score can be written explicitly as

$$\sum_{t=1}^{T} \left(2 \, \frac{\partial^2 \log p(y_t|y_{1:t-1})}{\partial {y_t}^2} + \left(\frac{\partial \log p(y_t|y_{1:t-1})}{\partial y_t} \right)^2 \right)$$

- Similarly to the log-score, the H-score is
  - **Proper**: the expected loss $\mathbb{E}_\star \left[\mathcal{H}(Y, p) \right]$ under $Y \sim p_\star$ is minimized at $p = p_\star$
  - ***m*-local**: $\mathcal{H}(y, p)$ is a function of y and $p(y), p'(y), \ldots, p^{(m)}(y)$ (with $m = 2$)
- The H-score has the additional advantage of being **0-homogeneous**
  - $\mathcal{H}(y, p)$ is unchanged when multiplying $p(y)$, $p'(y)$, ..., $p^{(m)}(y)$ by any $\lambda > 0$
- The H-score is the simplest scoring rule satisfying propriety, m-locality, and 0-homogeneity (*Parry, Dawid & Lauritzen, 2012*)
- **It can be extended to discrete observations via finite differences, while preserving all the above properties** (*Dawid, Parry & Lauritzen, 2012*)
- But it involves derivatives of typically intractable predictive densities $p(y_t|y_{1:t-1})$ at every $t \in \{1, \ldots, T\}$. This calls for sequential estimation.

Using H-scores is principled and justified non-asymptotically

- For univariate observations, the H-score can be written explicitly as

$$\sum_{t=1}^{T}\left(2\,\frac{\partial^2 \log p(y_t|y_{1:t-1})}{\partial {y_t}^2}+\left(\frac{\partial \log p(y_t|y_{1:t-1})}{\partial y_t}\right)^2\right)$$

- Similarly to the log-score, the H-score is
  - Proper: the expected loss $\mathbb{E}_\star\left[\mathcal{H}(Y,p)\right]$ under $Y\sim p_\star$ is minimized at $p=p_\star$
  - m-local: $\mathcal{H}(y,p)$ is a function of y and $p(y), p'(y), \dots, p^{(m)}(y)$ (with $m=2$)
- The H-score has the additional advantage of being 0-homogeneous
  - $\mathcal{H}(y,p)$ is unchanged when multiplying $p(y)$, $p'(y)$, ..., $p^{(m)}(y)$ by any $\lambda>0$
- The H-score is the simplest scoring rule satisfying propriety, m-locality, and 0-homogeneity (*Parry, Dawid & Lauritzen, 2012*)
- It can be extended to discrete observations via finite differences, while preserving all the above properties (*Dawid, Parry & Lauritzen, 2012*)
- But it involves derivatives of typically intractable predictive densities $p(y_t|y_{1:t-1})$ at every $t\in\{1,\dots,T\}$. This calls for sequential estimation.

How to estimate H-scores ?

H-scores via SMC if likelihoods can be evaluated

- By differentiating under the integral sign, the H-score turns out to be

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

where $\mathbb{E}_t$ denotes posterior expectations with respect to $\Theta \sim p(d\theta|y_{1:t})$

Estimation using SMC (*Chopin, 2002; Del Moral, Doucet & Jasra, 2006*)

H-scores can be consistently estimated using standard SMC samplers, as long as one can evaluate the incremental likelihoods $p(y_t|y_{1:t-1},\Theta)$

- However, incremental likelihoods $p(y_t|y_{1:t-1},\Theta)$ are typically intractable for general state-space models.

H-scores via SMC if likelihoods can be evaluated

- By differentiating under the integral sign, the H-score turns out to be

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

where $\mathbb{E}_t$ denotes posterior expectations with respect to $\Theta \sim p(d\theta|y_{1:t})$

Estimation using SMC (*Chopin, 2002; Del Moral, Doucet & Jasra, 2006*)

H-scores can be consistently estimated using standard SMC samplers, as long as one can evaluate the incremental likelihoods $p(y_t|y_{1:t-1},\Theta)$

- However, incremental likelihoods $p(y_t|y_{1:t-1},\Theta)$ are typically intractable for general state-space models.

H-scores via SMC if likelihoods can be evaluated

- By differentiating under the integral sign, the H-score turns out to be

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

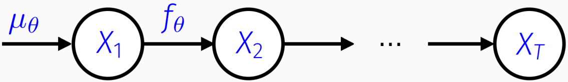
where $\mathbb{E}_t$ denotes posterior expectations with respect to $\Theta \sim p(d\theta|y_{1:t})$

Estimation using SMC (*Chopin, 2002; Del Moral, Doucet & Jasra, 2006*)

H-scores can be consistently estimated using standard SMC samplers, as long as one can evaluate the incremental likelihoods $p(y_t|y_{1:t-1},\Theta)$

- **However, incremental likelihoods $p(y_t|y_{1:t-1},\Theta)$ are typically intractable for general state-space models.**

What is so challenging about state-space models ?



- Unobserved Markov chain of latent states $X_1, \ldots, X_T$ with

$$X_1 \sim \mu_\theta \quad \text{and} \quad X_t \,|\, X_{t-1} \sim f_\theta(\cdot|X_{t-1}) \quad \text{for } t \geq 2$$

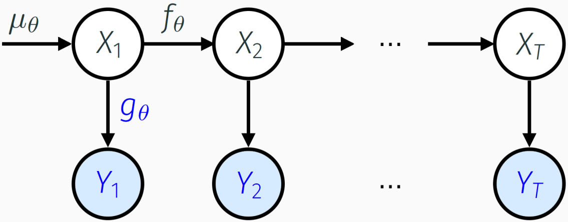
- Observations $Y_1, \ldots, Y_T$ conditionally independent given $X_1, \ldots, X_T$ with

$$Y_t \,|\, X_t \sim g_\theta(\cdot|X_t) \quad \text{for } t \geq 1$$

The likelihood is generally intractable

$$p(y_{1:T}|\theta) = \int \mu_\theta(x_1) \prod_{t=2}^{T} f_\theta(x_t|x_{t-1}) \prod_{t=1}^{T} g_\theta(y_t|x_t)\, dx_{1:T}$$

What is so challenging about state-space models ?



- Unobserved Markov chain of latent states $X_1, \ldots, X_T$ with

$$X_1 \sim \mu_\theta \quad \text{and} \quad X_t \mid X_{t-1} \sim f_\theta(\cdot|X_{t-1}) \quad \text{for } t \geq 2$$

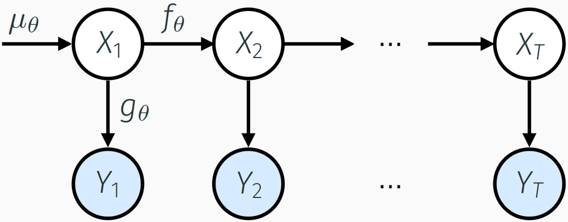
- Observations $Y_1, \ldots, Y_T$ conditionally independent given $X_1, \ldots, X_T$ with

$$Y_t \mid X_t \sim g_\theta(\cdot|X_t) \quad \text{for } t \geq 1$$

The likelihood is generally intractable

$$p(y_{1:T}|\theta) = \int \mu_\theta(x_1) \prod_{t=2}^{T} f_\theta(x_t|x_{t-1}) \prod_{t=1}^{T} g_\theta(y_t|x_t)\, dx_{1:T}$$

What is so challenging about state-space models ?



- Unobserved Markov chain of latent states $X_1, \ldots, X_T$ with

$$X_1 \sim \mu_\theta \quad \text{and} \quad X_t \,|\, X_{t-1} \sim f_\theta(\cdot|X_{t-1}) \quad \text{for } t \geq 2$$

- Observations $Y_1, \ldots, Y_T$ conditionally independent given $X_1, \ldots, X_T$ with

$$Y_t \,|\, X_t \sim g_\theta(\cdot|X_t) \quad \text{for } t \geq 1$$

The likelihood is generally intractable

$$p(y_{1:T}|\theta) = \int \mu_\theta(x_1) \prod_{t=2}^{T} f_\theta(x_t|x_{t-1}) \prod_{t=1}^{T} g_\theta(y_t|x_t)\, dx_{1:T}$$

H-scores via SMC^2 for general state-space models

- Under further integrability conditions, the H-score proves to equal

$$\sum_{t=1}^{T}\left(2\,\mathbb{E}_t\left[\frac{\partial^2 \log g_\Theta(y_t|X_t)}{\partial y_t^2}+\left(\frac{\partial \log g_\Theta(y_t|X_t)}{\partial y_t}\right)^2\right]-\left(\mathbb{E}_t\left[\frac{\partial \log g_\Theta(y_t|X_t)}{\partial y_t}\right]\right)^2\right)$$

where the expectations $\mathbb{E}_t$ are now with respect to the joint posterior distributions $(\Theta, X_t) \sim p(d\theta|y_{1:t})p(dx_t|y_{1:t},\theta)$

Estimation using SMC^2 (*Chopin, Jacob & Papaspiliopoulos, 2013*)

H-scores can be consistently estimated for state-space models, as long as one can simulate transitions from f_θ and evaluate g_θ

- SMC and SMC^2 can be used similarly in the case of discrete observations.

H-scores via SMC^2 for general state-space models

- Under further integrability conditions, the H-score proves to equal

$$\sum_{t=1}^{T}\left(2\,\mathbb{E}_t\left[\frac{\partial^2 \log g_\Theta(y_t|X_t)}{\partial y_t^2}+\left(\frac{\partial \log g_\Theta(y_t|X_t)}{\partial y_t}\right)^2\right]-\left(\mathbb{E}_t\left[\frac{\partial \log g_\Theta(y_t|X_t)}{\partial y_t}\right]\right)^2\right)$$

where the expectations $\mathbb{E}_t$ are now with respect to the joint posterior distributions $(\Theta, X_t) \sim p(d\theta|y_{1:t})p(dx_t|y_{1:t}, \theta)$

Estimation using SMC^2 (*Chopin, Jacob & Papaspiliopoulos, 2013*)

H-scores can be consistently estimated for state-space models, as long as one can simulate transitions from f_θ and evaluate g_θ

- SMC and SMC^2 can be used similarly in the case of discrete observations.

H-scores via SMC^2 for general state-space models

- Under further integrability conditions, the H-score proves to equal

$$\sum_{t=1}^{T}\left(2\,\mathbb{E}_t\left[\frac{\partial^2 \log g_\Theta(y_t|X_t)}{\partial y_t^2}+\left(\frac{\partial \log g_\Theta(y_t|X_t)}{\partial y_t}\right)^2\right]-\left(\mathbb{E}_t\left[\frac{\partial \log g_\Theta(y_t|X_t)}{\partial y_t}\right]\right)^2\right)$$

  where the expectations $\mathbb{E}_t$ are now with respect to the joint posterior distributions $(\Theta, X_t) \sim p(d\theta|y_{1:t})p(dx_t|y_{1:t},\theta)$

Estimation using SMC^2 (*Chopin, Jacob & Papaspiliopoulos, 2013*)

H-scores can be consistently estimated for state-space models, as long as one can simulate transitions from f_θ and evaluate g_θ

- **SMC and SMC^2 can be used similarly in the case of discrete observations.**

Asymptotic guarantees

Consistency of the H-score for model selection

- Assume the true data generating process is $Y_1, \ldots, Y_T \overset{\text{i.i.d.}}{\sim} p_\star$
- Consider two non-nested i.i.d. models M_1 and M_2 with respective posterior distributions concentrating around $\theta_1^\star$ and $\theta_2^\star$. Let $p_{\theta_j^\star} = p_j(dy|\theta_j^\star)$
- Under regularity conditions, the H-scores $\mathcal{H}_T$ of M_1 and M_2 satisfy

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

where $\Delta(p_\star, M_j) = \mathbb{E}_\star[\mathcal{H}(Y, p_{\theta_j^\star})] - \mathbb{E}_\star[\mathcal{H}(Y, p_\star)]$
with $\Delta(p_\star, M_j) \geq 0$ and $\Delta(p_\star, M_j) = 0$ if and only if $p_{\theta_j^\star} = p_\star$

- By analogy, the log-Bayes factor of M_1 against M_2 satisfies

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

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

- Similar results hold for dependent data and state-space models.

Consistency of the H-score for model selection

- Assume the true data generating process is $Y_1, ..., Y_T \overset{\text{i.i.d.}}{\sim} p_\star$
- Consider two non-nested i.i.d. models M_1 and M_2 with respective posterior distributions concentrating around $\theta_1^\star$ and $\theta_2^\star$. Let $p_{\theta_j^\star} = p_j(dy|\theta_j^\star)$
- Under regularity conditions, the H-scores $\mathcal{H}_T$ of M_1 and M_2 satisfy

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

where $\Delta(p_\star, M_j) = \mathbb{E}_\star[\mathcal{H}(Y, p_{\theta_j^\star})] - \mathbb{E}_\star[\mathcal{H}(Y, p_\star)]$
with $\Delta(p_\star, M_j) \geq 0$ and $\Delta(p_\star, M_j) = 0$ if and only if $p_{\theta_j^\star} = p_\star$

- By analogy, the log-Bayes factor of M_1 against M_2 satisfies

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

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

- Similar results hold for dependent data and state-space models.

Consistency of the H-score for model selection

- Assume the true data generating process is $Y_1, \ldots, Y_T \overset{\text{i.i.d.}}{\sim} p_\star$
- Consider two non-nested i.i.d. models M_1 and M_2 with respective posterior distributions concentrating around $\theta_1^\star$ and $\theta_2^\star$. Let $p_{\theta_j^\star} = p_j(dy|\theta_j^\star)$
- **Under regularity conditions, the H-scores $\mathcal{H}_T$ of M_1 and M_2 satisfy**

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

where $\Delta(p_\star, M_j) = \mathbb{E}_\star[\mathcal{H}(Y, p_{\theta_j^\star})] - \mathbb{E}_\star[\mathcal{H}(Y, p_\star)]$
with $\Delta(p_\star, M_j) \geq 0$ and $\Delta(p_\star, M_j) = 0$ if and only if $p_{\theta_j^\star} = p_\star$

- By analogy, the log-Bayes factor of M_1 against M_2 satisfies

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

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

- Similar results hold for dependent data and state-space models.

Consistency of the H-score for model selection

- Assume the true data generating process is $Y_1, ..., Y_T \overset{\text{i.i.d.}}{\sim} p_\star$
- Consider two non-nested i.i.d. models M_1 and M_2 with respective posterior distributions concentrating around $\theta_1^\star$ and $\theta_2^\star$. Let $p_{\theta_j^\star} = p_j(dy|\theta_j^\star)$
- Under regularity conditions, the H-scores $\mathcal{H}_T$ of M_1 and M_2 satisfy

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

where $\Delta(p_\star, M_j) = \mathbb{E}_\star[\mathcal{H}(Y, p_{\theta_j^\star})] - \mathbb{E}_\star[\mathcal{H}(Y, p_\star)]$
with $\Delta(p_\star, M_j) \geq 0$ and $\Delta(p_\star, M_j) = 0$ if and only if $p_{\theta_j^\star} = p_\star$

- **By analogy, the log-Bayes factor of M_1 against M_2 satisfies**

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

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

- Similar results hold for dependent data and state-space models.

Consistency of the H-score for model selection

- Assume the true data generating process is $Y_1, ..., Y_T \overset{\text{i.i.d.}}{\sim} p_\star$
- Consider two non-nested i.i.d. models M_1 and M_2 with respective posterior distributions concentrating around $\theta_1^\star$ and $\theta_2^\star$. Let $p_{\theta_j^\star} = p_j(dy|\theta_j^\star)$
- Under regularity conditions, the H-scores $\mathcal{H}_T$ of M_1 and M_2 satisfy

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

where $\Delta(p_\star, M_j) = \mathbb{E}_\star[\mathcal{H}(Y, p_{\theta_j^\star})] - \mathbb{E}_\star[\mathcal{H}(Y, p_\star)]$
with $\Delta(p_\star, M_j) \geq 0$ and $\Delta(p_\star, M_j) = 0$ if and only if $p_{\theta_j^\star} = p_\star$

- By analogy, the log-Bayes factor of M_1 against M_2 satisfies

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

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

- Similar results hold for dependent data and state-space models.

Illustration of consistency for i.i.d. models

- **Example 1**. Given simulated $y_1, \ldots, y_{1000} \sim \mathcal{N}(\mu_\star, \sigma_\star^2)$, we compare

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

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

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

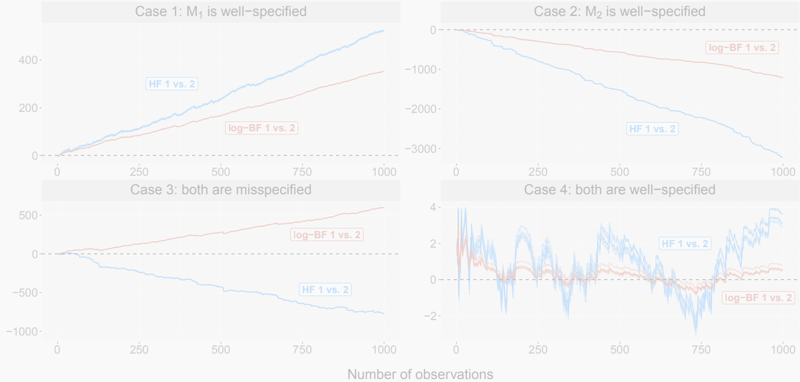


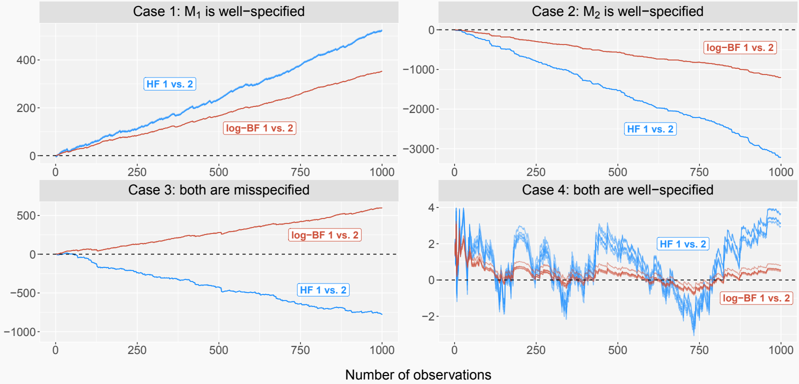
Illustration of consistency for i.i.d. models

- **Example 1**. Given simulated $y_1, \ldots, y_{1000} \sim \mathcal{N}(\mu_\star, \sigma_\star^2)$, we compare

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

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

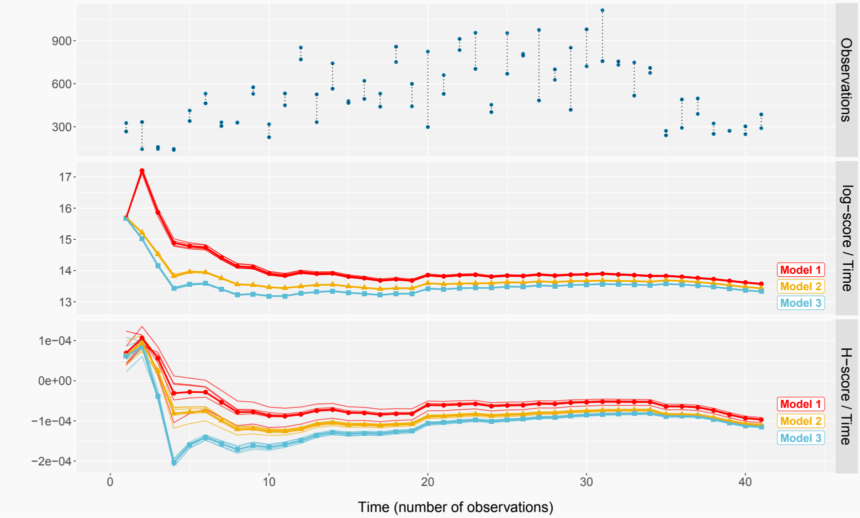
in the following four cases $(\mu_\star, \sigma_\star^2) = (1, 1), (0, 5), (4, 3), (0, 1)$.



Applications and discussion

Jumping back to kangaroos

- **Example 2.** Population dynamics of red kangaroos (*Knape & de Valpine, 2012*)



Summary

Advantages of using the H-score

- Robust to prior vagueness and allows for improper priors
- Justified non-asymptotically and generally consistent asymptotically
- Can be estimated sequentially for a wide class of parametric models by using SMC methods

Limitations and avenues for future research

- Requires additional smoothness conditions on the densities
- Tends to require a larger number of particles than the log-evidence for accurate estimation via SMC methods
- Extension to nonparametric models ?

- More details in the manuscript: arxiv.org/pdf/1711.00136
- 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, S. Lauritzen, and M. Parry.
Proper local scoring rules on discrete sample spaces.
The Annals of Statistics, 40 (1):593–608, 2012.

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

References (2/2)

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.

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.

F. B. Lempers.
Posterior Probabilities of Alternative Linear Models.
Rotterdam University Press, 1971.

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.

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

- **Example 3.** 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:

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

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:

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

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

Illustration of consistency for state-space models

- **Example 3**. Lévy-driven stochastic volatility models, given $T = 1000$ observations simulated from a single-factor model with $\lambda = 0.01$, $\xi = 0.5$, $\omega^2 = 0.0625$, $\mu = 0$, and $\beta = 0$ (*Barndorff-Nielsen & Shephard, 2001*)

Model 1 (single-factor)

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:

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

Model 2 (multi-factor)

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

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

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

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

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

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

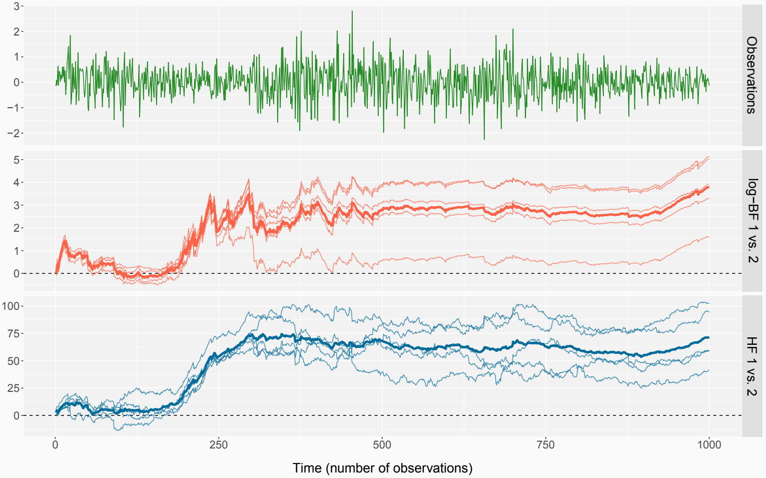
with independent priors:

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

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

Illustration of consistency for state-space models

- **Example 3**. Lévy-driven stochastic volatility models, given $T = 1000$ observations simulated from a single-factor model with $\lambda = 0.01$, $\xi = 0.5$, $\omega^2 = 0.0625$, $\mu = 0$, and $\beta = 0$ (*Barndorff-Nielsen & Shephard, 2001*)



Prequential vs. Batch approach

- Notice that, unlike for the log scoring rule, here we have:

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

- Batch version:
  - Easier to compute, as it only requires to estimate final evidence $p(y_{1:T})$
  - But typically inconsistent for model selection
- Prequential version:
  - Generally consistent for model selection
  - Requires to estimate all the intermediary predictive densities $p(dy_t|y_{1:t-1})$, but this can be achieved using algorithms such as SMC or SMC^2

Partial Bayes factors (Lempers, 1971)

- Split the data $y_{1:T}$ into a training set $y_{1:m}$ and another set $y_{m+1:T}$ for some choice of m
- Idea: condition on the training set to make the prior proper (or less vague) then compute the Bayes factor on the remaining data
- Essentially we replace the prior $p(\theta|M)$ by the posterior given the training set $p(\theta|y_{1:m}, M)$, and compute the usual Bayes factor on the remaining data set $y_{m+1:T}$
- The partial Bayes factor between Models M_1 and M_2 is defined as:

$$\frac{p(y_{m+1:T}|y_{1:m}, M_1)}{p(y_{m+1:T}|y_{1:m}, M_2)}$$

- Drawback: choice of m is a bit ad-hoc, undesirable to "waste" data for the training set especially in settings where the number of observations is small (e.g. Example 2 where $T = 41$)

Fractional Bayes factors (O'Hagan, 1995)

- In the setting of partial Bayes factors, if m and T are both large, the likelihood $p(y_{1:m}|\theta, M)$ of the training set will approximate (at least in the i.i.d. case) the full likelihood raised to a power $b \equiv m/T$
- For a given model M we define:

$$q_b(y_{1:T}|M) := \frac{\int p(\theta|M)p(y_{1:T}|\theta, M)d\theta}{\int p(\theta|M)p(y_{1:T}|\theta, M)^b d\theta}$$

  which approximates $p(y_{m+1:T}|y_{1:m}, M)$ for large m and T
- The fractional Bayes factor between Models M_1 and M_2 is defined as:

$$\frac{q_b(y_{1:T}|M_1)}{q_b(y_{1:T}|M_2)}$$

- Drawback: choice of b is a bit ad-hoc, not very principled for small sample size since the main justification relies on asymptotics