Model selection for state-space models

Stephane Shao*
October 4, 2016

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

*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

What is a state-space model ?

What is a state-space model ?

- Also known as **Hidden Markov model**
- Class of **time series models** used in various fields like econometrics, bioinformatics, signal processing, target tracking, epidemiology ...

- Unobserved Markov chain of latent states $X_{1:T} \equiv (X_1, \ldots, X_T)$ 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 $Y_{1:T} \equiv (Y_1, \ldots, Y_T)$ conditionally independent given $X_{1:T}$ with

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

- Prior distribution $p(\theta)$ on the parameter

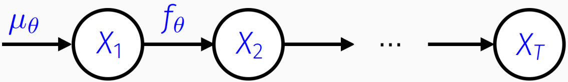
What is a state-space model ?

- Also known as **Hidden Markov model**
- Class of **time series models** used in various fields like econometrics, bioinformatics, signal processing, target tracking, epidemiology ...

- Unobserved Markov chain of latent states $X_{1:T} \equiv (X_1, ..., X_T)$ 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 $Y_{1:T} \equiv (Y_1, ..., Y_T)$ conditionally independent given $X_{1:T}$ with
$$Y_t \mid X_t \sim g_\theta(\cdot|X_t) \quad \text{for } t \geq 1$$
- Prior distribution $p(\theta)$ on the parameter

What is a state-space model ?

- Also known as **Hidden Markov model**
- Class of **time series models** used in various fields like econometrics, bioinformatics, signal processing, target tracking, epidemiology ...



- Unobserved Markov chain of latent states $X_{1:T} \equiv (X_1, ..., X_T)$ with

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

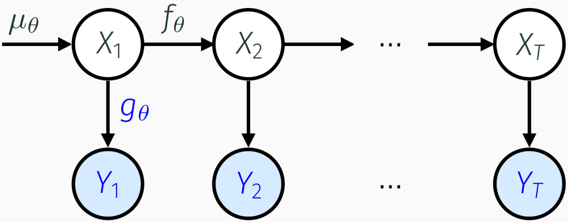
- Observations $Y_{1:T} \equiv (Y_1, ..., Y_T)$ conditionally independent given $X_{1:T}$ with

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

- Prior distribution $p(\theta)$ on the parameter

What is a state-space model ?

- Also known as **Hidden Markov model**
- Class of **time series models** used in various fields like econometrics, bioinformatics, signal processing, target tracking, epidemiology ...



- Unobserved Markov chain of latent states $X_{1:T} \equiv (X_1, \dots, X_T)$ with

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

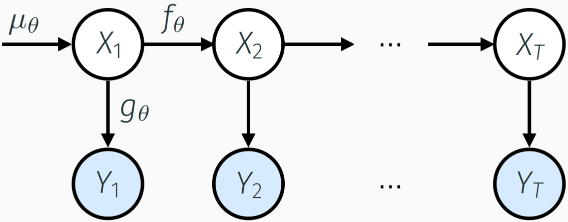
- Observations $Y_{1:T} \equiv (Y_1, \dots, Y_T)$ conditionally independent given $X_{1:T}$ with

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

- Prior distribution $p(\theta)$ on the parameter

What is a state-space model ?

- Also known as **Hidden Markov model**
- Class of **time series models** used in various fields like econometrics, bioinformatics, signal processing, target tracking, epidemiology ...



- Unobserved Markov chain of latent states $X_{1:T} \equiv (X_1, \dots, X_T)$ with

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

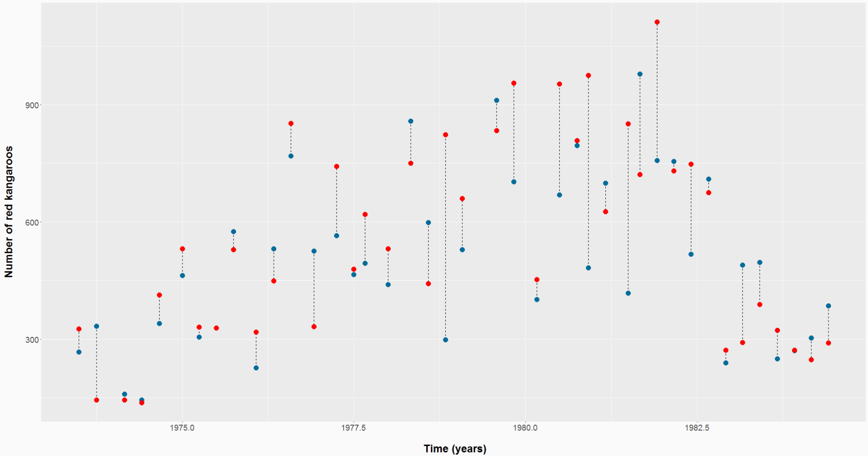
- Observations $Y_{1:T} \equiv (Y_1, \dots, Y_T)$ conditionally independent given $X_{1:T}$ with

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

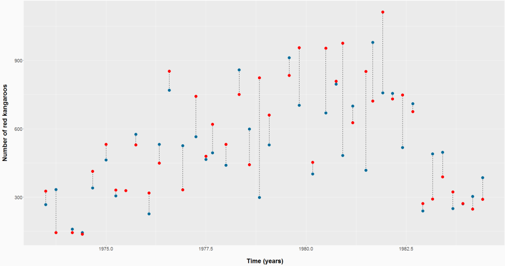
- Prior distribution $p(\theta)$ on the parameter

What does the data look like ?

- Counts of red kangaroos performed twice on 41 sampling occasions (Knape and Valpine, 2012)



What do the models look like ?



Model 1

$$X_1 \sim \mathrm{LN}(0,5)$$
$$\frac{dX_t}{X_t} = (r + \frac{\sigma^2}{2} - bX_t)\, dt + \sigma dW_t$$

$$Y_{1,t} \mid X_t \sim \mathrm{NegBin}(X_t, X_t + \tau X_t^2)$$
$$Y_{2,t} \mid X_t \sim \mathrm{NegBin}(X_t, X_t + \tau X_t^2)$$

$$b, \sigma, \tau \sim \mathrm{Unif}(0,10)$$
$$r \sim \mathrm{Unif}(-10,10)$$

Model 2

$$X_1 \sim \mathrm{LN}(0,5)$$
$$\frac{dX_t}{X_t} = (r + \frac{\sigma^2}{2})\, dt + \sigma dW_t$$

$$Y_{1,t} \mid X_t \sim \mathrm{NegBin}(X_t, X_t + \tau X_t^2)$$
$$Y_{2,t} \mid X_t \sim \mathrm{NegBin}(X_t, X_t + \tau X_t^2)$$

$$\sigma, \tau \sim \mathrm{Unif}(0,10)$$
$$r \sim \mathrm{Unif}(-10,10)$$

Model 3

$$X_1 \sim \mathrm{LN}(0,5)$$
$$\frac{dX_t}{X_t} = (\frac{\sigma^2}{2})\, dt + \sigma dW_t$$

$$Y_{1,t} \mid X_t \sim \mathrm{NegBin}(X_t, X_t + \tau X_t^2)$$
$$Y_{2,t} \mid X_t \sim \mathrm{NegBin}(X_t, X_t + \tau X_t^2)$$

$$\sigma, \tau \sim \mathrm{Unif}(0,10)$$

What is so challenging about state-space models ?

The likelihood is unavailable in closed form

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

which is typically an intractable high-dimensional integral ...

- We are interested in quantities of the form:

$$\mathbb{E}\left[\varphi(\Theta, X_t) \,\middle|\, y_{1:t}\right]$$

where $\mathbb{E}$ is with respect to the joint posterior distribution $p(\theta, x_{1:t}|y_{1:t})$

- The **SMC^2 algorithm** produces consistent estimators of such expectations assuming we known g_θ numerically and can simulate from f_θ (Chopin, Jacob, and Papaspiliopoulos, 2013)

What is so challenging about state-space models ?

The likelihood is unavailable in closed form

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

which is typically an intractable high-dimensional integral ...

- We are interested in quantities of the form:

$$\mathbb{E}\left[\varphi(\Theta, X_t) \,\middle|\, y_{1:t}\right]$$

where $\mathbb{E}$ is with respect to the joint posterior distribution $p(\theta, x_{1:t}|y_{1:t})$

- The **SMC^2 algorithm** produces consistent estimators of such expectations assuming we known g_θ numerically and can simulate from f_θ (Chopin, Jacob, and Papaspiliopoulos, 2013)

What is so challenging about state-space models ?

The likelihood is unavailable in closed form

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

which is typically an intractable high-dimensional integral ...

- We are interested in quantities of the form:

$$\mathbb{E}\Big[\varphi(\Theta, X_t) \,\Big|\, y_{1:t}\Big]$$

where $\mathbb{E}$ is with respect to the joint posterior distribution $p(\theta, x_{1:t}|y_{1:t})$

- The **SMC^2 algorithm** produces consistent estimators of such expectations assuming we known g_θ numerically and can simulate from f_θ (Chopin, Jacob, and Papaspiliopoulos, 2013)

Why not use Bayes factors ?

What can go wrong with Bayes factors ?

Bayes factors $\equiv$ choose the model M with the largest *evidence*

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

where $p(\theta|M)$ denotes the prior distribution of θ under model M

Sensitivity to the choice of prior

- Bayes factors do not allow for improper priors
- The evidence for any given model can be made arbitrarily small by making the prior distribution arbitrarily vague

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

What can go wrong with Bayes factors ?

Bayes factors $\equiv$ choose the model M with the largest *evidence*

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

where $p(\theta|M)$ denotes the prior distribution of θ under model M

Sensitivity to the choice of prior

- Bayes factors do not allow for improper priors
- The evidence for any given model can be made arbitrarily small by making the prior distribution arbitrarily vague

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

What can go wrong with Bayes factors ?

Bayes factors $\equiv$ choose the model M with the largest *evidence*

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

where $p(\theta|M)$ denotes the prior distribution of θ under model M

Sensitivity to the choice of prior

- Bayes factors do not allow for improper priors
- The evidence for any given model can be made arbitrarily small by making the prior distribution arbitrarily vague

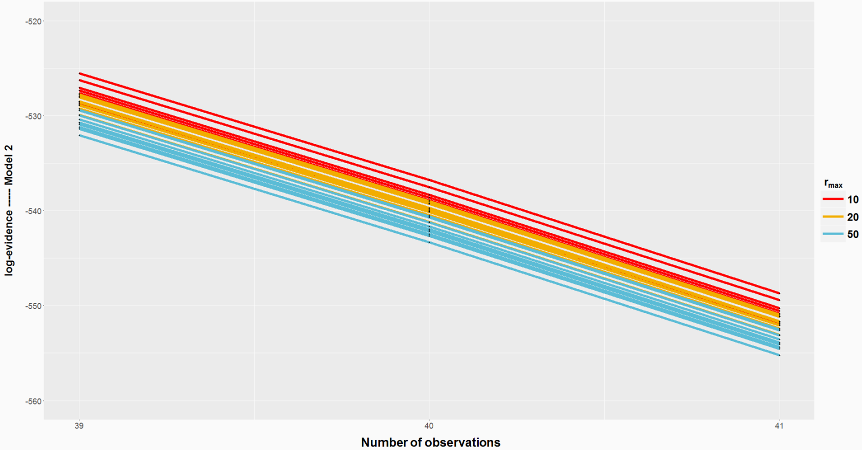
Yet, vague or improper priors often stem from reasonable approaches (genuine non-informativeness, Jeffreys prior, …)

Sensitivity of the Bayes Factor to vague priors

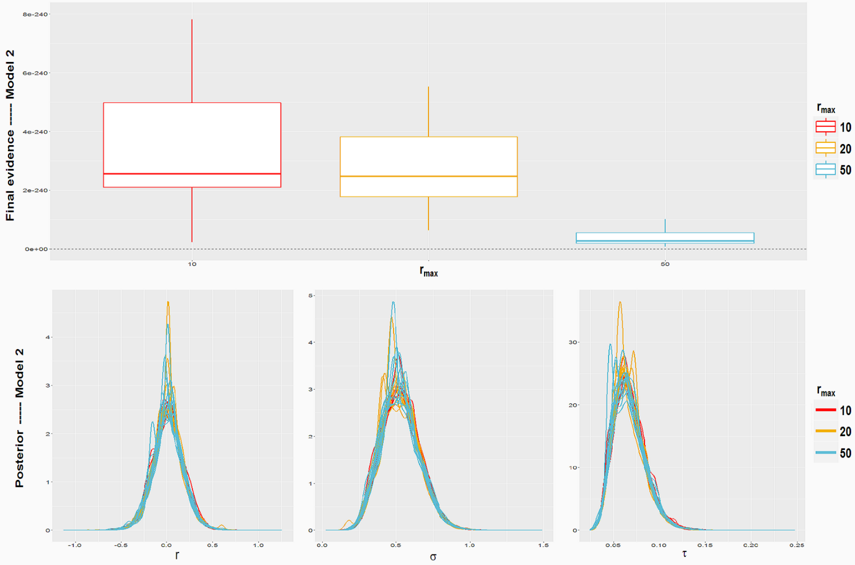
« [...] as pointed out by others, posterior model probabilities and Bayes factors can be sensitive to the priors on the parameters. This was the case for the logistic model M1. ***Under the alternative uniform priors over the interval (-100, 100) for r and (0, 100) for the other parameters the marginal density was a factor 10^3 times smaller than under the original prior****. »* (Knape and Valpine, 2012)

Sensitivity of the Bayes Factor to vague priors

« […] as pointed out by others, posterior model probabilities and Bayes factors can be sensitive to the priors on the parameters. This was the case for the logistic model M1. ***Under the alternative uniform priors over the interval (-100, 100) for r and (0, 100) for the other parameters the marginal density was a factor 10^3 times smaller than under the original prior.*** » (Knape and Valpine, 2012)



Sensitivity of the Bayes Factor to vague priors



A new criterion for model selection

Bayes factors as a decision rule

- Bayes factors choose the model M that maximizes the evidence :

$$p(y_{1:T}|M) = p(y_1|M)\, p(y_2|y_1, M)\, p(y_3|y_{1:2}, M) \ldots p(y_T|y_{1:T-1}, M)$$

- Hence it chooses the model minimizing $-\log(p(y_{1:T}|M))$ or equivalently :

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

- This is a particular case of a more general decision rule that chooses the model M minimizing the **prequential score** :

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

for a specific choice of **scoring rule** $\mathcal{S} : (\tilde{y}, q(dy)) \longmapsto -\log(q(\tilde{y}))$

- The scoring rule is a **loss function** that quantifies the performance of the model in terms of probabilistic predictions at each step

Bayes factors as a decision rule

- Bayes factors choose the model M that maximizes the evidence :

$$p(y_{1:T}|M) = p(y_1|M)\, p(y_2|y_1, M)\, p(y_3|y_{1:2}, M) \dots p(y_T|y_{1:T-1}, M)$$

- Hence it chooses the model minimizing $-\log(p(y_{1:T}|M))$ or equivalently :

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

- This is a particular case of a more general decision rule that chooses the model M minimizing the **prequential score** :

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

for a specific choice of **scoring rule** $\mathcal{S} : (\tilde{y}, q(dy)) \longmapsto -\log(q(\tilde{y}))$

- The scoring rule is a **loss function** that quantifies the performance of the model in terms of probabilistic predictions at each step

Bayes factors as a decision rule

- Bayes factors choose the model M that maximizes the evidence :

$$p(y_{1:T}|M) = p(y_1|M)\, p(y_2|y_1, M)\, p(y_3|y_{1:2}, M) \ldots p(y_T|y_{1:T-1}, M)$$

- Hence it chooses the model minimizing $-\log(p(y_{1:T}|M))$ or equivalently :

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

- This is a particular case of a more general decision rule that chooses the model M minimizing the **prequential score** :

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

  for a specific choice of **scoring rule** $\mathcal{S} : (\tilde{y},\ q(dy)) \longmapsto -\log(q(\tilde{y}))$

- The scoring rule is a **loss function** that quantifies the performance of the model in terms of probabilistic predictions at each step

Bayes factors as a decision rule

- Bayes factors choose the model M that maximizes the evidence :

$$p(y_{1:T}|M) = p(y_1|M)\, p(y_2|y_1, M)\, p(y_3|y_{1:2}, M) \dots p(y_T|y_{1:T-1}, M)$$

- Hence it chooses the model minimizing $-\log(p(y_{1:T}|M))$ or equivalently :

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

- This is a particular case of a more general decision rule that chooses the model M minimizing the **prequential score** :

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

  for a specific choice of **scoring rule** $\mathcal{S} : (\tilde{y},\ q(dy)) \longmapsto -\log(q(\tilde{y}))$

- The scoring rule is a **loss function** that quantifies the performance of the model in terms of probabilistic predictions at each step

What makes a good scoring rule ?

- Key idea: replace the log scoring rule by a different scoring rule (Dawid and Musio, 2015)

Propriety

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **proper** (resp. *strictly*) if the function $q \longmapsto \mathbb{E}_{Y \sim p^*}[\mathcal{S}(Y, q)]$ is minimized (resp. *uniquely*) by $q = p^*$

Localility of order m

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **m-local** if $\mathcal{S}(\tilde{y}, q)$ is only a function of $\tilde{y}$ and the first m derivatives of q all evaluated at $\tilde{y}$

Homogeneity of order h

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **h-homogeneous** if it satisfies $\mathcal{S}(\tilde{y}, \lambda q) = \lambda^h \mathcal{S}(\tilde{y}, q)$ for every $\tilde{y}$ and q, and every $\lambda > 0$

- 0-Homogeneity implies invariance to arbitrary scaling of the prior
- The log scoring rule is strictly proper and 0-local but not homogeneous

What makes a good scoring rule ?

- Key idea: replace the log scoring rule by a different scoring rule (Dawid and Musio, 2015)

Propriety

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **proper** (resp. *strictly*) if the function $q \longmapsto \mathbb{E}_{Y \sim p^*}[\mathcal{S}(Y, q)]$ is minimized (resp. *uniquely*) by $q = p^*$

Localility of order m

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **m-local** if $\mathcal{S}(\tilde{y}, q)$ is only a function of $\tilde{y}$ and the first m derivatives of q all evaluated at $\tilde{y}$

Homogeneity of order h

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **h-homogeneous** if it satisfies $\mathcal{S}(\tilde{y}, \lambda q) = \lambda^h \mathcal{S}(\tilde{y}, q)$ for every $\tilde{y}$ and q, and every $\lambda > 0$

- 0-Homogeneity implies invariance to arbitrary scaling of the prior
- The log scoring rule is strictly proper and 0-local but not homogeneous

What makes a good scoring rule ?

- Key idea: replace the log scoring rule by a different scoring rule (Dawid and Musio, 2015)

Propriety

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **proper** (resp. *strictly*) if the function $q \longmapsto \mathbb{E}_{Y \sim p^*}[\mathcal{S}(Y, q)]$ is minimized (resp. *uniquely*) by $q = p^*$

Localility of order m

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **m-local** if $\mathcal{S}(\tilde{y}, q)$ is only a function of $\tilde{y}$ and the first m derivatives of q all evaluated at $\tilde{y}$

Homogeneity of order h

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **h-homogeneous** if it satisfies $\mathcal{S}(\tilde{y}, \lambda q) = \lambda^h \mathcal{S}(\tilde{y}, q)$ for every $\tilde{y}$ and q, and every $\lambda > 0$

- 0-Homogeneity implies invariance to arbitrary scaling of the prior
- The log scoring rule is strictly proper and 0-local but not homogeneous

What makes a good scoring rule ?

- Key idea: replace the log scoring rule by a different scoring rule (Dawid and Musio, 2015)

Propriety

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **proper** (resp. *strictly*) if the function $q \longmapsto \mathbb{E}_{Y\sim p^*}\left[\mathcal{S}(Y, q)\right]$ is minimized (resp. *uniquely*) by $q = p^*$

Localility of order m

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **m-local** if $\mathcal{S}(\tilde{y}, q)$ is only a function of $\tilde{y}$ and the first m derivatives of q all evaluated at $\tilde{y}$

Homogeneity of order h

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **h-homogeneous** if it satisfies $\mathcal{S}(\tilde{y}, \lambda q) = \lambda^h \mathcal{S}(\tilde{y}, q)$ for every $\tilde{y}$ and q, and every $\lambda > 0$

- 0-Homogeneity implies invariance to arbitrary scaling of the prior
- The log scoring rule is strictly proper and 0-local but not homogeneous

What makes a good scoring rule ?

- Key idea: replace the log scoring rule by a different scoring rule (Dawid and Musio, 2015)

Propriety

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **proper** (resp. *strictly*) if the function $q \longmapsto \mathbb{E}_{Y \sim p^*}[\mathcal{S}(Y, q)]$ is minimized (resp. *uniquely*) by $q = p^*$

Localility of order m

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **m-local** if $\mathcal{S}(\tilde{y}, q)$ is only a function of $\tilde{y}$ and the first m derivatives of q all evaluated at $\tilde{y}$

Homogeneity of order h

A scoring rule $\mathcal{S}(\tilde{y}, q)$ is said to be **h-homogeneous** if it satisfies $\mathcal{S}(\tilde{y}, \lambda q) = \lambda^h \mathcal{S}(\tilde{y}, q)$ for every $\tilde{y}$ and q, and every $\lambda > 0$

- 0-Homogeneity implies invariance to arbitrary scaling of the prior
- The log scoring rule is strictly proper and 0-local but not homogeneous

Hyvärinen score

- Parry et al. (2012) characterized all the 0-homogeneous strictly proper m-local scoring rules: they only exist when m is a positive even integer
- Thus the "simplest" such scoring rule is the **Hyvärinen score**

$$\mathcal{S}_{\mathcal{H}}(\tilde{y}, q) := 2\,\frac{d^2 \log q(\tilde{y})}{dy^2} + \left(\frac{d \log q(\tilde{y})}{dy}\right)^2$$

- It can be extended to discrete observations as follows:

$$\mathcal{S}_{\mathcal{H}}(\tilde{y}, q) := 2\left(\frac{q(\tilde{y}+1)-q(\tilde{y})}{q(\tilde{y})} - \frac{q(\tilde{y})-q(\tilde{y}-1)}{q(\tilde{y}-1)}\right) + \left(\frac{q(\tilde{y}+1)-q(\tilde{y})}{q(\tilde{y})}\right)^2$$

New model selection criterion

Choose the model M that minimizes the **prequential Hyvärinen score**

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

Hyvärinen score

- Parry et al. (2012) characterized all the 0-homogeneous strictly proper m-local scoring rules: they only exist when m is a positive even integer
- Thus the "simplest" such scoring rule is the Hyvärinen score

$$\mathcal{S}_{\mathcal{H}}(\tilde{y}, q) := 2\,\frac{d^2 \log q(\tilde{y})}{dy^2} + \left(\frac{d \log q(\tilde{y})}{dy}\right)^2$$

- It can be extended to discrete observations as follows:

$$\mathcal{S}_{\mathcal{H}}(\tilde{y}, q) := 2\left(\frac{q(\tilde{y}+1)-q(\tilde{y})}{q(\tilde{y})} - \frac{q(\tilde{y})-q(\tilde{y}-1)}{q(\tilde{y}-1)}\right) + \left(\frac{q(\tilde{y}+1)-q(\tilde{y})}{q(\tilde{y})}\right)^2$$

New model selection criterion

Choose the model M that minimizes the **prequential Hyvärinen score**

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

Hyvärinen score

- Parry et al. (2012) characterized all the 0-homogeneous strictly proper m-local scoring rules: they only exist when m is a positive even integer
- Thus the "simplest" such scoring rule is the Hyvärinen score

$$\mathcal{S}_{\mathcal{H}}(\tilde{y}, q) := 2\,\frac{d^2 \log q(\tilde{y})}{dy^2} + \left(\frac{d \log q(\tilde{y})}{dy}\right)^2$$

- It can be extended to discrete observations as follows:

$$\mathcal{S}_{\mathcal{H}}(\tilde{y}, q) := 2\left(\frac{q(\tilde{y}+1)-q(\tilde{y})}{q(\tilde{y})} - \frac{q(\tilde{y})-q(\tilde{y}-1)}{q(\tilde{y}-1)}\right) + \left(\frac{q(\tilde{y}+1)-q(\tilde{y})}{q(\tilde{y})}\right)^2$$

New model selection criterion

Choose the model M that minimizes the **prequential Hyvärinen score**

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

Hyvärinen score

- Parry et al. (2012) characterized all the 0-homogeneous strictly proper m-local scoring rules: they only exist when m is a positive even integer
- Thus the "simplest" such scoring rule is the Hyvärinen score

$$\mathcal{S}_{\mathcal{H}}(\tilde{y}, q) := 2\,\frac{d^2 \log q(\tilde{y})}{dy^2} + \left(\frac{d \log q(\tilde{y})}{dy}\right)^2$$

- It can be extended to discrete observations as follows:

$$\mathcal{S}_{\mathcal{H}}(\tilde{y}, q) := 2\left(\frac{q(\tilde{y}+1)-q(\tilde{y})}{q(\tilde{y})} - \frac{q(\tilde{y})-q(\tilde{y}-1)}{q(\tilde{y}-1)}\right) + \left(\frac{q(\tilde{y}+1)-q(\tilde{y})}{q(\tilde{y})}\right)^2$$

New model selection criterion

Choose the model M that minimizes the **prequential Hyvärinen score**

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

Why does it work ?

Theoretical justifications and guarantees

- Principled approach that is justified for any finite sample size by the framework of Decision Theory (Bernardo and Smith, 2000)

- Consistency: when comparing the true model with any other misspecified model, we end up choosing† the true model as $T \longrightarrow +\infty$

† $\mathbb{P}^*$-almost surely, where $\mathbb{P}^*$ denotes the true data generating distribution of $(Y_t)_{t \geq 1}$

Theoretical justifications and guarantees

- Principled approach that is justified for any finite sample size by the framework of Decision Theory (Bernardo and Smith, 2000)
- Consistency: when comparing the true model with any other misspecified model, we end up choosing[†] the true model as $T \longrightarrow +\infty$

[†] $\mathbb{P}^*$-almost surely, where $\mathbb{P}^*$ denotes the true data generating distribution of $(Y_t)_{t\geq 1}$

How to implement it ?

Hyvärinen score as an expectation of known quantities

- Let's fix some arbitrary model (and drop the conditioning on M)
- The prequential Hyvärinen score turns out* to be exactly equal to:

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

where $\mathbb{E}_t$ denotes the expectation with respect to $(\Theta, X_t) \sim p(\theta, x_t|y_{1:t})$

- This only involves expectations with respect to the successive posterior distributions $p(\theta, x_t|y_{1:t})$ of known quantities

  $\longrightarrow$ We can use SMC^2 to estimate it consistently

- Similar approach holds for discrete observations

*After some non-trivial derivation.

Hyvärinen score as an expectation of known quantities

- Let's fix some arbitrary model (and drop the conditioning on M)
- The prequential Hyvärinen score turns out* to be exactly equal to:

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

where $\mathbb{E}_t$ denotes the expectation with respect to $(\Theta, X_t) \sim p(\theta, x_t | y_{1:t})$

- This only involves expectations with respect to the successive posterior distributions $p(\theta, x_t|y_{1:t})$ of known quantities

$\longrightarrow$ **We can use SMC^2 to estimate it consistently**

- Similar approach holds for discrete observations

*After some non-trivial derivation.

Hyvärinen score as an expectation of known quantities

- Let's fix some arbitrary model (and drop the conditioning on M)
- The prequential Hyvärinen score turns out* to be exactly equal to:

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

where $\mathbb{E}_t$ denotes the expectation with respect to $(\Theta, X_t) \sim p(\theta, x_t|y_{1:t})$

- This only involves expectations with respect to the successive posterior distributions $p(\theta, x_t|y_{1:t})$ of known quantities

$\longrightarrow$ **We can use SMC^2 to estimate it consistently**

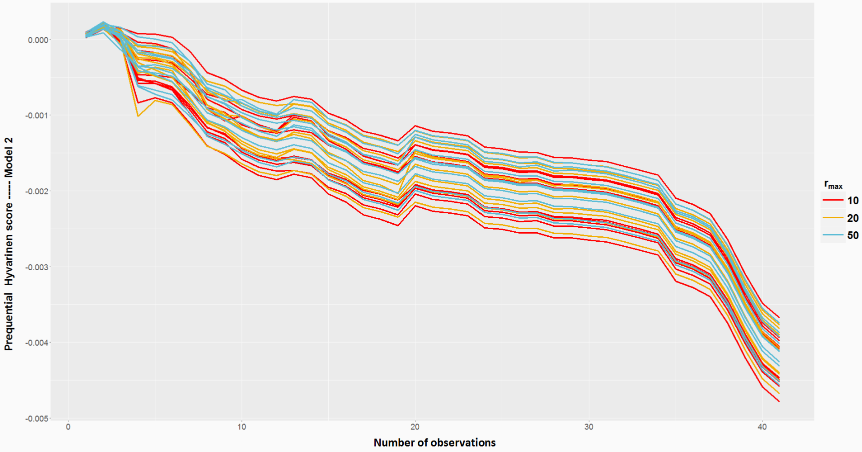
- Similar approach holds for discrete observations

*After some non-trivial derivation.

Applications and discussion

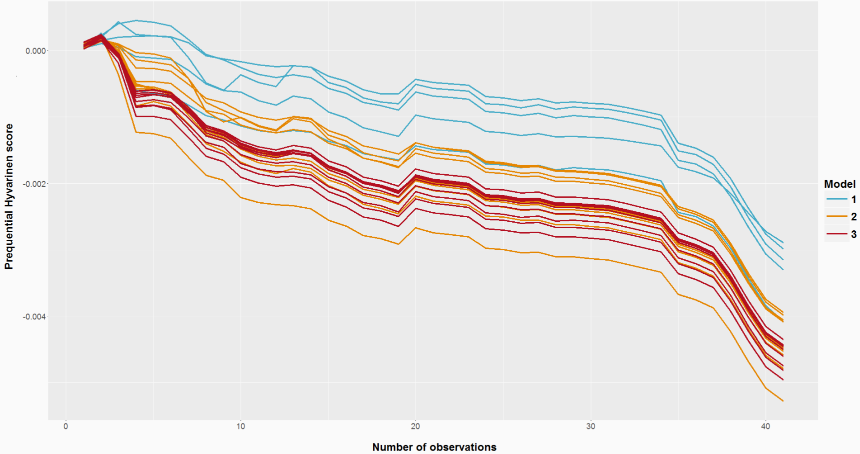
Jumping back to kangaroos

- The prequential Hyvärinen score is insensitive to arbitrary vagueness of the prior distribution (as expected)



Comparing all three models

- Lower = Better
- Need more particles (or more data) to choose between models 2 and 3



Summary

Advantages of prequential Hyvärinen score

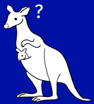
- Allows for improper priors
- Not sensitive to arbitrary vagueness of priors
- Can be estimated consistently in a sequential fashion via SMC^2 by only knowing g_θ numerically and being able to simulate from f_θ

Possible limitations

- Computational cost induced by SMC^2

- Further work : applications to stochastic volatility models, neuroscience data, epidemic models, and more ...
- R package on its way, for everyone to use

Questions ?



References (1/2)

J. M. Bernardo and A. F. M. Smith.
Bayesian Theory.
John Wiley & Sons, 2000.

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)

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.

Multivariate observations (continuous case)

- Let $y = (y_{(1)}, ..., y_{(d_y)})^\top \in \mathbb{R}^{d_y}$
- Then the Hyvärinen score is defined as:

$$S_{\mathcal{H}}(\tilde{y}, q) := 2\Delta_y \log q(\tilde{y}) + \|\nabla_y \log q(\tilde{y})\|^2$$

- Which is exactly equal to:

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

Multivariate observations (discrete case)

- Let $\tilde{y} \equiv (\tilde{y}_{(1)}, \ldots, \tilde{y}_{(d_y)})^\top$ taking finite values in $\mathbb{Y} := [\![a_1, b_1]\!] \times \ldots \times [\![a_{d_y}, b_{d_y}]\!]$ where $a_k, b_k \in \mathbb{Z} \cup \{-\infty, +\infty\}$ with $a_k < b_k$ for each k.
- Let $e^{(k)} \in \mathbb{Z}^{d_y}$ such that $e^{(k)}_{(j)} = \delta_{jk}$
- Then the discrete Hyvärinen score can be defined as:

$$S_{\mathcal{H}}(\tilde{y}, q) := \sum_{k=1}^{d_y} S_{\mathcal{B}_k}(\tilde{y}, q)$$

where:

$$S_{\mathcal{B}_k}(\tilde{y}, q) := \begin{cases} -2\left(\frac{q(\tilde{y}) - q(\tilde{y} - e^{(k)})}{q(\tilde{y} - e^{(k)})}\right) & \text{if } \tilde{y}_{(k)} = b_k \\ 2\left(\frac{q(\tilde{y} + e^{(k)}) - q(\tilde{y})}{q(\tilde{y})} - \frac{q(\tilde{y}) - q(\tilde{y} - e^{(k)})}{q(\tilde{y} - e^{(k)})}\right) + \left(\frac{q(\tilde{y} + e^{(k)}) - q(\tilde{y})}{q(\tilde{y})}\right)^2 & \text{if } a_k < \tilde{y}_{(k)} < b_k \\ 2\left(\frac{q(\tilde{y} + e^{(k)}) - q(\tilde{y})}{q(\tilde{y})}\right) + \left(\frac{q(\tilde{y} + e^{(k)}) - q(\tilde{y})}{q(\tilde{y})}\right)^2 & \text{if } \tilde{y}_{(k)} = a_k \end{cases}$$

Prequential vs. Batch approach

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

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

- "Batch" version*:
  - Easier to compute, only requires to estimate final evidence $p(y_{1:T}|M)$
  - But typically inconsistent
- Prequential version:
  - Generally consistent
  - Requires to estimate all the intermediary predictive $p(dy_t|y_{1:t-1}, M)$, but this can be achieved by using algorithms like SMC^2

*On the right hand side.

Prequential vs. Batch approach

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

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

- "Batch" version*:
  - Easier to compute, only requires to estimate final evidence $p(y_{1:T}|M)$
  - But typically inconsistent
- Prequential version:
  - Generally consistent
  - Requires to estimate all the intermediary predictive $p(dy_t|y_{1:t-1}, M)$, but this can be achieved by using algorithms like SMC[2]

*On the right hand side.

Prequential vs. Batch approach

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

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

- "Batch" version*:
  - Easier to compute, only requires to estimate final evidence $p(y_{1:T}|M)$
  - But typically inconsistent
- Prequential version:
  - Generally consistent
  - Requires to estimate all the intermediary predictive $p(dy_t|y_{1:t-1}, M)$, but this can be achieved by using algorithms like SMC^2

*On the right hand side.

Partial Bayes factors

- 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, not ideal to "waste" data for the training set especially in setting with few observations (cf. Red Kangaroos example where $T = 41$)

Fractional Bayes factors

- 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