TRACTABLE COMPUTATION OF BAYES FACTORS FOR ROBUST MODEL SELECTION IN THE PHYSICAL SCIENCES

Jack S. Hale

Probability & Statistics Seminar – 19th December 2024

Joint work with *Damian Mingo Ndiwago*, Remko Nijzink and Christophe Ley

TOPICS

- *Uncertainty quantification with a focus on problems in the physical sciences.*
- Finite element methods for elasticity problems on manifolds.
- Expressive mathematical software for the solution of partial differential equations.

Dr. Damian Mingo Ndiwago

Mrs. Sona Salehian Ghamsari

Mr. Andrey Latyshev

Luxembourg National Research Fund

OUTLINE

- Importance of the model selection problem.
- Motivating example in hydrology.
- The Bayesian parameter inference problem.
- By analogy, the Bayesian model selection problem.
- Recent work – robust computation of Bayes factor applied to hydrological models [Min+25].

ASK ANY MODELLER...

- I would like models that:
  - fit the data,
  - are simple (Occam's Razor),
  - and generalize beyond the data used for calibration.

ALL MODELS ARE WRONG…

Two questions:

1. Which modelling procedure, will, with enough data, identify the ‘true’ model?
2. *Based on the data, which model from a finite set of models, lies closest to the ‘true’ model?*

MODEL SELECTION CRITERIA

- Frequentist and Bayesian model selection criteria offer a principled way of answering the second question and selecting models.
- In particular, the Bayes factor, provides a principal way of selecting between models, and implicitly penalises model complexity.
- The Bayes factor is not widely used in practice to assess and select moderately complex models - difficult and expensive to compute.

MAGELA CREEK - LOCATION

Magela Creek - 600 km^2.

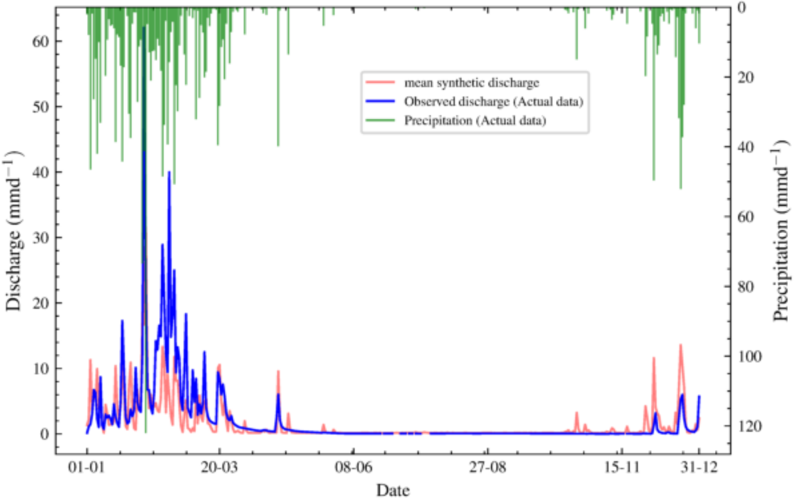
MAGELA CREEK - MAIN CHARACTERISTICS

Magela Creek - Summer rainfall, tropical.

MAGELA CREEK - DATA

Streamflow monitoring device. Source: Fine Art America

MAGELA CREEK - HYDROGRAPH



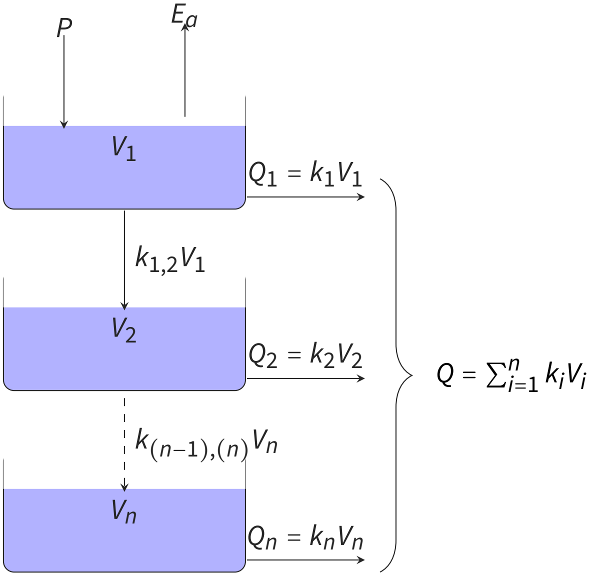
MODEL CONSTRUCTION IN HYDROLOGICAL SCIENCES

- Physically *inspired* conceptual models predominant.
  - Axiomatic physical principles or postulates, e.g. conservation of mass.
  - Empirical parametric models of system components in the absence of a complete or tractable theoretical framework e.g. transpiration, flow through ground, multiscale behaviour.
- Space of models is created by combining various existing components ‘suitable’ for the site.
- Model parameters calibrated against data.
- Success? - goodness of fit on held out data.

MAGELA CREEK - MODEL CONSTRUCTION

1. Approach hydrological modeller, Remko Nijzink.
2. After discussions on the site:
   - Large summer rainfall events – dominant fast runoff processes.
   - No snowfall or ice.
   - Equatorial - evapotranspiration likely to be important.
   - Uncertain number of slower processes (groundwater storage etc.)
3. Remko proposes something called the Hydrologiska Byråns Vattenbalansavdelning (HBV) approach.

MAGELA CREEK - HBV-LIKE MODELS



HBV-LIKE MODELS

$$
\begin{aligned}
(V_1)_t &= P - E_a - k_1 V_1, && && n = 1,\\
(V_1)_t &= P - E_a - k_1 V_1 - k_{1,2} V_1, && && n \geq 2,\\
(V_i)_t &= k_{(i-1),(i)} V_{i-1} - k_i V_i - k_{(i),(i+1)} V_i, && i = 2, \ldots, n-1, && n \geq 3,\\
(V_n)_t &= k_{(n-1),(n)} V_{n-1} - k_n V_n, && && n \geq 2,\\
V(t=0) &= \hat{V},\\
E_a &= \frac{E_p}{V_{\max}} V_1,\\
Q &= \sum_{i=1}^{n} k_i V_i.
\end{aligned}
$$

IN AN IDEAL WORLD...

Statement (The Bayesian model-parameter inference problem - [WHR12])

Define a space of models $\mathbb{M} = \{M_1, \ldots, M_n\}$ *each with its own parameter space* $\Theta_M \subset \mathbb{R}^{p_M}$ *with* $M \in \mathbb{M}$*. Given some data* $y \in \mathbb{R}^q$*, determine the posterior distribution (or its statistics)* $p(M, \theta_M \mid y)$ *on the joint model-parameter space* $\mathbb{M} \times \bigcup_{M \in \mathbb{M}} \Theta_M$*.*

$$n = \{2, 3, 4\},$$

$$\mathbb{M} = \{M_2, M_3, M_4\},$$

$$\Theta_M \subset \mathbb{R}^{3n},$$

$$\mathbb{R}^p \ni y = G_M(\theta_M) + \eta, \ \eta \sim N(0, \sigma^2).$$

START SIMPLE: THE PARAMETER INFERENCE PROBLEM

Statement (The Bayesian parameter inference problem - [WHR12])

Define a space of parameters $\Theta \subset \mathbb{R}^p$. *Given some data* $y \in \mathbb{R}^q$, *determine the posterior distribution (e.g. its statistics)* $p(\theta \mid y)$ *on the parameter space* Θ.

PARAMETER INFERENCE WORKFLOW

1. (Prior) Set the distribution on Θ

$$p(\theta).$$

2. (Likelihood) Acquire data y and assume model for y for fixed $\theta \in \Theta$

$$p(y \mid \theta).$$

3. Joint probability

$$p(\theta, y) = p(y \mid \theta)p(\theta).$$

4. (Posterior) Condition joint probability on data y

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}.$$

COMPUTATION [MFR24B]

- We would like to compute quantities such as posterior probabilities

$$\mathbb{E}\left[g(\theta) \mid y\right] = \int_{\Theta} g(\theta) p(\theta \mid y) \, \mathrm{d}\theta. \tag{2}$$

- In almost all interesting cases eq. (2) has no closed form solution.
- However, we can usually *compute* at fixed θ and y:

$$\log p(\theta \mid y) = \log p(y \mid \theta) + \log p(\theta) - \log p(y)$$

  where $p(y)$ is the unknown/hard-to-compute constant of proportionality.

THE WORK HORSE OF BAYES [MFR24B]

- The Metropolis-Hastings (et al.) algorithm can compute

$$\left\{\theta^{(1)},\ldots,\theta^{(M)}\right\} \xrightarrow{M\to\infty} p(\theta \mid y)$$

- and requires only evaluations of

$$\log p(\theta^{c} \mid y) - \log(\theta^{(i)} \mid y) = \log p(y \mid \theta^{c}) + \log p(\theta^{c}) \\ - \log p(y \mid \theta^{(i)}) - \log p(\theta^{(i)})$$

- The constant of proportionality $\log p(y)$ is gone!

GOOD NEWS

Claim

For some common problems (small data, moderate dimension parameter space, known likelihood etc.) it is possible to numerically solve the parameter inference problem [MFR24a].

WHAT ABOUT THE MODEL SELECTION PROBLEM?

Statement (The Bayesian model-parameter inference problem - [WHR12])

Define a space of models $\mathbb{M} = \{M_1, \ldots, M_n\}$ *each with its own parameter space* $\Theta_M \subset \mathbb{R}^{p_M}$ *with* $M \in \mathbb{M}$*. Given some data* $y \in \mathbb{R}^q$*, determine the posterior distribution* $p(M, \theta_M \mid y)$ *on the joint model-parameter space* $\mathbb{M} \times \bigcup_{M \in \mathbb{M}} \Theta_M$.

MODEL SELECTION WORKFLOW: PART I

1. *Hierarchical prior* Set a distribution on $\mathbb{M} \times \bigcup_{M \in \mathbb{M}} \Theta_M$

$$p(M) \text{ and } p(\theta_M \mid M).$$

2. *Likelihood* – Acquire data y and assume model for y on fixed $M \in \mathbb{M}$ and $\theta_M \in \Theta_M$

$$p(y \mid \theta_M, M).$$

3. *Joint probability*

$$p(y, \theta_M, M) = p(y \mid \theta_M, M) p(\theta_M \mid M) p(M) \tag{3}$$

MODEL SELECTION WORKFLOW: PART II

5. *Joint posterior* Condition joint probability on data y

$$p(M, \theta_M \mid y) = \frac{p(y \mid M, \theta_M) p(\theta_M \mid M) p(M)}{p(y)}.$$

Observations

- The dimension of the joint model-parameter space is larger.
- $\mathbb{M}$ is a discrete set.
- The size of the parameter space p_M varies across $\mathbb{M}$.

Problem

- *Standard* MCMC demands that the joint posterior has a density with respect to some underlying 'nice' measure – not the case.

MODEL SELECTION WORKFLOW: PART III

6. Marginalise across θ_M

$$p(M \mid y) = \int_{\Theta_M} p(M, \theta_M \mid y)\, d\theta_M$$

$$= \frac{p(M)}{p(y)} \overbrace{\int_{\Theta_M} p(y \mid M, \theta_M) p(\theta_M \mid M)\, d\theta_M}^{p(y \mid M)}.$$

7. Take two competing models $M_i \in \mathbb{M}$ and $M_j \in \mathbb{M}$ and write the posterior ratio

$$\frac{p(M_i \mid y)}{p(M_j \mid y)} = \frac{p(y \mid M_i)}{p(y \mid M_j)} \frac{p(M_i)}{p(M_j)} \frac{p(y)}{p(y)}.$$

MODEL SELECTION WORKFLOW: PART IV

- Under a non-informative prior on $\mathbb{M}$, the posterior ratio is equal to ratio of the marginal likelihoods (the Bayes factor).

$$\text{BF}_{ij} = \frac{p(M_i \mid y)}{p(M_j \mid y)} = \frac{p(y \mid M_i)}{p(y \mid M_j)}.$$

Challenge

If we can compute the marginal likelihood $p(y \mid M)$ then we can compute the Bayes factor – model selection.

$$p(y \mid M) = \int_{\Theta_M} p(y \mid M, \theta_M) p(\theta_M \mid M) \, d\theta_M.$$

SUMMARY SO FAR...

- The Bayesian parameter inference problem is straightforward.
- The Bayesian parameter-model inference problem is not.
- There is a ‘nearby’ quantity, the marginal likelihood $p(y \mid M_i)$ and hence the Bayes factor, which may be tractable.

NAIVE MONTE CARLO – NO FREE LUNCH [FW12]

- Let's take a look at the marginal likelihood $p(y \mid M)$

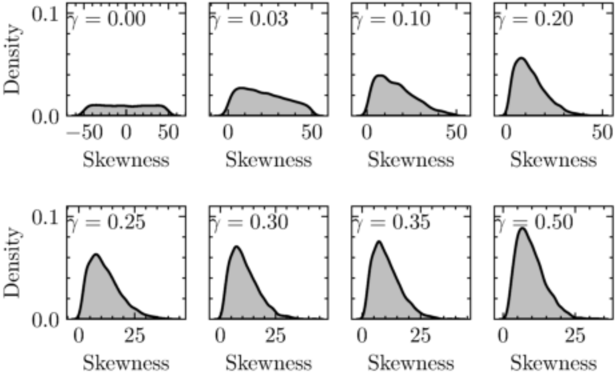
$$
\begin{aligned}
p(y \mid M) &= \int_{\Theta_M} \underbrace{p(y \mid M, \theta_M)}_{\text{likelihood}} \underbrace{p(\theta_M \mid M)}_{\text{prior}} \; d\theta_M, \\
&\approx \frac{1}{N} \sum_{i=1}^{N} p(y \mid M, \theta_M^{(i)}), \quad \theta_M^{(i)} \overset{\text{iid}}{\sim} p(\theta_M \mid M).
\end{aligned}
$$

- Prior weighted likelihood.
- Integrand may have sharp peaks or heavy tails and does not 'concentrate' like posterior.
- Likelihood evaluation expensive - 10 million+ evaluations for simple problems.

THERMODYNAMIC INTEGRAL: PART I [FP08]

- Define the *power posterior*. For inverse temperature parameter $\gamma \in [0, 1]$

$$p_\gamma(\theta_M \mid y, M) \propto p(y \mid \theta_M, M)^\gamma p(\theta_M \mid M)$$



THERMODYNAMIC INTEGRAL: PART II [FP08]

- The power posterior normalising constant $Z(\gamma)$ is

$$Z(\gamma) = \int_{\Theta_M} p(y \mid M, \theta_M)^{\gamma} p(\theta_M \mid M)\, d\theta_M$$

- Differentiate $\log Z(\gamma)$ with respect to γ using chain rule

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\gamma} \log Z(\gamma) &= \frac{1}{Z(\gamma)} \frac{\mathrm{d}}{\mathrm{d}\gamma} Z(\gamma) \\ &= \mathbb{E}_{p_{\gamma}(\theta_M \mid y, M)}[\log p(D \mid \theta_M, M)] \end{aligned}$$

THERMODYNAMIC INTEGRAL: PART III [FP08]

- Use second fundamental theory of calculus on the interval $\gamma \in [0, 1]$

$$\underbrace{\log Z(1)}_{\log p(y \mid M)} - \underbrace{\log Z(0)}_{\log \int_{\theta_M} p(\theta_M \mid M) = 0}$$

$$= \int_0^1 \mathbb{E}_{p_\gamma(\theta_M \mid y, M)} [\log p(y \mid \theta_M, M)] \, \mathrm{d}\gamma.$$

Result [FP08]

The logarithm of the marginal likelihood is equal to the thermodynamic integral

$$\log p(y \mid M) = \int_0^1 \mathbb{E}_{p_\gamma(\theta_M \mid y, M)} [\log p(y \mid \theta_M, M)] \, \mathrm{d}\gamma.$$

MAIN CONTRIBUTIONS OF RECENT WORK - [MIN+25]

- An algorithm *Replica Exchange Hamiltonian Monte Carlo + Thermodynamic Integration* that can simultaneously compute:
  - parameter estimates
  - marginal likelihood (hence, the Bayes factor).

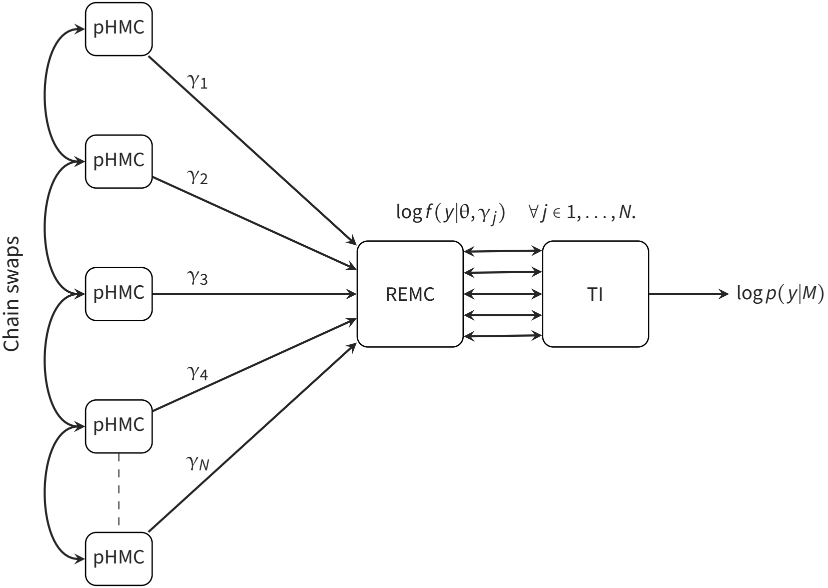
  for general Bayesian inference problems expressible in the probabilistic programming framework TFP with:
  - moderate parametric dimension $\dim(\theta_M) \sim 50$,
  - possibly multimodal and highly constrained posteriors.

ALGORITHMIC ASPECTS: I

$$\log p(y \mid M) = \int_0^1 \mathbb{E}_{p_\gamma(\theta_M \mid y, M)}[\log p(y \mid \theta_M, M)] \, \mathrm{d}\gamma$$

- *Thermodynamic integral*: Trapezoidal rule.
- *Power posterior*:
  - strong correlations: preconditioned Hamiltonian Monte Carlo.
  - multi-modality: replica exchange.
- *Likelihood*: Monte Carlo via Power posterior samples.

ALGORITHMIC ASPECTS: II

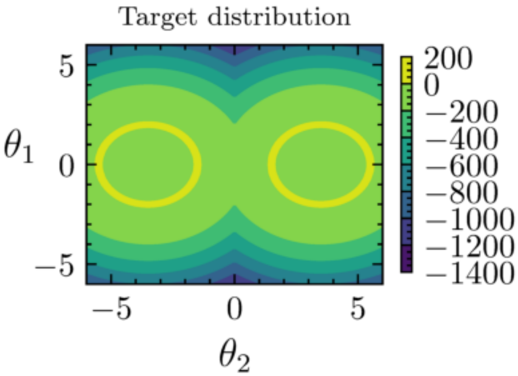


GAUSSIAN SHELLS: I

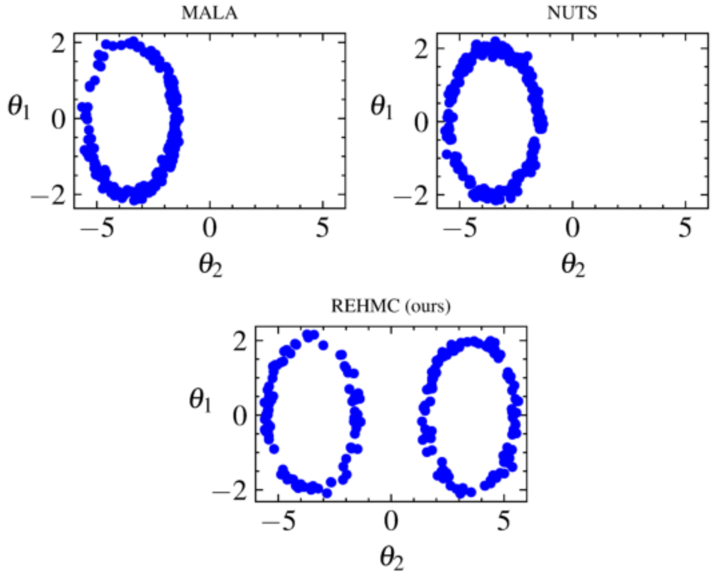
- Benchmark problem of two Gaussian Shells in moderate dimensions.
- Shells separated by region of low probability.
- Concentration of measure in high dimensions.

$$\ell(\theta) = \frac{1}{\sqrt{2\pi w_1^2}} \exp\left[- \frac{(||\theta_1 - c_1|| - r_1)^2}{2w_1^2} \right] + \frac{1}{\sqrt{2\pi w_2^2}} \exp\left[- \frac{(||\theta_2 - c_2|| - r_2)^2}{2w_2^2} \right].$$

GAUSSIAN SHELLS: II



GAUSSIAN SHELLS: III



GAUSSIAN SHELLS: IV

Dimensions	*Reference log $p(y)$	Estimated log $p(y)$
2	-1.75	-1.75 $\pm$ 0.003
5	-5.67	-5.68 $\pm$ 0.006
10	-14.59	-14.60 $\pm$ 0.006
20	-36.09	-36.12 $\pm$ 0.014
30	-60.13	-60.19 $\pm$ 0.025

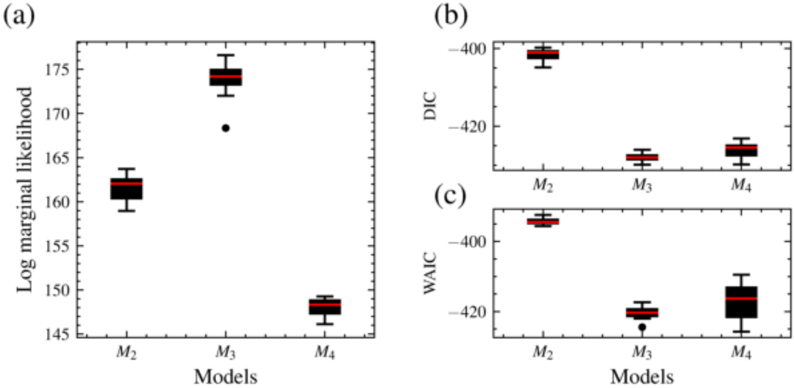
* As reported in Feroz et al. (2009)

MAGELA CREEK EXAMPLE - CONSISTENCY

Check that the Bayes Factor selects the parsimonious model when the 'true' model is in $\mathbb{M}$.

1. Generate data from the intermediate model M_3.
2. This leaves the simpler model M_2 which cannot fit M_3, and M_4, which due to the nested construction can, but with higher complexity.
3. Compute the marginal likelihood, DIC and WAIC for each model.

MAGELA CREEK EXAMPLE - CONSISTENCY



MAGELA CREEK EXAMPLE - REAL DATA

See which model the Bayes Factor selects when the ‘true’ model (i.e. reality) is not in $\mathbb{M}$.

1. Use real run-off data from Magela Creek.
2. All models in $\mathbb{M}$ are possible candidates.
3. Compute the marginal likelihood, DIC and WAIC for each model.

	M_2 (95 % CI)	M_3 (95 % CI)	M_4 (95 % CI)
k_1	0.724 (0.517, 0.940)	0.794 (0.0.574, 1.046)	1.169 (0.774, 1.520)
k_2	0.125 (0.081, 0.174)	0.242 (0.155, 0.344)	1.991 (1.192, 2.801)
k_3	-	0.157 (0.096, 0.221)	1.352 (0.720, 1.964)
k_4	-	-	1.067 (0.598, 1.546)
$k_{1,2}$	1.195 (0.838, 1.637)	1.923 (1.105, 2.889)	2.292 (1.367, 3.417)
$k_{2,3}$	-	0.511 (0.380, 0.648)	0.728 (0.463, 0.983)
$k_{3,4}$	-	-	0.826 (0.497, 1.136)
$\hat{V}_1$	1.030 (0.548, 1.530)	1.029 (0.566, 1.457)	1.140 (0.032, 2.893)
$\hat{V}_2$	1.017 (0.593, 1.549)	0.999 (0.582, 1.477)	0.861 (0.048, 2.239)
$\hat{V}_3$	-	0.997 (0.569, 1.523)	0.940 (0.041, 2.325)
$\hat{V}_4$	-	-	1.082 (0.060, 2.768)
$V_{\max}$	1.139 (0.808, 1.474)	0.912 (0.657, 1.201)	0.796 (0.549, 1.057)
σ^2	5.289 (4.694, 5.830)	5.273 (4.739, 5.828)	5.847 (5.212, 6.499)
$\log p(y\|M)$	-506.259	-529.483	-608.181
$\log \text{BF}_{23}$	23.224	-	-
$\log \text{BF}_{24}$	101.922	-	-
DIC	940.352	940.397	969.722
WAIC	946.536	946.512	979.932

SUMMARY

- For modellers in the physical sciences model structure, and consequently model selection, is as an important problem as parameter inference.
- Full Bayes on $\mathbb{M} \bigcup_{M \in \mathbb{M}} \Theta_M$ is still intractable, but Bayes factors provide a reasonable alternative for model selection.
- We propose a performant methodology for robustly computing the marginal likelihood, and hence the Bayes factor.
- It is not hard to find cases where information theoretic approaches fail and Bayes factor succeeds - that robustness (still) comes at a computational cost.

[FP08] N. Friel and A. N. Pettitt. “Marginal Likelihood Estimation via Power Posteriors”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70.3 (July 2008), pp. 589–607. DOI: 10.1111/j.1467-9868.2007.00650.x. (Visited on 12/17/2024).

[FW12] Nial Friel and Jason Wyse. “Estimating the evidence – a review”. en. In: *Statistica Neerlandica* 66.3 (2012), pp. 288–308. ISSN: 1467-9574. DOI: 10.1111/j.1467-9574.2011.00515.x. (Visited on 12/17/2024).

[MFR24a] Gael M. Martin, David T. Frazier, and Christian P. Robert. “Approximating Bayes in the 21st Century”. In: *Statistical Science* 39.1 (Feb. 2024). DOI: 10.1214/22-STS875.

[MFR24b] Gael M. Martin, David T. Frazier, and Christian P. Robert. “Computing Bayes: From Then ’Til Now’”. In: *Statistical Science* 39.1 (Feb. 2024). DOI: 10.1214/22-STS876.

[Min+25] Damian N. Mingo et al. “Selecting a conceptual hydrological model using Bayes’ factors computed with Replica Exchange Hamiltonian Monte Carlo and Thermodynamic Integration”. English. In: *Geoscientific Model Development* (2025), pp. 1–45. DOI: 10.5194/egusphere-2023-2865.

[WHR12] Ernst Wit, Edwin van den Heuvel, and Jan-Willem Romeijn. “‘All models are wrong...’: an introduction to model uncertainty”. In: *Statistica Neerlandica* 66.3 (2012), pp. 217–236. DOI: 10.1111/j.1467-9574.2012.00530.x.