A sampling-based implementation of Bayesian model-selection of Stephan et al. (2009)

Alireza Modirshanechi

The Bayesian model selection with random effects, as proposed by Stephan et al. (2009), has become the standard approach to model selection in psychology, cognitive science, and neuroscience (Wilson and Collins, 2019). The standard implementation is based on approximate, variational inference (MacKay, 2003). Here, we propose an alternative, efficient, and asymptotically exact implementation based on Markov chain Monte Carlo (MCMC) sampling. This implementation has been previously used in Modirshanechi et al. (2025a,b).

1 Original generative model of Stephan et al. (2009)

We have N subjects, and K candidate models. Main variables:

$$\begin{aligned} D_i &= \text{data of subject } i \in \{1, \dots, N\} \\ M_i \in \{1, ..., K\} &= \text{model index assigned to subject } i \in \{1, \dots, N\} \\ R_k \in [0, 1] &= \text{frequency of model } k \in \{1, ..., K\}. \end{aligned} \tag{1}$$

Given the prior hyper-parameter $\alpha \in \mathbb{R}_+^K$, the data generative model of Stephan et al. (2009) is

$$\begin{aligned} R_{1:K} &\sim \text{Dirichlet}(\alpha) & \\ M_i | R_{1:K} &\sim \text{Categorical}(R_{1:K}) & \text{independently for } i \in \{1, \dots, N\}, \\ D_i | M_{1:N}, R_{1:K} &\sim P_{M_i}, & \text{independently for } i \in \{1, \dots, N\}, \end{aligned} \tag{2}$$

where P_k is pre-specified and denotes the data distribution according model $k \in \{1, \dots, K\}$.

2 Inference problem

The goal of inference is to find

$$\log P(R_{1:K}, M_{1:N} | D_{1:N}) = \log P(R_{1:K}, M_{1:N}, D_{1:N}) - \underbrace{\log P(D_{1:N})}_{\text{constant}}, \tag{3}$$

where

$$\begin{aligned} \log P(R_{1:K}, M_{1:N}, D_{1:N}) &= \log P(D_{1:N} | R_{1:K}, M_{1:N}) + \log P(M_{1:N} | R_{1:K}) + \log P(R_{1:K}) \\ &= \sum_{i=1}^{N} L_{i,M_i} + \sum_{k=1}^{K} C_k(M_{1:N}) \log R_k + \log \text{Dir}(R_{1:K}; \alpha), \end{aligned} \tag{4}$$

and we defined

$$L_{i,M_i} := \log P_{M_i}(D_i) \quad \text{and} \quad C_k(M_{1:N}) := |\{i : M_i = k\}|. \tag{5}$$

Stephan et al. (2009) use variational inference (MacKay, 2003) to find the approximation

$$P(R_{1:K}|D_{1:N}) \approx \text{Dirichlet}(R_{1:K};\alpha').$$

Given this approximation, the main statistics of inference in Stephan et al. (2009) are the expected model frequency

$$\mathbb{E}[R_k|D_{1:N}] \tag{6}$$

and the exceedance probability

$$P[R_k > R_{k'}\ \forall k' \neq k|D_{1:N}] \tag{7}$$

for $k \in \{1,\ldots,K\}$.

3 MCMC-based inference

In this section, we use MCMC sampling (Efron and Hastie, 2016) to estimate the full posterior $\log P(R_{1:K}, M_{1:N}|D_{1:N})$. Specifically, we use the Metropolis-Hastings algorithm (Efron and Hastie, 2016) with a base chain tailored to the model-selection problem.

The MCMC procedure describes sampling from a Markov chain defined by the vector

$$X^{(t)} = \left(R^{(t)}_{1:K}, M^{(t)}_{1:N}\right) \tag{8}$$

at time $t \in \mathbb{N}$. The base chain below describes the evolution $X^{(t)}$ independently of the data $D_{1:N}$. The acceptance probability introduced late will be used to correct the samples of the base chain such that the marginal distribution of $X^{(t)}$ asymptotically converges to $P(R_{1:K}, M_{1:N}|D_{1:N})$.

Base chain. The base chain describes $P(X^{(t+1)}|X^{(t)})$ and has 3 hyper-parameters:

$$\begin{aligned} N_{\text{change}} \in \{1,\ldots,N\} &= \text{number of subjects whose models may change from } M^{(t)} \text{ to } ^{(t+1)} \\ \epsilon \in \mathbb{R}_+ &= \text{controlling the tendency of } R^{(t+1)} \text{ being sampled close to uniform} \\ N_{\text{scale}} \in \mathbb{R}_+ &= \text{controlling the extent of dependence of } R^{(t+1)} \text{ on } R^{(t)}. \end{aligned} \tag{9}$$

We specify

$$P_{\text{base}}(X^{(t+1)}|X^{(t)}) = P_{\text{base}}(M^{(t+1)}|M^{(t)}; N_{\text{change}})P_{\text{base}}(R^{(t+1)}|M^{(t)}; \epsilon, N_{\text{scale}}) \tag{10}$$

where

$$P_{\text{base}}(M^{(t+1)} = m|M^{(t)}; N_{\text{change}}) = \begin{cases} c > 0 & \text{if } |\{i : m_i \neq M_i^{(t)}\}| \leq N_{\text{change}} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

and

$$P_{\text{base}}(R^{(t+1)} = r|M^{(t)}; \epsilon, N_{\text{scale}}) = \text{Dirichlet}\left(r; \left\{\epsilon + \frac{C_k(M^{(t)}_{1:N})}{N_{\text{scale}}}\right\}_{k=1}^{K}\right) \tag{12}$$

with $C_k(M^{(t)}_{1:N})$ defined in Eq. 5. *This base distribution corresponds to the sampling procedure in Lines 6-9 in Algorithm 1.*

Algorithm 1 Pseudocode for MCMC-based inference

```
1: Set the prior parameter $\alpha$.
2: Set the MCMC hyper parameters $N_{\text{change}}$, $N_{\text{scale}}$, and $\epsilon$ as well as the number of samples $T$.
3: Specify data likelihood $L_{i,k}$ for all $i \in \{1, \dots, N\}$ and $k \in \{1, \dots, K\}$.
4: Set $X^{(0)} = x^*$
5: for $t = 0 : T - 1$ do
6:     Sample a sub-set of size $N_{\text{change}}$ of subjects' indexes: $\{i_1, \dots, i_{N_{\text{change}}}\}$.
7:     For every $i \in \{i_1, \dots, i_{N_{\text{change}}}\}$, sample $M_i^{(t+1)}$ from Uniform($\{1, \dots, K\}$).
8:     For every $i \notin \{i_1, \dots, i_{N_{\text{change}}}\}$, set $M_i^{(t+1)} \leftarrow M_i^{(t)}$.
9:     Sample $R^{(t+1)}$ from Dirichlet($\alpha^{(t)}$) with $\alpha_k^{(t)} = \epsilon + C_k(M_{1:N}^{(t)})/N_{\text{scale}}$.
10:    if $t = 0$ then
11:        Set $X^{(t+1)} \leftarrow (R^{(t+1)}, M^{(t+1)})$
12:    else
13:        Set $X^{(t+1)} \leftarrow (R^{(t+1)}, M^{(t+1)})$ and evaluate $a^{(t+1)}$ using Eq. 13.
14:        Accept $X^{(t+1)}$ with probability $a^{(t+1)}$; otherwise set $X^{(t+1)} \leftarrow X^{(t)}$.
15:    end if
16: end for
```

Acceptance probabilities. Following the Metropolis-Hastings algorithm (Efron and Hastie, 2016), the acceptance probability of sample $X^{(t+1)}$ proposed by the base-chain is given by

$$a^{(t+1)} = \min\left\{1\ ,\ \frac{\pi\big(X^{(t+1)}\big)}{\pi\big(X^{(t)}\big)} \cdot \frac{\phi\big(X^{(t+1)} \to X^{(t)}\big)}{\phi\big(X^{(t)} \to X^{(t+1)}\big)}\right\} \tag{13}$$

where

$$\pi\big(X\big) = P(R_{1:K}, M_{1:N}, D_{1:N}) \tag{14}$$

is given by Eq. 4 and

$$\phi\big(X \to X'\big) = P_{\text{base}}(X'|X) \tag{15}$$

is given by Eq. 10, Eq. 11, and Eq. 12. *The acceptance step corresponds to the sampling procedure in Lines 13-14 in Algorithm 1.*

Initial sample. To set the starting point for $t = 1$, we first define

$$m_i^* = \arg\max_{k \in \{1, \dots, K\}} L_{i,k} \quad i \in \{1, \dots, N\}, \tag{16}$$

and

$$r_k^* = \frac{C_k(m_{1:N}^*)}{N}, \tag{17}$$

where $L_{i,k}$ and C_k are defined in Eq. 5. Then, the distribution of the initial point $X^{(1)}$ is given by

$$P_{\text{initial}}(X^{(1)} = x) = P_{\text{base}}(x|x^*) \tag{18}$$

with $x^* = (r^*, m^*)$. Alternatively, $X^{(1)}$ can also be sampled from its prior in Eq. 2.

Typical statistics of interest. The typical statistics of interest in the Bayesian model-selection of Stephan et al. (2009) can be estimated via our MCMC samples $x^{(1:T)}$:

1. Expected model frequency:
$$\mathbb{E}[R_k|D_{1:N}] \approx \bar{r}_k = \frac{\sum_{t=1}^{T} r_k^{(t)}}{T}. \tag{19}$$

2. Posterior variance of model frequency:
$$\mathrm{Var}[R_k|D_{1:N}] \approx \frac{\sum_{t=1}^{T} \left(r_k^{(t)} - \bar{r}_k\right)^2}{T-1}. \tag{20}$$

3. Exceedance probability (XP):
$$P[R_k > R_{k'} \; \forall k' \neq k | D_{1:N}] \approx \frac{\sum_{t=1}^{T} \mathbb{1}_{r_k^{(t)} > r_{k'}^{(t)} \; \forall k' \neq k}}{T}. \tag{21}$$

4. Individual model probability:
$$P[M_i = k|D_{1:N}] \approx \frac{\sum_{t=1}^{T} \mathbb{1}_{m_i^{(t)} = k}}{T}. \tag{22}$$

4 Inference with accounting for null hypothesis

In an extension of the model-selection approach of Stephan et al. (2009), Rigoux et al. (2014) propose the addition of the null hypothesis H_0 to the generative model. Given the prior hyper-parameter $\alpha \in \mathbb{R}_+^K$, the data generative model of Rigoux et al. (2014) is

$$\begin{aligned}
H_0 &\sim \text{Uniform}(\{0,1\}) \\
R_{1:K}|H_0 &\sim \begin{cases} \delta_{\{\frac{1}{K}\}_{k=1}^{K}} & \text{if } H_0 = 1 \\ \text{Dirichlet}(\alpha) & \text{if } H_0 = 0 \end{cases} \\
M_i|R_{1:K}, H_0 &\sim \text{Categorical}(R_{1:K}) && \text{independently for } i \in \{1,\dots,N\}, \\
D_i|M_{1:N}, R_{1:K}, H_0 &\sim P_{M_i}, && \text{independently for } i \in \{1,\dots,N\},
\end{aligned} \tag{23}$$

where δ denotes the Dirac delta distribution, and p_k for $k \in \{1,\dots,K\}$ are pre-specified.

Correction with respect to the null hypothesis. If we denote the probability distribution under the generative model in Eq. 2 by P and under the generative model in Eq. 23 by P_+, then

$$\mathbb{E}_{P_+}[f(R_{1:K})|D_{1:N}] = P_+(H_0 = 1|D_{1:N}) f\big(\{1/K\}_{k=1}^{K}\big) + P_+(H_0 = 0|D_{1:N}) \mathbb{E}_P[f(R_{1:K})|D_{1:N}] \tag{24}$$

for any measurable function $f : \mathbb{R}^K \to \mathbb{R}$.

The key point of Eq. 24 is that having $P_+(H_0 = 1|D_{1:N})$ enables us to correct the statistics evaluated from inference on P by accounting for the possibility of $H_0 = 1$. Hence, Rigoux et al. (2014) called this quantity *Bayesian omnibus risk (BOR)*. In particular, Rigoux et al. (2014) was interested in the *protected exceedance probability (PXP)* defined as

$$\underbrace{P_+[R_k > R_{k'} \; \forall k' \neq k|D_{1:N}]}_{\text{PXP}} = \frac{\text{BOR}}{K} + (1 - \text{BOR}) \underbrace{P[R_k > R_{k'} \; \forall k' \neq k|D_{1:N}]}_{\text{XP; see Eq. 21}}. \tag{25}$$

Algorithm 2 Pseudocode for MC-based estimation of BOR

```
Set the prior parameter $\alpha$.
Set the number of samples $T$.
Specify data likelihood $L_{i,k}$ for all $i \in \{1, \dots, N\}$ and $k \in \{1, \dots, K\}$.
Evaluate $P_+(D_{1:N}|H_0 = 1) = P(D_{1:N}|R = \{1/K\}_{k=1}^K)$ using Eq. 30
for $t = 1 : T$ do
    Sample $r^{(t)} \sim \text{Dirichlet}(\alpha)$.
    Evaluate $p^{(t)} = P(D_{1:N}|R = r^{(t)})$ using Eq. 30.
end for
Estimate $P_+(D_{1:N}|H_0 = 0) \approx \sum_{t=1}^T p^{(t)}/T$.
Estimate BOR using Eq. 26.
```

MC-based estimation of BOR. BOR is given by

$$\text{BOR} := P_+(H_0 = 1|D_{1:N}) = \frac{P_+(D_{1:N}|H_0 = 1)}{P_+(D_{1:N}|H_0 = 0) + P_+(D_{1:N}|H_0 = 1)}. \tag{26}$$

Hence, in order to estimate BOR, we evaluate

$$P_+(D_{1:N}|H_0 = 1) = P(D_{1:N}|R = \{1/K\}_{k=1}^K) \tag{27}$$

and use Monte-Carlo sampling to estimate

$$P_+(D_{1:N}|H_0 = 0) \approx \frac{1}{T} \sum_{t=1}^T P(D_{1:N}|R = r^{(t)}) \quad \text{with } r^{(1:T)} \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(\alpha). \tag{28}$$

Marginalization over M. The challenge in evaluating Eq. 27 and Eq. 28 is that we need to evaluate the marginal probability $P(D_{1:N}|R = r)$. Here, we show how this can be done analytically:

$$\begin{aligned} P(D_{1:N}|R = r) &= \sum_{m_{1:N} \in \{1,\dots,K\}^N} P(D_{1:N}|M_{1:N} = m_{1:N}, R = r) P(M_{1:N} = m_{1:N}|R = r) \\ &= \sum_{m_{1:N} \in \{1,\dots,K\}^N} \prod_{i=1}^N P(D_i|M_i = m_i) r_{m_i} \end{aligned} \tag{29}$$

which can be further simplified

$$P(D_{1:N}|R = r) = \prod_{i=1}^N \left(\sum_{k=1}^K r_k \exp L_{i,k} \right), \tag{30}$$

which can be efficiently computed using simple tricks to handle numerical errors.

References

B. Efron and T. Hastie. *Computer age statistical inference.* Cambridge University Press, 2016.

D. J. MacKay. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

A. Modirshanechi, P. Dayan, and E. Schulz. An integrative framework for the human sense of control. *PsyArXiv*, 2025a. doi: 10.31234/osf.io/cnkyz_v1.

A. Modirshanechi, W.-H. Lin, H. A. Xu, M. H. Herzog, and W. Gerstner. Even if suboptimal, novelty drives human exploration. *bioRxiv*, 2025b. doi: 10.1101/2022.07.05.498835.

L. Rigoux, K. E. Stephan, K. J. Friston, and J. Daunizeau. Bayesian model selection for group studies—revisited. *NeuroImage*, 84:971–985, 2014. doi: 10.1016/j.neuroimage.2013.08.065.

K. E. Stephan, W. D. Penny, J. Daunizeau, R. J. Moran, and K. J. Friston. Bayesian model selection for group studies. *NeuroImage*, 46(4):1004–1017, 2009. doi: 10.1016/j.neuroimage.2009.03.025.

R. C. Wilson and A. G. Collins. Ten simple rules for the computational modeling of behavioral data. *eLife*, 8:e49547, 2019. doi: 10.7554/eLife.49547.