Bayesian model choice

Jean-Michel Marin

University of Montpellier
Faculty of Sciences

HAX918X / 2024-2025

1. Bayesian discrimination between models

2. Bayesian Model Averaging

3. Difficulties with the Bayesian model choice paradigm

Bayesian discrimination between models

When are comparing models with indices $k = 1, 2, \ldots, J$, we introduce a model indicator $\mathfrak{M}$ taking values in $\{1, 2, \ldots, J\}$ and representing the index of the “true” model

If $\mathfrak{M} = k$, the data $\mathscr{D}_n$ are generated from a statistical model $\mathfrak{M}_k$ with likelihood $\ell_k(\theta_k|\mathscr{D}_n)$ and parameter $\theta_k \in \Theta_k$

Bayesian discrimination between models

Bayes procedures will depend on the posterior probabilities in the model space

$$\mathbb{P}^{\pi}(\mathfrak{M} = k | \mathscr{D}_n)$$

Bayesian discrimination between models

The prior π is defined over the collection of model indices $\{1, 2, \ldots, J\}$, and, conditionally on the model index $\mathfrak{M}$, on the corresponding parameter space Θ_k

Choice of the prior model probabilities $\mathbb{P}^\pi(\mathfrak{M} = k)$

- in some cases, there is experimental or subjective evidence about those probabilities,
- typically, we are forced to settle for equal weights $\mathbb{P}^\pi(\mathfrak{M} = k) = 1/J$

Bayesian discrimination between models

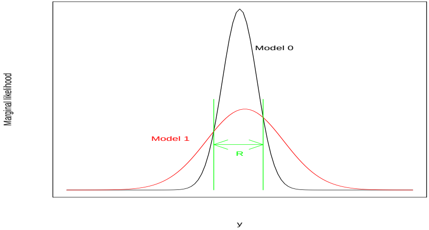
A key quantity, the integrated likelihood, also called the evidence

$$\mathbb{P}^{\pi}(\mathfrak{M} = k|\mathscr{D}_n) \propto \mathbb{P}^{\pi}(\mathfrak{M} = k) \int \ell_k(\theta_k|\mathscr{D}_n)\pi_k(\theta_k)\, \mathrm{d}\theta_k$$

$\mathbb{P}^{\pi}(\mathfrak{M} = k|\mathscr{D}_n)$ **is the core object in Bayesian model choice, the default procedure is to select the model with the highest posterior probability**

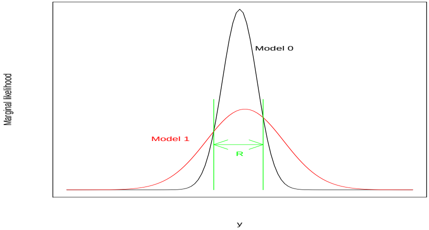
Bayesian discrimination between models

Why Bayesian inference embodies Occam's razor?



A simple model, like Model 0, makes only a limited range of predictions; a more powerful model, like Model 1, that has, for example, more free parameters, is able to predict a greater variety of data sets

Bayesian discrimination between models



Suppose that equal prior probabilities have been assigned to the two models. Then, if the data set falls in region R, the less powerful model will be the more probable model

The marginal likelihood corresponds to a penalized likelihood!

The BIC information criterium comes from an asymptotic Laplace approximation of the evidence

Bayesian discrimination between models

Bayesian test and model choice: the same problem

For instance, given a single observation $x|\theta \sim \mathscr{N}(\theta, 1)$

If $\theta \sim \mathscr{N}(0, 1)$, the posterior distribution $\theta|x \sim \mathscr{N}(x/2, 1/2)$

If the question of interest is to decide whether θ is negative or positive, we can directly compute

$$\begin{aligned} \mathbb{P}^\pi(\theta < 0|x) &= \mathbb{P}^\pi\left(\sqrt{2}(\theta - x/2) < -\sqrt{2}x/2|x\right) \\ &= \Phi\left(-x/\sqrt{2}\right) \end{aligned}$$

where Φ is the standard gaussian cdf

Bayesian discrimination between models

This computation does not seem to follow from the principles we just stated but it is only a matter of perspective

Let model 1 be the model such that $\theta < 0$ and model 2 the model such that $\theta > 0$

Using the fact that $\theta \sim \mathscr{N}(0, 1)$, we can derive the priors on both models from the original prior

Let π_1 be the prior distribution of θ under model 1, we have

$$\pi_1(\theta) = 2\frac{\exp(-\theta^2/2)}{\sqrt{2\pi}}\,\mathbb{I}_{\theta<0}$$

a truncated gaussian distribution

Bayesian discrimination between models

Let π_2 be the prior distribution of μ under model 2, we have

$$\pi_2(\theta) = 2\frac{\exp(-\theta^2/2)}{\sqrt{2\pi}}\,\mathbb{I}_{\theta>0}$$

another truncated gaussian distribution

Moreover, using the fact that $\mu \sim \mathcal{N}(0,1)$, we deduce

$$\mathbb{P}^\pi(\theta<0) = \mathbb{P}^\pi(\theta>0) = 1/2$$

Bayesian discrimination between models

$$\int_{\mathbb{R}} f(x|\theta)\pi_1(\theta)d\theta =$$
$$\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right) \frac{2}{\sqrt{2\pi}} \exp(-\theta^2/2)d\theta =$$
$$\frac{2\exp(-x^2/4)}{\sqrt{2\pi}\sqrt{2}} \int_{-\infty}^{0} \frac{\sqrt{2}}{\sqrt{2\pi}} \exp\left(-(\theta - x/2)^2\right) d\theta =$$
$$\frac{2\exp(-x^2/4)}{\sqrt{2\pi}\sqrt{2}} \mathbb{P}^{\pi}(\theta < 0|x)$$

And, with exactly the same type of calculations

$$\int_{\mathbb{R}} f(x|\theta)\pi_2(\theta)d\theta = \frac{2\exp(-x^2/4)}{\sqrt{2\pi}\sqrt{2}} \mathbb{P}^{\pi}(\theta > 0|x)$$

Bayesian discrimination between models

Then

$$\mathbb{P}^\pi(\mathfrak{M}=1|x) = \frac{(1/2)\frac{2\exp(-x^2/4)}{\sqrt{2\pi}\sqrt{2}}\mathbb{P}^\pi(\theta<0|x)}{(1/2)\frac{2\exp(-x^2/4)}{\sqrt{2\pi}\sqrt{2}}\left(\mathbb{P}^\pi(\theta<0|x)+\mathbb{P}^\pi(\theta>0|x)\right)}$$

$$\mathbb{P}^\pi(\mathfrak{M}=1|x) = \frac{\mathbb{P}^\pi(\theta<0|x)}{\mathbb{P}^\pi(\theta<0|x)+\mathbb{P}^\pi(\theta>0|x)}$$

$$\mathbb{P}^\pi(\mathfrak{M}=1|x) = \mathbb{P}^\pi(\theta<0|x) = \Phi\left(-x/\sqrt{2}\right) = \mathbb{P}^\pi(\theta<0|x)$$

Bayesian test and model choice: the same answer

Bayesian discrimination between models

The Bayes factor

$$B_{21}^{\pi}(\mathscr{D}_n) = \frac{\mathbb{P}^{\pi}(\mathfrak{M} = 2|\mathscr{D}_n)/\mathbb{P}^{\pi}(\mathfrak{M} = 1|\mathscr{D}_n)}{\mathbb{P}^{\pi}(\mathfrak{M} = 2)/\mathbb{P}^{\pi}(\mathfrak{M} = 1)}$$

While this quantity is a simple one-to-one transform of the posterior probability, it can be used for Bayesian model choice without first resorting to a determination of the prior weights of both models

$$B_{21}^{\pi}(\mathscr{D}_n) = \frac{\int_{\Theta_2} \ell_2(\theta_2|\mathscr{D}_n)\pi_2(\theta_2)\,\mathrm{d}\theta_2}{\int_{\Theta_1} \ell_1(\theta_1|\mathscr{D}_n)\pi_1(\theta_1)\,\mathrm{d}\theta_1} = \frac{m_2(\mathscr{D}_n)}{m_1(\mathscr{D}_n)}$$

Bayesian Model Averaging

The posterior probabilities in the model space can be used to average over the decisions coming from different models

Suppose that we are interested in the prediction of z and that, for model k, the predictive distribution of z is $g_k(z|\mathscr{D}_n)$

The average predictive of z is

$$\sum_{k=1}^{J} \mathbb{P}^\pi(\mathfrak{M} = k|\mathscr{D}_n) g_k(z|\mathscr{D}_n)$$

Difficulties with the Bayesian model choice paradigm

Prior difficulties:

- When we have prior informations, how to choose the prior distributions on the parameters of each model in a compatible way? What about the prior distribution in the models's space?
- When we do not have any prior information, **we can not use improper prior distribution.** Indeed, in that case, the models's posterior probabilities are only defined up to some arbitrary constants. How to choose the various prior distributions?

Difficulties with the Bayesian model choice paradigm

Computational difficulties:

- How to approximate the various posterior probabilities?
- How to approximate the evidences?
- When the number of models in consideration is huge, how to explore the models's space?