V - Model Choice

The goal of this short chapter is to present the Bayesian answer to the important problem of selecting one model among a set of $M \in \bar{\mathbb{N}}$ competing models:

$$\mathcal{M}_i = \{f_i(\cdot|\theta_i),\, \theta_i \in \Theta_i \subset \mathbb{R}^{d_i}\}, \quad i = 1, ..., M.$$

Model choice can be seen as an extension of hypothesis testing since testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is equivalent to choosing between the model

$$\mathcal{M}_0 = \{f(\cdot|\theta),\, \theta \in \Theta_0\}$$

and the model

$$\mathcal{M}_1 = \{f(\cdot|\theta),\, \theta \in \Theta_1\}.$$

However, the inference in this chapter is on much 'bigger' objects than in Chapter 4 since we are now dealing with models rather than parameters. As a consequence of this, and as briefly explained below, the Bayesian solution of model choice is usually hard to justify from a purely Bayesian perspective.

Model choice as an estimation problem

The standard Bayesian solution to model choice consists to extend the prior modelling from parameters to models by considering the index of the model $\mu \in \{1, \ldots, M\}$ as an additional parameter to estimate.

More precisely, let

$$\Theta = \cup_{i=1}^{M} \{i\} \times \Theta_i$$

be the parameter space, $\pi_i(\theta_i)$ be a prior distribution on Θ_i and $(p_1, \ldots, p_M)$ be the prior distribution of μ.

Then, using Bayes theorem, the posterior distribution of μ given the observation x is given by

$$\begin{aligned}\pi(i|x) &= \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\sum_{j=1}^{M} p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j} \\ &= \frac{p_i\, m_i(x)}{\sum_{j=1}^{M}\, p_j m_j(x)}, \quad i = 1, \ldots, M\end{aligned}$$

and we can use the estimator δ^π derived in Chapter 2 to estimate μ.

Typically, the MAP (i.e. the posterior mode) is used so that, in this case, the estimator $\delta^\pi : \mathcal{X} \to \{1, \ldots, M\}$ is defined by

$$\delta^\pi(x) \in \operatorname*{argmax}_{i \in \{1,\ldots,M\}} \{p_i\, m_i(x)\}, \quad x \in \mathcal{X}.$$

Remark: The posterior distribution $\pi(i|x)$ is usually hard to compute (even with advanced Monte Carlo techniques).

Model choice as a testing problem

While the previous approach treats the problem of model choice as an estimation problem, the approach described below treats this problem as a testing problem.

As in Chapter 4, models $\mathcal{M}_1$ and $\mathcal{M}_2$ can be compared using the Bayes factor:

$$B_{12}^{\pi} = \frac{\pi(\{1\}|x)}{\pi(\{2\}|x)} \frac{p_2}{p_1} = \frac{\int_{\Theta_1} f_1(x|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}{\int_{\Theta_2} f_2(x|\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2} = \frac{m_1(x)}{m_2(x)}.$$

Remarks:

- Assume that for any i, j model $\mathcal{M}_i$ is preferred to model $\mathcal{M}_j$ when $B_{ij}^{\pi} > 1$. Then, since

  $$B_{ij}^{\pi} = B_{ik}^{\pi} B_{kj}^{\pi}$$

  the resulting model ordering is transitive.

- The quantity $m_i(x)$ is called the evidence of model i.

- This approach is useful only when M is small because it requires to compute $m_i(x)$ for all $i = 1, \ldots, M$.

- The difficulties with this approach are the same as for hypothesis testing (Chapter 4), namely that improper and vague prior densities should me avoided.

Some comments on Bayesian model choice

The Bayesian solution of model choice is hard to justify from a purely Bayesian perspective.

Indeed,

- In the estimation approach of model choice, there should be some coherence in the choice of $(p_1, \ldots, p_M)$. For instance, if $\mathcal{M}_1 = \mathcal{M}_2 \cup \mathcal{M}_3$ we should have $\max(p_2, p_3) \leq p_1 \leq p_2 + p_3$. The construction of such a prior distribution is therefore complicated when M is large.
- In the testing approach of model choice, the Bayes factor does not depend on the prior probabilities $(p_1, \ldots, p_M)$ but one need to specify the thresholds $\tilde{a}_{ij}$ such that model i is preferred to model j when $B^{\pi}_{ij} > \tilde{a}_{ij}$. As for the prior probabilities $(p_1, \ldots, p_M)$, there should be some coherence in the choice of the $\tilde{a}_{ij}$'s and therefore the construction of these bounds is complicated when M is large.

For these reasons,

1. In practice we usually choose the model $\mu^* \in \operatorname{argmax}_{i \in 1:M} m_i(x)$ (which amounts to set $p_i = \frac{1}{M}$ for all i in the estimation approach and $\tilde{a}_{ij} = 1$ for all i, j in the testing approach).
2. Bayesian model choice is often justified using asymptotic arguments, as explained in the rest of this chapter.

Asymptotic expansion of the evidence

Let $X_1, \ldots, X_n$ be i.i.d. random variables with common density function $\tilde{f}(\cdot|\theta)$,

$$l_n(\theta) = \sum_{i=1}^{N} \log \tilde{f}(X_i|\theta), \quad \theta \in \Theta$$

be the log-likelihood function, $X^{(n)} = (X_1, \ldots, X_n)$ and $\hat{\theta}_n$ be the MLE of θ.

Then, under some regularity conditions (see Chapter 6),

$$\log m(X^{(n)}) = l_n(\hat{\theta}_n) - \frac{d}{2} \log n + \mathcal{O}_{\mathbb{P}}(1)$$

so that, as in the frequentist approach, the criterion used to carry out Bayesian model choice penalizes the number of parameters d.

Recall that the Bayesian information criterion (BIC) is defined by

$$BIC_n = -2l_n(\hat{\theta}_n) + d \log n$$

and therefore

$$\log m(X^{(n)}) = -\frac{BIC_n}{2} + \mathcal{O}_{\mathbb{P}}(1).$$

Consequently, selecting the model $i \in \{1, \ldots, M\}$ having the largest evidence $m_i(x)$ is asymptotically equivalent to choosing the model that minimizes the BIC criterion.

Important remark: Since it is known that (under suitable assumptions), the BIC criterion chooses the ‘true’ model with probability one as the number observations n tends to infinity, the above expansion of $\log m(X^{(n)})$ shows that selecting the model having the highest evidence is an asymptotically ‘correct’ procedure.

An example

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}_1(\theta_0, 1)$ random variables for some $\theta_0 \in \mathbb{R}$ and $f(\cdot|\theta)$ be the p.d.f. of $\mathcal{N}_n(\theta, I_n)$ distribution, with $\theta \in \mathbb{R}$.

We consider the two following two models for $X^{(n)} = (X_1, \ldots, X_n)$

$$\mathcal{M}_1 = \{f(\cdot|0)\}, \quad \mathcal{M}_2 = \{f(\cdot|\theta),\, \theta \in \mathbb{R} \setminus \{0\}\}$$

and we assume that $\pi_1(\theta) = \mathbf{1}_{\{0\}}(\theta)$ while $\pi_2(\theta)$ is the density of the $\mathcal{N}_1(\mu_0, \sigma_0^2)$ distribution.

Then, with $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,

$$\log B_{12}^{\pi}(X^{(n)}) = -\frac{n\bar{X}_n^2}{2} + \frac{n(\bar{X}_n - \mu_0)^2}{2\sigma_0^2(n + 1/\sigma_0^2)} + \frac{1}{2}\log\big(n\sigma_0^2 + 1\big),$$

so that, if $\theta_0 = 0$ and a $n \to +\infty$

$$\log B_{12}^{\pi}(X^{(n)}) \to +\infty \quad \text{(in probability)}$$

at speed $\log n$ while, if $\theta_0 \neq 0$,

$$\lim_{n\to+\infty} \log B_{12}^{\pi}(X^{(n)}) = -\infty \quad \text{(almost surely)}$$

at speed n.

Proof of these results: See Problem Sheet 3, Problem 4.