Applied Bayesian Data Analysis — Chapter 10

Kim Albertsson

CERN and Luleå University of Technology

kim.albertsson@ltu.se

November 19, 2019

Chapter 10

Model comparison and Hierarchical Modelling

Introduction

Occam's razor: Simple models are preferred.

How to measure? Model comparison!

(Can be) cast as a hierarchical modelling problem over a categorical variable.

Note: Model, for model comparison, given by $\theta, p(\mathbf{y}|\theta)$ and $p(\theta)$. Sensitive to choice of $p(\theta)$!

Note: A categorical variable is a discrete variable lacking *order*. I.e. the output can be shuffled without a change in semantics.

General Formula and the Bayes Factor

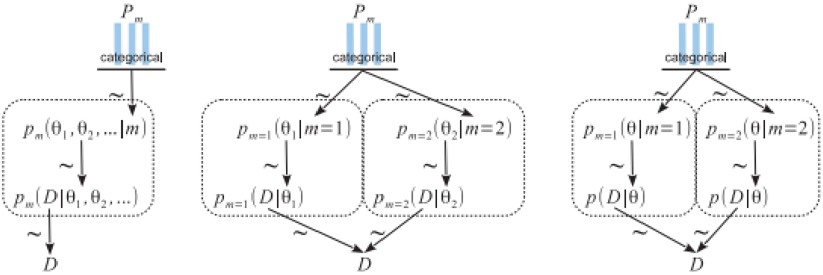
Introduce categorical varibale $m = 1, 2, \dots$ for model.

$$\begin{aligned} \text{Posterior } & p_m(\theta_1, \theta_2, \dots, m|\mathbf{y}) \\ \text{Likelihood } & p_m(\mathbf{y}|\theta_m, m) \\ \text{Prior } & p_m(\theta_m|m) \\ \text{Model (hyper-)prior } & p(m) \end{aligned}$$

Note: Having both subscript and argument above is for mathematical convenience.

$$p_m(\cdot, M = m') = \begin{cases} p_m(\cdot) \text{ if } m = m' \\ 0, \text{ otherwise} \end{cases}$$

Model Comparison as a Single Hierarchical Model



Left Joint model space.

Middle Expanded example of joint space when number of models is two.

Right Special case: Compare only priors.

Posterior (I)

With

$$p_m(\cdot, M = m') = \begin{cases} p_m(\cdot) \text{ if } m = m' \\ 1, \text{ otherwise} \end{cases} ,$$

and by definition

$$\begin{aligned} p(\theta_1, \theta_2, \ldots, 1|\mathbf{y}) &\propto p_1(\mathbf{y}|\theta_1, 1)p(\theta_1, 1) \\ p(\theta_1, \theta_2, \ldots, 2|\mathbf{y}) &\propto p_2(\mathbf{y}|\theta_2, 2)p_2(\theta_2, 2). \end{aligned}$$

Thus the posterior is proportional to (loop over all models)

$$\begin{aligned} p(\theta_1, \theta_2, \ldots, m|\mathbf{y}) &\propto p(\mathbf{y}|\theta_1, \theta_2, \ldots, m)p(\theta_1, \theta_2, \ldots, m) \\ &\propto p(\mathbf{y}|\theta_1, \theta_2, \ldots, m)p(\theta_1, \theta_2, \ldots|m)p(m) \\ &\propto p(m)\prod_{m^*} p_{m^*}(\mathbf{y}|\theta_{m^*}, m)p_{m^*}(\theta_{m^*}|m). \end{aligned}$$

Posterior (II)

Finally

$$p(\theta_1, \theta_2, \ldots, m|\mathbf{y}) = \frac{p(m) \prod_{m^*} p_{m^*}(\mathbf{y}|\theta_{m^*}, m) p_{m^*}(\theta_{m^*}|m)}{\sum_{m'} \int d\theta_{m'} p(m') \prod_{m^*} p_{m^*}(\mathbf{y}|\theta_{m^*}, m') p_{m^*}(\theta_{m^*}|m')} \qquad \text{(10.2*)}$$

Contrast to eqn. in book (handwavy!)

$$p(\theta_1, \theta_2, \ldots, m|\mathbf{y}) = \frac{\prod_m p_m(\mathbf{y}|\theta_m, m) p_m(\theta_m|m) p(m)}{\sum_m \int d\theta_m \; \prod_m p_m(\mathbf{y}|\theta_m, m) p_m(\theta_m|m) p(m)} \qquad \text{(10.2)}$$

Alternative Posterior

Alternative formulation

$$\begin{aligned} p(\theta_1, \theta_2, \dots, m|\mathbf{y}) &\propto p(\mathbf{y}|\theta_1, \theta_2, \dots, m) p(\theta_1, \theta_2, \dots, m) \\ &\propto p(\mathbf{y}|\theta_1, \theta_2, \dots, m) p(\theta_1, \theta_2, \dots |m) p(m) \\ &\propto \sum_{m^*} 1_{m^*}(m) p_{m^*}(\mathbf{y}|\theta_{m^*}) p_{m^*}(\theta_{m^*}) p(m) \\ &\propto p_m(\mathbf{y}|\theta_m) p_m(\theta_m) p(m). \end{aligned}$$

Thus

$$p(\theta_1, \theta_2, \dots, m|\mathbf{y}) = \frac{p_m(\mathbf{y}|\theta_m) p_m(\theta_m) p(m)}{\sum_{m'} \int d\theta_{m'} \; p_{m'}(\mathbf{y}|\theta_{m'}) p_{m'}(\theta_{m'}) p(m')}. \qquad (10.2^*)$$

For me, this is clearer.

Model posterior

Model posterior:

$$p(m|\mathbf{y}) = \frac{p(\mathbf{y}|m)p(m)}{\sum_{m'} p(\mathbf{y}|m')p(m')}, \tag{10.3}$$

where

$$p(\mathbf{y}|m) = \int d\theta_m \, p_m(\mathbf{y}|\theta_m, m) p_m(\theta_m|m). \tag{10.4}$$

Key point: A model is a tuple: `(likelihood, prior)`! (We marginalise across both likelihood and prior to arrive at the model posterior.)

Bayes Factor

Consider the relative posterior probabilities (posterior odds):

$$\frac{p(m_0|\mathbf{y})}{p(m_1|\mathbf{y})} = \underbrace{\frac{p(\mathbf{y}|m_0)}{p(\mathbf{y}|m_1)}}_{\mathrm{BF}} \frac{p(m_0)}{p(m_1)} \underbrace{\frac{/\sum_{m'} p(\mathbf{y}|m')p(m')}{/\sum_{m'} p(\mathbf{y}|m')p(m')}}_{=1} \tag{10.5}$$

The Bayes Factor (BF) is ratio of the probabilities of the data (under the two models).

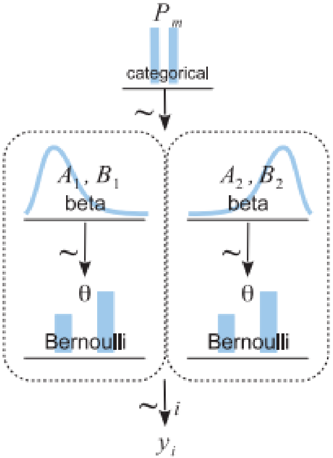
$$\text{posterior odds} = \mathrm{BF} \cdot \text{prior odds}$$

"How much the prior odds change given the data"

Example of two factories of coins

For the following we'll use model of two factories having different expected means of the coin bias.

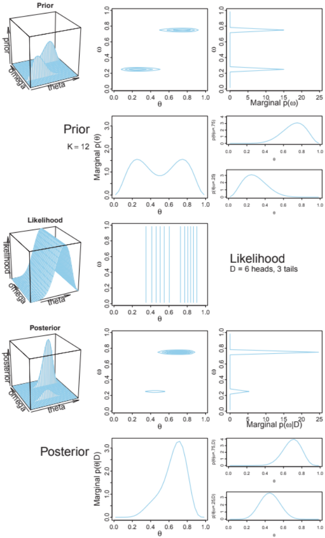
Goal: Find out what factory produced the coin.



10.2.2 Solution by grid approximation

Discretisize joint model (possible b.c. small)

Yields Bayes factor estimate similar to analytical method: $\sim 5:1$.



10.2.1 Solution by formal analysis

We can calculate the Bayes factor with

$$p(\mathbf{y}|m) = p(z, N) = \frac{B(z + a_m, N - z + b_m)}{B(a_m, b_m)} \tag{10.6}$$

(But use logarithm rewrite for numerical stability!)

$$\frac{p(m_0|\mathbf{y})}{p(m_1|\mathbf{y})} = \frac{p(\mathbf{y}|m_0)}{p(\mathbf{y}|m_1)} \frac{p(m_0)}{p(m_1)} = \frac{0.000499}{0.002339} \frac{0.5}{0.5} = 0.213$$

Note: In this analysis, posterior only over m! Use previously developed techniques for estimating θ!

10.3 Solution by MCMC

Idea 1: Estimate $p(\mathbf{y}|m)$ for each model independently and calculate Bayes factor. Use variation of

$$\begin{aligned} p(\mathbf{y}) &= \int d\theta \, p(\mathbf{y}|\theta) p(\theta) \\ &\approx \frac{1}{N} \sum_{\theta_i \sim p(\theta)}^{N} p(\mathbf{y}|\theta_i). \end{aligned}$$

Problematic for complex models.

Idea 2: Cast problem as a hierarchical one, including m. Calculate bayes factor by simple statistic. Can have convergence problems due to autocorrelation.

10.3.1 Non-hierarchical MCM computation of each mdoel's marginal likelihood

Idea 1: Estimate $p(\mathbf{y}|m)$ for each model independently and calculate Bayes factor.

$$p(\mathbf{y}) = \int d\theta\, p(\mathbf{y}|\theta)p(\theta) \approx \frac{1}{N} \sum_{\theta_i \sim p(\theta)}^{N} p(\mathbf{y}|\theta_i).$$

However, usually $p(\mathbf{y}|\theta)$ is nearly zero for many values of θ. Use trick

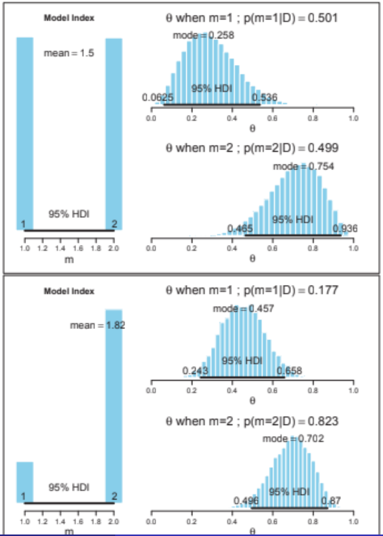
$$\begin{aligned} \frac{1}{p(\mathbf{y})} &= \frac{p(\theta|\mathbf{y})}{p(\mathbf{y}|\theta)p(\theta)} \\ &= \int d\theta\, \frac{h(\theta)}{p(\mathbf{y}|\theta)p(\theta)} p(\theta|\mathbf{y}) \approx \frac{1}{N} \sum_{\theta_i \sim p(\theta)}^{N} \frac{h(\theta)}{p(\mathbf{y}|\theta_i)p(\theta_i)}. \end{aligned}$$

Note: For numerical statbility, choose $h(\theta_i)$ close to posterior, but can be tricky for high-dimensional models.

10.3.2 Hierarchical MCMC computation of relative model probability

$$\frac{p(m_0|\mathbf{y})}{p(m_1|\mathbf{y})} = \frac{0.177}{0.823} = 0.215$$

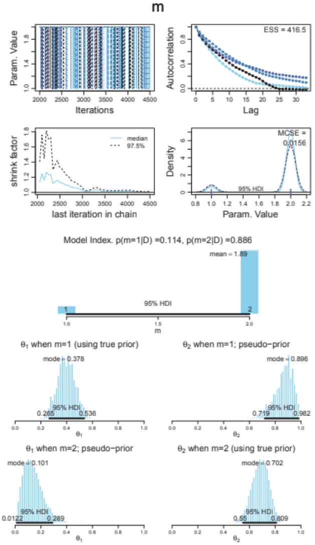
However, problems with autocorrelation and skewed visitation of m.



Poor jumping when not using pseudo-priors

When $M = 1$, θ_1 is sampled from it's posterior, but θ_2 is sampled from its prior.

When sampling m (lockstep) a jump is likely to be rejected since θ_2 does not describe the data well.



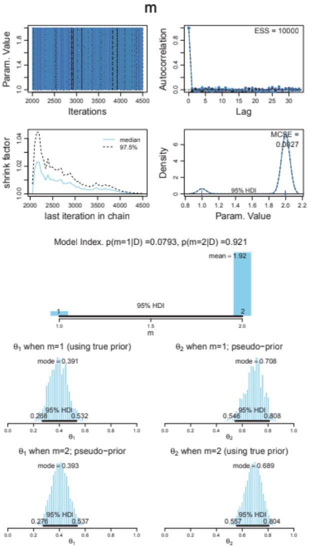
Better jumping with pseudo-priors

Introduce pseudo-priors, which change the sampling distribution when m does not correspond to the model.

If chosen close to posterior, jump is likely. Choose by iterative process.

$$\frac{p(m_0|\mathbf{y})}{p(m_1|\mathbf{y})} = \frac{0.114}{0.886} = 0.129 \text{ (!?)}$$

Q: Why does this not change the final distribution?



10.4 Prediciton: Model Averaging

The full posterior is the best description available.

$$p(\hat{y}|D, M = b) = \int d\theta_b \, p_b(\hat{y}|\theta_b, M = b) p_b(\theta_b|\mathbf{y}, M = b)$$

vs.

$$p(\hat{y}|D) = \sum_m \int d\theta_m \, p_m(\hat{y}|\theta_m, M = m) p_m(\theta_m|\mathbf{y}, M = m) p(M = m)$$

10.5 Model complexity naturally accounted for

Bayesian modeling naturally accounts for model complexity.

Complex models, with large paramter space, $\implies$ probability density spread out across probabilties.

Key: Prior important!

$$\frac{p(m_0|\mathbf{y})}{p(m_1|\mathbf{y})} = \frac{p(\mathbf{y}|m_0)}{p(\mathbf{y}|m_1)} \frac{p(m_0)}{p(m_1)}$$

$$p(\mathbf{y}|m) = \int d\theta_m \, p_m(\mathbf{y}|\theta_m, m) p_m(\theta_m|m). \tag{10.4}$$

10.6 Extreme sensitivity to prior distribution

Vague priors with vastly different density distributions.

$$\text{beta}(1, 1)$$
$$\text{beta}(.1, .1)$$

Marginalise across likelihood and prior.

$$p(\mathbf{y}|m) = \int d\theta_m \, p_m(\mathbf{y}|\theta_m, m) p_m(\theta_m|m). \tag{10.4}$$

10.6.1 Priors of different models should be equally informed

Prior should model our beliefs about distribution under data.

Use parts of the dataset to inform priors. This overwhelms any "vague" prior.

Calculate the posterior for each model individually ($\sim 10\%$ of data) and use that as prior for model comparison.