Chapter 10: Model comparison and Hierachical Modelling

Shubham Gupta

May 14, 2019

1 Introduction

- When we have multiple models describing the same data, we need to assign credibilities to each model.
- Bayesian model comparison reallocates credibility across models given the data.
- Model comparison $\implies$ bayesian estimatation of hierachical models where the top-level is the index of the models.

2 Bayes Factor

2.1 General Formula

- Assume we have data D with parameters θ.
- Prior distribution is $p(\theta)$
- Parameter m to specify the index of the model.
- Hence, we will get
$$likelihood = p_m(y|\theta_m, m).$$
$$prior = p(\theta_m|m).$$
- Priors have different subscripts because they might have different distributions for each model.
- Assume each model is given a prior probability of $p(\theta)$. Then, for all possible models $\theta_1, \theta_2 \dots m$, we have:
$$p(\theta_1, \theta_2 \dots |D) = \frac{P(D|\theta_1, \theta_2 \dots m) * p(\theta_1, \theta_2 \dots m)}{\sum_m \int d\theta_m p(D|\theta_1, \theta_2 \dots m) p(\theta_1, \theta_2 \dots m)}.$$

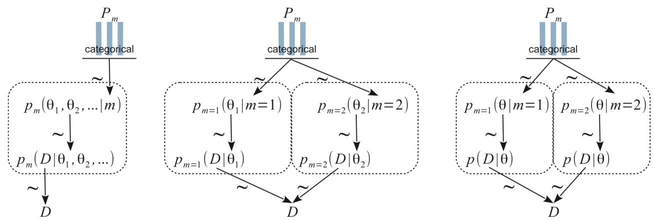

Figure 1: Model comparisons as a hierachical model

- To get the relative probabilities of the models, we will divide their posterior outputs.

$$\frac{p(m=1|D)}{p(m=2|D)} = \frac{p(D|m=1) * p(m=1)/\sum_m P(D|m) * p(m)}{p(D|m=2) * p(m=2)/\sum_m P(D|m) * p(m)}.$$

- The above equation is called the Bayes Factor
- We can use the below table for reference on figuring out when to report a model is better than the alternative model.

K	dHart	bits	Strength of evidence
$< 10^0$	0	—	Negative (supports M_2)
10^0 to $10^{1/2}$	0 to 5	0 to 1.6	Barely worth mentioning
$10^{1/2}$ to 10^1	5 to 10	1.6 to 3.3	Substantial
10^1 to $10^{3/2}$	10 to 15	3.3 to 5.0	Strong
$10^{3/2}$ to 10^2	15 to 20	5.0 to 6.6	Very strong
$> 10^2$	> 20	> 6.6	Decisive

Figure 2: bayes factor

3 Head biased vs tail biased factories

- Two factories that produce head-biased and tail biased factories. Given we have seen some tosses, which factory did the coin come from?
- We have the following hierachy

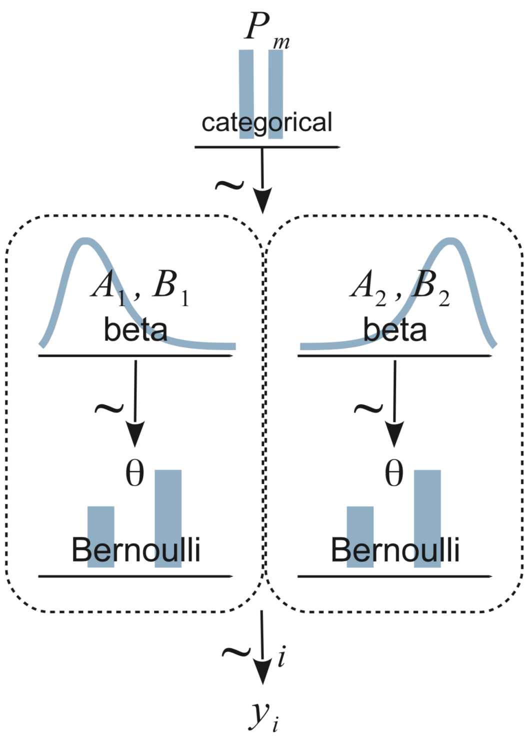

Figure 3: Coin Model Hierachy

3.1 MCMC Method: Individual models

- Main formula to compute the probability of the data is:

$$\frac{1}{p(D)} = \frac{1}{N} \sum_{n=\theta_i \tilde{p}(\theta|D)}^{N} \frac{h(\theta_i)}{p(D|\theta_i)p(\theta_i)}.$$

- $h(\theta_i$ is a probability density function. There is a complex derivation to this formula, which we are skipping for now. Refer to page 275.

3.2 MCMC Method: Hierachical model

- Similar to pymc3 models you have used. Use a index 'm' to indicate for which model are the parameters being specified.
- The chain will be highly corelated with model index.
- Chains will linger on one model for a long time. It will take a lot of iterations for the chains to explore both models equally.
- We can solve this using pseudo-priors.

3.3 Why chains get stuck?

- At a step for which m=1, θ_1 is used to descibe the data. However, θ_2 is not bounded and is sampled randomly from prior. Vice-versa for m=2.
- The prior might be very far away from the posterior. Because the other $\theta_{m=other}$ is a poor description of the data, the chain rarely jumps to it.
- **Solution**: For the parameter currently not being used, make it mimic the posterior. This way, it will always stay in the credible range of values.

3.3.1 Values for pseudopriors

- Do initial run with pseudo prior set to true prior. Note characteristics of marginal distribution of the posterior.
- Set pseudoprior values that mimic the current posterior. Run analysis and do step 1 again. Repeat this analysis if the pseudo-prior values are very different from the previous values you had.

3.4 Model Averaging

- Instead of picking the best model that we got from the posterior analysis, we should instead do a weighted average of all the models.
- This is because our initial hierachical model took into account the posterior distributions from all models, rather than one single model.
- We can obtain this by using the following formula for the weighted averages:

  $$\sum_m p(\hat{y}|D, m)p(m|D).$$

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

  This is called **model averaging**.

3.5 Accounting for model complexity

- Complex models can find relationships between more variable BUT are more sensitive to noise.
- We need a way to measure model complexity, since the noisy data will always prefer more complex models(overfitting).
- Simple models can win if the data is in the same range the prior. This is because in simple models, the prior is restricted to a specific parameter space. In complex models, because the prior is spread over a very large space, the posterior distributions are not as strong as the simple model.

3.6 Comparing nested models

- Suppose there is a model with many parameters that can describe the data very well.
- Now, we define some restrictions on the above parameters(lower bound 0, setting them equal to each other, etc.)
- Such a model can be ”nested” in the original model.
- Here, the Bayesian model comparison will prefer the restricted model, because the full-model has a large prior diluted parameter space.
- **If prior probability of model is 0, then posterior probability is also 0.**

3.7 Sensitive to prior distribution

- Bayesian model comparison is extremely sensitive to the prior distribution.
- Different priors and yield different Bayes factor’s.
- For uninformative priors, prefer **Haldane’s prior**
- For Haldane’s prior, the parameters are very close to 0. It is still a beta distribution.
$$a = b = 0.01.$$