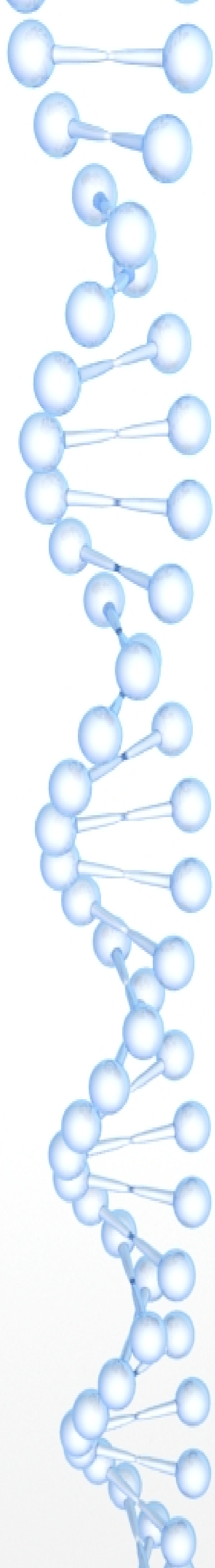
Bayesian Inference 2: Bayesian Model Selection

Blake Pantoja

MCMC Coffee, Episode 7

Material Reference:
Statistics, Data Mining, and Machine Learning in Astronomy
Željko Ivezić, Andrew J. Connolly, Jacob T. VanderPlas & Alexander Gray



From Last Time

The **Bayes' theorem**

Bayes' rule is not controversial

The frequentist vs. Bayesian controversy sets in when we apply Bayes' rule to the likelihood function $p(D|M)$ to obtain Bayes' theorem:

Posterior probability = **Likelihood** × **Prior** / **Evidence**

$$p(M, \boldsymbol{\theta} | D, I) = \frac{p(D | M, \boldsymbol{\theta}, I)\, p(M, \boldsymbol{\theta} | I)}{p(D | I)}$$

D = data
M = model
$\theta = \theta_1, \ldots, \theta_k$ parameters of the model
I = any other information

An "improved belief" is proportional to the product of an "initial belief" and the probability that the "initial belief" generated the observed data.

Odds Ratio

- Before we assumed model to be true. What if we can't?
- We consider the odds ratio between two different models we consider
- $$O_{21} \equiv \frac{p(M_2|D, I)}{p(M_1|D, I)}.$$

Evidence and Posterior Probability

- We can integrate over model parameter space spanned by theta
- $$E(M) \equiv p(D|M, I) = \int p(D|M, \theta, I)\, p(\theta|M, I)\, d\theta$$
- ^ Evidence or global likehood
- *Evidence != data. From physics literature*
- With this and prior probability of model, we can get the posterior probability of model M given data D
- These are probabilities, not PDFs!

$$p(M|D, I) = \frac{p(D|M, I)\, p(M|I)}{p(D|I)},$$

Bayes Factor

- Hardest term to calculate: p(D|I). Drops out though:

$$O_{21} = \frac{E(M_2)\, p(M_2|I)}{E(M_1)\, p(M_1|I)} = B_{21}\, \frac{p(M_2|I)}{p(M_1|I)}.$$

- B21 is the Bayes factor

$$B_{21} = \frac{\int p(D|M_2, \boldsymbol{\theta}_2, I)\, p(\boldsymbol{\theta}_2|M_2, I)\, d\boldsymbol{\theta}_2}{\int p(D|M_1, \boldsymbol{\theta}_1, I)\, p(\boldsymbol{\theta}_1|M_1, I)\, d\boldsymbol{\theta}_1}.$$

- Theta1 and Theta2 can be very different parameter spaces

Interpreting the Odds Ratio

- O21 > 100 : Decisive (100x M2 is more probable than M1)
- O21 > 10 : Strong
- O21 > 3 : Can't say much

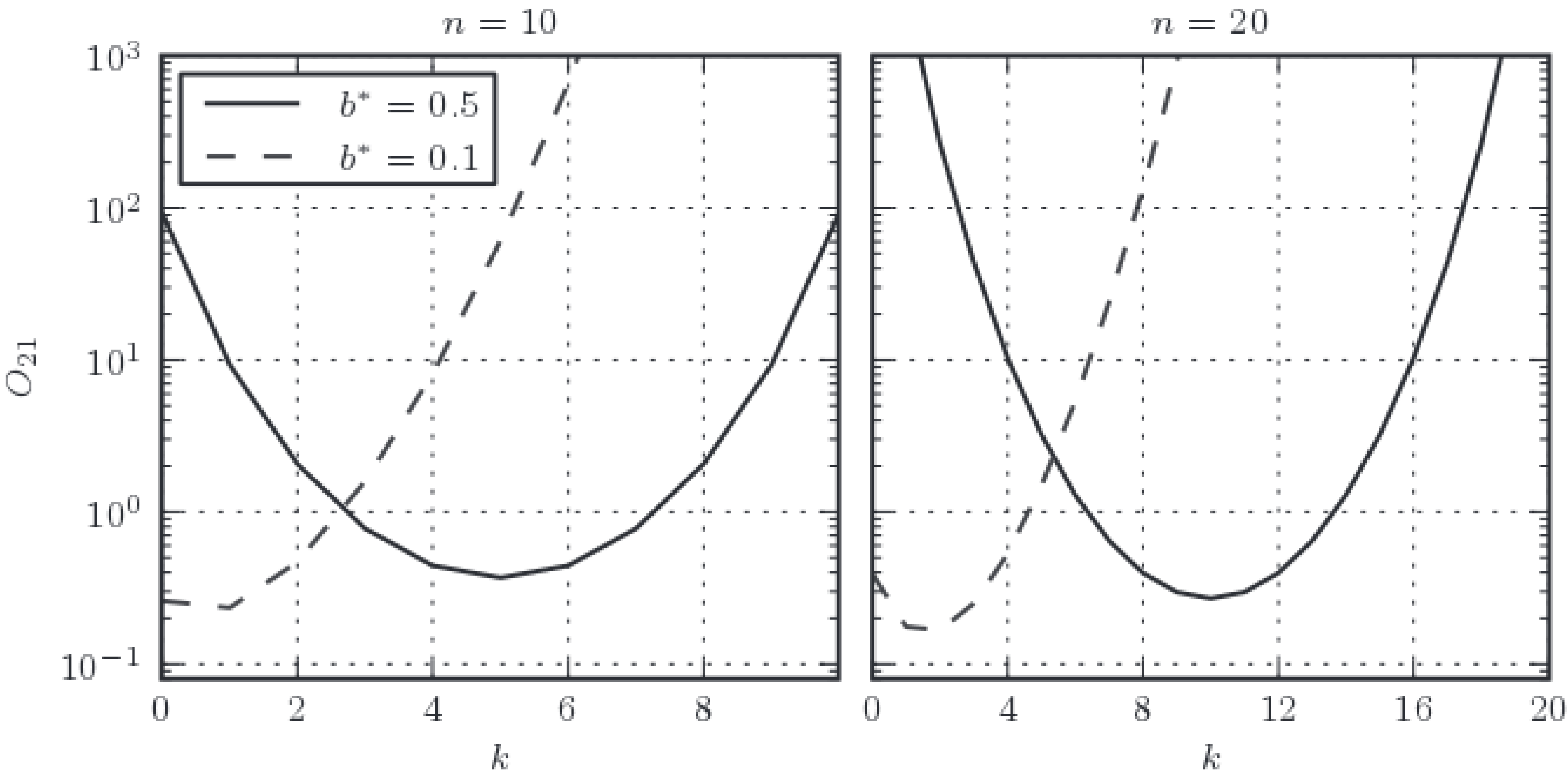
Example: Coin Toss

- M1: known heads probability of b*
- prior is del(b – b*)
- Binomial dist:

$$p(k|b, N) = \frac{N!}{k!(N-k)!} b^k (1-b)^{N-k}$$

$$O_{21} = \int_0^1 \left(\frac{b}{b_*}\right)^k \left(\frac{1-b}{1-b_*}\right)^{N-k} db.$$

- M2: b unknown. Uniform prior in 0-1 range
- O21 = sqrt(pi/(2N)), k = b*N, b=0.5
- N=10000 needed for 1% precision with odds 1/80 for O21



b* = 0.5 (fair) and b* = 0.1.
If k/N = 0.1, O21 is ~9 for N=10,
O21 ~ 263 for N=20
Iif b* = k/N and b* = 0.5 we need N>157 for strong evidence
N=10000 needed for 1% precision with odds 1/80

Hypothesis Testing

- In classical, we reject a null hypothesis based on confidence intervals
- In Bayesian, we cannot reject without alternate hypothesis!
- Let's consider p(M1) + p(M2) = 1. M1 is “null” hypothesis
- Equal priors:
- $$O_{21} = B_{21} = \frac{p(D|M_1)}{p(D|M_2)}.$$
- For coin, let's say that k=16 heads out of N=20 tries
- If sigma=2.24, it is 2.7*sig away, above 0.05 confidence interval leading us to reject.
- In Bayesian, we have alternate hypothesis that there is an unknown probability
- By comparing, we can O21 of 10 in favor of an unfair coin

Occam's Razor

- Choose the simplest model in fair agreement with data
- Let's say prior pdf is flat in range:
- $$-\Delta_\mu/2 < \mu < \Delta_\mu/2, \text{ and thus } p(\mu|I) = 1/\Delta_\mu.$$
- Data is more informative than prior when sigma_mu << Delta for Gaussian data with width sigma_mu
- Global likelihood:
- $$E(M) \approx \sqrt{2\pi}\, L^0(M) \frac{\sigma_\mu}{\Delta_\mu}.$$
- E<<L0 when sigma_mu << Delta_mu

Occam's Razor

- Each parameter has a multiplicative penalty proportional to sigma/Delta, where sigma ~ Delta gives no penality (unconstrained)
- The odds ratio can justify an additional parameter if the penalty is offset a larger maximum likelihood L0 or by prior model probability ratio p(M2|I) / p(M1|I)

Occam's Razor

- For coin flip example from what we had before:
-
-

$$O_{21} = \frac{E(M_2)}{E(M_1)} \approx \sqrt{2\pi}\,\sigma_b \left(\frac{b^0}{b_*}\right)^k \left(\frac{1-b^0}{1-b_*}\right)^{N-k},$$

As M2 has a free parameter, we can see its favor decrease with posterior PDF width (sigma)

- It increases if b0 is far from b*