Bayesian Model Selection (between Signal Models)

15 Oct 2021

Reed Essick

Learning Objectives

How to test for the presence/absence of effects

Hierarchical models for population inference

Selection Effects

~~Scaling of combined results w/ numbers of detections~~

A More In-Depth Discussion of Bayes Factors.

We've introduced the Bayes Factor (evidence ratio) several times now but have not discussed its interpretation in great detail. Consider two alternative hypotheses : A & B.
The Bayes factor between them based on data h is

$$B^A_B = \frac{p(h|A)}{p(h|B)}$$

or the ratio of the probabilities for observing the recorded data assuming each hypothesis. The Bayes Factor therefore prefers hypotheses under which it is more likely to see what we actually saw.

B^A_B also maps a high-dimensional space onto the real line, where it is easier to set thresholds.

When asking whether a value of B is "significant", people often use the following rule of thumb (see Kass & Raftery, <u>Bayes Factors</u>, Journal of American Statistical Association, Vol 90, No. 430 (1995))

$\log_{10} B$	B	$\ln B$	B	
0 - 1/2	1-3.2	0-2	1-3	not worth more than a bare mention
1/2 - 1	3.2-10	2-6	3-20	substantial/positive
1-2	10-100	6-10	20-150	strong
>2	>100	>10	>150	decisive

HOWEVER! Bayes Factors can be tricky as they combine many modeling assumptions into a single number. As such, <u>This table should only be used as a rough guideline.</u> Robust significance estimates rely on actually measuring the distrib. of B in the presence/absence of signals.

A common situation that can be tricky involves <u>NESTED MODELS</u>. Consider two parametrized models where one contains "extra parameters" but is identical to the other if those params are zero that is, models are described by

A : $\vec{\theta} \oplus \vec{\phi}$

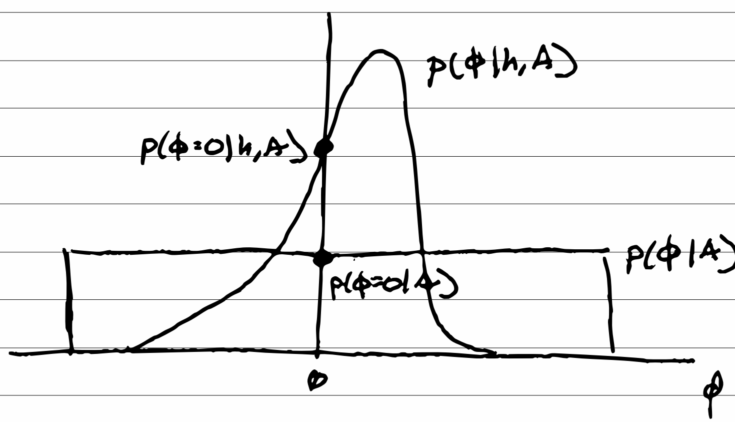
B : $\vec{\theta}$ OR $\vec{\theta} \oplus \vec{\phi} \mid \underline{\underline{\vec{\phi} = 0}}$

In this case, we can use the Savage-Dickey Density Ratio to quickly compute B^A_B.

From last week's notes, we have that

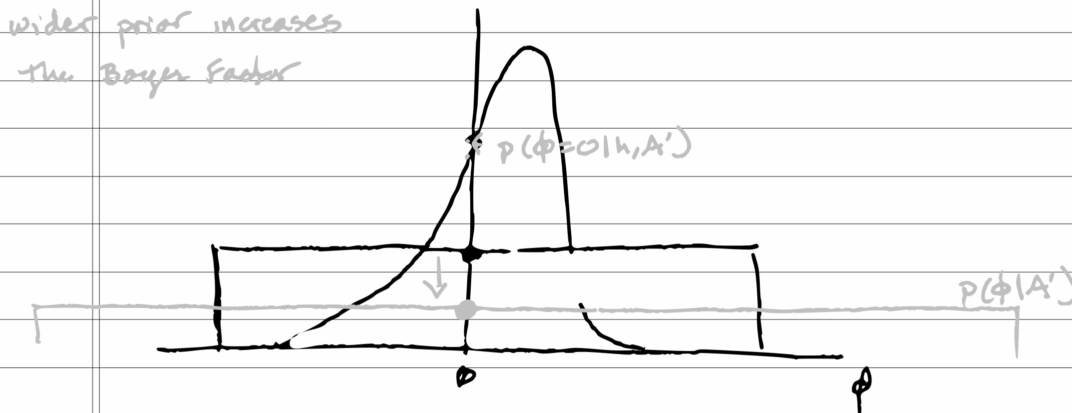
$$B^A_B = \left[\frac{p(\vec{\phi} = 0 \mid h, A)}{p(\vec{\phi} = 0 \mid A)} \right]^{-1}$$

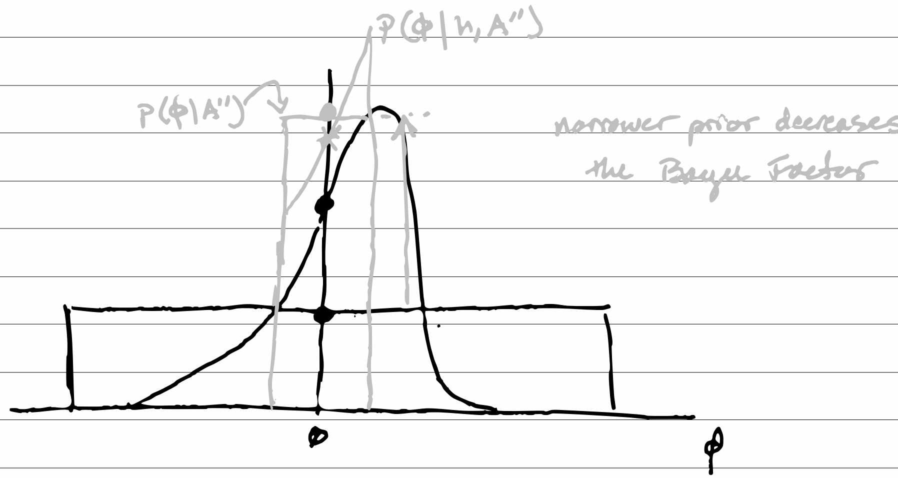
This can be visualized via



If the posterior @ $\phi = 0$ is higher than the prior, then the data will tend to favor hypothesis B.

However, let's consider a few alternative priors for ϕ





These effects are often referred to as <u>Occam Factors</u> and are associated w/ the "prior volume" assumed. If you allow for big deviations, then the fact that you didn't see any is very significant. If you allow for only small deviations, then the fact that you saw small deviations is not significant.

In general, given the fact that the posterior depends on the prior, one may ask if there is a way to use the Bayes factor to infer the "correct" prior. This is often called HIERARCHICAL BAYESIAN INFERENCE.

The basic idea is that we have some observations & we believe the "true parameters" of each observation are iid from the same distribution. We then simultaneously infer the properties of individual events & their population.

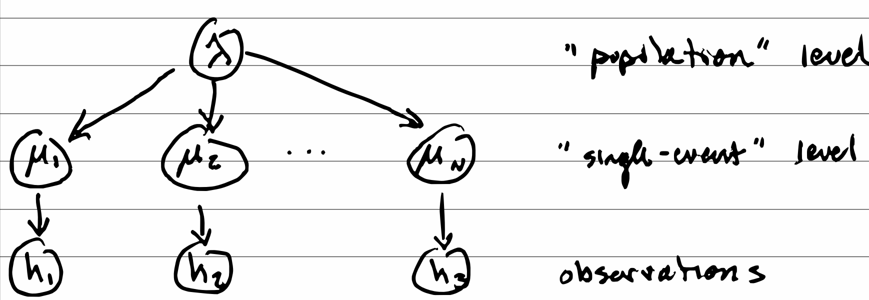
Consider
h_i : data observed for event i
μ_i : parameters of event i
λ : population parameters

$$p(\{\mu_i\}, \lambda \mid \{h_i\}) \propto p(\lambda) \prod_i \left[p(d_i \mid \mu_i)\, p(\mu_i \mid \lambda) \right]$$

prior for pop. params

prior for indiv. params given population

Graphically, we have



Now, if we only care about the population (the "correct prior" for the single-event inference) then we can write

$$\frac{p(\lambda_A \mid \{d_i\})}{p(\lambda_B \mid \{d_i\})} = \frac{p(\lambda_A) \prod_i \left[\int d\mu_i \, p(d_i \mid \mu_i) \, p(\mu_i \mid \lambda_A) \right]}{p(\lambda_B) \prod_i \left[\int d\mu_i \, p(d_i \mid \mu_i) \, p(\mu_i \mid \lambda_B) \right]}$$

$$= \left(\frac{p(\lambda_A)}{p(\lambda_B)} \right) \prod_i \left[B^A_B(d_i) \right]$$

posterior odds

prior odds

product of single-event Bayes factors

Therefore, the hierarchical population analysis is just the systematic comparison of different single-event priors to see which would be most likely to have produced the observed data.

These expressions are all you need if you know that you detect every possible event. However, if your survey only detects/records some fraction of the possible number of events (N_{obs}) &/or the number of events is a random variable, then we must account for this (ie, infer both the <u>rate</u> & the <u>population</u>)

This is often done by assuming an INHOMOGENEOUS POISSON PROCESS:

$$p(R, \lambda, \{\mu_i\} \mid \{h_i\}, N) \propto$$

$$\Big[\underbrace{p(R)}_{\text{prior for rate}} \underbrace{R^N e^{-R\mathcal{E}(\lambda)}}_{\substack{\text{poisson likelihood} \\ \text{for } N \text{ given an} \\ \text{expected number: } R\mathcal{E}(\lambda)}}\Big] \Big[\underbrace{p(\lambda)}_{\substack{\text{prior for} \\ \text{population}}} \prod_i \underbrace{p(h_i|\mu_i)\, p(\mu_i|\lambda)}_{\substack{\text{single-event} \\ \text{posterior given} \\ \text{the population}}}\Big]$$

often referred to as the "selection effects"

$$\mathcal{E}(\lambda) = \int d\mu \underbrace{\Big[\int_{\text{Detectable}} dh\, p(h|\mu)\Big]}_{P(\det|\mu)} p(\mu|\lambda)$$

the prob. that a signal from the pop. would produce <u>detectable data</u>

Often, if we are only interested in λ, we will assume $p(R) \sim 1/R$ and marginalize

$$\underbrace{p(\lambda \mid \{h_i\}, N) \propto p(\lambda) \prod_i \left[\frac{\int d\mu_i\, p(h_i|\mu_i)\, p(\mu_i|\lambda)}{\mathcal{E}(\lambda)}\right]}_{}$$

we can then sample from this distrib. for λ

If we want a posterior over R as well, we note

$$p(R \mid \lambda, N) \propto p(R) (R\mathcal{E})^N e^{-R\mathcal{E}}$$

so we can quickly sample for R for each posterior sample λ

COMPUTATIONAL TECHNIQUES: Importance Sampling

Single-event Likelihoods

Draw single-event posterior samples w/ a fixed prior ("PE") for each event. Then

$$\int d\mu\, p(d|\mu)p(\mu|\lambda) \propto \frac{\int d\mu\, p(d|\mu)p(\mu|\lambda)}{\int d\mu\, p(d|\mu)p(\mu|\lambda_{ref})}$$

$$= \int d\mu \left(\frac{p(d|\mu)p(\mu|\lambda_{ref})}{\int d\mu\, p(d|\mu)p(\mu|\lambda_{ref})}\right)\frac{p(\mu|\lambda)}{p(\mu|\lambda_{ref})}$$

$$\approx \frac{1}{N}\sum_i \left[\frac{p(\mu_i|\lambda)}{p(\mu_i|\lambda_{ref})}\right] \Bigg| \; \mu_i \sim p(\mu|d_i,\lambda_{ref})$$

Generate a large set of simulated signals from a reference population (λ_{inj}) & try to detect them. If we detect N_{det} out of N_{inj} total injections, then

$$\mathcal{E}(\lambda) = \int d\mu\, P(det|\mu)p(\mu|\lambda)$$

$$= \int d\mu\, P(det|\mu)p(\mu|\lambda_{inj})\left[\frac{p(\mu|\lambda)}{p(\mu|\lambda_{inj})}\right]$$

$$\approx \frac{1}{N_{inj}}\sum_i^{N_{det}} \frac{p(\mu_i|\lambda)}{p(\mu_i|\lambda_{inj})}$$