7: Evaluating, comparing and expanding models

10/16/19

Issues with the introduced criteria

- It is difficult to evaluate the differences of the information among all models. The information scales as sample size grows.
- There exists bias in evaluating the predictive performance of the selected model. Selection procedure can strongly overfit the data when comparisons are made for a large number of models.

Bayes Factors

Bayes factors are another way to compare models, two at a time. You compare each model's prior predictive distribution/marginal likelihood/integrated likelihood/evidence:

Bayes Factors

$$
\begin{aligned}
B_{2,1} &= \frac{p(y \mid H_2)}{p(y \mid H_1)} \\
&= \frac{\int p(y \mid \theta_2, H_2) p(\theta_2 \mid H_2) \mathrm{d}\theta_2}{\int p(y \mid \theta_1, H_1) p(\theta_1 \mid H_1) \mathrm{d}\theta_1}
\end{aligned}
$$

assuming $0 < p(y \mid H_i) < \infty$

Models do not have to be nested, and the parameters can be of varying dimension.
Unlike frequentist hypothesis testing, it measures the *strength* of one hypothesis over another.

Bayes Factors

$B_{2,1} = \frac{p(y|H_2)}{p(y|H_1)}$

$\log_{10}(B_{10})$	B_{10}	Evidence against H_0
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
>2	>100	Decisive

From
http://www.andrew.cmu.edu/user/kk3n/simplicity/KassRaftery1995.pdf

Bayes Factors

The reason they call it a Bayes factor is because

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}$$

Bayes Factors

The reason they call it a Bayes factor is because

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}$$

$$\begin{aligned}
\text{posterior odds} &= \frac{p(H_2 \mid y)}{p(H_1 \mid y)} \\
&= \frac{p(y \mid H_2)p(H_2)/p(y)}{p(y \mid H_1)p(H_1)/p(y)} && \text{(Bayes rule)} \\
&= \frac{p(y \mid H_2)}{p(y \mid H_1)} \frac{p(H_2)}{p(H_1)} \\
&= \text{Bayes factor} \times \text{prior odds}
\end{aligned}$$

Bayes Factors

You should not use improper priors when you calculate Bayes factors because

$$p(y \mid H_1) = \int p(y \mid \theta_1, H_1) p(\theta_1 \mid H_1) \mathrm{d}\theta_1$$

is not a density (homework question), and the normalizing constant will be ambiguous.

Bayes Factors

Even noninformative proper priors can be “biased” towards one of the hypotheses.

Consider the following example of the **Jeffreys-Lindley's paradox**:

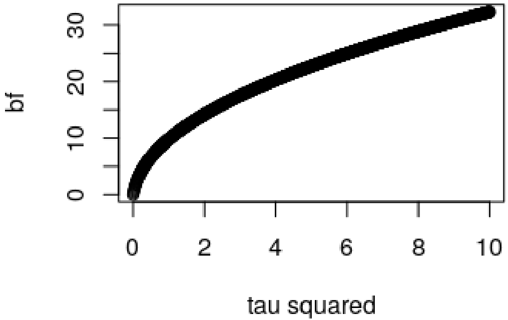
1. under H_1: $\theta = 0$ with prior probability 1
2. $p(\bar{y} \mid H_1) = (2\pi)^{-1/2} n^{1/2} \exp\left[-\frac{n}{2}\bar{y}^2\right]$
3. $p(\theta \mid H_2) = N(0, \tau^2)$
4. $p(\bar{y} \mid H_2) = \int p(\bar{y} \mid \theta, H_2) p(\theta \mid H_2) \mathrm{d}\theta =$
   $[2\pi(\tau^2 + n^{-1})]^{-1/2} \exp\left[-\frac{1}{2(\tau^2+n^{-1})}\bar{y}^2\right]$

so

$$B_{1,2} = (n\tau^2 + 1)^{1/2} \exp\left[-\frac{\bar{y}^2}{2}\left(n - \frac{1}{(\tau^2 + n^{-1})}\right)\right]$$

Bayes Factors: The Jeffreys-Lindley's paradox

Say $\bar{y} = 1.5$ and $n = 10$. Then our p-value for the null is $2.101436e - 06$, but



Different decisions based on whether we are frequentist or Bayesian?!

Bayes Factors

If you can't derive $p(y \mid H_i)$, then it must be approximated. Noticing that the joint $p(y \mid \theta_i, H_i)p(\theta_i \mid H_i)$ is an unnormalized target, here is the justification behind importance sampling:

$$
\begin{aligned}
p(y \mid H_i) &= \int p(y \mid \theta_i, H_i)p(\theta_i \mid H_i)\mathrm{d}\theta_i \\
&= \int \frac{p(y \mid \theta_i, H_i)p(\theta_i \mid H_i)}{q(\theta_i)} q(\theta_i)\mathrm{d}\theta_i \\
&\leftarrow \frac{1}{S}\sum_{s=1}^{S} \frac{p(y \mid \theta_i^s, H_i)p(\theta_i^s \mid H_i)}{q(\theta_i^s)}
\end{aligned}
$$

where $\theta_i^s \sim q(\theta_i)$.

Bayes Factors

Under certain conditions, the **Bayesian Information Criterion** or **Schwarz Information Criterion** approximates the log of integrated likelihood.

$$BIC(H_i) = \log p(y \mid \hat{\theta}, H_i) - k \log(n)$$

where n is the number of data points, and k is the dimension of θ.

You don't even need to specify a prior. However, BIC requires the knowledge of number of parameters, which can be a hard to obtain in complicated models.