Statistics for Astronomers: Lecture 16, 2019.04.25

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: The Jeffreys prior

One example of a non-informative prior that is also invariant over transformation of the random variable (the form of the dependence on the variable doesn't change).

$\pi_J(\theta) \propto \sqrt{\mathcal{I}(\theta)}$, where $\mathcal{I}(\theta)$ is the Fisher information.

In the multidimensional case, replace $\mathcal{I}$ above with the determinant of the Fisher information matrix.

As previously noted, the Fisher information is related to the variance. This form of prior is ideal for scale parameters.

Invariance: Let ψ be some function of θ (*e.g.*, if θ is the probability of a coin flip resulting in a head, then $\psi = \dfrac{\theta}{1-\theta}$, the odds ratio, is a function of θ). We then have

$$\pi_J(\psi) = \pi_J(\theta)\left|\frac{d\theta}{d\psi}\right| \propto \sqrt{\mathcal{I}(\theta)\left(\frac{d\theta}{d\psi}\right)^2} = \sqrt{\mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial\theta}\right)^2\right]\left(\frac{d\theta}{d\psi}\right)^2} = \sqrt{\mathbb{E}\left[\left(\frac{d\theta}{d\psi}\frac{\partial \ln \mathscr{L}}{\partial\theta}\right)^2\right]}$$
$$= \sqrt{\mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial\psi}\right)^2\right]} = \sqrt{\mathcal{I}(\psi)}.$$

The form of the dependence on the parameter is the same, regardless of whether it is θ or ψ.

Recall: Bayesian inference using Jeffreys priors

1. Compute the likelihood, find its logarithm (base e usually easier to deal with, but any base OK in principle).
2. Differentiate the log-likelihood wrt the parameter(s) in question.
3. At this point, decide whether it's easier to compute π_J by squaring the first derivative or by obtaining the second derivative.
4. Compute the expectation value based on your choice in the previous step. Remember that the expectation value is a weighted average over the data, so that any parameters are treated as constants.
5. Once π_J is obtained, multiply it with the likelihood to estimate the posterior distribution.
6. Depending on your application, normalise the posterior.
7. Sanity check: compute prior [if not improper] and posterior means, compare with sample mean. Compare the prior and posterior distributions with the data, compare the CI with the CrI/HPD interval.

Recall: Jeffreys prior for Bernoulli and normal distributions

Bernoulli trial (θ be the probability of "success"):
Likelihood for this problem: $\mathscr{L}(\theta) = \mathrm{Beta}(x+1, 2-x)$.
Jeffreys prior: $\pi_J(\theta) \propto \dfrac{1}{\sqrt{\theta(1-\theta)}} = \mathrm{Beta}\left(\dfrac{1}{2}, \dfrac{1}{2}\right)$.
Posterior: $p(\theta|\mathrm{data}) \propto \mathscr{L}(\theta)\pi_J(\theta) = \mathrm{Beta}(x+1, 2-x) \times \mathrm{Beta}\left(\dfrac{1}{2}, \dfrac{1}{2}\right) = \mathrm{Beta}\left(x+\dfrac{1}{2}, \dfrac{3}{2}-x\right)$.
When, as in this case, the prior and the posterior belong to the same family of distributions, the prior is said to be conjugate to the likelihood.

Univariate normal distribution: priors for μ and σ.
Likelihood: $\mathscr{L}(x|\mu,\sigma) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right]$.
Jeffreys priors: $\pi_J(\mu) \propto \mathrm{constant}$, $\pi_J(\sigma) \propto \dfrac{1}{\sigma}$.
Uniform prior for location parameter (μ),
Logarithmic prior for scale parameter (σ) (because uniform prior for $\ln\sigma$).
$\Longrightarrow$ if large dynamic range in parameter space, use log prior.

Model selection

"Is my data better fit by a Gaussian than a parabola?"
Which model results in a higher likelihood (likelihood ratio)?
Log version: which model gives the lower χ^2?

Bayesian version: compute the ratio of the posterior probabilities – the posterior odds ratio.

"The χ^2 for a cubic polynomial model is much lower than for a linear model!!!!1ONE1!!!!"
But the cubic model has more complexity which must be accounted for.

Occam's razor

Simpler solutions are more likely to be correct than complex ones.

Prefer the simplest solution unless there is sufficient evidence for a more complex one.

The Bayes setup naturally penalises complexity. We can also penalise likelihoods via information criteria such as the BIC or AIC.

Bayes' Theorem (model selection version)

$$\underbrace{p(M|D,I)}_{\text{posterior predictive prob.}} = \frac{\overbrace{p(D|M,I)}^{\text{sampling prob. for } D} \times \overbrace{p(M|I)}^{\text{prior prob.}}}{\underbrace{p(D|I)}_{\text{prior predictive prob.}}} = \frac{\overbrace{\mathscr{L}(M|I)}^{\text{Global likelihood of } M} \times \overbrace{p(M|I)}^{\text{prior prob.}}}{\underbrace{p(D|I)}_{\text{prior predictive prob.}}}$$

"Global likelihood" because $\mathscr{L}(M,I)$ is marginalised over each parameter:

$$\mathscr{L}(M|I) = \prod_{j=1}^{N_{\rm par}} \int_{\theta_j} d\theta_j \; \overbrace{p(\theta_j|M,I)}^{\text{prior prob. for } \theta_j} \; \mathscr{L}(\theta_j|M,I)$$

Model complexity

(based on Section 3.5 in P. Gregory's "Bayesian Logical Data Analysis for the Physical Sciences")

For a single-parameter model,

$$\mathscr{L}(M|I) = \int_\theta d\theta \; \overbrace{p(\theta|M,I)}^{\text{prior prob. for } \theta} \; \mathscr{L}(\theta|M,I) = \mathscr{L}(\hat{\theta}_{\mathrm{MLE}}|M,I)\;\Omega_\theta$$

Where $\hat{\theta}_{\mathrm{MLE}}$ is the value of θ at which the likelihood is maximised (*i.e.*, $\hat{\theta}_{\mathrm{MLE}}$ is the MLE for that likelihood).

Ω_θ (called the Occam Factor or Occam Penalty) ≤ 1.

N parameters: likelihood can be written as a product of N such Ω values, each ≤ 1.
Ω can therefore be thought of as a penalty for model complexity, or a penalty for the fraction of the parameter space ruled out by the likelihood.

Bayesian inference therefore naturally incorporates a quantitative version of Occam's razor, penalising complex models in favour of simpler ones.

Information criteria

Recall the definition of the Occam penalty:

$$\mathscr{L}(M|I) = \int_\theta d\theta \; \overbrace{p(\theta|M,I)}^{\text{prior prob. for } \theta} \; \mathscr{L}(\theta|M,I) = \mathscr{L}(\hat{\theta}_{\mathrm{MLE}}|M,I)\;\Omega_\theta$$

Information criteria are similar penalties combined with the maximum value of the likelihood.
If k is the number of parameters in a model, then

Akaike Information Criterion: $\mathrm{AIC} = 2k - 2\ln \mathscr{L}(\hat{\theta}_{\mathrm{MLE}})$
Bayesian Information Criterion: $\mathrm{BIC} = k\ln N - 2\ln \mathscr{L}(\hat{\theta}_{\mathrm{MLE}})$

By these definitions, the model with the lowest AIC/BIC (note the negative sign for the maximum likelihood) should be preferred.

Model selection using the odds ratio

Which model is better, M_1 or M_2?

The odds ratio, O_{12}, in favour of M_1 over M_2, is the ratio of the posterior probabilities:

$$O_{12} = \frac{p(M_1|D,I)}{p(M_2|D,I)} = \overbrace{\frac{\mathscr{L}(M_1)}{\mathscr{L}(M_2)}}^{\text{Bayes' Factor}} \times \overbrace{\frac{\pi(M_1|I)}{\pi(M_2|I)}}^{\text{prior odds ratio}}$$

Bayes Factor = ratio of global likelihoods.

Jaynes' scale: $O_{12} < 3$: "not worth a mention";
> 10: "strong evidence for M_1";
> 100: "decisive evidence for M_1".

For a given dataset, the odds ratio depends only on the models (effect of data averaged out).

Illustration: coin tosses (from Ivezić/AstroML)

Toss a coin N times. Result: k heads. $M_1: \theta \sim \delta(\theta - \theta_0)$, M_2: $\theta \sim U(0,1)$.

Admission of ignorance: $\pi(M_1|I) = \pi(M_2|I)$.

$$\mathscr{L}(M_1) \propto \int d\theta\ \delta(\theta-\theta_0)\ \theta^k(1-\theta)^{N-k} = \theta_0^k(1-\theta_0)^{N-k}$$

$$\mathscr{L}(M_2) \propto \int_0^1 d\theta\ 1\ \theta^k(1-\theta)^{N-k}$$

$$O_{21} = \frac{\mathscr{L}(M_2)}{\mathscr{L}(M_1)} = \int_0^1 d\theta \left(\frac{\theta}{\theta_0}\right)^k \left(\frac{1-\theta}{1-\theta_0}\right)^{N-k}$$
$$= \frac{\Gamma[N+2]}{\Gamma[k+1]\Gamma[N-k+1]}\theta_0^{-k}(1-\theta_0)^{k-N}.$$

Plot $\ln O_{21}$ as function of k for $N = 20, \theta_0 = 0.5$, $N = 40, \theta_0 = 0.5$, and $N = 40, \theta_0 = 0.2$.

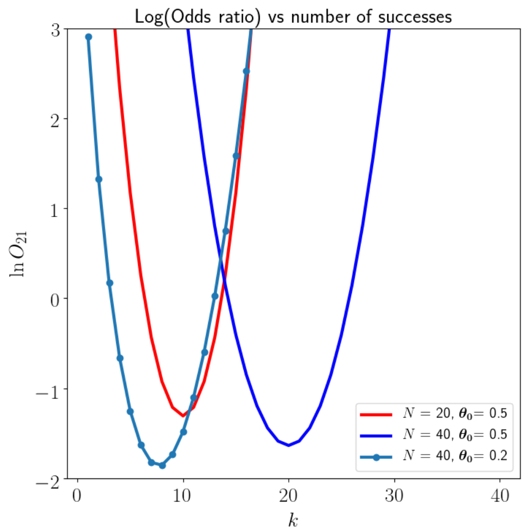
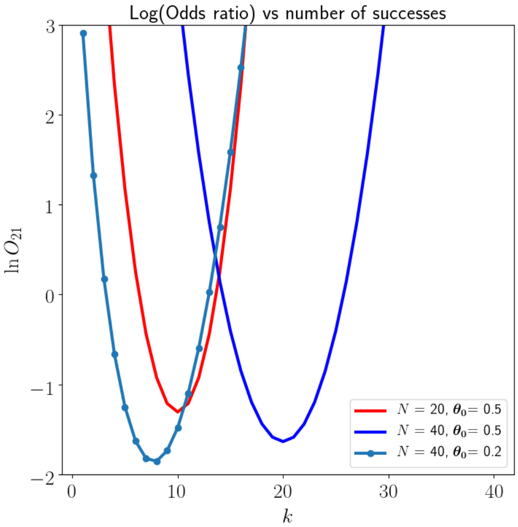


Illustration: coin tosses (contd.)



For $\theta_0 = 0.5$, the maximum value of O_{21} is $\sqrt{\dfrac{\pi}{2N}}$.
For "strong" evidence in favour of M_2 ($O_{21} < 0.1$), $N \geq 160$.

"Decisive" evidence: $N \approx 15,000$.
At this point, relative uncertainty in probability of fairness

$$= \frac{\sqrt{N\theta_0(1-\theta_0)}}{N\theta_0} \approx 0.8\%.$$

In Bayesian hypothesis testing, we use the odds ratio to favour one model instead of another. For the coin-toss example, the null hypothesis may have been that the coin is fair (*i.e.*, the probability of heads is known, and it is 0.5, which was M_1). If we observe for $N = 20$ that $k = 16$, then (see plot) $O_{21} \geq 10$ ("strong" evidence for unfairness).