Statistics for Astronomers: Lecture 15, 2019.04.22

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: Priors

(from Ivezić et al.)

In terms of information, priors can be informative or "non-informative".

Informative prior
Specific information about parameter(s). Progressively increasing amounts of data $\Longrightarrow$ posterior is evidence-dominated.

Example: "Data from the past ten years suggests that there is a 2% change of rain in Morelia today between 2 and 3 PM."

Non-informative prior
Vague information about parameters, typically based on general principles/objective information (also called objective prior). "Light" modification to observations $\Longrightarrow$ posterior is likelihood-dominated.

Example: "The flux from this star is non-negative" $(0 \leq F < \infty)$.
This is also an example of an improper prior, as it does not integrate to unity.
However, we are still OK if the resulting posterior is well-defined
(bus example from last week – $p(\tau|I) \propto 1/\tau, t \leq \tau < \infty$).

The Principle of Indifference is a classic example of an uninformative prior.

Recall: Bayesian point/location and interval estimates

Once $p(\theta|\mathrm{data})$ is computed, we can compute the location estimates (mean, median, mode).

For example, the Bayesian estimator of the parameter mean is $\bar{\theta} = \int d\theta\ \theta\ p(\theta|\mathrm{data})$.

We can also compute Bayesian interval estimates, also called posterior intervals or credible intervals (abbreviated in these lectures as CrI).

One example of a $100(1-\alpha)\%$ CrI is $[a, b]$ such that

$$\int_{-\infty}^{a} d\theta\ p(\theta|\mathrm{data}) = \int_{b}^{\infty} d\theta\ p(\theta|\mathrm{data}) = \alpha/2.$$

Another type of CrI is the highest posterior density (HPD) interval, defined as the narrowest interval that contains $100(1-\alpha)\%$ of the posterior probability.

Recall: Numerical computation of HPD interval

1. Obtain N random deviates $x[i]$ drawn from the posterior density distribution.
2. Sort them in ascending order.
3. For each $x[i]$, find the point that is $w = (1-\alpha)N$ points away.
4. Compute the widths $w[i] = x[w+i] - x[i]$.
5. Find the location $i = i_0$ corresponding to the smallest width. The HPD interval is then $(x[i_0], x[i_0 + w])$.

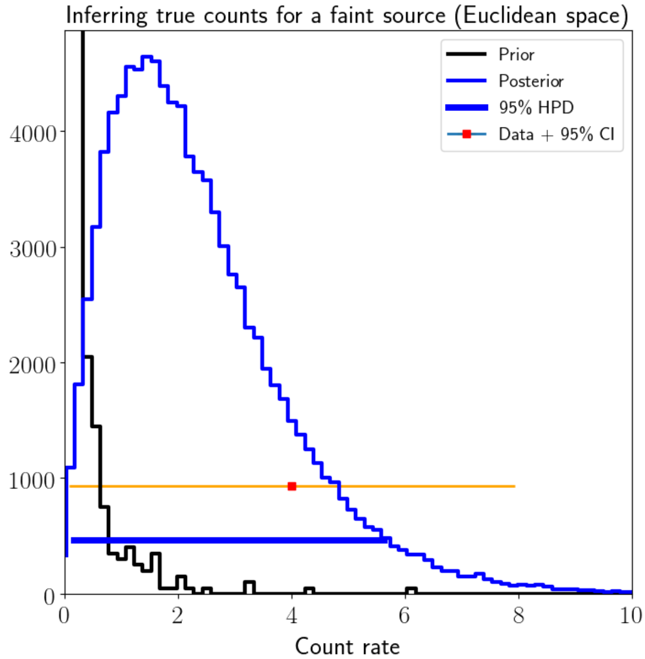
Write your own script! You'll need it for your research if you're using Bayesian inference.

Recall: Prior-dominated posterior

Prior: $p(S) \propto S^{-5/2}$.
Likelihood of obtaining data $S_{\rm obs} = 4$ from Poissonian uncertainties acting on S:
$\mathscr{L}(S) \propto S^{S_{\rm obs}} \exp\Big[-S\Big] = S^4 \exp\Big[-S\Big]$.
Posterior $p(S|S_{\rm obs}) \propto S^{3/2} \exp\Big[-S\Big]$

$$= \mathrm{Gamma}\left(\frac{5}{2}, 1\right).$$

$\Longrightarrow$ Mean: 5/2; Mode: 3/2. Can also compute HPD (homework).



Example of prior-dominated posterior

from Andreon, "Bayesian Methods for the Physical Sciences".

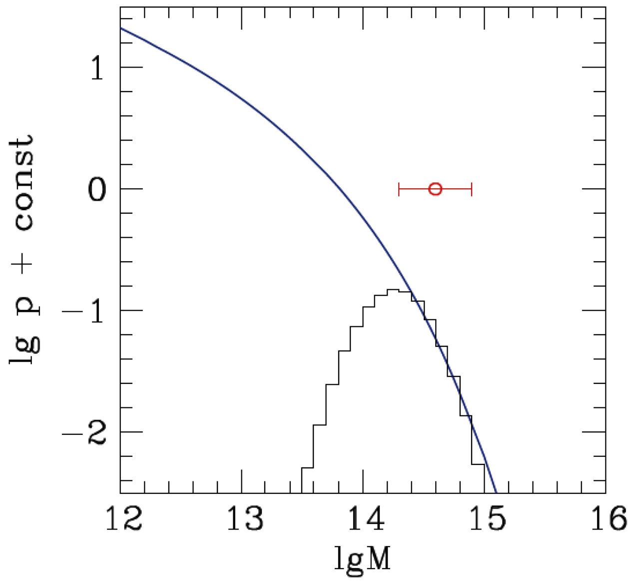
Andreon et al. 2009 mass measurement for most distant ($z \geq 2$) galaxy cluster, JKCS041.

Mass estimate important to constrain parameters of ΛCDM model.

Observation: $\log M/M_\odot = 14.6 \pm 0.3$.
Prior: Schechter mass function.

Prior changes drastically near observed value, similar to previous example.

Posterior mean is therefore lower than observed value: $\log M/M_\odot = 14.3 \pm 0.3$.
(lower by 2x!)



Recall: Maximum likelihood and Fisher information

The variance associated with the MLE estimate is bounded below by the reciprocal of the Fisher information (Cramér-Rao Bound).

The square-root of the reciprocal of the Fisher information is therefore a lower bound to the standard deviation of the MLE estimate.

Given the likelihood $\mathscr{L}(\theta)$, the Fisher information is given by

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial \theta}\right)^2\right] \text{ (under some regularity conditions) } = -\mathbb{E}\left[\frac{\partial^2 \ln \mathscr{L}}{\partial \theta^2}\right]$$

The Fisher information is related to the curvature of the log-likelihood near the MLE value.

The Jeffreys prior

One example of a non-informative priorthat 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.

Jeffreys prior example: a Bernoulli trial

Let θ be the probability of "success" in a Bernoulli trial. We perform one trial and obtain a value $X = x$.

The likelihood associated with this observation is $\mathscr{L}(\theta) \propto \theta^x (1-\theta)^{1-x} = \mathrm{Beta}(x+1, 2-x)$
$\implies \ln \mathscr{L} = x \ln \theta + (1-x) \ln (1-\theta) \implies \dfrac{\partial \ln \mathscr{L}}{\partial \theta} = \dfrac{x}{\theta} - \dfrac{1-x}{1-\theta} = \dfrac{x-\theta}{\theta(1-\theta)}$.

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial \theta}\right)^2\right] = \frac{1}{\theta(1-\theta)}$$

$\implies \pi_J(\theta) \propto \dfrac{1}{\sqrt{\theta(1-\theta)}} = \mathrm{Beta}\Big(\dfrac{1}{2}, \dfrac{1}{2}\Big)$; prior mean: $\dfrac{1/2}{1/2 + 1/2} = 0.5$ as expected.

Posterior: $p(\theta|\mathrm{data}) \propto \mathscr{L}(\theta)\pi_J(\theta) = \mathrm{Beta}(x+1, 2-x) \times \mathrm{Beta}\Big(\dfrac{1}{2}, \dfrac{1}{2}\Big) = \mathrm{Beta}\Big(x + \dfrac{1}{2}, \dfrac{3}{2} - x\Big)$.
Posterior mean: $0.5(x + 0.5) = 0.5$ (sample mean) $+ 0.5$ (prior mean). Effective sample size: 2.

Note that the posterior and prior are both Beta distributions. In such a case, we say that the Beta distribution is the conjugate prior to a Bernoulli likelihood. The Beta distribution is also conjugate to binomial likelihoods.

Jeffreys priors for a univariate normal distribution

$$\mathscr{L}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad \implies \ln \mathscr{L} = -\ln \sigma - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2.$$

$$\frac{\partial \ln \mathscr{L}}{\partial \mu} = \left(\frac{x-\mu}{\sigma^2}\right).$$

$$\mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial \mu}\right)^2\right] = \mathbb{E}\left[\left(\frac{x-\mu}{\sigma^2}\right)^2\right] = \frac{1}{\sigma^2}\mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^2\right] = \frac{1}{\sigma^2}\mathbb{E}[z] = \frac{1}{\sigma^2} \propto \mathrm{constant}.$$

$\implies$ uniform prior for μ.

$$\frac{\partial \ln \mathscr{L}}{\partial \sigma} = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}; \qquad \frac{\partial^2 \ln \mathscr{L}}{\partial \sigma^2} = \frac{1}{\sigma^2} - 3\frac{(x-\mu)^2}{\sigma^4}.$$

$$\mathbb{E}\left[-\frac{\partial^2 \ln \mathscr{L}}{\partial \sigma^2}\right] = \frac{1}{\sigma^4}\mathbb{E}\left[3\left(\frac{x-\mu}{\sigma}\right)^2 - 1\right] \propto \frac{1}{\sigma^2}.$$

Therefore, the prior for σ is $\pi_J(\sigma) \propto \dfrac{1}{\sigma}$
$\implies$ logarithmic prior for σ.

More on priors

See https://bit.ly/2KW92Pt for a good discussion of the applicability of this procedure to problems in fundamental physics.