Statistics for Astronomers: Lecture 17, 2019.05.02

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: 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:

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

Recall: Model selection and Occam's Razor

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. For a single-parameter model,

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

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

Recall: Information criteria and the posterior odds ratio

Information criteria are related to the Occam Penalty. If k is the number of parameters in a model, then

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

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

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".

Multivariate posteriors
(from Andrew Gelman et al., "Bayesian Data Analysis", 3ed.)

In most of the problems you will deal with in research, $\vec{\boldsymbol{\theta}} = (\theta_1, \theta_2, \cdots, \theta_{N_{\rm par}})$ with $N_{\rm par} > 1$.

Definition (Joint, conditional, and marginal posteriors)

$p(\vec{\boldsymbol{\theta}}|\mathrm{data})$ – joint posterior distribution for all the parameters.

$p(\theta_1|\theta_2, \cdots, \theta_{N_{\rm par}}, \mathrm{data})$ – conditional posterior for θ_1 at fixed values of all other components of $\vec{\boldsymbol{\theta}}$ and data.

$p(\theta_1|\mathrm{data})$ – marginal posterior for θ_1, marginalised over all other parameters.

Illustration: normal posterior, joint distribution

For data $\sim \mathcal{N}(\mu, \sigma^2)$, with uniform priors for μ and $\ln \sigma$, the joint posterior distribution is

$$p(\mu, \sigma^2|\mathrm{data}) \propto \sigma^{-(N+2)} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^2\right].$$

Use $\dfrac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \mathrm{Var}(x) + (\mu - \bar{x})^2$:

$$p(\mu, \sigma^2|\mathrm{data}) \propto \sigma^{-(N+2)} \exp\left[-\frac{1}{2}\frac{\mathrm{Var}(x)}{(\sigma/\sqrt{N})^2}\right] \exp\left[-\frac{1}{2}\left(\frac{\mu - \bar{x}}{\sigma/\sqrt{N}}\right)^2\right].$$

(contd.) normal posterior, conditional distributions

$$p(\mu, \sigma^2|\text{data}) \propto \sigma^{-(N+2)} \exp\left[-\frac{1}{2}\frac{\text{Var}(x)}{(\sigma/\sqrt{N})^2}\right] \exp\left[-\frac{1}{2}\left(\frac{\mu-\bar{x}}{\sigma/\sqrt{N}}\right)^2\right]$$

$p(\mu|\sigma^2, \text{data})$ obtained by treating σ as fixed in the above equation:

$$p(\mu|\sigma^2, \text{data}) \propto \exp\left[-\frac{1}{2}\left(\frac{\mu-\bar{x}}{\sigma/\sqrt{N}}\right)^2\right] = \mathcal{N}(\bar{x}, \sigma^2/N).$$

$p(\sigma^2|\mu^2, \text{data})$ obtained by treating μ as fixed instead:

$$p(\sigma^2|\mu^2, \text{data}) \propto (\sigma^2)^{-(N+2)/2} \exp\left[-\frac{1}{2}\frac{Var(x)+(\mu-\bar{x})^2}{(\sigma/\sqrt{N})^2}\right]$$

Defining $y = \dfrac{(\sigma/\sqrt{N})^2}{Var(x)+(\mu-\bar{x})^2}$,

$p(\sigma^2|\mu^2, \text{data}) \propto y^{-(N+2)/2} \exp\left[-\frac{1}{2y}\right]$, which is the Inverse-$\chi^2$ distribution for degree N.

If $z \sim \chi^2(N), z^{-1} \sim \text{Inv-}\chi^2(N)$.

$\implies p(\sigma^2|\mu, \text{data}) = N\Big(Var(x)+(\mu-\bar{x})^2\Big)\ \text{Inv-}\chi^2(N).$

(contd). normal posterior, marginal distribution for μ

$$p(\mu, \sigma^2|\text{data}) \propto \sigma^{-(N+2)} \exp\left[-\frac{1}{2}\frac{\text{Var}(x)}{(\sigma/\sqrt{N})^2}\right] \exp\left[-\frac{1}{2}\left(\frac{\mu-\bar{x}}{\sigma/\sqrt{N}}\right)^2\right]$$

$$p(\mu|\text{data}) \propto \int_0^\infty d\sigma^2 p(\mu, \sigma^2|\text{data}) = \int_0^\infty \frac{d\sigma^2}{\sigma^2}(\sigma^2)^{-N/2} \exp\left[-\frac{1}{2}\frac{Var(x)+(\mu-\bar{x})^2}{(\sigma/\sqrt{N})^2}\right].$$

As before, define $y = \dfrac{(\sigma/\sqrt{N})^2}{Var(x)+(\mu-\bar{x})^2}$:

$$p(\mu|\text{data}) \propto \int_0^\infty \frac{dy}{y}\left[\frac{y}{Var(x)+(\mu-\bar{x})^2}\right]^{N/2} \exp[-y] \propto \left[Var(x)+(\mu-\bar{x})^2\right]^{-N/2}.$$

Recall: $\text{Var}(x) = \dfrac{N-1}{N}s^2$

$\implies p(\mu|\text{data}) \propto \left[1+\frac{1}{N-1}\left(\frac{\mu-\bar{x}}{s/\sqrt{N}}\right)^2\right]^{-N/2} \propto t(N-1)$ (Student's t for $N-1$ dof).

$\implies p(\mu|\text{data}) = \bar{x} + \dfrac{s}{\sqrt{N}}t(N-1).$

(contd). normal posterior, marginal distribution for σ^2

$$p(\mu, \sigma^2|\text{data}) \propto \sigma^{-(N+2)} \exp\left[-\frac{1}{2}\frac{\text{Var}(x)}{(\sigma/\sqrt{N})^2}\right] \exp\left[-\frac{1}{2}\left(\frac{\mu-\bar{x}}{\sigma/\sqrt{N}}\right)^2\right]$$

$$p(\sigma^2|\text{data}) \propto \int_{-\infty}^{\infty} d\mu \; p(\mu, \sigma^2|\text{data})$$

$$= \sigma^{-(N+2)} \exp\left[-\frac{1}{2}\frac{\text{Var}(x)}{(\sigma/\sqrt{N})^2}\right] \int_{-\infty}^{\infty} d\mu \exp\left[-\frac{1}{2}\left(\frac{\mu-\bar{x}}{\sigma/\sqrt{N}}\right)^2\right]$$

$\propto (\sigma^2)^{-(N+1)/2} \exp\left[-\frac{1}{2}\frac{\text{Var}(x)}{(\sigma/\sqrt{N})^2}\right]$; therefore $p(\sigma^2|\text{data}) = N\ \text{Var}(x)\ \text{Inv-}\chi^2(N-1)$.

Summary: if the data is drawn from a normal distribution, with non-informative priors for μ and σ^2, the posterior is such that
For known σ^2, μ is distributed normally about the sample mean, with variance σ^2/N.
For known μ, σ^2 has an Inverse-χ^2 distribution with degree equal to the sample size.
For unknown σ^2, μ has a Student's t distribution around the sample mean.
For unknown μ, σ^2 has an Inverse-χ^2 distribution with degree equal to the sample size minus 1.

For the last two cases, the unknown parameter is a nuisance parameter that has been marginalised over.

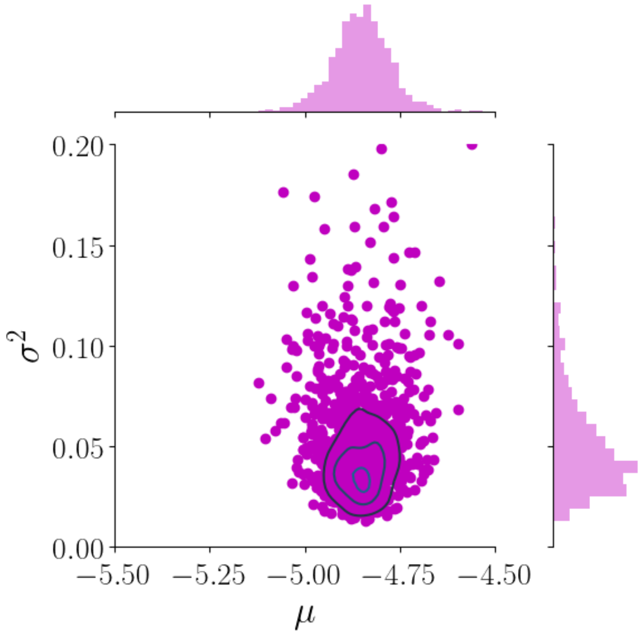
(contd). Sampling and visualising the posterior

To sample the posterior, note that
$p(\mu, \sigma^2|\text{data}) = p(\mu|\sigma^2, \text{data})p(\sigma^2|\text{data}) = p(\sigma^2|\mu, \text{data})p(\mu|\text{data})$.

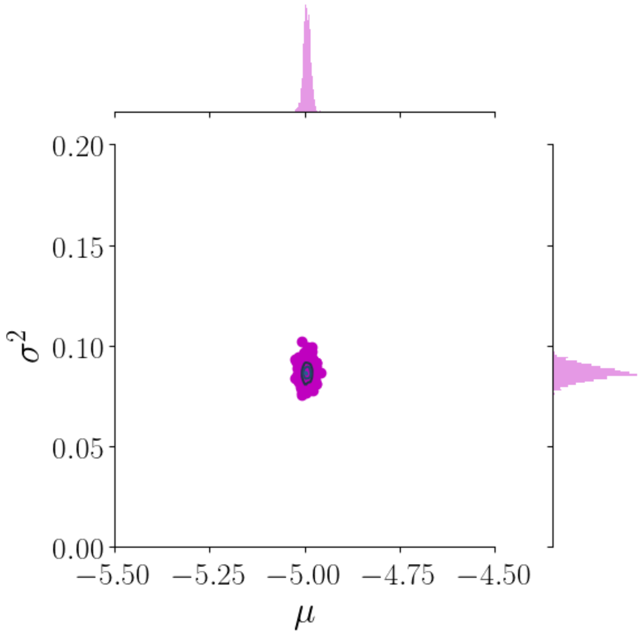
One way: we can first sample σ from the distribution for $p(\sigma^2|\text{data})$, then use those values to sample μ from the distribution for $p(\mu|\sigma^2, \text{data})$.
Other way: μ first then σ^2.

Activity:
1) Generate data: draw $N_{\text{data}} = 10$ deviates from a normal distribution with $\mu = -5.0$ and $\sigma = 0.3$.
2) Draw $N_{\text{samples}} = 1000$ values from the marginal posterior for σ^2, use these to draw the same number of values from the conditional distribution for μ.
3) Plot one histogram each for the distribution of the resulting μ values and σ^2 values (these are the marginalised distributions, since they don't care about the value of the other parameter).

(contd). Visualising the posterior via `seaborn.jointplot`



$N_{\rm data} = 10$, $N_{\rm samples} = 1000$.



$N_{\rm data} = 1000$, $N_{\rm samples} = 1000$.

Posterior predictive distribution

Given a set of observations (data) and the resulting posterior for the model ("data is drawn from a normal distribution"), predict the pdf of future data values.
For the problem discussed in this lecture,

$$p(\text{future data}|\text{data}) = \int\int d\mu\ d\sigma^2\ \underbrace{p(\mu,\sigma^2|\text{data})}_{\text{joint posterior}}\ \overbrace{p(\text{future data}|\mu,\sigma^2,\text{data})}^{\sim \mathcal{N}(\mu,\sigma^2)}$$

To simulate this distribution, first draw μ, σ^2 from their joint pdf then draw new data values from $\mathcal{N}(\mu,\sigma^2)$.

We expect that the new data point be distributed around $\bar{x}$, the mean of the current dataset.

The expected variance is $\sigma^2 + \sigma^2/N = (1+1/N)\sigma^2$.

In fact, the posterior predictive pdf for the new data point is a Student's t distribution with location $\bar{x}$, scale $\sigma\sqrt{1+1/N}$, and degree $N-1$.