Monte Carlo Markov Chain

Likelihoods, Priors, Model comparison

Simone Paradiso

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Likelihoods and priors

Binomial data - Beta prior

Let’s consider a prior distribution and a likelihood:

$$\pi(p; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$$

$$\mathscr{L} = \binom{n}{x} p^x (1-p)^{n-x}$$

Likelihoods and priors

Binomial data - Beta prior

Let’s consider a prior distribution and a likelihood:

$$\pi(p; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \longleftarrow \text{Beta}(a, b)$$

$$\mathscr{L} = \binom{n}{x} p^x (1-p)^{n-x} \longleftarrow \text{Binom}(n, x)$$

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Likelihoods and priors

Binomial data - Beta prior

Let's consider a prior distribution and a likelihood:

$$\pi(p; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \longleftarrow \text{Beta}(a, b)$$

$$\mathscr{L} = \binom{n}{x} p^x (1-p)^{n-x} \longleftarrow \text{Binom}(n, x)$$

If we compute the posterior, it is $\mathscr{L} \cdot \pi$:

$$P(p \,|\, x) = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{x+a-1}(1-p)^{n-x+b-1} \propto \text{Beta}(x+a, n-x+b)$$

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Likelihoods and priors

Binomial data - Beta prior

Let's consider a prior distribution and a likelihood:

$$\pi(p; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \longleftarrow \text{Beta}(a, b)$$

$$\mathscr{L} = \binom{n}{x} p^x (1-p)^{n-x} \longleftarrow \text{Binom}(n, x)$$

If we compute the posterior, it is $\mathscr{L} \cdot \pi$:

$$P(p \,|\, x) = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{x+a-1}(1-p)^{n-x+b-1} \propto \boxed{\text{Beta}(x+a, n-x+b)}$$

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Likelihoods and priors

Normal data - Normal prior

Let’s consider normally distributed Data, with a normal prior on the mean (we keep the variance fixed here):

$$\mathscr{L}(d) = N(\mu, \sigma) \propto e^{-\frac{(\mu - d)^2}{2\sigma^2}}$$

$$\pi(\mu) \propto e^{-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}}$$

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Likelihoods and priors

Normal data - Normal prior

Let's consider normally distributed Data, with a normal prior on the mean (we keep the variance fixed here):

$$\mathscr{L}(d) = N(\mu, \sigma) \propto e^{-\frac{(\mu - d)^2}{2\sigma^2}}$$

$$\pi(\mu) \propto e^{-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}}$$

The posterior distribution is then:

$$P(\mu \,|\, d) = \mathscr{L} \cdot \pi \propto \exp - \left(\frac{(\mu - d)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)$$

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Likelihoods and priors

Normal data - Normal prior

Let's consider normally distributed Data, with a normal prior on the mean (we keep the variance fixed here):

$$\mathscr{L}(d) = N(\mu, \sigma) \propto e^{-\frac{(\mu - d)^2}{2\sigma^2}}$$

$$\pi(\mu) \propto e^{-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}}$$

The posterior distribution is then:

$$P(\mu \,|\, d) = \mathscr{L} \cdot \pi \propto \boxed{\exp - \left(\frac{(\mu - d)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)}$$

Convince yourself this is still a Gaussian…

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Conjugate priors

Definition

We have given a couple of examples for which the couple of likelihood and prior gives a posterior with the same analytical form of the prior.

Conjugate priors

Definition

We have given a couple of examples for which the couple of likelihood and prior gives a posterior with the same analytical form of the prior.

In other words, a family of distributions is said to be **conjugate to the given distribution** if whenever the prior is in the conjugate family, so is the posterior, regardless of the observed value of the data.

Conjugate priors

Definition

We have given a couple of examples for which the couple of likelihood and prior gives a posterior with the same analytical form of the prior.

In other words, a family of distributions is said to be **conjugate to the given distribution** if whenever the prior is in the conjugate family, so is the posterior, regardless of the observed value of the data.

- If the data distribution is binomial, then the conjugate family of distributions is Beta.

Conjugate priors

Definition

We have given a couple of examples for which the couple of likelihood and prior gives a posterior with the same analytical form of the prior.

In other words, a family of distributions is said to be **conjugate to the given distribution** if whenever the prior is in the conjugate family, so is the posterior, regardless of the observed value of the data.

- If the data distribution is binomial, then the conjugate family of distributions is Beta.

- If the data distribution is normal with known variance, the conjugate family of distributions is normal.

The Bayesian guy

"A **subjective Bayesian** is a person who really buys the Bayesian philosophy."

Someone on Internet about Bayesian Statistics.

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Improper priors

Loosening priors

Priors are somewhat important in Bayesian statistics as they contribute to the overall uncertainty of the quantities we are interested in.

It is not uncommon that people don’t really bother about priors in their Bayesian analyses. When the signal in our data is large, the likelihood outweighs the prior and it becomes subdominant…

So PAY ATTENTION TO IT!

Improper priors

Loosening priors

However, we may have enough S/N in our data to allow for less “physics motivated” priors.

In such cases, we could choose priors for **mathematical convenience** rather to accurately express uncertainties.

We would use priors with a very spread distribution, that represent “extreme” uncertainty. Something we might define “vague” or “diffuse’, despite it is not a rigorous mathematical definition…

Those very loose priors are called **improper priors.**

Improper priors

Normal example

Let's consider the normal likelihood with normal prior case:

$$\mathscr{L}(d) = N(\mu, \sigma) \propto e^{-\frac{(\mu - d)^2}{2\sigma^2}} \qquad \pi(\mu) \propto e^{-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}}$$

In the limit of the prior variance $\sigma_0^2 \to \infty$ the prior tends towards a constant unitary value.

In this case, the posterior is therefore just the likelihood:

$$P(\theta \,|\, d) = \mathscr{L}(d)\pi(\theta) = \mathscr{L}(d)$$

Improper priors

Frequentist and Bayesian agreement

Interestingly, the Bayesian with such a prior agrees with the Frequentist:

The maximum likelihood estimatore (MLE) is $\hat{\mu}_n = \bar{x}_n$ and we know the exact sampling distribution of the MLE is:

$$\hat{\mu}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Where we are interested in determining μ, whereas the variance is fixed.

Improper priors

Frequentist and Bayesian agreement

From a frequentist point of view, the confidence interval would computed from:

$$\hat{\mu} - \mu \sim \mathcal{N}(0, \frac{\sigma^2}{n})$$

The Bayesian sees $\hat{\mu}$ as fixed by the data and μ as a random quantity, whereas the Frequentist tells μ is fixed by the model, and $\hat{\mu}$ is random… **but both agree on the distribution of $\hat{\mu} - \mu$ with this improper prior definition!**

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Improper priors

Be very very mindful about this..

We this kind of priors improper because they don't really make sense from a mathematical point of view.

Let's look at the Bayes' rule $P(\theta \mid d) = \mathscr{L}(d)\pi(\theta)$.

Here the prior is a probability distribution, whereas **we just defined a prior whose integral diverges** (can't be normalized, and therefore does not live in the mathematical space of probability distributions).

However the posterior results in something that can be normalized!

Improper priors

Be very very mindful about this..

We are using the Bayes rule's form but not the content; some people say we are using the **formal Bayes rule**.

There is no guarantee that an improper prior leads to a posterior that does integrate. If it is not the case, our result makes no sense.

Improper priors

Summary

Improper priors are very questionable.

- Subjective Bayesians think they are nonsense. They do not correctly describe the uncertainty of anyone.
- Everyone has to be careful using them, because they don’t always yield proper posteriors. Everyone agrees improper posteriors are nonsense.
- Because the joint distribution of data and parameters is also improper, paradoxes arise. These can be puzzling.

However they are widely used and need to be understood.

“Objective” Bayesian Inference

Flat prior

The most obvious “default” prior is flat (constant), which seems to express no preference for any parameter value over any other.

If the parameter space is unbounded, then the prior is improper.

The real issue with flat priors is that they are flat for only one parameterization.

“Objective” Bayesian Inference

Reparameterization

Recall the change-of-variable formulas; in many dimensions:

$$f_Y(y) = f_X(h(y)) \cdot |\det(\nabla h(y))|$$

Where $x = h(y)$ and there is the Jacobian term $\nabla h(y)$.

Let’s use a flat prior, and change variable into $\varphi = \theta^2$:

$$h(\varphi) = \varphi^{1/2} \quad h'(\varphi) = 1/(2\sqrt{\varphi})$$

The prior on φ now reads: $\pi_\Psi(\varphi) = \pi_\Theta(h(\varphi)) \cdot \frac{1}{2}\sqrt{\varphi} = \frac{1}{2\sqrt{\varphi}}$

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

“Objective” Bayesian Inference

Jeffreys prior

Another choice revolves around the Fisher information $I(\theta)$. The Jeffreys prior is therefore defined as:

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

This is objective because **any reparameterization yields the Jeffreys prior for that parameter**. Can you see why?

If the parameter space is unbounded, then the Jeffreys prior is usually improper.

“Objective” Bayesian Inference

Jeffreys prior

In a multivariate case, for a parameter vector $\vec{\theta}$ and Fisher matrix $\mathbf{I}(\vec{\theta})$:

$$\pi(\vec{\theta}) \propto \sqrt{\det(\mathbf{I}(\vec{\theta}))}$$

With the same property of the univariate Jeffreys prior: any reparameterization yields the Jeffreys prior for that parameter.

Bayesian quantities

Point estimates

Point estimates are not really interesting from a Bayesian point of view because in principle we have access to the full parameter posterior distribution, which gives a whole lot of information —much more than a single compressed number.

Bayesian quantities

Point estimates

Point estimates are not really interesting from a Bayesian point of view because in principle we have access to the full parameter posterior distribution, which gives a whole lot of information —much more than a single compressed number.

However, since we are here learning about Bayesian statistics, we might want to learn about some number people refer to when quoting results —even in a Bayesian analysis.

Bayesian quantities

Point estimates

Point estimates are not really interesting from a Bayesian point of view because in principle we have access to the full parameter posterior distribution, which gives a whole lot of information —much more than a single compressed number.

However, since we are here learning about Bayesian statistics, we might want to learn about some number people refer to when quoting results —even in a Bayesian analysis.

We usually compute Mean, Median and Mode (usually referred as best-fit in cosmology) of our parameter posterior distribution.

Bayesian quantities

Point estimates: Best-fit

Let’s focus a second on the best-fit, because has a few connections with the frequentist approach.

Bayesian quantities

Point estimates: Best-fit

Let’s focus a second on the best-fit, because has a few connections with the frequentist approach.

From a frequentist point of view, finding the mode is basically a maximum likelihood estimation by differentiating wrt the parameters. In the Bayesian framework, we differentiate the variable, but the variable is the parameter itself!

Bayesian quantities

Point estimates: Best-fit

Let's focus a second on the best-fit, because has a few connections with the frequentist approach.

From a frequentist point of view, finding the mode is basically a maximum likelihood estimation by differentiating wrt the parameters. In the Bayesian framework, we differentiate the variable, but the variable is the parameter itself!

Long story short it is a maximum likelihood where instead of maximizing $\mathscr{L}(d\,|\,\theta)$ we maximize $\mathscr{L}(d\,|\,\theta)\pi(\theta)$, **the posterior**.

Bayesian quantities

Credible intervals

Guess where the Bayesian gets the information on uncertainty from…

From a Bayesian point of view, the interval estimate is inferred from the posterior distribution.

The Bayesian does not quite like the term “confidence interval”, for that is associated with a frequentist notion and refers to the confidence one has that repeating an experiment many times, one gets the parameter within a certain interval (remember that the parameter is a number in the frequentist approach).

In the Bayesian approach “**credible interval**” sounds much more appropriate, because **it describes the probability of a parameter (that is a PDF) to be within some interval.**

Model selection

Let’s include an additional term

When giving the Bayes rule, we always implicitly assumed a model to our data, and dropped the relative notation. However, we should now make it explicit that this is something that affects the analysis overall:

$$P(\theta \mid d, M) = \frac{P(d \mid \theta, M) P(\theta \mid M)}{\int P(d \mid \theta, M) P(\theta \mid M) d\theta}$$

So, our parameters actually depend (of course) on the model choice. The question is, what if we had different models, and wanted to test data to see which model better describes them?

Model selection

Bayes factor

Let $\mathscr{M}$ be a finite or countable set of models. For each model $m \in \mathscr{M}$ we have a prior probability $\pi(m)$. Each model has a parameter space Θ_m with its prior $\pi(\theta \mid m)$, with $\theta \in \Theta_m$.

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Model selection

Bayes factor

Let $\mathscr{M}$ be a finite or countable set of models. For each model $m \in \mathscr{M}$ we have a prior probability $\pi(m)$. Each model has a parameter space Θ_m with its prior $\pi(\theta \,|\, m)$, with $\theta \in \Theta_m$.

The spaces Θ_m can have different dimension and within model parameters' priors must be normalized.

Model selection

Bayes factor

Let $\mathscr{M}$ be a finite or countable set of models. For each model $m \in \mathscr{M}$ we have a prior probability $\pi(m)$. Each model has a parameter space Θ_m with its prior $\pi(\theta \,|\, m)$, with $\theta \in \Theta_m$.

The spaces Θ_m can have different dimension and within model parameters' priors must be normalized.

The data likelihood then is, for each model:

$$\mathscr{L}(d) = P(d \,|\, m, \theta)$$

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Model selection

Bayes factor

The unnormalized posterior distribution is given by:

$$P(\theta \,|\, d, m) = P(d \,|\, m, \theta) P(\theta \,|\, m) P(m)$$

Model selection

Bayes factor

The unnormalized posterior distribution is given by:

$$P(\theta \,|\, d, m) = P(d \,|\, m, \theta) P(\theta \,|\, m) P(m)$$

To obtain the conditional distribution of the data given the model, we must integrate out the model parameters:

$$P(d \,|\, m) = \int_{\Theta_m} P(d \,|\, m, \theta) P(\theta \,|\, m) P(m) d\theta = P(m) \int_{\Theta_m} P(d \,|\, m, \theta) P(m \,|\, \theta) d\theta$$

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Model selection

Bayes factor

The unnormalized posterior distribution is given by:

$$P(\theta \,|\, d, m) = P(d \,|\, m, \theta) P(\theta \,|\, m) P(m)$$

To obtain the conditional distribution of the data given the model, we must integrate out the model parameters:

$$P(d \,|\, m) = \int_{\Theta_m} P(d \,|\, m, \theta) P(\theta \,|\, m) P(m) d\theta = P(m) \int_{\Theta_m} P(d \,|\, m, \theta) P(m \,|\, \theta) d\theta$$

This gives the unnormalized posterior of the model, the normalized one is:

$$P(m \,|\, d) = \frac{P(d \,|\, m)}{\sum_{i \in \mathscr{M}} P(d \,|\, m_i)}$$

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Model selection

Bayes factor

Let's define the quantity:

$$b(d\,|\,m) = \int_{\Omega_m} P(d\,|\,\theta, m)P(\theta\,|\,m)d\theta$$

So now $P(d\,|\,m) \propto b(d\,|\,m)P(m)$.

Model selection

Bayes factor

Let's define the quantity:

$$b(d\,|\,m) = \int_{\Omega_m} P(d\,|\,\theta, m)P(\theta\,|\,m)d\theta$$

So now $P(d\,|\,m) \propto b(d\,|\,m)P(m)$.

The ratio of the posterior distribution for two different models m_1, m_2 is:

Model selection

Bayes factor

Let's define the quantity:

$$b(d\,|\,m) = \int_{\Omega_m} P(d\,|\,\theta, m)P(\theta\,|\,m)d\theta$$

So now $P(d\,|\,m) \propto b(d\,|\,m)P(m)$.

The ratio of the posterior distribution for two different models m_1, m_2 is:

$$\frac{P(m_1\,|\,d)}{P(m_2\,|\,d)} = \frac{P(d\,|\,m_1)}{P(d\,|\,m_2)} = \frac{b(d\,|\,m_1)P(m_1)}{b(d\,|\,m_2)P(m_2)}$$

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Model selection

Bayes factor

Let's define the quantity:

$$b(d\,|\,m) = \int_{\Omega_m} P(d\,|\,\theta, m) P(\theta\,|\,m) d\theta$$

So now $P(d\,|\,m) \propto b(d\,|\,m) P(m)$.

The ratio of the posterior distribution for two different models m_1, m_2 is:

$$\boxed{\frac{P(m_1\,|\,d)}{P(m_2\,|\,d)}} = \frac{P(d\,|\,m_1)}{P(d\,|\,m_2)} = \frac{b(d\,|\,m_1)P(m_1)}{b(d\,|\,m_2)P(m_2)}$$

Posterior odds

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Model selection

Bayes factor

Let's define the quantity:

$$b(d\,|\,m) = \int_{\Omega_m} P(d\,|\,\theta, m) P(\theta\,|\,m) d\theta$$

So now $P(d\,|\,m) \propto b(d\,|\,m) P(m)$.

The ratio of the posterior distribution for two different models m_1, m_2 is:

$$\frac{P(m_1\,|\,d)}{P(m_2\,|\,d)} = \frac{P(d\,|\,m_1)}{P(d\,|\,m_2)} = \frac{b(d\,|\,m_1)P(m_1)}{b(d\,|\,m_2)P(m_2)}$$

Posterior odds

Prior odds

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Model selection

Bayes factor

Let's define the quantity:

$$b(d\,|\,m) = \int_{\Omega_m} P(d\,|\,\theta, m) P(\theta\,|\,m) d\theta$$

So now $P(d\,|\,m) \propto b(d\,|\,m) P(m)$.

The ratio of the posterior distribution for two different models m_1, m_2 is:

$$\frac{P(m_1\,|\,d)}{P(m_2\,|\,d)} = \frac{P(d\,|\,m_1)}{P(d\,|\,m_2)} = \frac{b(d\,|\,m_1)}{b(d\,|\,m_2)} \frac{P(m_1)}{P(m_2)}$$

Posterior odds | **Bayes factor** | Prior odds

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Model selection

Bayes factor

The prior odds tells how the prior compares the probability of the models. **The Bayes factor tells us how the data shifts that comparison going from prior to posterior via Bayes rule.**

Model selection

Bayes factor

The prior odds tells how the prior compares the probability of the models. **The Bayes factor tells us how the data shifts that comparison going from prior to posterior via Bayes rule.**

The Bayes factor is very handy when it comes to compare two models, because it allows to quantify the preference of the data for one model or the other, without needing to compute the integral.

Model selection

Bayes factor

The prior odds tells how the prior compares the probability of the models. **The Bayes factor tells us how the data shifts that comparison going from prior to posterior via Bayes rule.**

The Bayes factor is very handy when it comes to compare two models, because it allows to quantify the preference of the data for one model or the other, without needing to compute the integral.

Bear in mind that the normalization is the real problem in Bayesian model selection. We need to explore a large parameter space and we need to compute its value in the whole phase space.