37262 Mathematical Statistics

Lecture 11

What Does a “Good” MCMC Simulation Look Like?

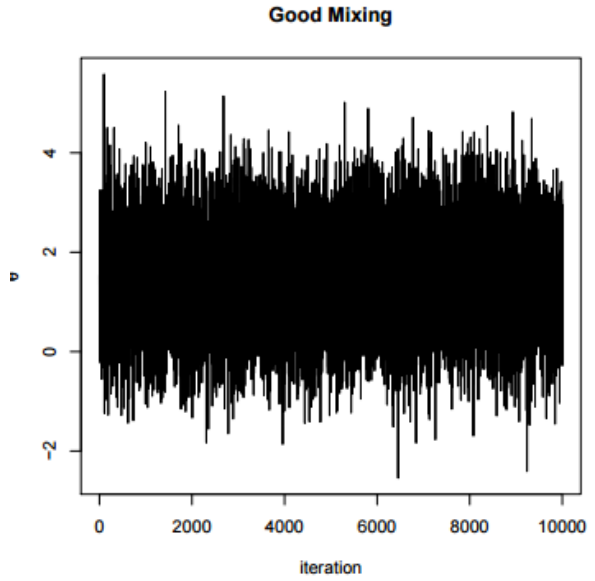
- The Markov Chain Monte Carlo method allows us to generate samples from a distribution $\pi(x)$ by constructing a Markov Chain whose equilibrium distribution is $\pi(x)$ and sampling from its long-run distribution instead.

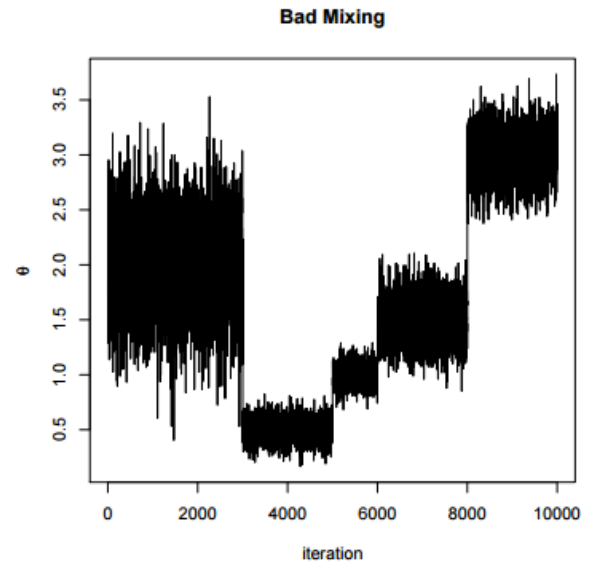
- In a perfect world, we would have a Markov Chain which was aperiodic and ergodic and we would run the chain for a very long time before sampling each individual point, restarting the chain after each observation.

- Clearly this is hugely wasteful in terms of time and computer resources, so we prefer to be able to use all (or almost all) positions of the chain once we are content that it has reached equilibrium (i.e. after a “burn in” period.)

MCMC Diagnostics

- If we want to generate independent samples from an equilibrium, we want to see that, ideally, every state is reachable immediately after every other state.
- A **traceplot** shows which state the chain is in at each iteration.
- We say a chain is **mixing** well if no obvious patterns are observable in the traceplot.
- If the samples from the equilibrium distribution were truly independent, then there would be perfect mixing with all iterations having the same probability of showing a given state, irrespective of what position the chain was in at the previous iteration.

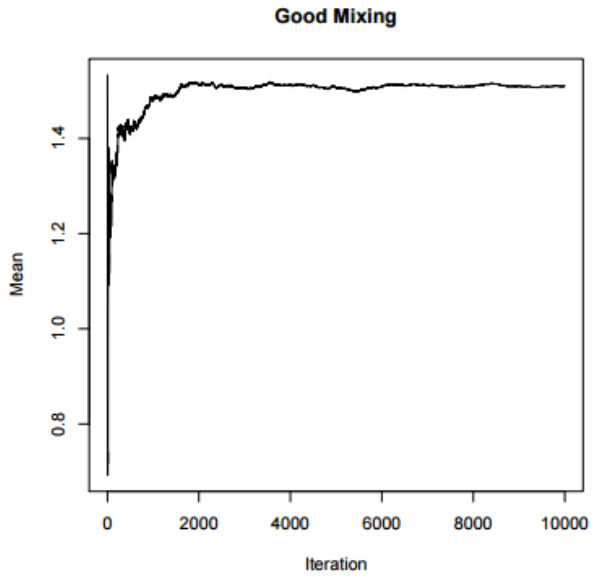


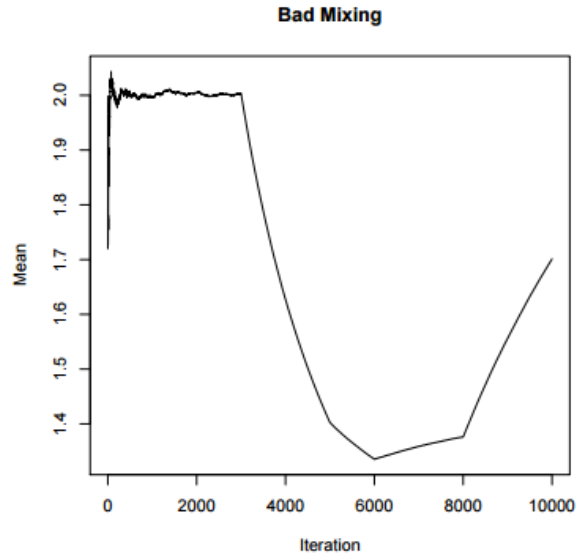


MCMC Diagnostics

- A **running mean** plot can also be used to identify obvious flaws in the mixing.

- Plotting the mean of all iterations up to and including the current step should show no strong signal once the burn in period has passed.

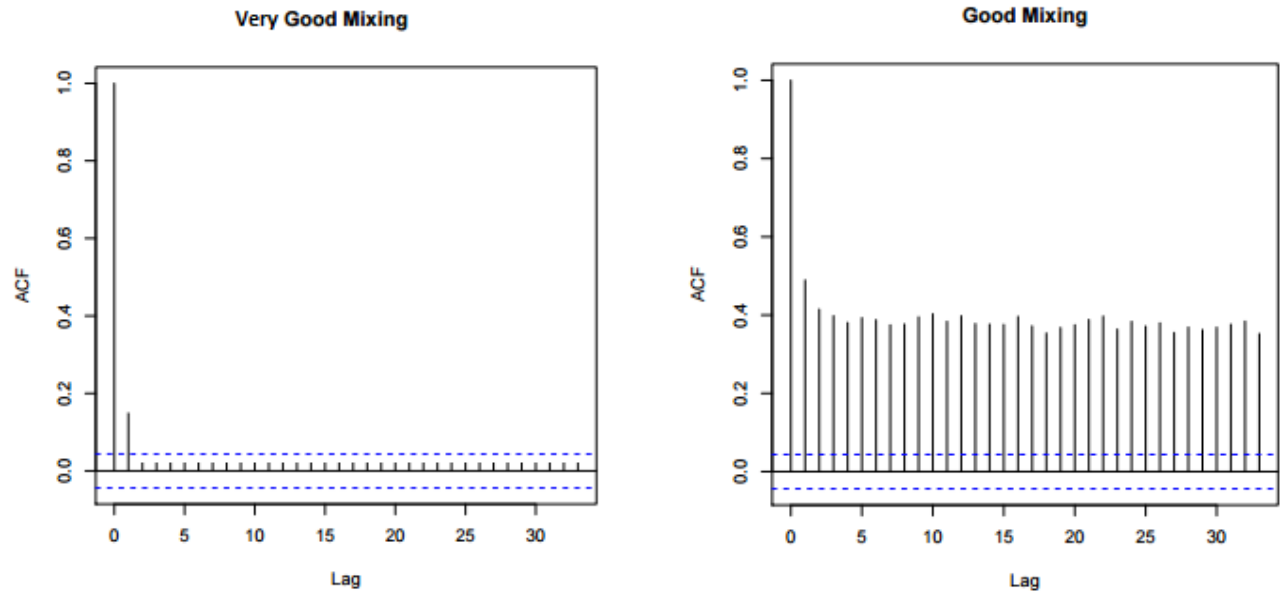




MCMC Diagnostics

- Problems with MCMC mixing can also show up in an **autocorrelation** plot.

- The *k*th autocorrelation for a sequence with mean and variance μ, σ^2 respectively is $\rho_k = \frac{E\left[(X_n - \mu)(X_{n+k} - \mu)\right]}{\sigma^2}$.



MCMC and Bayes' Theorem

- Recall the Metropolis-Hastings algorithm. Given proposal distribution $g(x)$, a realisation is generated from $\pi(x)$ by accepting the move proposed by $g(x)$, x_p, with probability

$$A(x_p|x) = \min\left\{1, \frac{\pi(x_p)g(x|x_p)}{\pi(x)g(x_p|x)}\right\}.$$

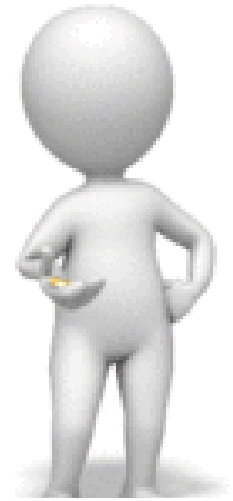
- One of the main strengths of MCMC is that we do not need to be able to calculate $\pi(x)$ as the acceptance probability only relies on the ratio $\frac{\pi(x_p)}{\pi(x)}$, which is often easier.

- As we will see, this allows us to sidestep one of the most difficult parts of Bayesian inference.

- Especially when we are dealing with multivariate probabilities, this property is extremely useful.

Bayesian Inference

- Bayesian inference uses Bayes' Theorem to update the belief around a parameter as more evidence is gathered.



- Consider flipping a single coin repeatedly and recording its outcome. We do not know what the probability θ that the coin will lands Heads is, but rather will have an initial belief which we update each time more evidence is observed. Call the outcome of the *i*th flip x_i.

- We have a **prior distribution** $\pi(\theta)$ which describes our belief in the parameter before any observations have been gathered.

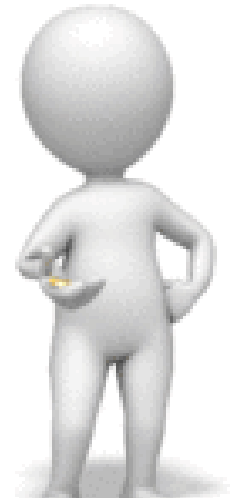
- After observations $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$ have been gathered, we update our belief of the value of θ. (Assume $x_i = 1$ if the coin lands Heads and 0 if it lands Tails.)

Bayesian Inference

- We call our initial belief the prior distribution $\pi(\theta)$.
- From this, we calculate the **likelihood** of seeing the evidence which we did, assuming θ, $f(\boldsymbol{X}|\theta)$
- Bayes' Theorem then gives $f(\theta|\boldsymbol{X}) = \dfrac{f(\boldsymbol{X}|\theta)\pi(\theta)}{f(\boldsymbol{X})} = \dfrac{f(\boldsymbol{X}|\theta)\pi(\theta)}{\int_{\Theta} f(\boldsymbol{X},\theta)d\theta}$ where $f(\theta|\boldsymbol{X})$ is the **posterior distribution**.
- The posterior density is equal to the prior multiplied by the likelihood, divided by the integral of this over all possible values of θ.
- Without knowing the denominator, this is sometimes said as "posterior $\propto$ prior $\times$ likelihood."

Bayesian Inference

- Consider the problem of describing our uncertainty about the probability θ that a coin will land Heads when flipped.
- Initially, the experimenter believes the coin is fair, or close to fair and will update this belief based on future observations.
- The prior belief on the probability is described by $\Theta \sim N(0.5, \sigma^2)$ where σ^2 is some known constant.
- (This is not a very good prior as does put some belief onto probabilities of below 0 or above 1, which is clearly not sensible.)



Bayesian Inference

- Bayes' Theorem gives $f(\theta|\boldsymbol{X}) = \dfrac{f(\boldsymbol{X}|\theta)f(\theta)}{f(\boldsymbol{X})} = \dfrac{f(\boldsymbol{X}|\theta)f(\theta)}{\int_{\Theta} f(\boldsymbol{X},\theta)d\theta}$
- The probability that, say, $\boldsymbol{X} = (x_1, x_2, ..., x_n) = (0,1,1,0,0)$ for a known θ is $f(\boldsymbol{X}|\theta) = (1-\theta)\theta\theta(1-\theta)(1-\theta) = (1-\theta)^3\theta^2$.
- The numerator is trivial to calculate. If $\Theta \sim N(0.5, \sigma^2)$, then $f(\theta) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta-0.5)^2}{2\sigma^2}}$.
- The denominator is, however, much more problematic $f(\boldsymbol{X}) = \int_{-\infty}^{\infty} (1-\theta)^3\theta^2 \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta-0.5)^2}{2\sigma^2}} d\theta$.

Bayesian Inference

- If, however, we wished to know what the posterior distribution of the parameter θ was, we could sample from this easily by MCMC.

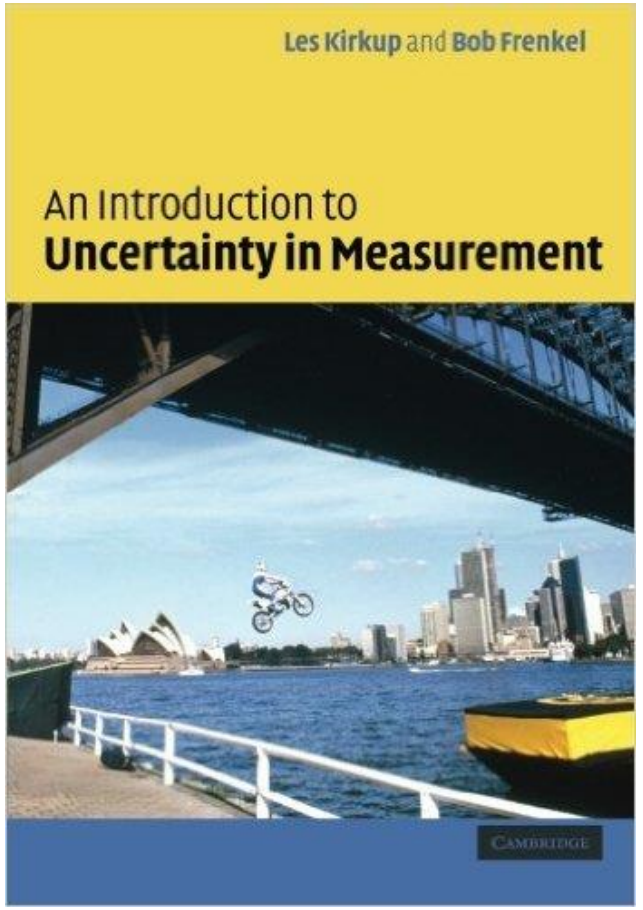
- We only need to know the ratio $\frac{\pi(\theta_p)}{\pi(\theta)}$ and this is equal to

$$\frac{(1-\theta_p)^3\theta_p{}^2\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(\theta_p-0.5)^2}{2\sigma^2}}}{(1-\theta)^3\theta^2\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(\theta-0.5)^2}{2\sigma^2}}} = \frac{(1-\theta_p)^3\theta_p{}^2e^{-\frac{(\theta_p-0.5)^2}{2\sigma^2}}}{(1-\theta)^3\theta^2e^{-\frac{(\theta-0.5)^2}{2\sigma^2}}}$$

- As more and more evidence becomes available, for a well-defined Bayesian framework, the posterior distribution should converge ever more closely to the information provided by the observations.

Bayesian Inference

- As models (and resulting probability density functions or probability mass functions) increase in complexity, the advantages of using MCMC become ever greater.
- Consider drawing samples from $X \sim N(\mu,\sigma^2)$ where neither μ nor σ^2 are definitely known.
- We might, though, know what these parameters should be, plus or minus some uncertainties.
- For example, we might say $\mu \sim N(10,0.5)$ and $\sigma^2 \sim N(0.1,0.0001)$.
- How would we update these beliefs as we observe more and more realisations of X?



Bayesian Inference

- Given observations of X, $\boldsymbol{X} = (x_1, x_2, ..., x_n)$, our posterior belief of the values of μ and σ^2 is

$$f(\mu,\sigma^2|\boldsymbol{X}) = \frac{f(\boldsymbol{X}|\mu,\sigma^2)f(\mu,\sigma^2)}{f(\boldsymbol{X})} = \frac{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi(0.5)}}e^{-\frac{(\mu-10)^2}{2(0.5)}}\frac{1}{\sqrt{2\pi(0.01)}}e^{-\frac{(\sigma^2-0.1)^2}{2(0.0001)}}}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi(0.5)}}e^{-\frac{(\mu-10)^2}{2(0.5)}}\frac{1}{\sqrt{2\pi(0.01)}}e^{-\frac{(\sigma^2-0.1)^2}{2(0.0001)}}d\mu d\sigma^2}$$

- The denominator is hardly trivial to calculate…

- Using Metropolis-Hastings, however, we would only ever need to calculate the numerator, which requires no integration and into which proposed values for μ and σ^2 could simply be plugged.

Bayesian Inference: Conjugate Priors

- In some cases, we can choose a prior which gives a simple form for the posterior distribution which can sidestep the need for MCMC (or similar) methods.

- Consider again our flipping of a coin which lands Heads with probability θ for some unknown θ.

- Given θ , the likelihood of a given sequence of outcomes $\boldsymbol{X} = (x_1, x_2, ..., x_n)$ is $f(\boldsymbol{X} \,|\, \theta) = \theta^k (1-\theta)^{n-k}$ where k is the number of times Heads was observed.

- If we do not know θ , but have no reason to believe (before any observations) that it is not close to 0.5, we may choose a prior distribution which has expectation 0.5.

Bayesian Inference: Conjugate Priors

- Recall the beta distribution, with density function $f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$ for $\alpha, \beta > 0$ which takes values on [0,1].

(Note $\Gamma(z)$ is the gamma function $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$. For integer values of z, $\Gamma(z) = (z-1)!$ and for all z, $\Gamma(z+1) = z\Gamma(z)$)

- The expectation of this is easy to calculate

$$E(Y) = \int_0^1 y \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}\, dy = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 y^{\alpha}(1-y)^{\beta-1}\, dy = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}$$

$$= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)} = \frac{\alpha}{\alpha+\beta}$$

Bayesian Inference: Conjugate Priors

- When examining Bernoulli trials (and hence Binomial distributions), it is often desirable to use a beta prior on the probability parameter.

- Since the likelihood is $f(\boldsymbol{X}|\theta) = \theta^k(1-\theta)^{n-k}$, then setting a prior of $f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

  gives a posterior proportional to $\theta^k(1-\theta)^{n-k} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$

- This therefore gives a posterior which is also beta distributed, $\theta|\boldsymbol{X} \sim Beta(\alpha+k, \beta+n-k)$.

- This has expectation $\frac{\alpha+k}{\alpha+\beta+n}$.

- If we set a large α and β , this is $\approx \frac{\alpha}{\alpha+\beta}$, however as n and k get bigger, it is $\approx \frac{k}{n}$.

Bayesian Inference: Conjugate Priors

- We say that a distribution is a **conjugate prior** for a given likelihood function if it results in a posterior distribution which is of the same form (usually with updated parameters.)

- We have already seen that if $X \sim Bern(\theta)$ then $\theta \sim Beta(\alpha, \beta)$ is a conjugate prior since we get a posterior distribution $\theta | \boldsymbol{X} \sim Beta(\alpha + k, \beta + n - k)$ (where $\boldsymbol{X}$ contains k 1s and $n - k$ 0s.)

- Other conjugate priors include, for $X \sim N(\mu, \sigma^2)$ where σ^2 is known but μ is uncertain, $\cdot \mu \sim N(m, s^2)$.

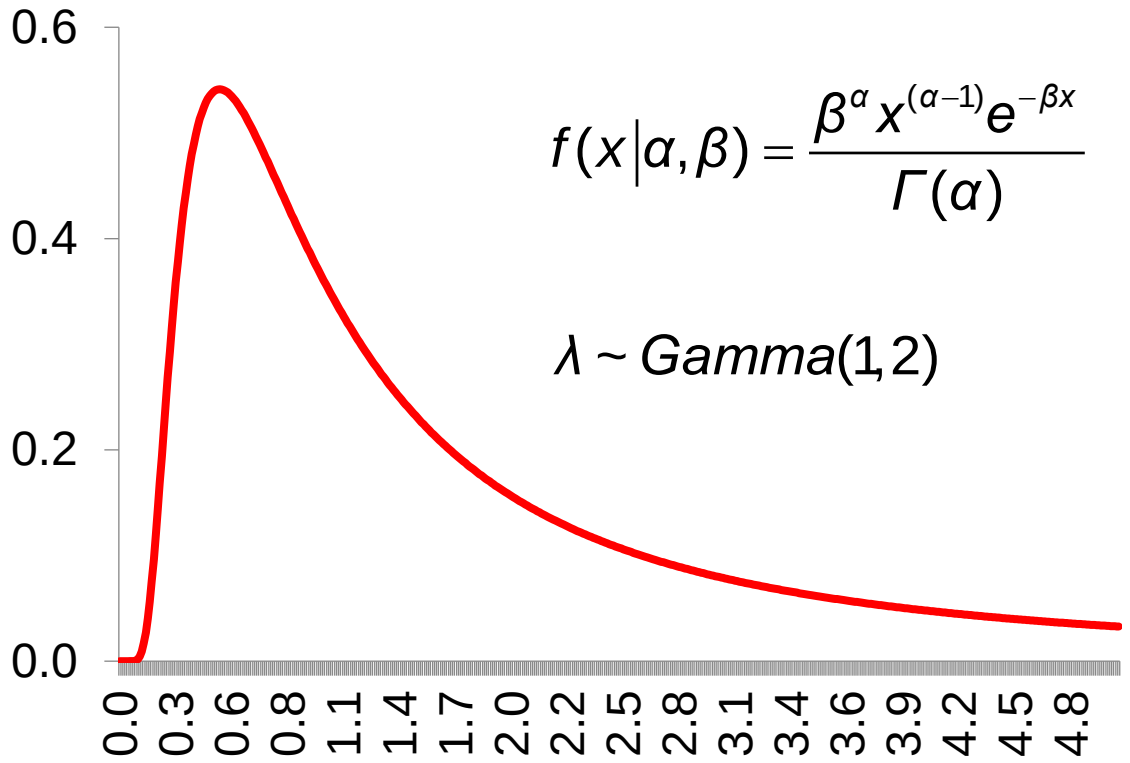
- Although not simple to derive, this gives a posterior of

$$\mu \left| \boldsymbol{X} \sim N\left(\left(\left(\frac{m}{s^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i \right) \left(\frac{1}{\sigma^2} + \frac{n}{s^2} \right)^{-1} \right), \left(\frac{1}{\sigma^2} + \frac{n}{s^2} \right)^{-1} \right) \right.$$

Conjugate Priors: Example

- Suppose that it is known that faults arise in a system at the instants of a Poisson process where the rate of this process is not known.

- This means that the time between successive arrivals are independent realisations of an $\exp(\lambda)$ variable for some unknown $\lambda > 0$.



- Initially, the prior belief is that $\lambda \sim Gamma(1,2)$ i.e. it is thought that λ is probably around 0.5, but that there is some uncertainty around this value. As observations become available, this belief is updated.

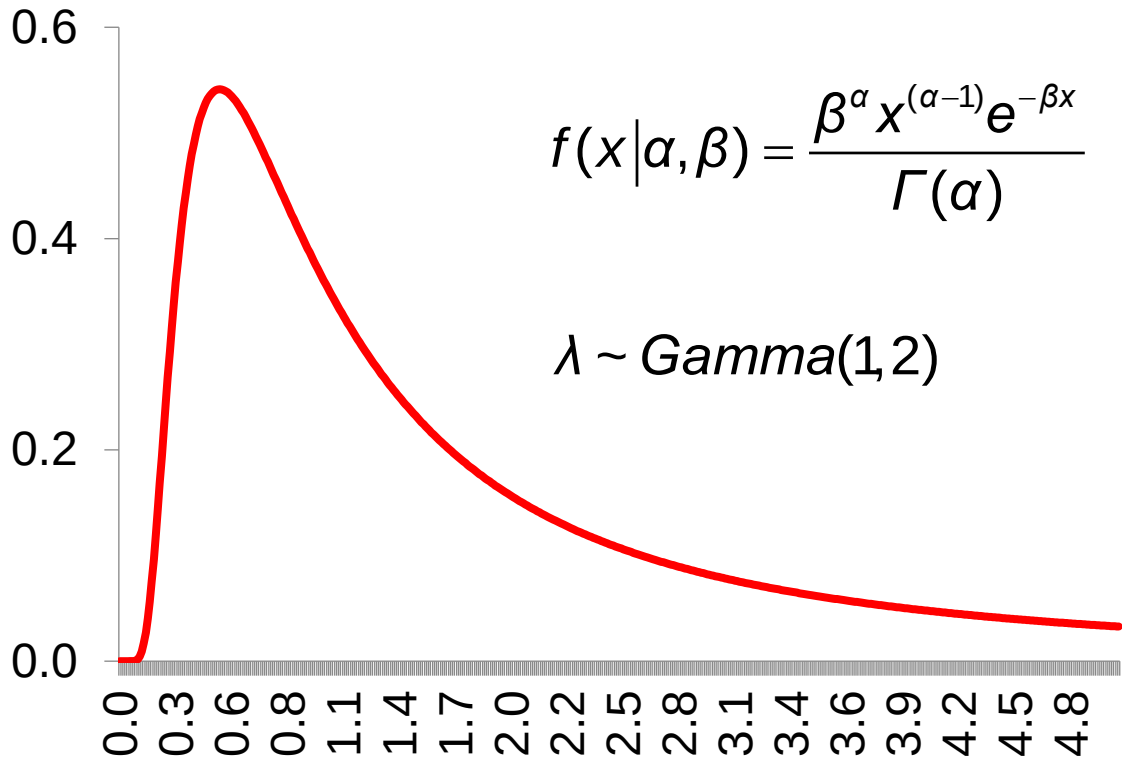
Conjugate Priors: Example

- Suppose now that observations of the interarrival times $w_1, w_2, \ldots, w_n$ are recorded.

- If λ were known, we could work out the likelihood of obtaining the observations we did.

$$L(w_1, w_2, \ldots, w_n \mid \lambda) = \left[\lambda e^{-\lambda w_1}\right]\left[\lambda e^{-\lambda w_2}\right]\ldots\left[\lambda e^{-\lambda w_n}\right]$$

$$= \lambda^n \left[e^{-\lambda(w_1 + w_2 + \ldots + w_n)}\right]$$



- Bayes' theorem gives $f(\lambda \mid w_1, w_2, \ldots, w_n) = \dfrac{L(w_1, w_2, \ldots, w_n \mid \lambda) f(\lambda)}{f(w_1, w_2, \ldots, w_n)}$.

Conjugate Priors: Example

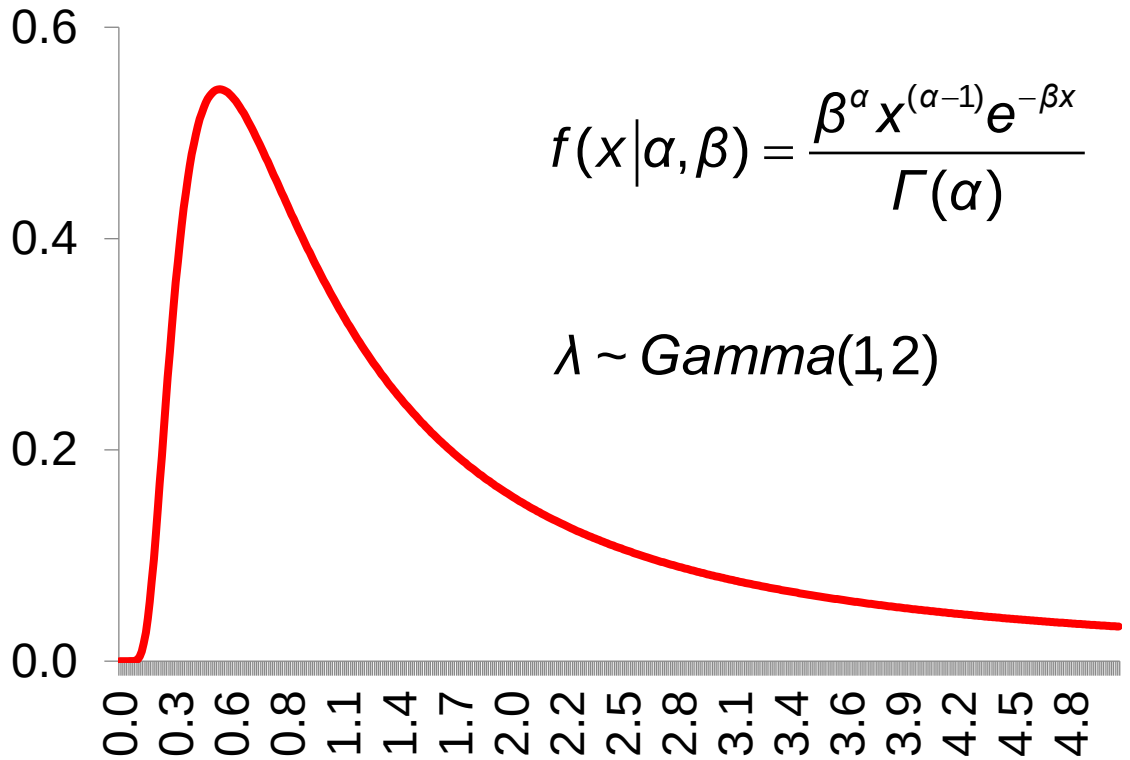
- $$f(\lambda|w_1,w_2,\ldots,w_n)=\frac{L(w_1,w_2,\ldots,w_n|\lambda)f(\lambda)}{f(w_1,w_2,\ldots,w_n)}$$

$$=\frac{\lambda^n\left[e^{-\lambda(w_1+w_2+\ldots+w_n)}\right]\left[\frac{\beta^\alpha\lambda^{(\alpha-1)}e^{-\beta\lambda}}{\Gamma(\alpha)}\right]}{f(w_1,w_2,\ldots,w_n)}$$

- This tells us that, building in the information from the observations, the posterior belief is

$$f(\lambda|w_1,w_2,\ldots,w_n)=\frac{\lambda^n\left[e^{-\lambda(w_1+w_2+\ldots+w_n)}\right]\left[\beta^\alpha\lambda^{(\alpha-1)}e^{-\beta\lambda}\right]}{f(w_1,w_2,\ldots,w_n)\Gamma(\alpha)}$$

$$=\frac{\lambda^{(n+\alpha)-1}\left[e^{-\lambda(\beta+w_1+w_2+\ldots+w_n)}\right]\left[\beta^\alpha\right]}{f(w_1,w_2,\ldots,w_n)\Gamma(\alpha)}.$$



$$f(x|\alpha,\beta)=\frac{\beta^\alpha x^{(\alpha-1)}e^{-\beta x}}{\Gamma(\alpha)}$$

$$\lambda\sim Gamma(1,2)$$

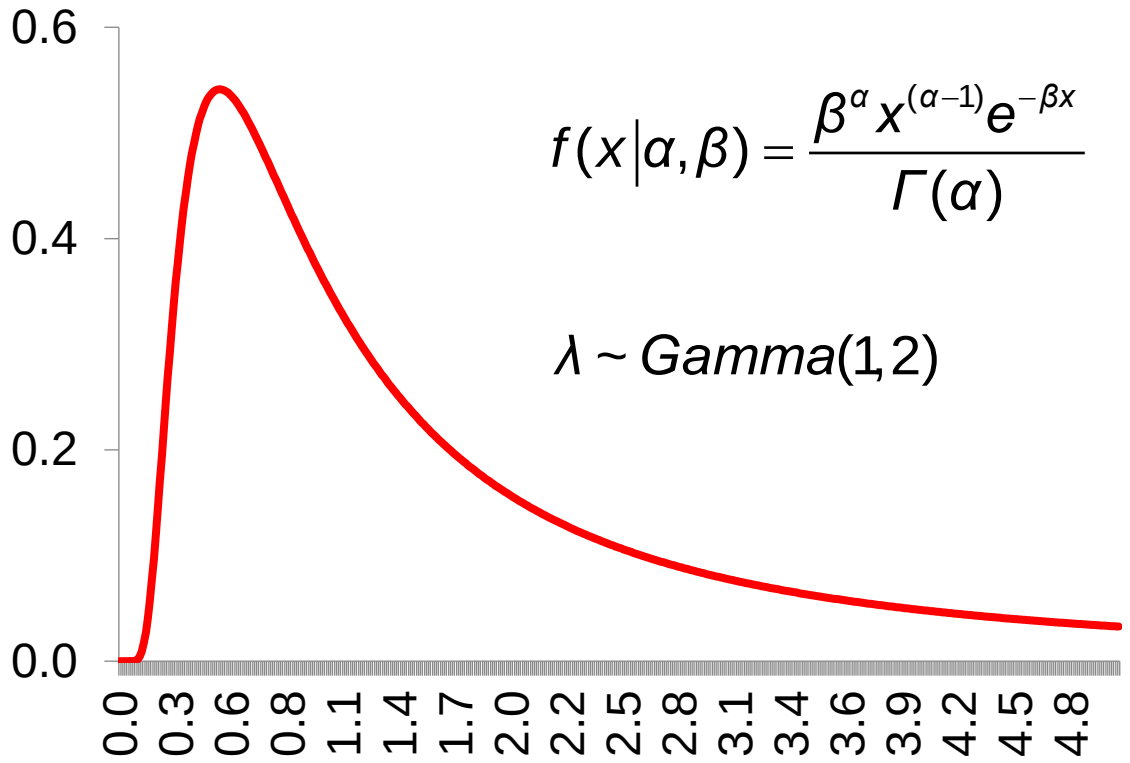
Conjugate Priors: Example

- Since the posterior distribution

$$f(\lambda|w_1,w_2,\ldots,w_n)=\frac{\lambda^{(n+\alpha)-1}\left[e^{-\lambda(\beta+w_1+w_2+\ldots+w_n)}\right]\left[\beta^{\alpha}\right]}{f(w_1,w_2,\ldots,w_n)\Gamma(\alpha)}$$

is also of a form proportional to $\lambda^{(\alpha-1)}e^{-\beta\lambda}$ we note that our posterior distribution after observing the data is

$$\lambda \sim Gamma\left(n+\alpha,\sum_{j=1}^{n}w_j+\beta\right).$$



Conjugate Priors: Example

- For example, if we had observations

  $w_1 = 0.5$

  $w_2 = 0.3$

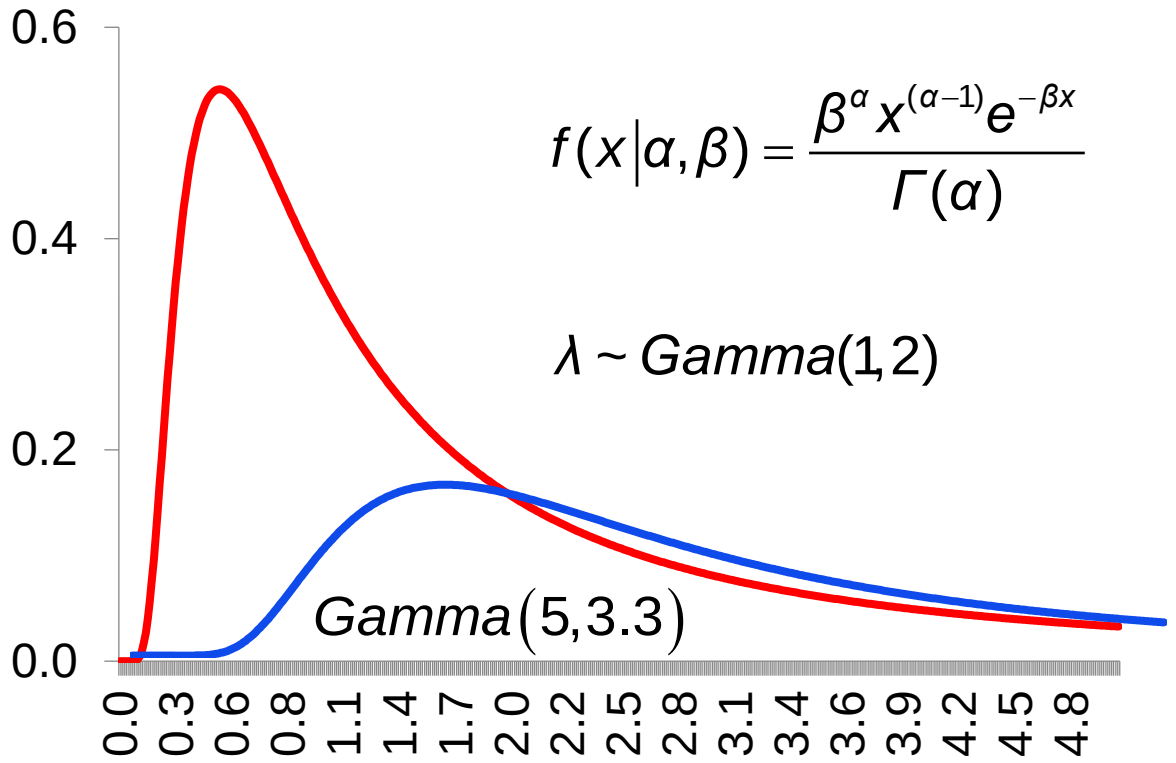
  $w_3 = 0.3$

  $w_4 = 0.2$

  our posterior belief about λ would be

$$\lambda \sim Gamma\left(n + \alpha, \sum_{j=1}^{n} w_j + \beta\right)$$

  i.e. $\lambda \sim Gamma(5, 3.3)$



- As the expectation of a $Gamma(\alpha,\beta)$ variable is $\frac{\alpha}{\beta}$, our smaller than expected observations of w led to our belief in the rate parameter increasing.

Conjugate Priors: Example

- A negative binomial variable with range $\{r, r+1, r+2, \ldots\}$ can be considered as the sum of r independent identically distributed geometric random variables. It describes how many independent identically distributed Bernoulli variables must be observed until r successes (1s) have been observed.

- For $N \sim NegBin(r, p)$, the probability mass function is

$$f(n) = P(N = n) = \begin{cases} \dfrac{(n-1)!}{(r-1)!(n-r)!} p^r (1-p)^{n-r} & n \in \{r, r+1, r+2, \ldots\} \\ 0 & \text{otherwise} \end{cases}$$

Conjugate Priors: Example

- Consider a series of observations of $N \sim NegBin(r, p)$, $\boldsymbol{N} = (n_1, n_2, ..., n_k)$.

- The likelihood of this sample (assuming r is known) would then be

$$L(\boldsymbol{N} \mid p) = \prod_{i=1}^{m} \frac{(n_i - 1)!}{(r-1)!(n_i - r)!} p^r (1-p)^{n_i - r}.$$ (assuming each $n_i \in \{r, r+1, r+2, ...\}$)

- We now consider quantifying our uncertainty about the value of p. Initially, we place a prior belief on p, $P \sim Beta(\alpha, \beta)$ since this only takes values on [0,1].

- $f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1}(1-p)^{\beta - 1}$ for $\alpha, \beta > 0$

- We can show that this is, in fact, a conjugate prior.

Conjugate Priors: Example

- $f(p \mid \boldsymbol{N}) \propto L(\boldsymbol{N} \mid p) f(p)$ hence $f(p \mid \boldsymbol{N}) \propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \prod_{i=1}^{m} \frac{(n_i-1)!}{(r-1)!(n_i-r)!} p^r (1-p)^{n_i-r}$
- $f(p \mid \boldsymbol{N}) \propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{i=1}^{m} \frac{(n_i-1)!}{(r-1)!(n_i-r)!} p^{\alpha-1} p^{\sum_{i=1}^{m} r} (1-p)^{\beta-1} (1-p)^{\sum_{i=1}^{m} n_i - r} \propto p^{\alpha+mr-1} (1-p)^{\beta+\sum_{i=1}^{m} n_i - mr - 1}$
- This gives a posterior density of $Beta\left(\alpha+mr, \beta+\sum_{i=1}^{m} n_i - mr\right)$
- As both the prior and posterior distribution are beta distributed, the prior is a conjugate prior for negative binomial outcomes.