Elements of Bayesian Inference

January 23, 2019
The parameter as a random variable

- So far we have seen the *frequentist* approach to statistical inference where inferential statements about $\theta$ are interpreted in terms of repeated sampling.
- In contrast the Bayesian approach treats $\theta$ as a *random variable* taking values in $\Theta$.
- The researcher’s information and beliefs about the possible values of $\theta$, before observing the data, are summarised by a *prior distribution* $p(\theta)$.
- When data $X = x$ are observed, this is combined with the prior to obtain the *posterior distribution* $\pi(\theta|x)$.
- Bayesian methods are appropriate in many applications: e.g. spam filters, speech recognition, bioinformatics, machine learning, and many more.
Bayesian Inference: Introduction

In the Bayesian approach to inference, parameters are treated as random variables and hence have a probability distribution.

**Prior information** about $\theta$ is combined with information from sample data to estimate the distribution of $\theta$.

This distribution contains all the available information about $\theta$ so should be used for making estimates or inferences.

We have prior information about $\theta$ given by the **prior distribution** $p(\theta)$ and information from sample data given by the **likelihood** $L(\theta, x) = f(x; \theta)$.

By Bayes Theorem the conditional distribution of $\theta$ given $X = x$ is

$$q(\theta; x) = \frac{f(x; \theta)p(\theta)}{h(x)} = \frac{L(\theta; x)p(\theta)}{h(x)} \propto L(\theta; x)p(\theta)$$

where $h(x)$ is the marginal distribution of $x$. We call $q(\theta; x)$ the **posterior distribution**.
Suppose I have 3 coins in my pocket:

1. biased 3:1 in favour of tails
2. a fair coin
3. biased 3:1 in favour of heads

I randomly select one coin and flip it once, observing a head. What is the probability that I have chosen coin 3?

Let $X = 1$ denote the event that I observe a head and $X = 0$ if a tail.

The probability of a head $\theta = (0.25, 0.5, 0.75)$ having equal prior probability.

The probability mass function $p(x|\theta) = \theta^x(1 - \theta)^{1-x}$ so $L(\theta; x) = P(x = 1|\theta) = \theta$
Inference about a discrete parameter

<table>
<thead>
<tr>
<th>Coin</th>
<th>$\theta$</th>
<th>Prior $p(\theta)$</th>
<th>Likelihood $L(\theta; x)$</th>
<th>Un-normalised Posterior $p(\theta)L(\theta; x)$</th>
<th>Posterior $q(\theta; x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.25</td>
<td>1/3</td>
<td>0.25</td>
<td>0.083</td>
<td>0.167</td>
</tr>
<tr>
<td>2</td>
<td>0.50</td>
<td>1/3</td>
<td>0.50</td>
<td>0.167</td>
<td>0.333</td>
</tr>
<tr>
<td>3</td>
<td>0.75</td>
<td>1/3</td>
<td>0.75</td>
<td>0.250</td>
<td>0.500</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td>1</td>
<td>0.75</td>
<td>$h(x) = 0.5$</td>
<td>1</td>
</tr>
</tbody>
</table>

The marginal distribution, or normalising constant $h(x) = \sum_i p(\theta_i)L(\theta_i; x)$.

So observing a head on a single toss of the coin means that there is now a 50% probability that the chance of heads is 0.75 and only a 16.7% probability that the chance of heads is 0.25.
Bayesian inference: How did it start?

In 1763, Reverend Thomas Bayes wrote:

PROBLEM.

*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.

In modern language, given $x \sim \text{Bin}(n, \theta)$, what is $P(\theta_1 < \theta < \theta_2 | x, n)$?
Prior Distributions

The prior distribution $p(\theta)$ quantifies information about $\theta$ prior to any (further) data being gathered.

Sometimes $p(\theta)$ can be constructed on the basis of past data.

More commonly, $p(\theta)$ must be based on an expert’s experience and personal judgement.

Example

Suppose that the proportions $\theta$ of defective items in a large manufactured lot is unknown. The prior distribution assigned to $\theta$ is $U(0, 1)$, i.e.

$$p(\theta) = \begin{cases} 1 & \text{for } 0 < \theta < 1 \\ 0 & \text{otherwise.} \end{cases}$$
Example

Suppose that the lifetimes of lamps of a certain type are to be observed. Let $X$ be the lifetime of any lamp and let $X \sim \text{Exp}(\beta)$, where $\beta$ is unknown. On the basis of previous experience the prior distribution of $\beta$ is taken as a gamma distribution with mean $r/\lambda = 0.0002$ and standard deviation $r/\lambda^2 = 0.0001$, i.e. $\text{Gamma}(r = 4, \lambda = 20000)$

$$p(\beta) = \begin{cases} \frac{20000^4}{3!} \beta^3 e^{-20000\beta}, & \beta > 0 \\ 0 & \text{otherwise.} \end{cases}$$
Prior Distributions

Example

A medical researcher was questioned about $\theta$, the proportion of asthma sufferers who would be helped by a new drug. She thought that

$$P[\theta > 0.3] = P[\theta < 0.3]$$

i.e. that the median $\theta_{0.5} = 0.3$.

Similarly, she thought that

$$\theta_{0.25} = 0.2 \quad \text{and} \quad \theta_{0.75} = 0.45$$

From tables giving quantiles of beta distributions, the researcher’s opinion could be represented by $\text{Beta}(\alpha = 2, \beta = 4)$ for which

$$\theta_{0.25} = 0.194, \quad \theta_{0.5} = 0.314 \quad \text{and} \quad \theta_{0.75} = 0.454$$
Parameter Estimation

Suppose we wish to estimate a parameter $\theta$. We define a loss function

$$L_s(\theta, \hat{\theta})$$

which measures the loss from taking $\hat{\theta}$ as an estimator when the true value is $\theta$.

The Bayes estimator minimizes the expected loss

$$E \left[ L_s(\theta, \hat{\theta}) | x \right] = \int_{-\infty}^{\infty} L_s(\theta, \hat{\theta}) q(\theta; x) d\theta$$

for the observed value of $x$.

The form of the Bayes estimator depends on the loss function that is used and the prior that is assigned to $\theta$.

For example, if the loss is the absolute error

$$L_s(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$$

then the Bayes estimator $\hat{\theta}_B(x)$ is the median of the posterior distribution.

For other loss functions the minimum might have to be numerically estimated.
Exercise

A commonly used loss function is the squared error loss function

\[ L_s(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2 \]

Show that the corresponding Bayes estimator \( \hat{\theta}_B \) is equal to the mean of the posterior distribution.

\[ \hat{\theta}_B(x) = E[\theta|x] \]

What is the minimum expected loss?
Continuing exercise 2

Example

Suppose that a random sample of \( n \) items is taken from the lot of manufactured items. Let

\[
X_i = \begin{cases} 
1 & \text{if } \text{'i' th item is defective} \\
0 & \text{otherwise.} 
\end{cases}
\]

then \( X_1, \ldots, X_n \) is a sequence of Bernoulli trials with parameter \( \theta \). The pdf of each \( X_i \) is

\[
f(x|\theta) = \begin{cases} 
\theta^x (1 - \theta)^{1-x} & \text{for } x = 0, 1 \\
0 & \text{otherwise.} 
\end{cases}
\]
Example

Then the joint pdf of \( X_1, \ldots, X_n \) is

\[
f_n(x|\theta) = \theta^{\sum x_i}(1 - \theta)^{n-\sum x_i}.
\]

Since the prior pdf \( p(\theta) \) is uniform it follows that

\[
f_n(x|\theta)p(\theta) = \theta^{\sum x_i}(1 - \theta)^{n-\sum x_i}, \quad 0 \leq \theta \leq 1.
\]

This is proportional to a beta distribution with parameters \( \alpha = y + 1 \) and \( \beta = n - y + 1 \), where \( y = \sum x_i \). Therefore the posterior has pdf

\[
q(\theta|x) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1 - \theta)^{n-y}, \quad 0 \leq \theta \leq 1.
\]
Example 3 contd.

Example

Suppose that the lifetimes $X_1, \ldots, X_n$ of a random sample of $n$ lamps are recorded. The pdf of each $x_i$ is

$$f(x_i, \beta) = \begin{cases} \beta e^{-\beta x_i} & x > 0, \\ 0 & \text{otherwise}. \end{cases}$$

The joint pdf of $x_1, \ldots, x_n|\beta$ is $f(x|\beta) = \beta^n e^{-\beta y}$, where $y = \sum_{i=1}^n x_i$. With a gamma specified for $p(\beta)$ we have

$$f(x|\beta)p(\beta) \propto \beta^{n+3} e^{-(y+20000)\beta}$$

where a factor that is constant w.r.t. $\beta$ has been omitted. The RHS is proportional to a $\text{Gamma}(n + 4, y + 20000)$, hence

$$q(\beta|x) = \frac{(y + 20000)^{n+4}}{(n + 3)!} \beta^{n+3} e^{-(y+20000)\beta}.$$
A **conjugate prior distribution** when combined with the likelihood function, produces a posterior distribution in the same family as the prior.

If we find a conjugate prior distribution which adequately fits our prior beliefs regarding $\theta$, we should use it because it will simplify computations considerably.
Sampling from a Bernoulli Distribution

Suppose $X_1, \ldots, X_n$ are a random sample from

$$\text{Be}(\theta), \quad 0 < \theta < 1.$$ 

Let $p(\theta)$ be

$$\text{Beta}(\alpha, \beta)$$

Then $q(\theta|x)$ is

$$\text{Beta}(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i)$$

**Proof:** analogous to Example 2 (note $U(0, 1) \equiv \text{Beta}(1, 1)$).
Suppose we are interested in the true mortality risk $\theta$ in a hospital $H$ which is about to try a new operation. The average hospital mortality rate is around 10%, but mortality rates in different hospitals vary from around 3% to around 20%. Hospital $H$ has no deaths in their first 10 operations. What should we believe about $\theta$?

Let $X_i = 1$ if the $i$th patient dies in $H$ (zero otherwise) $i = 1, \cdots, n$. Then

$$q(\theta|x) = \text{Beta}(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i), \ 0 < \theta < 1.$$
In practice we need to find a $\text{Be}(\alpha, \beta)$ prior distribution that matches the information from other hospitals.

A $\text{Be}(3, 27)$ prior distribution has mean 0.1 and $P(0.03 < \theta < 0.20) = 0.9$.

The data is $\sum_i x_i = 0$, $n = 10$, so the posterior distribution $q(\theta|x) = \text{Beta}(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i) = \text{Be}(3, 37)$, which has posterior mean $E(\theta|x) = 3/40 = 0.075$.

Even though nobody has died so far, the MLE $\hat{\theta} = \sum_{i=1}^n x_i/n = 0$ (i.e. it is impossible that any will ever die) does not seem plausible.
install.packages("LearnBayes")
library(LearnBayes)
prior = c( a= 3, b = 27 ) # beta prior
data = c( s = 0, f = 10 ) # s events out of f trials
triplot(prior,data)
Exercise

Suppose $X_1, \ldots, X_n$ are a random sample from

$$\text{Poisson}(\theta), \quad \theta > 0.$$ 

Let $p(\theta)$ be

$$\text{Gamma}(\alpha, \beta)$$

Show that $q(\theta|x)$ is

$$\text{Gamma}\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n\right)$$
Sampling from a Normal Distribution with known $\sigma^2$

Suppose $X_1, \ldots, X_n$ are a random sample from $N(\theta, \sigma^2)$ with $\sigma^2$ known.

Let $p(\theta)$ be $N(\phi, \tau^2)$.

Then $q(\theta|x)$ is

$$N \left( \frac{\phi \sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2}, \left( \frac{\sigma^2 + n\tau^2}{\sigma^2 \tau^2} \right)^{-1} \right)$$
Proof.

\[ q(\theta|x) \propto p(\theta)L(\theta; x) \]

\[ = (2\pi\tau^2)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \left( \frac{\theta - \phi}{\tau} \right)^2 \right\} \]

\[ \times \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \left( \frac{x_i - \theta}{\sigma} \right)^2 \right\} \]

\[ \therefore q(\theta|x) \propto \exp \left\{ -\frac{1}{2} \left[ \theta^2 \left( \frac{1}{\tau^2} + \frac{n}{\sigma^2} \right) - 2\theta \left( \frac{\phi}{\tau^2} + \frac{\sum x_i}{\sigma^2} \right) \right] \right\} \]

\[ = \exp \left\{ -\frac{\sigma^2 + n\tau^2}{2\sigma^2\tau^2} \left[ \theta - \frac{\phi\sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2} \right]^2 \right\} + \text{constant} \]

i.e. \( q(\theta|x) \) is the pdf of \( \mathcal{N} \left( \frac{\phi\sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2}, \left( \frac{\sigma^2 + n\tau^2}{\sigma^2\tau^2} \right)^{-1} \right) \) as required. \qed
Note that the posterior variance

\[
\left( \frac{\sigma^2 + n \tau^2}{\sigma^2 \tau^2} \right)^{-1} = \left( \frac{1}{\tau^2} + \frac{n}{\sigma^2} \right)^{-1}
\]

i.e. the reciprocal of the sum of the reciprocals of the prior variance and the variance of the sample mean, respectively.

Because of this reciprocal relationship

\[
\text{precision} = \frac{1}{\text{variance}}
\]

is sometimes quoted instead of the variance.
The posterior mean

$$\frac{\phi \sigma^2 + n \bar{x} \tau^2}{\sigma^2 + n \tau^2} = \frac{\phi}{\tau^2} + \frac{\bar{x}}{\sigma^2/n} = \frac{1}{\tau^2} + \frac{1}{\sigma^2/n}$$

i.e. a weighted average of the prior mean $\phi$ and the sample mean $\bar{x}$, with weights proportional to the prior precision and the precision of the sample mean.

This type of relationship holds for several sampling distributions when a conjugate prior is used.
Sampling from an Exponential Distribution

Suppose $X_1, \ldots, X_n$ are a random sample from $\text{Exp}(\theta), \; \theta > 0$.

Let $p(\theta)$ be $\text{Gamma}(\alpha, \beta)$

Then $q(\theta|x)$ is

$$\text{Gamma}\left(\alpha + n, \beta + \sum_{i=1}^{n} x_i\right)$$

**Proof:** see Example 3.
An **uninformative** or “flat” prior has $p(\theta) = \text{constant \ \forall \ \theta}$. 

Such priors can sometimes be obtained as special or limiting cases of conjugate priors, e.g.

- using a Beta$(1, 1) = U(0, 1)$ prior for the Bernoulli parameter
- letting $\tau^2 \rightarrow \infty$ in the $N(\phi, \tau^2)$ prior for the mean of $N(\mu, \sigma^2)$ with known $\sigma^2$.

If the prior is fairly constant over the range of $\theta$ for which the likelihood is appreciable, then approximately

$$q(\theta|X) \propto L(\theta|X)$$

and inference becomes equivalent to ML estimation.
Problems with Uninformative Priors

If the prior range of $\theta$ is infinite, an uninformative prior cannot integrate to 1. Such an improper prior may lead to problems in finding a “proper” posterior.

We usually have some prior knowledge of $\theta$ and if we do we should use it, rather than claiming ignorance.
Jeffrey’s Prior

Another issue is whether an informative prior should be flat for $\theta$ or some function of $\theta$, say $\theta^2$ or $\ln \theta$.

One solution is to construct a prior which is flat for a function $\phi(\theta)$ whose Fisher information $I_{\phi}$ is constant. This leads to Jeffrey’s prior which is proportional to

$$I_{\theta}^{\frac{1}{2}} = E \left[ \left( \frac{\partial \ln[L(\theta; x)]}{\partial \theta} \right)^2 \right]^{\frac{1}{2}} = \left[ -E \left( \frac{\partial^2 \ln[L(\theta; x)]}{\partial \theta^2} \right) \right]^{\frac{1}{2}}.$$

Example

Suppose we are sampling from a Bernoulli distribution so that the likelihood is binomial. Show that in this case, Jeffrey’s prior is proportional to

$$[\theta(1 - \theta)]^{-\frac{1}{2}}$$

which is a Beta $\left( \frac{1}{2}, \frac{1}{2} \right)$ distribution.
Bayesian inference treats unknown parameters as variables with a probability distribution. **Hierarchical models** exploit the flexibility this gives.

Consider the following **three-stage hierarchical model**. The data $x$ have density

$$f(x; \theta) \quad \text{[stage 1]}$$

where $\theta$ is unknown. Denote the prior distribution for $\theta$ by

$$p(\theta; \psi) \quad \text{[stage 2]}$$

where $\psi$ is also unknown, with prior distribution

$$g(\psi) \quad \text{[stage 3]}.$$
**Example**

Suppose $\theta_1, \ldots, \theta_k$ are the average reading abilities of 7-year-old children in each of $k$ different schools. Samples of 7-year-olds are to be given reading tests to estimate $\theta_i$. Let $X_{i1}, \ldots, X_{in_i}$ be the reading abilities of a sample of $n_i$ children from school $i$. Suppose

$$X_{ij} \sim N(\theta_i, \sigma^2) \quad [\text{stage 1}]$$

where $\sigma^2$ is the same for all schools and is assumed known.
Example

Then let each $\theta_i$ be normally distributed, so that

$$p(\theta_1, \ldots, \theta_k; \psi) = \prod_{i=1}^{k} (2\pi \tau^2)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\tau^2} (\theta_i - \phi)^2 \right\} \quad [\text{stage 2}]$$

where $\psi = (\phi, \tau)$. Finally assign an uninformative prior to $\phi$ and $\tau^2$

$$g(\phi, \tau^2) = g_1(\phi)g_2(\tau^2) = \text{constant} \quad [\text{stage 3}].$$

Note that information about some of the $\theta_i$ provides information about the remainder (borrowing strength).

For example, if we had data from $k - 1$ schools, then we could estimate the average reading abilities $\theta_1, \ldots, \theta_{k-1}$. Hence we could estimate $\phi$ and $\tau^2$, the mean and variance of the distribution of $\theta_k$. 
As usual

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$ 

Here the unknown parameters are $\theta$ and $\psi$ and their prior is 

$$p(\theta; \psi)g(\psi).$$

The sampling distribution of the data depends only on $\theta$, not $\psi$

$$L(\theta, \psi; x) = f(x; \theta).$$

Hence the joint posterior distribution of the parameters is 

$$q(\theta, \psi; x) \propto f(x; \theta)p(\theta; \psi)g(\psi).$$

From the joint posterior

$$q(\theta, \psi; x) \propto f(x; \theta)p(\theta; \psi)g(\psi)$$

we can obtain the posterior density of $\theta$ by integrating w.r.t. $\psi$.

Further integration yields marginal distributions of individual components of $\theta$.

For many hierarchical models the relevant integrations cannot be done analytically and the standard approach is to use MCMC to do the integration via computation.
Suppose $\theta$ is a vector of unknown parameters. As usual

$$q(\theta; x) \propto L(\theta; x)p(\theta)$$

and our aim of inference is to obtain $q(\theta; x)$. We can say that

$$q(\theta; x) = cL(\theta; x)p(\theta)$$

with

$$c = \left\{ \int L(\theta; x)p(\theta)d\theta \right\}^{-1}$$

We may be able to use Monte Carlo integration methods, e.g. importance sampling to find $c$, but this is often difficult. MCMC methods allows us to sample from $q(\theta; x)$ without knowing $c$. 

Inference using Markov Chain Monte Carlo (MCMC) Methods
Let $q_i(\theta_i|\theta\setminus i, x)$ denote the posterior probability density of $\theta_i$ given values of $\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k$.

The Gibbs sampler requires that for each $i = 1, \ldots, k$ these conditional densities are ones that we can easily sample from.

The algorithm aims to obtain a random sample from $q(\theta|x)$ by iteratively and successively sampling from the individual $q_i(\theta_i|\theta\setminus i, x)$. 
Gibbs Sampling

Initialize \( \theta \), i.e. find starting values \( \theta_i^{(1)} \), \( i = 1, \ldots, k \).

For \( j = 1, \ldots, M \)

1. Draw \( \theta_1^{(j+1)} \) from \( q_1(\theta_1 | \theta_{\setminus 1}^{(j)}, x) \); \( \theta_{\setminus 1}^{(j)} = (\theta_2^{(j)}, \theta_3^{(j)}, \ldots, \theta_k^{(j)}) \).
2. Draw \( \theta_2^{(j+1)} \) from \( q_2(\theta_2 | \theta_{\setminus 2}^{(j)}, x) \); \( \theta_{\setminus 2}^{(j)} = (\theta_1^{(j+1)}, \theta_3^{(j)}, \ldots, \theta_k^{(j)}) \).
3. \( \ldots \)
4. Draw \( \theta_k^{(j+1)} \) from \( q_k(\theta_k | \theta_{\setminus k}^{(j)}, x) \); \( \theta_{\setminus k}^{(j)} = (\theta_1^{(j+1)}, \theta_2^{(j+1)}, \ldots, \theta_{k-1}^{(j+1)}) \).
5. Put \( \theta^{(j+1)} = (\theta_1^{(j+1)}, \theta_2^{(j+1)}, \ldots, \theta_k^{(j+1)}) \). Set \( j = j + 1 \).
Gibbs Sampling

As $j \to \infty$, under suitable regularity conditions, the limiting distribution of the vector $\theta^{(j)}$ is the required posterior $q(\theta; x)$, i.e. for large $j$, $\theta^{(j)}$ is a random observation from $q(\theta; x)$.

The sequence $\theta^{(1)}, \theta^{(2)}, \ldots$ is one realisation of a Markov Chain, since the probability of $\theta^{(j+1)}$ is only dependent on $\theta^{(j)}$.

We need to run the Markov chain until it has converged to its stationary distribution — all observations from the burn-in phase are discarded.
Gibbs Sampling

Suppose we have generated a large random sample $\theta^{[1]}, \theta^{[2]}, \ldots, \theta^{[n]}$ using the Gibbs sampler.

Inferences about a single $\theta_i$ would be based on this sample. For example the posterior mean and variance of $\theta_i$ given $x$ would be

$$\bar{\theta}_i = \frac{1}{n} \sum_j \theta_i^{[j]}$$

and

$$\frac{1}{n} \sum_j (\theta_i^{[j]} - \bar{\theta}_i)^2$$

using the ML estimator for the sample variance.

The output from MCMC can also be used to approximate integrals, e.g. for marginalisation.
Exercise

Let $X_1, \ldots, X_n$ be a random sample. Consider the following hierarchical Bayesian model:

\[
X_1, \ldots, X_n \sim N(\theta, \sigma^2) \quad \sigma^2 \text{ is known},
\]
\[
\theta \sim N(0, \tau^2),
\]
\[
\frac{1}{\tau^2} \sim \text{Gamma}(a, b) \quad \text{where } a, b \text{ are known}.
\]

Derive the Gibbs sampler for this model.
Solution: First let’s derive

\[ g(\theta|x, \tau^2) \propto f(x|\theta)p(\theta|\tau^2)p(\tau^{-2}) \]

We can omit all parts that do not depend on \( \theta \):

\[ g(\theta|x, \tau^2) \propto f(x|\theta)p(\theta|\tau^2) \]

The RHS is equivalent to the likelihood \( \times \) prior in the normal-normal model seen earlier. So \( g(\theta|x, \tau^2) \) is the pdf

\[
N \left( \frac{\phi \sigma^2 + n \bar{x} \tau^2}{\sigma^2 + n \tau^2}, \left( \frac{\sigma^2 + n \tau^2}{\sigma^2 \tau^2} \right)^{-1} \right)
\]

Similarly,

\[ g(\tau|x, \theta) \propto p(\theta|\tau^2)p(\tau^2). \]
Thus

\[ g(\tau|x, \theta) \propto \frac{1}{\tau} \exp \left\{ -\frac{1}{2} \frac{\theta^2}{\tau^2} \right\} \left( \frac{1}{\tau^2} \right)^{a-1} \exp \left\{ -\frac{1}{\tau^2} \frac{1}{b} \right\} \]

\[ \propto \left( \frac{1}{\tau^2} \right)^{a + \frac{1}{2} - 1} \exp \left\{ -\frac{1}{\tau^2} \left[ \frac{\theta^2}{2} + \frac{1}{b} \right] \right\} \]

which specifies the pdf of

\[ \text{Gamma} \left( a + \frac{1}{2}, \frac{\theta^2}{2} + \frac{1}{b} \right) \]
Thus the Gibbs sampler for this model is:

For \( i = 1, \ldots, m \) sample from the following conditionals

\[
\theta(i) | \tau^2(i-1), x \sim N \left( \frac{\phi \sigma^2 + n \bar{x} \tau^2}{\sigma^2 + n \tau^2}, \left( \frac{\sigma^2 + n \tau^2}{\sigma^2 \tau^2} \right)^{-1} \right)
\]

\[
\frac{1}{\tau^2(i)} \left| \theta(i), x \sim \text{Gamma} \left( a + \frac{1}{2}, \frac{\theta^2}{2} + \frac{1}{b} \right) \right.
\]
The Metropolis-Hastings Algorithm is an extension of Gibbs sampling for non-standard conditional densities $q_i(\theta_i|\theta_{\setminus i}^{(j)}, x)$, which cannot be sampled from directly.

The basic procedure is the same, but now acceptance/rejection sampling is used to draw $\theta_i^{(j+1)}$ from $q_i(\theta_i|\theta_{\setminus i}^{(j)}, x)$. 
Exercise

Let \( X_1, \ldots, X_n \) be a random sample. Consider the following hierarchical Bayesian model of failure rates:

\[
X_1, \ldots, X_n \sim \text{Poisson}(\theta_i t_i) \quad \text{for fixed time } t_i,
\]
\[
\theta_i \sim \text{Gamma}(\alpha, \beta),
\]
\[
\alpha \sim \text{Exp}(a_0) \quad \text{for known } a_0,
\]
\[
\beta \sim \text{Gamma}(c, b_0) \quad \text{for known } c, b_0,
\]

where \( \alpha \) and \( \beta \) are independent. Show that the conditionals for \( \theta_i \) and \( \beta \) are Gamma distributions, whilst

\[
P(\alpha|\beta, \theta) \propto \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n \left(\prod_{i=1}^n \theta_i\right)^{\alpha-1} \exp(-a_0 \alpha).
\]