Bayesian Statistics

Andrew Andreas

July 1, 2024

1 Introduction

Please report any errors to andrew.andreas0@gmail.com

2 Background

Bayesian statistics is a branch of statistics that is based on the Bayesian interpretation of probability whereby probability expresses a degree of belief in an event. Our uncertainties are encapsulated by probability distributions which are updated after obtaining new data. This is done via **Bayes theorem**.

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{x})} \tag{1}$$

$$f(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \tag{2}$$

- $f(\boldsymbol{\theta})$: The **Prior** distribution which encapsulates the beliefs *prior* to observing any data
- $f(\mathbf{x}|\boldsymbol{\theta})$: The **Likelihood** which encapsulates the information in the data we observe and can be thought of as the evidence of $\mathbf{x}$ given that $\boldsymbol{\theta}$ is true
- $f(\mathbf{x})$: The normalizing constant, which is the marginal distribution of the data, ensures that the posterior integrates to 1
- $f(\boldsymbol{\theta}|\mathbf{x})$: The **Posterior** distribution which encapsulates our beliefs **after** having observed the data

3 Prior Distributions

A prior distribution captures the beliefs about $\boldsymbol{\theta}$ before any data are available, that is they encode the beliefs about its plausible values. Some useful definitions are listed below

Precision is the reciprocal of the variance and thus higher precision means smaller variance

$$\tau = \frac{1}{\mathbb{V}}$$

Strong prior is one in which the precision is higher and thus the density will be concentrated around a smaller range of values.

Weak prior is one with low precision and the density will be wider and flatter, encoding our uncertainty around the range of values for $\boldsymbol{\theta}$

Improper Prior is one which is not actually a valid probability density function and is the result of a (very) weak prior. Each of the examples below is not a proper density as it violates some aspect of the distribution, mainly parameter values.

- $U(a, \infty)$
- $U(-\infty, b)$
- $U(-\infty, \infty)$

- $N(a, \infty)$
- Gamma$(0, 0)$

Improper priors often still yield valid posterior densities and encode (very) weak information, hence their usefulness in some instances. However, it must be noted that improper priors do not always yield valid posterior densities.

Non-informative Priors do not provide any information about the parameter, $\boldsymbol{\theta}$. These priors are flat across their entire range of possible values.

3.1 Common Priors

Common prior distributions include the following four distributions because they exhibit useful qualities such as **Conjugacy**.

- The Uniform Distribution, $U(a_3, b_3)$
- The Normal Distribution, $N(a_1, b_1)$
- The Gamma Distribution, Gamma(a_2, b_2)
- The Beta Distribution, Beta(a_4, b_4)

When we first introduced Bayes theorem we simplified the equation by ignoring the normalizing constant.

$$f(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta})$$

If we treat all terms not involving $\boldsymbol{\theta}$ as constants, it can simplify our calculations because the density core of the posterior distribution will either be a recognized distribution from which we know the form of the normalizing constant or it will not be. In the latter case we will then have to calculate the marginal posterior density.

Example 1 Consider independent random variables modelled by the Poisson distribution with mean $\lambda > 0$, then the likelihood is given by

$$L(\lambda) = \lambda^{n\bar{x}} \exp(-n\lambda)$$

If we use a Gamma(a, b) prior for λ, which is appropriate since a RV following a Gamma distribution can only take on positive values, then its density core is given by

$$f(\lambda) \propto \lambda^{a-1} \exp(-b\lambda)$$

So that the posterior distribution can be computed as

$$\begin{aligned} f(\lambda|x) &\propto L(\lambda) f(\lambda) \\ f(\lambda|x) &\propto \lambda^{n\bar{x}} \exp(-n\lambda) \times \lambda^{a-1} \exp(-b\lambda) \\ &\propto \lambda^{n\bar{x}+a-1} \exp(-\lambda(n+b)) \end{aligned}$$

Notice that the posterior density core is the form of a Gamma$(n\bar{x} + a, n + b)$ distribution whose normalizing constant we know.

4 Conjugate Priors

We alluded to the idea of conjugacy earlier when we mentioned some common prior distributions but we will formalise the definition here.

Conjugacy For a family of distributions, F, these are said to be **closed under sampling** from $f(\mathbf{x}|\boldsymbol{\theta})$ if for every prior in F then the posterior distribution $f(\boldsymbol{\theta}|\mathbf{x})$ is also in F.

The benefits of using conjugate models are as follows

- The posterior distribution is a standard distribution and is hence well understood

- The posterior is easy to compute analytically as it is of a standard form
- We can easily compare the posterior and prior given they are from the same family of distributions

We are going to list some common conjugate models below. In the appendix there will be a summary of some of the key probability distributions which are used extensively here.

4.1 The beta/binomial model

Suppose that a random variable X is modelled by a binomial distribution, $X \sim B(n, \theta)$

$$f(x|\theta) \propto \theta^x (1-\theta)^{n-x}$$

and a $Beta(a, b)$ prior is used for θ

$$f(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Then the posterior distribution, $f(\theta|\mathbf{x})$, is given by

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto \theta^{a-1}(1-\theta)^{b-1} \times \propto \theta^x (1-\theta)^{n-x} \\ &= \theta^{a+x-1}(1-\theta)^{b+n-x-1} \end{aligned}$$

which is a Beta$(a+x, b+n-x)$ distribution. We also know that the Beta distribution has mean given by

$$\mathbb{E}(\theta) = \frac{a}{a+b}$$

4.1.1 The Posterior Mean

Then we can consider the mean of the posterior distribution to be a weighted average of the observed mean, x/n, and the prior mean, where for increasing n the posterior mean tends towards the mean of the likelihood

$$\begin{aligned} \mathbb{E}(\theta|x) &= \frac{a+x}{a+b+n} \\ &= \frac{a}{a+b+n} + \frac{x}{a+b+n} \\ &= \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{x}{n} \\ &= t\frac{a}{a+b} + (1-t)\frac{x}{n} \\ &\quad \text{where } t = \frac{a+b}{a+b+n} \text{ and } (1-t) = \frac{n}{a+b+n} \end{aligned}$$

which highlights the relationship of increasing sample size on the weighting towards to observed mean.

4.1.2 The Posterior Variance

The variance of the Beta distribution is given by

$$\mathbb{V}(\theta) = \frac{ab}{(a+b+1)(a+b)^2}$$

meaning the posterior variance is

$$\mathbb{V}(\theta|x) = \frac{(a+x)(b+n-x)}{(a+b+n-1)(a+b+n)^2}$$

which we can express in the form

$$\begin{aligned} \mathbb{V}(\theta|x) =& \frac{a+x}{a+b+n}\frac{b+n-x}{(a+b+n)(a+b+n+1)} \\ =& \frac{\mathbb{E}(\theta|x)(1-\mathbb{E}(\theta|x))}{a+b+n-1} \end{aligned}$$

where we see that in the limit as $n \to \infty$, $\mathbb{V}(\theta|x) \to 0$

4.2 The gamma/Poisson model

We have actually already shown the conjugacy of the gamma/Poisson model in **Example 1** but we will restate the result here and go on to derive some useful properties.

For independent random variables modelled by the Poisson distribution,

$$X_i \sim \text{Poisson}(\lambda)$$

then if a Gamma prior is used, such that

$$\lambda \sim \text{Gamma}(a, b)$$

then the posterior distribution for λ is also a Gamma distribution

$$f(\lambda|\mathbf{x}) \sim \text{Gamma}(a + n\bar{x}, b + n)$$

4.2.1 The Posterior Mean

The prior mean for the Gamma distribution is given by

$$\mathbb{E}(\lambda) = \frac{a}{b}$$

and the observed (sample) mean is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Then the posterior mean is therefore given by

$$\begin{aligned}
\mathbb{E}(\lambda|\mathbf{x}) =& \frac{a + n\bar{x}}{b + n} \\
=& \frac{a}{b + n} + \frac{n\bar{x}}{b + n} \\
=& \frac{a}{b}\frac{b}{b + n} + \frac{n}{b + n}\bar{x} \\
=& \mathbb{E}(\lambda)t + \bar{x}(1 - t) \\
& \text{where } t = \frac{b}{b + n} \text{ and } (1 - t) = \frac{n}{b + n}
\end{aligned}$$

Again we can see that as the sample size increases then the posterior mean tends towards the observed mean.

4.2.2 The Posterior Variance

The variance for the Gamma distribution is given by

$$\mathbb{V}(\lambda) = \frac{a}{b^2}$$

Therefore the posterior mean would be

$$\mathbb{V}(\lambda) = \frac{a + n\bar{x}}{(b + n)^2}$$

Given the squared dependence of the sample size in the numerator then we see that in the limit as $n \to \infty$, $\mathbb{V}(\theta|x) \to 0$

4.3 The normal-gamma/normal model

Suppose that independent random variables $X_i, i = 1, ...n$ are modelled by a normal distribution with unknown mean and unknown precision

$$X_i \sim N(\mu, \frac{1}{\tau})$$

and the prior distribution turns out to be the normal-gamma distribution

$$(\mu, \tau) \sim \text{Ngamma}(a, b, c, d) \propto \tau^{1/2} \exp(-\frac{\tau}{2b}(\mu - a)^2)\tau^{c-1} \exp(-d\tau)$$
$$\mu|\tau \sim N(a, \frac{b}{\tau}) \text{ and } \tau \sim \text{Gamma}(c, d)$$

Note the dependence between μ and τ in the conditional normal distribution for μ. Ideally there will be independence between parameters but in this case no such conjugate prior exists.

Then the posterior distribution is also a normal-gamma distribution with parameters a_1, b_1, c_1, d_1, so that

$$a_1 = \frac{a + bn\bar{x}}{1 + bn}, \qquad b_1 = \frac{b}{1 + bn}, \qquad c_1 = c + n/2$$
$$d_1 = d + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{n(\bar{x} - a)^2}{2(1 + bn)}$$

Note that the case where both the mean and variance of the observed data are assumed unknown is in reality the most likely case. However, in the appendix there is a derivation of two (more unrealistic cases)

- **Case 1: The normal/normal model** μ is assumed unknown while the variance is known $\sigma = \sigma_0^2$
- **Case 2: The gamma/normal model** $\mu = \mu_0$ is assumed known while the precision, τ, is unknown

4.3.1 Marginal prior for μ

The margin distribution is computed by integrating out the variables not of interest, in this case τ

$$\begin{aligned} f(\mu) &= \int_{\tau=0}^{\infty} f(\mu, \tau) d\tau \\ &= \int_{\tau=0}^{\infty} \propto \tau^{1/2} \exp(-\frac{\tau}{2b}(\mu - a)^2)\tau^{c-1} \exp(-d\tau) d\tau \\ &= \int_{\tau=0}^{\infty} \tau^{c-1/2} \exp\Big[-\tau\Big(\frac{1}{2b}(\mu - a)^2 + d\Big)\Big] d\tau \end{aligned}$$

which is the density core of a $\text{Gamma}\Big(c + 1/2, \frac{(\mu-a)^2+d}{2b}\Big)$ distribution.

The general form of the normalizing constant for a $\text{Gamma}(a, b)$ distribution is $b^a/\Gamma(a)$, where $\Gamma(a)$ is the gamma function. So for this gamma distribution it is

$$\frac{\Gamma(c + 1/2)}{\left(\frac{(\mu-a)^2}{2b} + d\right)^{c+1/2}}$$

This tells us that the marginal prior is equal to the inverse of the normalizing constant

$$\begin{aligned} f(\mu) &= \int_{\tau=0}^{\infty} \tau^{c-1/2} \exp\Big[-\tau\Big(\frac{1}{2b}(\mu - a)^2 + d\Big)\Big] d\tau \\ &= \frac{\left(\frac{(\mu-a)^2}{2b} + d\right)^{c+1/2}}{\Gamma(c + 1/2)} \\ &\propto \left(\frac{(\mu - a)^2}{2b} + d\right)^{-(c+1/2)} \end{aligned}$$

While we now have the form of the marginal prior, we don't yet know its distribution. We can however find it and it turns out to be a relocated and rescaled t distribution. To see this, first

recall that the prior for precision was a Gamma(c, d) with mean c/d

$$\left(\mathbb{E}(\tau)\frac{(\mu-a)^2}{2bc}+1\right)^{-(2c+1)/2}$$

Comparing this to the t distribution with density $f(x)$

$$f(x) \propto (1+\frac{x^2}{v})^{-(v+1)/2}$$

we see that the degrees of freedom is $v = 2c$ and the mean of the distribution has been shifted by a and the distribution scaled by $b/\mathbb{E}(\tau)$.

We can therefore conclude that the marginal prior of the normal-gamma / normal model follows a relocated and rescaled t distribution

$$\mu \sim t\Big(2c; a, \frac{b}{\mathbb{E}(\tau)}\Big)$$

4.3.2 Marginal Posterior for $\mu|\mathbf{x}$, $\tau|\mathbf{x}$

The posterior distribution $\mu, \tau|\mathbf{x}$ is a normal-gamma distribution as mentioned here.

Since we have just determined that a Ngamma joint prior results in a relocated and rescaled t distribution as a marginal prior for μ, we can easily determine that the marginal posterior for $\mu|\mathbf{x}$ and $\tau|\mathbf{x}$ will be

$$\begin{aligned}\mu|\mathbf{x} &\sim t(2c_1, a_1, \frac{b_1}{\mathbb{E}(\tau|\mathbf{x})})\\ \tau|\mathbf{x} &\sim \text{ Gamma}(c_1, d_1)\end{aligned}$$

5 The Natural Exponential Family

The random variable X has a distribution from the **natural exponential family (NEF)** with a one-dimensional parameter θ if its density can be expressed in the form

$$f(x|\theta) = \exp\big(xa(\theta) - b(\theta) + c(x)\big)$$

where f denotes either the pdf or pmf and the support of f does not depend on θ

In practice it is easier to compare the log form of the NEF with prospective distributions

$$(x|\theta) = xa(\theta) - b(\theta) + c(x)$$

5.1 The Natural Exponential Family Model

Suppose independent random variables $X_i, i = 1, ..., n$ are modelled by the NEF with likelihood

$$f(x|\mathbf{x}) = \exp\Big[n\bar{x}a(\theta) - nb(\theta) + \sum_{i=1}^{n} c(x_i)\Big]$$

Then the conjugate prior has the form

$$f(\theta) \propto \exp\big(\alpha(\theta) - \beta b(\theta)\big)$$

and the posterior is given by

$$f(\theta|\mathbf{x}) \propto \exp\Big[(n\bar{x}+\alpha)a(\theta) - (n+\beta)b(\theta)\Big]$$

5.1.1 The Poisson Distribution

The Poisson distribution with pmf

$$f(x|\lambda) = \frac{\exp(-\lambda)\lambda^x}{x!}$$
$$\log(f(x|\lambda)) = x\log(\lambda) - \lambda - \log(x!)$$

where

$$a(\lambda) = \log(\lambda), \qquad b(\lambda) = \lambda, \qquad c(x) = -\log(x!)$$

5.1.2 The Normal Distribution

The normal distribution with known variance σ_0^2 can be decomposed into the following formula

$$f(x|\mu) = -\frac{\mu^2}{2\sigma_0^2} + \frac{x\mu}{\sigma_0^2} - \frac{x^2}{2\sigma_0^2} - \log(\sqrt{2\pi\sigma_0^2})$$

which when compared directly with the NEF pdf yields

$$a(\mu) = -\frac{\mu^2}{2\sigma_0^2}, \qquad b(\mu) = \frac{\mu}{\sigma_0^2}, \qquad c(x) = -\frac{x^2}{2\sigma_0^2} - \log(\sqrt{2\pi\sigma_0^2})$$

5.1.3 The Bernoulli Distribution

The Bernoulli distribution with PMF

$$f(x|\theta) = \begin{cases} \theta, & x = 1 \\ 1-\theta, & x = 0 \end{cases}$$

can be rewritten as a binomial distribution with $n = 1$

$$f(x|\theta) = \theta^x(1-\theta)^{1-x} \qquad x = 0, 1$$

then taking logarithms

$$\begin{aligned} \log\big(f(x|\theta)\big) &= x\log\theta - x\log(1-\theta) + \log(1-\theta) \\ &= x\log\Big(\frac{\theta}{1-\theta}\Big) + \log(1-\theta) \end{aligned}$$

So that

$$a(\theta) = \log\Big(\frac{\theta}{1-\theta}\Big), \qquad b(\theta) = -\log(1-\theta), \qquad c(x) = 0$$

which proves that the Bernoulli distribution is a member of the NEF.

5.1.4 The Binomial Distribution

If we consider a random variable $X = \sum_{i=1}^n X_i$ which is a sum of n independent Bernoulli random variables then, X follows a binomial distribution with parameters (n, θ). Therefore we can think of the binomial distribution, with density core given by

$$f(x|\theta) \propto \theta^x(1-\theta)^{n-x}$$

then we can show that the binomial distribution is a member of the NEF if we extend the case of the Bernoulli trial to n independent observations $X_i, i = 1, ..., n$ then the log likelihood becomes

$$\begin{aligned} \log\big(f(\mathbf{x}|\theta)\big) &= \sum_{i=1}^n \Big[x_i\log\Big(\frac{\theta}{1-\theta}\Big) + \log(1-\theta)\Big] \\ &= n\bar{x}\log\Big(\frac{\theta}{1-\theta}\Big) + n\log(1-\theta) \\ &= n\bar{x}\log\theta - n\bar{x}\log\big(1-\theta\big) + n\log(1-\theta) \\ &= n\bar{x}\log\theta + n(1-\bar{x})\log(1-\theta) \\ &= \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})} \\ &= \theta^x(1-\theta)^{n-x} \end{aligned}$$

5.1.5 The gamma/Poisson model

The gamma/Poisson model has posterior density that follows a Gamma$(a+n\bar{x}, b+n)$ distribution with log density core

$$\begin{aligned} log\big(f(\lambda|\mathbf{x})\big) =& (a+n\bar{x}-1)\log\lambda - (b+n)\lambda \\ =& (a+n\bar{x}-1)a(\lambda) - (b+n)b(\lambda) \end{aligned}$$

so that the form of $f(\lambda|\mathbf{x})$ is

$$f(\lambda|\mathbf{x}) = \exp\Big[(a+n\bar{x}-1)a(\lambda) - (b+n)b(\lambda)\Big]$$

where if we compare to the general form of the NEF yields $\alpha_1 = a-1$ and $\beta_1 = b$

5.1.6 The beta/binomial model

The log of the core of the beta/binomial posterior density is given by

$$\begin{aligned} \log\big(f(\theta|\mathbf{x})\big) &= (a+x-1)\log\Big(\frac{\theta}{1-\theta}\Big) - (a+b+n-2)\log\big(1-\theta\big) \\ &= (a+x-1)a(\theta) - (a+b+n-2)b(\theta) \end{aligned}$$

noting that $n\bar{x}$ in the Bernoulli formulation corresponds to x in the binomial formulation, so that the form of $f(\theta|\mathbf{x})$ is

$$f(\theta|\mathbf{x}) = \exp\Big[(a+n\bar{x}-1)a(\theta) - (a+b+n-2)b(\theta)\Big]$$

where if we compare to the general form of the NEF yields $\alpha_2 = a-1$ and $\beta_2 = a+b-2$

6 Bayesian Inference

Bayesian Inference is based on **Statistical Decision Theory** which we will first view through the lens of using the posterior distribution to estimate the value of the unknown parameter. We will then extend this idea to the prediction of future observations using the **predictive distribution**.

6.1 Statistical Decision Theory

The general framework for Bayesian inference follows the following principles

- Define a set of D of possible decisions that can be taken
- Define a loss function, $L(d,\theta)$, which is a function of both the decision and the parameter. The loss function should provide a numerical loss for for each possible decision, D.
- Find the decision, D, which minimizes the posterior expected loss

$$\mathbb{E}\Big[L(d,\theta)|\mathbf{x}\Big] = \int_{\theta\in\Omega} L(d,\theta)f(\theta|\mathbf{x})d\theta$$

Note that we take expectation over the posterior distribution because we do not know the true value of the parameter θ so we use the posterior distribution which represents the cumulative knowledge about θ including our prior beliefs and after having observed the data.

6.1.1 Loss Functions

There are a multitude of loss functions to choose from and the most appropriate choice depends on the problem at hand.
A few common loss functions are listed below and the decision which minimizes them is also derived

6.1.2 The Squared Loss

The squared loss function has the form

$$L(d, \theta) = (d - \theta)^2$$

for $d \in D$ and the parameter θ and is an appropriate choice for real-valued functions in instances where distance from the optimal decision matters.

The expected squared loss function is then given by

$$\mathbb{E}\Big[L(d, \theta)|\mathbf{x}\Big] = \mathbb{E}\Big[(d - \theta)^2|\mathbf{x}\Big]$$

and to minimize the expected squared loss function we can expand the bracket and use the definition of the variance $\mathbb{V}\big[\theta\big] = \mathbb{E}\big[\theta^2\big] - \mathbb{E}\big[\theta\big]^2$

$$\begin{aligned}
\mathbb{E}\Big[(d - \theta)^2|\mathbf{x}\Big] =& \mathbb{E}\Big[(d^2 - 2d\theta + \theta^2)|\mathbf{x}\Big] \\
=& \mathbb{E}\Big[d^2|\mathbf{x}\Big] - 2d\mathbb{E}\Big[\theta|\mathbf{x}\Big]\theta + \mathbb{E}\Big[\theta^2|\mathbf{x}\Big] \\
=& d^2 - 2d\mathbb{E}\Big[\theta|\mathbf{x}\Big]\theta + \mathbb{V}\Big[\theta|\mathbf{x}\Big] + \mathbb{E}\Big[\theta|\mathbf{x}\Big]^2 \\
=& (d - \mathbb{E}\Big[\theta|\mathbf{x}\Big])^2 + \mathbb{V}\Big[\theta|\mathbf{x}\Big]
\end{aligned}$$

and from this formulae we can see that the best decision is the posterior mean $\mathbb{E}\Big[\theta|\mathbf{x}\Big]$. Note the similarity here between the the decomposition of the mean squared error and the squared loss which should come as no surprise given that the mean squared error is just the average squared loss.

6.1.3 The Absolute Loss Function

The absolute loss function has the form

$$L(d, \theta) = |d - \theta| \tag{3}$$

for $d \in D$ and the parameter θ. Unlike the quadratic loss, the loss is a linear function and so penalizes larger losses less severely. However, it is still an appropriate choice for real-valued functions where positive and negative differences aren't distinguished.

As the absolute value is given by

$$|d - \theta| = \begin{cases} (d - \theta) & d > \theta \\ (\theta - d) & d < \theta \end{cases}$$

we must therefore consider the loss over both cases. The expected absolute loss function is then given by

$$\begin{aligned}
\mathbb{E}\Big[L(d, \theta)|\mathbf{x}\Big] =& \mathbb{E}\Big[|d - \theta||\mathbf{x}\Big] \\
=& \mathbb{E}\Big[(d - \theta)|\mathbf{x}\Big] + \mathbb{E}\Big[(\theta - d)|\mathbf{x}\Big]
\end{aligned}$$

If we look at the definition of the absolute value we see that in each case θ is bounded from above or below by the decision value, d. This then defines the bounds of the integral definition of the expected value

$$\mathbb{E}\Big[(d - \theta)|\mathbf{x}\Big] + \mathbb{E}\Big[(\theta - d)|\mathbf{x}\Big] = \int_{-\infty}^{d} df(\theta|\mathbf{x})d\theta - \int_{-\infty}^{d} \theta f(\theta|\mathbf{x})d\theta + \int_{d}^{\infty} \theta f(\theta|\mathbf{x})d\theta - \int_{d}^{\infty} df(\theta|\mathbf{x})d\theta$$

From here we want to find the decision, d, which minimizes the expected loss. The full derivation is quite lengthy and involves application of integration by parts followed by the product rule of

differentiation so we will just state the result here

$$\begin{aligned}\frac{d}{dd}\mathbb{E}\Big[L(d,\theta)|\mathbf{x}\Big] &= \int_{-\infty}^{d} f(\theta|\mathbf{x})d\theta - \int_{d}^{\infty} f(\theta|\mathbf{x})d\theta + 2df(d|\mathbf{x}) - 2df(d|\mathbf{x}) = 0\\ &= \int_{-\infty}^{d} f(\theta|\mathbf{x})d\theta - \int_{d}^{\infty} f(\theta|\mathbf{x})d\theta = 0\\ &= F(d|\mathbf{x}) - (1 - F(d|\mathbf{x})) = 0\\ &= 2F(d_0|\mathbf{x}) - 1 = 0\\ &= F(d_0|\mathbf{x}) = \frac{1}{2}\end{aligned}$$

which suggests that the decision which results in a stationary point is the posterior median. To confirm this stationary point is a minimum we must look to its second derivative

$$\frac{d^2}{dd^2}\mathbb{E}\Big[L(d,\theta)|\mathbf{x}\Big]\Big|_{d=d_0} = 2f(d_0|\mathbf{x})$$

and since a probability density function by definition is positive over its support, then we can confirm that d_0 is a minimum point

Restating the above result we see that the **posterior median** is the estimate of θ which minimizes the posterior expected absolute loss.

6.1.4 The $0-1$ Loss Function

The $0-1$ loss function has the form

$$L(d,\theta) = \begin{cases} 1, & |d-\theta| > \alpha \\ 0, & \text{otherwise} \end{cases}$$

where α is some predetermined acceptable error value. The $0-1$ loss function states that within a certain error bound, α, we incur zero loss and anything outside this error range should be penalised uniformly with error 1.

Again if we look at the expected posterior loss function and consider the case where we have loss 1, this occurs in two cases

- When $d - \theta > \alpha \longrightarrow \theta < d - \alpha$
- When $\theta - d > \alpha \longrightarrow \theta > d + \alpha$

which then gives us the bounds of integration for θ in each case

$$\begin{aligned}\mathbb{E}\Big[L(d,\theta)|\mathbf{x}\Big] &= \int_{-\infty}^{d-\alpha} (1)f(\theta|\mathbf{x})d\theta + \int_{d+\alpha}^{\infty} (1)f(\theta|\mathbf{x})d\theta + \int_{d-\alpha}^{d+\alpha} (0)f(\theta|\mathbf{x})d\theta\\ &= \int_{-\infty}^{d-\alpha} f(\theta|\mathbf{x})d\theta + \int_{d+\alpha}^{\infty} f(\theta|\mathbf{x})d\theta\\ &= F(d-\alpha|\mathbf{x}) + (1 - F(d+\alpha|\mathbf{x}))\end{aligned}$$

The next step is to use a Taylor series expansion of $F(d-\alpha|\mathbf{x})$ and $F(d+\alpha|\mathbf{x})$ around $F(d|x)$ excluding any terms of the order greater than α. This is appropriate since we expect α to be small. For $F(d+\alpha|\mathbf{x})$, the Taylor series expansion yields

$$\begin{aligned}F(d+\alpha|\mathbf{x}) &\approx F(d|x) + F'(d|\mathbf{x})(d+\alpha-d) + ...\\ &= F(d|x) + \alpha f(d|\mathbf{x})\end{aligned}$$

Similarly, the Taylor series expansion for $F(d-\alpha|\mathbf{x})$ is

$$F(d-\alpha|\mathbf{x}) \approx F(d|x) - \alpha f(d|\mathbf{x})$$

Substituting our Taylor series expansions into the formula for the expected $0-1$ loss

$$\begin{aligned}\mathbb{E}\Big[L(d,\theta)|\mathbf{x}\Big] &= F(d-\alpha|\mathbf{x}) + (1 - F(d+\alpha|\mathbf{x}))\\ &\approx (F(d|x) - \alpha f(d|\mathbf{x})) + 1 - (F(d|x) + \alpha f(d|\mathbf{x}))\\ &= 1 - 2\alpha f(d|\mathbf{x})\end{aligned}$$

and in order to minimize this term we want this to be as close to zero as possible. Since α is fixed, this means that we want the decision, d, that maximises the posterior density which is the **posterior mode**.

6.2 From Point Estimation to Interval Estimation

Sometimes we don't care to make a point estimate for a parameter but instead we will estimate an interval for $\theta \in \Omega$, which in Bayesian inference is termed a **credible interval**.

Since the general procedure of finding an optimal solution relies on minimizing the posterior expected loss, then if $\theta \in d$ we would incur no loss and if $\theta \notin d$ we should incur a non-zero loss. If the loss function is defined in this way then a trivial solution would be an interval on the whole of Ω though this is uninformative.

On the other hand, a small credible interval is much more informative than a larger credible interval. So there should be some loss associated with the width of the credible interval d so that wider intervals result in a larger loss. A general form for this loss function is

$$L(d, \theta) = c(d, \theta) + w(d)$$

where $c(d, \theta)$ is the **coverage term** and $w(d)$ is the **width**. Example functions for the coverage and width are given below, though different functions can also be used

$$c(d, \theta) = \begin{cases} 0 & \theta \in d \\ 1 & \theta \notin d \end{cases}$$
$$w(d) = \int_{\theta \in d} 1 d\theta$$

The posterior expected loss is then given by

$$\begin{aligned}
\mathbb{E}\Big[L(d, \theta)|\mathbf{x}\Big] =& \mathbb{E}\Big[c(d, \theta) + w(d)|\mathbf{x}\Big] \\
=& \mathbb{E}\Big[c(d, \theta)|\mathbf{x}\Big] + w(d) \\
=& \int_{\theta \in d} 0 \times f(d|\mathbf{x}) d\theta + \int_{\theta \notin d} 1 \times f(d|\mathbf{x}) d\theta + w(d) \\
=& \int_{\theta \notin d} f(d|\mathbf{x}) d\theta + w(d) \\
=& P(\theta \notin d) + w(d)
\end{aligned}$$

There are two main ways of minimizing the posterior expected loss:

- **A fixed width solution** where for a fixed width, $w(d)$, we find the interval, d, which minimizes $P(\theta \notin d)$
- **A fixed coverage solution** where for fixed coverage $c(d, \theta)$ we find the smallest interval d so to minimize $w(d)$

It turns out that the both of these approaches are minimized when our interval is the **highest posterior density interval (HPD interval)**. This is an interval, d, such that no point outside d has higher posterior density

$$f(d_{HPD}|\mathbf{x}) \geq f(d^*|\mathbf{x}) \quad \forall d^* \notin d_{HPD}$$

6.3 Nuisance Parameters

This subsection briefly describes how we deal with complex multivariate distributions when we are interested in just a subset of the parameters. This utilizes concepts that have already been mentioned, namely the **marginal distribution**, and a good example of nuisance parameters is in our computation of the marginal priors. The general framework is restated below for clarity as we have not done so prior.

Let $\boldsymbol{\theta} = \{\boldsymbol{\phi}^{\boldsymbol{T}}\boldsymbol{\psi}^{\boldsymbol{T}}\}$ be a d-dimensional parameter vector where $\boldsymbol{\phi}$ is a p-dimensional parameter vector which is of interest and $\boldsymbol{\psi}$ is a $(d-p)$-dimensional vector of nuisance parameters. Then the marginal posterior distribution for $\boldsymbol{\phi}$ is given by

$$f(\boldsymbol{\phi}|\mathbf{x}) = \int_{\boldsymbol{\psi}\in\mathcal{A}_{\boldsymbol{\psi}}} f(\boldsymbol{\phi},\boldsymbol{\psi}|\mathbf{x})d\boldsymbol{\psi}$$

where integration over the support $\mathcal{A}_{\boldsymbol{\psi}}$ of the posterior for $\boldsymbol{\psi}$ and is of dimension $d-p$

6.4 Predictive Inference

Often we care about making predictions about future observations, rather than just the parameters and we can do so utilising the predictive distribution $f(\mathbf{y}|\mathbf{x})$.

In this setting, there are two unknowns: the future observations $\mathbf{Y}$ and the parameters $\boldsymbol{\theta}$ and these unknowns can be reflected by a joint posterior distribution $f(\mathbf{y},\boldsymbol{\theta}|\mathbf{x})$ and since we are only interested in predicting future observations we can treat $\boldsymbol{\theta}$ as a nuisance parameter and just consider the marginal posterior for $\mathbf{y}$

$$f(\mathbf{y}|\mathbf{x}) = \int_{\boldsymbol{\theta}\in\Omega} f(\mathbf{y},\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta}$$

Note that since we assume $\boldsymbol{\theta}$ is continuous then the integration is applicable even when $\mathbf{Y}$ is discrete.

When we assume that the future observations $\mathbf{Y}$ are independent of the observed data $\mathbf{X}$, then we can simplify our expression further

$$\begin{aligned} f(\mathbf{y}|\mathbf{x}) &= \int_{\boldsymbol{\theta}\in\Omega} f(\mathbf{y},\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}\in\Omega} f(\mathbf{y}|\boldsymbol{\theta},\mathbf{x})f(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}\in\Omega} f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta} \end{aligned}$$

where $f(\boldsymbol{\theta}|\mathbf{x})$ is the posterior distribution for the parameter given the data, which we have been working with previously and $f(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood of $\boldsymbol{\theta}$ given $\mathbf{y}$.

Inference for $\mathbf{Y}$ can be conducted in the same manner as for $\boldsymbol{\theta}$ - namely we decide on a loss function and solve for $\mathbf{y}$ so to minimize the posterior expected loss $\mathbb{E}\Big[L(d,y)\Big]_{\mathbf{Y}|\mathbf{X}}$

Example 2 Suppose that x is the number of successes in a sequence of n independent trials and that a conjugate Bayesian analysis using the beta / Binomial model resulted in a Beta posterior distribution for $\theta|x$ with density

$$f(\theta|x) = \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)}\theta^{a+x-1}(1-\theta)^{b+n-x-1} \quad 0<\theta<1$$

Then suppose that we are interested in making m future observations of the same model and we are interested in Y the future number of successes. This means that the likelihood $f(y|\theta)$ follows a binomial distribution so that

$$f(y|\theta) = \binom{m}{y}\theta^y(1-\theta)^{m-y}$$

Then the predictive distribution is then given by

$$\begin{aligned}
f(y|x) =& \int_{\theta\in\Omega} f(y|\theta)f(\theta|x)d\theta \\
=& \int_{\theta\in\Omega} \binom{m}{y}\theta^{y}(1-\theta)^{m-y}\theta^{a+x-1}(1-\theta)^{b+n-x-1}d\theta \\
=& \binom{m}{y}\frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)}\int_{\theta\in\Omega}\theta^{a+x+y-1}(1-\theta)^{b+n+m-y-x-1}d\theta \\
=& \binom{m}{y}\frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)}\frac{\Gamma(a+x+y)\Gamma(b+n+m-y-x)}{\Gamma(a+b+n+m)} \\
\propto& \frac{\Gamma(a+x+y)\Gamma(b+n+m-y-x)}{m!(m-y)!}
\end{aligned}$$

Note that this is a discrete distribution with a non-standard form. However, we can compare the odds of the binary outcomes quite easily by considering the ratio of an event $Y = 1$ to its complement $Y \neq 1$

$$\frac{P(Y=1|x)}{P(Y=0|x)} = \frac{\Gamma(a+x+1)\Gamma(b+n+m-1-x)}{\Gamma(a+x)\Gamma(b+n+m-x)}$$

If say, $m = 1$, then this simplifies to

$$\frac{P(Y=1|x)}{P(Y=0|x)} = \frac{a+x}{b+n-x}$$

7 Bayes Factors

The Bayes factor for an event A against its complement A^c is denoted $B(\mathbf{x}, A)$ and is given by

$$B(\mathbf{x}, A) = \frac{f(\mathbf{x}|A)}{f(\mathbf{x}|A^c)}$$

The Bayes factor measures the odds for an event given the data - it is a ratio of the likelihoods. If we consider Bayes theorem again, and solve for the likelihood then we can also interpret the Bayes factor as the ratio of the posterior and prior odds

$$\begin{aligned}
f(A|\mathbf{x}) &\propto f(\mathbf{x}|A)f(A) \\
f(\mathbf{x}|A) &\propto \frac{f(A|\mathbf{x})}{f(A)}
\end{aligned}$$

Then we can rewrite $B(x, A)$ as ratio of the posterior and prior odds

$$B(\mathbf{x}, A) = \frac{f(A|\mathbf{x})}{f(A^c|\mathbf{x})}\frac{f(A^c)}{f(A)}$$

7.1 Model Comparison

The Bayes factor is a useful quantity when we are looking for evidence for and against a belief. For example, A could denote the belief of the value of a parameter $\theta = \theta_1$ or it could denote the belief that the data come from a specific distribution $f_1(x|\theta)$. Simply evaluate the ratio of the likelihoods under the different scenarios.

8 Appendix

8.1 The normal/normal model

If we assume that μ is the parameter of interest with σ_0^2 is known, then the density of the likelihood is just a function of μ

$$f(\mathbf{x}|\mu) = \frac{1}{(2\pi\sigma_0^2)^{n/2}}\exp\Big[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big]$$

then we can rewrite

$$\sum_{i=1}^{n}(x_i-\mu)^2=\sum_{i=1}^{n}(x_i+\bar{x})^2+n(\bar{x}-\mu)^2$$

then if we consider only the density core (so excluding terms not involving μ) we can rewrite this as

$$L(\mu)\propto\frac{1}{(2\pi\sigma_0^2)^{n/2}}\exp\Big[-\frac{n}{2\sigma_0^2}(\bar{x}-\mu)^2\Big]$$

We know that μ can take any value on the real line so a normal prior is appropriate here and is of the form $\mathrm{N}(a,b)$ distribution

$$f(\mu)=\frac{1}{\sqrt{2}}\exp\Big[-\frac{1}{2b}(\mu-a)^2\Big]$$

So then the posterior is computed using Bayes theorem and yields

$$\begin{aligned}f(\mu|\mathbf{x})\propto&\exp\Big[-\frac{n}{2\sigma_0^2}(\bar{x}-\mu)^2-\frac{1}{2b}(\mu-a)^2\Big]\\=&\exp\Big[-\frac{nb}{2b\sigma_0^2}(\bar{x}^2-2\bar{x}\mu+\mu^2)-\frac{\sigma_0^2}{2b\sigma_0^2}(\mu^2-2a\mu+a^2)\Big]\end{aligned}$$

Then if we remove any terms not involving μ we can rewrite this as

$$\begin{aligned}f(\mu|\mathbf{x})\propto&\exp\Big[-\frac{1}{2b\sigma_0^2}\Big((nb+\sigma_0^2)\mu^2-2\mu(\bar{x}nb+2a\sigma_0^2)\Big)\Big]\\=&\exp\Big[-\frac{1}{2b_1}\big(\mu^2-2\mu a_1\big)\Big]\end{aligned}$$

where

$$a_1=\frac{\bar{x}nb+2a\sigma_0^2}{nb+\sigma_0^2}\quad\text{and}\quad b_1=\frac{\sigma_0^2 b}{nb+\sigma_0^2}$$

then we can complete the square so that we get a density of the form (note that the above form ignores the term only in a_1)

$$f(\mu|\mathbf{x})\propto\exp\Big[-\frac{1}{2b_1}(\mu-a_1)^2\Big]$$

which tells us that the posterior density for the normal/normal model with known variance σ_0^2 is a $\mathrm{N}(a_1,b_1)$

8.2 The gamma/normal model

In the case where random variables are modelled by a normal distribution with known mean μ_0 and unknown precision τ, then the log likelihood is

$$l(\tau)=C_1-n\log\Big(\frac{1}{\sqrt{\tau}}\Big)-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2$$

meaning that the likelihood can be written as

$$L(\tau)\propto\tau^{n/2}\exp\Big[-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\Big]$$

Then the prior we use for τ is one which can only take positive values, since the variance and hence precision must be greater than zero. The gamma(c,d) distribution turns out to be the appropriate prior with density

$$\tau^{c-1}\exp\big[-d\tau\big]$$

Then the posterior is given by Bayes theorem

$$\begin{aligned} f(\tau|\mathbf{x}) \propto & \tau^{n/2} \exp\Big[-\frac{\tau}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \Big] \times \tau^{c-1} \exp\big[-d\tau \big] \\ & = \tau^{n/2+c-1} \exp\Big[-\tau \big(\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 + d \big) \Big] \end{aligned}$$

which as we expect is a Gamma(c_1, d_1) distribution where

$$c_1 = n/2 + c \quad \text{and} \quad d_1 = \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 + d$$