Lecture Notes

MSH3 – Advanced Bayesian Inference

Lecturer

Dr. John T. Ormerod

School of Mathematics & Statistics F07

University of Sydney

(w) 02 9351 5883

(e) john.ormerod(at) sydney.edu.au

Outline of MSH3 – Monte Carlo Methods

- ☐ The Bayesian Inferential Paradigm
- ☐ Random Variable Generation
- ☐ Monte Carlo Integration
- ☐ Markov Chains
- ☐ Markov Chain Monte Carlo
- ☐ Analytic Approximations

Lecture 1 - Content

- Bayesian inference preliminaries
- Bayesian versus Frequentist inference
- Examples
- Priors
- Bayesian point estimation, confidence intervals and model selection

Bayesian Inference Preliminaries

Almost all of methods at the University of Sydney concern what is referred to as *frequentist inference*. A major alternative to frequentist inference is *Bayesian inference* named after Reverend Thomas Bayes (1701–1761).

For much of the 20th century Bayesian inference was heavily criticised, initially most prominently by Fisher (grand-daddy of Statistics).

In the early years of Statistics there was a non-negligible probability that if Bayesian statistician and a frequentist statistician were in the same room then a fist fight would ensue.

Several landmark papers in the 1980s and early 1990s showed that Markov Chain Monte Carlo methods were applicable to many Bayesian problems making many practical problems computationally tractable.

Reverend Thomas Bayes (1701 - 1761)

- Born in Hertfordshire (London, England),
- was a Presbyterian minister,
- studied: theology and mathematics,
- best known for Essay Towards Solving a Problem in the Doctrine of Chances
- where Bayes' Theorem was first proposed.
- Words: Bayes' rule, Bayes' Theorem, Bayesian Statistics.

In frequentist inference the properties of estimators are judged on their behaviour if the experiment leading to the recorded data were repeated over and over.

Wasserman (2003) states the following postulates of frequentist inference:

F1. Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.

F2. Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameters.

F3. Statistical procedure should be designed to have well-defined long run frequency properties. For example, a 95% confidence interval should trap the true vale of the parameter with a limiting frequency at least 95%.

An alternative, perhaps as popular, method for conducting statistical inferences is *Bayesian Inference*. Bayesian inferential procedures are often far simpler to develop and use and have been applied to problems where equivalent frequentist procedures would be difficult to use.

Wasserman (2003) states the following postulates of Bayesian inference:

B1. Probability describes degree of belief, not limiting frequency. As such we can make probability statements about lot of things, not just data which are subject to variation. For example, I might say that "the probability that Albert Einstein drank a cup of tea on August 1, 1948" is 0.35. This does not refer to any limiting frequency. If reflects my strength of belief that the proposition is true.

B2. We can make probability statements about parameters, even though they are fixed constants.

B3. We make inferences about a parameter θ by producing the probability distribution for θ. Inferences, such as point estimates and interval estimates, may be extracted from this distribution.

The Bayes Song

There's no Theorem like Bayes Theorem
Like no theorem we know
Everything about it is appealing
Every thing about it is a wow
Let out that a priori feeling
You've been concealing right up to now.
...
There's no Theorem like Bayes Theorem
Like no theorem we know

George E. P. Box (1919–)

From the above statements we can see that the reason that Bayesian inference is so controversial because is because it adopts a subjective notion of probability. Consequently, Bayesian inferences may be considered subjective rather than objective and subject to the user's biases.

Bayesian methods do not make guarantees on the long term performance of procedures.

Putting philosophical issues aside Bayesian methods for many aspects of inference have been developed analogously to those developed for frequentist inference. Hence, Bayesian methodology has its own versions of point estimation, confidence intervals and hypothesis testing. Lets now get down to how Bayesian inference is performed.

Bayesian Inference

Bayesian inference consists of calculating a distribution or distributions that describe the parameters of a model. As one might expect this is determined via Bayes theorem which, we remind the reader, may be stated, for discrete random variables X, as:

$$\begin{aligned}\mathbb{P}(\Theta = \theta_i | X = x) &= \frac{\mathbb{P}(X = x | \Theta = \theta_i) \times \mathbb{P}(\Theta = \theta_i)}{\sum_{i=1}^{k} \mathbb{P}(X = x | \Theta = \theta_i) \times \mathbb{P}(\Theta = \theta_i)} \\ &= \frac{\mathbb{P}(X = x, \Theta = \theta_i)}{\mathbb{P}(X = x)}\end{aligned}$$

where $\Theta \in \{\theta_1, \ldots, \theta_k\}$.

As one might expect the version for continuous variables uses density functions in place of probabilities,

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{\int f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)d\theta} = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)}. \tag{1}$$

where the joint distribution is defined by

$$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)$$

and the marginal distribution of X is defined by

$$f_X(x) = \int f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)d\theta = \int f_{X,\Theta}(x,\theta)d\theta. \tag{2}$$

We will use the convention that all integrals without limits denote integrals over the range of all allowable values of the integrated parameter. For example, if $X \sim$ Uniform$(0,1)$, then $\int_0^1 f_X(x)dx \equiv \int f_X(x)dx$.

Note also that instead of $f_X(x;\theta)$ we have written $f_{X|\Theta}(x|\theta)$. This reflects the Bayesian philosophy that parameters, in this case θ, are random quantities.

Ingredients of Bayesian Inference

There are various ingredients used to calculate the above probabilities using Bayes theorem. Attached to these ingredients is the following terminology:

1. We choose a statistical model $f_{X|\Theta}(x|\theta)$, the *model distribution*, that reflects our beliefs about x given a particular value of θ.
2. We choose a probability density $f_{\Theta}(\theta)$, called a *prior distribution*, that expresses our beliefs about a parameter θ, before we have seen any data.
3. After observing data $x_1, \ldots, x_n$, we update or modify our beliefs according to this observed data. Mathematically this amounts to calculating the *posterior distribution*, which, using Bayes theorem, is given by

$$f_{\Theta|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n) = f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{X},\Theta}(\mathbf{x},\theta)}{f_{\mathbf{X}}(\mathbf{x})}$$

Note that $f_{\mathbf{X},\Theta}(\mathbf{x},\theta)$ is referred to as the *joint distribution* while

$$f_{\mathbf{X}}(\mathbf{x}) = \int f_{\mathbf{X},\Theta}(\mathbf{x},\theta)d\theta$$

is referred to as the *marginal distribution*. If we have n iid observations $x_1,\ldots,x_n$, then we replace $f_{X|\Theta}$ in (1) with

$$f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_1|\theta) = \prod_{i=1}^{n} f_{X|\Theta}(x_i|\theta) = \mathcal{L}_n(\theta)$$

where $\mathcal{L}_n(\theta)$ is the familiar likelihood function. The posterior density can then be re-expressed as:

$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{\mathcal{L}_n(\theta)f_\Theta(\theta)}{\int \mathcal{L}_n(\theta)f_\Theta(\theta)d\theta} = \frac{\mathcal{L}_n(\theta)f_\Theta(\theta)}{c_n} \propto \mathcal{L}_n(\theta)f_\Theta(\theta)$$

where $c_n = \int \mathcal{L}_n(\theta)f_\Theta(\theta)d\theta$ is also called the *normalising constant* of the posterior distribution. Note that while c_n does depend on $\mathbf{x}$ it does not depend on θ.

The main practical difficulty in Bayesian inference is either in the calculation of the normalising constant c_n (which may be used to obtain the posterior distribution) or simulating values from $f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$. We will consider these issues later.

In order to simplify notation we will, from now on, remove subscripting of densities, i.e. $f_{X|\Theta}(x|\theta)$ will be written as $f(x|\theta)$. We will rely on context to remove ambiguity, for example, the functions $f_{X|\Theta}(x|\theta) \equiv f(x|\theta)$ and $f_X(x) \equiv f(x)$ are different densities. This convention is used in almost all papers, books and other literature in Bayesian statistics.

Example [Bernoulli Model with Uniform Prior]: Let

$$X_1, \ldots, X_n \sim \text{Bernoulli}(p).$$

Suppose that we use $f(p) = 1$, $0 \leq p \leq 1$, i.e. the Uniform$(0, 1)$ distribution, as our prior for p. Suppose we then observe the values $x_1, \ldots, x_n$, then,

$$f(p|\mathbf{x}) \propto \mathcal{L}_n(p) f(p) = p^s (1-p)^{n-s} = p^{(s+1)-1} (1-p)^{(n-s+1)-1} \tag{3}$$

where $s = \sum_{i=1}^{n} x_i$ is the number of successes. By Bayes' theorem, the posterior has the form

$$f(p|\mathbf{x}) = \frac{p^s (1-p)^{n-s}}{\int_0^1 p^s (1-p)^{n-s} dp} \tag{4}$$

since any constants with respect to p cancel from the numerator and denominator. Now, recall that a random variable has a Beta distribution with parameters α and β if its probability density function is given by

$$f(p; \alpha; \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1} (1-p)^{\beta - 1}.$$

Hence, since $f(p;\alpha;\beta)$ is a density, i.e. $\int_0^1 f(p;\alpha;\beta)dp = 1$ for all α, β we have

$$\int_0^1 p^{\alpha-1}(1-p)^{\beta-1} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

for all α, β. So, if $\alpha = s+1$ and $\beta = n-s+1$ we have

$$f(\mathbf{x}) = \int_0^1 p^s(1-p)^{n-s}dp = \frac{\Gamma(s+1)\Gamma(n-s+1)}{\Gamma(n+2)},$$

so that

$$f(p|\mathbf{x}) = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)}p^{(s+1)-1}(1-p)^{(n-s+1)-1}.$$

or

$$p|\mathbf{x} \sim \text{Beta}(s+1, n-s+1).$$

Alternatively, we can see by examining (3) that $f(p|\mathbf{x}) \propto p^{(s+1)-1}(1-p)^{(n-s+1)-1}$ is proportional to a Beta density with parameters $\alpha = s+1$ and $\beta = n-s+1$. Using this method we obtain precisely the same solution without actually having to calculate the integral $\int_0^1 p^{\alpha-1}(1-p)^{\beta-1}$.

Example [Bernoulli Model with Beta Prior]: Consider the previous example. Suppose that in the example that instead of using a uniform prior for p, we use the prior $p \sim \text{Beta}(\alpha, \beta)$. If you repeat the calculations used in the example, see Exercises, you will see that

$$p|\mathbf{x} \sim \text{Beta}(\alpha + s, \beta + n - s).$$

Note that the uniform prior is a special case of a Beta prior with $\alpha = \beta = 1$.

Example [Poisson Model with Gamma Prior]: Let x follow a Poisson distribution with rate θ so that

$$f(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad \text{for} \;\; x = 0, 1, 2, \ldots,$$

and for a vector of iid observations $\mathbf{x} = (x_1, \ldots, x_n)$, the likelihood function is

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} \propto \theta^s e^{-n\theta}$$

where $s = \sum_{i=1}^{n} x_i$. Suppose that we use the Gamma distribution as a prior for θ parameter, i.e. $\theta \sim \text{Gamma}(\alpha, \beta)$ for some strictly positive constants α and β, so

that

$$f(\theta;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}.$$

Note that

$$f(\theta|\mathbf{x}) = \frac{\mathcal{L}_n(\theta)f_\Theta(\theta)}{\int_0^\infty \mathcal{L}_n(\theta)f_\Theta(\theta)d\theta} = \frac{\theta^{\alpha+s-1}e^{-(\beta+n)\theta}}{\int_0^\infty \theta^{\alpha+s-1}e^{-(\beta+n)\theta}d\theta} \tag{5}$$

Now, since $f(\theta;\alpha,\beta)$ is a density for all α, β we have $\int_0^\infty \theta^{\alpha-1}e^{-\beta\theta} = \frac{\Gamma(\alpha)}{\beta^\alpha}$ for all α, β. Hence,

$$f(\mathbf{x}) = \int_0^\infty \theta^{\alpha+s-1}e^{-(\beta+n)\theta} = \frac{\Gamma(\alpha+s)}{(\beta+n)^{\alpha+s}}$$

so that

$$f(\theta|\mathbf{x}) = \frac{(\beta+n)^{\alpha+s}}{\Gamma(\alpha+s)}\theta^{\alpha+s-1}e^{-(\beta+n)\theta}, \quad \theta > 0,$$

or $\theta|\mathbf{x} \sim \text{Gamma}(\alpha+s,\beta+n)$. Again, we can see by examining (5) that $f(\theta|\mathbf{x})$ is proportional to a Gamma density with parameters $\widetilde{\alpha} = \alpha + s$ and $\widetilde{\beta} = \beta + n$.

Example [Normal with Known Variance and Normal Prior on the Mean]: Let $X_1, \ldots, X_n|\theta \sim N(\theta, \sigma^2)$ where, for simplicity, we assume that σ^2 is known. Suppose we take as a prior $\theta \sim N(a, b^2)$. It can be shown, see Exercises, that the posterior for θ is

$$\theta|\mathbf{x} \sim N(\overline{\theta}_n, \tau^2)$$

where

$$\overline{\theta}_n = w\overline{x} + (1-w)a, \quad w = \frac{\frac{1}{\mathsf{se}^2}}{\frac{1}{\mathsf{se}^2} + \frac{1}{b^2}}, \quad \frac{1}{\tau^2} = \frac{1}{\mathsf{se}^2} + \frac{1}{b^2}$$

and $\mathsf{se} = \sigma/\sqrt{n}$ is the standard error of the maximum likelihood estimator $\overline{x}$.

Priors

An important question which arises in Bayesian inference is: how does one choose the prior distribution $f(\theta)$? As with many aspects of Statistics there are several ways, and reasons for choosing, different prior distributions. Prior distributions may be categorised into several different, potentially overlapping, categories. We will cover the following types of priors:

- Conjugate
- Informative
- Non-informative
- Improper
- and Jeffery's (objective) prior.

Priors can also be used to address identifiability problems of parameters and can be used to induce desirable properties in Bayesian estimators.

Conjugate Priors

One of the key difficulties in Bayesian inference is the calculation of the posterior distribution. The calculation of the posterior distribution involves the calculation on an integral. In the examples we have covered so far all of the expression for the integrals required in the calculation of the posterior distribution were known, i.e. the integrals were tractable.

The reason the posterior distributions were calculable in the examples we have seen so far is due to a special relationship between the model distributions and the prior distributions. This relationship is referred to as *conjugacy*. Two distributions are said to be conjugate if the model distribution and the prior distribution share a common functional form. More formally:

Definition: If $\mathcal{F}$ is a class if model distributions $f(\mathbf{x}|\theta)$, and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ so *conjugate* for $\mathcal{F}$ if

$$f(\theta|\mathbf{x}) \in \mathcal{P} \quad \text{for all} \quad f(\cdot|\theta) \in \mathcal{F} \quad \text{and} \quad f(\cdot) \in \mathcal{P}.$$

This definition is purposefully vague: Choosing $\mathcal{P}$ to be the class of all distributions then $\mathcal{P}$ is always conjugate to $\mathcal{F}$ regardless of which class of model distributions are used. More interesting are *natural* conjugate prior families, which arise by taking $\mathcal{P}$ to be the set of all densities having the same functional form as the likelihood.

Conjugate prior distributions have the practical advantage of computational convenience and of being interpretable as additional data. It can be shown that, in general, the exponential family of distributions are the only class of distributions that have natural conjugate prior distributions, since apart from certain irregular cases, the only distributions having a fixed number of sufficient statistics for all n are of the exponential type.

Definition: The class $\mathcal{F}$ is an exponential family if all its members are of the form

$$f(x_i|\theta) = a(x_i)b(\theta)\exp\left\{c(\theta)^T d(x_i)\right\}.$$

The factors $c(\theta)$ and $d(x_i)$ are, in general, vectors of equal dimension to that of θ whereas the functions $a(x_i)$ are $b(\theta)$ scalar function. The vector $c(\theta)$ is called the 'natural parameter' of the family $\mathcal{F}$.

The likelihood corresponds to a sequence $\mathbf{x} = (x_1, \ldots, x_n)$ of iid observations is

$$\mathcal{L}_n(\theta) = \left[\prod_{i=1}^{n} a(x_i)\right] b(\theta)^n \exp\left\{c(\theta)^T \left[\sum_{i=1}^{n} d(x_i)\right]\right\}.$$

For all n and x, this has a fixed form (as a function of θ):

$$\mathcal{L}_n(\theta) \propto b(\theta)^n \exp\left\{c(\theta)^T t(\mathbf{x})\right\} \quad \text{where} \quad t(\mathbf{x}) = \sum_{i=1}^{n} d(x_i).$$

where $t(\mathbf{x})$ is a sufficient statistic for θ, because the likelihood for θ depends on the data $\mathbf{x}$ only through the value of $t(\mathbf{x})$.

Suppose that the prior density is of the form $f(\theta) \propto b(\theta)^\eta \exp\left\{c(\theta)^T \boldsymbol{\nu}\right\}$ for some constants η and $\boldsymbol{\nu}$, then the posterior density is of the form

$$f(\theta|\mathbf{x}) \propto b(\theta)^{\eta+n} \exp\left\{c(\theta)^T (\boldsymbol{\nu} + t(\mathbf{x}))\right\}$$

which shows that this choice of prior is conjugate. Hence, if the normalising constant of the prior is known, then the normalising constant for the posterior density will also be known. **This of great practical importance!**

Unfortunately, only a small number of model/prior distribution combinations are conjugate. Some of the most common model/prior distribution combinations are listed below:

Likelihood ($x_1, \ldots, x_n \sim$)	Prior	Posterior ($\theta \vert \mathbf{x}$)
Bernoulli(θ)	$\theta \sim$ Beta(α, β)	Beta($\alpha + n\overline{x}, \beta + n - n\overline{x}$)
Poisson(θ)	$\theta \sim$ Gamma(α, β)	Gamma($\alpha + n\overline{x}, \beta + n$)
$N(\theta, \tau^{-1})$ with τ known	$\theta \sim N(b, c^{-1})$	$N\left(\frac{cb+n\tau\overline{x}}{c+n\tau}, \frac{1}{c+n\tau}\right)$
Gamma(k, θ) with k known	$\theta \sim$ Gamma(α, β)	Gamma($\alpha + nk, \beta + n\overline{x}$)

although many other conjugacy combinations are available.

In practice keeping to models which use conjugate distribution can be quite limiting. A prior distribution which is not conjugate is called a *non-conjugate prior*. Unfortunately, using a non-conjugate prior, even if it has a simple functional form, can lead to awkward numerical problems as the following example demonstrates.

Example [Non-conjugate Prior]: Suppose, for example, that $X_1, \ldots, X_n \sim$ Poisson(θ) random variables, and our beliefs are that θ definitely lies in the range $[1, 2]$, i.e. there is a Uniform$(1, 2)$ prior on θ, with $f(\theta) = \log(\theta)/(2\log(2) - 1)$, $1 \leq \theta \leq 2$. Then the normalising constant is

$$\int_1^2 \frac{\log(\theta)}{2\log(2) - 1} \exp\{-n\theta\} \theta^{n\overline{x}} d\theta.$$

This integral can only be evaluated numerically, for example by using the trapezoid rule, Simpson's rule or some other method.

Informative Priors

One school of thought, called **subjectivism** says that the prior should reflect our subjective opinion about θ (before that data is collected).

There are some cases where there is little alternative but to use subjective priors. For example, there are some cases where expert opinion suggests that certain crimes, for example sexual assault, are under-reported. In such circumstances relying on the data alone may provide misleading results to those where priors are developed using expert opinion.

While this may be possible or even desirable in some cases but it is impractical in complicated problems, especially if there are many parameters in the model. Moreover, injecting subjective opinion into the analysis is contrary to the goal of making scientific inference as objective as possible.

Hence, an obvious alternative is to try to define some sort of "noninformative prior".

Noninformative and Improper Priors

Noninformative priors come under a variety of names. Here "noninformative" is used as to mean the opposite of informative. However, as we will later see, there is a sense in which all priors are informative and so researchers often use synonyms for noninformative such as flat, diffuse or vague to described priors of this type.

The goal of noninformative priors is to choose the prior $f(\theta)$ so that no value of θ has a larger value of $f(\theta)$ than any other value, i.e. we choose the prior $f(\theta)$ so that θ is selected as objectively as possible by the data. This leads to the flat prior

$$f(\theta) \propto \text{constant}.$$

In the Bernoulli example, taking $f(p) = 1$ leads to $p|\mathbf{x} \sim \text{Beta}(s+1, n-s+1)$. After looking at some of the properties of the posterior this choice seemed to give reasonable results.

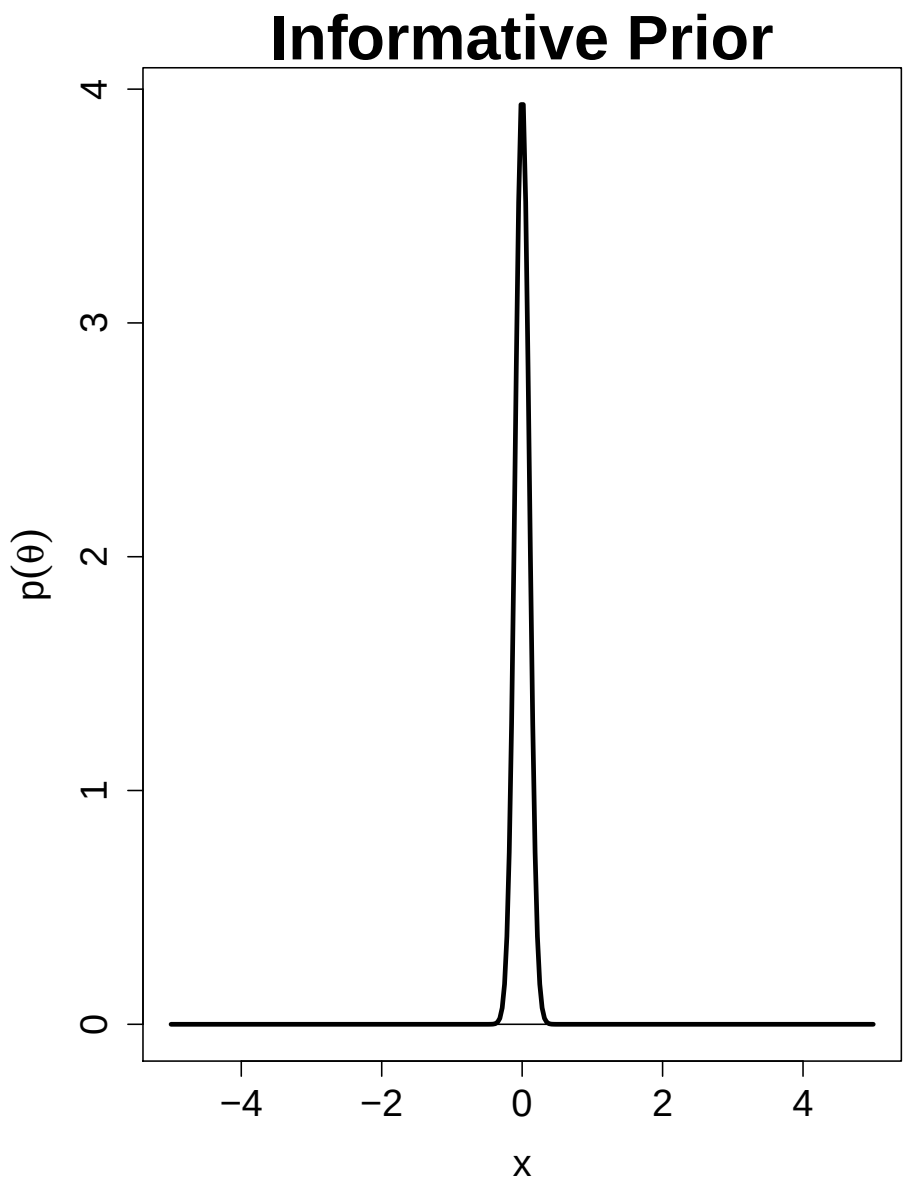
However, for more general problems, in particular when the domain of θ is not bounded, the choice $f(\theta) \propto$ constant is called an *improper prior* since

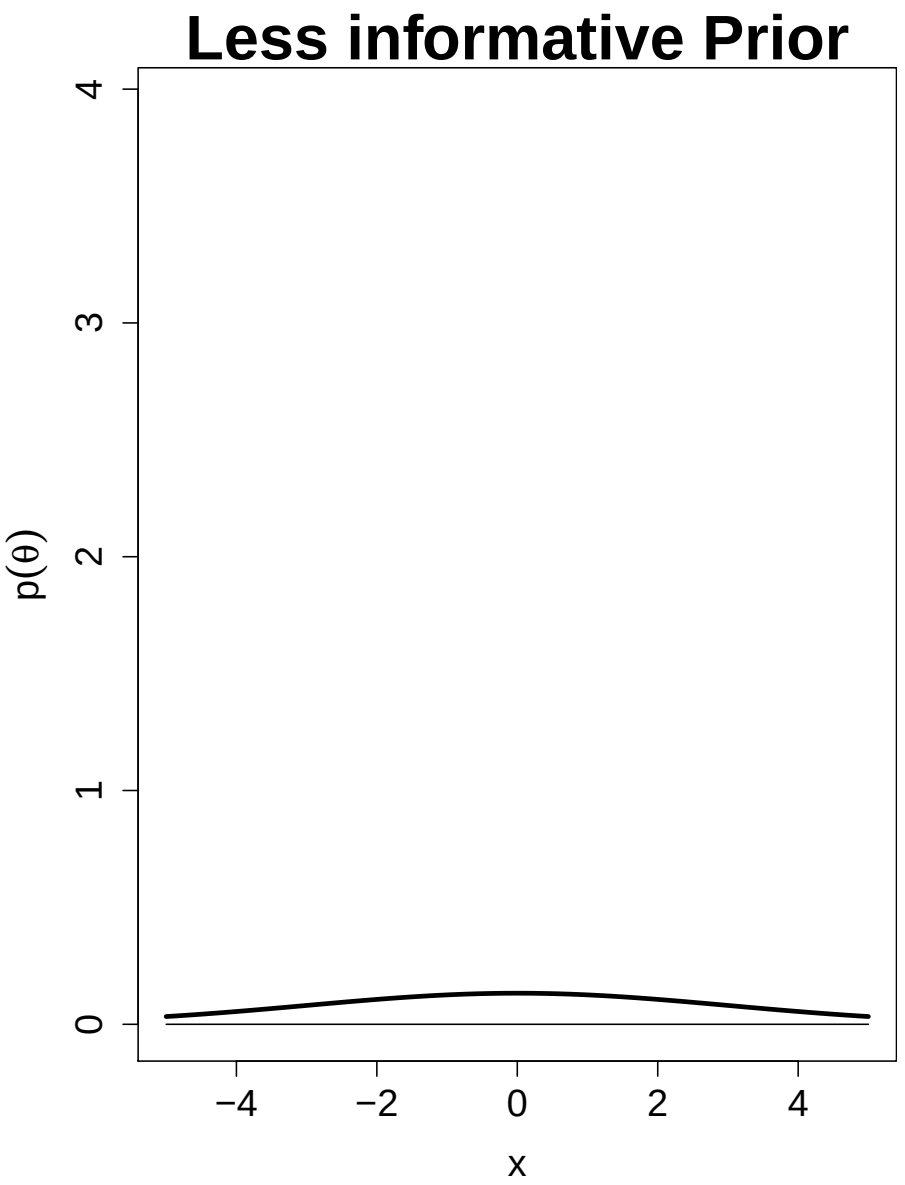
$$\int f(\theta) d\theta \quad \text{is not defined}$$

and as such $f(\theta)$ is not a probability density in the usual sense. Nevertheless, we can still formally carry out Bayes theorem and compute the posterior density by multiplying the prior by the likelihood and calculating the normalising constant, i.e.

$$f(\theta|\mathbf{x}) = \frac{\mathcal{L}_n(\theta) f(\theta)}{\int \mathcal{L}_n(\theta) f(\theta) d\theta} = \frac{\mathcal{L}_n(\theta)}{\int \mathcal{L}_n(\theta) d\theta}.$$

However, the density $f(\theta|\mathbf{x})$ may or may not be a proper density. The following two examples demonstrate this problem.


Informative Prior
p(θ)
0
1
2
3
4
−4
−2
0
2
4
x


Less informative Prior
p(θ)
0
1
2
3
4
−4
−2
0
2
4
x

Example: Let $X_1, \ldots, X_n \sim$ Bernoulli(p). Suppose we use the prior $p \sim$ Beta$(0, 0)$, i.e.

$$f(p) \propto p^{-1}(1-p)^{-1}, \ 0 \leq p \leq 1.$$

then it is easy to verify, see Exercises, that $f(p)$ is no longer a proper density. Furthermore, it can be show that when $s = 0$ or $s = n$ then the posterior distribution is also improper.

Example: Suppose that $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ where the variance σ^2 is known. Suppose that we use the improper prior $f(\theta) \propto 1$. It can be shown, see Exercises, that the posterior distribution is given by

$$\theta|\mathbf{x} \sim N\left(\overline{\mathbf{x}}, \frac{\sigma^2}{n}\right).$$

This is the distribution of the maximum likelihood estimator for a normal sample when the variance is known.

If the resulting posterior is not a well defined probability function alternative priors should be sought. One such alternative is to seek a proper diffuse or vague prior. We saw in the example of a normal model with known variance and a normal prior on the mean that as $b^2 \to \infty$ we obtained $\theta|\mathbf{x} \sim N(\widehat{\theta}_n, \text{se}^2)$ the distribution of the maximum likelihood estimator.

Hence, for continuous parameters defined on the whole real line, one option is

$$\theta \sim N(a, b^2)$$

for some constant a and a suitably large value of b^2. For positive continuous parameters is is common to use

$$\theta \sim \text{Gamma}(a, b) \qquad \text{or equivalently} \qquad \theta^{-1} \sim \text{Inverse-Gamma}(a, b).$$

where a and b are small positive constants, i.e. 10^{-4}.

However, for either case many other proper diffuse or vague priors could be used in practice.

Jeffery's Priors

Jeffery's priors arise due to the following problem in Bayesian inference: priors are not, in general, transformation invariant. This is demonstrated in the following example.

Example: Let $X \sim$ Bernoulli(p) and suppose we use the flat prior $p \sim$ Uniform(0,1). This flat prior presumably represents our lack of information about p before the experiment. Now let $\psi = \log(p/(1-p))$. This is a transformation of p and the resulting distribution of ψ is given by

$$f(\psi) = \frac{e^{\psi}}{(1+e^{\psi})^2}$$

which is the (standard) *logistic distribution*. The distribution $f(\psi)$ is not flat. But if we are ignorant about p then we should also be ignorant of ψ. Hence, we should be using a flat prior for ψ. This is a contradiction. In short, the notion of a flat prior is not well defined because a flat prior on a parameter does not imply a flat prior on a transformed version of the parameter, i.e. flat priors are not *transformation invariant*.

This is a property which is not shared my maximum likelihood estimators and is one of the criticisms levelled against Bayesian inference.

Jeffery came up with a rule which avoids this problem by choosing

$$f(\theta) \propto \mathcal{I}(\theta)^{1/2}$$

where $I(\theta)$ is the Fisher information function for θ. This rule turns out to be transformation invariant. There are other various reasons for thinking that this prior might be a useful prior but we will not go into details here.

Example: Consider the Bernoulli model with parameter p. It can be shown that, see Exercises, that

$$I(p) = \frac{1}{p(1-p)}.$$

Hence, Jefferys' rule says to use the prior

$$f(p) \propto \sqrt{I(p)} = p^{-1/2}(1-p)^{-1/2}.$$

This turns out to be a Beta$(1/2, 1/2)$ density and is very close to the uniform density.

Note for multiparameter problems, i.e. where θ is not a one-dimensional parameter, the Jefferys' prior is defined to be

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

where $|\mathbf{A}|$ denotes the determinant of a square matrix $\mathbf{A}$ and $I(\boldsymbol{\theta})$ is the Fisher information matrix.

Using the Posterior Distribution

Now that we have seen a few examples of how to calculate Posterior distributions you might be asking: why is the posterior distribution so important?

In Bayesian inference the posterior distribution is, roughly speaking, analogous to the distribution of an estimator $\widehat{\theta}$ in frequentist inference.

We can use the posterior distribution to:

- Summarise various aspects of the posterior distribution of a parameter. For example, to calculate the mean, mode and variance.
- Calculate "confidence intervals" for a parameter of interest.
- To perform model selection or "hypothesis testing".

Bayesian Point Estimation

There are a number of ways to summarise features of the posterior distribution. For example, centre of the posterior distribution may be summarised my its mean, mode or median. Each of these could be used as a Bayesian point estimate. The posterior mean by be written as:

$$\overline{\theta}_n = \mathbb{E}(\theta|\mathbf{x}) = \int \theta f(\theta|\mathbf{x}) d\theta = \frac{\int \theta \mathcal{L}_n(\theta) f(\theta) d\theta}{\int \mathcal{L}_n(\theta) f(\theta) d\theta}.$$

whereas the posterior mode is given by

$$\widetilde{\theta}_n = \text{argmax}_\theta \left\{ f(\theta|\mathbf{x}) \right\}.$$

Either $\overline{\theta}_n$ or $\widetilde{\theta}_n$ might be thought of as Bayesian point estimators. The spread of the posterior distribution might be summarised, for example, by its variance or standard deviation. The posterior variance by be written as:

$$\text{Var}(\theta|\mathbf{x}) = \int (\theta - \mathbb{E}(\theta|\mathbf{x}))^2 f(\theta|\mathbf{x}) d\theta.$$

Example [Bernoulli Model with Uniform Prior – Continued]: Reminder:

$$p|\mathbf{x} \sim \text{Beta}(s+1, n+1-s).$$

The posterior mean of p, i.e. the mean of $p|\mathbf{x}$, is given by $\mathbb{E}(p|\mathbf{x})$, which, using properties of the Beta distribution, is given by

$$\mathbb{E}(p|\mathbf{x}) = \frac{s+1}{n+2} \approx \frac{s}{n} \qquad \text{for large } n$$

The posterior variance of p is given by

$$\text{Var}(p|\mathbf{x}) = \frac{(s+1)(n-s+1)}{(n+2)^2(n+3)} \approx \frac{s(n-s)}{n^3} \qquad \text{for large } n.$$

The maximum likelihood estimator for p when observations are binomial is s/n and it can be shown that $\text{Var}(\widehat{p}) \approx s(n-s)/n^3$. This would suggest, at least for this particular example, that the posterior distribution $p|\mathbf{x}$ has similar asymptotic properties to the maximum likelihood estimator from frequentist statistics.

Example [Bernoulli Model with Beta Prior – Continued]: Reminder $p|\mathbf{x} \sim \text{Beta}(s+\alpha, n-s+\beta)$. It is sometimes useful to look at the posterior mean as a function of the maximum likelihood estimator and the prior mean. For this example,

$$\mathbb{E}[p|\mathbf{x}] = \frac{\alpha + s}{\alpha + \beta + n}$$

it can be shown, see Exercises, that

$$\begin{aligned}\mathbb{E}[p|\mathbf{x}] &= \left(\frac{n}{\alpha+\beta+n}\right)\frac{s}{n} + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right) \\ &= \left(\frac{n}{\alpha+\beta+n}\right)\widehat{p} + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)p_0\end{aligned}$$

where the maximum likelihood estimator is given by $\widehat{p} = s/n$ and the prior mean is $p_0 = \alpha/(\alpha+\beta)$. From the above equation we can see that as $n \to \infty$ we have $\mathbb{E}[p|\mathbf{x}] \to \widehat{p}$, i.e. the prior mean loses influence on the posterior mean as we observe more and more data. However, when $n = 0$ we have no observed x values and all we have to rely on is the prior mean p_0. Hence, we see that $\mathbb{E}[p|\mathbf{x}]$ is a tradeoff between the maximum likelihood estimator and the prior.

Example [Poisson Model with Gamma Prior – Continued]: Reminder $\theta|\mathbf{x} \sim$ Gamma$(\alpha + s, \beta + n)$. The posterior mean of θ, i.e. the mean of $\theta|\mathbf{x}$, is given by $\mathbb{E}(\theta|\mathbf{x})$, which, using properties of the Gamma distribution, is given by

$$\mathbb{E}(\theta|\mathbf{x}) = \frac{\alpha + s}{\beta + n} \approx \frac{s}{n} \qquad \text{for large } n.$$

Again, the maximum likelihood estimator for θ when observations are Poisson is s/n. Hence, the posterior mean of θ, $\mathbb{E}(\theta|\mathbf{x})$, is an asymptotically consistent estimator of θ.

Asymptotic Properties

In the above examples we saw that the posterior mean exhibited desirable asymptotic properties. In fact it can be shown that in general, under appropriate regularity conditions, that for any parameter θ,

$$\left(\frac{\theta - \mathbb{E}(\theta|\mathbf{x})}{\sqrt{\mathsf{Var}(\theta|\mathbf{x})}} \middle| \mathbf{x} \right) \rightarrow N(0,1)$$

which is often used to justify approximating the posterior distribution with a normal distribution. This is more formally expressed in the following theorem:

Theorem: Let $\widehat{\theta}_n$ be the maximum likelihood estimator and let $\widehat{se} = 1/\sqrt{nI(\widehat{\theta}_n)}$. Under appropriate regularity conditions, the posterior is approximately normal with mean $\widehat{\theta}_n$ and standard deviation $\widehat{se}$, i.e.

$$\theta|\mathbf{x} \sim N(\widehat{\theta}_n, \widehat{se})$$

Hence, $\overline{\theta}_n \approx \widehat{\theta}_n$.

Example [Normal with Known Variance and Normal Prior on the Mean – Continued]: Reminder:

$$\theta|\mathbf{x} \sim N(\overline{\theta}_n, \tau^2)$$

where $\mathsf{se} = \sigma/\sqrt{n}$,

$$\overline{\theta}_n = w\overline{x} + (1-w)a, \quad w = \frac{\frac{1}{\mathsf{se}^2}}{\frac{1}{\mathsf{se}^2} + \frac{1}{b^2}}, \quad \frac{1}{\tau^2} = \frac{1}{\mathsf{se}^2} + \frac{1}{b^2}$$

Note that $w \to 1$ and $\tau/\mathsf{se} \to 1$ as $n \to \infty$. So for large n, the posterior is approximately $N(\overline{x}, \mathsf{se}^2)$ as stated in the above theorem. Note the same is true if n is fixed but we allow $b^2 \to \infty$ which corresponds to letting the prior becoming flat.

Bayesian Confidence Intervals

We may also use the posterior distribution to obtain a Bayesian interval estimates. One particular approach is to find constants L and R such that

$$\int_{-\infty}^{L} f(\theta|\mathbf{x})d\theta = \int_{R}^{\infty} f(\theta|\mathbf{x})d\theta = \alpha/2$$

where α is a specified level, i.e. 0.05. Let $C = (L, R)$ then

$$\mathbb{P}(\theta \in C|\mathbf{x}) = \int_{L}^{R} f(\theta|\mathbf{x})d\theta = 1 - \alpha.$$

The interval C is a $1-\alpha$ *posterior interval* or *credible interval* or sometimes *credible region*.

Example [Normal with Known Variance and Normal Prior on the Mean – Continued]: Reminder $\theta|\mathbf{x} \sim N(\overline{\theta}_n, \tau^2)$. Suppose we wish to use the above method to find a 95% credible interval for this example. This means we want to find an interval $C = (c, d)$ such that $\mathbb{P}(\theta \in C|\mathbf{X}) = 0.95$ for some constants c and d, i.e.

$$\mathbb{P}(\theta < c|\mathbf{X}) = 0.025 \quad \text{and} \quad \mathbb{P}(\theta > d|\mathbf{X}) = 0.025.$$

Hence, c satisfies

$$\mathbb{P}(\theta < c|\mathbf{X}) = \mathbb{P}\left(\frac{\theta - \overline{\theta}_n}{\tau} < \frac{c - \overline{\theta}_n}{\tau}\middle|\mathbf{X}\right) = \Phi\left(\frac{c - \overline{\theta}_n}{\tau}\right) = 0.025.$$

We know that $\Phi(-1.96) \approx 0.025$. Hence, $(c - \overline{\theta}_n)/\tau \approx -1.96$ so $c \approx \overline{\theta}_n - 1.96\tau$. Using similar arguments $d = \overline{\theta}_n + 1.96\tau$. So a 95% credible interval is given by

$$\overline{\theta}_n \pm 1.96\tau.$$

Since, $\overline{\theta}_n \approx \widehat{\theta}_n$ and $\tau \approx$ se, the 95% credible interval is approximated by $\widehat{\theta}_n \pm 1.96$se which is the frequentist confidence interval.

General Credible Intervals

In general credible intervals are more formally defined as follows:

Definition: Suppose that $\theta \in \Omega$ and C is a subset of Ω. Then C is a $1-\alpha$ credible interval for θ if

$$\mathbb{P}(\theta \in C|\mathbf{X}) = 1 - \alpha.$$

Note:

- By the above definition the $1-\alpha$ credible interval initially described above is by no means the only $1-\alpha$ credible interval.
- The initially described credible interval aims to make the two tail areas to be of size $\alpha/2$. This particular strategy works well when *the posterior density is unimodal*. In most situations in practice the posterior density is unimodal and so such a strategy is sensible.

For a given value of α we want to divide the possible values of θ into 'more plausible' and 'less plausible' values. Suppose that θ_1 and θ_2 are two values of θ. It is natural to say that θ_1 is more plausible than θ_2 if

$$f(\theta_1|\mathbf{x}) > f(\theta_2|\mathbf{x}).$$

This suggests the requirement,

$$f(\theta_1|\mathbf{x}) > f(\theta_2|\mathbf{x}) \qquad \text{for all} \quad \theta_1 \in C \quad \text{and} \quad \theta_2 \notin C.$$

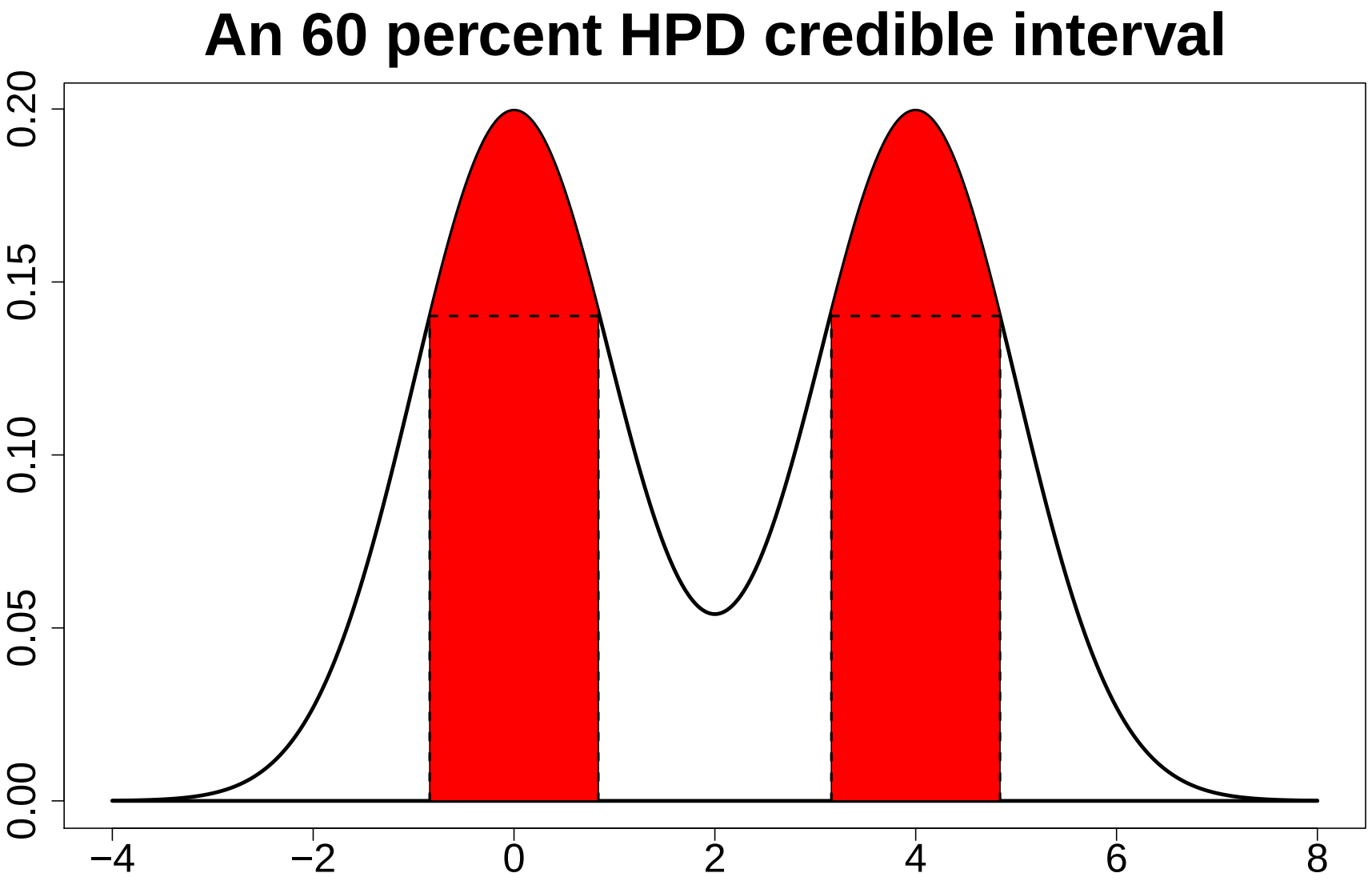
so that any value of θ included in C is at least as probable as any excluded value.

Definition: A $1-\alpha$ credible interval C that satisfies

$$f(\theta_1|\mathbf{x}) > f(\theta_2|\mathbf{x}) \qquad \text{for all} \quad \theta_1 \in C \quad \text{and} \quad \theta_2 \notin C.$$

is called a $1-\alpha$ *highest posterior density (HPD) credible interval*.

Note that if $f(\theta_1|\mathbf{x})$ is symmetric, unimodal and continuous over the real line then the interval $C \in (L, R)$ has the property $f(L|\mathbf{x}) = f(R|\mathbf{x})$.


An 60 percent HPD credible interval
0.00
0.05
0.10
0.15
0.20
−4
−2
0
2
4
6
8

Bayesian Model Selection

Model selection from a Bayesian perspective is quite a subtle topic. Like it frequentist inference there are numerous methods for selecting a model. Some Bayesian methods are analogous to frequentist counterparts (albeit with a different interpretation).

Frequentist	Bayesian
Hypothesis Testing	Bayes Factors
AIC or BIC	Deviance Information Criterion (DIC)
Penalised likelihood	Sparsity inducing priors

We will only consider the briefest of introductions here.

Bayesian Hypothesis Testing

Suppose that one of two hypotheses H_0 and H_1 are true. Let $\mathbb{P}(H_i)$ denote the prior probability that H_i is the true hypothesis, and, after sample data $\mathbf{x}$ has been observed, $\mathbb{P}(H_i|\mathbf{x})$ denotes the posterior probability that H_i is true given the observed data. Bayesian hypothesis testing aims determine the *posterior odds*,

$$Q^* = \frac{\mathbb{P}(H_0|\mathbf{x})}{\mathbb{P}(H_1|\mathbf{x})}.$$

Hence, the conclusion from a Bayesian analysis might be a statement of the form

> "H_0 is Q^* times more likely to be true than H_1."

Alternatively, as $\mathbb{P}(H_0|\mathbf{x}) + \mathbb{P}(H_1|\mathbf{x}) = 1$, we might conclude that

> "$Q^*/(1+Q^*)$ and $1/(1+Q^*)$ are the probabilities that H_0 and that H_1 is true respectively."

Notice that, unlike the hypothesis testing approaches in frequentest inference, in Bayesian hypothesis testing the hypotheses have equal status, i.e. in the frequentist approach the null hypothesis is retained unless there is evidence against it. Now,

$$\mathbb{P}(H_0|\mathbf{x}) \propto \mathbb{P}(H_0) f(\mathbf{x}|H_0)$$

so

$$Q^* = \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} \frac{f(\mathbf{x}|H_0)}{f(\mathbf{x}|H_1)} = Q \times B$$

where

$$Q = \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}$$

are the prior odds of the competing hypotheses and

$$B = \frac{f(\mathbf{x}|H_0)}{f(\mathbf{x}|H_1)}$$

is called the *Bayes factor*.

The prior odds, Q, represent our beliefs, before collecting the data, as to which hypothesis is true. Often set to 1 to represent impartiality between the hypotheses, so that each hypothesis is considered to be equally likely, a priori. Primary interest is centred on the Bayes factor, B, since this determines how the data have changed our beliefs as to which hypothesis is true.

This approach can be used in a variety of situations. Here, however, we restrict our attention to the situation where $f(\mathbf{x}|\theta)$ is the model distribution for each hypothesis, but the hypotheses differ in the values they specify for θ. The cases to be considered are:

1. Both H_0 and H_1 are simple hypotheses.
2. Both H_0 and H_1 are composite hypotheses.
3. H_0 is a simple hypothesis and H_1 is a composite hypothesis.

Both H_0 and H_1 are Simple Hypotheses

If both H_0 and H_1 are simple hypotheses, they have the form

$$H_0 \colon \theta = \theta_0 \quad \textit{vs} \quad H_1 \colon \theta = \theta_1$$

and the Bayes factor is just $B = f(\mathbf{x}|\theta_0)/f(\mathbf{x}|\theta_1)$ which coincides with the likelihood ratio in frequentist inference.

Example: Suppose that $X_1, X_2, \ldots, X_n$ are observations form an exponential distribution with parameter θ. Then the Bayes factor is given by

$$B = \frac{\theta_0^n \exp\{-n\theta_0 \overline{x}\}}{\theta_1^n \exp\{-n\theta_1 \overline{x}\}} = \left(\frac{\theta_0}{\theta_1}\right)^n \exp\{(\theta_1 - \theta_0) n \overline{x}\}.$$

For particular values of θ_0, θ_1, n and $\overline{x}$ we can calculate the Bayes factor.

Both H_0 and H_1 are Composite Hypotheses

If both H_0 and H_1 are composite hypotheses, we have

$$H_0 : \theta \in \omega \quad \textit{vs} \quad H_1 : \theta \in \Omega - \omega$$

where $\theta \in \Omega$ and ω is some subset of Ω. For each hypothesis a prior distribution must be specified. Denote these prior densities by $f_k(\theta|H_k)$, $k \in \{0, 1\}$. Then

$$f(\mathbf{x}|H_0) = \int_\omega f(\mathbf{x}|\theta) f_0(\theta|H_0) d\theta$$

and

$$f(\mathbf{x}|H_1) = \int_{\Omega-\omega} f(\mathbf{x}|\theta) f_1(\theta|H_1) d\theta$$

(similarly for $f(\mathbf{x}|H_0)$) and

$$B = \frac{f(\mathbf{x}|H_0)}{f(\mathbf{x}|H_1)} = \frac{\displaystyle\int_\omega f(\mathbf{x}|\theta) f_0(\theta|H_0) d\theta}{\displaystyle\int_{\Omega-\omega} f(\mathbf{x}|\theta) f_1(\theta|H_1) d\theta}$$

If $f_0(\theta|H_0) = f_1(\theta|H_1) = f(\theta)$ then

$$B = \frac{\int_\omega \frac{f(\mathbf{x}|\theta)f(\theta)}{f(\mathbf{x}} d\theta}{\int_{\Omega-\omega} \frac{f(\mathbf{x}|\theta)f(\theta|H)}{f(\mathbf{x}} d\theta}$$

$$= \frac{\int_\omega f(\theta|\mathbf{x}) d\theta}{\int_{\Omega-\omega} f(\theta|\mathbf{x}) d\theta}$$

Example: As part of a quality inspection program, five components are selected at random from a batch of components and tested. The number of components that fail, X, follows the binomial distribution, $X \sim \text{Binomial}(5, \theta)$, and from base experience θ has a Beta distribution $f(\theta) = 30\theta(1-\theta)^4$, $0 \leq \theta \leq 1$. For one batch, no failure were found when five of its components were tested. For this batch we wish to test the hypotheses

$$H_0 : \theta \leq 0.2 \quad \textit{vs} \quad H_1 : \theta > 0.2.$$

The observed value of X is 0. This leads to the posterior distribution

$$f(\theta|X) = 11\theta(1-\theta)^9.$$

Thus,

$$Q^* = \frac{\int_0^{0.2} 11\theta(1-\theta)^9 d\theta}{\int_{0.2}^1 11\theta(1-\theta)^9 d\theta} = \frac{0.6779}{0.3221} = 2.10.$$

Note that for this problem we used the same priors for both hypotheses. In general the priors use for both hypotheses could be different.

H_0 is a Simple Hypothesis and H_1 is a Composite Hypothesis

If H_0 is a simple hypothesis and H_1 is a composite hypothesis, we might have

$$H_0 : \theta = \theta_0 \quad \textit{vs} \quad H_1 : \theta \neq \theta_0$$

Combining the results from previous two subsections we obtain

$$B = \frac{f(\mathbf{x}|\theta_0)}{\int_{\theta \neq \theta_0} f(\mathbf{x}|\theta) f_1(\theta|H_1) d\theta}.$$

Example: Suppose that we have $X_1, \ldots, X_n$ from a Poisson distribution with mean θ. Suppose that we wish to test the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ and the prior distribution for θ is the distribution

$$f(\theta|H_1) = \alpha_2^{(\alpha_1+1)} \theta^{\alpha_1} \exp\{-\alpha_2 \theta\} / \Gamma(\alpha_1 + 1).$$

To obtain the Bayes factor or we require $f(\mathbf{x}|\theta_0) = \theta_0^{n\overline{x}} \exp\{-n\theta_0\} / \prod_{i=1}^n x_i!$ and

$$\int_{\theta \neq \theta_0} f(\mathbf{x}|\theta_1) f_1(\theta|H_1) d\theta = \int_0^\infty \left[\frac{\theta^{n\overline{x}} \exp\{-n\theta\}/}{\prod_{i=1}^n x_i!} \right] \left[\frac{\alpha_2^{(\alpha_1+1)} \theta^{\alpha_1} \exp\{-\alpha_2\theta\}}{\Gamma(\alpha_1+1)} \right] d\theta$$

$$= \frac{\alpha_2^{(\alpha_1+1)} \Gamma(\alpha_1 + n\overline{x} + 1)}{(n+\alpha_2)^{\alpha_1+n\overline{x}+1} \Gamma(\alpha_1+1)(\prod_{i=1}^n x_i!)} \times \int_0^\infty \frac{(n+\alpha_2)^{\alpha_1+n\overline{x}+1} \theta^{\alpha_1+n\overline{x}} \exp\{-(n+\alpha_2)\theta\} d\theta}{\Gamma(\alpha_1 + n\overline{x} + 1)}$$

where the fact that $\theta \neq \theta_0$ has been ignored, as this does not affect the value of the integral. Thus (see Exercises),

$$B = \frac{\theta_0^{n\overline{x}} \exp(-n\theta_0)(n+\alpha_2)^{\alpha_1+n\overline{x}+1} \Gamma(\alpha_1+1)}{\alpha_2^{\alpha_1+1} \Gamma(\alpha_1 + n\overline{x} + 1)}.$$

As a specific illustration suppose that the random sample consists of six observations: $3, 1, 6, 2, 5, 2$, $\theta_0 = 2.0$, $\alpha_1 = 2.6$, $\alpha_2 = 0.6$ and that we believe both hypotheses are equally true then substitution into the formula gives $B = 0.77$ so the posterior odds are 0.385 and we have firmer belief that H_1 is the true hypothesis.

Deviance Information Criterion

The Deviance Information Criterion (DIC), is a way of scoring models, similar in spirit to the AIC and BIC criterion (which you may have heard about), in order to determine which models are preferable. The DIC is defined by:

$$\text{DIC} = -2\log f(\mathbf{x}|\widetilde{\boldsymbol{\theta}}) + 2P_D$$

where $\mathbf{x}$ is the vector of observed data, $\widetilde{\boldsymbol{\theta}}$ is some Bayesian point estimate of $\widetilde{\boldsymbol{\theta}}$ (usually either the poterior mean, median or mode) and P_D is an estimate of the dimension of the model and is given by

$$P_D = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}\left[-2\log f(\mathbf{x}|\boldsymbol{\theta})\right] + 2\log f(\mathbf{x}|\widetilde{\boldsymbol{\theta}}).$$

Hence,

$$\text{DIC} = -4\mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}\left[\log f(\mathbf{x}|\boldsymbol{\theta})\right] + 2\log f(\mathbf{x}|\widetilde{\boldsymbol{\theta}}).$$

Note:

- The DIC was derived for Bayesian models and DIC behaves more similarly to AIC than BIC.
- Bayes factors only work when proper priors are used! DIC can be used even if improper priors are used.
- The DIC is often calculated automatically by packages such as `WinBUGS`.
- P_D is not invariant to reparametrisation!
- Can integrate out a subset of parameters, i.e. could start with $f(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\xi})$ and calculate the DIC based on

$$f(\mathbf{x}|\boldsymbol{\theta}) = \int f(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\xi}) f(\boldsymbol{\xi}).$$

  In this case $\boldsymbol{\theta}$ is called the *focus* for the DIC.

Example [Normal with Known Variance and Normal Prior on the Mean]: Let $X_1, \ldots, X_n|\theta \sim N(\theta, \sigma^2)$ where, for simplicity, we assume that σ^2 is known. Suppose we take as a prior $\theta \sim N(a, b^2)$. Then, $\theta|\mathbf{x} \sim N(\mu, \tau^2)$ where

$$\mu = w\overline{x} + (1-w)a, \quad w = \frac{1}{\mathsf{se}^2}\left(\frac{1}{\mathsf{se}^2} + \frac{1}{b^2}\right)^{-1}, \quad \frac{1}{\tau^2} = \frac{1}{\mathsf{se}^2} + \frac{1}{b^2}$$

and $\mathsf{se} = \sigma/\sqrt{n}$ is the standard error of the maximum likelihood estimator $\overline{x}$. Let $\widetilde{\boldsymbol{\theta}} = E(\theta|\mathbf{x}) = \mu$ then the log-"likelihood" term is $\log f(x|\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$ and

$$\begin{aligned}
\mathbb{E}_{\theta|\mathbf{x}}\left[\log f(x|\theta)\right] &= -\tfrac{n}{2}\log(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\mathbb{E}_{\theta|\mathbf{x}}\left[\sum_{i=1}^n (x_i - \theta)^2\right] \\
&= -\tfrac{n}{2}\log(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\mathbb{E}_{\theta|\mathbf{x}}\left[n\tau^2 + \sum_{i=1}^n (x_i - \mu)^2\right].
\end{aligned}$$

$$\text{Hence,} \quad P_D = \frac{n\tau^2}{\sigma^2} = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{b^2}} = 1 - \left(1 + \frac{bn}{\sigma^2}\right)^{-1}.$$

Sparsity Inducing Priors

Definition: A sparsity inducing prior is any prior with a discontinuous derivative at zero.

There has been a lot of recent activity concerning sparsity inducing priors. These priors encourage point estimators of their corresponding parameters towards 0. Estimators with these priors can have the following properties:

- Consistency.
- Robustness.
- If the true value of the parameter is 0 then these estimators can converge to 0 at a faster rate than the MLE.

Example: One of the first examples of a sparsity inducing prior in the literature is the famous spike and slab prior. Suppose

$$y_i|\boldsymbol{\beta}, \sigma^2 \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \sigma^2).$$

A spike and slab prior for β_i uses:

$$\beta_i|\gamma_i \sim \gamma_i N(0, \sigma^2_\beta) + (1 - \gamma_i)\delta_0$$

and $\gamma_i|\rho \sim$ Bernoulli(ρ) with prior, $\rho \sim$ Beta(A, B) where δ_0 is a point mass at 0 and σ^2_β, A and B are constants.

Note that $f(\gamma_i|\mathbf{y})$ can be interpreted as the posterior probability that the covariate x_i contributes to the model.

Tongue in Cheek Proof that all Frequentists are Bayesian

- Consider any model $f(\mathbf{x}|\boldsymbol{\theta})$.
- Let $f(\boldsymbol{\theta}) \propto 1$.
- Let $\widehat{\boldsymbol{\theta}}_{\text{Bayes}} = \text{mode}(f(\boldsymbol{\theta}|\mathbf{x}))$
- Then

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{\text{Bayes}} &= \text{argmax}_{\theta}(f(\boldsymbol{\theta}|\mathbf{x})) \\ &= \text{argmax}_{\theta}(f(\mathbf{x}|\boldsymbol{\theta})) \\ &= \widehat{\theta}_{\text{MLE}}.\end{aligned}$$

Summary

Why Bayesian inference is good for you:

- Has a solid decision-theoretic framework.
- Intuitively combines the prior distribution (prior beliefs and/or experience) with the likelihood (experiment) to obtain the posterior distribution (accumulated information).
- The plug-in principle is avoided (uncertainty is properly propagated through the model).
- Newly developed MCMC methods may computations tractable for *practically all models*. Software (such as `WinBUGS`) is available for this.
- Focus shifts from model estimation to model appropriateness.