Bayesian Inference I

Loukia Meligkotsidou,
National and Kapodistrian University of Athens

MSc in Biostatistics,
Department of Mathematics and School of Medicine

Outline of the course

This course provides theory and practice of the Bayesian approach to statistical inference. Applications are performed with the statistical package R.

Topics:

- Bayesian Updating through Bayes' Theorem
- Prior Distributions
- Multi-parameter Problems
- Decision Theory and Bayesian Inference
- Prediction
- The Gibbs Sampler

Unit 3: Specifying Priors

The computational difficulties arise in using Bayes' Theorem when it is necessary to evaluate the normalizing constant

$$\int f(\theta)f(x|\theta)d\theta.$$

Unit 3: Specifying Priors

The computational difficulties arise in using Bayes' Theorem when it is necessary to evaluate the normalizing constant

$$\int f(\theta)f(x|\theta)d\theta.$$

Suppose $X_1, \ldots, X_n$ are independent Poisson(θ) r.v.s, and our beliefs about θ are that it lies in $[0, 1]$ and all values are equally likely: $f(\theta) = 1;\ 0 \leq \theta \leq 1$ and $f(\theta|x) \propto \exp(-n\theta)\theta^{\Sigma x_i}$. Then

$$\int_0^1 \exp(-n\theta)\theta^{\Sigma x_i} d\theta,$$

and this integral can only be evaluated numerically.

Unit 3: Specifying Priors

The computational difficulties arise in using Bayes' Theorem when it is necessary to evaluate the normalizing constant

$$\int f(\theta) f(x|\theta) d\theta.$$

Suppose $X_1, \ldots, X_n$ are independent Poisson(θ) r.v.s, and our beliefs about θ are that it lies in $[0, 1]$ and all values are equally likely: $f(\theta) = 1;\ 0 \leq \theta \leq 1$ and $f(\theta|x) \propto \exp(-n\theta)\theta^{\Sigma x_i}$. Then

$$\int_0^1 \exp(-n\theta)\theta^{\Sigma x_i} d\theta,$$

and this integral can only be evaluated numerically.

So, even simple choices of priors can lead to awkward **numerical problems**. But, we have seen cases in which we were able to identify a prior for which the posterior was in the same family of distributions as the prior; such priors are called *conjugate priors*.

An Example. Gamma Sample

Let $X_1, \dots X_n$ be independent variables having the $Gamma(k, \theta)$ distribution, where k is known. Then

$$f(x_i \mid \theta) = \frac{1}{\Gamma(k)} \theta^k x_i^{k-1} e^{-\theta x_i} \propto \theta^k e^{-\theta x_i}$$

$$\text{So,} \quad f(x \mid \theta) \propto \prod_{i=1}^{n} \theta^k e^{-\theta x_i} = \theta^{nk} \exp\{-\theta \Sigma x_i\}.$$

An Example. Gamma Sample

Let $X_1, \ldots X_n$ be independent variables having the $Gamma(k, \theta)$ distribution, where k is known. Then

$$f(x_i \mid \theta) = \frac{1}{\Gamma(k)} \theta^k x_i^{k-1} e^{-\theta x_i} \propto \theta^k e^{-\theta x_i}$$

$$\text{So,} \quad f(x \mid \theta) \propto \prod_{i=1}^{n} \theta^k e^{-\theta x_i} = \theta^{nk} \exp\{-\theta \Sigma x_i\}.$$

Now, studying this form, regarded as a function of θ suggests we could take a prior of the form

$$f(\theta) \propto \theta^{p-1} \exp\{-q\theta\}$$

that is, $\theta \sim Gamma(p, q)$.

An Example. Gamma Sample

Let $X_1, \ldots X_n$ be independent variables having the $Gamma(k, \theta)$ distribution, where k is known. Then

$$f(x_i \mid \theta) = \frac{1}{\Gamma(k)} \theta^k x_i^{k-1} e^{-\theta x_i} \propto \theta^k e^{-\theta x_i}$$

$$\text{So,} \quad f(x \mid \theta) \propto \prod_{i=1}^{n} \theta^k e^{-\theta x_i} = \theta^{nk} \exp\{-\theta \Sigma x_i\}.$$

Now, studying this form, regarded as a function of θ suggests we could take a prior of the form

$$f(\theta) \propto \theta^{p-1} \exp\{-q\theta\}$$

that is, $\theta \sim Gamma(p, q)$. Then by Bayes' Theorem

$$f(\theta|x) \propto \theta^{p+nk-1} \exp\{-(q + \Sigma x_i)\theta\},$$

and so $\theta|x \sim Gamma(p + nk, q + \sum x_i)$.

Conjugate Priors

Provided they are not in direct conflict with our prior beliefs, and provided such a family can be found, the simplicity induced by using a conjugate prior is compelling.

Conjugate Priors

Provided they are not in direct conflict with our prior beliefs, and provided such a family can be found, the simplicity induced by using a conjugate prior is compelling.

The only case where conjugates can be easily obtained is for data models within the *exponential family*. That is,

$$f(x|\theta) = h(x)g(\theta)\exp\{t(x)c(\theta)\}$$

for functions h, g, t and c such that

$$\int f(x|\theta)dx = g(\theta)\int h(x)\exp\{t(x)c(\theta)\}dx = 1.$$

This might seem restrictive, but in fact includes the exponential distribution, the Poisson distribution, the gamma distribution with known shape parameter, the binomial distribution, the normal distribution with known variance and many more.

Obtaining Conjugate Priors

Given a random sample $x = (x_1, x_2, \ldots, x_n)$ from this general distribution, the likelihood for θ is then

$$
\begin{aligned}
f(x \mid \theta) &= \prod_{i=1}^{n} \{h(x_i)\} g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i) c(\theta)\} \\
&\propto g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i) c(\theta)\}.
\end{aligned}
$$

Obtaining Conjugate Priors

Given a random sample $x = (x_1, x_2, \ldots, x_n)$ from this general distribution, the likelihood for θ is then

$$
\begin{aligned}
f(x \mid \theta) &= \prod_{i=1}^{n}\{h(x_i)\} g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i) c(\theta)\} \\
&\propto g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i) c(\theta)\}.
\end{aligned}
$$

Thus, if we choose a prior of the form $f(\theta) \propto g(\theta)^d \exp\{b\, c(\theta)\}$,

Obtaining Conjugate Priors

Given a random sample $x = (x_1, x_2, \ldots, x_n)$ from this general distribution, the likelihood for θ is then

$$\begin{aligned} f(x \mid \theta) &= \prod_{i=1}^{n} \{h(x_i)\} g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i) c(\theta)\} \\ &\propto g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i) c(\theta)\}. \end{aligned}$$

Thus, if we choose a prior of the form $f(\theta) \propto g(\theta)^d \exp\{b\, c(\theta)\}$,

$$\begin{aligned} f(\theta|x) &\propto f(\theta) f(x \mid \theta) \\ &\propto g(\theta)^d \exp\{b\, c(\theta)\} \times g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i) c(\theta)\} \end{aligned}$$

Obtaining Conjugate Priors

Given a random sample $x = (x_1, x_2, \ldots, x_n)$ from this general distribution, the likelihood for θ is then

$$
\begin{aligned}
f(x \mid \theta) &= \prod_{i=1}^{n}\{h(x_i)\} g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i) c(\theta)\} \\
&\propto g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i) c(\theta)\}.
\end{aligned}
$$

Thus, if we choose a prior of the form $f(\theta) \propto g(\theta)^d \exp\{b\, c(\theta)\}$,

$$
\begin{aligned}
f(\theta|x) &\propto f(\theta) f(x \mid \theta) \\
&\propto g(\theta)^d \exp\{b\, c(\theta)\} \times g(\theta)^n \exp\{\sum_{i=1}^{n} t(x_i) c(\theta)\} \\
&= g(\theta)^{n+d} \exp\{[b + \sum_{i=1}^{n} t(x_i)] c(\theta)\} = g(\theta)^D \exp\{B c(\theta)\}
\end{aligned}
$$

Example 1. Binomial Sample

A binomial random variable has pdf

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

Example 1. Binomial Sample

A binomial random variable has pdf

$$\begin{aligned} f(x|\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}(1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^x \end{aligned}$$

Example 1. Binomial Sample

A binomial random variable has pdf

$$\begin{aligned} f(x|\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}(1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^x \\ &= \binom{n}{x}(1-\theta)^n \exp\{x\log(\frac{\theta}{1-\theta})\}. \end{aligned}$$

So, $h(x) = \binom{n}{x}$, $g(\theta) = 1-\theta$, $t(x) = x$, and $c(\theta) = \log(\frac{\theta}{1-\theta})$.

Example 1. Binomial Sample

A binomial random variable has pdf

$$\begin{aligned} f(x|\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}(1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^x \\ &= \binom{n}{x}(1-\theta)^n \exp\{x\log(\frac{\theta}{1-\theta})\}. \end{aligned}$$

So, $h(x) = \binom{n}{x}$, $g(\theta) = 1-\theta$, $t(x) = x$, and $c(\theta) = \log(\frac{\theta}{1-\theta})$. Thus, we construct a conjugate prior with the form

$$f(\theta) \propto (1-\theta)^d \exp\{b\log(\frac{\theta}{1-\theta})\}$$

Example 1. Binomial Sample

A binomial random variable has pdf

$$\begin{aligned} f(x|\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}(1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^x \\ &= \binom{n}{x}(1-\theta)^n\exp\{x\log(\frac{\theta}{1-\theta})\}. \end{aligned}$$

So, $h(x)=\binom{n}{x}$, $g(\theta)=1-\theta$, $t(x)=x$, and $c(\theta)=\log(\frac{\theta}{1-\theta})$.
Thus, we construct a conjugate prior with the form

$$\begin{aligned} f(\theta) &\propto (1-\theta)^d\exp\{b\log(\frac{\theta}{1-\theta})\} \\ &= (1-\theta)^{d-b}\theta^b = (1-\theta)^{\alpha-1}\theta^{\beta-1} \end{aligned}$$

which is a member of the beta family of distributions.

Example 2. Normal Mean

Let $X_1, ..., X_n$ be a random sample from the $N(\theta, \sigma^2)$ distribution with σ^2 known. Then,

$$\begin{aligned} f(x|\theta) &\propto \exp\left\{-\frac{n\theta^2}{2\sigma^2} + \frac{\theta \sum x_i}{\sigma^2}\right\} = \left[\exp\left\{-\frac{\theta^2}{2\sigma^2}\right\}\right]^n \exp\left\{\frac{\theta \sum x_i}{\sigma^2}\right\} \\ &= g(\theta)^n \exp\{\sum_{i=1}^n t(x_i)c(\theta)\}, \end{aligned}$$

where, $g(\theta) = \exp\left\{\frac{-\theta^2}{2\sigma^2}\right\}$, $t(x_i) = x_i$, and $c(\theta) = \frac{\theta}{\sigma^2}$.

Example 2. Normal Mean

Let $X_1, ..., X_n$ be a random sample from the $N(\theta, \sigma^2)$ distribution with σ^2 known. Then,

$$f(x|\theta) \propto \exp\left\{-\frac{n\theta^2}{2\sigma^2} + \frac{\theta \sum x_i}{\sigma^2}\right\} = \left[\exp\left\{-\frac{\theta^2}{2\sigma^2}\right\}\right]^n \exp\left\{\frac{\theta \sum x_i}{\sigma^2}\right\}$$

$$= g(\theta)^n \exp\{\sum_{i=1}^n t(x_i)c(\theta)\},$$

where, $g(\theta) = \exp\left\{\frac{-\theta^2}{2\sigma^2}\right\}$, $t(x_i) = x_i$, and $c(\theta) = \frac{\theta}{\sigma^2}$.

Conjugate prior:

$$f(\theta) \propto g(\theta)^d \exp\{bc(\theta)\} \propto \exp\left\{-\frac{d\theta^2}{2\sigma^2}\right\} \exp\left\{\frac{b\theta}{\sigma^2}\right\}$$

Example 2. Normal Mean

Let $X_1, \ldots, X_n$ be a random sample from the $N(\theta, \sigma^2)$ distribution with σ^2 known. Then,

$$\begin{aligned} f(x|\theta) &\propto \exp\left\{-\frac{n\theta^2}{2\sigma^2} + \frac{\theta\sum x_i}{\sigma^2}\right\} = \left[\exp\left\{-\frac{\theta^2}{2\sigma^2}\right\}\right]^n \exp\left\{\frac{\theta\sum x_i}{\sigma^2}\right\} \\ &= g(\theta)^n \exp\{\sum_{i=1}^n t(x_i)c(\theta)\}, \end{aligned}$$

where, $g(\theta) = \exp\left\{\frac{-\theta^2}{2\sigma^2}\right\}$, $t(x_i) = x_i$, and $c(\theta) = \frac{\theta}{\sigma^2}$.

Conjugate prior:

$$\begin{aligned} f(\theta) &\propto g(\theta)^d \exp\{bc(\theta)\} \propto \exp\left\{-\frac{d\theta^2}{2\sigma^2}\right\} \exp\left\{\frac{b\theta}{\sigma^2}\right\} \\ &= \exp\left\{-\frac{d\theta^2}{2\sigma^2} + \frac{b\theta}{\sigma^2}\right\} = \exp\left\{-\frac{\theta^2}{2D^2} + \frac{B\theta}{D^2}\right\}, \quad [N(B, D^2)]. \end{aligned}$$

Mixtures of Priors

An Example. When a coin is tossed, then almost invariably there is a 0.5 chance of it coming up heads. However, if the coin is spun on a table, it is often the case that slight imperfections in the edge of the coin cause it to have a tendency to prefer either heads or tails. Taking this into account, we may wish to give the probability θ of the coin coming up heads a prior distribution which favours values around either 0.3 or 0.7 say.

Mixtures of Priors

An Example. When a coin is tossed, then almost invariably there is a 0.5 chance of it coming up heads. However, if the coin is spun on a table, it is often the case that slight imperfections in the edge of the coin cause it to have a tendency to prefer either heads or tails. Taking this into account, we may wish to give the probability θ of the coin coming up heads a prior distribution which favours values around either 0.3 or 0.7 say.

That is, our prior beliefs may be reasonably represented by a *bimodal* distribution (or even *trimodal* if we wish to give extra weight to the unbiased possibility, $\theta = 0.5$).

Mixtures of Priors

An Example. When a coin is tossed, then almost invariably there is a 0.5 chance of it coming up heads. However, if the coin is spun on a table, it is often the case that slight imperfections in the edge of the coin cause it to have a tendency to prefer either heads or tails. Taking this into account, we may wish to give the probability θ of the coin coming up heads a prior distribution which favours values around either 0.3 or 0.7 say.

That is, our prior beliefs may be reasonably represented by a *bimodal* distribution (or even *trimodal* if we wish to give extra weight to the unbiased possibility, $\theta = 0.5$).

Our likelihood model for the number of heads in n spins will be Binomial: $X|\theta \sim Binomial(n, \theta)$ and so the conjugate prior is the beta family. However, no member of this family is multimodal.

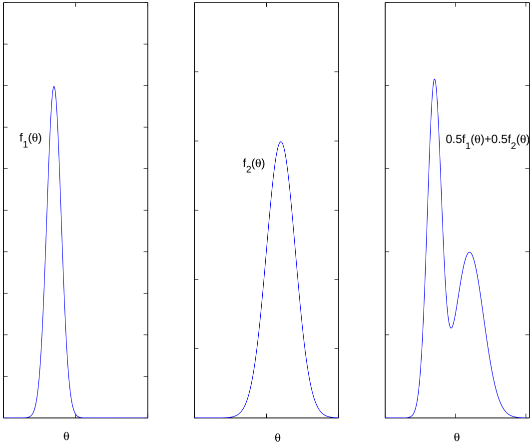
Mixtures of Priors

An Example. When a coin is tossed, then almost invariably there is a 0.5 chance of it coming up heads. However, if the coin is spun on a table, it is often the case that slight imperfections in the edge of the coin cause it to have a tendency to prefer either heads or tails. Taking this into account, we may wish to give the probability θ of the coin coming up heads a prior distribution which favours values around either 0.3 or 0.7 say.

That is, our prior beliefs may be reasonably represented by a *bimodal* distribution (or even *trimodal* if we wish to give extra weight to the unbiased possibility, $\theta = 0.5$).

Our likelihood model for the number of heads in n spins will be Binomial: $X|\theta \sim Binomial(n, \theta)$ and so the conjugate prior is the beta family. However, no member of this family is multimodal.
One solution is to use *mixtures* of conjugate distributions!

A Mixture of two Distributions



Mixtures of Priors

The mixtures of conjugate priors will also be a conjugate prior family for the likelihood model!

Mixtures of Priors

The mixtures of conjugate priors will also be a conjugate prior family for the likelihood model!

Suppose $f_1(\theta), \ldots, f_k(\theta)$ are all conjugate distributions for θ, leading to posterior distributions $f_1(\theta|x), \ldots, f_k(\theta|x)$.

Now consider the family of mixture distributions:

$$f(\theta) = \sum_{i=1}^{k} p_i f_i(\theta),$$

where $0 \leq p_i \leq 1$, $i = 1, \ldots, k$ and $\sum_{i=1}^{k} p_i = 1$.

The Posterior

Then,

$$\begin{aligned} f(\theta|x) &\propto f(\theta)f(x|\theta) \\ &= \sum_{i=1}^{k} p_i f_i(\theta) f(x|\theta), \text{but} \;\; f_i(\theta|x) = \frac{f_i(\theta)f(x|\theta)}{f_i(x)} \end{aligned}$$

The Posterior

Then,

$$\begin{aligned} f(\theta|x) &\propto f(\theta)f(x|\theta) \\ &= \sum_{i=1}^{k} p_i f_i(\theta)f(x|\theta), \text{but } \ f_i(\theta|x) = \frac{f_i(\theta)f(x|\theta)}{f_i(x)} \\ &= \sum_{i=1}^{k} p_i f_i(x)f_i(\theta|x), \text{hence } \ f(\theta|x) = \sum_{i=1}^{k} p_i^* f_i(\theta|x), \end{aligned}$$

where $p_i^* \propto p_i f_i(x)$. So the posterior is in the same mixture–family. Notice though that the mixture proportions in the posterior p_i^* generally will be different from those in the prior.

The Posterior

Then,

$$\begin{aligned} f(\theta|x) &\propto f(\theta)f(x|\theta) \\ &= \sum_{i=1}^{k} p_i f_i(\theta) f(x|\theta), \text{but} \;\; f_i(\theta|x) = \frac{f_i(\theta)f(x|\theta)}{f_i(x)} \\ &= \sum_{i=1}^{k} p_i f_i(x) f_i(\theta|x), \text{hence} \;\; f(\theta|x) = \sum_{i=1}^{k} p_i^* f_i(\theta|x), \end{aligned}$$

where $p_i^* \propto p_i f_i(x)$. So the posterior is in the same mixture–family. Notice though that the mixture proportions in the posterior p_i^* generally will be different from those in the prior.

Finite mixtures of conjugate priors can be made arbitrarily close to *any* prior distribution. However, it may be possible to represent one's prior beliefs more succinctly using non–conjugate priors.

Improper Priors

Let $X_1, \ldots, X_n \sim N(\theta, \tau^{-1})$, τ known, $\theta \sim N(b, c^{-1})$.
The posterior is $\theta|x \sim N(\frac{cb+n\tau\bar{x}}{c+n\tau}, \frac{1}{c+n\tau})$.

Improper Priors

Let $X_1, \ldots, X_n \sim N(\theta, \tau^{-1})$, τ known, $\theta \sim N(b, c^{-1})$.
The posterior is $\theta | x \sim N(\frac{cb+n\tau\bar{x}}{c+n\tau}, \frac{1}{c+n\tau})$.

The strength of our prior beliefs about θ are determined by the variance, or equivalently the precision, c, of the normal prior.

A large value of c corresponds to very strong prior beliefs; on the other hand small values of c reflect very weak prior information.

Improper Priors

Let $X_1, \ldots, X_n \sim N(\theta, \tau^{-1})$, τ known, $\theta \sim N(b, c^{-1})$.
The posterior is $\theta|x \sim N(\frac{cb+n\tau\overline{x}}{c+n\tau}, \frac{1}{c+n\tau})$.

The strength of our prior beliefs about θ are determined by the variance, or equivalently the precision, c, of the normal prior.

A large value of c corresponds to very strong prior beliefs; on the other hand small values of c reflect very weak prior information.

Now, suppose our prior beliefs about θ were so weak that we let $c \to 0$. Then simply enough, the posterior distribution becomes $N(\overline{x}, \frac{1}{n\tau})$, or in the more familiar notation: $N(\overline{x}, \frac{\sigma^2}{n})$. Thus we seemingly obtain a perfectly valid posterior distribution through this limiting procedure.

Improper Priors

Consider, though, what's happening to the prior as $c \to 0$. In effect, we obtain a $N(b, \infty)$ prior, which is not a genuine, 'proper' distribution.

Improper Priors

Consider, though, what's happening to the prior as $c \to 0$. In effect, we obtain a $N(b, \infty)$ prior, which is not a genuine, 'proper' distribution.

In fact, as $c \to 0$, the distribution of $N(b, c^{-1})$ becomes increasingly flatter, so that in any interval $-K \leq \theta \leq K$, provided c is sufficiently close to 0, we have approximately

$$f(\theta) \propto 1; \quad -K \leq \theta \leq K.$$

Improper Priors

Consider, though, what's happening to the prior as $c \to 0$. In effect, we obtain a $N(b, \infty)$ prior, which is not a genuine, 'proper' distribution.

In fact, as $c \to 0$, the distribution of $N(b, c^{-1})$ becomes increasingly flatter, so that in any interval $-K \leq \theta \leq K$, provided c is sufficiently close to 0, we have approximately

$$f(\theta) \propto 1; \quad -K \leq \theta \leq K.$$

But this cannot be valid, in the limit as $c \to 0$, over the whole real line $\mathcal{R}$, because

$$\int_{\mathcal{R}} f(\theta) d\theta = \infty.$$

Improper Priors

The posterior $N(\overline{x}, \frac{\sigma^2}{n})$, obtained by letting $c \to 0$ in the standard conjugate analysis, cannot arise through the use of any proper prior distribution. It does arise however by formal use of the prior specification $f(\theta) \propto 1$, which is an example of what is termed an *improper* prior distribution.

Improper Priors

The posterior $N(\overline{x}, \frac{\sigma^2}{n})$, obtained by letting $c \to 0$ in the standard conjugate analysis, cannot arise through the use of any proper prior distribution. It does arise however by formal use of the prior specification $f(\theta) \propto 1$, which is an example of what is termed an *improper* prior distribution.

So, is it valid to use a posterior distribution obtained by specifying an improper prior to reflect vague knowledge?
The use of improper prior distributions is considered to be acceptable in the following sense.

Improper Priors

The posterior $N(\overline{x}, \frac{\sigma^2}{n})$, obtained by letting $c \to 0$ in the standard conjugate analysis, cannot arise through the use of any proper prior distribution. It does arise however by formal use of the prior specification $f(\theta) \propto 1$, which is an example of what is termed an *improper* prior distribution.

So, is it valid to use a posterior distribution obtained by specifying an improper prior to reflect vague knowledge?
The use of improper prior distributions is considered to be acceptable in the following sense.

If we chose c to be any value other than zero, we would have obtained a perfectly proper prior. Thus, we could choose c arbitrarily close to zero and obtain a posterior arbitrarily close to the one we actually obtained by using the improper prior $f(\theta) \propto 1$.

Representation of Ignorance

We saw that attempting to represent ignorance within the standard conjugate analysis of a Normal mean led to the concept of improper priors.

Representation of Ignorance

We saw that attempting to represent ignorance within the standard conjugate analysis of a Normal mean led to the concept of improper priors.

Another fundamental problem of the prior $f(\theta) \propto 1$ is that it is **not invariant** in 1–1 transformations of the parameter!

Representation of Ignorance

We saw that attempting to represent ignorance within the standard conjugate analysis of a Normal mean led to the concept of improper priors.

Another fundamental problem of the prior $f(\theta) \propto 1$ is that it is **not invariant** in 1–1 transformations of the parameter!

Consider that we might have specified a prior $f_\Theta(\theta)$ for a parameter θ in a model. It is quite reasonable to decide to use instead the parameter $\phi = 1/\theta$. For example, θ may be the variance and ϕ the precision of a Normal distribution. By probability theory the corresponding prior density for ϕ must be given by

$$\begin{aligned} f_\Phi(\phi) &= f_\Theta(\theta) \times \left| \frac{d\theta}{d\phi} \right| \\ &= f_\Theta(1/\phi) \frac{1}{\phi^2}. \end{aligned}$$

Jeffreys' Prior

If we wished to express our ignorance about θ by choosing $f_\Theta(\theta) \propto 1$, then we are forced to take $f_\Phi(\phi) \propto 1/\phi^2$. But if we are ignorant about θ, we are surely equally ignorant about ϕ, and so might equally have made the specification $f_\Phi(\phi) \propto 1$. Thus, prior ignorance as represented by uniformity, is not preserved under re-parameterisation.

Jeffreys' Prior

If we wished to express our ignorance about θ by choosing $f_\Theta(\theta) \propto 1$, then we are forced to take $f_\Phi(\phi) \propto 1/\phi^2$. But if we are ignorant about θ, we are surely equally ignorant about ϕ, and so might equally have made the specification $f_\Phi(\phi) \propto 1$. Thus, prior ignorance as represented by uniformity, is not preserved under re-parameterisation.

There is one way of using the log likelihood $\ell(\theta) = \log f(x \mid \theta)$, to specify a prior which *is* consistent across 1–1 parameter transformations. This is the 'Jeffreys' prior', and is based on the concept of Fisher information:

$$I(\theta) = -E\left\{\frac{d^2\,\ell(\theta)}{d\theta^2}\right\} = E\left\{\left(\frac{d\,\ell(\theta)}{d\theta}\right)^2\right\}.$$

Then, the Jeffreys' prior is defined as $J_\Theta(\theta) \propto |I(\theta)|^{1/2}$.

The Invariance Property

Proposition. $J_\Phi(\phi) = J_\Theta(\theta)\left|\frac{d\theta}{d\phi}\right|$

Substituting the definition of Jeffrey's prior's, and squaring, we need to verify that

$$I(\phi) = I(\theta)\left|\frac{d\theta}{d\phi}\right|^2.$$

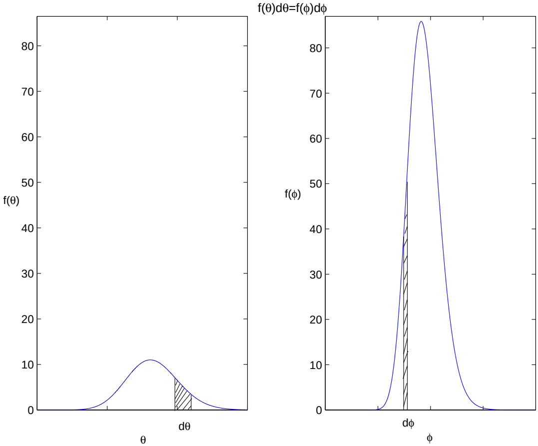
Proof. We have $\ell_\Phi(\phi) = \ell_\Theta(\theta(\phi))$ and

$$\frac{d\ell_\Phi(\phi)}{d\phi} = \frac{d\ell_\Theta(\theta)}{d\theta}\frac{d\theta(\phi)}{d\phi}.$$

Therefore

$$I(\phi) = E\left\{\left(\frac{d\,\ell(\phi)}{d\phi}\right)^2\right\} = E\left\{\left(\frac{d\,\ell(\theta)}{d\theta}\frac{d\theta}{d\phi}\right)^2\right\} = \left(\frac{d\theta}{d\phi}\right)^2 I_\Theta(\theta).$$

Plots of Jeffreys' Prior for a Parameter θ and for $\phi = 1/\theta$



Example. Binomial Sample

Suppose $X|\theta \sim Binomial(n, \theta)$. Then,

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

and

$$\ell(\theta) = \log(f(x|\theta)) = x \log(\theta) + (n-x) \log(1-\theta) + c.$$

Example. Binomial Sample

Suppose $X|\theta \sim Binomial(n, \theta)$. Then,

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

and

$$\ell(\theta) = \log(f(x|\theta)) = x\log(\theta) + (n-x)\log(1-\theta) + c.$$

So,

$$\frac{d\ell(\theta)}{d\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

and

$$\frac{d^2\ell(\theta)}{d\theta^2} = \frac{-x}{\theta^2} - \frac{(n-x)}{(1-\theta)^2},$$

Example. Binomial Sample

Then, Fisher's Information:

$$
\begin{aligned}
I(\theta) &= \frac{n\theta}{\theta^2} + \frac{(n - n\theta)}{(1-\theta)^2} = n\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right) \\
&= n\left(\frac{1-\theta+\theta}{\theta(1-\theta)}\right) = n\theta^{-1}(1-\theta)^{-1},
\end{aligned}
$$

since $E(x) = n\theta$.

Example. Binomial Sample

Then, Fisher's Information:

$$
\begin{aligned}
I(\theta) &= \frac{n\theta}{\theta^2} + \frac{(n - n\theta)}{(1-\theta)^2} = n\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right) \\
&= n\left(\frac{1-\theta+\theta}{\theta(1-\theta)}\right) = n\theta^{-1}(1-\theta)^{-1},
\end{aligned}
$$

since $E(x) = n\theta$.

Jeffreys' prior:

$$J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

which in this case is the $Beta(\frac{1}{2}, \frac{1}{2})$ distribution.

Example. Binomial Sample

Then, Fisher's Information:

$$
\begin{aligned}
I(\theta) &= \frac{n\theta}{\theta^2} + \frac{(n - n\theta)}{(1-\theta)^2} = n\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right) \\
&= n\left(\frac{1-\theta+\theta}{\theta(1-\theta)}\right) = n\theta^{-1}(1-\theta)^{-1},
\end{aligned}
$$

since $E(x) = n\theta$.

Jeffreys' prior:

$$J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

which in this case is the $Beta(\frac{1}{2}, \frac{1}{2})$ distribution.

Which is Jeffreys' prior for $\phi = 1/\theta$ in this case?