Part II

Bayesian Statistics

Chapter 6

Bayesian Statistics: Background

In the frequency interpretation of probability, the probability of an event is limiting proportion of times the event occurs in an infinite sequence of independent repetitions of the experiment. This interpretation assumes that an experiment can be repeated!

Problems with this interpretation:

- Independence is defined in terms of probabilities; if probabilities are defined in terms of independent events, this leads to a circular definition.
- How can we check whether experiments were independent, without doing more experiments?
- In practice we have only ever a finite number of experiments.

6.1 Subjective probability

Let $P(A)$ denote your personal probability of an event A; this is a numerical measure of the strength of your degree of belief that A will occur, in the light of available information.

Your personal probabilities may be associated with a much wider class of events than those to which the frequency interpretation pertains. For example:

- non-repeatable experiments (e.g. that England will win the World Cup next time);
- propositions about nature (e.g. that this surgical procedure results in increased life expectancy).

All subjective probabilities are conditional, and may be revised in the light of additional information. Subjective probabilities are assessments in the light of incomplete information; they may even refer to events in the past.

6.2 Axiomatic development

Coherence states that a system of beliefs should avoid internal inconsistencies. Basically, a quantitative, coherent belief system must behave as if it was governed by a subjective probability distribution. In particular this assumes that all events of interest can be compared.

Note: different individuals may assign different probabilities to the same event, even if they have identical background information.

See Chapters 2 and 3 in *Bernardo and Smith* for fuller treatment of foundational issues.

6.3 Bayes Theorem and terminology

Bayes Theorem: Let $B_1, B_2, \ldots, B_k$ be a set of mutually exclusive and exhaustive events. For any event A with $P(A) > 0$,

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{k} P(A|B_j)P(B_j)}.$$

Equivalently we write

$$P(B_i|A) \propto P(A|B_i)P(B_i).$$

Terminology: $P(B_i)$ is the called *prior probability* of B_i; $P(A|B_i)$ is called the *likelihood* of A given B_i; $P(B_i|A)$ is called the *posterior probability* of B_i; $P(A)$ is called the *predictive probability* of A implied by the likelihoods and the prior probabilities.

Example: Two events. Assume we have two events B_1, B_2, then

$$\frac{P(B_1|A)}{P(B_2|A)} = \frac{P(B_1)}{P(B_2)} \times \frac{P(A|B_1)}{P(A|B_2)}.$$

If the data is relatively more probable under B_1 than under B_2, our belief in B_1 compared to B_2 is increased, and conversely.

If in addition $B_2 = B_1^c$, then

$$\frac{P(B_1|A)}{P(B_1^c|A)} = \frac{P(B_1)}{P(B_1^c)} \times \frac{P(A|B_1)}{P(A|B_1^c)}.$$

It follows that:

posterior odds = prior odds $\times$ likelihood ratio.

6.4 Parametric models

A Bayesian statistical model consists of

1. A parametric statistical model $f(x|\theta)$ for the data x, where $\theta \in \Theta$ a parameter; x may be multidimensional.
2. A prior distribution $\pi(\theta)$ on the parameter.

Note: The parameter θ is now treated as random!

The *posterior distribution* of θ given x is

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

Shorter, we write: $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$, or

posterior $\propto$ prior $\times$ likelihood .

Example: normal distribution. Suppose that θ has the normal $\mathcal{N}(\mu, \sigma^2)$-distribution. Then, just focussing on what depends on θ,

$$\begin{aligned}
\pi(\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(\theta-\mu)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(\theta-\mu)^2\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}(\theta^2 - 2\theta\mu + \mu^2)\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\theta^2 + \frac{\mu}{\sigma^2}\theta\right\}.
\end{aligned}$$

Conversely, if there are constants $a > 0$ and b such that

$$\pi(\theta) \propto \exp\left\{-a\theta^2 + b\theta\right\}$$

then it follows that θ has the normal $\mathcal{N}\left(\frac{b}{2a}, \frac{1}{2a}\right)$-distribution.

6.4.1 Nuisance parameters

Let $\theta = (\psi, \lambda)$, where λ is a nuisance parameter. Then $\pi(\theta|x) = \pi((\psi, \lambda)|x)$. We calculate the *marginal posterior* of ψ:

$$\pi(\psi|x) = \int \pi(\psi, \lambda|x) d\lambda$$

and continue inference with this marginal posterior. Thus we just integrate out the nuisance parameter.

6.4.2 Prediction

The *(prior) predictive distribution* of x on the basis π is

$$p(x) = \int f(x|\theta)\pi(\theta) d\theta.$$

Suppose that data x_1 is available, and we want to predict additional data:

$$\begin{aligned} p(x_2|x_1) &= \frac{p(x_2,x_1)}{p(x_1)} \\ &= \frac{\int f(x_2,x_1|\theta)\pi(\theta)d\theta}{\int f(x_1|\theta)\pi(\theta)d\theta} \\ &= \int f(x_2|\theta,x_1)\frac{f(x_1|\theta)\pi(\theta)}{\int f(x_1|\theta)\pi(\theta)d\theta}d\theta \\ &= \int f(x_2|\theta)\frac{f(x_1|\theta)\pi(\theta)}{\int f(x_1|\theta)\pi(\theta)d\theta}d\theta \\ &= \int f(x_2|\theta)\pi(\theta|x_1)d\theta. \end{aligned}$$

Note that x_2 and x_1 are assumed conditionally independent given θ. They are **not**, in general, unconditionally independent.

Example: Billard ball. A billard ball W is rolled from left to right on a line of length 1 with a uniform probability of stopping anywhere on the line. It stops at p. A second ball O is then rolled n times under the same assumptions, and X denotes the number of times that the ball O stopped on the left of W. Given X, what can be said about p?

Our prior is $\pi(p) = 1$ for $0 \le p \le 1$; our model is

$$P(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, \ldots, n.$$

We calculate the predictive distribution, for $x = 0, \ldots, n$

$$\begin{aligned} P(X = x) &= \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp \\ &= \binom{n}{x} B(x+1, n-x+1) \\ &= \binom{n}{x} \frac{x!(n-x)!}{(n+1)!} = \frac{1}{n+1}, \end{aligned}$$

where B is the beta function,

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

We calculate the posterior distribution:

$$\pi(p|x) \quad \propto \quad 1 \times \binom{n}{x} p^x (1-p)^{n-x} \propto p^x (1-p)^{n-x},$$

so

$$\pi(p|x) \quad = \quad \frac{p^x (1-p)^{n-x}}{B(x+1, n-x+1)};$$

this is the $Beta(x+1, n-x+1)$-distribution. In particular the posterior mean is

$$\begin{aligned} E(p|x) &= \int_0^1 p \frac{p^x (1-p)^{n-x}}{B(x+1, n-x+1)} dp \\ &= \frac{B(x+2, n-x+1)}{B(x+1, n-x+1)} = \frac{x+1}{n+2}. \end{aligned}$$

(For comparison: the mle is $\frac{x}{n}$.) Further, $P(O$ stops to the left of W on the next roll $|x)$ is Bernoulli-distributed with probability of success $E(p|x) = \frac{x+1}{n+2}$.

Example: exponential model, exponential prior. Let $X_1, \ldots, X_n$ be a random sample with density $f(x|\theta) = \theta e^{-\theta x}$ for $x \geq 0$, and assume $\pi(\theta) = \mu e^{-\mu\theta}$ for $\theta \geq 0$; and some known μ. Then

$$f(x_1, \ldots, x_n|\theta) = \theta^n e^{-\theta \sum_{i=1}^n x_i}$$

and hence the posterior distribution is

$$\pi(\theta|\mathbf{x}) \quad \propto \quad \theta^n e^{-\theta \sum_{i=1}^n x_i} \mu e^{-\mu\theta} \propto \theta^n e^{-\theta(\sum_{i=1}^n x_i + \mu)},$$

which we recognize as $Gamma(n+1, \mu + \sum_{i=1}^n x_i)$.

Example: normal model, normal prior. Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\theta, \sigma^2)$, where σ^2 is known, and assume that the prior is normal, $\pi(\theta) \sim \mathcal{N}(\mu, \tau^2)$, where μ, τ^2 is known. Then

$$f(x_1, \ldots, x_n|\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\left\{ -\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \theta)^2}{\sigma^2} \right\},$$

so we calculate for the posterior that

$$\begin{aligned}\pi(\theta|\mathbf{x}) &\propto exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2}+\frac{(\theta-\mu)^2}{\tau^2}\right)\right\}\\ &=: e^{-\frac{1}{2}M}.\end{aligned}$$

We can calculate (Exercise)

$$\begin{aligned}M &= a\left(\theta-\frac{b}{a}\right)^2-\frac{b^2}{a}+c,\\ a &= \frac{n}{\sigma^2}+\frac{1}{\tau^2},\\ b &= \frac{1}{\sigma^2}\sum x_i+\frac{\mu}{\tau^2},\\ c &= \frac{1}{\sigma^2}\sum x_i^2+\frac{\mu^2}{\tau^2}.\end{aligned}$$

So it follows that the posterior is normal,

$$\pi(\theta|x)\sim\mathcal{N}\left(\frac{b}{a},\frac{1}{a}\right).$$

Exercise: the predictive distribution for x is $\mathcal{N}(\mu,\sigma^2+\tau^2)$.

Note: The posterior mean for θ is

$$\mu_1 = \frac{\frac{1}{\sigma^2}\sum x_i+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}.$$

If τ^2 is very large compared to σ^2, then the posterior mean is approximately $\overline{x}$. If σ^2/n is very large compared to τ^2, then the posterior mean is approximately μ. = The posterior variance for θ is

$$\phi = \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} < \min\left\{\frac{\sigma^2}{n},\tau^2\right\},$$

which is smaller than the original variances.

6.4.3 Credible intervals

A $(1-\alpha)$ *(posterior) credible interval* is an interval of $\theta-$values within which $1-\alpha$ of the posterior probability lies.

In the above example:

$$P\left(-z_{\alpha/2} < \frac{\theta-\mu_1}{\sqrt{\phi}} < z_{\alpha/2}\right) = 1-\alpha$$

is a $(1-\alpha)$ *(posterior) credible interval* for θ.

The equality is correct conditionally on x, but the r.h.s. does not depend on x, so the equality is also unconditionally correct.

If $\tau^2 \to \infty$ then $\phi - \frac{\sigma^2}{n} \to 0$, and $\mu_1 - \overline{x} \to 0$, and

$$P\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha,$$

which gives the usual $100(1-\alpha)\%$ confidence interval in frequentist statistics.

Note: The interpretation of credible intervals is different to confidence intervals: In frequentist statistics, $\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ applies before $\mathbf{x}$ is observed; the randomness relates to the distribution of $\mathbf{x}$, in Bayesian statistics:,the credible interval is conditional on the observed $\mathbf{x}$; the randomness relates to the distribution of θ.

Chapter 7

Bayesian Models

7.1 Sufficiency

As before, a sufficient statistic captures all the useful information in the data.

Definition: A statistic $t = t(x_1, \ldots, x_n)$ is *parametric sufficient* for θ if

$$\pi(\theta|\mathbf{x}) = \pi(\theta|t(\mathbf{x})).$$

Note: then we also have

$$p(x_{new}|\mathbf{x}) = p(x_{new}|t(\mathbf{x})).$$

Factorization Theorem: $t(\mathbf{x})$ is parametric sufficient if and only if

$$\pi(\theta|\mathbf{x}) = \frac{h(t(\mathbf{x}), \theta)\pi(\theta)}{\int h(t(\mathbf{x}), \theta)\pi(\theta)d\theta}$$

for some function h.

Recall: For classical suffiency, the factorization theorem gave as necessary and sufficient condition that

$$f(\mathbf{x}, \theta) = h(t(\mathbf{x}), \theta)g(\mathbf{x}).$$

Theorem: Classical sufficiency is equivalent to parametric sufficiency.

To see this: assume classical sufficiency, then

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)d\theta} \\ &= \frac{h(t(\mathbf{x}),\theta)g(\mathbf{x})\pi(\theta)}{\int h(t(\mathbf{x}),\theta)g(\mathbf{x})\pi(\theta)d\theta} \\ &= \frac{h(t(\mathbf{x}),\theta)\pi(\theta)}{\int h(t(\mathbf{x}),\theta)\pi(\theta)d\theta}\end{aligned}$$

depends on $\mathbf{x}$ only through $t(\mathbf{x})$, so $\pi(\theta|\mathbf{x}) = \pi(\theta|t(\mathbf{x}))$. Conversely, assume parametric sufficiency, then

$$\frac{f(\mathbf{x}|\theta)\pi(\theta)}{f(\mathbf{x})} = \pi(\theta|\mathbf{x}) = \pi(\theta|t(\mathbf{x})) = \frac{f(t(\mathbf{x})|\theta)\pi(\theta)}{f(t(\mathbf{x}))}$$

and so

$$f(\mathbf{x}|\theta) = \frac{f(t(\mathbf{x})|\theta)}{f(t(\mathbf{x}))} f(\mathbf{x})$$

which implies classical sufficiency.

Example: Recall that a *k-parameter exponential family* is given by

$$f(x|\theta) = \exp\left\{\sum_{i=1}^{k} \phi_i(\theta)h_i(x) + c(\theta) + d(x),\right\}, \quad x \in \mathcal{X},$$

where $c(\theta)$ is chosen such that $\int f(x|\theta)\,dx = 1$. The family is called *regular* if $\mathcal{X}$ does not depend on θ; otherwise it is called *non-regular*. In k-parameter exponential family models,

$$t(\mathbf{x}) = \left(n, \sum_{j=1}^{n} h_1(x_j), \ldots, \sum_{j=1}^{n} h_k(x_j)\right)$$

is sufficient.

7.2 Exchangeability

$X_1, \ldots, X_n$ are *(finitely) exchangeable* if

$$P(X_1 \in E_1, \ldots, X_n \in E_n) = P(X_{\sigma(i)} \in E_1, \ldots, X_{\sigma(n)} \in E_n)$$

for any permutation σ of $\{1, 2, \ldots, n\}$, and any (measureable) sets $E_1, \ldots, E_n$.

An infinite sequence $X_1, X_2, \ldots$ is *exchangeable* if every finite sequence is (finitely) exchangeable.

Intuitively: a random sequence is exchangeable if the random quantities do not arise, for example, in a time ordered way.

Every independent sequence is exchangeable, but NOT every exchangeable sequence is independent.

Example: A simple random sample from a finite population (sampling without replacement) is exchangeable, but not independent.

Theorem (de Finetti). If $X_1, X_2, \ldots$ is *exchangeable*, with probability measure P, then there exists a prior measure Q on the set of all distributions (on the real line) such that, for any n, the joint distribution function of $X_1, \ldots X_n$ has the form

$$\int \prod_{i=1}^{n} F(x_i) dQ(F),$$

where the integral is over all distributions F, and

$$Q(E) = \lim_{n \to \infty} P(F_n \in E),$$

where $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leq x)$ is the empirical c.d.f. of $X_1, X_2, \ldots X_n$.

Thus each exchangeable sequence arises from a 2-stage randomization:
(a) pick F according to Q;
(b) conditional on F, the observations are i.i.d. .

De Finetti's Theorem tells us that subjective beliefs which are consistent with (infinite) exchangeability must be of the form
(a) There are beliefs (a priori) on the "parameter" F; representing your expectations for the behaviour of $X_1, X_2, \ldots$;
(b) conditional on F the observations are i.i.d. .
In Bayesian statistics, we can think of the prior distribution on the parameter θ as such an F. If a sample is i.i.d. given θ, then the sample is exchangeable.

For further reading on de Finetti's Theorem, see Steffen Lauritzen's graduate lecture at

`http://www.stats.ox.ac.uk/ steffen/teaching/grad/definetti.pdf`

Example: exchangeable Bernoulli random variables. Suppose $X_1, X_2, \ldots$ are exchangeable 0-1 variables. Then the distribution of X_i is uniquely defined by $p = P(X_i = 1)$; the set of all probability distributions on $\{0, 1\}$ is equivalent to the interval $[0, 1]$. Hence Q puts a probability on $[0, 1]$; de Finetti' Theorem gives that

$$p(x_1, \ldots, x_n) = \int p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} dQ(p).$$

Furthermore, $Q(p) = P(Y \leq p)$, where, (in probability),

$$Y = lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Not all finitely exchangeable sequences can be imbedded in an infinite exchangeable sequence.

Exercise: X_1, X_2 such that

$$P(X_i = 1, X_2 = 0) = P(X_1 = 0, X_2 = 1) = \frac{1}{2}$$

cannot be embedded in an exchangeable sequence X_1, X_2, X_3.

Example: Meta-Modelling. Assume that our sample enjoys both exchangeability and spherical symmetry: $X_1, \ldots, X_n$ enjoy *spherical symmetry* if, for any orthogonal matrix A, the distribution of $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$ is the same as that of $A(X_1 - \overline{X}, \ldots, X_n - \overline{X})$.

Proposition (*Bernardo and Smith, pp.183-*)
Let $X_1, X_2, \ldots$ be exchangeable, and suppose that, for each n, $X_1, \ldots, X_n$ enjoy *spherical symmetry*. Then the joint density at $(x_1, \ldots, x_n)$ has the form

$$\int_{\mathbf{R}\times\mathbf{R}^+} \prod_{i=1}^{n} \phi\left(\frac{x_i - \mu}{\sigma}\right) \frac{1}{\sigma} dQ(\mu, \sigma)$$

where ϕ is the standard normal density, and Q is some distribution on $\mathbf{R}\times\mathbf{R}^+$. Interpretation: Q gives a joint prior on (μ, σ^2). The proposition then says that, conditional on μ, σ^2 the data are i.i.d. $\mathcal{N}(\mu, \sigma^2)$.

Chapter 8

Prior Distributions

Let Θ be a parameter space. How do we assign prior probabilities on Θ? Recall: we need a coherent belief system.

In a *discrete parameter space* we assign subjective probabilities to each element of the parameter space, which is in principle straightforward. In a *Continuous parameter space* there are three main approaches:

1. 1. Histogram approach: Say, Θ is an interval of $\mathbf{R}$. We can discretize Θ, assess the total mass assigned to each subinterval, smooth the histogram.

2. 2. Relative likelihood approach: Again, say, Θ is an interval of $\mathbf{R}$. We assess the relative likelihood that θ will take specific values. This relative likelihood is proportional to the prior density. If Θ is unbounded, normalization can be an issue.

3. 3. Particular functional forms: conjugate priors

8.1 Conjugate priors and regular exponential families

A family $\mathcal{F}$ of prior distributions for θ is *closed under sampling* from a model $f(x|\theta)$ if for every prior distribution $\pi(\theta) \in \mathcal{F}$, the posterior $\pi(\theta|x)$ is also in $\mathcal{F}$. When this happens, the common parametric form of the prior and posterior are called a *conjugate prior family* for the problem. Then we also

say that the family $\mathcal{F}$ of prior distributions is *conjugate* to this class of models $\{f(x|\theta), \theta \in \Theta\}$. Often we abuse notation and call an element in the family of conjugate priors a *conjugate prior* itself.

Example: Normal. We have seen already: if $X \sim \mathcal{N}(\theta, \sigma^2)$ with σ^2 known; and if $\theta \sim \mathcal{N}(\mu, \tau^2)$, then the posterior for θ is also normal. Thus the family of normal distributions forms a conjugate prior family for this normal model.

If

$$f(x|\theta) = \exp\left\{\sum_{i=1}^{k} \phi_i(\theta)h_i(x) + c(\theta) + d(x),\right\}, \quad x \in \mathcal{X},$$

is in a regular k-parameter exponential family, then the family of priors of the form

$$\pi(\theta|\tau) = (K(\tau))^{-1} exp\{\sum_{i=1}^{k} \tau_i \phi_i(\theta + \tau_0 c(\theta)\},$$

where $\tau = (\tau_0, \ldots, \tau_k)$ is such that

$$K(\tau) = \int_\Theta \exp\{\sum_{i=1}^{k} \tau_i \phi_i(\theta) + \tau_0 c(\theta)\} d\theta < \infty,$$

is a conjugate prior family. The parameters τ are called *hyperparameters*.

Example: Bernoulli distribution: Beta prior. For a Bernoulli random sample,

$$f(x|\theta) \;=\; \theta^x(1-\theta)^{1-x} = \exp\left\{x \log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right\};$$

We have seen that we can choose $k = 1$ and

$$\begin{aligned} h(x) &= x, \quad \phi(\theta) = \log\left(\frac{\theta}{1-\theta}\right), \\ c(\theta) &= \log(1-\theta), \quad d(x) = 0. \end{aligned}$$

Then

$$\begin{aligned} \pi(\theta|\tau) &\propto \exp\left\{\tau_1 \log\left(\frac{\theta}{1-\theta}\right) + \tau_0 \log(1-\theta)\right\} \\ &= \frac{1}{K(\tau_0, \tau_1)} \theta^{\tau_1}(1-\theta)^{\tau_0 - \tau_1}. \end{aligned}$$

This density will have a finite integral if and only if $\tau_1 > -1$ and $\tau_0 - \tau_1 > -1$, in which case it is the $Beta(\tau_1 + 1, \tau_0 - \tau_1 + 1)$-distribution. Thus the family of Beta distributions forms a conjugate prior family for the Bernoulli distribution.

Example: Poisson distribution: Gamma prior. Here,

$$d(x) = -\log(x!), \quad c(\theta) = -\theta, \quad h(x) = x, \quad \phi(\theta) = \log\theta,$$

and an element of the family of conjugate priors is given by

$$\pi(\theta|\tau) = \frac{1}{K(\tau_0, \tau_1)}\theta^{\tau_1}e^{-\theta\tau_0}.$$

The density will have a finite integral if and only if $\tau_1 > -1$ and $\tau_0 > 0$, in which case it is the $Gamma(\tau_1 + 1, \tau_0)$ distribution.

Example: Normal, unknown variance: normal-gamma prior. For the normal distribution with mean μ, we let the *precision* be $\lambda = \sigma^{-2}$, then

$$\begin{aligned} d(x) &= -\frac{1}{2}\log(2\pi), \quad c(\mu, \lambda) = -\frac{\lambda\mu^2}{2} + \frac{1}{2}\log\lambda \\ h(x) &= (x, x^2), \quad \phi(\mu, \lambda) = \left(\mu\lambda, -\frac{1}{2}\lambda\right) \end{aligned}$$

and an element of the family of conjugate priors is given by

$$\begin{aligned} &\pi(\mu, \lambda|\tau_0, \tau_1, \tau_2) \\ &\quad \propto \lambda^{\frac{\tau_0}{2}} exp\left\{-\frac{1}{2}\lambda\mu^2\tau_0 + \lambda\mu\tau_1 - \frac{1}{2}\lambda\tau_2\right\} \\ &\quad \propto \lambda^{\frac{\tau_0+1}{2}-1} exp\left\{-\frac{1}{2}\left(\tau_2 - \frac{\tau_1^2}{\tau_0}\right)\lambda\right\} \lambda^{\frac{1}{2}} exp\left\{-\frac{\lambda\tau_0}{2}\left(\mu - \frac{\tau_1}{\tau_0}\right)^2\right\}. \end{aligned}$$

The density can be interpreted as follows: Use a $Gamma((\tau_0+1)/2, (\tau_2 - \tau_1^2/\tau_0)/2)$ prior for λ. Conditional on λ, we have a normal $\mathcal{N}(\tau_1/\tau_0, 1/(\lambda\tau_0))$ for μ. This is called a *normal-gamma distribution* for (μ, λ); it will have a finite integral if and only if $\tau_2 > \tau_1^2/\tau_0$ and $\tau_0 > 0$.

Fact: In **regular k-parameter exponential family models**, the family of conjugate priors is closed under sampling. Moreover, if $\pi(\theta|\tau_0, \ldots, \tau_k)$ is in the above conjugate prior family, then

$$\begin{aligned}&\pi(\theta|x_1, \ldots, x_n, \tau_0, \ldots, \tau_k)\\ &\quad = \quad \pi(\theta|\tau_0 + n, \tau_1 + H_1(\mathbf{x}), \ldots, \tau_k + H_k(\mathbf{x})),\end{aligned}$$

where

$$H_i(\mathbf{x}) = \sum_{j=1}^{n} h_i(x_j).$$

Recall that $(n, H_1(\mathbf{x}), \ldots, H_k(\mathbf{x}))$ is a sufficient statistic. Also note that indeed the posterior is of the same parametric form as the prior.

The posterior predictive density for future observations $\mathbf{y} = (y_1, \ldots, y_m)$ given the data $\mathbf{x} = (x_1, \ldots, x_n)$ is

$$\begin{aligned}&p(\mathbf{y}|x_1, \ldots, x_n, \tau_0, \ldots, \tau_k)\\ &\quad = \quad p(\mathbf{y}|\tau_0 + n, \tau_1 + H_1(\mathbf{x}), \ldots, \tau_k + H_k(\mathbf{x}))\\ &\quad = \quad \frac{K(\tau_0 + n + m, \tau_1 + H_1(\mathbf{x}, \mathbf{y}), \ldots, \tau_k + H_k(\mathbf{x}, \mathbf{y}))}{K(\tau_0 + n, \tau_1 + H_1(\mathbf{x}), \ldots, \tau_k + H_k(\mathbf{x}))} \exp\left\{\sum_{\ell=1}^{m} d(y_\ell)\right\},\end{aligned}$$

where

$$H_i(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{n} h_i(x_j) + \sum_{\ell=1}^{m} h_i(y_\ell).$$

This form is particularly helpful for inference: The effect of the data $x_1, \ldots, x_n$ is that the labelling parameters of the posterior are changed from those of the prior, $(\tau_0, \ldots, \tau_k)$ by simply adding the sufficient statistics

$$(t_0, \ldots, t_k) = \left(n, \sum_{i=1}^{n} h_1(x_j), \ldots, \sum_{i=1}^{n} h_k(x_j)\right)$$

to give the parameters $(\tau_0 + t_0, \ldots, \tau_k + t_k)$ for the posterior.

Mixtures of priors from this conjugate prior family also lead to a simple analysis (Exercise).

8.2 Noninformative priors

Often we would like a prior that favours no particular values of the parameter over others. If is Θ finite with $|\Theta| = n$, then we just put mass $\frac{1}{n}$ at each parameter value. If Θ is infinite, there are several ways in which one may seek a noninformative prior.

8.2.1 Improper priors

These are priors which do not integrate to 1. They are interpreted in the sense that *posterior* $\propto$ *prior* $\times$ *likelihood*. Often they arise as natural limits of proper priors. They have to be handled carefully in order not to create paradoxes!

Example: Binomial, Haldane's prior. Let $X \sim Bin(n, p)$, with n fixed, and let π be a prior on p. *Haldane's prior* is given by

$$\pi(p) \propto \frac{1}{p(1-p)};$$

note that $\int_0^1 \pi(p)\, dp = \infty$. This prior gives as marginal density

$$p(x) \propto \int_0^1 (p(1-p))^{-1} \binom{n}{x} p^x (1-p)^{n-x} dp,$$

which is not defined for $x = 0$ or $x = n$. For all other values of x we obtain the $Beta(x, n-x)$-distribution. The posterior mean is

$$\frac{x}{x+n-x} = \frac{x}{n}.$$

For $x = 0$ or n: heuristically,

$$\pi \approx \pi_{\alpha,\beta} \sim Beta(\alpha, \beta),$$

where $\alpha, \beta > 0$ are small. Then the posterior is $Beta(\alpha + x, \beta + n - x)$. Now let $\alpha, \beta \downarrow 0$: then $\pi_{\alpha,\beta}$ converges to the distribution π. Hence the posterior mean converges to x/n for *all* values of x.

8.2.2 Noninformative priors for location parameters

Let $\Theta, \mathcal{X}$ be subsets of Euclidean space. Suppose that $f(x|\theta)$ is of the form $f(x-\theta)$: this is called a *location density*, θ is called *location parameter*. An example is the normal distribution with known variance; θ is the mean.

For a noninformative prior for θ, suppose that we observe $y = x + c$, where c fixed. If $\eta = \theta + c$ then y has density $f(y|\eta) = f(y-\eta)$, and so a noninformative prior should be the same (as we assume that we have the same parameter space).

Call π the prior for the (x, θ)-problem, and π^* the prior for the (y, η)-problem. Then we want that for any (measureable) set A

$$\int_A \pi(\theta)d\theta \;=\; \int_A \pi^*(\theta)d\theta = \int_{A-c} \pi(\theta)d\theta = \int_A \pi(\theta - c)d\theta$$

yielding

$$\pi(\theta) = \pi(\theta - c)$$

for all θ. Hence $\pi(\theta) = \pi(0)$ is constant; usually we choose $\pi(\theta) = 1$ for all θ. This is an improper prior.

8.2.3 Noninformative priors for scale parameters

A (one-dimensional) *scale density* is a density of the form

$$f(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$$

for $\sigma > 0$; σ is called *scale parameter*. An example is the normal distribution with zero mean; σ is the standard deviation.

For a noninformative prior for σ, we use a similar argument as above. Consider $y = cx$, for $c > 0$; put $\eta = c\sigma$, then y has density $f(y|\eta) = \frac{1}{\eta} f\left(\frac{x}{\sigma}\right)$. The noninformative priors for σ and η should be the same (assuming the same parameter space), so we want that for any (measureable) set A that $\pi(A) = \pi(A/c)$, i.e.

$$\int_A \pi(\sigma)d\sigma \;=\; \int_{c^{-1}A} \pi(\sigma)d\sigma = \int_A \pi(c^{-1}\sigma)c^{-1}d\sigma$$

yielding

$$\pi(\sigma) = c^{-1}\pi(c^{-1}\sigma)$$

for all $\sigma > 0$; $\pi(c) = c^{-1}\pi(1)$; hence $\pi(\sigma) \propto \frac{1}{\sigma}$. Usually we choose $\pi(\sigma) = \frac{1}{\sigma}$. This is an improper prior.

8.2.4 Jeffreys Priors

Example: Binomial model. Let $X \sim Bin(n, p)$.A plausible noninformative prior would be $p \sim U[0, 1]$, but then $\sqrt{p}$ has higher density near 1 than near 0. Thus it seems that "ignorance" about p leads to "knowledge" about $\sqrt{p}$??? We would like the prior to be invariant under reparametrization.

Recall: the *Fisher information* is given by

$$I(\theta) = E_\theta \left[\left(\frac{\partial \log f(x|\theta)}{\partial \theta} \right)^2 \right]$$

and under regularity this equals

$$-E_\theta \left(\frac{\partial^2 \log f(x|\theta)}{\partial^2 \theta} \right).$$

Under the same regularity assumptions, we define the *Jeffreys prior* as

$$\pi(\theta) \propto I(\theta)^{\frac{1}{2}}.$$

This prior may or may not be improper.

To see that this prior is invariant under reparametrisation, let h be monotone and differentiable. The chain rule gives

$$I(\theta) = I(h(\theta)) \left(\frac{\partial h}{\partial \theta} \right)^2.$$

For Jeffreys prior $\pi(\theta)$, we have

$$\pi(h(\theta)) \;=\; \pi(\theta) \left| \frac{\partial h}{\partial \theta} \right|^{-1} \propto I(\theta)^{\frac{1}{2}} \left| \frac{\partial h}{\partial \theta} \right|^{-1} = I(h(\theta))^{\frac{1}{2}};$$

thus the prior is indeed invariant under reparametrization.

The Jeffreys prior favours values of θ for which $I(\theta)$ is large. Hence minimizes the effect of the prior distribution relative to the information in the data.

If θ is multivariate, the Jeffreys prior is

$$\pi(\theta) \propto [detI(\theta)]^{\frac{1}{2}},$$

which is still invariant under reparametrization.

Exercise: The above non-informative priors for scale and location correspond to Jeffreys priors.

Example: Binomial model, Jeffreys prior. Let $X \sim Bin(n,p)$; where n is known, so that $f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$. Then we have alreayd calculated that

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

Thus the Jeffreys prior is

$$\pi(p) \propto (p(1-p))^{-\frac{1}{2}},$$

which we recognize as the $Beta(1/2, 1/2)$-distribution. This is a proper prior.

Example: Normal model, Jeffreys prior. Let $X \sim \mathcal{N}(\mu, \sigma^2)$, where $\theta = (\mu, \sigma)$. We abbreviate

$$\psi(x, \mu, \sigma) = -\log\sigma - \frac{(x-\mu)^2}{2\sigma^2};$$

then

$$\begin{aligned} I(\theta) &= -E_\theta \begin{pmatrix} \frac{\partial^2}{\partial\mu^2}\psi(x,\mu,\sigma) & \frac{\partial^2}{\partial\mu\partial\sigma}\psi(x,\mu,\sigma) \\ \frac{\partial^2}{\partial\mu\partial\sigma}\psi(x,\mu,\sigma) & \frac{\partial^2}{\partial\sigma^2}\psi(x,\mu,\sigma) \end{pmatrix} \\ &= -E_\theta \begin{pmatrix} -\frac{1}{\sigma^2} & \frac{2(x-\mu)}{\sigma^3} \\ \frac{2(x-\mu)}{\sigma^3} & \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \end{aligned}$$

So

$$\pi(\theta) \propto \left(\frac{1}{\sigma^2} \times \frac{2}{\sigma^2}\right)^{\frac{1}{2}} \propto \frac{1}{\sigma^2}$$

is the Jeffreys prior.

Note: $\mathcal{N}(\mu, \sigma^2)$ is a location-scale density, so we could take a uniform prior for μ, $1/\sigma$-prior for σ; this would yield

$$\pi(\theta) = \frac{1}{\sigma}.$$

Thie prior is *not* equal to the Jeffreys prior.

8.2.5 Maximum Entropy Priors

The discrete case

Assume first that Θ is discrete. The *entropy* of π is defined as

$$\mathcal{E}(\pi) = -\sum_{\Theta} \pi(\theta_i) log(\pi(\theta_i)),$$

where $0 \log 0 = 0$. It measures the amount of uncertainty in an observation. If Θ is finite, with n elements, then $\mathcal{E}(\pi)$ is largest for the uniform distribution, and smallest if $\pi(\theta_i) = 1$ for some $\theta_i \in \Theta$.

Suppose that we are looking for a prior π, and we would like to take partial information in terms of functions $g_1, \ldots, g_m$ into account, where this partial information can be written as

$$E_\pi g_k(\theta) = \mu_k, \quad k = 1, \ldots, m.$$

We would like to choose the distribution with the maximum entropy under these constraints: This distribution is

$$\tilde{\pi}(\theta_i) = \frac{\exp\left(\sum_{k=1}^m \lambda_k g_k(\theta_i)\right)}{\sum_i \exp\left(\sum_{k=1}^m \lambda_k g_k(\theta_i)\right)},$$

where the λ_i are determined by the constraints.

Example: prior information on the mean. Let $\Theta = \{0, 1, 2, \ldots\}$. The prior mean of θ is thought to be 5. We write this as a constraint: $m = 1, g_1(\theta) = \theta, \mu_1 = 5$. So

$$\tilde{\pi}(\theta) \;=\; \frac{e^{\lambda_1 \theta}}{\sum_{j=0}^{\infty} e^{\lambda_1 j}} = \left(e^{\lambda_1}\right)^\theta \left(1 - e^{\lambda_1}\right)$$

is the maximum-entropy prior, where the last equality assumes that $\lambda_1 < 0$. We recognize it as the distribution of (geometric - 1). Its mean is $e^{-\lambda_1} - 1$, so setting the mean equal to 5 yields $e^{\lambda_1} = \frac{1}{6}$, and $\lambda_1 = -\log 6$. So the maximum-entropy prior under the constraint that the mean is 5 is

$$\pi(\theta) = \left(\frac{1}{6}\right)^{\theta} \frac{5}{6}; \quad \theta = 0, 1, \ldots$$

The continuous case

If Θ *is continuous*, then $\pi(\theta)$ is a density. The *entropy* of π *relative* to a particular reference distribution with density π_0 is defined as

$$\mathcal{E}(\pi) \;=\; -E_\pi\left(\log\frac{\pi(\theta)}{\pi_0(\theta)}\right) = -\int_\Theta \pi(\theta) log \frac{\pi(\theta)}{\pi_0(\theta)} d\theta.$$

(A discrete Θ corresponds to π_0 being discrete uniform.) As reference distribution π_0 we usually choose the "natural" invariant noninformative prior.

Now assume that we have partial information in terms of functions $g_1, \ldots, g_m$:

$$\int g_k(\theta)\pi(\theta)d\theta = \mu_k, \quad k = 1, \ldots, m.$$

We choose the distribution with the maximum entropy under these constraints (when it exists):

$$\tilde{\pi}(\theta) = \frac{\pi_0(\theta)\exp\left(\sum_{k=1}^m \lambda_k g_k(\theta)\right)}{\int_\Theta \pi_0(\theta)\exp\left(\sum_{k=1}^m \lambda_k g_k(\theta)\right) d\theta},$$

where the λ_i are determined by the constraints.

Example: location parameter, known mean, known variance. Let $\Theta = \mathbf{R}$, and let θ be location parameter; we choose as reference prior $\pi_0(\theta) = 1$. Suppose that mean and variance are known:

$$g_1(\theta) = \theta, \mu_1 = \mu; \quad g_2(\theta) = (\theta - \mu)^2, \mu_2 = \sigma^2.$$

Then we choose

$$\tilde{\pi}(\theta) \;=\; \frac{\exp\left(\lambda_1\theta + \lambda_2(\theta-\mu)^2\right)}{\int_{-\infty}^{\infty}\exp\left(\lambda_1\theta + \lambda_2(\theta-\mu)^2\right) d\theta} \propto \exp(\lambda_1\theta + \lambda_2\theta^2) \propto \exp(\lambda_2(\theta-\alpha)^2),$$

for a suitable α (here the λ's may not be the same). So $\tilde{\pi}$ is normal; the constraints give $\tilde{\pi}$ is $\mathcal{N}(\mu, \sigma^2)$.

Example: location parameter, known mean. Suppose that in the previous example only the prior mean, not the prior variance, is specified. Then

$$\tilde{\pi}(\theta) \;=\; \frac{\exp(\lambda_1\theta)}{\int_{-\infty}^{\infty}\exp(\lambda_1\theta)\,d\theta},$$

and the integral is infinite, so the distribution does not exist.

8.3 Additional Material: Bayesian Robustness

To check how much the conclusions change for different priors, we can carry out a sensitivity analysis.

Example: Normal or Cauchy? Suppose $\Theta = \mathbf{R}$, and $X \sim \mathcal{N}(\theta, 1)$, where θ is known to be either normal or Cauchy. We calculate the posterior means under both models:

obs. x	post. mean ($\mathcal{N}$)	post. mean (C)
0	0	0
1	0.69	0.55
2	1.37	1.28
4.5	3.09	4.01
10	6.87	9.80

For small x the posterior mean does not change very much, but for large x it does.

Example: Normal model; normal or contaminated class of priors. Assume that $X \sim \mathcal{N}(\theta, \sigma^2)$, where σ^2 known, and $\pi_0 \sim \mathcal{N}(\mu, \tau^2)$. An alternative class of priors is Γ, constructed as follows: Let

$$Q = \{q_k; q_k \sim \mathcal{U}|(\mu - k, \mu + k)\},$$

then put

$$\Gamma = \{\pi : \pi = (1-\epsilon)\pi_0 + \epsilon q, \text{ some } q \in Q\}.$$

The Γ is called an *ϵ-contamination class of priors*. Let $C = (c_1, c_2)$ be an interval. Put
$P_0 = P(\theta \in C|x, \pi_0), \quad Q_k = P(\theta \in C|x, q_k)$
Then (by Bayes' rule) for $\pi \in \Gamma$ we hvae

$$P(\theta \in C|x) \;\; = \;\; \lambda_k(x)P_0 + (1-\lambda_k(x))Q_k,$$

where

$$\lambda_k(x) = \left(1 + \frac{\epsilon}{1-\epsilon} \times \frac{p(x|q_k)}{p(x|\pi_0)}\right)^{-1},$$

and $p(x|q)$ is the predictive density of x when the prior is q. The predictive density $p(x|\pi_0)$ is $\mathcal{N}(\mu, \sigma^2 + \tau^2)$,

$$p(x|q_k) = \int_{\mu-k}^{\mu+k} \frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)\frac{1}{2k}d\theta$$

and

$$Q_k = \frac{1}{p(x|q_k)}\int_{c^*}^{c^{**}} \frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)\frac{1}{2k}d\theta$$

where $c^* = \max\{c, \mu-k\}$ and $c^{**} = \min\{c_2, \mu+k\}$ (ϕ is the standard normal density).

Numerical example: Let $\epsilon = 0.1, \sigma^2 = 1, \tau^2 = 2, \mu = 0, x = 1$, and $C = (-0.93, 2.27)$ is the 95% credible region for π_0. Then we calculate

$$\inf_{\pi\in\Gamma} P(\theta \in C|x, \pi) = 0.945,$$

achieved at $k = 3.4$, and

$$\sup_{\pi\in\Gamma} P(\theta \in C|x, \pi) = 0.956,$$

achieved at $k = 0.93$. So in this sense the inference is very robust.

Chapter 9

Posterior Distributions

9.1 Point estimates

Some natural summaries of distributions are the mean, the median, and the mode. The mode is most likely value of θ, so it is the *maximum Bayesian likelihood estimator*. For summarizing spread, we use the inter-quartile range (IQR), the variance, etc.

9.2 Interval estimates

If π is the density for a parameter $\theta \in \Theta$, then a region $C \subset \Theta$ such that

$$\int_C \pi(\theta) d\theta = 1 - \alpha$$

is called a $100(1-\alpha)\%$ *credible region* for θ with respect to π. If C is an interval: it is called a *credible interval*. Here we take credible regions with respect to the posterior distribution; we abbreviate the posterior by π, abusing notation.

A $100(1-\alpha)\%$ credible region is not unique. Often it is natural to give the smallest $100(1-\alpha)\%$ credible region, especially when the region is an interval. We say that $C \subset \Theta$ is a $100(1-\alpha)\%$ *highest posterior density region* (HPD) with respect to π if

(i) $\int_C \pi(\theta) d\theta = 1 - \alpha$; and

(ii) $\pi(\theta_1) \geq \pi(\theta_2)$ for all $\theta_1 \in C, \theta_2 \notin C$ except possibly for a subset of Θ having π-probability 0.
A $100(1-\alpha)\%$ HPD has minimum volume over all $100(1-\alpha)\%$ credible regions.

The full posterior distribution itself is often more informative than credible regions, unless the problem is very complex.

9.3 Asymptotics

When n is large, then under suitable regularity conditions, the posterior is approximately normal, with mean the m.l.e. $\hat{\theta}$, and variance $(nI(\hat{\theta}))^{-1}$. This asymptotics requires that the prior is non-zero in a region surrounding the m.l.e..

Since, under regularity, the m.l.e. is consistent, it follows that if the data are i.i.d. from $f(x|\theta_0)$ and if the prior is non-zero around θ_0, then the posterior will become more and more concentrated around θ_0. In this sense Bayesian estimation is automatically consistent.