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.

Continuous parameter space

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. 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. Particular functional forms: conjugate priors

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: 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.

Regular k-parameter exponential family

If

$$f(x|\theta) = f(x)g(\theta)exp\{\sum_{i=1}^{k} c_i\phi_i(\theta)h_i(x)\}, \quad x \in \mathcal{X}$$

then the family of priors of the form

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

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

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

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

Example: Bernoulli distribution: Beta prior

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

SO

$$\begin{aligned} f(x) &= 1, \quad g(\theta) = 1-\theta, \quad h(x) = x, \\ \phi(\theta) &= \log\left(\frac{\theta}{1-\theta}\right), \quad c = 1 \\ \pi(\theta|\tau) &\propto (1-\theta)^{\tau_0}exp\left\{\tau_1\log\left(\frac{\theta}{1-\theta}\right)\right\} \\ &= \frac{1}{K(\tau_0,\tau_1)}\theta^{\tau_1}(1-\theta)^{\tau_0-\tau_1}. \end{aligned}$$

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

$$f(x) = (x!)^{-1}, \quad g(\theta) = e^{-\theta}$$

$$h(x) = x, \quad \phi(\theta) = \log\theta, \quad c = 1$$

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

$$f(x) = (2\pi)^{-\frac{1}{2}}, \quad g(\mu, \lambda) = \sqrt{\lambda} e^{-\frac{\lambda \mu^2}{2}}$$

$$h(x) = (x, x^2), \quad \phi(\mu, \lambda) = (\mu\lambda, \lambda)$$

$$c_1 = 1, \quad c_2 = -\frac{1}{2}$$

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

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

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

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

where

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

Recall: $(n, H_1(x), \ldots, H_k(x))$ is a sufficient statistic.

Note that indeed the posterior is of the same parametric form as the prior.

The predictive density for future observations $\mathbf{y} = y_1, \ldots, y_m$ is

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

where

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

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

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

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 (see the next homework sheet).

Noninformative priors

Often we would like a prior that favours no particular values of the parameter over others.

If 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.

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

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

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 $x = 0, n$: think

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

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 π. Note that the posterior mean converges to x/n for *all* values of x.

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*.

For a noninformative prior for θ: suppose that we observe $y = x + c$, where c fixed. If $\eta = \theta + c$ then y has density $f(y-\eta)$, and so a noninformative priors 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

$$\begin{aligned} \int_A \pi(\theta) d\theta &= \int_A \pi^*(\theta) d\theta = \int_{A-c} \pi(\theta) d\theta \\ &= \int_A \pi(\theta - c) d\theta \end{aligned}$$

yielding

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

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

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*.

Consider $y = cx$, for $c > 0$; put $\eta = c\sigma$, then y has density $\frac{1}{\eta} f\left(\frac{x}{\sigma}\right)$.

Noninformative priors for σ and η should be the same (assuming the same parameter space), so we want that for any (measureable) set A

$$\begin{aligned}\int_A \pi(\sigma)d\sigma &= \int_{c^{-1}A} \pi(\sigma)d\sigma \\ &= \int_A \pi(c^{-1}\sigma)c^{-1}d\sigma\end{aligned}$$

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.

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 "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.

Reparametrization:

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

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

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.

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

$$\frac{\partial^2 log f(x|p)}{\partial^2 p} = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}.$$

Take expectation and multiply by minus 1:

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

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.

If θ is multivariate, the Jeffreys prior is

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

which is still invariant under reparametrization.

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.

Maximum Entropy Priors

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 $0log0 = 0$). 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 we are looking for a prior π, taking partial information in terms of functions $g_1, \ldots, g_m$ into account, where this partial in-

formation 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

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

is the maximum-entropy prior. 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 $\lambda_1 = -\log 6$.

If Θ *is continuous*:

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

$$\begin{aligned} \mathcal{E}(\pi) &= -E_\pi \left(\log \frac{\pi(\theta)}{\pi_0(\theta)} \right) \\ &= -\int_\Theta \pi(\theta) log \frac{\pi(\theta)}{\pi_0(\theta)} d\theta \end{aligned}$$

Θ discrete corresponds to π_0 discrete uniform.

How do we choose π_0? 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

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

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; so

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

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

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.