Hierarchical models These are models where the prior has parameters which again have a probability distribution.

1. Data x have a density $f(x;\theta)$.
2. The prior distribution of θ is $\pi(\theta;\psi)$.
3. ψ has a prior distribution $g(\psi)$, for $\psi \in \Psi$.
4. We can work with posterior $\pi(\theta,\psi|x) \propto f(x;\theta)\pi(\theta;\psi)g(\psi)$, or

$$\begin{aligned}\pi(\theta|x) &= \int_\Psi \pi(\theta,\psi|x)d\psi \\ &\propto f(x;\theta)\int_\Psi \pi(\theta;\psi)g(\psi)d\psi \\ &\propto f(x;\theta)\pi(\theta)\end{aligned}$$

The hierarchical model simply specifies the prior for θ indirectly

$$\pi(\theta) = \int \pi(\theta;\psi)g(\psi)d\psi.$$

An example is when $\theta = (\theta_1, \ldots, \theta_k)$ and the parameters are each associated with a subpopulation and have an exchangeable distribution (the labels $i = 1, 2, ..., k$ can be permuted without changing the prior for θ).

Example For $i = 1, 2, ..., k$ we make n_i observations $X_{i,1}, X_{i,2}, ..., X_{i,n_i}$ on population i, with $X_{ij} \sim N(\vartheta_i, \sigma^2)$. The ϑ_i are the unknown means for observations on the i'th population but σ^2 is known.

Suppose the prior model for the ϑ_i is iid normal, $\vartheta_i \sim N(\varphi, \tau^2)$.
If $\psi = (\varphi, \tau^2)$

$$\pi(\theta_1, \ldots, \theta_k; \psi) = \prod_{i=1}^{k} (2\pi\tau^2)^{-1/2} \exp\left\{-\frac{1}{2\tau^2}(\theta_i - \varphi)^2\right\},$$

Now we need a prior for φ and τ^2. Suppose we take

$$g(\varphi, \tau^2) = \text{constant},$$

so all possible φ, τ^2 equally likely *a priori*. [careful!]

The joint posterior of the parameters is

$$\pi(\theta, \psi|x) \propto f(x;\theta)\pi(\theta|\psi)g(\psi)$$

Integration with respect to ψ gives the posterior density of θ.

Here $g(\psi)$ is constant, so the joint posterior distribution is

$$\begin{aligned}\pi(\theta, \varphi, \tau^2|x) \;\propto\; & \left[\prod_{i=1}^{k} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(x_{ij}-\theta_i)^2\right\}\right] \\ \times\; & \left[\prod_{i=1}^{k} \tau^{-1}\exp\left\{-\frac{1}{2\tau^2}(\theta_i-\varphi)^2\right\}\right]\end{aligned}$$

Integrate out wrt φ and τ^2 to obtain the marginal posterior distribution of θ. Integrating the last factor wrt φ gives a term proportional to

$$\tau^{1-k} \exp\left\{-\frac{1}{2\tau^2}\sum(\theta_i - \bar{\theta})^2\right\}$$

Then the integral wrt τ gives a term proportional to

$$\left[\sum(\theta_i - \bar{\theta})^2\right]^{1-k/2}$$

Thus the posterior distribution of θ is

$$\begin{aligned}
\pi(\theta|x) &\propto \left[\prod_{i=1}^{k} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(x_{ij} - \theta_i)^2\right\}\right] \cdot \left[\sum(\theta_i - \bar{\theta})^2\right]^{1-k/2} \\
&\propto \left[\prod_{i=1}^{k} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k} n_i(\theta_i - \bar{x}_i)^2\right\}\right] \cdot \left[\sum(\theta_i - \bar{\theta})^2\right]^{1-k/2}
\end{aligned}$$

where $\bar{x}_i = \sum_j x_{ij}/n_i$.

Let $\widehat{\theta}_j$ be the MAP estimate for θ_j (posterior mode) and put $\widehat{\theta}^* = \sum \widehat{\theta}_j/k$. Differentiate $\pi(\theta|x)$ wrt θ_j and set to equal zero to get

$$\widehat{\theta}_j = (\omega_1 \bar{x}_j + \omega_2 \widehat{\theta}^*)/(\omega_1 + \omega_2),$$

where $\omega_1 = (\sigma^2/n_j)^{-1}$ and

$$\omega_2 = \left[\sum (\widehat{\theta}_i - \widehat{\theta}^*)^2/(k-2)\right]^{-1}.$$

If the θ_j were unrelated $\bar{x}_j$ would be the estimate of θ_j. The model modifies the estimate by pulling it towards the mean of the estimated θ_is.

Constructing priors

Subjective Priors: Write down a distribution representing prior knowledge about the parameter before the data is available. If possible, build a model for the parameter. If different scientists have different priors or it is unclear how to represent prior knowledge as a distribution, then consider several different priors. Repeat the analysis and check that conclusions are insensitive to priors representing 'different points of view'.

Non-Subjective Priors: Several approaches offer the promise of an 'automatic' and even 'objective' prior. We list some suggestions below (Uniform, Jeffreys, MaxEnt). In practice, if one of these priors conflicted prior knowledge, we wouldnt use it. These approaches can be useful to complete the specification of a prior distribution, once subjective considerations have been taken into account.

Uniform priors Priors which take $\pi(\theta) \propto$ constant (ie uniform on Θ) are sometimes called 'non-informative'.

1) This is misleading. Ignorance about θ should imply ignorance about $g(\psi) = \theta$, but $\pi(g(\psi))|g'|$ may not be uniform. Such priors may be highly informative with respect to certain hypotheses and may dominate the likelihood in parameter estimation. This is particularly important for high dimensional parameter vectors.

Example (for point 1): for $0 \leq \vartheta \leq 1$, and $Y \sim U(0,1)$, consider $H_0 : \vartheta > a$ with $a = 0.999$. Now the prior $\pi(\theta) = 1$ weights strongly against H_0 and we expect the posterior to assign H_0 a low probability. If we were considering $H_0 : \vartheta > a$ as a serious hypothesis, a prior assigning probability one half to H_0 will typically better represent the prior state of knowledge.

2) The posterior distribution may not exist: if $\pi(\theta)$ is constant then

$$\pi(\theta|x) = L(\theta; x) / \int_{\Theta} L(\theta; x) d\theta$$

and $\int_{\Theta} L(\theta; x) d\theta$ may not be finite. Such distributions are called improper and all inference is meaningless.

Example (for point 2): $X \sim \text{Exp}(1/\mu)$ and $N = \mathbb{I}(X < 1)$ yielding $N = n$ with $n \in \{0, 1\}$. Suppose we observe $n = 0$. Now

$$L(\mu; n) = \exp(-1/\mu)$$

so if we take $\pi(\mu) \propto 1$ for $\mu > 0$ we have

$$\pi(\mu|n) \propto \exp(-1/\mu)$$

which is improper, as $\pi(\mu|n) \to 1$ as $\mu \to \infty$ so $\int_0^{\infty} \exp(-1/\mu) d\mu$ cannot exist.

3) If the prior is uniform (and possibly improper itself) but the posterior is proper, and it can be shown that all parameter values being equally probable a priori is a fair representation of prior knowledge, then the uniform prior may be useful.

Example (for point 3): we have made occasional use of uniform priors in the examples above. In the very first example (θ was N, the number of red individuals in a population of size m) the uniform prior was natural. The prior on φ, τ^2 in the hierarchical example was improper, but the posterior is proper.

Jeffreys' Priors Jeffreys reasoned as follows. If we have a rule for constructing priors it should lead to the same distribution if we apply it to θ or some other parameterization ψ with $g(\psi) = \theta$. Jeffreys took

$$\pi(\theta) \propto \sqrt{I_\theta} \qquad \text{where} \qquad I_\theta = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right]$$

is the Fisher information. Now if $g(\psi) = \theta$ then

$$\pi_\Psi(\psi) \propto \pi(g(\psi))|g'(\psi)|,$$

so Jeffreys rule should yield $\pi_\Psi(\psi) \propto \sqrt{I_{g(\psi)}}|g'(\psi)|$. The rule gives $\pi_\Psi(\psi) \propto \sqrt{I_\psi}$. But $I_\psi = g'(\psi)^2 I_g(\psi)$, so

$$\sqrt{I_\psi} = \sqrt{I_{g(\psi)}}|g'(\psi)|$$

and the rule is consistent in this respect.

It is sometimes desirable (on subjective grounds) to have a prior which is invariant under reparameterization.

Example $X_1, \ldots, X_n$ is a random sample from $N(\mu, \theta)$, with μ known. Then

$$\ell = \log L(\theta, x) = -\frac{n}{2}\log(2\pi\theta) - \frac{1}{2\theta}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$I_\theta = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] = -\mathbb{E}\left[\frac{n}{2\theta^2} - \frac{2}{2\theta^3}\sum(x_i - \mu)^2\right] = -\frac{n}{2\theta^2} + \frac{n}{\theta^2} = \frac{n}{2\theta^2}$$

Jeffrey's prior

$$\pi(\theta) \propto \left(\frac{n}{2\theta^2}\right)^{1/2} \propto \frac{1}{\theta}$$

The result is scale invariant - if for example σ has units of length, then the prior puts the same weight on $1cm \leq \sigma \leq 10cm$ as it does on $1km \leq \sigma \leq 10km$, so the prior is non-informative with respect to statements about scale. Notice the prior $1/\theta$ is improper.

Exercise: verify that the posterior in the example,

$$\pi(\theta|x) \propto \theta^{-n/2-1} \exp(-\sum_i (\mu - x_i)^2/2\theta), \quad \theta > 0,$$

is proper (Inverse-Gamma density $\theta^{-\alpha-1}e^{-\beta/\theta}, \alpha, \beta > 0$).

Exercise: Binomial (n, θ). Verify

$$I_\theta = \frac{n}{\theta(1-\theta)}$$

so Jeffreys' prior is $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ which is Beta$(\frac{1}{2}, \frac{1}{2})$.

Higher dimensions

If $\Theta \subset \mathbb{R}^k$, and $\ell(\theta; X) = \log(f(X; \theta))$, the Fisher information

$$[I_\theta]_{i,j} = -\mathbb{E}_\theta \left(\frac{\partial^2 \ell(\theta; X)}{\partial \theta_i \partial \theta_j} \right)$$

satisfies

$$-\mathbb{E}_\theta \left(\frac{\partial^2 \ell(\theta; X)}{\partial \theta_i \partial \theta_j} \right) = \mathbb{E}_\theta \left(\frac{\partial \ell(\theta; X)}{\partial \theta_i} \frac{\partial \ell(\theta; X)}{\partial \theta_j} \right)$$

subject to regularity conditions. A k-dimensional Jeffreys' prior

$$\pi(\theta) \propto |I_\theta|^{1/2}$$

($|A| \equiv \det(A)$) is invariant under 1-1 reparameterization.

Exercise Verify 1 to 1 $g(\psi) = \theta$ in R^k gives $\pi_\Psi(\psi) = \sqrt{I_{g(\psi)}} \left| \frac{\partial \theta^T}{\partial \psi} \right|$.

Maximum Entropy Priors (MaxEnt)

Choose a density $\pi(\theta)$ which maximizes the entropy

$$\phi[\pi] = -\int_\Theta \pi(\theta)\log\pi(\theta)d\theta$$

over functions $\pi(\theta)$ subject to constraints on π. This is a Calculus of Variations problem.

Example The distribution π maximizing $\phi[\pi]$ over all densities π on $\Theta = R$, subject to

$$\int_0^\infty \pi(\theta)d\theta = 1, \quad \int_0^\infty \theta\pi(\theta)d\theta = \mu, \quad \text{and} \quad \int_0^\infty (\theta-\mu)^2\pi(\theta)d\theta = \sigma^2,$$

(normalized with $\mathbb{E}\vartheta = \mu$ and $\mathsf{Var}(\vartheta) = \sigma^2$) is the normal density

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(\theta-\mu)^2/2\sigma^2}.$$

This is a special case of the following Theorem.

Theorem The density $\pi(\theta)$ that maximizes $\phi(\pi)$, subject to

$$\mathbb{E}[t_j(\theta)] = t_j,\ j = 1, \ldots, p$$

takes the p-parameter exponential family form

$$\pi(\theta) \propto \exp\left\{\sum_{i=1}^{p} \lambda_i t_i(\theta)\right\}$$

for all $\theta \in \Theta$, where $\lambda_1, \ldots, \lambda_p$ are determined by the constraints. (For the proof see Leonard and Hsu).

Example In the normal case $t_1(\theta) = \theta$, $t_1 = \mu$, $t_2(\theta) = (\theta - \mu)^2$, $t_2 = \sigma^2$ gives $\pi(\theta) \propto \exp(\lambda_1\theta + \lambda_2(\theta - \mu)^2)$. Impose the constraints to get $\lambda_1 = 0$ and $\lambda_2 = -1/2\sigma^2$.

Example Suppose prior probabilities are specified so that

$$P(a_{j-1} < \vartheta \leq a_j) = \phi_j, j = 1, \ldots, p$$

with $\sum_j \phi_j = 1$ and

$$\vartheta \in (a_0, a_p),\ a_0 \leq a_1 \cdots \leq a_p \leq a_p.$$

We find the maximum entropy distribution subject to these conditions. The conditions are equivalent to

$$\mathbb{E}[t_j(\vartheta)] = \phi_j, j = 1, \ldots, p$$

where $t_j(\vartheta) = \mathbb{I}[a_{j-1} < \vartheta \leq a_j]$. The posterior density of ϑ is

$$\pi(\theta) \propto \exp\left\{\sum_{j=1}^{p} \lambda_j \mathbb{I}[a_{j-1} < \theta \leq a_j]\right\},\ a_0 \leq \theta \leq a_p$$

where $\lambda_1, \ldots, \lambda_p$ are determined by the conditions. $\pi(\theta)$ is a histogram, with intervals $(a_0, a_1], (a_1, a_2], \ldots, (a_{p-1}, a_p]$.

Credible intervals

Let $\theta \in \Theta$ and S_x be a subset of Θ. Let $\pi(\theta|x)$ be a posterior probability distribution for θ given data x. Then S_x is a posterior $1-\alpha$ credible region for θ if

$$\int_{S_x} \pi(\theta \mid x) d\theta = 1 - \alpha$$

When we report the results of a Bayesian analysis, we can report that 'the probability that the unknown true parameter is in S_x, given the prior and data, is $1-\alpha$'.

Exercise: Revisit (GJJ Ex 5.7), which we discussed in connection with classical confidence intervals. Take the improper prior $\pi(\theta) \propto 1, \theta \in R$. Show the posterior is proper, and give a $1-\alpha$ credible region for θ.

Highest posterior density (HPD) credible region

$$\pi(\theta_1 \mid x) \geq \pi(\theta_2 \mid x), \text{ for all } \theta_1 \in S_x, \theta_2 \notin S_x$$

Credible and HPD regions need not be unique.

Theorem The hypervolume (volume in the parameter space) of a $1-\alpha$ HPD credible region is as small as that of any $1-\alpha$ credible region.

Example $X_1, \ldots, X_n$ is a random sample from $N(\mu, \theta)$ with μ known. Take $\pi(\theta) \propto \theta^{-1}$, non-informative with respect to scale.

$$L(\theta; x) = (2\pi)^{-n/2}\theta^{-n/2} \exp\left[-\sum(x_i - \mu)^2/2\theta\right],$$

We would naturally get a credible interval from the quantiles of an Inverse Gamma ($\alpha = n/2$, $\beta = \sum(x_i - \mu)^2/2$).

It is slightly more convenient to read off the posterior distribution

$$\frac{\sum(x_i - \mu)^2}{\vartheta} \sim \chi^2_n$$

since tables of χ^2_n values are more readily available.

Suppose $n = 15$, $\sum(x_i - \mu)^2 = 12.6$. We want an interval estimate of the variance θ. Equal tailed 0.025 and 0.975 critical values of χ^2_{15} are 6.262 and 27.49.

$$\begin{aligned} 0.025 &= P(6.262 > 12.6/\theta) \\ &= P(12.6/\theta > 27.49) \end{aligned}$$

A 95% equal tailed credible interval for θ is (12.6/27.49,12.6/6.262), that is (0.458,2.012).

For the 95% HPD interval we want an interval (a, b) for which

$$a^{-17/2} \exp[-12.6/(2a)] = b^{-17/2} \exp[-12.6/(2b)]$$

so that

$$\pi(a \mid x) = \pi(b \mid x)$$

and for which

$$\int_{12.6/b}^{12.6/a} f(y)dy = 0.95,$$

where f is the χ^2_{15} density. A numerical search gives (0.377,1.769) as the 95% HPD credible interval, 10% shorter than the equal-tailed interval.