7.4. Choosing priors

Important issue – how to reflect the degree of prior belief? How subjective is the choice?

Notation: family of priors $\{\pi_\lambda : \lambda \in \Lambda\}$, selecting a prior means selecting $\lambda_0 \in \Lambda$. ... λ is a hyperparameter.

7.4.1 Conjugate priors

Def. The family of priors $\{\pi_\lambda : \lambda \in \Lambda\}$ for the parameter θ of the model $\{f_\theta : \theta \in \Omega\}$ is conjugate, if for all $s \in S$ and $\lambda \in \Lambda$ the posterior $\pi_\lambda\{\cdot | s\} \in \{\pi_\lambda : \lambda \in \Lambda\}$

Example: conjugate families we described earlier.

7.4.2. Elicitation –

– means using the statistician's beliefs about θ to specify the prior in $\{\pi_\lambda : \lambda \in \Lambda\}$ that reflects these beliefs.

Ex. 7.4.2. Location Normal

Data – $N(\mu, \sigma_0^2)$; restrict to $N(\mu_0, \tau_0^2) : \mu_0 \in \mathbb{R}^1, \tau_0^2 > 0\}$

the priors for μ. So, $\lambda = (\mu_0, \tau_0^2)$, 2 df.

E.g. we can ask an expert to specify two quantiles of the prior distr. for μ $\Rightarrow$ defines the prior. We may want to specify μ_0 such that μ is as likely to greater than as less than μ_0, so that μ_0 is the median of the prior. We may also want to specify v_0 s.t. there is 99% certainty that the true $\mu < v_0$. –

– that is 0.99 quantile of the prior.

7.4.2 Empirical Bayes.

Here, the main idea is to base the choice of λ_0 on the data s. Note this approach seems to violate the principle of conditional probability. Indeed, in this case the posterior distr. of θ given s is not, in general, the conditional distribution of θ given data

One possibility is to compute the prior predictive $m_\lambda(s)$ for the data s, and then base the choice of λ on these values.

Note that the prior predictive is like a likelihood function for λ so that it is logical to select λ that maximizes $m_\lambda(s)$.

Ex. 7.4.3. Bernoulli model

$(X_1, \to X_n) \sim Ber(\theta)$; $\theta \sim Beta(\lambda, \lambda)$ for some $\lambda > 0$

The prior is symmetric around $\frac{1}{2}$. Check that the prior mean is $\frac{1}{2}$ and the prior variance is $\frac{\lambda^2}{(2\lambda+1)(2\lambda)^2} = \frac{1}{4(2\lambda+1)} \to 0$

as $\lambda \to \infty$. Choosing large λ makes for a very precise prior.

Then, $m_\lambda(x_1, \to x_n) = \frac{\Gamma(2\lambda)}{\Gamma^2(\lambda)} \int_0^1 \theta^{n\bar{x}+\lambda-1} (1-\theta)^{n(1-\bar{x})+\lambda-1} d\theta =$

$= \frac{\Gamma(2\lambda)}{\Gamma^2(\lambda)} \frac{\Gamma(n\bar{x}+\lambda)\Gamma(n(1-\bar{x})+\lambda)}{\Gamma(n+2\lambda)}$. Maximization

w.r.t. to λ has to be numerical.

E.g. $n = 20$, $n\bar{x} = 2.3$ – the number of 1's out of 20.

See Fig. 7.4.1 – max is approx. $\lambda = 2.3$

7.4.4. Hierarchical Bayes.

Here, yet another (hyper) prior is put on λ.

Then, the prior for θ becomes $\pi(\theta) = \int_\Lambda \pi_\lambda(\theta) w(\lambda) d\lambda$

Typically, default choices are made for w.

The posterior density of θ is

$\pi(\theta|s) = \frac{f_\theta(s) \int_\Lambda \pi_\lambda(\theta) w(\lambda) d\lambda}{m(s)} = \int_\Lambda \frac{f_\theta(s)\, \pi_\lambda(\theta)\, m_\lambda(s) w(\lambda)}{m_\lambda(s)\, m(s)} d\lambda$

where $m(s) = \int_\Lambda \int_\Omega f_\theta(s) \pi_\lambda(\theta) w(\lambda) d\theta d\lambda = \int_\Lambda m_\lambda(s) w(\lambda) d\lambda$ and

$m_\lambda(s) = \int f_\theta(s) \pi_\lambda(\theta) d\theta$ [λ is continuous with a prior density given by w]

Note that the posterior density of λ is $m_\lambda(s)\omega(\lambda)$

while $\frac{f_\theta(s)\pi_\lambda(\theta)}{m_\lambda(s)}$ is the posterior density of $m(s)$

θ given λ. Thus, we can use $\pi(\theta|s)$ for inference about θ and $\frac{m_\lambda(s)\omega(\lambda)}{m(s)}$ for inference about λ.

Ex. 7.4.2 Location-scale normal -

Data $x \sim N(\mu, \sigma^2)$ where $\mu \sim N(\mu_0, \tau_0^2)$

$\frac{1}{\sigma^2} \sim \text{Gamma}(\alpha_0, \beta_0)$.

Just need to have one more prior $\omega(\mu_0, \tau_0^2, \alpha_0, \beta_0)$ -

7.4.5 Improper Priors and non Informativity.

Where to stop the chain of priors in hierarchical Bayes? One possibility is to stop at a non informative prior (also, default prior or reference prior) Just how to express that ignorance is often difficult to say. In some cases, the default prior is improper.

$\int \pi(\theta) d\theta = +\infty$. How to interpret this, is unclear.

It also implies that (s, θ) no longer has a joint prob distribution; the use of the principle of conditional probability to justify our inference based on posterior also doesn't work. Need to make sure that at least the posterior is proper.

Ex. 7.4.5 Location - Normal model.

Let $(x_1, \ldots, x_n) \sim N(\mu, \sigma_0^2)$; choice $\pi(\mu) = 1$

$\Rightarrow$ the posterior density of μ is $\propto e^{-\frac{n}{2\sigma_0^2}(x-\mu)^2}$

$\Rightarrow \mu|s \sim N\left(\bar{x}, \frac{\sigma_0^2}{n}\right)$. This is the same as the limiting posterior in Ex. 7.1.2 as $\tau_0 \to \infty$.

One common method of default prior selection is to use Jeffreys' prior $\pi^{J_2}(\theta)$ for $\theta \in R^d$. π is dependent on the model. It is often improper

Ex. 7.4.6. Location normal model ...

Here $a(\theta) = \frac{\sqrt{n}}{\theta_0}$ — the same posterior is obtained as in Ex. 7.4.5. This is because the Jeffreys' prior is a constant w.r.t. θ.