1 Priors

A prior is a sword and Achilles heel of Bayesian statistics. Priors are carriers of prior information that is coherently incorporated via Bayes theorem to the inference. At the same time, parameters are unobservable, and prior specification is subjective in nature. Subjectivity of specifying the prior is fundamental objection of rabid frequentists to the Bayesian approach.

Frequentists, attacking Bayesians for subjectivity are not saints of objectivity themselves. The very elicitation of a model (likelihood) and loss function is highly subjective, and Bayesians merely divide the necessary subjectivity to two sources - that from the model and from the prior. Since the prior and loss are **not separable** in decision theoretic statistical inference (Herman Rubin (1987)) it follows that a decision theoretic frequentist and a Bayesian are equally subjective.

Figure 1: Importance of Priors

Being subjective for an engineer is not a bad thing! Being subjective does not mean being nonscientific, as critics of Bayesian statistic often insinuate. On the contrary – vast amount scientific information coming from theoretical and physical models is guiding specification of priors and merging such information with the data for better inference. Examples are abundant and in this course you will see many instances. In the last several decades Bayesian research also focused on priors that are un-informative and robust as an answer to criticism that Bayesian inference is overly sensitive to the choice of a prior. We will cover several such paradigms.

The nature and sources of priors carry deep philosophical load as well and are subject of discussion and disagreements even among Bayesians.

In this handout we will discuss various priors. Some of them are historic (uniform priors of Laplace and Bayes), some of them mathematically convenient (conjugate priors; Raiffa and Schlaifer, 1961), and some robust and noninformative (Jefreys, improper, reference, etc. priors. We will talk as well about hierarchical priors and priors over various domains (Dirichlet, Wavelets, etc).

1.1 Uniform Priors of Bayes and Laplace

In modern mathematical language, the Bayes' essay dealt with the following problem.

[year 1764] A billiard ball W is rolled on a line $[0, 1]$, with a uniform probability of stopping anywhere. It stops at p. A second ball O is then rolled n times under the same conditions and X times stopped on the left of W. Given X what can we say about p.

In particular, Bayes was interested in $P(a < p < b)$ for $0 \leq a < b \leq 1$. In our notation the solution is given by $\mathcal{B}e(X+1, n-X+1)$ as the posterior of Binomial/Uniform Bayesian model, but at that time the Reverend had difficulties of evaluating $P(a < p < b)$.

Laplace first gave example of abstract Bayesian modeling.

[year 1773] In the box there are N black and white cards. One card is selected and it was white. What is the probability that the proportion of white cards p is equal to p_0.

Figure 2: Pierre-Simon Laplace, born 23 March 1749 in Beaumont-en-Auge, Normandy; Died: 5 March 1827 in Paris

Laplace assumes that all possible proportions $\{2/N, 3/N, \ldots, (N-1)/N\}$ are equally likely. Then he proceed to find the posterior distribution of p using Bayes theorem,

$$P(p = p_0 | \text{one card turned white}) = \frac{p_0 \cdot \frac{1}{N-2}}{\sum_{p=2/N}^{(N-1)/N} p \cdot \frac{1}{N-2}} = \frac{p_0}{N(N-1)/2 - 1}.$$

1.2 Conjugate Priors

When the posterior remains in the same family as the prior, and the effect of likelihood is to "update" the prior parameters but not to change its functional form, we say that such priors are **conjugate** with the likelihood. The conjugacy is popular because of its mathematical ease, once the conjugate pair likelihood/prior is found, the posterior is found easily. In the BC[1] and pre-MCMC erae the conjugate priors have been extensively used (and misused) precisely because of the computational convenience. Nowadays, the general agreement is that simple conjugate analysis is of limited practical value since, given the likelihood, the conjugate prior has limited modeling capability.

There are many univariate and multivariate instances of conjugacy. The following table provides several cases, some of which we already dealt with in the previous handouts.

[1]For some BC era signifies *Before Christ,* rather than *Before Computers.*

Likelihood	Prior	Posterior
$X\|\theta \sim \mathcal{N}(\theta, \sigma^2)$	$\theta \sim \mathcal{N}(\mu, \tau^2)$	$\theta\|X \sim \mathcal{N}(\frac{\tau^2}{\sigma^2+\tau^2}X + \frac{\sigma^2}{\sigma^2+\tau^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$
$X\|\theta \sim \mathcal{B}(n, \theta)$	$\theta \sim \mathcal{B}e(\alpha, \beta)$	$\theta\|X \sim \mathcal{B}e(\alpha + x, n - x + \beta)$
$X_1, \ldots, X_n\|\theta \sim \mathcal{P}(\theta)$	$\theta \sim \mathcal{G}a(\alpha, \beta)$	$\theta\|X_1, \ldots, X_n \sim \mathcal{G}a(\sum_i X_i + \alpha, n + \beta)$.
$X_1, \ldots, X_n\|\theta \sim \mathcal{NB}(m, \theta)$	$\theta \sim \mathcal{B}e(\alpha, \beta)$	$\theta\|X_1, \ldots, X_n \sim \mathcal{B}e(\alpha + mn, \beta + \sum_{i=1}^n x_i)$
$X \sim \mathcal{G}(n/2, 2\theta)$	$\theta \sim \mathcal{IG}(\alpha, \beta)$	$\theta\|X \sim \mathcal{IG}(n/2 + \alpha, (x/2 + \beta^{-1})^{-1})$
$X_1, \ldots, X_n\|\theta \sim \mathcal{U}(0, \theta)$	$\theta \sim \mathcal{P}a(\theta_0, \alpha)$	$\theta\|X_1, \ldots, X_n \sim \mathcal{P}a(\max\{\theta_0, x_1, \ldots, x_n\}\alpha + n)$
$X\|\theta \sim \mathcal{N}(\mu, \theta)$	$\theta \sim \mathcal{IG}(\alpha, \beta)$	$\theta\|X \sim \mathcal{IG}(\alpha + 1/2, \beta + (\mu - X)^2/2)$
$X\|\theta \sim \mathcal{G}a(\nu, \theta)$	$\theta \sim \mathcal{G}a(\alpha, \beta)$	$\theta\|X \sim \mathcal{G}a(\alpha + \nu, \beta + x)$

Exhibiting conjugate priors can be systematic if the likelihood belongs to the exponential (sometimes called Koopmans-Darmois) family. The exponential family is quite broad – and this conjugate methodology is quite developed.

Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ are observations from the exponential family,

$$X_i|\theta \sim f(x_i|\theta) = A(\theta)e^{T^*(x_i)B(\theta)}\Psi(x_i).$$

The likelihood is

$$\ell(\theta) = \prod_{i=1}^n f(x_i|\theta) = [A(\theta)]^n e^{TB(\theta)} H(\mathbf{x}),$$

where $T = \sum_i T^*(x_i)$ and $H(\mathbf{x}) = \prod_{i=1}^n \Psi(x_i)$.

Claim: The natural conjugate prior is proportional to

$$[A(\theta)]^p e^{qB(\theta)}.$$

It is easy to check that for such prior the posterior is proportional to

$$[A(\theta)]^{n+p} e^{(T+q)B(\theta)}.$$

Example: Consider binomial/beta model: $X|\theta \sim \mathcal{B}(n, \theta)$ and $\theta \sim \mathcal{B}e(\alpha, \beta)$.

(i) The likelihood can be thought as the product of n Bernoulli's $Y_1, \ldots, Y_n|\theta \sim \mathcal{B}er(\theta)$, $\prod_{i=1}^n \theta^{Y_i}(1-\theta)^{1-Y_i} = \prod_{i=1}^n (1-\theta)e^{Y_i \log \frac{\theta}{1-\theta}}$, with $\sum_i Y_i = X$.

$$\ell(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{n}{x}\left(\frac{\theta}{1-\theta}\right)^x (1-\theta)^n = (1-\theta)^n \cdot e^{x[\log \frac{\theta}{1-\theta}]} \cdot \binom{n}{x}.$$

Here $A(\theta) = (1-\theta)$, $T^* = y$, $T = \sum_i T^*$, $B(\theta) = \log \frac{\theta}{1-\theta}$, and $H(x) = \binom{n}{x}$. The likelihood can be thought as the product of n Bernoulli's $\prod_{i=1}^n (1-\theta)e^{Y_i \log \frac{\theta}{1-\theta}}$. (ii) The prior is proportional to

$$\theta^{\alpha-1}(1-\theta)^{\beta-1} = (1-\theta)^{\alpha+\beta-2}\left(\frac{\theta}{1-\theta}\right)^{\alpha-1} = (1-\theta)^{\alpha+\beta-2}e^{(\alpha-1)\log \frac{\theta}{1-\theta}}.$$

Here $A(\theta)^p = (1-\theta)^{\alpha+\beta-2}$ and $q = \alpha - 1$.

(ii) The posterior is proportional to

$$(1-\theta)^{n+\alpha+\beta-2}e^{(x+\alpha-1)\log \frac{\theta}{1-\theta}} = \theta^{x+\alpha-1}(1-\theta)^{n+\alpha+\beta-2-x-\alpha+1} = \theta^{x+\alpha-1}(1-\theta)^{n+\beta-x-1}$$

which is proportional to the density of $\mathcal{B}e(\alpha + x, n - x + \beta)$ distribution.

1.3 Location/Scale Family: Improper Priors

If the parameter of interest θ is a location parameter, i.e., if

$$X|\theta \sim f(x-\theta),$$

then invariance principle can justify selection of a prior. If the prior is to be invariant with respect to translation, the functional form of π should be selected so,

$$\pi(\theta) = \pi(\theta - \theta_0), \text{ for all } \theta_0.$$

This is possible only if $\pi(\theta)$ is constant. Thus the invariance prior for the location parameter is $\pi(\theta) = c$.

This prior is improper, that is, $\pi(\theta) = c, \ -\infty < \theta < \infty$, is not a density. We will see that, although the prior is improper, the resulting posterior could be proper (bonafide density). A sufficient condition is that the marginal distribution $m(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$ is finite for any x.

Figure 3: Flat Prior, $\pi(\theta) = 1$.

Prior $\pi(\theta) = c$ is often termed *flat* prior, and selection of constant is irrelevant since in inference, the constants from the prior and prior predictive (marginal) cancel.

Example: Let $X|\theta \sim \mathcal{N}(\theta, \sigma^2)$ be the likelihood. Parameter θ is location parameter and σ^2 is assumed known. If the prior on $\theta = 1$ then the posterior is $\theta|X \sim \mathcal{N}(X, 1)$ and the posterior mean X coincides with the frequentist estimator.

If the parameter of interest θ is a scale parameter, i.e., if

$$X|\theta \sim \frac{1}{\theta} f(\frac{x}{\theta}),$$

then invariance prior on θ should be scale invariant, $\forall c > 0 \ \pi(\theta) = \frac{1}{c}\pi(\frac{\theta}{c})$.

A choice for π that satisfies this invariance requirement is

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

Note that $\pi(\theta) = \frac{1}{\theta}$ is improper since $\int_0^\infty \frac{1}{\theta} d\theta = \infty$.

Example: Let $X_1, \ldots, X_n|\theta \sim \mathcal{N}(\mu, \sigma^2)$. Parameter σ is the scale parameter and μ is assumed known. If the prior on σ is $\pi(\sigma) = 1/\sigma$ then the posterior depends on sufficient statistics $s^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$, $\sigma|X_1, \ldots X_n \propto \sigma^{-(n+1)} e^{-\frac{ns^2}{2\sigma^2}} = (1/\sigma^2)^{(n-1)/2+1} e^{-\frac{ns^2}{2\sigma^2}} [\sim \mathcal{IG}(a, b)] \sim \mathcal{IG}((n-1)/2, ns^2/2)$ and the posterior mean is $b/(a-1) = \frac{1}{n-3}\sum_{i=1}^n (X_i - \mu)^2$ (See handout 0 and second parametrization of inverse gamma). The posterior mode is $b/(a+1) = \frac{1}{n+1}\sum_{i=1}^n (X_i - \mu)^2$.

1.4 Jefreys' Priors

Consider the likelihood $f(x|\theta)$ and its Fisher Information $I(\theta) = -E^{X|\theta}\left(\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}\right)$. The Fisher Information measures sensitivity of an estimator in the neighborhood on MLE since, it is proportional to the expected curvature of the likelihood at the MLE.

Jeffreys suggested

$$\pi(\theta) \propto det(I(\theta))^{1/2},$$

as a prior for the likelihood $f(x|\theta)$. The selection of a square root is based on the invariance principle. Let $\phi = h(\theta)$, and h be an invertible function with inverse function $\theta = g(\phi)$. Then

$$\pi(\phi) = \pi(g(\phi)) \left|\frac{dg(\phi)}{d\phi}\right| = \pi(\theta) \left|\frac{d\theta}{d\phi}\right|. \tag{1}$$

But,

$$I(\phi) = -E^{X|\phi}\left(\frac{\partial^2 \log f(x|\phi)}{\partial \phi^2}\right) = -E^{X|\theta}\left(\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} \cdot \left|\frac{d\theta}{d\phi}\right|^2\right) = I(\theta)\left|\frac{d\theta}{d\phi}\right|^2.$$

Thus,

$$I^{1/2}(\phi) = I^{1/2}(\theta) \cdot \left|\frac{d\theta}{d\phi}\right|,$$

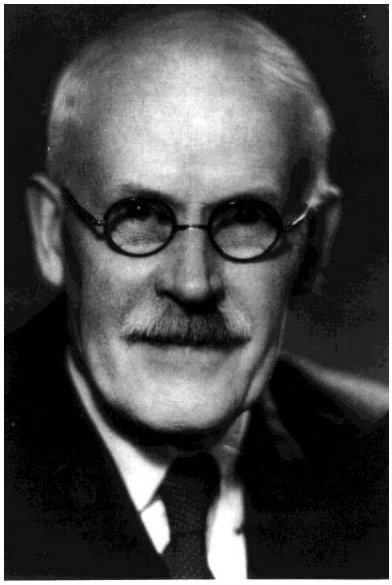
as in (1).

Example: Let $X_1, \ldots, X_n$ have likelihood proportional to $e^{-n(\bar{x}-\theta)^2/(2\sigma^2)}$, for σ^2 known. Since $\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} = -n/\sigma^2$, the Jeffreys prior is $\pi(\theta) \propto (n/\sigma^2)^{1/2} \propto 1$. If, for the same data, the mean μ were known but the variance was parameter of interest, the likelihood would be

$$f(x|\theta) = \theta^{n/2} e^{-s/(2\theta)},$$

where $s = \sum_i (x_i - \mu)^2$ is the sufficient statistics. Then, $\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} = n/(2\theta^2) - s/\theta^3$. Since $E(s|\theta) = n\theta$, the Jeffreys prior is $\pi(\theta) = \frac{1}{\theta}$.

Exercise: Show that for $[X|\theta] \sim \mathcal{P}oi(\theta)$ the Jeffreys prior is $\pi(\theta) = \frac{1}{\sqrt{\theta}}$.



(a)

(b)

Figure 4: (a) Jeffreys, Sir Harold 1891–1989; (b) Jeffreys' Priors

Solution: $I(\theta) = -E^{X|\theta}\frac{\partial^2}{\partial\theta^2}\log fIx|\theta) = -E^{X|\theta}\frac{\partial^2}{\partial\theta^2}(-\theta + x\log\theta - \log(x!)) = -E^{X|\theta}\frac{\partial}{\partial\theta}(1 - x/\theta) = \frac{EX}{\theta^2} = \frac{1}{\theta}$.

Assume that the number of accidents at particular intersection in one month is a Poisson random variable with unknown mean θ. In a particular month $X = 4$ accidents are observed. What is the frequentist estimator for θ. What is the Bayes estimator for θ under Jeffreys' prior.

Exercise: Show that $\pi(\theta) \propto \theta^{1/2}(1-\theta)^{1/2}$ is Jeffreys' prior for $[X|\theta] \sim \mathcal{B}in(n, \theta),\ n$ known. One criticism of Jeffreys' prior is that it violates the likelihood principle. If $n \sim \mathcal{NB}(x, \theta)$, then the Jeffreys prior on θ is $\pi_1(\theta) \propto \theta^{-1}(1-\theta)^{-1/2} \neq \pi(\theta)$.

Exercise: Read on definition of Jeffreys' prior in the case of multiple parameters. Check that if both parameters in $\mathcal{N}(\mu, \sigma^2)$ are of interest, the Jeffreys prior is $\pi(\mu, \sigma^2) = \frac{1}{\sigma^2}$.

Exercise: Consider the Maxwell distribution with density

$$f(x|\theta) = \sqrt{2/\pi}\,\theta^{3/2}x^2\exp\left\{-\frac{\theta x}{2}\right\},\ x \geq 0, \theta > 0.$$

(i) Find Jeffreys prior for θ.
(ii) Find transformation of this parameter for which the corresponding prior is uniform.

Exercise: Let (X_1, X_2, X_3) have trinomial distribution with parameter $(\theta_1, \theta_2, \theta_3),\ \theta_1 + \theta_2 + \theta_3 = 1$,

$$f(x_1, x_2, x_3|\theta_1, \theta_2, \theta_3) = \frac{(x_1 + x_2 + x_3)!}{x_1!x_2!x_3!}\,\theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3}.$$

Derive Jeffreys prior for $(\theta_1, \theta_2, \theta_3)$.

Exercise: The important characterizing property of Jeffreys' priors is their invariance with respect to 1-1 transformations. This implies

$$\sqrt{I(\phi)}d\phi = \sqrt{I(\theta)}d\theta.$$

For $X \sim \mathcal{N}(\theta, 1)$ Jeffreys' prior on θ is $\pi(\theta) \propto 1$. Let $\phi(\theta) = e^{\theta}$. Show that Jeffreys' prior on ϕ is $\pi(\phi(\theta)) = e^{-\theta}$, $\theta \in \mathbb{R}$.

References

[1] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer Verlag.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory,* Division of Research. Boston. Graduate School of Business Administration, Harvard University.

[2] Robert, C. (2001). *Bayesian Choice*, Second Edition, Springer Verlag.

[3] Rubin, H. (1987). A Weak System of Axioms for "Rational" Behavior and the NonSeparability of Utility from Prior. *Statistics and Decisions*, **5**, 47–58.