Unit 2: Prior Distributions

Course Units

- Introduction to Bayesian Statistics
- Prior Distributions
- Simple Models
- Bayesian Asymptotics
- Hierarchical Modeling
- Bayesian Computation
- Model Assessment and Comparison
- Regression Modeling
- Bayesian Nonparametrics
- Survival Analysis and Missing Data
- Clinical Trials and Bayesian Design

Outline of the Unit

General principles

Conjugate Priors

Non-informative Priors

- Jeffreys' Prior

Informative Priors

Considerations in Choice of Prior

1. Mathematical convenience → “conjugate” priors
2. Express lack of information → $\left.\begin{array}{l}\text{non-informative}\\ \text{reference}\end{array}\right\}$ priors
3. Express specific information about parameters → “informative”
   - 3.1 elicitation
   - 3.2 based on previous experiments

Two interpretations of prior distributions:

1. Population interpretation: prior represents a population of possible parameter values, from which the current value has been drawn.
2. State of knowledge: prior expresses our knowledge of the parameter as if the value of the parameter is a random realization from the prior distribution

Remarks:

- In many problems there is no relevant population from which the value could have been drawn
- The prior should include all plausible values because if the prior assigns zero probability to a particular value(s), then the posterior will as well.
- Asymptotically, as we collect more data, the influence of the prior generally goes to zero, provided the number of parameters does not grow with the sample size (and other relevant regularity criteria).
- For random effects and processes, influence of prior is generally important, resulting in shrinkage or smoothing.

Conjugate Prior Distributions

A prior is conjugate for a family of distributions if the prior and the posterior are of the same family.

Example #1:

$$y \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2) \qquad \sigma^2 \text{ known}$$

$$\theta \sim \mathcal{N}(\theta_0, \sigma_0^2)$$

$$\theta|y \sim \mathcal{N}\left(\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\left[\frac{\theta_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}\right], \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$$

$\therefore$ Normal prior for θ is a conjugate prior.

Example #2:

$$y \sim \text{Bin}(n, \theta)$$

$$\theta \sim \text{Beta}(\alpha, \beta)$$

$$\theta|y \sim \text{Beta}(y + \alpha, n - y + \beta)$$

$\therefore$ Beta prior for θ is a conjugate prior.

Summary of conjugate pairs

Family	Conjugate Prior for θ
Bin (n, θ)	$\theta \sim$ Beta (α, β)
$\mathcal{P}(\theta)$	$\theta \sim \mathcal{G}(\alpha, \beta)$
$\mathcal{N}(\theta, \sigma^2)$; σ^2 known	$\theta \sim \mathcal{N}(\theta_0, \sigma_0^2)$
$\mathcal{G}(\alpha, \theta)$; α known	$\theta \sim \mathcal{G}(\delta_0, \gamma_0)$
$\mathcal{N}(\alpha, \theta)$; α known	$\theta^{-1} \sim \mathcal{G}(\delta_0, \gamma_0)$

- The key is that the likelihood, considered in terms of the parameter(s), has a kernel in the same form as the prior distribution.
- Conjugate prior distributions can be interpreted in terms of additional pseudo-data and prior 'sample size'.
- Conjugate priors are useful as building blocks in more complicated models, even though the full model won't be conjugate.
- Exponential families have conjugate priors in general.
- Conjugate priors can be either informative or non-informative.

Non-informative Prior Distributions

a.k.a. vague prior, flat prior, reference prior
A prior distribution is non-informative if the prior is “flat” relative to the likelihood function.
$\Rightarrow$ it has minimal impact on the posterior distribution of θ

Examples:

- If $0 \leq \theta \leq 1$, then $\theta \sim \mathcal{U}(0,1)$ is a non-informative prior for θ.
- If $-\infty < \theta < \infty$, then if $\theta \sim \mathcal{N}(\theta_0, \sigma_0^2)$, and $\sigma_0^2 \rightarrow \infty$, we get a non-informative prior. We can pick σ_0^2 large enough so that the prior has little influence on the posterior.

Locally Uniform Prior:
A prior which does not change very much over the region in which the likelihood is appreciable and does not assume large values outside that range.
If locally uniform:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{\int P(y|\theta)P(\theta)d\theta} \approx \frac{L(\theta|y)}{\int L(\theta|y)} = \text{standardized likelihood}$$

Reference Priors

Reference Prior: a prior which is convenient to use as a standard.

Remarks:

- Choice of prior to characterize where little is known a-priori is a matter of dispute - see Kass and Wasserman (1996): JASA: 91:1343-1370.
- Bayes' Postulate $\Rightarrow$ where such knowledge is lacking concerning the nature of the prior, it might be regarded as uniform.
- Application of Bayes' Postulate to different transformations of $\theta \longrightarrow$ posterior distributions from same data which are inconsistent. $P(\theta)$ uniform, however $P(\log\theta)$ is not uniform.
- Example: prior distributions for variance components
  - $\eta = \log\sigma^2$: $\quad P(\eta) \propto 1 \Rightarrow P(\sigma^2) \propto \frac{1}{\sigma^2}$
  - $P(\sigma^2) \propto 1 \Rightarrow P(\eta) \propto \exp(\eta)$

Location-invariant prior

Location-invariant prior: a prior that is invariant to the choice of the origin (e.g., inference about temperature in Celsius or Kelvin)

- Suppose we observe X, whose density depends only on $x - \theta$, i.e., θ is a location parameter and $X - \theta$ is a pivot
- The origin can be thought to be arbitrary, so we might have measured $Y = X + c$, where the density of Y depends on $y - \eta$, and $\eta = \theta + c$
- If we think the priors for θ and for η should be the same (same structure to the problem, so they should have the same non-informative prior), then

$$P(\theta \in A) = P^*(\eta \in A)$$

- Substituting $\eta = \theta + c$, this implies that $P(\theta \in A) = P(\theta \in A - c)$ (i.e., just shifting the elements in A), a location-invariant prior.
- This is equivalent to $P(\theta) = P(\theta - c)$ for all θ, so for $\theta = c$, we have $P(c) = P(0)$ but c is arbitrary, so we have $P(\theta)$ is a constant, which is improper.

Scale-invariant prior

Scale-invariant prior: a prior that is invariant to the scale of measurement (e.g., inference in meters vs. feet)

- Suppose we observe X, whose density depends only on $\frac{1}{\theta}f(\frac{x}{\theta})$, i.e., θ is a scale parameter
- The scaling can be thought to be arbitrary, so we might have measured $Y = cX$, where the density of Y depends on $\frac{1}{\eta}f(\frac{y}{\eta})$, and $\eta = c\theta$
- Again thinking that the priors for θ and η should be the same:

$$P(\theta \in A) = P^*(\eta \in A)$$

- Substituting $\eta = c\theta$, this implies that $P(\theta \in A) = P(\theta \in c^{-1}A)$ (i.e., just scaling the elements in A), a scale-invariant prior.

$$\int_A P(\theta)d\theta = \int_{c^{-1}A} P(\theta)d\theta = \int_A c^{-1}P(c^{-1}\theta)d\theta$$

- This implies that $P(\theta) = c^{-1}P(c^{-1}\theta)$ for all θ, so for $\theta = c$, we have $P(c) = c^{-1}P(1)$, which indicates that the prior scales as $P(\theta) \propto 1/\theta$, which is improper.

Propriety

Recall: A prior $P(\theta)$ is improper if:

$$\int_{\Theta} P(\theta) d\theta = \infty$$

- In other words, the kernel of the prior, the part that involves θ, must integrate to a finite number, in which case we can compute the appropriate normalizing constant.
- Improper priors can lead to proper posteriors.
- Improper prior may result in an improper posterior $\Rightarrow$ cannot make inference with improper posteriors

Example

$$y_1, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$$

Let $P(\theta) \propto 1$.
What is posterior distribution for θ?

$$\begin{aligned} P(\theta|y) &\propto \exp\left\{-\frac{1}{2}\sum_i (y_i - \theta)^2\right\} \times 1 \\ &= \exp\left\{-\frac{1}{2}\sum_i (y_i - \theta)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[n\theta^2 - 2\theta\Sigma y_i\right]\right\} \\ &\propto \exp\left\{-\frac{n}{2}[\theta - \bar{y}]^2\right\} \end{aligned}$$

$$\therefore\ \theta|y \sim \mathcal{N}\left(\bar{y}, \frac{1}{n}\right)$$

Uniform improper prior ⇒ Proper posterior

Note that this posterior is also the limiting posterior under a sequence of proper prior distributions, $P(\theta) = \mathcal{N}(\theta_0, \sigma_0^2)$ with $\sigma_0^2 \to \infty$ (see Unit 1).

Comments on improper priors

- Improper priors are often used because they generally yield non-informative priors.
- If you use an improper prior you need to be sure that the posterior is proper.
- Can be viewed as limits of proper priors, with the corresponding posterior being the limit of the posteriors corresponding to those priors.
- Can be justified practically in cases in which the data are informative about that parameter (likelihood dominates the prior): prior is unimportant enough that it doesn't require specifying our ignorance exactly.
    - If this is not the case, we need to be more careful and improper priors are a bad choice.

Choosing non-informative priors

Some motivations for non-informative prior choice include:

- Priors that give posteriors corresponding to frequentist point estimates
    - e.g., uniform priors for location parameters
    - $P(\theta, \sigma^2) \propto \frac{1}{\sigma^2}$ for the normal model
- Priors that have an attractive interpretation
    - e.g., priors based on well-known distributions with parameters selected to give non-informativeness
    - $P(\theta) = \mathcal{N}(\theta_0, \sigma_0^2)$
        - let σ_0^2 be large for a non-informative but proper prior
        - or $\sigma_0^2 \to \infty$ for an improper prior, $P(\theta) \propto 1$
- Priors with convenient analytic forms for deriving the posterior
    - e.g., conjugate priors with parameters selected to give non-informativeness

Jeffreys' Rule

$\boxed{\text{Jeffreys' Rule}}$ - a rule for the choice of a non-informative prior [Sir Harold Jeffreys].

Idea Any rule for determining $P(\theta)$ should yield an equivalent result if applied to the transformed parameter; i.e., if we apply the rule to θ to get $P(\theta)$ and then calculate $P(\phi)$ where $\phi = h(\theta)$, we should get the same prior as applying the rule to ϕ to get $P(\phi)$ directly.

Recall Let $P(y|\theta) =$ density of $y|\theta$

$$\underbrace{I(\theta)}_{\text{Fisher Information}} = -E\left[\frac{\partial^2 \log P(y|\theta)}{\partial \theta^2}\right]$$

In the case of θ multivariate:

$$\underbrace{I(\theta)}_{\text{Fisher Information}} = -E\left[\frac{\partial^2 \log P(y|\theta)}{\partial \theta_i \partial \theta_j}\right]_{p \times p}$$

Definition: Jeffreys' Prior

$$P(\theta) \propto |I(\theta)|^{1/2}$$

where $|\cdot|$ is the determinant.

Jeffreys' Prior: Justification

- Jeffreys' prior is invariant to transformation. Consider $\phi = h(\theta)$. To show this, we want to show that

  $$|I(\phi)|^{1/2} = \left|I(h^{-1}(\phi))\right|^{1/2} \left|\frac{\partial h^{-1}(\phi)}{\partial \phi}\right|$$

  If h is one-to-one, we can use the chain rule applied to

  $$I(\phi) = E_{Y|\Phi}\left(\left(\frac{\partial \log P(y|\phi)}{\partial \phi}\right)^2\right)$$

  showing that the prior one gets when applying the rule directly to ϕ is the same as obtained in applying the rule to θ and then transforming.
- Jeffreys' prior is locally uniform and non-informative (see Box and Taio 1973; Section 1.3)

Remarks

- Jeffreys' prior can be improper for *many* models.
- Jeffreys' prior can give strange results for multivariate θ and modifications have been proposed (see Box and Tiao, 1973)
- Jeffreys' prior violates the likelihood principle.
- Jeffreys' prior gives an automated method for finding a non-informative prior for any parametric model, $P(y|\theta)$.

Example 1

Example #1: Suppose we have

$y_1, y_2, \ldots, y_n \sim \text{Binomial}(1, \theta)$

What is Jeffreys' prior?

Solution:

$$P(y|\theta) \propto \theta^y (1-\theta)^{1-y} \quad \begin{array}{l} \theta \in [0,1] \\ y = 0, 1 \end{array}$$

(i.e., assume a single observation)

$$\begin{aligned}
\log P(y|\theta) &= y \log \theta + (1-y)\log(1-\theta) \\
\frac{\partial \log P(y|\theta)}{\partial \theta} &= \frac{y}{\theta} - \frac{(1-y)}{(1-\theta)} \\
\frac{\partial^2 \log P(y|\theta)}{\partial \theta^2} &= -\frac{y}{\theta^2} - \frac{(1-y)}{(1-\theta)^2} \\
I(\theta) &= -E\left[\frac{\partial^2 \log P(y|\theta)}{\partial \theta^2}\right] \\
&= \frac{E(y)}{\theta^2} + \frac{E(1-y)}{(1-\theta)^2} \\
&= \frac{1}{\theta} + \frac{1}{(1-\theta)} = \frac{1}{\theta(1-\theta)}
\end{aligned}$$

Example 1 (cont'd)

$\therefore$ Jeffreys' prior is:

$$\begin{aligned} P(\theta) &\propto I(\theta)^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{\theta(1-\theta)}} \\ &= \theta^{-1/2}(1-\theta)^{-1/2} \end{aligned}$$

Note, we recognize this as a Beta$(1/2, 1/2)$.
$\therefore$ Jeffreys' prior is proper for the Binomial model.

Note This non-informative prior is uniform in $\phi = \sin^{-1}\sqrt{\theta}$ (the usual variance-stabilizing transformation). This is a special case of a more general result that Jeffreys' prior ensures that if there is a transformation of the parameter that behaves like a location parameter with respect to a sufficient statistic, the prior in that parameterization will be uniform. Why? Priors for transformations will be consistent with this prior. Variance stabilization is basically a way to ensure that the transformed random variable behaves like a location parameter.

Example 2

Example 2: Let $y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$

Suppose θ, σ^2 unknown

Derive Jeffreys' prior for θ and σ^2.

Solution: $P(y|\theta, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\dfrac{1}{2\sigma^2}(y-\theta)^2\right\}$

Example (cont'd)

Taking expectations:

....

$$\begin{aligned} \therefore \quad P(\theta, \sigma^2) &\propto |I(\theta, \sigma)|^{1/2} \\ &= \sqrt{\frac{1}{2(\sigma^2)^3}} \\ &\propto (\sigma^2)^{-3/2} \end{aligned}$$

$\therefore$ Jeffreys' joint prior for (θ, σ^2) is improper and is equivalent to $P(\theta, \sigma) \propto \sigma^{-2}$.

What would we have gotten if we had applied Jeffreys' rule separately to θ and σ^2?

Case study

Some non-informative priors for the binomial setting:

1. Jeffreys' prior is Beta$(\frac{1}{2}, \frac{1}{2})$
2. On the scale of natural parameter, logit(θ), a uniform distribution is equivalent to Beta$(0, 0)$.
    - posterior mean is $\hat{\theta} = \frac{y}{n}$

    When does this give an improper posterior?
3. Bayes' principle of uniformity suggests $\mathcal{U}(0, 1) =$ Beta$(1, 1)$ (in terms of the kernel, no prior observations)
    - posterior mode is $\hat{\theta} = \frac{y}{n}$
    - posterior mean is shifted as if we used the MLE but with one additional success and one additional failure

In large samples these will give very similar results.

Summary of Jeffreys' Priors

	Non-Informative	
Family	Prior	Posterior
$\text{Bin}(n, \theta)$	$P(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$	$\theta^{y-\frac{1}{2}}(1-\theta)^{n-y-\frac{1}{2}}$
$\mathcal{P}(\theta)$	$P(\theta) \propto \theta^{-\frac{1}{2}}$	$\theta^{n\bar{y}-1/2} \exp(-n\theta)$
$\mathcal{N}(\theta, \sigma^2)$ σ^2 known	$P(\theta) \propto 1$	$\exp\left\{\frac{-n(\theta-\bar{y})^2}{2\sigma^2}\right\}$
$\mathcal{N}(\theta, \sigma^2)$ θ known	$P(\sigma^2) \propto \frac{1}{\sigma^2}$	$\frac{1}{(\sigma^2)^{n/2+1}} \exp\left\{\frac{-\Sigma(y_i-\theta)^2}{2\sigma^2}\right\}$
$\mathcal{N}(\theta, \sigma^2)$	$P(\theta, \sigma^2) \propto \frac{1}{(\sigma^2)^{3/2}}$	$\frac{1}{(\sigma^2)^{(n+3)/2}} \exp\{-\frac{1}{2\sigma^2}\,(n(\theta-\bar{y})^2$ $+\Sigma(y_i-\bar{y})^2)\}$

Informative Priors

- An informative prior is one that is not dominated by the likelihood, and therefore has an impact on the posterior.
- Example:
  Suppose $y_1, \ldots, y_{10} \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 10)$
  Let $\theta \sim \mathcal{N}(0, 1)$.
  Then the posterior distribution is

  $$\begin{aligned} P(\theta|y) &\propto P(y|\theta)P(\theta) \\ &\propto \exp\left(-\frac{1}{2}\frac{n}{10}(\theta - \bar{y})^2\right) \exp\left(-\frac{1}{2}\theta^2\right) \end{aligned}$$

  Note that prior and likelihood contribute the same amount of information. How do I see this?

Choosing informative priors

- Usually the main challenge is to choose parameter values once a reasonable distributional family is chosen.
- Some obvious choices:
  - $\theta \in \Re^1$: normal or t distribution
  - $\theta \in (0, \infty)$: gamma distribution (e.g., for precision parameters), inverse gamma for variance components, lognormal
  - $\theta \in (0, 1)$: beta distribution
- Once a prior is chosen, one can see if other summaries of the prior are consistent with one's prior beliefs: e.g., quantiles, mode.
- Mixtures can be used if a single component from a familiar form is not sufficient.
- It's hard to specify multivariate prior distributions in part because of the difficulty in coming up with good summary measures and in part because of the limited range of multivariate forms.
- Hierarchical models are one approach to deal with this problem: introducing conditional independences to ease specification.

Elicitation

- Elicitation is the process of gathering expert opinion through specially designed methods of verbal or written communication. This can be accomplished through individual interviews, interactive groups, or Delphi situations (feedback from separated individuals).
- The prior then represents the beliefs of the community as a whole, not one's personal beliefs.
- There are a number of biases in people's probability assessments.
- The trick is generally to elicit opinions about quantities that are interpretable to experts and then translate to parameter values. E.g., this might involve mean and standard deviation or quantiles or predictive distributions.

Elicitation (cont'd)

- The elicitation process itself can be biased. E.g., clinicians may be overly optimistic, or if one samples trial investigators, they may be more optimistic than clinicians in general.
- Some references:
  - Spiegelhalter DJ, Freedman LS, Parmar MKB (1993). Applying Bayesian thinking in drug development and clinical trials. *Statistics in Medicine*, 12, 1501-1511.
  - Kadane and Wolfson (1997). Experiences in Elicitation. The Statistician 46: 1-17.
  - Garthwaite et al. (2005) Statistical Methods for Eliciting Probability Distributions. JASA 100:680.

Example:

$$y \sim \text{Bin}(n, \theta)$$

$$P(\theta) \equiv \text{Beta}(\alpha, \beta)$$

How to specify values of hyperparameters?
Let $\theta \sim \text{Beta}(r\mu_0, r(1-\mu_o))$

$$\left.\begin{aligned} E(\theta) &= \mu_0 \\ V(\theta) &= \tfrac{\mu_0(1-\mu_0)}{r+1} \end{aligned}\right\} \text{elicit via experts}$$

- $r - 2 =$ prior sample size.

Summary

- Conjugate
- Non-informative ⟶ Jeffreys' is one possibility
- Informative

Caution

- If prior and likelihood conflict, the tails of the distributions can be very important, leading to non-robust results.
- Automated rules may not be appropriate.
- Selecting a prior that is always vague can be misguided, particularly if the parameter is not well-identified by the data.
- See the Kass & Wasserman (1996) article for a review.
- There's an ongoing debate about subjective vs. objective Bayes, where objective Bayes is based on reference distributions. See the recent issue of the journal Bayesian Analysis (2006, vol 1, issue 3) for some papers on this.