STAT 425: Introduction to Bayesian Analysis

Marina Vannucci

Rice University, USA

Fall 2017
Lecture 7: Prior Types

- Subjective and objective priors
- Non-informative priors
- Jeffreys’ priors
- Diffuse priors
- Bayes factors for hypothesis testing
Choice of Prior Distribution

What is the interpretation of probability?

- **Objective Probability**
  As normative and objective representations of what is rational to believe about a parameter, usually in a state of ignorance (dice, coin, etc.)

- **Frequentist probability**
  Long run frequency when a process is repeated

- **Subjective Probability**
  Personal judgment about an event. But, one must be “coherent”.

Two general types of Priors:

- **Subjective Priors**
  Priors chosen to reflect expert opinion or personal beliefs

- **Objective Priors**
  Priors chosen to let the data (i.e., likelihood) dominate the posterior distribution, and hence inference. These are generally determined based on the sampling model in use.
Subjective Priors

Ways of Specifying

- For single events:
  Relative probability of A to “not A” (as for lotteries/betting)

- Continuous quantities: prob. distribution
  divide range into intervals, and assign probabilities for each (as for events, above) or specify percentiles.
  Then, fit a smooth probability density.

- Examples: (i) $\theta \sim \text{Beta}(\alpha, \beta)$. From expert: mean=0.3 and s.d.=0.1. Then solve for $\alpha$ and $\beta$ and find Beta(9.2,13.8). (ii) $\theta \sim N(\mu, s)$. From expert: an educated guess of 25 and a 95% confidence that $\theta$ is between 10 and 40. Then, $\theta \sim N(25, 30/4)$.

Subjective priors may be classified into two classes:

- Conjugate Priors
- Non-conjugate priors
Non-informative or “Objective” Priors

Goal: Choose priors which reflect a state of “no-information” about $\theta$. Central problem: specifying a prior distribution for a parameter about which nothing is known.

For example, if $\theta$ can only have a finite set of values, it seems natural to assume all values equally likely a priori. Such specifications fall into the general class of noninformative priors. Roughly speaking a prior is non-informative if it has minimal impact on the posterior, i.e. it does not change much over the region where the likelihood is appreciable.

This notion was first used by Bayes(1763), and Laplace(1812) in Astronomy. A formal development is given in Box and Tiao (1973) and in Kass and Wasserman (1996, JASA). Still the notion is not adequately well-defined.

Two types:

- Reference or “default” priors
- Diffuse priors
Reference Priors

Example: \( x = (x_1, \ldots, x_n) \) from \( N(\theta, \sigma^2) \), \( \theta \) unknown, \( \sigma \) known.

What is the prior that reflects a state of no information about \( \theta \) and lets the likelihood dominate the posterior?

The “flat” prior: \( \pi(\theta) = 1 \) for \(-\infty < \theta < \infty\)

Some features of the flat prior:

- Non-informative: All values of \( \theta \) are equally likely. No information.
- Likelihood dominates posterior: \( p(\theta|x) \propto L(\theta)\pi(\theta) = L(\theta) \)
- May not be a probability distribution, i.e. \( \int \pi(\theta)d\theta = \infty \) (“improper prior”). Not a major concern if posterior is proper.

Ex: \( X \sim \text{Poisson}(\lambda) \), \( 0 < \lambda < \infty \), \( \lambda \sim \text{Ga}(\alpha, \beta) \). Then \( \alpha = 1, \beta \to 0 \) gives \( \pi(\lambda) \sim 1 \) but with a proper posterior \( \lambda|x \sim \text{Ga}(\alpha + t, n) \) (also, example above).

In complex models, however, posterior may also be improper and one cannot make inference.

- Other “default” priors: \( \pi(\lambda) \propto \frac{1}{\lambda^{1/2}} \) [i.e. \( \lambda \sim \text{Ga}(1/2, \beta \to 0) \)] and \( \pi(\lambda) \propto \frac{1}{\lambda} \) [i.e. \( \lambda \sim \text{Ga}(0, \beta \to 0) \)] (both with proper posterior).
Change of variable

Another concern with reference priors: The “constant/flat rule” prior is not transformation invariant.

If $\tau = g(\theta)$ is a monotone function of $\theta$ with inverse $g^{-1}(\tau) = \theta$ then the density of $\tau$ is obtained from the density of $\theta$ as

$$p(\tau) = p(g^{-1}(\tau)) \left| \frac{\partial g^{-1}(\tau)}{\partial \tau} \right|,$$

with Jacobian $J = \frac{\partial g^{-1}(\tau)}{\partial \tau}$

Ex: $\pi(\theta) = 1$ and $\tau = e^\theta$. Implied $\pi(\tau) = 1/\tau$, no-longer a uniform.

Binomial model with $\theta \sim Beta(1, 1)$ uniform for $0 \leq \theta \leq 1$

$\theta \sim Beta(0, 0)$ uniform for log odds, $\tau = \log[\theta/(1 - \theta)]$

$\theta \sim Beta(1, -1)$ uniform for odds, $\tau = \theta/(1 - \theta)$

A density cannot be simultaneously uniform in all transformations. Problem with uniform distribution as “non-informative” (no information on a transformation of $\theta$ does not imply no information on $\theta$).

Remedy: Jeffreys’ priors.
Jeffreys’ Priors
Let \( p(x|\theta) \) be a (single-parameter) probability model. The Fisher information is defined as

\[
I(\theta) = -E_{x|\theta} \left[ \frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} \right]
\]

and the Jeffreys’ prior is \( \pi(\theta) \propto \sqrt{I(\theta)} \). Properties:

- Locally uniform, i.e., flat in the region of the likelihood.
- Transformation invariant, i.e., all parameterizations lead to same prior
  \[
  \pi(\theta) = \sqrt{I(\theta)} \quad \text{and} \quad \pi(\tau) = \sqrt{I(\tau)} \quad \text{with} \quad \tau = g(\theta)
  \]
  same prior obtained whether (i) apply Jeffreys rule to get \( \pi(\theta) \) and then transform to get \( \pi(\tau) \) or (ii) transform to \( \tau \) and then apply Jeffreys rule to get \( \pi(\tau) \).
- Caution: Posterior may not be proper.
- Does not work well for multi-parameter models, \( \theta = (\theta_1, \ldots, \theta_k) \), \( I(\theta) \) is a matrix of partial derivatives and \( \pi(\theta) \propto \sqrt{|I(\theta)|} \).
Examples

Binomial: \( x \sim \text{Binomial}(n, \theta) \),
\[
\log(p(x|\theta)) = x \log(\theta) + (n - x) \log(1 - \theta), \quad E[x|\theta] = n\theta,
\]
\[
\frac{\partial^2 \log(p(x|\theta))}{\partial \theta^2} = -\frac{x}{\theta^2} - \frac{(n - x)}{(1 - \theta)^2}, \quad I(\theta) = \frac{n\theta}{\theta^2} + \frac{(n - n\theta)}{(1 - \theta)^2} = \frac{n}{\theta(1 - \theta)}
\]

Jeffreys’ prior is \( \pi(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2} \), i.e., \( \theta \sim \text{Beta}(1/2, 1/2) \) (conjugate, proper)

Poisson: \( x \sim \text{Poisson}(\lambda) \),
\[
\log(p(x|\lambda)) = \lambda + x \log(\lambda) - \log(x!) \quad \text{(single obs)}, \quad E[x|\lambda] = \lambda
\]
\[
\frac{\partial^2 \log(p(x|\lambda))}{\partial \lambda^2} = -\frac{x}{\lambda^2}, \quad I(\lambda) = \frac{1}{\lambda}
\]

Jeffreys’ prior is \( \pi(\lambda) \propto \lambda^{-1/2} \), i.e., \( \theta \sim \text{Ga}(1/2, \beta \to 0) \) (conjugate, improper with proper posterior)
Diffuse Priors
Motivation is to use a proper prior with “little” info about the parameter.
- Conjugate priors are proper
- The spread of a prior is a measure of the amount of uncertainty about the parameter expressed by the prior.
So, choose a conjugate prior with large standard deviation as a diffuse or weakly informative prior.
Examples: (i) $X \sim \text{Poisson}(\lambda)$, $0 < \lambda < \infty$, $\lambda \sim \text{Ga}(\alpha, \beta)$. A diffuse prior is $\text{Ga}(\alpha, 1)$ with $\alpha$ the prior expectation.

$$(ii) \quad X \sim N(\theta, \sigma^2) \text{ with } \sigma^2 \text{ known. A diffuse prior for } \theta \text{ is } N(\mu_0, 1000).$$

HTTP://WWW.AIACCESS.NET/E_GM.HTM