Bayesian Statistics

Prior Specification

Nan Lin
Department of Mathematics
Washington University in St. Louis

Prior

- Priors are carriers of prior information that is coherently incorporated via Bayes theorem to the inference
- Parameters are unobservable, and prior specification is subjective in nature
- Subjectivity of specifying the prior is fundamental objection of frequentists to the Bayesian approach
  - For frequentists, the elicitation of a model (likelihood) and loss function is highly subjective
  - Bayesians divide the necessary subjectivity to two sources - that from the model and from the prior
- Being subjective does not mean being nonscientific.

Prior specification

- Elicited prior
- Conjugate prior
- Non-informative prior

Elicited Prior

- Usually, intuitive for discrete-valued parameter
- For continuous-valued parameter, discretize into intervals → histogram prior
  - Approximation to the true continuous prior
  - Have to be on a bounded support
- Alternatively, one may decide the form of the prior distribution first, and then identify the value of the hyperparameters
  - Method-of-moments estimation based on (mean, variance) or quantiles
  - Difficulty 1: none of the well known distribution may match the prior knowledge
  - Difficulty 2: Two different distributions can be very similar, e.g. $Cauchy(0,1)$ and $N(0,2.19)$
- Interactive graphical approach

Conjugate prior

- When the posterior remains in the same distributional family as the prior, i.e. 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.

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

Conjugate prior for exponential family

- Suppose $X = (X_1, \ldots, X_n)$ are observations from the exponential family, that is

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

- Likelihood

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

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

- Conjugate prior is proportional to $[A(\theta)]^p e^{qB(\theta)}$.
- Posterior is proportional to $[A(\theta)]^{n+p} e^{(T+q)B(\theta)}$.

Mixture of conjugate priors

- Suppose $\pi_1(\theta)$ and $\pi_2(\theta)$ are two conjugate priors for the models $p(x|\theta)$, and they lead to posterior distributions $p_1(\theta|x)$ and $p_2(\theta|x)$, respectively.
- Suppose a mixture prior is formed as

$$\pi(x|\theta) = \alpha\pi_1(\theta) + (1-\alpha)\pi_2(\theta)$$

- Posterior is also a mixture of $p_1(\theta|x)$ and $p_2(\theta|x)$.

Noninformative prior

- Flat (uniform) prior
  - Not invariant under reparametrization
  - Improper prior: location/scale family

- The Jeffereys prior
  - Invariant under reparametrization

Improper prior for location/scale family

- If the parameter of interest θ is a location parameter, i.e., if $X|\theta \sim p(x-\theta)$, then invariance principle can justify selection of a prior. → $\pi(\theta) = c$
- If the parameter of interest θ is a scale parameter, i.e., if $X|\theta \sim \frac{1}{\theta} f\left(\frac{x}{\theta}\right)$, then invariance principle can justify selection of a prior. → $\pi(\theta) = \frac{1}{\theta}$
- For the location-scale family (normal, t, Cauchy), $f(x|\theta) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, where $\theta = (\mu, \sigma)$, the most common approach is to use the two noninformative priors given above in concert with "prior independence". → $\pi(\mu, \sigma) = \frac{1}{\sigma}, \mu \in R, \sigma > 0.$

Jeffereys prior

- Based on invariance principle, Jefferey proposed

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

- Fisher information
  - Suppose the likelihood is $f(x|\theta)$, then the Fisher information is

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

Example Consider a naive Jeffreys prior for a two-parameter Gaussian: $X \sim N(\mu, \sigma^2)$, and let $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$. We take derivatives to compute the Fisher information matrix:

$$I(\theta) = -\mathbb{E}_\theta \begin{pmatrix} \frac{1}{\sigma^2} & \frac{2(X-\mu)}{\sigma^2} \\ \frac{2(X-\mu)}{\sigma^2} & \frac{3}{\sigma^4}(X-\mu)^2 - \frac{1}{\sigma^2} \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix}$$

since $\mathbb{E}_\theta(X-\mu) = 0$ and $\mathbb{E}_\theta(X-\mu)^2 = \sigma^2$. Therefore

$$\pi_J(\theta) = |I(\theta)|^{1/2} \propto \frac{1}{\sigma^2}.$$

This is different from the earlier suggestion based on "prior independence", and it turns out to have poor convergence properties.

Jeffreys himself proposed using the prior $\pi(\theta) \propto \frac{1}{\sigma}$, which is a product of the separate priors for μ and σ. This prior is better motivated and gives better results as well, and typically called the "reference prior".

Reference Prior

- The idea behind reference priors is to formalize what exactly we mean by an "uninformative prior":
  - it is a function that maximizes some measure of distance or divergence between the posterior and prior, as data observations are made.
  - Any of several possible divergence measures can be chosen, for example the Kullback-Leibler divergence or the Hellinger distance.
  - By maximizing the divergence, we allow the data to have the maximum effect on the posterior estimates.