Chapter 4. Bayesian Analysis

♠ The Posterior Distribution

• Definition and Determination

The posterior distribution of θ given $\boldsymbol{x}$ will be denoted $\pi(\theta|\boldsymbol{x})$, and is defined to the conditional distribution of θ given the sample observation $\boldsymbol{x}$.

Notice that the joint density for θ and $\boldsymbol{X}$ is

$$h(\boldsymbol{x}, \theta) = \pi(\theta) f(\boldsymbol{x}|\theta),$$

and the marginal density for $\boldsymbol{X}$ is

$$m(\boldsymbol{x}) = \int_{\Theta} f(\boldsymbol{x}|\theta) dF^{\pi}(\theta).$$

Thus, the posterior distribution is

$$\pi(\theta|\boldsymbol{x}) = \frac{h(\boldsymbol{x}, \theta)}{m(\boldsymbol{x})},$$

provided that $m(\boldsymbol{x}) \neq 0$.

Lemma 1: *Let T denote a sufficient statistic for $\boldsymbol{\theta}$. Assume $m(t)$ (the marginal distribution of t) is greater than zero, and that the factorization theorem holds. Then, if $T(\boldsymbol{x}) = t$,*

$$\pi(\theta|\boldsymbol{x}) = \pi(\theta|t) = \frac{\pi(\theta)g(t|\theta)}{m(t)}.$$

Proof: Using the factorization theorem, we have

$$f(\boldsymbol{x}|\theta) = h(\boldsymbol{x})g(T(x)|\theta),$$

and

$$m(\boldsymbol{x}) = \int_{\Theta} \pi(\theta)h(\boldsymbol{x})g(T(x)|\theta)d\theta.$$

Hence

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta)h(\boldsymbol{x})g(t|\theta)}{\int_{\Theta} \pi(\theta)h(\boldsymbol{x})g(T(x)|\theta)d\theta} = \frac{\pi(\theta)g(t|\theta)}{m(t)}.$$

Example 1: Assume $X \sim N(\theta, \sigma^2)$, where θ is unknown and σ^2 is known. Let $\pi(\theta)$ be a $N(\mu, \tau^2)$ density. Then

$$\theta | x \sim N(\mu(x), 1/\rho),$$

where

$$\mu(x) = \frac{\sigma^2}{\sigma^2 + \tau^2}\mu + \frac{\tau^2}{\sigma^2 + \tau^2}x = x - \frac{\sigma^2}{\sigma^2 + \tau^2}(x - \mu),$$

and $\rho = \text{precision} = \frac{1}{\sigma^2} + \frac{1}{\tau^2}$.

The derivation of $\pi(\theta|x)$ directly follows from straightforward calculus.

Example 2: Assume a sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ from a $N(\theta, \sigma^2)$ distribution is to be taken (σ^2 known) and $\theta \sim N(\mu, \tau)$. Then, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a sufficient statistic and

$$\overline{X}|\theta \sim N(\theta, \frac{\sigma^2}{n}).$$

Using Lemma 1 and Example 1 with σ^2 replaced by $\frac{\sigma^2}{n}$, we obtain

$$\theta|\bar{x} \sim N(\mu(\bar{x}), 1/\rho),$$

where

$$\mu(\bar{x}) = \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \tau^2}\mu + \frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2}\bar{x} = \bar{x} - \frac{\sigma^2}{\sigma^2 + n\tau^2}(\bar{x} - \mu),$$

and $\rho = \text{precision} = \frac{n}{\sigma^2} + \frac{1}{\tau^2}$.

• **Conjugate Families**

Let $\mathcal{F}$ denote the class of density functions $f(x|\theta)$ (indexed by θ). A class $\mathcal{P}$ of prior distributions is said to be a *conjugate family* for $\mathcal{F}$ if $\pi(\theta|x)$ is in the class $\mathcal{P}$ for all $x \in \mathcal{X}$ and $\pi \in \mathcal{P}$.

Examples of conjugate priors include (i) (Normal mean) $X|\theta \sim N(\theta, \sigma^2)$ and conjugate prior $\theta \sim N(\mu, \tau^2)$; (ii) (Normal variance) $X|\theta \sim N(\mu, \theta)$ and conjugate prior $\theta \sim \mathcal{IG}(\alpha, \beta)$; (iii) (Poisson) $X|\theta \sim \mathcal{P}(\theta)$ and conjugate prior $\theta \sim \mathcal{G}(\alpha, \beta)$; (iv) (Gamma) $X|\theta \sim \mathcal{G}(\nu, \theta)$ and conjugate prior $\theta \sim \mathcal{G}(\alpha, \beta)$; and (v) (Binomial) $X|\theta \sim \mathcal{B}(n, \theta)$ and conjugate prior $\theta \sim \mathcal{B}e(\alpha, \beta)$.

• **Improper Prior**

The analysis leading to the posterior distribution can formally be carried out even if $\pi(\theta)$ is an improper prior.

For example, suppose $\pi(\theta) = 1$. Then we can take

$$\pi_n(\theta) = \frac{1}{2n} I_{(-n,n)}(\theta)$$

and then the posterior can be defined as

$$\pi(\theta|\boldsymbol{x}) = \lim_{n \longrightarrow \infty} \frac{f(x|\theta)\pi_n(\theta)}{\int_{-\infty}^{\infty} f(x|\theta)\pi_n(\theta)d\theta}.$$

♠ Bayesian Inference

• Estimation

The common point estimates of θ include

- **The Largest Posterior Mode**, $\hat{\theta}$, which is also termed as the *generalized maximum likelihood estimate*:

$$\pi(\hat{\theta}|\boldsymbol{x}) = \sup_{\theta \in \Theta} \pi(\theta|\boldsymbol{x}).$$

- **Posterior Mean**:

$$E^{\pi(\theta|\boldsymbol{x})}[\theta] = \int_{\Theta} \theta \pi(\theta|\boldsymbol{x}) d\theta.$$

- **Posterior Median**, $\tilde{\theta}$:

$$P^{\pi(\theta|\boldsymbol{x})}(\theta \geq \tilde{\theta}) \geq \frac{1}{2}$$

and

$$P^{\pi(\theta|\boldsymbol{x})}(\theta \leq \tilde{\theta}) \geq \frac{1}{2}.$$

Example 3: Assume

$$f(x|\theta) = e^{-(x-\theta)} I_{[\theta,\infty)}(x),$$

and

$$\pi(\theta) = \frac{1}{\pi(1+\theta^2)}.$$

Then,

$$\pi(\theta|x) = \frac{e^{-(x-\theta)} I_{[\theta,\infty)}(x)}{m(x)\pi(1+\theta^2)}.$$

To find $\hat{\theta}$ maximizing $\pi(\theta|x)$, note first that only $\theta \leq x$ need be considered. For such θ,

$$\frac{d}{d\theta} \log \pi(\theta|x) = 1 - \frac{2\theta}{1+\theta^2} = \frac{(\theta-1)^2}{1+\theta^2} > 0.$$

Thus, $\log \pi(\theta|x)$ (so as $\pi(\theta|x)$) is increasing for $\theta \leq x$. It follows that $\pi(\theta|x)$ is maximized at $\hat{\theta} = x$, which is thus the generalized maximum likelihood estimate of θ.

Note that the closed forms of the posterior mean and median are not available for this case.

Example 4: A not uncommon situation is to observe $X \sim N(\theta, \sigma^2)$ (for simplicity assume σ^2 is known), where θ is a measure of some clearly positive quantity. The classical estimate of θ is x, which is clearly unsuitable when x turns out to be negative. A reasonable way of developing an alternative estimate (assuming no specific prior knowledge is available) is to use the noninformative prior $\pi(\theta) = I_{(0,\infty)}(\theta)$ (since θ is a location parameter). The resulting posterior is

$$\pi(\theta|x) = \frac{\exp\{-(\theta - x)^2/(2\sigma^2)\} I_{(0,\infty)}(\theta)}{\int_0^\infty \exp\{-(\theta - x)^2/(2\sigma^2)\} d\theta}.$$

The posterior mean is

$$\begin{aligned} E^{\pi(\theta|x)}[\theta] =& \frac{\int_0^\infty \theta \exp\{-(\theta - x)^2/(2\sigma^2)\} d\theta}{\int_0^\infty \exp\{-(\theta - x)^2/(2\sigma^2)\} d\theta} \\ \overset{\eta = \frac{\theta - x}{\sigma}}{=}& \frac{\int_{-(x/\sigma)}^\infty (\sigma\eta + x) \exp\{-\eta^2/2\} \sigma d\eta}{\int_{-(x/\sigma)}^\infty \exp\{-\eta^2/2\} \sigma d\eta} \\ =& x + \frac{(2\pi)^{-1}\sigma \int_{-(x/\sigma)}^\infty \eta \exp\{-\eta^2/2\} d\eta}{1 - \Phi(-x/\sigma)} \\ =& x + \frac{(2\pi)^{-1}\sigma \exp\{-x^2/(2\sigma^2)\}}{1 - \Phi(-x/\sigma)}, \end{aligned}$$

where Φ is the $N(0,1)$ cdf.

• **Estimation Error**

The Bayesian measure of the precision of an estimate is (in one dimension) the posterior variance of the estimate, which is defined as follows.

> If θ is a real valued parameter with posterior distribution $\pi(\theta|\boldsymbol{x})$, and δ is the estimate of θ, then the *posterior variance of* δ is
>
> $$V_\delta^\pi(\boldsymbol{x}) = E^{\pi(\theta|\boldsymbol{x})}[(\theta - \delta)^2].$$
>
> When δ is the posterior mean
>
> $$\mu^\pi(\boldsymbol{x}) = E^{\pi(\theta|\boldsymbol{x})}[\theta],$$
>
> then
>
> $$V^\pi(\boldsymbol{x}) = V_{\mu^\pi(\boldsymbol{x})}^\pi(\boldsymbol{x})$$
>
> will be called simply the *posterior variance* (and it is indeed the variance of θ for the distribution $\pi(\theta|x)$). The *posterior standard deviation* is $\sqrt{V^\pi(\boldsymbol{x})}$.

• **Multivariate Estimation**

Bayesian estimation of a vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ is also straightforward. The generalized maximum likelihood estimate (the posterior mode) is often a reasonable estimate, although existence and uniqueness difficulties are more likely to be encountered in the multivariate case. The posterior mean

$$\boldsymbol{\mu}^{\pi}(\boldsymbol{x}) = (\mu_1^{\pi}(\boldsymbol{x}), \mu_2^{\pi}(\boldsymbol{x}), \ldots, \mu_p^{\pi}(\boldsymbol{x}))' = E^{\pi(\boldsymbol{\theta}|\boldsymbol{x})}[\boldsymbol{\theta}]$$

is a very attractive Bayesian estimate, and its precision can be described by the *posterior covariance matrix*

$$V^{\pi}(\boldsymbol{x}) = E^{\pi(\boldsymbol{\theta}|\boldsymbol{x})}[(\boldsymbol{\theta} - \boldsymbol{\mu}^{\pi}(\boldsymbol{x}))'(\boldsymbol{\theta} - \boldsymbol{\mu}^{\pi}(\boldsymbol{x}))].$$

For a general estimate $\boldsymbol{\delta}$ of $\boldsymbol{\theta}$, can be shown to be

$$\begin{aligned} V_{\boldsymbol{\delta}}^{\pi}(\boldsymbol{x}) =& E^{\pi(\boldsymbol{\theta}|\boldsymbol{x})}[(\boldsymbol{\theta} - \boldsymbol{\delta})'(\boldsymbol{\theta} - \boldsymbol{\delta})] \\ =& V^{\pi}(\boldsymbol{x}) + (\boldsymbol{\mu}^{\pi}(\boldsymbol{x}) - \boldsymbol{\delta})'(\boldsymbol{\mu}^{\pi}(\boldsymbol{x}) - \boldsymbol{\delta}). \end{aligned}$$

Again, it is clear that the posterior mean "minimizes" $V_{\boldsymbol{\delta}}^{\pi}(\boldsymbol{x})$.

Example 5: Suppose $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \Sigma)$ and $\boldsymbol{\theta} \sim N_p(\boldsymbol{\mu}, A)$. Here $\boldsymbol{\mu}$ is a known p-vector, and Σ and A are know $(p \times p)$ positive definite matrix. It can be shown that $\pi(\boldsymbol{\theta}|x)$ is a $N_p(\boldsymbol{\mu}^\pi(\boldsymbol{x}), V^\pi(\boldsymbol{x}))$ density, where the posterior mean is given by

$$\boldsymbol{\mu}^\pi(\boldsymbol{x}) = \boldsymbol{x} - \Sigma(\Sigma + A)^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$$

and the posterior covariance matrix by

$$V^\pi(\boldsymbol{x}) = (A^{-1} + \Sigma^{-1})^{-1} = \Sigma - \Sigma(A + \Sigma)^{-1}\Sigma.$$

- **Credible Sets**
  - A $100(1-\alpha)\%$ *credible set* for θ is a subset C of Θ such that

$$1-\alpha \leq P(C|\boldsymbol{x}) = \int_C dF^{\pi(\theta|\boldsymbol{x})}(\theta)$$

$$= \begin{cases} \int_C \pi(\theta|\boldsymbol{x})d\theta & \text{(continuous case),} \\ \sum_{\theta \in C} \pi(\theta|\boldsymbol{x}) & \text{(discrete case).} \end{cases}$$

  - The $100(1-\alpha)\%$ Highest Posterior Density (HPD) credible set for θ is the subset C of Θ of the form

$$C = C(k(\alpha)) = \{\theta \in \Theta : \ \pi(\theta|\boldsymbol{x}) \geq k(\alpha)\},$$

    where $k(\alpha)$ is the largest constant such that

$$P(C(k(\alpha))|\boldsymbol{x}) \geq 1-\alpha.$$

Example 1 (continued): In Example 1, we assume $X \sim N(\theta, \sigma^2)$, where θ is unknown and σ^2 is known. Then, we obtain

$$\theta | x \sim N(\mu(x), 1/\rho),$$

where

$$\mu(x) = \frac{\sigma^2}{\sigma^2 + \tau^2}\mu + \frac{\tau^2}{\sigma^2 + \tau^2}x = x - \frac{\sigma^2}{\sigma^2 + \tau^2}(x - \mu),$$

and $\rho = \text{precision} = \frac{1}{\sigma^2} + \frac{1}{\tau^2}$.

Since the normal distribution is symmetric, it is clear that the $100(1-\alpha)\%$ HPD credible set (HPD interval) is given by

$$C(k(\alpha)) = \left(\mu(x) + z(\alpha/2)\rho^{-\frac{1}{2}}, \mu(x) - z(\alpha/2)\rho^{-\frac{1}{2}}\right),$$

where $z(\alpha)$ is the α-fractile of a $N(0,1)$ distribution.

• **Properties of HPD Interval**

A HPD interval has two main properties:

(a) the density for every point inside the interval is greater than that for every point outside the interval; and

(b) for a given probability content $(1-\alpha)$ the interval is of the shortest length.

When $\pi(\theta|\boldsymbol{x})$ is symmetric and unimodal, the Bayesian credible interval $(\theta^{(\alpha/2)}, \theta^{(1-\alpha/2)})$, where $\theta^{(\alpha/2)}$ and $\theta^{(1-\alpha/2)}$ denote the $(\alpha/2)$-fractile and $(1-\alpha/2)$-fractile of $\pi(\theta|\boldsymbol{x})$, is also an HPD interval. However, when $\pi(\theta|\boldsymbol{x})$ is not symmetric, $(\theta^{(\alpha/2)}, \theta^{(1-\alpha/2)})$ is not an HPD interval in general, and in this case, an HPD interval is more desirable, since it displays more desired features of the posterior distribution than a credible interval.

• **Monte Carlo Estimation of HPD Intervals**

Density Estimation Based Method

— *Reference*: Box and Tiao (1992, book), Wei and Tanner (1990, Biometrika), Tanner (1996, book), or Hyndman (1996, American Statistician)

— *Algorithm*

(a) Assume that a random (or dependent) sample $\{\theta_i,\ i = 1, 2, \ldots, n\}$ is available from the posterior distribution $\pi(\theta|\boldsymbol{x})$.

(b) Compute $\xi_i = \pi(\theta_i|\boldsymbol{x})$ for $i = 1, 2, \ldots, n$.

(c) Choose $\hat{k}(\alpha) = \xi_{(j)}$, where $\xi_{(j)}$ is the j^{th} smallest of $\{\xi_i\}$, $j = [\alpha n]$, and $[\alpha n]$ denotes the integer part of αn.

(d) Compute $C(\hat{k}(\alpha)) = \{\theta : \quad \pi(\theta|\boldsymbol{x}) \geq \hat{k}(\alpha)\}$.

— *Disadvantages:* Completely known $\pi(\theta|\boldsymbol{x})$ and difficult to compute $C(\hat{k}(\alpha))$.

A Nonparametric Monte Carlo Approach

— *Reference*: Chen and Shao (1999, JCGS).

– *Algorithm*

Step 1. Obtain a random or dependent sample $\{\theta_i,\ i = 1, 2, \ldots, n\}$ from $\pi(\theta|\boldsymbol{x})$.

Step 2. Sort $\{\theta_i,\ i = 1, 2, \ldots, n\}$ to obtain the ordered values:

$$\theta_{(1)} \leq \theta_{(2)} \leq \cdots \leq \theta_{(n)}.$$

Step 3. Compute $100(1-\alpha)\%$ credible intervals

$$C_j(n) = (\theta_{(j)}, \theta_{(j+[(1-\alpha)n])})$$

for $j = 1, 2, \ldots, n - [(1-\alpha)n]$.

Step 4. The $100(1-\alpha)\%$ HPD interval is the one, denoted by $C_{j^*}(n)$, with **the smallest interval width** among all credible intervals.

Asymptotic Consistency

Theorem 1: *Assume that* $\{\theta_i,\ i = 1, 2, \ldots, n\}$ *be an ergodic Markov chain Monte Carlo or random sample from* $\pi(\theta|\boldsymbol{x})$ *and* $\pi(\theta|\boldsymbol{x})$ *is unimodal. Then,*

$$C_{j^*}(n) \to C(k(\alpha))\ a.s.\ as\ n \to \infty.$$

The proof of this theorem is very technical, and it can be found in Chen and Shao (1999).