557: MATHEMATICAL STATISTICS II
BAYESIAN INFERENCE AND DECISION MAKING

1 Bayesian Inference

1.1 Introduction and Terminology

In Bayesian analysis, θ is treated as a random variable with a **prior** density encapsulating the beliefs about θ before the data are collected. Denoting by $\pi_0(\theta)$ the prior density, by Bayes Theorem,

$$\pi_n(\theta) = \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi_0(\theta)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi_0(\theta)}{\int f_{\mathbf{X}|\theta}(\mathbf{x}|t)\pi_0(t)dt} \tag{1}$$

The denominator in (1), $f_{\mathbf{X}}(\mathbf{x})$ is sometimes termed the **marginal likelihood** of the data $\mathbf{x}$, and does not depend on θ. The term $\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})$ in (1) is the **posterior distribution** of θ given $\mathbf{x}$, which encapsulates the beliefs about θ in light of the data. Note that for the posterior calculation in (1)

$$\pi_n(\theta) \propto f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi_0(\theta)$$

ignoring terms that do not involve θ; $f_{\mathbf{X}}(\mathbf{x})$ for fixed $\mathbf{x}$ acts as a normalizing constant.

1.2 A Variance Lemma

The following inequality proves that the expected posterior variance is smaller than the prior variance, and gives a general version of the result for vector parameters. For two $k \times k$ matrices A and B, write $A \geq B$ if $A - B$ is **non-negative definite**, that is, if $\mathbf{x}^\top(A-B)\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^k$.

Lemma 1.1 *For any two vector random variables* $\mathbf{X}$ $(k_1 \times 1)$ *and* $\mathbf{Y}$ $(k_2 \times 1)$ *having some joint probability structure,*

$$Var_{\mathbf{X}}[\mathbf{X}] \geq \mathbb{E}_{\mathbf{Y}}[Var_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}=\mathbf{y}]]$$

where both sides are $k_1 \times k_1$ *matrices.*

Proof. This result follows straightforwardly from the iterated variance result:

$$\text{Var}_{\mathbf{X}}[\mathbf{X}] \;=\; \mathbb{E}_{\mathbf{Y}}[\text{Var}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}=\mathbf{y}] + \text{Var}_{\mathbf{Y}}[\mathbb{E}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}=\mathbf{y}]]$$

To see this, denote the marginal and conditional expectations of $\mathbf{X}$ by

$$\mu^X = \mathbb{E}_{\mathbf{X}}[\mathbf{X}] \qquad\qquad \mu^{X|y} = \mathbb{E}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}=\mathbf{y}]$$

and then define

$$\mu^{X|Y} = \mathbb{E}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}]$$

as the **random variable** formed by conditioning on $\mathbf{Y} = \mathbf{y}$ as $\mathbf{y}$ varies according to $f_{\mathbf{Y}}$. We have that

$$\begin{aligned}
\text{Var}_{\mathbf{X}}[\mathbf{X}] &= \mathbb{E}_{\mathbf{X}}[(\mathbf{X}-\mu^X)(\mathbf{X}-\mu^X)^\top] = \mathbb{E}_{\mathbf{Y}}[\mathbb{E}_{\mathbf{X}|\mathbf{Y}}[(\mathbf{X}-\mu^X)(\mathbf{X}-\mu^X)^\top]|\mathbf{Y}=\mathbf{y}] \\
&= \mathbb{E}_{\mathbf{Y}}[\mathbb{E}_{\mathbf{X}|\mathbf{Y}}[(\mathbf{X}-\mu^{X|y}+\mu^{X|y}-\mu^X)(\mathbf{X}-\mu^{X|y}+\mu^{X|y}-\mu^X)^\top]|\mathbf{Y}=\mathbf{y}] \\
&= \mathbb{E}_{\mathbf{Y}}[\mathbb{E}_{\mathbf{X}|\mathbf{Y}}[(\mathbf{X}-\mu^{X|y})(\mathbf{X}-\mu^{X|y})^\top]|\mathbf{Y}=\mathbf{y}] + 2\mathbb{E}_{\mathbf{Y}}[\mathbb{E}_{\mathbf{X}|\mathbf{Y}}[(\mathbf{X}-\mu^{X|y})(\mu^{X|y}-\mu^X)^\top]|\mathbf{Y}=\mathbf{y}] \\
&\quad + \mathbb{E}_{\mathbf{Y}}[\mathbb{E}_{\mathbf{X}|\mathbf{Y}}[(\mu^{X|y}-\mu^X)(\mu^{X|y}-\mu^X)^\top]|\mathbf{Y}=\mathbf{y}]
\end{aligned}$$

The second expectation is zero, as in the interior expectation

$$\mathbb{E}_{\mathbf{X}|\mathbf{Y}}[(\mathbf{X}-\mu^{X|y})(\mu^{X|y}-\mu^{X})^{\top}|\mathbf{Y}=\mathbf{y}] = \mathbb{E}_{\mathbf{X}|\mathbf{Y}}[(\mathbf{X}-\mu^{X|y})|\mathbf{Y}=\mathbf{y}](\mu^{X|y}-\mu^{X})^{\top} = 0.$$

Therefore

$$\begin{aligned}
\mathrm{Var}_{\mathbf{X}}[\mathbf{X}] &= \mathbb{E}_{\mathbf{Y}}[\mathrm{Var}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}=\mathbf{y}]] + \mathbb{E}_{\mathbf{Y}}[\mathbb{E}_{\mathbf{X}|\mathbf{Y}}[(\mu^{X|y}-\mu^{X})(\mu^{X|y}-\mu^{X})^{\top}]|\mathbf{Y}=\mathbf{y}] \\
&= \mathbb{E}_{\mathbf{Y}}[\mathrm{Var}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}=\mathbf{y}]] + \mathrm{Var}_{\mathbf{Y}}[\mathbb{E}_{\mathbf{X}|\mathbf{Y}}[\mu^{X|y}|\mathbf{Y}=\mathbf{y}]] \\
&= \mathbb{E}_{\mathbf{Y}}[\mathrm{Var}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}=\mathbf{y}] + \mathrm{Var}_{\mathbf{Y}}[\mu^{X|Y}]
\end{aligned}$$

Thus, using this iterated expectation argument, and denoting by $\mathrm{Var}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}]$ the random variable formed by constructing $\mathrm{Var}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}=\mathbf{y}]$ as $\mathbf{y}$ varies according to $f_{\mathbf{Y}}$, we have

$$\begin{aligned}
\mathrm{Var}_{\mathbf{X}}[\mathbf{X}] &= \mathbb{E}_{\mathbf{Y}}[\mathrm{Var}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}]] + \mathrm{Var}_{\mathbf{Y}}[\mathbb{E}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}]] \\
&\geq \mathbb{E}_{\mathbf{Y}}[\mathrm{Var}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X}|\mathbf{Y}]]
\end{aligned}$$

as the second term is non-negative definite. ■

Corollary : The variance of the posterior distribution, $\mathrm{Var}_{\theta|\mathbf{X}}[\theta|\mathbf{X}=\mathbf{x}]$ satisfies

$$\mathrm{Var}_{\theta}[\theta] \geq \mathbb{E}_{\mathbf{X}}\left[\mathrm{Var}_{\theta|\mathbf{X}}[\theta|\mathbf{X}]\right]$$

that is, the expected posterior variance no greater than the prior variance. ■

1.3 Bayesian Updating

The Bayesian calculation in (1) acts sequentially, that is, for data $\mathbf{x}_1$

$$\pi_{\theta|\mathbf{X}_1}(\theta|\mathbf{x}_1) = \frac{f_{\mathbf{X}|\theta}(\mathbf{x}_1|\theta)\pi_0(\theta)}{f_{\mathbf{X}}(\mathbf{x}_1)} = \frac{f_{\mathbf{X}|\theta}(\mathbf{x}_1|\theta)\pi_0(\theta)}{\displaystyle\int f_{\mathbf{X}|\theta}(\mathbf{x}_1|t)\pi_0(t)\,dt} \tag{2}$$

contains the information about θ in light of the data $\mathbf{x}_1$ and prior assumptions. If new (independent and identically distributed to $\mathbf{x}_1$) data $\mathbf{x}_2$ become available, then the posterior for θ in light of the combined data $(\mathbf{x}_1, \mathbf{x}_2)$ is

$$\pi_{\theta|\mathbf{X}_1,\mathbf{X}_2}(\theta|\mathbf{x}_1,\mathbf{x}_2) = \frac{f_{\mathbf{X}|\theta}(\mathbf{x}_1,,\mathbf{x}_2|\theta)\pi_0(\theta)}{f_{\mathbf{X}}(\mathbf{x}_1,\mathbf{x}_2)} = \frac{f_{\mathbf{X}|\theta}(\mathbf{x}_1,\mathbf{x}_2|\theta)\pi_0(\theta)}{\displaystyle\int f_{\mathbf{X}|\theta}(\mathbf{x}_1,\mathbf{x}_2|t)\pi_0(t)\,dt}.$$

But note also that

$$\pi_{\theta|\mathbf{X}_1,\mathbf{X}_2}(\theta|\mathbf{x}_1,\mathbf{x}_2) = \frac{f_{\mathbf{X}|\theta}(\mathbf{x}_2|\theta)\pi_{\theta|\mathbf{X}_1}(\theta|\mathbf{x}_1)}{f_{\mathbf{X}}(\mathbf{x}_2|\mathbf{x}_1)}$$

where $\pi_{\theta|\mathbf{X}_1}(\theta|\mathbf{x}_1)$ is the posterior for θ from (2), and

$$f_{\mathbf{X}}(\mathbf{x}_2|\mathbf{x}_1) = \frac{f_{\mathbf{X}}(\mathbf{x}_1,\mathbf{x}_2)}{f_{\mathbf{X}}(\mathbf{x}_1)} = \frac{\displaystyle\int f_{\mathbf{X}|\theta}(\mathbf{x}_1|\theta)\pi_0(\theta)\,d\theta}{\displaystyle\int f_{\mathbf{X}|\theta}(\mathbf{x}_1,\mathbf{x}_2|t)\pi_0(t)\,dt} \tag{3}$$

1.4 Sufficiency

Direct from (1), if $\mathbf{T}(\mathbf{X})$ is a sufficient statistic for θ, it follows that

$$\pi_n(\theta) = \frac{f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi_0(\theta)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{g(\mathbf{T}(\mathbf{x}),\theta)h(\mathbf{x})\pi_0(\theta)}{f_{\mathbf{X}}(\mathbf{x})} = \left[\frac{h(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right] g(\mathbf{T}(\mathbf{x}),\theta)\pi_0(\theta)$$

where $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = g(\mathbf{T}(\mathbf{x}),\theta)h(\mathbf{x})$ by the Neyman factorization result. Thus the posterior distribution of θ only depends on the data through $\mathbf{T}(\mathbf{x})$.

Lemma 1.2 If $\mathbf{T}(\mathbf{X})$ is a sufficient statistic for θ (in the classical sense) then

$$\pi_n(\theta|\mathbf{x}) = \pi_n(\theta|\mathbf{T}(\mathbf{x}))$$

for all prior specifications $\pi_0(\theta)$.

Proof. By definition

$$f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = f_{\mathbf{X},\mathbf{T}|\theta}(\mathbf{x},\mathbf{t}|\theta)$$

if $\mathbf{t} = \mathbf{T}(\mathbf{x})$, and zero otherwise. Thus, by sufficiency,

$$f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = f_{\mathbf{X}|\mathbf{T}}(\mathbf{x}|\mathbf{t})f_{\mathbf{T}|\theta}(\mathbf{t}|\theta)$$

and hence

$$\pi_n(\theta) \propto f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta) \quad = \quad f_{\mathbf{X}|\mathbf{T}}(\mathbf{x}|\mathbf{t})f_{\mathbf{T}|\theta}(t|\theta)\pi_0(\theta) \propto f_{\mathbf{T}|\theta}(\mathbf{t}|\theta)\pi_0(\theta) \propto \pi_{\theta|\mathbf{T}}(\theta|\mathbf{t})$$

with the constant of proportionality equal to one, as both sides must integrate to one. ■

Statistic $\mathbf{T}(\mathbf{X})$ is sufficient for θ in the Bayesian sense if

$$\pi_n(\theta|\mathbf{x}) \propto f_{\mathbf{T}|\theta}(\mathbf{t}|\theta)\pi_0(\theta).$$

Lemma 1.3 Statistic $\mathbf{T}(\mathbf{X})$ is sufficient in the Bayesian sense if and only if it is sufficient in the Classical sense.

Proof. For the if, see the previous Lemma. For the only if, suppose

$$\pi_n(\theta) = f_{\mathbf{T}|\theta}(\mathbf{t}|\theta)\pi_0(\theta)h(\mathbf{x}).$$

By (1),

$$\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi_0(\theta)}{f_{\mathbf{X}}(\mathbf{x})} \Longrightarrow \frac{f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{\pi_n(\theta|\mathbf{x})}{\pi(\theta)} = f_{\mathbf{T}|\theta}(\mathbf{t}|\theta)h(\mathbf{x}).$$

Hence

$$f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = f_{\mathbf{T}|\theta}(\mathbf{t}|\theta)h(\mathbf{x})f_{\mathbf{X}}(\mathbf{x}) = g(\mathbf{t},\theta)h^*(\mathbf{x})$$

where $g(\mathbf{t},\theta) = f_{\mathbf{T}|\theta}(\mathbf{t}|\theta)$ and $h^*(\mathbf{x}) = h(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})$. Thus $\mathbf{T}(\mathbf{X})$ is sufficient in the classical sense. ■

2 Construction of Prior Distributions

In the Bayesian formulation, the prior density plays an important role. There are several methods via which the prior can be specified quantitatively; from historical or training data; by subjective assessment, similar to the subjective assessment of probabilities in elementary probability theory; by matching to a desired functional form; or in a **non-informative** or **vague** specification, where the prior probability is supposedly spread 'evenly' across the parameter space.

2.1 Conjugate Priors

For some models, a **conjugate prior** can be chosen; this prior combines with the likelihood in such a way to give an analytically tractable posterior calculation. Consider a class of distributions $\mathcal{F}$ indexed by parameter θ

$$\mathcal{F} = \left\{ f_{X|\theta}(x|\theta) : \theta \in \Theta \right\}$$

A class $\mathcal{P}$ of prior distributions for θ is a conjugate family for $\mathcal{F}$ if the posterior distribution for θ resulting from data $\mathbf{x}$ is an element of $\mathcal{P}$ for all $f_{X|\theta} \in \mathcal{F}$, $\pi_\theta \in \mathcal{P}$ and $\mathbf{x} \in \mathcal{X}$.

Example 2.1 *Suppose that $f_{X|\theta}(x|\theta)$ is an Exponential Family distribution*

$$f_{X|\theta}(x|\theta) = h(x)c(\theta)\exp\left\{\sum_{j=1}^{k} t_j(x)w_j(\theta)\right\}$$

so that for a random sample of size n

$$\mathscr{L}_n(\theta) = h(\mathbf{x})\{c(\theta)\}^n \exp\left\{\sum_{j=1}^{k} T_j(\mathbf{x})w_j(\theta)\right\} \tag{4}$$

where $\mathbf{T}(\mathbf{x}) = (T_1(\mathbf{x}), \ldots, T_k(\mathbf{x}))^\top$ and

$$T_j(\mathbf{x}) = \sum_{i=1}^{n} t_j(x_i).$$

Suppose that

$$\pi_0(\theta) = d(\alpha, \beta)\{c(\theta)\}^\alpha \exp\left\{\sum_{j=1}^{k} \beta_j w_j(\theta)\right\} \tag{5}$$

*where α and $\beta = (\beta_1, \ldots, \beta_k)^\top$ are **hyperparameters**. Combining equations (4) and (5) yields the posterior distribution up to proportionality as*

$$\begin{aligned}
\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x}) &\propto \{c(\theta)\}^{\alpha+n} \exp\left\{\sum_{j=1}^{k}[\beta_j + T_j(\mathbf{x})]w_j(\theta)\right\} \\
&= \{c(\theta)\}^{\alpha^\star} \exp\left\{\sum_{j=1}^{k} \beta_j^\star w_j(\theta)\right\}
\end{aligned}$$

The normalizing constant can be deduced to be $d(\alpha + n, \beta + \mathbf{T}(\mathbf{x}))$, and hence the posterior distribution has the same functional form as the prior, but with parameters updated to

$$\alpha^\star = \alpha + n \qquad \beta^\star = (\beta_1^\star, \ldots, \beta_k^\star)^\top = (\beta_1 + T_1(\mathbf{x}), \ldots, \beta_1 + T_1(\mathbf{x}))^\top.$$

2.2 Ignorance Priors

A non-informative prior expresses **prior ignorance** about the parameter of interest.

- If $\Theta = \{\theta_1, \ldots, \theta_k\}$ (that is, θ is known to take one of a finite number of possible values). Then a non-informative prior places equal probability on each value, that is,

$$\pi_0(\theta) = \frac{1}{k} \qquad \theta \in \Theta.$$

- If Θ is a **bounded region**, then a natural non-informative prior is **constant** on Θ.
- If the parameter space Θ is uncountable and unbounded, however, a non-informative prior specification is more difficult to construct. A naive prior specification would be to set $\pi_0(\theta)$ to be a constant, although this prior does not give a valid probability measure as it does not integrate to 1 over Θ.

A prior distribution $\pi_0(\theta)$ for parameter θ is termed **improper** if it does not integrate to 1.

Even for improper priors, (1) can be used to compute the posterior density, which itself will often not be improper (that is, the posterior is **proper**) However, if $\phi = \mathbf{g}(\theta)$ is a transformation of θ that may of inferential interest, then by elementary transformation results, including the Jacobian of the transform $J(\theta \to \phi)$, it follows that

$$\pi_0(\theta) = c \qquad \Longrightarrow \qquad \pi_\phi(\phi) = c \times J(\theta \to \phi)$$

which may **not** be constant, and hence a **non-uniform** prior on ϕ results. This is perhaps unsatisfactory, and so the following procedure may be preferable.

2.3 Jeffreys' Prior

Consider the prior $\pi_0(\theta)$ for parameter θ in probability model $f_{X|\theta}(x|\theta)$ determined by

$$\pi_0(\theta) \propto \{|\mathcal{I}_\theta(\theta)|\}^{1/2}$$

where $\mathcal{I}_\theta(\theta)$ is the *Fisher Information*,

$$\mathcal{I}_\theta(\theta) = \mathbb{E}_X\left[U(X;\theta)U(X;\theta)^\top;\theta\right] = -\mathbb{E}_X\left[\Psi(X;\theta);\theta\right]$$

and $|\mathcal{I}_\theta(\theta)|$ indicates the absolute value of the determinant of $\mathcal{I}_\theta(\theta)$, and $U(X;\theta)$ is the $k \times 1$ vector *score function* with jth element

$$U_j(X;\theta) = \frac{\partial}{\partial \theta_j} \log f_{X|\theta}(x|\theta) \qquad j = 1, \ldots, k$$

and $\Psi(X;\theta)$ is the $k \times k$ matrix of second partial derivatives with (j,l)th element

$$\frac{\partial^2}{\partial \theta_j \partial \theta_l} \log f_{X|\theta}(x|\theta)$$

Example 2.2 $Binomial(m, \theta)$. We have

$$\begin{aligned}
\log f_{X|\theta}(x|\theta) &= \log \binom{m}{x} + x \log \theta + (m-x)\log(1-\theta) \\
U(x;\theta) &= \frac{x}{\theta} - \frac{(m-x)}{(1-\theta)} \\
\Psi(x;\theta) &= -\frac{x}{\theta^2} - \frac{(m-x)}{(1-\theta)^2}
\end{aligned}$$

so therefore

$$\mathcal{I}_\theta(\theta) = -\mathbb{E}_{X|\theta}\left[-\frac{X}{\theta^2} - \frac{(m-X)}{(1-\theta)^2}\right] = \frac{m\theta}{\theta^2} + \frac{m(1-\theta)}{(1-\theta)^2} = \frac{m}{\theta(1-\theta)}$$

and hence

$$\pi_0(\theta) \propto |\mathcal{I}_\theta(\theta)|^{1/2} = \{\theta(1-\theta)\}^{-1/2}$$

Lemma 2.1 Jeffreys' prior is invariant under 1-1 transformations, that is, if $\phi = \phi(\theta)$, then the prior for ϕ obtained by reparameterization from θ to ϕ in the prior for θ, is precisely Jeffreys' prior for ϕ.

Proof. Let $\phi = \phi(\theta)$ be a 1-1 transformation. Denote by $\ell_\theta(x;\theta)$ and $\ell_\phi(x;\phi)$ the log pdfs in the two parameterizations. Then by the rules of partial differentiation

$$\frac{\partial \ell_\phi}{\partial \phi_j} = \sum_{l=1}^{k} \frac{\partial \ell_\theta}{\partial \theta_l}\frac{\partial \theta_l}{\partial \phi_j} \qquad j = 1, \ldots, k$$

so that

$$U(X;\phi) = \Lambda(\theta,\phi)U(X;\theta)$$

where $\Lambda(\theta,\phi)$ is the $k \times k$ matrix with (j,l)th element

$$\frac{\partial \theta_l}{\partial \phi_j}$$

In fact, $\Lambda(\theta,\phi)$ is just the Jacobian of the transformation from θ to ϕ, $J(\theta \to \phi)$. Hence

$$\mathcal{I}_\phi(\phi) = \Lambda(\theta,\phi)\mathcal{I}_\theta(\theta)\Lambda(\theta,\phi)^\top$$

and so

$$|\mathcal{I}_\phi(\phi)| = |\Lambda(\theta,\phi)\mathcal{I}_\theta(\theta)\Lambda(\theta,\phi)^\top| = |\Lambda(\theta,\phi)|^2|\mathcal{I}_\theta(\theta)|$$

and

$$|\mathcal{I}_\phi(\phi)|^{1/2} = |\Lambda(\theta,\phi)||\mathcal{I}_\theta(\theta)|^{1/2}.$$

Thus

$$\pi_0(\phi) \propto |\mathcal{I}_\phi(\phi)|^{1/2} = |\Lambda(\theta,\phi)||\mathcal{I}_\theta(\theta)|^{1/2} = |\Lambda(\theta,\phi)|\pi_0(\theta)$$

and Jeffreys' prior for ϕ is identical to the one that would be obtained by constructing Jeffreys' prior for θ and reparameterizing to ϕ. ■

Example 2.3 $Binomial(m,\theta)$. Suppose that $\phi = \theta/(1-\theta)$ (so that $\theta = \phi/(1+\phi)$). Then

$$\begin{aligned}
\log f_X(x;\phi) &= \log\binom{m}{x} + x\log\phi - m\log(1+\phi) \\
U(x;\phi) &= \frac{x}{\phi} - \frac{m}{(1+\phi)} \\
\Psi(x;\phi) &= -\frac{x}{\phi^2} + \frac{m}{(1+\phi)^2}
\end{aligned}$$

so therefore

$$\mathcal{I}_\phi(\phi) = -\mathbb{E}_X\left[-\frac{X}{\phi^2} + \frac{m}{(1+\phi)^2};\phi\right] = \frac{m\phi}{(1+\phi)\phi^2} - \frac{m}{(1+\phi)^2} = \frac{m}{\phi(1+\phi)^2}$$

and hence

$$\pi_0(\phi) \propto |\mathcal{I}_\phi(\phi)|^{1/2} = \{\phi(1+\phi)^2\}^{-1/2}.$$

Now, recall that Jeffreys' prior for θ takes the form

$$\pi_0(\theta) \propto \{\theta(1-\theta)\}^{-1/2}$$

The Jacobian of the transformation from θ to ϕ is $(1+\phi)^2$, and thus using the univariate transformation theorem

$$\pi_0(\phi) \propto \{\phi/(1+\phi)^2\}^{-1/2}(1+\phi)^2 = \{\phi(1+\phi)^2\}^{-1/2}$$

matching the result found above.

2.4 Location and Scale Parameters: Invariance priors

Consider the one-dimensional case: θ is a **location parameter** if

$$f_X(x;\theta) = f_X(x-\theta).$$

θ is a **scale parameter** if

$$f_X(x;\theta) = \frac{1}{\theta} f_X\left(\frac{x}{\theta}\right)$$

A 'non-informative' prior can be constructed using invariance principles in the location and scale cases.

- For a location parameter, for a non-informative prior, it is required to have, for set $A \subset \Theta$

  $$\int_A \pi_0(\theta)\, d\theta = \int_{A_c} \pi_0(\theta)\, d\theta$$

  where $A_c = \{\theta : \theta - c \in A\}$ for scalar c. Therefore, for all c, we must have

  $$\int_{A_c} \pi_0(\theta)\, d\theta = \int_A \pi_0(\theta - c)\, d\theta \Longrightarrow \pi_0(\theta) = \pi_0(\theta - c) \Longrightarrow \pi_0(\theta) = \text{constant}.$$

- For a scale parameter, it is required to have, for arbitrary set $A \subset \Theta$

  $$\int_A \pi_0(\theta)\, d\theta = \int_{A_c} \pi_0(\theta)\, d\theta$$

  where now $A_c = \{\theta : c\theta \in A\}$ for scalar c. Therefore, for all c, we must have

  $$\int_{A_c} \pi_0(\theta)\, d\theta = \int_A c\pi_0(c\theta)\, d\theta \Longrightarrow \pi_0(\theta) = c\pi_0(c\theta) \Longrightarrow \pi_0(\theta) \propto \frac{1}{\theta}$$

  This follows by the usual 'scale invariance' definition: a function $f(x)$ is scale invariant if

  $$f(cx) \propto f(x)$$

  and all scale invariant functions are power laws; for some $\alpha > 0$,

  $$f(x) \propto x^{-\alpha}.$$

  Here, the condition $\pi_0(\theta) = c\pi_0(c\theta)$ means that we must have $\alpha = 1$.

3 Bayesian Inference and Optimal Decision Making

The key components of a **decision problem** are as follows;

- a **decision** d is to be made, and the decision is selected from some set $\mathcal{D}$ of alternatives.
- a true **state of nature**, $\upsilon(\theta)$, lying in set Υ, defined by the data generating model, $F_X(x;\theta)$.
- a **loss function**, $L(\mathsf{d},\upsilon)$, for decision d and state υ, which records the loss (or penalty) incurred when the true state of nature is υ and the decision made is d.

We aim to select the decision to **m**inimize the expected loss

3.1 Bayesian Estimation via Decision Theory

In an estimation context, the decision is the **estimate** of the parameter, and the true state of nature is the true value of the parameter, $\upsilon(\theta) \equiv \theta$. If data $x_1,\ldots,x_n$ are available, the optimal decision will intuitively become a function of the data. Suppose now that the decision in light of the data is now in the form of an estimate, denoted $\mathsf{d}(\mathbf{x}) = \widehat{\theta}_n$, say, with associated loss $L(\widehat{\theta}_n,\theta)$

(i) The **frequentist risk** or **loss** associated with decision $\mathsf{d}(\mathbf{X})$ (given by estimator $\widehat{\theta}_n$ is the expected loss associated with $\mathsf{d}(\mathbf{X})$, with the expectation taken over the distribution of $\mathbf{X}$ given θ

$$R_{\mathsf{d}}(\theta) = \mathbb{E}_{\mathbf{X}|\theta}\left[L(\widehat{\theta}_n,\theta)\right] = \int_{\mathcal{X}} L(\widehat{\theta}_n,\theta) f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\, d\mathbf{x}$$

(ii) The **Bayes risk** is the expected risk $R_{\mathsf{d}}(\theta)$ associated with $\mathsf{d}(\mathbf{X})$, with the expectation taken over the prior distribution of θ

$$\begin{aligned} R(\mathsf{d}) = \mathbb{E}_\theta[R_{\mathsf{d}}(\theta)] = \mathbb{E}_\theta\left[\mathbb{E}_{\mathbf{X}|\theta}\left[L(\widehat{\theta}_n,\theta)\right]\right] &= \int_\Theta \left\{\int_{\mathcal{X}} L(\widehat{\theta}_n,\theta) f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\, d\mathbf{x}\right\} \pi_0(\theta)\, d\theta \\ &= \int_\Theta \int_{\mathcal{X}} L(\widehat{\theta}_n,\theta) f_{\mathbf{X}}(\mathbf{x}) \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\mathbf{x}\, d\theta \\ &= \int_{\mathcal{X}} \left\{\int_\Theta L(\widehat{\theta}_n,\theta)\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta\right\} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \end{aligned}$$

where, by equation (1), $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi_0(\theta) = f_{\mathbf{X}}(\mathbf{x})\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})$.

(iii) With prior $\pi_0(\theta)$ and fixed data $\mathbf{x}$ the optimal Bayesian decision, termed the **Bayes rule** is

$$\widehat{\mathsf{d}}_B = \underset{\mathsf{d}\in\mathcal{D}}{\operatorname{argmin}}\, R(\mathsf{d})$$

so that, for the Bayes estimate $\widehat{\theta}_{nB}$

$$\widehat{\theta}_{nB} = \underset{\mathsf{d}\in\mathcal{D}}{\operatorname{argmin}} \int_{\mathcal{X}} \left\{\int_\Theta L(\widehat{\theta}_n,\theta)\pi_n(\theta|\mathbf{x})\, d\theta\right\} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} = \underset{\widehat{\theta}_n\in\Theta}{\operatorname{argmin}} \int_\Theta L(\widehat{\theta}_n,\theta)\pi_n(\theta|\mathbf{x})\, d\theta$$

as only the inner integral depends on the decision and the data. That is, the decision that minimizes the Bayes risk minimizes **posterior expected loss** in making decision d, with expectation taken with respect to the posterior distribution $\pi_n(\theta|\mathbf{x})$.

3.2 Results for Different Loss Functions

(I) Under **squared-error loss**

$$L(\widehat{\theta}_n, \theta) = (\widehat{\theta}_n - \theta)^2$$

the Bayes rule for estimating θ is

$$\widehat{\mathsf{d}}_B(\mathbf{x}) = \widehat{\theta}_{nB}(\mathbf{x}) = \mathbb{E}_{\theta|\mathbf{X}}\left[\theta|\mathbf{X} = \mathbf{x}\right] = \int \theta \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta$$

that is, the **posterior expectation**. The expected posterior loss for any Bayes estimate $\widehat{\theta}_n$ is

$$\int L(\widehat{\theta}_n, \theta)\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = \int (\widehat{\theta}_n - \theta)^2 \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta$$

which needs to be minimized with respect to $\widehat{\theta}_n$. Write $t = \widehat{\theta}_n$: then

$$\frac{d}{dt}\left\{\int (t-\theta)^2\, \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta\right\} = \int \frac{d}{dt}\left\{(t-\theta)^2\right\} \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = \int 2\,(t-\theta)\, \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta$$

and equating this to zero gives

$$\int (t-\theta)\, \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = 0 \qquad \Longrightarrow \qquad t = \int \theta \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = \mathbb{E}_{\theta|\mathbf{X}}\left[\theta|\mathbf{X} = \mathbf{x}\right]$$

and hence the optimal $t = \widehat{\theta}_n$ is the posterior expectation as stated.

(II) Under **absolute error loss**

$$L(\widehat{\theta}_n, \theta) = |\widehat{\theta}_n - \theta|$$

the Bayes estimate for θ is the solution of

$$\int_{-\infty}^{\widehat{\theta}_n} \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = \frac{1}{2}$$

that is, it is the **posterior median**. The expected posterior loss is

$$\int L(\widehat{\theta}_n, \theta)\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = \int |\widehat{\theta}_n - \theta| \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta$$

which needs to be minimized with respect to $\widehat{\theta}_n$. Set $t = \widehat{\theta}_n$. Then

$$\int |t-\theta|\, \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = \int_{-\infty}^{t} (t-\theta)\, \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta + \int_{t}^{\infty} (\theta - t)\, \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \tag{6}$$

Differentiating with respect to t the first term using the product rule yields

$$\begin{aligned}\frac{d}{dt}\left\{\int_{-\infty}^{t} (t-\theta)\, \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta\right\} &= \frac{d}{dt}\left\{t\int_{-\infty}^{t} \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta - \int_{-\infty}^{t} \theta \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta\right\} \\ &= t\pi_{\theta|\mathbf{X}}(t|\mathbf{x}) + \int_{-\infty}^{t} \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta - t\pi_{\theta|\mathbf{X}}(t|\mathbf{x}).\end{aligned}$$

Similarly

$$\frac{d}{dt}\left\{\int_{t}^{\infty} (\theta - t)\, \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta\right\} = -t\pi_{\theta|\mathbf{X}}(t|\mathbf{x}) - \int_{t}^{\infty} \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta + t\pi_{\theta|\mathbf{X}}(t|\mathbf{x})$$

Thus, equating the derivative of equation (6) to zero yields

$$\int_{-\infty}^{t} \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta - \int_{t}^{\infty} \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = 0$$

so that

$$\int_{-\infty}^{t} \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = \int_{t}^{\infty} \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = \frac{1}{2}$$

and hence the optimal $t = \widehat{\theta}_n$ is the posterior median.

(III) Under **zero-one loss**

$$L(\mathsf{d}(\mathbf{x}), \theta) = \begin{cases} 0 & \mathsf{d}(\mathbf{x}) = \theta \\ 1 & \mathsf{d}(\mathbf{x}) \neq \theta \end{cases}$$

the Bayes rule for estimating θ is

$$\widehat{\mathsf{d}}_B(\mathbf{x}) = \widehat{\theta}_{nB}(\mathbf{x}) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})$$

that is, the **posterior mode**. To see this, note that the expected posterior loss is

$$\int L(\widehat{\theta}_n, \theta) \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = \int_{\Theta \setminus \widehat{\theta}_n} \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta$$

which needs to be minimized with respect to the choice of $\widehat{\theta}_n$. Consider the loss function

$$L_\delta(\widehat{\theta}_n, \theta) = \begin{cases} 0 & \widehat{\theta}_n \in (\theta - \delta, \theta + \delta) \\ 1 & \widehat{\theta}_n \notin (\theta - \delta, \theta + \delta) \end{cases}$$

for $\delta \geq 0$. That is, the loss is zero if $|\theta_n - \theta| < \delta$, and one otherwise. The expected loss is therefore

$$\int L_\delta(\widehat{\theta}_n, \theta) \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = \int_{\Theta \setminus (\widehat{\theta}_n - \delta, \widehat{\theta}_n + \delta)} \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = 1 - \Pr[\theta \in (\widehat{\theta}_n - \delta, \widehat{\theta}_n + \delta)|\mathbf{x}].$$

Thus we need to choose $\widehat{\theta}_n$ so that

$$\Pr[\theta \in (\widehat{\theta}_n - \delta, \widehat{\theta}_n + \delta)|\mathbf{x}]$$

is as large as possible, that is, we need to choose $\widehat{\theta}_n$ as the centre of the highest posterior probability region of width 2δ. As $\delta \longrightarrow 0$, this interval shrinks to be the posterior mode, as stated.