557: MATHEMATICAL STATISTICS II
BAYESIAN INFERENCE AND DECISION MAKING

1 Bayesian Inference

1.1 Introduction and Terminology

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

$$\pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x}) = \frac{f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})}{f_{\underset{\sim}{X}}(\underset{\sim}{x})} = \frac{f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})}{\int f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})d\underset{\sim}{\theta}} \tag{1}$$

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

$$\pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x}) \propto f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})$$

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

1.2 A Variance Lemma

The following inequality proves that the posterior variance is larger 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

$$\underset{\sim}{x}^{\mathsf{T}}(A - B)\underset{\sim}{x} \geq 0 \qquad \text{for all } \underset{\sim}{x} \in \mathbb{R}^k$$

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

$$Var_{f_{\underset{\sim}{X}}}[\underset{\sim}{X}] \geq E_{f_{\underset{\sim}{Y}}}[Var_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[\underset{\sim}{X}|\underset{\sim}{Y} = \underset{\sim}{y}]]$$

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

Proof. Denote the marginal and conditional expectations of $\underset{\sim}{X}$ by

$$\underset{\sim}{\mu}^{X} = \mathrm{E}_{f_{\underset{\sim}{X}}}[\underset{\sim}{X}] \qquad \underset{\sim}{\mu}^{X|y} = \mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[\underset{\sim}{X}|\underset{\sim}{Y} = \underset{\sim}{y}]$$

and then define

$$\underset{\sim}{\mu}^{X|Y} = \mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[\underset{\sim}{X}|\underset{\sim}{Y}]$$

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

$$\begin{aligned}
\mathrm{Var}_{f_{\underset{\sim}{X}}}[\underset{\sim}{X}] &= \mathrm{E}_{f_{\underset{\sim}{X}}}[(\underset{\sim}{X} - \underset{\sim}{\mu}^{X})(\underset{\sim}{X} - \underset{\sim}{\mu}^{X})^{\mathsf{T}}] = \mathrm{E}_{f_{\underset{\sim}{Y}}}[\mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[(\underset{\sim}{X} - \underset{\sim}{\mu}^{X})(\underset{\sim}{X} - \underset{\sim}{\mu}^{X})^{\mathsf{T}}]|\underset{\sim}{Y} = \underset{\sim}{y}] \\
&= \mathrm{E}_{f_{\underset{\sim}{Y}}}[\mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[(\underset{\sim}{X} - \underset{\sim}{\mu}^{X|y} + \underset{\sim}{\mu}^{X|y} - \underset{\sim}{\mu}^{X})(\underset{\sim}{X} - \underset{\sim}{\mu}^{X|y} + \underset{\sim}{\mu}^{X|y} - \underset{\sim}{\mu}^{X})^{\mathsf{T}}]|\underset{\sim}{Y} = \underset{\sim}{y}] \\
&= \mathrm{E}_{f_{\underset{\sim}{Y}}}[\mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[(\underset{\sim}{X} - \underset{\sim}{\mu}^{X|y})(\underset{\sim}{X} - \underset{\sim}{\mu}^{X|y})^{\mathsf{T}}]|\underset{\sim}{Y} = \underset{\sim}{y}] + \\
&\quad 2\mathrm{E}_{f_{\underset{\sim}{Y}}}[\mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[(\underset{\sim}{X} - \underset{\sim}{\mu}^{X|y})(\underset{\sim}{\mu}^{X|y} - \underset{\sim}{\mu}^{X})^{\mathsf{T}}]|\underset{\sim}{Y} = \underset{\sim}{y}] \\
&\quad + \mathrm{E}_{f_{\underset{\sim}{Y}}}[\mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[(\underset{\sim}{\mu}^{X|y} - \underset{\sim}{\mu}^{X})(\underset{\sim}{\mu}^{X|y} - \underset{\sim}{\mu}^{X})^{\mathsf{T}}]|\underset{\sim}{Y} = \underset{\sim}{y}]
\end{aligned}$$

The second expectation is zero, as in the interior expectation

$$\mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[(\underset{\sim}{X}-\underset{\sim}{\mu}^{X|y})(\underset{\sim}{\mu}^{X|y}-\underset{\sim}{\mu}^{X})^{\mathsf{T}}|\underset{\sim}{Y}=\underset{\sim}{y}]=\mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[(\underset{\sim}{X}-\underset{\sim}{\mu}^{X|y})|\underset{\sim}{Y}=\underset{\sim}{y}](\underset{\sim}{\mu}^{X|y}-\underset{\sim}{\mu}^{X})^{\mathsf{T}}=0.$$

Therefore

$$\begin{aligned}\mathrm{Var}_{f_{\underset{\sim}{X}}}[\underset{\sim}{X}] &= \mathrm{E}_{f_{\underset{\sim}{Y}}}[\mathrm{Var}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[\underset{\sim}{X}|\underset{\sim}{Y}=\underset{\sim}{y}]]+\mathrm{E}_{f_{\underset{\sim}{Y}}}[\mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[(\underset{\sim}{\mu}^{X|y}-\underset{\sim}{\mu}^{X})(\underset{\sim}{\mu}^{X|y}-\underset{\sim}{\mu}^{X})^{\mathsf{T}}]|\underset{\sim}{Y}=\underset{\sim}{y}]\\ &= \mathrm{E}_{f_{\underset{\sim}{Y}}}[\mathrm{Var}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[\underset{\sim}{X}|\underset{\sim}{Y}=\underset{\sim}{y}]]+\mathrm{Var}_{f_{\underset{\sim}{Y}}}[\mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[\underset{\sim}{\mu}^{X|y}|\underset{\sim}{Y}=\underset{\sim}{y}]]\\ &= \mathrm{E}_{f_{\underset{\sim}{Y}}}[\mathrm{Var}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[\underset{\sim}{X}|\underset{\sim}{Y}=\underset{\sim}{y}]+\mathrm{Var}_{f_{\underset{\sim}{Y}}}[\underset{\sim}{\mu}^{X|Y}]\end{aligned}$$

Thus, using this iterated expectation argument, and denoting by

$$\mathrm{Var}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[\underset{\sim}{X}|\underset{\sim}{Y}]$$

the random variable formed by constructing $\mathrm{Var}_{f_{\underset{\sim}{X}|\underset{\sim}{Y}}}[\underset{\sim}{X}|\underset{\sim}{Y}=\underset{\sim}{y}]$ as $\underset{\sim}{y}$ varies according to $f_{\underset{\sim}{Y}}$, we have

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

as the second term is non-negative definite. ■

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

$$\mathrm{Var}_{\pi_{\underset{\sim}{\theta}}}\left[\underset{\sim}{\theta}\right]\geq \mathrm{E}_{f_{\underset{\sim}{X}}}\left[\mathrm{Var}_{\pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}}\left[\underset{\sim}{\theta}|\underset{\sim}{X}\right]\right]$$

that is, the prior variance is at least as big as expected posterior variance. ■

1.3 Bayesian Updating

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

$$\pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x}_1)=\frac{f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}_1|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})}{f_{\underset{\sim}{X}}(\underset{\sim}{x}_1)}=\frac{f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}_1|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})}{\displaystyle\int f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}_1|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})\,d\underset{\sim}{\theta}} \tag{2}$$

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

$$\pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x}_1,\underset{\sim}{x}_2)=\frac{f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}_{1,},\underset{\sim}{x}_2|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})}{f_{\underset{\sim}{X}}(\underset{\sim}{x}_1,\underset{\sim}{x}_2)}=\frac{f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}_1,\underset{\sim}{x}_2|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})}{\displaystyle\int f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}_1,\underset{\sim}{x}_2|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})\,d\underset{\sim}{\theta}}.$$

But note also that

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

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

$$f_{\underset{\sim}{X}}(\underset{\sim}{x}_2|\underset{\sim}{x}_1)=\frac{f_{\underset{\sim}{X}}(\underset{\sim}{x}_1,\underset{\sim}{x}_2)}{f_{\underset{\sim}{X}}(\underset{\sim}{x}_1)}=\frac{\displaystyle\int f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}_1|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})\,d\underset{\sim}{\theta}}{\displaystyle\int f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}_1,\underset{\sim}{x}_2|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})\,d\underset{\sim}{\theta}} \tag{3}$$

1.4 Sufficiency

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

$$\pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x}) = \frac{f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})}{f_{\underset{\sim}{X}}(\underset{\sim}{x})} = \frac{g(\underset{\sim}{T}(\underset{\sim}{x}),\underset{\sim}{\theta})h(\underset{\sim}{x})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})}{f_{\underset{\sim}{X}}(\underset{\sim}{x})} = \left[\frac{h(\underset{\sim}{x})}{f_{\underset{\sim}{X}}(\underset{\sim}{x})}\right] g(\underset{\sim}{T}(\underset{\sim}{x}),\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})$$

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

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

$$\pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x}) = \pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{T}(\underset{\sim}{x}))$$

for all prior specifications $\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})$.

Proof. By definition

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

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

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

and hence

$$\begin{aligned}
\pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x}) \propto f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta}) &= f_{\underset{\sim}{X}|\underset{\sim}{T}(\underset{\sim}{X})}(\underset{\sim}{x}|\underset{\sim}{t}) f_{\underset{\sim}{T}(\underset{\sim}{X})|\underset{\sim}{\theta}}(t|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta}) \\
&\propto f_{\underset{\sim}{T}(\underset{\sim}{X})|\underset{\sim}{\theta}}(\underset{\sim}{t}|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta}) \\
&\propto \pi_{\underset{\sim}{\theta}|\underset{\sim}{T}(\underset{\sim}{X})}(\underset{\sim}{\theta}|\underset{\sim}{t})
\end{aligned}$$

with the constant of proportionality equal to one, as both sides must integrate to one. ■
Statistic $\underset{\sim}{T}(\underset{\sim}{X})$ is sufficient for $\underset{\sim}{\theta}$ in the Bayesian sense if

$$\pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x}) \propto f_{\underset{\sim}{T}(\underset{\sim}{X})|\underset{\sim}{\theta}}(\underset{\sim}{t}|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})$$

Lemma 1.3 Statistic $\underset{\sim}{T}(\underset{\sim}{X})$ 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

$$\pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x}) = f_{\underset{\sim}{T}(\underset{\sim}{X})|\underset{\sim}{\theta}}(\underset{\sim}{t}|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})h(\underset{\sim}{x}).$$

By (1),

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

Hence

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

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

2 Construction of Prior Distributions

In the Bayesian formulation, the prior density $\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})$ 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 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 $\underset{\sim}{\theta}$

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

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

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

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

so that for a random sample of size n

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

where $\underset{\sim}{T}(\underset{\sim}{x}) = \big(T_1(\underset{\sim}{x}), \ldots, T_k(\underset{\sim}{x})\big)^{\mathsf{T}}$ and

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

Suppose that

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

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

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

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

$$\alpha^\star = \alpha + n \qquad \underset{\sim}{\beta}^\star = (\beta_1^\star, \ldots, \beta_k^\star)^{\mathsf{T}} = (\beta_1 + T_1(\underset{\sim}{x}), \ldots, \beta_1 + T_1(\underset{\sim}{x}))^{\mathsf{T}}.$$

2.2 Ignorance Priors

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

- If $\Theta = \{\underset{\sim}{\theta}_1, \ldots, \underset{\sim}{\theta}_k\}$ (that is, $\underset{\sim}{\theta}$ 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,

$$\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta}) = \frac{1}{k} \qquad \underset{\sim}{\theta} \in \Theta.$$

- 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 $\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})$ 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 $\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})$ for parameter $\underset{\sim}{\theta}$ 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 $\underset{\sim}{\phi} = \underset{\sim}{g}(\underset{\sim}{\theta})$ is a transformation of $\underset{\sim}{\theta}$ that may of inferential interest, then by elementary transformation results, including the Jacobian of the transform $J(\underset{\sim}{\theta} \to \underset{\sim}{\phi})$, it follows that

$$\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta}) = c \qquad \Longrightarrow \qquad \pi_{\underset{\sim}{\phi}}(\underset{\sim}{\phi}) = c \times J(\underset{\sim}{\theta} \to \underset{\sim}{\phi})$$

which may **not** be constant, and hence a **non-uniform** prior on $\underset{\sim}{\phi}$ results. This is perhaps unsatisfactory, and so the following procedure may be preferable.

2.3 Jeffreys' Prior

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

$$\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta}) \propto |I\left(\underset{\sim}{\theta}\right)|^{1/2}$$

where $I\left(\underset{\sim}{\theta}\right)$ is Fisher's Information, and $|I\left(\underset{\sim}{\theta}\right)|$ indicates the absolute value of the determinant of $I\left(\underset{\sim}{\theta}\right)$. Recall that

$$I(\underset{\sim}{\theta}) = \mathrm{E}_{f_{X|\underset{\sim}{\theta}}}\left[\underset{\sim}{S}(X;\underset{\sim}{\theta})\underset{\sim}{S}(X;\underset{\sim}{\theta})^{\mathsf{T}}\right] = -\mathrm{E}_{f_{X|\underset{\sim}{\theta}}}\left[\Psi(X;\underset{\sim}{\theta})\right]$$

and also that $\underset{\sim}{S}(X;\underset{\sim}{\theta})$ is the $k \times 1$ vector score function with jth element

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

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

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

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

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

so therefore

$$I(\theta) = -E_{f_{X|\theta}}\left[-\frac{X}{\theta^2} - \frac{(m-X)}{(1-\theta)^2}\right] = \frac{m\theta}{\theta^2} + \frac{m(1-\theta)}{(1-\theta)^2} = \frac{m}{\theta(1-\theta)}$$

and hence

$$\pi_{\underset{\sim}{\theta}}(\theta) \propto |I(\theta)|^{1/2} = \{\theta(1-\theta)\}^{-1/2}$$

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

Proof. Let $\underset{\sim}{\phi} = \underset{\sim}{\phi}(\underset{\sim}{\theta})$ be a 1-1 transformation. Denote by $\ell_\theta(x|\underset{\sim}{\theta})$ and $\ell_\phi(x|\underset{\sim}{\phi})$ 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

$$\underset{\sim}{S}(X; \underset{\sim}{\phi}) = \Lambda(\underset{\sim}{\theta}, \underset{\sim}{\phi}) \underset{\sim}{S}(X; \underset{\sim}{\theta})$$

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

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

In fact, $\Lambda(\underset{\sim}{\theta}, \underset{\sim}{\phi})$ is just the Jacobian of the transformation from $\underset{\sim}{\theta}$ to $\underset{\sim}{\phi}$, $J(\underset{\sim}{\theta} \to \underset{\sim}{\phi})$. Hence

$$I_\phi(\underset{\sim}{\phi}) = \Lambda(\underset{\sim}{\theta}, \underset{\sim}{\phi}) I_\theta(\underset{\sim}{\theta}) \Lambda(\underset{\sim}{\theta}, \underset{\sim}{\phi})^{\mathsf{T}}$$

and so

$$|I_\phi(\underset{\sim}{\phi})| = |\Lambda(\underset{\sim}{\theta}, \underset{\sim}{\phi}) I_\theta(\underset{\sim}{\theta}) \Lambda(\underset{\sim}{\theta}, \underset{\sim}{\phi})^{\mathsf{T}}| = |\Lambda(\underset{\sim}{\theta}, \underset{\sim}{\phi})|^2 |I_\theta(\underset{\sim}{\theta})|$$

and

$$|I_\phi(\underset{\sim}{\phi})|^{1/2} = |\Lambda(\underset{\sim}{\theta}, \underset{\sim}{\phi})| |I_\theta(\underset{\sim}{\theta})|^{1/2}.$$

Thus

$$\pi_{\underset{\sim}{\phi}}(\underset{\sim}{\phi}) \propto |I_\phi(\underset{\sim}{\phi})|^{1/2} = |\Lambda(\underset{\sim}{\theta}, \underset{\sim}{\phi})| |I_\theta(\underset{\sim}{\theta})|^{1/2} = |\Lambda(\underset{\sim}{\theta}, \underset{\sim}{\phi})| \pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})$$

and Jeffreys' prior for $\underset{\sim}{\phi}$ is identical to the one that would be obtained by constructing Jeffreys' prior for $\underset{\sim}{\theta}$ and reparameterizing to $\underset{\sim}{\phi}$. ■

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

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

so therefore

$$I(\phi) = -E_{f_{X|\phi}}\left[-\frac{X}{\phi^2} + \frac{m}{(1+\phi)^2}\right] = \frac{m\phi}{(1+\phi)\phi^2} - \frac{m}{(1+\phi)^2} = \frac{m}{\phi(1+\phi)^2}$$

and hence

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

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

$$\pi_{\underset{\sim}{\theta}}(\theta) \propto \{\theta(1-\theta)\}^{-1/2}$$

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

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

matching the result found above.

2.4 Location and Scale Parameters

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

$$f_{X|\theta}(x|\theta) = f_X(x-\theta).$$

θ is a **scale parameter** if

$$f_{X|\theta}(x|\theta) = \frac{1}{\theta} f_X\left(\frac{x}{\theta}\right)$$

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

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

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

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

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

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

where now $A_c = \{\theta : c\theta \in A\}$ for scalar c. But, therefore, for all c,

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

3 Bayesian Inference and Decision Making

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

- a **decision** δ is to be made, and the decision is selected from some set $\mathcal{D}$ of alternatives.
- a true **state of nature**, υ, lying in set Υ
- a **loss function**, $\mathcal{L}(\delta, \upsilon)$, for decision δ and state υ, which records the loss (or penalty) incurred when the true state of nature is υ and the decision made is δ.

Broadly, we aim to select the decision to minimize the loss. In the presence of uncertainty, we minimize the **expected loss**.

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. The minimization of expected loss proceeds in slightly different fashion dependent on whether data are observed;

(a) With prior $\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})$ and no data, the **expected loss** (or **Bayes loss**) is defined as

$$\mathrm{E}_{\pi_{\underset{\sim}{\theta}}}\left[\mathcal{L}(\delta, \underset{\sim}{\theta})\right] = \int \mathcal{L}(\delta, \underset{\sim}{\theta}) \pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})\, d\underset{\sim}{\theta}$$

The optimal Bayesian decision is

$$\widehat{\delta}_B = \underset{\delta \in \mathcal{D}}{\operatorname{argmin}}\, \mathrm{E}_{\underset{\sim}{\theta}}\left[\mathcal{L}(\delta, \underset{\sim}{\theta})\right]$$

that is, the decision that minimizes the Bayes loss.

(b) If data are available, the optimal decision will intuitively become a function of the data. Suppose now that the decision in light of the data is denoted $\delta(\underset{\sim}{x})$ (a function from $\mathcal{X}$ to $\mathcal{D}$, and the associated loss is $\mathcal{L}(\delta(\underset{\sim}{x}), \underset{\sim}{\theta})$

The **risk** associated with decision $\delta(\underset{\sim}{X})$ is the expected loss associated with $\delta(\underset{\sim}{X})$, with the expectation taken over the distribution of $\underset{\sim}{X}$ given $\underset{\sim}{\theta}$

$$R_\delta(\underset{\sim}{\theta}) = \mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{\theta}}}\left[\mathcal{L}(\delta(\underset{\sim}{X}), \underset{\sim}{\theta})\right] = \int_\Theta \mathcal{L}(\delta(\underset{\sim}{x}), \underset{\sim}{\theta}) f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}|\underset{\sim}{\theta})\, d\underset{\sim}{x}$$

The **Bayes risk** is the expected risk $R_\delta(\underset{\sim}{\theta})$ associated with $\delta(\underset{\sim}{X})$, with the expectation taken over the prior distribution of $\underset{\sim}{\theta}$

$$\begin{aligned} R(\delta) = \mathrm{E}_{\pi_{\underset{\sim}{\theta}}}[R_\delta(\underset{\sim}{\theta})] = \mathrm{E}_{\pi_{\underset{\sim}{\theta}}}\left[\mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{\theta}}}\left[\mathcal{L}(\delta(\underset{\sim}{X}), \underset{\sim}{\theta})\right]\right] &= \int_\Theta \left\{ \int_{\mathcal{X}} \mathcal{L}(\delta(\underset{\sim}{x}), \underset{\sim}{\theta}) f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}|\underset{\sim}{\theta})\, d\underset{\sim}{x} \right\} \pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})\, d\underset{\sim}{\theta} \\ &= \int_\Theta \int_{\mathcal{X}} \mathcal{L}(\delta(\underset{\sim}{x}), \underset{\sim}{\theta}) f_{\underset{\sim}{X}}(\underset{\sim}{x}) \pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x})\, d\underset{\sim}{x}\, d\underset{\sim}{\theta} \\ &= \int_{\mathcal{X}} \left\{ \int_\Theta \mathcal{L}(\delta(\underset{\sim}{x}), \underset{\sim}{\theta}) \pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x})\, d\underset{\sim}{\theta} \right\} f_{\underset{\sim}{X}}(\underset{\sim}{x})\, d\underset{\sim}{x} \end{aligned}$$

where, by equation (1), $f_{\underset{\sim}{X}|\underset{\sim}{\theta}}(\underset{\sim}{x}|\underset{\sim}{\theta})\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta}) = f_{\underset{\sim}{X}}(\underset{\sim}{x})\pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x})$. With prior $\pi_{\underset{\sim}{\theta}}(\underset{\sim}{\theta})$ and fixed data $\underset{\sim}{x}$ the optimal Bayesian decision, termed the **Bayes rule** is $\widehat{\delta}_B = \underset{\delta \in \mathcal{D}}{\operatorname{argmin}}\, R(\delta)$ so that

$$\widehat{\delta}_B = \underset{\delta \in \mathcal{D}}{\operatorname{argmin}} \int_{\mathcal{X}} \left\{ \int_\Theta \mathcal{L}(\delta(\underset{\sim}{x}), \underset{\sim}{\theta}) \pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x})\, d\underset{\sim}{\theta} \right\} f_{\underset{\sim}{X}}(\underset{\sim}{x})\, d\underset{\sim}{x} = \underset{\delta \in \mathcal{D}}{\operatorname{argmin}} \int_\Theta \mathcal{L}(\delta(\underset{\sim}{x}), \underset{\sim}{\theta}) \pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x})\, d\underset{\sim}{\theta}$$

as only the inner integral depends on δ. That is, the decision that minimizes the Bayes risk minimizes **posterior expected loss** in making decision δ, with expectation taken with respect to the posterior distribution $\pi_{\underset{\sim}{\theta}|\underset{\sim}{X}}(\underset{\sim}{\theta}|\underset{\sim}{x})$.

Results for Different Loss Functions in the One Parameter case

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

$$\mathcal{L}(\delta(\underset{\sim}{x}), \theta) = \big(\delta(\underset{\sim}{x}) - \theta\big)^2$$

the Bayes rule for estimating θ is

$$\widehat{\delta}_B(\underset{\sim}{x}) = \widehat{\theta}_B(\underset{\sim}{x}) = \mathrm{E}_{\pi_{\theta|\underset{\sim}{X}}}\left[\theta|\underset{\sim}{X} = \underset{\sim}{x}\right] = \int \theta \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta$$

that is, the **posterior expectation**. The expected posterior loss is

$$\int \mathcal{L}(\delta(\underset{\sim}{x}), \theta)\pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta = \int \big(\delta(\underset{\sim}{x}) - \theta\big)^2 \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta$$

which needs to be minimized with respect to $\delta(\underset{\sim}{x})$. Write $\delta = \delta(\underset{\sim}{x})$. Then

$$\frac{d}{d\delta}\left\{\int (\delta - \theta)^2 \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta\right\} = \int \frac{d}{d\delta}\left\{(\delta - \theta)^2\right\} \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta = \int 2\,(\delta - \theta)\, \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta$$

and equating this to zero gives

$$\int (\delta - \theta)\, \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta = 0 \qquad \Longrightarrow \qquad \delta = \int \theta \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta = \mathrm{E}_{\pi_{\theta|\underset{\sim}{X}}}\left[\theta|\underset{\sim}{X} = \underset{\sim}{x}\right]$$

and hence the optimal δ is the posterior expectation as stated.

(II) Under **absolute error loss**

$$\mathcal{L}(\delta(\underset{\sim}{x}), \theta) = \big|\delta(\underset{\sim}{x}) - \theta\big|$$

the Bayes rule for estimating θ is the solution of

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

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

$$\int \mathcal{L}(\delta(\underset{\sim}{x}), \theta)\pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta = \int \big|\delta(\underset{\sim}{x}) - \theta\big| \, \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta$$

which needs to be minimized with respect to $\delta(\underset{\sim}{x})$. Set $\delta = \delta(\underset{\sim}{x})$. Then

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

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

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

Similarly

$$\frac{d}{d\delta}\left\{\int_{\delta}^{\infty} (\theta - \delta)\, \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta\right\} = -\delta \pi_{\theta|\underset{\sim}{X}}(\delta|\underset{\sim}{x}) - \int_{\delta}^{\infty} \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta + \delta \pi_{\theta|\underset{\sim}{X}}(\delta|\underset{\sim}{x})$$

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

$$\int_{-\infty}^{\delta} \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta - \int_{\delta}^{\infty} \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta = 0$$

so that

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

and hence the optimal δ is the posterior median.

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

$$\mathcal{L}(\delta(\underset{\sim}{x}), \theta) = \begin{cases} 0 & \delta(\underset{\sim}{x}) = \theta \\ 1 & \delta(\underset{\sim}{x}) \neq \theta \end{cases}$$

the Bayes rule for estimating θ is

$$\widehat{\delta}_B(\underset{\sim}{x}) = \widehat{\theta}_B(\underset{\sim}{x}) = \underset{\theta \in \Theta}{\text{argmax}}\, \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})$$

that is, the **posterior mode**. The expected posterior loss is

$$\int \mathcal{L}(\delta(\underset{\sim}{x}), \theta) \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta = \int_{\Theta \setminus \delta(\underset{\sim}{x})} \pi_{\theta|\underset{\sim}{X}}(\theta|\underset{\sim}{x})\, d\theta$$

which needs to be minimized with respect to $\delta(\underset{\sim}{x})$. The optimal $\delta(\underset{\sim}{x})$ is thus the posterior mode as stated