1.5 Bayesian point estimation

The problem consists in producing a point summary of the posterior distribution. Possible choices: posterior mode, mean and median

- posterior mode

$$\begin{aligned}\hat{\theta} &= \underset{\theta \in \Theta}{\operatorname{argmax}}\ \pi(\theta \mid \boldsymbol{x}) \\ &= \underset{\theta \in \Theta}{\operatorname{argmax}}\ f(\boldsymbol{x} \mid \theta)\ \pi(\theta)\end{aligned}$$

Remark:

1. no need to know $m(\boldsymbol{x})$ to compute $\hat{\theta}$
2. If $\pi(\theta)$ is (approximately) constant, $\hat{\theta}$ coincides (approximately) with the MLE of θ
3. Therefore, the MLE can be perceived as a Bayesian estimate, but the interpretation is quite different
4. If $\hat{\theta}$ is the posterior mode of θ and $\psi = g(\theta)$, then the posterior mode of ψ is

not $g(\hat{\theta})$, since (with $h = g^{-1}$ one-to-one)

$$\pi^{\star}(\psi \mid \boldsymbol{x}) = |h'(\psi)| \; \pi(h(\psi) \mid \boldsymbol{x})$$

2. <u>posterior mean</u>:

$$\hat{\theta} = E[\theta \mid \boldsymbol{x}] = \int_{\Theta} \theta \; \pi(\theta \mid \boldsymbol{x}) \; d\theta$$

3. <u>posterior median</u>:

$$\hat{\theta} : P(\theta \geq \hat{\theta} \mid \boldsymbol{x}) \geq 1/2 \text{ e } P(\theta \leq \hat{\theta} \mid \boldsymbol{x}) \geq 1/2$$

which in the continuous case means

$$\hat{\theta} : P(\theta \leq \hat{\theta} \mid \boldsymbol{x}) = 1/2$$

In a particular situation, how do we choose between these estimates and even other summaries of $\pi(\theta \mid \boldsymbol{x})$?

- Without additional details, the choice may boil down to how they are to compute
- a formal justification requires the ingredients of statistical decision theory:
- Loss function: $L(a, \hat{\theta})$ denotes the loss on incurs when estimating θ by $\hat{\theta}$ and the true value of θ is a
- Frequent choices: $L(a, \hat{\theta}) = (\hat{\theta} - a)^2$; $L(a, \hat{\theta}) = |\hat{\theta} - a|$
- Frequently used criterion: pick the estimate that minimizes the posterior risk:

$$r(\hat{\theta}) = E[L(\theta, \hat{\theta}) \mid \boldsymbol{x}] = \int_{\Theta} L(\theta, \hat{\theta})\ \pi(\theta \mid \boldsymbol{x})\ d\theta$$

  resulting in the so-called Bayes estimate, $\hat{\theta}^B$

- If $L(a, \hat{\theta}) = (\hat{\theta} - a)^2$, then $\hat{\theta}^B$ is the posterior mean; if $L(a, \hat{\theta}) = |\hat{\theta} - a|$, then $\hat{\theta}^B$ is the posterior median

Example 1.1 *Let*$X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} B(1,\theta)$*; a priori* $\theta \sim Be(a,b)$, $a, b > 0$ *known.*

We saw that $\theta \mid \boldsymbol{x} \sim \mathrm{Be}\,(t+a, n-t+b)$ where $t = \sum x_i$.

Hence, the posterior mean of θ is

$$\begin{aligned}\hat{\theta} &= \frac{t+a}{t+a+n-t+b} = \frac{t+a}{a+b+n} \\ &= \frac{a+b}{a+b+n}\;\frac{a}{a+b} + \left(1 - \frac{a+b}{a+b+n}\right)\;\frac{t}{n}\end{aligned}$$

which corresponds to the weighted average of the prior mean of θ (given by $a/(a+b)$) and the sample mean (given by t/n).

Remarks:

- As $n \to +\infty$ with t/n fixed, $\hat{\theta} \to t/n$, the MLE of θ
- When $t = 0$ or $t = n$, the MLE of θ are respectively 0 and 1; that does not happen with the posterior mean: $a/(a+b+n)$ e $(a+n)/(a+b+n)$, respectively

Example 1.2 *Let* $X_1, \ldots, X_n \mid \mu \stackrel{iid}{\sim} N(\mu, 1)$.

We saw that the Jeffreys prior in this case is $\pi^J(\mu) \propto 1$ and leads to $\mu \mid \boldsymbol{x} \sim \mathrm{N}(\bar{x}, 1/n)$. Hence, the posterior mean and median of μ coincide with the MLE. Notice that

$$E[\mu \mid \boldsymbol{x}] = \bar{x}$$

$$E[\bar{X} \mid \mu] = \mu$$

The conjugate family in this case is normal $\mu \sim \mathrm{N}(m_0, v_0^2)$ and

$$\mu \mid \boldsymbol{x} \sim \mathrm{N}(m_n, v_n^2)$$

where $v_n^2 = (n + 1/v_0^2)^{-1}$ and

$$m_n = \frac{n}{n + 1/v_0^2}\,\bar{x} + \frac{1/v_0^2}{n + 1/v_0^2}\,m_0$$

Note that the posterior mean is the weighted average of the sample mean and of the prior mean of μ with the weights proportional to the associated precisions (ie, inverse of the variances). When $v_0^2 \to +\infty$, $E[\theta \mid \boldsymbol{x}] \to \bar{x}$, that is, the sampling information dominates.

1.6 Bayesian prediction

The goal here is to predict a random quantity Y whose distribution involves θ using $x_1, \ldots, x_n$, the observed value of random sample from $f(x \mid \theta)$

How? Obtaining the distribution of $Y \mid x_1, \ldots, x_n$:

$$\begin{aligned} f(y \mid \boldsymbol{x}) &= \int_\Theta f(y, \theta \mid \boldsymbol{x}) \; d\theta \\ &= \int_\Theta f(y \mid \boldsymbol{x}, \theta) \; \pi(\theta \mid \boldsymbol{x}) \; d\theta \end{aligned}$$

known as the (posterior) predictive distribution of Y.

In most cases, Y is statistically independent of $X_1, \ldots, X_n$ dado θ, resulting in

$$f(y \mid \boldsymbol{x}) = \int_\Theta f(y \mid \theta) \; \pi(\theta \mid \boldsymbol{x}) \; d\theta$$

Remarks

- Note how elegant and general this solution is
- The frequentist solution is often to use $f(y \mid \hat{\theta})$, where $\hat{\theta}$ an estimate ofθ: we proceed as if the estimate was the true value of the parameter
- the Bayesian solution incorporates the uncertainty associated with the true value of θ

Example 1.3 *Let* $X_1, \ldots, X_n \mid \theta \stackrel{iid}{\sim} B(1, \theta)$*; a priori* $\theta \sim Be(a, b)$, $a, b > 0$ *are known.*

We saw that $\theta \mid \boldsymbol{x} \sim \mathrm{Be}\,(t + a, n - t + b)$ where $t = \sum x_i$.

We want to predict the $(n+1)$-th Bernoulli trial, which is statistically independent of $X_1, \ldots, X_n$, and which we denote by Y

$$
\begin{aligned}
f(y \mid \boldsymbol{x}) &= \int_0^1 f(y \mid \theta)\ \pi(\theta \mid \boldsymbol{x})\ d\theta \\
&= \int_0^1 \theta^y (1-\theta)^{1-y}\ \frac{1}{B(t+a, n-t+b)}\ \theta^{t+a-1}\ (1-\theta)^{n-t+b-1}\ d\theta \\
&= \frac{B(t+a+y, n-t+b+1-y)}{B(t+a, n-t+b)}\ , \quad y = 0, 1
\end{aligned}
$$

(Beta-Bernoulli distribution)

Cont:

Using $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ e $\Gamma(x+1) = x\ \Gamma(x)$, we obtain

$$P(Y = 1 \mid \boldsymbol{x}) = f(1 \mid \boldsymbol{x}) = \frac{t+a}{n+a+b}$$

Simpler solution:

$$\begin{aligned} P(Y = 1 \mid \boldsymbol{x}) &= E[I_{\{1\}}(Y) \mid \boldsymbol{x}] \\ &= E[\ E[I_{\{1\}}(Y) \mid \theta, \boldsymbol{x}] \mid \boldsymbol{x}] \\ &= E[\ E[I_{\{1\}}(Y) \mid \theta] \mid \boldsymbol{x}] \\ &= E[\ P(Y = 1 \mid \theta) \mid \boldsymbol{x}] \\ &= E[\theta \mid \boldsymbol{x}] \\ &= \frac{t+a}{n+a+b} \end{aligned}$$

Example 1.4 *Let* $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} Ex(\theta)$ *with* $\theta \sim G(a, b)$*,* $a, b > 0$ *known.*

It is easy to see that $\theta \mid \boldsymbol{x} \sim G(n + a, b + t)$ where $t = \sum x_i$. Suppose that we want to predict the next observation, $Y = X_{n+1}$, which is statistically independent of the previous. Hence,

$$
\begin{aligned}
f(y \mid \boldsymbol{x}) &= \int_0^{+\infty} f(y \mid \theta)\ \pi(\theta \mid \boldsymbol{x})\ d\theta \\
&= (n+a)\ \left(\frac{b+t}{b+y+t}\right)^{n+a}\ \left(\frac{1}{b+y+t}\right)\ , \quad y > 0
\end{aligned}
$$

(Gamma-Gamma distribution.)

We do not necessarily need $f(y \mid \boldsymbol{x})$ to obtain point estimates of Y:

$$\begin{aligned}
E[Y \mid \boldsymbol{x}] &= E[\ E[Y \mid \boldsymbol{x}, \theta] \mid \boldsymbol{x}] \\
&= E[\ E[Y \mid \theta] \mid \boldsymbol{x}] \\
&= E[1/\theta \mid \boldsymbol{x}] \\
&= \int_0^{+\infty} \frac{1}{\theta}\ \pi(\theta \mid \boldsymbol{x})\ d\theta \\
&= \ \cdots \\
&= \frac{b+t}{n+a-1}
\end{aligned}$$

Example 1.5 *Let* $X_1, \ldots, X_n \mid \mu \overset{iid}{\sim} N(\mu, 1)$*.*

We saw that in this case the Jeffreys prior is $\pi^J(\mu) \propto 1$ leading to $\mu \mid \boldsymbol{x} \sim \mathrm{N}(\bar{x}, 1/n)$. Suppose we want to predict the average of the next m observations, $\bar{Y} = \sum_{j=1}^{m} X_{n+j}/m$.

$$\begin{aligned} E[\bar{Y} \mid \boldsymbol{x}] &= E[\ E[\bar{Y} \mid \boldsymbol{x}, \mu] \mid \boldsymbol{x}] \\ &= E[\ E[\bar{Y} \mid \mu] \mid \boldsymbol{x}] \end{aligned}$$

Since $\bar{Y} \mid \mu \sim \mathrm{N}(\mu, 1/m)$, we get

$$E[\bar{Y} \mid \boldsymbol{x}] = E[\mu \mid \boldsymbol{x}] = \bar{x}$$

1.7 Bayesian interval estimation

We now want an interval summary of the posterior distribution

Definition 1.1 *We say that* $R(\boldsymbol{x}) = (a(\boldsymbol{x}), b(\boldsymbol{x})) \subset \Theta \subset \mathbb{R}$ *is a* $(1-\alpha)$ *posterior credible interval for* θ *if*

$$P(\theta \in R(\boldsymbol{x}) \mid \boldsymbol{x}) = P(a(\boldsymbol{x}) < \theta < b(\boldsymbol{x}) \mid \boldsymbol{x}) = 1 - \alpha$$

Remarks:

- If $\pi(\theta \mid \boldsymbol{x})$ is continuous

$$P(\theta \in R(\boldsymbol{x}) \mid \boldsymbol{x}) = \int_{a(\boldsymbol{x})}^{b(\boldsymbol{x})} \pi(\theta \mid \boldsymbol{x})\, d\theta = 1 - \alpha$$

- Recall that $C(\boldsymbol{X})$ is a $(1-\alpha)$ random confidence interval for θ if

$$P(\theta \in C(\boldsymbol{X}) \mid \theta) = 1 - \alpha \quad \forall\, \theta \in \Theta$$

but about the observed interval, $C(\boldsymbol{x})$, we can only state that

$$P(\theta \in C(\boldsymbol{x}) \mid \theta) = \begin{cases} 1 & \text{if } \theta \in (\boldsymbol{x}) \\ 0 & \text{otherwise} \end{cases}$$

hence the need for the concept of “confidence”

Example 1.6 *Let* $X_1, \ldots, X_n \mid \mu \overset{iid}{\sim} N(\mu, 1)$ *and suppose we use Jeffreys prior for* μ, $\pi^J(\mu) \propto 1$. *We know that* $\mu \mid \boldsymbol{x} \sim N(\bar{x}, 1/n)$.

Obtain the $(1-\alpha)$ credible interval for μ:

- There are infinitely many intervals $(a(\boldsymbol{x}), b(\boldsymbol{x}))$ such that $P(a(\boldsymbol{x}) < \mu < b(\boldsymbol{x}) \mid \boldsymbol{x}) = 1 - \alpha$.
- For simplicity, we often obtain central credible intervals i.e. $P(\theta < a(\boldsymbol{x}) \mid \boldsymbol{x}) = P(\theta > b(\boldsymbol{x}) \mid \boldsymbol{x}) = \alpha/2$
- The HPD (highest posterior density) credible interval: obtain c such that $R(\boldsymbol{x}) = \{\theta : \pi(\theta \mid \boldsymbol{x}) \geq c\}$ and $P(\theta \in R(\boldsymbol{x})) = 1 - \alpha$
- in this case, the two intervals coincide: $\bar{x} \pm \frac{1}{\sqrt{n}}\, z_{\alpha/2}$, where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$
- note that this coincides with the typical confidence interval for μ

Cont:

- In general, credible and confidence intervals will not coincide and in any case their interpretation is different
- Example: $n = 30$, $\bar{x} = 25$ and $1 - \alpha = 0.9$, which implies $z_{\alpha/2} = 1.64$. Hence, the credible interval is $(24.70, 25.30)$
- For the credible interval we can state that

$$P(\mu \in (24.70, 25.30) \mid \boldsymbol{x}) = 0.9$$

whereas for the confidence interval we can only state that

$$P(\mu \in (24.70, 25.30) \mid \mu) = I_{(24.70,25.30)}(\mu)$$

Example 1.7 *Let* $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} Ex(\theta)$ *and assume* $\theta \sim G(a, b)$, $a, b > 0$ *known.*

We know that $\theta \mid \boldsymbol{x} \sim \mathrm{G}(a+n, b+t)$ where $t = \sum x_i$.

- It is easy to see that the $(1-\alpha)$ HPD is $(\underline{\theta}, \bar{\theta})$ satisfying

$$\mathrm{G}(\underline{\theta} \mid a+n, b+t) = \mathrm{G}(\bar{\theta} \mid a+n, b+t)$$

$$\int_{\underline{\theta}}^{\bar{\theta}} \mathrm{G}(\theta \mid a+n, b+t)\ d\theta = 1-\alpha$$

- The central credible interval is $(\underline{\theta}, \bar{\theta})$ such that

$$P(\theta > \bar{\theta} \mid \boldsymbol{x}) = \alpha/2$$

$$P(\theta < \underline{\theta} \mid \boldsymbol{x}) = \alpha/2$$

  which means that

$$\bar{\theta} = \frac{1}{2(b+t)} F^{-1}_{\chi^2(2(n+a))}(1-\alpha/2)$$

$$\underline{\theta} = \frac{1}{2(b+t)} F^{-1}_{\chi^2(2(n+a))}(\alpha/2)$$

- Example: if $n = 10$, $t = 10$, $a = b = 1$, $1 - \alpha = 0.99$, $F^{-1}_{\chi^2(22)}(0.005) = 8.643$, $F^{-1}_{\chi^2(22)}(0.995) = 42.796$, which leads to the credible interval $(0.39, 1.95)$