STAT 4010 – Bayesian Learning

Tutorial 7

Spring 2022

Cheuk Hin (Andy) CHENG (Email | Homepage)

Di SU (Email | Homepage)

1 Region Estimation

Definition 1. *Given a prior distribution $\theta \sim \pi(\theta)$, and $\alpha \in (0,1)$. Then*

1. *A set $\widehat{I}$ is said to an $100(1-\alpha)\%$* ***credible set*** *if $\mathrm{P}(\theta \in \widehat{I} \mid x) \geq 1-\alpha$.*
2. *The $100(1-\alpha)\%$* ***HPD credible set*** *(where HPD stands for highest posterior density) is*
$$\widehat{I}_{\mathrm{HPD}} = \{\theta \in \Theta : \pi(\theta \mid x) \geq k\}$$
where k is the largest constant such that $\mathrm{P}\left(\theta \in \widehat{I}_{\mathrm{HPD}} \mid x\right) \geq 1-\alpha$.
Assume the posterior is continuous. Then the HPD credible set is the Bayes estimator under the loss
$$L(\theta, \widehat{I}) = k|\widehat{I}| + \mathbb{1}(\theta \notin \widehat{I}), \tag{1.1}$$
which measures both the coverage and the width of $\hat{I}$.
3. *The $100(1-\alpha)\%$* ***equal tailed credible set*** *is*
$$\widehat{I}_{\mathrm{ET}} = \left[Q_{\alpha/2}(\theta \mid x), \quad Q_{1-\alpha/2}(\theta \mid x)\right]$$
where $Q_p(\theta \mid x)$ is the pth quantile of $[\theta \mid x]$.

Lemma 1.1. *Suppose the PDF $\theta \mapsto \pi(\theta \mid x)$ is (i) symmetric about $\theta = \theta_0$, (ii) continuous, and (iii) strictly increasing in $(-\infty, \theta_0]$ and is strictly decreasing in $[\theta_0, +\infty)$. Then, given the same credible level $1-\alpha$,*
$$\widehat{I}_{\mathrm{HPD}} = \widehat{I}_{\mathrm{ET}}$$

Definition 2. *Let $g(\cdot)$ be a strictly increasing and continuous function, and $\phi = g(\theta)$. A principle for constructing region estimator is invariant under the transformation g if*
$$\left[\widehat{L}_\phi, \widehat{U}_\phi\right] = \left[g\left(\widehat{L}_\theta\right), g\left(\widehat{U}_\theta\right)\right]$$
where $\left[\widehat{L}_\phi, \widehat{U}_\phi\right]$ and $\left[\widehat{L}_\theta, \widehat{U}_\theta\right]$ are region estimators for ϕ and θ, respectively.

Lemma 1.2. *Suppose that $[\theta \mid x]$ is a continuous RV.*

1. *The "principle of HPD" for constructing credible sets is NOT invariant to continuous re-parametrization.*
2. *The "principle of equal tails" for constructing credible sets is invariant to continuous re-parametrization. (Because of the invariance of the quantile function)*

Example 1.1. (The Bayes estimator of the confidence index) Under the quadratic loss

$$L(\theta, \widehat{\alpha}) = \{\widehat{\alpha} - \mathbb{1}(\theta \in I)\}^2,$$

where $\widehat{\alpha} \in [0, 1]$, the Bayes estimator is given by $\widehat{\alpha}_\pi = \mathrm{P}(\theta \in I \mid x)$.

Remark 1.1. (Riemann Sum) Let $f : [a, b] \to \mathbb{R}$ be a function defined on a closed interval $I = [a, b]$ of the real numbers, $\mathbb{R}$, and

$$P = \{[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n]\},$$

be a partition of I, where

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b.$$

A Riemann sum S of f over I with partition P is defined as

$$S = \sum_{i=1}^{n} f(x_i^*) \Delta x_i$$

where $\Delta x_i = x_i - x_{i-1}$ and $x_i^* \in [x_{i-1}, x_i]$. This limiting value, if it exists, is defined as the definite Riemann integral of the function over the domain,

$$\int_a^b f(x)dx = \lim_{\|\Delta x\| \to 0} \sum_{i=1}^{n} f(x_i^*) \Delta x_i.$$

Thus, to find $\int_a^b f(x)\mathrm{d}x$, we can divide I as $I = [a, a+h] \cup [a+h, a+2h] \cup \cdots \cup [a+mh, b]$ where h is a step size that is close to zero, and $m = \lfloor (a-b)/h \rfloor - 1$. Then we have the following numerical approximation

$$\int_a^b f(x)\mathrm{d}x \approx \sum_{i=0}^{m} f(a+ih)h \approx (a-b)\bar{f}_i, \tag{1.2}$$

where $\bar{f}_i$ is the average of $\{f(a+ih)\}_{i=0,\cdots,m}$. Riemman sum is named after nineteenth century German mathematician Bernhard Riemann.

Example 1.2. (Wild guess - A4 Fall 2019). After grading the mid-term exam, Keith would like to analyze how well the students did in the multiple-choice questions (MCQs). Let $x_{ij} = \mathbb{1}$ (the i th student answer the j MCQ correctly), $i = 1, \ldots, n$, $\quad j = 1, \ldots, m$, where $n = 42$ and $m = 8$. Consider the model

$$[x_{ij} \mid \theta] \overset{\text{IID}}{\sim} \text{Bern}(\theta), \quad i = 1, \ldots, n, \quad j = 1, \ldots, m,$$

and the prior $\theta \sim \pi(\theta)$ defined by

$$\theta \sim \text{Beta}(\alpha, \beta),$$

where $\alpha = \beta = 1/2$. Denote the entire dataset by $X = \{x_{ij} : 1 \le i \le n, 1 \le j \le m\}$. The dataset MCQ.txt can be downloaded from the course website. Because of the privacy concern, not less than 10% of the data have been replaced by imputed values such that the structure of the privacy-protected dataset is close to that of the actual one.

Consider $\widehat{I} \in \{[a,b] : 0 \le a \le b \le 1\}$, and the loss

$$L(\theta, \widehat{I}) = k|\widehat{I}| + \mathbb{1}(\theta \notin \widehat{I})$$

where $|\widehat{I}|$ denotes the width of the interval $\widehat{I}$, and $k = 1$ is fixed.

1. Compute the Bayes estimator $\widehat{I}_\pi$. What is the credible level $\mathrm{P}\left(\theta \in \widehat{I}_\pi \mid X\right)$ of $\widehat{I}_\pi$?
2. Compute the 95% HPD credible interval $\widehat{I}_{\text{HPD}}$ of θ.
3. Compute the 95% equal-tailed credible interval $\widehat{I}_{\text{ET}}$ of θ.

(Hints: (a) Example 5.5. (b): Example 5.7. (c) Example 5.1.)

Solution:

1. By Example 2.13,
$$[\theta \mid x_{ij}] \sim Beta(\alpha + S_n, \beta + n - S_n),$$
where $S_n = \sum_{i=1}^n \sum_{j=1}^m x_{ij}$. After some calculation, we get $\alpha + S_n = 140.5$, $\beta + n - S_n = 196.5$. By Example 5.5, we have that
$$\hat{I}_\pi = \hat{I}_{HPD},$$
with $k = 1$. By R, we have $\hat{I}_{HPD} = [0.355, 0.480]$ is the Bayes estimator up to 3 decimal places. Its credible level is $P(\theta \in \hat{I}_\pi \mid X) = P(\theta \in \hat{I}_{HPD} \mid X) = 0.980$.

```
### Step 1: Initialization
data = read.table("MCQ.txt", header = T);
X = as.matrix(data);
a = 0.5; b=0.5;
n = dim(X)[1];
m = dim(X)[2];
alpha.b = sum(X) + a;
beta.b = n*m - sum(X) + b
# alpha.b = 140.5
# beta.b = 196.5

### Step 2: Find Bayes estimator
theta = seq(0, 1, length = 5001)
# notice that the posterior has a closed form
d = dbeta(theta, alpha.b, beta.b)
L = theta[which.min(abs(d[1:which.max(d)]-1))]
U = theta[which.min(abs(d[which.max(d)+1:length(d)]-1))+which.max(d)]
L
[1] 0.355
U
[1] 0.4796

### Step 3: Find Credible level
pbeta(U,alpha.b,beta.b)-pbeta(L,alpha.b,beta.b)
[1] 0.9800575
```

2. Using R, $\hat{I}_{HPD} = [0.4164, 0.470]$ with credible level: 0.950.

```r
# Method 1
{
    alpha = 0.05
    ### step 1: values of posterior at different values of theta in [a,b]
    theta = seq(from=0.3, to=0.5, length=501)
    d = dbeta(theta, alpha.b, beta.b)

    ### step 2: find the theta that satisfy the credibility requirement
    # O stores indices
    O = order(d,decreasing=TRUE)
    # confidence w.r.t. all interval candidates

    ##---Learning Moment: use Riemann sum to approximate integral---##
    step = theta[2]-theta[1]
    conf = cumsum(d[O]*step)

    # check confidence condition
    N = sum(conf<(1-alpha))+1
    selected.index = O[1:N]
    selected = theta[selected.index]
    # results
    k = d[O[N]]
    selected[1]
    selected[N]
    alpha.hat = conf[N]
    alpha.hat

    ### step 3: plot
    plot(theta,d, type="l", lwd=2, col="red4",
         xlab=expression(theta),
         ylab=expression(pi(theta~"|"~italic(x[1:n]))))
    abline(v=selected,col="pink")
    abline(h=c(0,k),v=c(a,b),lty=3, lwd=.75)

    ### exploration
    abline(v = theta[O[1:10]],col='yellow')
    abline(v = theta[O[1:50]],col='yellow')
    abline(v = theta[O[1:100]],col='yellow')
    abline(v = theta[O[1:150]],col='yellow')

    abline(h = d[O[10]],lty = 2,col='orange')
    abline(h = d[O[50]],lty = 2,col='orange')
    abline(h = d[O[100]],lty = 2,col='orange')
    abline(h = d[O[150]],lty = 2,col='orange')

    conf[1:50]

}

# Method 2
{
    ### search for all possible HPD set (with different credible level)
    i.max = which.max(d)
    out = array(NA, dim=c(i.max,3))
    colnames(out) = c("Lower-bound","Upper-bound","Credible-level")
    for(i.left in 1:i.max){
        delta = abs(d[i.left]-d[-(1:i.max)])
        out[i.left,1] = theta[i.left]
```

```
        i.right = i.max+which.min(delta)
        out[i.left,2] = theta[i.right]
        out[i.left,3] = pbeta(out[i.left,2],alpha.b,beta.b)-pbeta(out[i.left,1],alpha.b,beta.b)
}

    ### select the HPD set with a desired creible level
    (result = out[which.min(abs(out[,3]-0.95)),])

    ### plot
    abline(v=c(result[1:2]),col="blue4",lwd=3)

    # exploration
    plot(theta,d, type="l", lwd=2, col="red4",
         xlab=expression(theta),
         ylab=expression(pi(theta~"|"~italic(x[1:n]))))
    abline(h=c(0,k),v=c(a,b),lty=3, lwd=.75)
    abline(v=c(result[1:2]),col="blue4",lwd=3)
    abline(v = c(out[50,1],out[50,2]),col='blue',lty=4)
    abline(v = c(out[100,1],out[100,2]),col='blue',lty=4)
    abline(v = c(out[150,1],out[150,2]),col='blue',lty=4)

}

##---Learning Moment: use Riemann sum to approximate integral---##
riemann = function(f, left, right, step){
    theta = seq(from = left,to = right,by = step)
    f.all = f(theta)
    sum1 = sum(f.all*step)
    sum2 = mean(f.all)*(right-left)
    sum3 = pbeta(right,alpha.b,beta.b)-pbeta(left,alpha.b,beta.b)
    out = array(c(sum1,sum2,sum3),dimnames=list(c('Riemann Sum','Sample average','pbeta')))
    out
}
post = function(x) dbeta(x,alpha.b,beta.b)
l=0.3
r=0.5
riemann(f = post,left = l,right = r,step = 0.01)
```

Remark 1.2. Method 1 is better than Method 2 in the sense that Method 1 starts with the shortest intervals containing the MAP, and needs to consider less candidates. In assignments and projects, write a `HPD` function with suitable inputs and outputs.

3. Using R, we have $\hat{I}_{ET} = [0.365, 0.470]$.

```
    qbeta(c(0.025,0.975),alpha.b,beta.b)
    [1] 0.3648527 0.4699094
```

2 Theoretical Justification

This section shows that the Bayesian methods studied in previous chapters are theoretically sensible.

Definition 3. *Given any DGP $f_\star(x)$ and model $\mathscr{F} = \{f(x \mid \theta) : \theta \in \Theta\}$. Denote the expectation and variance under the DPG $f_\star(x)$ by* $\mathrm{E}_\star$ *and* $\mathrm{Var}_\star$*. Define*

$$\theta_\star = \underset{\theta\in\Theta}{\arg\max}\mathrm{E}_\star \{\log f(x_1 \mid \theta)\},$$

and

$$I_\star = \left[\mathrm{Var}_\star\left\{\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x_1 \mid \theta)\right\}\right]_{\theta=\theta_\star} \quad J_\star = \left[-\mathrm{E}_\star\left\{\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x_1 \mid \theta)\right\}\right]_{\theta=\theta_\star},$$

provided that the expectations exist. The quantities $I_\star$ and $J_\star$ are called Fisher information. If $\mathscr{F}$ well specifies $f_\star$, then $\theta_\star = \theta_0$ and $I_\star = J_\star$, where θ_0 is the true DGP parameter.

Theorem 2.1. *(Consistency of posterior). Assume regularity conditions (RCs). If n is large enough, then*

$$\widehat{\theta}_{\mathrm{MLE}} \approx \theta_\star \quad \textit{and} \quad [\theta \mid x_{1:n}] \approx \theta_\star.$$

Theorem 2.2. *(Asymptotic distributions of posterior). Assume RCs. If n is large enough, then*

$$\widehat{\theta}_{\mathrm{MLE}} \approx \mathrm{N}\left(\theta_\star, \frac{J_\star^{-1}I_\star J_\star^{-1}}{n}\right) \quad \textit{and} \quad [\theta \mid x_{1:n}] \approx \mathrm{N}\left(\widehat{\theta}_{\mathrm{MLE}}, \frac{J_\star^{-1}}{n}\right).$$

If the model is well-specified, the precision of Bayesian framework and frequentist framework are consistent.

Theorem 2.3. *(Asymptotic representation of posterior mean). Assume RCs. If n is large enough, then*

$$\mathrm{E}(\theta \mid x_{1:n}) \approx \widehat{\theta}_{\mathrm{MLE}}.$$

Remark 2.1. Intuitively, if the model is well-specified, we will have the following approximation

$$[\theta \mid x_{1:n}] \text{``}\approx\text{''} \widehat{\theta}_{\mathrm{MLE}} \approx \theta_\star = \theta_0.$$

Be careful that $\hat{\theta}_{MLE}$ is the maximizer $\log f(x \mid \theta)$, whereas $\theta_\star$ is the maximizer of $\log f(x \mid \theta)$ after taken expectation w.r.t. $f_\star(x)$. In practice, $f_\star$ may not be known.

Theorem 2.4. *We have the following bi-directional relation*

$$\begin{aligned} & x_{1:n} \text{ are exchangeable with joint density } f(x_{1:n}) \\ \iff \quad & \exists\theta \in \Theta, f(x \mid \theta), \pi(\theta) \text{ s.t. } \begin{cases} [x_{1:n} \mid \theta] \overset{IID}{\sim} f(x_{1:n} \mid \theta) \\ \theta \sim \pi(\theta). \end{cases} \end{aligned}$$

The direction "$\Longrightarrow$" is stated in theorem 6.5. De Finiti Theorem, and the direction "$\Longleftarrow$" is given in proposition 6.4.

Remark 2.2. Bayesian model enables us to work with IID data, by "extracting out" the dependence within exchangeable data.

Proof. (of Proposition 6.4.) Denote the PDF of $y_{1:n}$ by $f(x_1, \ldots, x_n)$. For any permutation σ of $\{1, \ldots, n\}$, we have

$$\begin{aligned}
&f(x_1, \ldots, x_n) \\
&\qquad = \int_\Theta \pi(\theta) f(x_1, \ldots, x_n \mid \theta)\, \mathrm{d}\theta && \because \text{By the definition of marginal PDF} \\
&\qquad = \int_\Theta \pi(\theta) \prod_{i=1}^n f(x_i \mid \theta)\, \mathrm{d}\theta && \because y_{1:n} \text{ are conditionally independent given } \theta \\
&\qquad = \int_\Theta \pi(\theta) \prod_{i=1}^n f\left(x_{\sigma(i)} \mid \theta\right) \mathrm{d}\theta && \because \text{Product is commutative} \\
&\qquad = f\left(x_{\sigma(1)}, \ldots, x_{\sigma(n)}\right) && \because \text{By the definition of marginal PDF}
\end{aligned}$$

Thus, $x_{1:n}$ are exchangeable.