BAYESIAN STATISTICS

Chapter 5

Instructor: Seokho Lee

Hankuk University of Foreign Studies

5. Bayesian Inference

5.1. Point estimation

In the traditional statistics, the maximum likelihood estimation (MLE) is used to estimate θ.

In Bayesian statistics, the **posterior mode** is used for the **point estimation** for θ.

Suppose $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is the random sample and the posterior is $h(\theta|\mathbf{X})$. The posterior mode is defined as

$$\hat{\theta} = \underset{\theta \in \Theta}{arg\ max}\ h(\theta|\mathbf{X}).$$

Example (5-1)

Suppose $X \sim N(\theta, 1)$ and $\theta \sim N(0, 1)$. Find the posterior mode $\hat{\theta}$.

(solution) From the chapter 4, the posterior distribution is $\mathrm{N}(x/2, 1/2)$. Thus,

$$h(\theta|X) = \frac{1}{\sqrt{\pi}} e^{-\frac{1}{2}\frac{(\theta - x/2)^2}{1/2}}, \quad -\infty < \theta < \infty.$$

$\hat{\theta} = X/2$ which maximizes $h(\theta|X)$.

Example (5-2)

The random sample, $X_1, X_2, \ldots, X_n$, are obtained from Poisson(θ). The prior is assumed to be the conjugate distribution Gamma(α, β). What is the posterior mode of θ?

(solution) From the chapter 4, the posterior distribution is Gamma$(n\bar{x} + \alpha, 1/(n + 1/\beta))$. Thus,

$$h(\theta|\mathbf{x}) = C \cdot \theta^{n\bar{x}+\alpha-1} e^{-\theta(n+1/\beta)}$$

where C is the constant term. To find out the maximizer of $h(\theta|\mathbf{x})$, take the logarithm

$$\log h(\theta|\mathbf{x}) = (n\bar{x} + \alpha - 1) \log \theta - (n + 1/\beta)\theta + \log C.$$

And take the derivative with respect to θ and set to zero. Then

$$\begin{aligned} &\tfrac{\partial \log h(\theta|\mathbf{x})}{\partial \theta} = \tfrac{n\bar{x}+\alpha-1}{\theta} - (n + 1/\beta) = 0 \\ \Rightarrow \quad &\hat{\theta} = \tfrac{n\bar{x}+\alpha-1}{n+1/\beta}. \end{aligned}$$

Note that

$$\hat{\theta} = w_n \bar{x} + (1 - w_n)(\alpha - 1)\beta$$

where $0 < w_n = n/(n + 1/\beta) < 1$. This posterior mode is the form of weight average between the mle ($\bar{x}$) and the prior mode ($(\alpha - 1)\beta$).

There are some practical problems with the posterior mode:

- The posterior mode often does not exist
- The posterior mode is often far away from the center of the posterior
- The posterior is often multimodal

Instead of the posterior mode, the **posterior mean** or the **posterior median** are used.

The posterior mean is defined as

$$E(\theta|\mathbf{X}) = \begin{cases} \sum_{\vartheta\in\Theta} \vartheta h(\vartheta|\mathbf{X}) & \mathbf{X} \text{ is discrete} \\ \int_\Theta \vartheta h(\vartheta|\mathbf{X}) d\vartheta & \mathbf{X} \text{ is continuous.} \end{cases}$$

Example (5-1 continue)

Find the estimator θ using the posterior mean and the posterior median.

(solution) Since the posterior $\mathrm{N}(x/2, 1/2)$ is symmetric and bell-shaped, the mean, median and mode are the same. Thus, the estimators from the posterior mean and the posterior median are $x/2$.

Example (5-2 continue)

Find the estimator θ using the posterior mean.

(solution) The posterior mean is

$$E(\theta|\mathbf{X}) = \frac{n\bar{X}+\alpha}{n+1/\beta}.$$

Note that

$$E(\theta|\mathbf{X}) = w_n\bar{x} + (1 - w_n)\alpha\beta$$

with $w_n = n/(n + 1/\beta)$. So the posterior mean is the weighted average of the mle and the prior mean. (show it)

5.2. Credible region

A $100(1-\alpha)\%$ **credible region** or **credible set**, C, satisfies

$$1-\alpha \leq \Pr(\theta \in C|\mathbf{x}) = \begin{cases} \sum_{\vartheta \in C} h(\vartheta|\mathbf{x}) & \theta \text{ is discrete} \\ \int_C h(\vartheta|\mathbf{x}) d\vartheta & \theta \text{ is continuous.} \end{cases}$$

The credible region is analogous to the confidence interval from the traditional statistics.

A credible region is not uniquely defined. To choose the best, one may find the credible region as narrow as possible.

The **highest posterior density (HPD)** credible region, C, has the narrowest region and satisfies

1. $\Pr(\theta \in C|\mathbf{x}) \geq 1-\alpha$.
2. For all $\theta_1 \in C$ and $\theta_2 \notin C$, $h(\theta_1|\mathbf{x}) \geq h(\theta_2|\mathbf{x})$.

If the posterior is symmetric and bell-shaped, then HPD credible region is symmetric about the posterior mode.

Example (5-1 continue)

Find the $100(1-\alpha)\%$ *HPD credible region for* θ*.*

(solution) Since the posterior is $\mathrm{N}(x/2, 1/2)$, $100(1-\alpha)\%$ HPD credible region is

$$C = \left(\frac{X}{2} - \frac{z_{\alpha/2}}{\sqrt{2}}, \frac{X}{2} + \frac{z_{\alpha/2}}{\sqrt{2}} \right)$$

where z_α is the quantity satisfying $\int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = \alpha$.

Example (5-3)

$X \sim N(\theta, 1)$ *and the prior of* θ *is* $\pi(\theta) = 1$*. Find the* $100(1-\alpha)\%$ *HPD credible region for* θ*.*

(solution) Since the posterior is $\mathrm{N}(x, 1)$, $100(1-\alpha)\%$ HPD credible region is

$$C = (X - z_{\alpha/2}, X + z_{\alpha/2})$$

Suppose $h(\theta|\mathbf{x})$ is continuous on θ. Consider

$$C_\kappa = \{\theta \in \Theta | h(\theta|\mathbf{x}) \geq \kappa\}.$$

If we can find κ such that

$$\Pr(\theta \in C_\kappa|\mathbf{x}) = \int_{C_\kappa} h(\vartheta|\mathbf{x}) d\vartheta = 1 - \alpha,$$

then C_κ is the $100(1-\alpha)\%$ HPD credible region. κ can be found numerically.

Some difficulties with HPD credible region:

- In general, the HPD credible region is hard to compute.
- If the mass of the posterior is concentrated on the tail parts, the HPD credible region is located on the tail, so that its interpretation is not natural.
- If the posterior is multimodal, the HPD credible region may be a union of several disjoint regions.

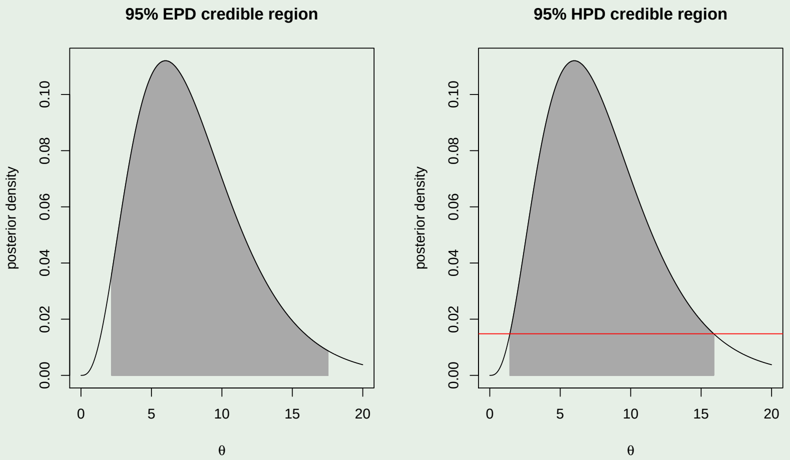
Instead of the HPD credible region, one may use the **equal-tail posterior density** (**EPD**) credible region. The $100(1-\alpha)\%$ EPD credible region is defined as the interval (a, b) satisfying

$$\begin{cases} \sum_{\vartheta \leq a} h(\vartheta|\mathbf{x}) = \frac{\alpha}{2} = \sum_{\vartheta \geq b} h(\vartheta|\mathbf{x}) & \text{if } \theta \text{ is discrete} \\ \int_{-\infty}^{a} h(\vartheta|\mathbf{x}) d\vartheta = \frac{\alpha}{2} = \inf_{b}^{\infty} h(\vartheta|\mathbf{x}) d\vartheta & \text{if } \theta \text{ is continuous.} \end{cases}$$

Example (HPD and EPD)

```
> x<-seq(0,20,by=0.001)
> den<-dgamma(x,shape=4,scale=2)
> par(mfrow=c(1,2))
> #### EPD credible region
> plot(x,den,type='l',col='white',xlab=expression(theta),
ylab='posterior density',main='95% EPD credible region')
> ( interval<-c(qgamma(0.025,shape=4,scale=2),qgamma(0.975,shape=4,scale=2)) )
[1]  2.179731 17.534546
> x.epd<-seq(interval[1],interval[2],by=0.001)
> lines(x.epd,dgamma(x.epd,shape=4,scale=2),type='h',col='darkgray')
> lines(x,den)
> #### HPD credible region
> k<-0.01
> for(k in seq(0,0.1,by=0.0001)){
+  idx.x<-which(den>k)
+  left<-pgamma(min(x[idx.x]),shape=4,scale=2)
+  right<-pgamma(max(x[idx.x]),shape=4,scale=2)
+  p<-right-left
+  if(p<=0.95) break
+ }
> c(p,min(x[idx.x]),max(x[idx.x]))
[1]  0.9499173  1.4260000 15.8920000
> plot(x,den,type='l',col='white',xlab=expression(theta),
ylab='posterior density',main='95% HPD credible region')
> lines(x[idx.x],den[idx.x],col='darkgray',type='h')
> lines(x,den)
> abline(h=k,col='red')
> dev.copy2pdf(file='fig5-1.pdf')
```

Example (HPD and EPD, continue)



5.3. Predictive inference

A motor company, of course, is interested in the average sales, θ, per a month. As well as this, the company is also interested in predicting how many cars will be sold in the coming month. In this case, we need to **predict** a random variable of 'the number of cars sold in the next month.'

The prediction uses the **predictive probability distribution**.

Consider the prediction of the unobserved random variable $Z \sim g(z|\theta)$ based on the observed value x of a random variable $X \sim f(x|\theta)$. X and Z are assumed to be independent. Then, the **predictive density** $p(z|x)$ is defined as

$$p(z|x) = \begin{cases} \sum_{\vartheta \in \Theta} g(z|\vartheta)h(\vartheta|x) & \text{if } \theta \text{ is discrete} \\ \int_\Theta g(z|\vartheta)h(\vartheta|x)d\vartheta & \text{if } \theta \text{ is continuous.} \end{cases}$$

Example (5-4)

Let X be the number of cars sales persons during a week and assume $X \sim Poisson(\theta)$. If the new employee sold one car in his/her first week. What is the probability that he/she will sell one car in the next week.

(solution) If Z is the number of cars he/she will sell in the next week, then Z also follows Poisson distribution. Thus, we would like to compute $p(z=1|x=1)$. Consider the conjugate prior Gamma$(\alpha=2, \beta=1)$ for θ. The posterior becomes

$$\theta|x \sim \text{Gamma}(\alpha+x, (1+1/\beta)^{-1}).$$

Therefore, the predictive distribution is

$$\begin{aligned} p(z|x) &= \int_0^\infty \frac{e^{-\vartheta}\vartheta^z}{z!} \cdot \frac{1}{\Gamma(\alpha+x)(1+1/\beta)^{-\alpha}} \vartheta^{\alpha+x-1} e^{\vartheta(1+1/\beta)} d\vartheta \\ &= \frac{\Gamma(z+\alpha+x)}{z!\Gamma(\alpha+x)} \cdot \frac{(1+1/\beta)^\alpha}{(2+1/\beta)^{z+\alpha+x}}. \end{aligned}$$

So, $p(1|1) = \frac{\Gamma(1+2+1)}{1!\Gamma(2+1)} \cdot \frac{(1+1/.5)^1}{(2+1/.5)^{1+2+1}} \approx .1055$.

5.4. Bayesian decision theory

Every decision risks the **loss**, meaning 'opportunity loss' or 'regrets'.

Consider a car sales problem. Suppose that a is the number of cars the car-dealer orders to the factory, and θ is the number of cars the consumers will purchase. If $a > \theta$ then the extra cars will remain in stock. If $a < \theta$ then the dealer cannot sell the extra cars. Both of cases become a "loss" to the dealer. The optimal case is, of course, $a = \theta$.

The more different a and θ, the more gross the loss. The commonly-used **loss functions**, representing this loss, are

1. **squared-error loss function**: $L(\theta, a) = (\theta - a)^2$.
2. **absolute-error loss function**: $L(\theta, a) = |\theta - a|$.
3. **0-1 loss function**: $L(\theta, a) = \begin{cases} 0, & \theta = a \\ 1, & \theta \neq a. \end{cases}$
4. **linear loss function**: $L(\theta, a) = \begin{cases} K_1(\theta - a), & \theta \geq a \\ K_2(a - \theta), & \theta < a. \end{cases}$

In the statistical sense, θ is the unknown parameter and a is the estimator of θ.

To minimize the loss, Bayesian statistics choses a which minimizes the **posterior expected loss**, $E[L(\theta, a)|\mathbf{X}]$, called **posterior risk**:

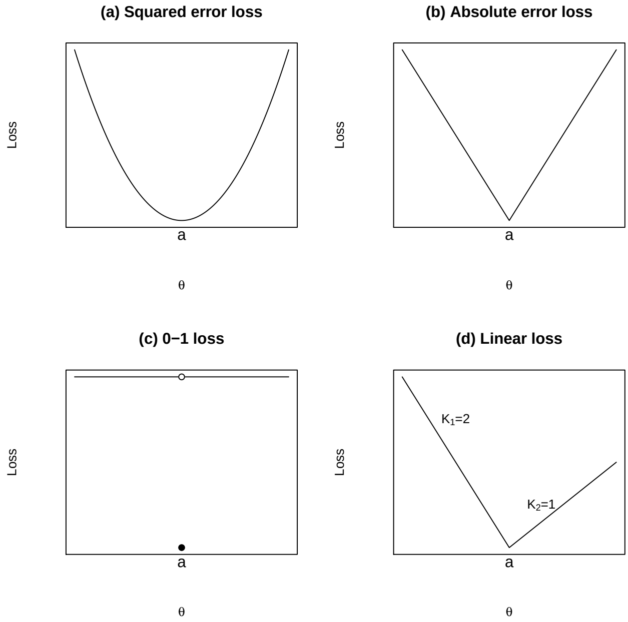
$$E[L(\theta, a)|\mathbf{X}] = \begin{cases} \sum_{\vartheta \in \Theta} L(\vartheta, a) h(\vartheta|\mathbf{X}) & \text{if } \theta \text{ is discrete} \\ \int_{\Theta} L(\vartheta, a) h(\vartheta|\mathbf{X}) d\vartheta & \text{if } \theta \text{ is continuous.} \end{cases}$$

The minimizer of the posterior risk is called **Bayes rule** or **Bayes estimator**, and denoted by $\delta_\pi(\mathbf{X})$.

The Bayes rule $\delta_\pi(\mathbf{X})$ differs depending on the loss function used. If we use the square-error loss function, then

$$\frac{d}{da} E[(\theta - a)^2|\mathbf{X}] = \frac{d}{da} \int_{\Theta} (\theta - a)^2 h(\theta|\mathbf{x}) d\theta = 2a - 2E(\theta|\mathbf{x}) = 0$$

gives $a = E(\theta|\mathbf{X})$. Thus, the Bayes rule is $\delta_\pi(\mathbf{X}) = E(\theta|\mathbf{X})$, which is the posterior mean.


(a) Squared error loss
Loss
a
θ
(b) Absolute error loss
Loss
a
θ
(c) 0−1 loss
Loss
a
θ
(d) Linear loss
Loss
K1=2
K2=1
a
θ

```
> a<-0
> theta<-seq(-1,1,by=0.01)
> loss1<-(theta-a)^2
> loss2<-abs(theta-a)
> loss3<-ifelse(theta==a,0,1)
> loss4<-ifelse(theta>a,1*(theta-a),2*(a-theta))
> par(mfrow=c(2,2))
> plot(theta,loss1,type='l',xlab=expression(theta),ylab='Loss',
main='(a) Squared error loss',xaxt='n',yaxt='n')
> mtext('a',side=1)
> plot(theta,loss2,type='l',xlab=expression(theta),ylab='Loss',
main='(b) Absolute error loss',xaxt='n',yaxt='n')
> mtext('a',side=1)
> plot(theta,rep(1,length(theta)),ylim=c(0,1),type='l',xlab=expression(theta),ylab='Loss',
main='(c) 0-1 loss',xaxt='n',yaxt='n')
> mtext('a',side=1)
> points(a,1,pch=19,col='white'); points(a,1)
> points(a,0,pch=19)
> plot(theta,loss4,type='l',xlab=expression(theta),ylab='Loss',
main='(d) Linear loss',xaxt='n',yaxt='n')
> mtext('a',side=1)
> text(-0.5,1.5,expression(paste(K[1],'=2',sep='')))
> text(0.3,0.5,expression(paste(K[2],'=1',sep='')))
> dev.copy2pdf(file='fig5-2.pdf')
```

loss function	Bayes rule ($\delta_\pi(X)$)
$L(\theta, a) = (\theta - a)^2$	posterior mean
$L(\theta, a) = \lvert\theta - a\rvert$	posterior median
$L(\theta, a) = \begin{cases} 0, & \theta = a \\ 1, & \theta \neq a \end{cases}$	posterior mode
$L(\theta, a) = \begin{cases} K_1(\theta - a), & \theta \geq a \\ K_2(a - \theta), & \theta < a \end{cases}$	$\frac{K_1}{K_1+K_2}$-posterior fractile

Example (5-5)

Suppose $X \sim N(\theta, 1)$ and assume the prior for θ is $\theta \sim N(0, 1)$. When we observe $X = 3$, what is the Bayes estimator which minimizes the posterior risk based on the square-error loss?

(solution) Since the posterior becomes $\theta|X \sim \mathrm{N}(X/2, 1/2)$, the Bayes estimate minimizing the posterior risk using the square-error loss is $3/2 = 1.5$.

Example (5-5 continue)

Suppose that it is twice more dangerous to underestimate θ than to overestimate. In such case, it makes sense to use the risk based on the linear loss function of $K_1 = 2$ and $K_2 = 1$. Find out the Bayes estimate.

(solution) The Bayes estimate of θ is the $2/3$-fractile of the posterior $\mathrm{N}(3/2, 1/2)$. Thus

$$3/2 + 0.43\sqrt{1/2} \approx 1.80.$$

(convention in this class) Hereafter, unless there is a specification of the loss function, the Bayes estimator is obtained under the square-error loss function.