STAT 4010 – Bayesian Learning

Tutorial 1

Spring 2022

Cheuk Hin (Andy) CHENG (Email | Homepage)

Di SU (Email | Homepage)

1 Review

1.1 Comparison of Bayesian and Frequentist Philosophy

The differences between the two schools can be summarized in the following table.

	Bayesian	Frequentist
View on probability	Degree of belief (Subjective)	Limiting relative frequencies (Objective)
θ is a	Random variable	Unknown constant
Infer on θ by	Modeling a probability distribution	Procedures with well-defined long run frequency properties
Common methods	Bayes estimator, MAP,Bayes factor	MME,MLE,C.I.

*This is a summary from Ch.11, All of Statistics (Wasserman 2003).
*An insightful discussion between Frequentist and Bayesian methods is given in example 1.7 from lecture note 1.

1.2 Posterior Calculation

Let Θ be the parameter space. Define the sampling density be $f(x \mid \theta)$ and the prior density be $f(\theta)$ (these are given).

- Prior predictive: $f(x) = \int_\Theta f(x,\theta)\mathrm{d}\theta = \int_\Theta f(x \mid \theta)f(\theta)\mathrm{d}\theta$
- Posterior: $f(\theta \mid x_{1:n}) \propto f(x_{1:n} \mid \theta)f(\theta)$
- Posterior predictive:

$$\begin{aligned} f(x_{n+1} \mid x_{1:n}) &= \int_\Theta f(\theta \mid x_{1:n})f(x_{n+1} \mid x_{1:n},\theta)\mathrm{d}\theta \\ &\overset{*}{=} \int_\Theta f(\theta \mid x_{1:n})f(x_{n+1} \mid \theta)\mathrm{d}\theta \end{aligned}$$

*The second equality holds if x_{n+1} and $x_{1:n}$ are conditional independent given θ.

Remark 1.1. The building blocks of Bayesian inference is the prior $f(\theta)$ and the sampling distribution $f(x_{1:n} \mid \theta)$. They are assumed by us based on our own belief. After observing data, inference is conducted through the following formula

$$\begin{aligned} f(\theta \mid x_{1:n}) &= \frac{f(x_{1:n} \mid \theta) f(\theta)}{f(x_{1:n})} \\ &\propto f(x_{1:n} \mid \theta) f(\theta) \end{aligned} \tag{1.1}$$

We call $f(x_{1:n} \mid \theta) f(\theta)$ the **kernel** of the probability density function. The above formula will be used intensively through out the whole course.

Remark 1.2. There are two reasons why we only care about the proportional part in 1.1,

1. The denominator $f(x_{1:n})$ is a "constant" with respect to θ, because it doesn't depend on θ. On the other hand, $f(\theta \mid x_{1:n})$ is the density of θ given the data, it is not a density of $x_{1:n}$. Hence we can "discard" $f(x_{1:n})$ from this density.

2. Using the following relation, once we know $f(x_{1:n} \mid \theta) f(\theta)$, we immediately know $f(x_{1:n})$.

$$f(x_{1:n}) = \int_{\Theta} f(x_{1:n} \mid \theta) f(\theta) \mathsf{d}\theta \tag{1.2}$$

Relation 1.2 comes from the fact that $f(\theta \mid x_{1:n})$ is a density function. We have

$$\int_{\Theta} f(\theta \mid x_{1:n}) \mathsf{d}\theta = \int_{\Theta} \frac{f(x_{1:n} \mid \theta) f(\theta)}{f(x_{1:n})} \mathrm{d}\theta = 1.$$

Remark 1.3. The posterior $f(\theta \mid x_{1:n})$ is a function of θ, and it depends on $x_{1:n}$ as well.

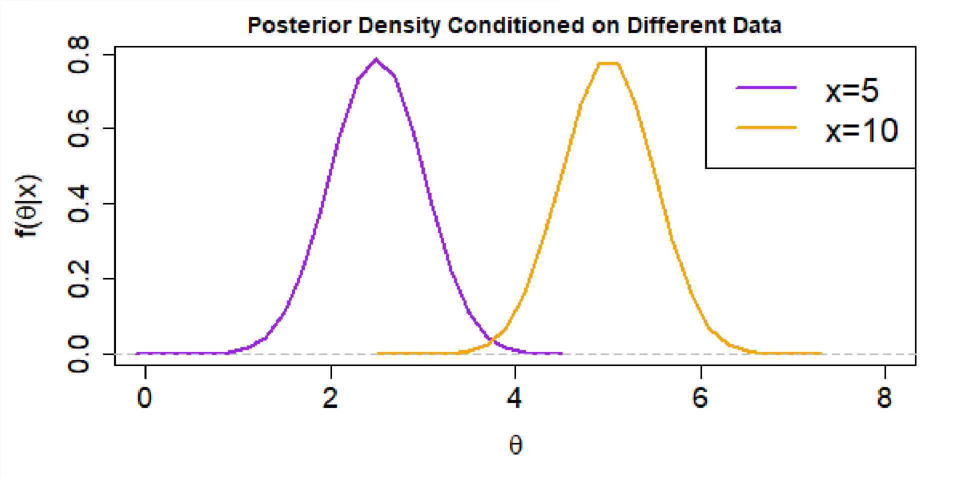
1. First of all, $f(x_{1:n}, \theta)$ is a function of both $x_{1:n}$ and θ.

2. Given $x_{1:n}$, θ is a random variable, and $f(\theta \mid x_{1:n})$ is the density function of θ only. It also depends on $x_{1:n}$ because $f(\theta \mid X_{1:n} = x_{1:n}) = f(X_{1:n} = x_{1:n}, \theta) / f_X(X_{1:n} = x_{1:n})$.

Experiment: Revisit Example 1.3. The model is $[x \mid \theta] \overset{\text{IID}}{\sim} \mathrm{N}(\theta, \sigma^2), \theta \sim \mathrm{N}(\theta_0, \tau_0^2)$. Assume $\sigma^2 = 1, \theta_0 = 0, \tau_0^2 = 1$. Then $[\theta \mid x] \sim \mathrm{N}(x/2, 1/2)$.

- When $x = 5$, $[\theta \mid x = 5] \sim \mathrm{N}(5/2, 1/2)$. It is represented in the following figure in purple.

- When $x = 10$, $[\theta \mid x = 10] \sim \mathrm{N}(5, 1/2)$. It is represented in the following figure in orange.



Takeaway: The data decides which curve the posterior density is.

Example 1.1. Let x denote the random variable of interest. The following table summarizes the density of common distribution (more details are given in chapter 2 lecture note).

Distribution	Kernel (in red)
$\mathrm{N}(\mu, \sigma^2)$	$(2\pi\sigma^2)^{-1/2} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$
$\theta\mathrm{Exp}(1)$	$\frac{1}{\theta}e^{-x/\theta}\mathbb{1}_{\{x>0\}}$
$\mathrm{Ga}(\alpha)/\beta$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}\mathbb{1}_{\{x>0\}}$
$\beta/\mathrm{Ga}(\alpha)$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}e^{-\beta/x}\mathbb{1}_{\{x>0\}}$
χ^2_k	$\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}\mathbb{1}_{\{x>0\}}$
$\mathrm{Po}(\theta)$	$e^{-\theta}\frac{\theta^x}{x!}\mathbb{1}_{\{x=0,1,\ldots\}}$
$\mathrm{Bin}(m, \theta)$	$\binom{m}{x}\theta^x(1-\theta)^{m-x}\mathbb{1}_{\{x=0,1,\ldots\}}$
$\mathrm{Beta}(\alpha, \beta)$	$\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}_{\{x\in(0,1)\}}$

Table 1: Density of Common Distribution

Example 1.2. Let Θ be the support of θ and the density of θ be $f(\theta) = cK(\theta)$, where $K(\theta)$ is the probability kernel and c is a constant (with respect to θ). Then, we have,

$$1 = \int_\Theta f(\theta)\mathrm{d}\theta \Rightarrow \frac{1}{c} = \int_\Theta K(\theta)\mathrm{d}\theta.$$

For example, if $\theta \sim Ga(\alpha)/\beta$, then $c = \beta^\alpha/\Gamma(\alpha)$ and $K(\theta) = \theta^{\alpha-1}e^{-\beta\theta}$. Therefore,

$$\int_0^\infty \theta^{\alpha-1}e^{-\beta\theta}\mathrm{d}\theta = \frac{1}{c} = \frac{\Gamma(\alpha)}{\beta^\alpha},$$

where $\Gamma(\alpha) = (\alpha-1)!$.

And if $\theta \sim \text{Beta}(\alpha, \beta)$, then $c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$, $K(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}$, and

$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\mathrm{d}\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Takeaway: Some integrals can be found without direct calculation. Try to find the integral of the kernels of other common distributions.

Example 1.3. (Poisson-Gamma Model) Consider the following model,

$$\begin{aligned} x_{1:n} \mid \theta &\overset{\text{IID}}{\sim} \text{Po}(\theta) \\ \theta &\sim \text{Ga}(\alpha)/\beta \end{aligned}$$

1. Find the prior predictive density, posterior distribution given $x_{1:n}$ and also the posterior predictive.
2. Find also $\mathsf{E}[\ln\theta \mid x_{1:4}]$, where $x_{1:4} = (6, 7, 5, 4)$ and take $(\alpha, \beta) = (12, 2)$.

Solution:

1. For prior predictive, we have

$$\begin{aligned} f(x) &= \int_0^\infty f(x \mid \theta)f(\theta)\mathrm{d}\theta \\ &= \int_0^\infty \frac{e^{-\theta}\theta^x}{x!}\mathbb{1}_{\{x=0,1,\ldots\}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\mathrm{d}\theta \\ &= \frac{\beta^\alpha}{x!\Gamma(\alpha)}\mathbb{1}_{\{x=0,1,\ldots\}}\int_0^\infty \theta^{\alpha+x-1}e^{-\theta(\beta+1)}\mathrm{d}\theta \\ &= \frac{\beta^\alpha}{x!\Gamma(\alpha)} \cdot \frac{\Gamma(a+x)}{(\beta+1)^{\alpha+x}}\mathbb{1}_{\{x=0,1,\ldots\}} \end{aligned}$$

Next, for the posterior

$$\begin{aligned} f(\theta \mid x_{1:n}) &\propto f(\theta)f(x_{1:n} \mid \theta) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\mathbb{1}_{\{\theta>0\}}\prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!}\mathbb{1}_{\{x_i=0,1,\ldots\}} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\mathbb{1}_{\{\theta>0\}}e^{-n\theta}\theta^{\sum_{i=1}^n x_i}\prod_{i=1}^n \frac{1}{x_i!}\mathbb{1}_{\{x_i=0,1,\ldots\}} \\ &\propto \theta^{\alpha+\sum_{i=1}^n x_i-1}e^{-\theta(\beta+n)}\mathbb{1}_{\{\theta>0\}}. \end{aligned}$$

Therefore, $\theta \mid x_{1:n}$ follows $Ga(\alpha_n)/\beta_n$, where $\alpha_n = \alpha + \sum_{i=1}^n x_i$ and $\beta_n = \beta + n$.

Finally for the posterior predictive, since x_{n+1} and $x_{1:n}$ are conditionally independent given θ, we have by similar calculation in prior predictive

$$\begin{aligned} f(x_{n+1} \mid x_{1:n}) &= \int_0^\infty f(\theta \mid x_{1:n})f(x_{n+1} \mid \theta)\mathrm{d}\theta \\ &= \frac{\beta_n^{\alpha_n}}{x_{n+1}!\Gamma(\alpha_n)} \cdot \frac{\Gamma(\alpha_n + x_{n+1})}{(\beta_n+1)^{\alpha_n+x_{n+1}}}\mathbb{1}_{\{x_{n+1}=0,1,\ldots\}} \end{aligned}$$

Notice that we can replace the parameters α, β in the prior predictive by α_n, β_n to quickly get the posterior predictive.

2. Note that the log of gamma distribution is not a standard distribution, thus the posterior mean of $\ln\theta$ does not have close form (this is common in Bayesian analysis). Luckily, by law of larger numbers, we can use simulation to approximate the value $\mathsf{E}(\ln\theta \mid x_{1:4})$ by sample average, with the following R code:

```r
##Data
x = c(6,7,5,4)

##Parameters
a = 12
b = 2

##Compute a_4 and b_4
a_4 = a + sum(x)
b_4 = b + 4

##Simulate log theta
set.seed(4010)
nRep = 2^20
theta = rgamma(nRep,shape = a_4, rate = b_4)
ln_theta = log(theta)

##Compute sample mean as an estimate of log theta given data
mean(ln_theta)
##the answer is about 1.72.
```

1.3 Commonly Used Models and Representation

Example 1.4. The following distributions will be encountered throughout the course.

Distribution	Representation	
$\mathrm{N}(\mu, \sigma^2)$	$x = \mu + \sigma z$	$z \sim \mathrm{N}(0,1)$
$\theta\mathrm{Exp}(1)$	$x = \theta z$	$z \sim \mathrm{Exp}(1)$
$\mathrm{Ga}(\alpha)/\beta$	$x = \frac{1}{\beta}\sum_{i=1}^{\alpha} z_i$	$z_i \overset{\text{IID}}{\sim} \mathrm{Ga}(1) = \mathrm{Exp}(1)$
$\beta/\mathrm{Ga}(\alpha)$	$x = 1/z$	$z \sim \mathrm{Ga}(\alpha)/\beta$
χ^2_k	$x = \sum_{i=1}^{k} z_i$	$z_i \overset{\text{IID}}{\sim} \chi^2_1$
$\mathrm{Po}(\theta)$	$x = \sum_{i=1}^{\theta} z_i$	$z_i \overset{\text{IID}}{\sim} \mathrm{Po}(1)$
$\mathrm{Bin}(m, \theta)$	$x = \sum_{i=1}^{m} z_i$	$z_i \overset{\text{IID}}{\sim} \mathrm{Bern}(\theta)$
$\mathrm{Beta}(\alpha, \beta)$	$x = \frac{z_\alpha}{z_\alpha + z_\beta}$	$z_j \overset{\text{IID}}{\sim} \mathrm{Ga}(j)$

Table 2: Representation of Common Distribution

The following examples show how representation would be a helpful tool.

Example 1.5. Let $X, Y \sim \text{Ga}(1)$. Find the distribution of $Z = \frac{X}{X+Y}$.

Solution: By representation,

$$Z \overset{d}{=} \frac{\text{Ga}(1)}{\text{Ga}(1) + \text{Ga}(1)} \overset{d}{=} \text{Beta}(1, 1).$$

That is $Z \sim \text{Beta}(1, 1)$.

Takeaway: Representation technique helps us to determine the distributions of random variables.

Example 1.6. (Optional*) Let $[X \mid Y] \sim \text{N}(Y, Y^2)$ and $Y \sim \text{Unif}\ (0, 1)$. Prove that $X/Y \perp\!\!\!\perp Y$.

Solution: We can jointly represent (X, Y) as

$$\begin{cases} X = Y + YZ \\ Y = U \end{cases} \quad \Rightarrow \quad \begin{cases} X/Y = 1 + Z \\ Y = U \end{cases}$$

where $Z \sim \text{N}(0, 1)$ and $U \sim \text{Unif}(0, 1)$ are independent. Note that

- X/Y depends only on Z and Y depends only on U, and
- Z and U are independent.

We conclude that X/Y and Y are independent.

*This example is from STAT4003 Lecture Note 1 (Keith 2020).

1.4 R Tips

Example 1.7. There are several methods/tricks that will be helpful in this course.

1. Four density-related functions in `R`. Take the normal distribution for an example.

   (a) The function `rnorm(nRep,mean,sd)` samples from N(`mean,sd`) for `nRep` times.

   (b) The function `qnorm(prob,mean,sd)` returns the 100%×`prob`-th quantile of N(`mean,sd`).

   (c) The function `pnorm(q,mean,sd)` returns the value of F(`q`) where F is the CDF of N(`mean,sd`).

   (d) The function `dnorm(q,mean,sd)` returns the value of f(`q`) where f is the PDF of N(`mean,sd`).

   The functions for other distributions are similarly used by changing `norm` to `binom,beta,exp,gamma,t,chisq`,etc., and specifying the corresponding parameters as arguments.

2. Usually, we want to find the expectation or variance of a complicated variables. Analytical results may be difficult to derive, instead, we can use Monte Carlo method to

find them based on a large number of samples from that distribution.

3. Remember to use `set.seed(4010)` to generate reproducible samples. You can change `4010` to other numbers. It will help you to debug your codes and to justify your results.

4. To debug, comment out your codes section by section and run again. If the bug disappears, then the bugs are from the section you just commented out.

5. If you want to achieve something in `R` but have no clue at all, you can always find references by Googling it. The best way to succeed in `R` programming is to practice.

Experiment:

```
#sampling standard normal
set.seed(4010)
x = rnorm(10^4,0,1) #x is a sample of 10^4 IID standard normal random
    variables
z = qnorm(0.975,0,1) # z = 1.959964
p = pnorm(1.96,0,1) # p = 0.9750021
d = dnorm(0,0,1) # d = 0.3989423

# using Monte Carlo to find quantities related to a certain random
    variable
expectation = mean(x)  # expectation = -0.0100194
variance = var(x) # variance = 1.006437
quart = quantile(x,0.25) # quart = -0.6778637
prob = mean(x<1.96) # prob = 0.975
```

2 Remarks on Assignment 1

Please check the platform first if you have questions.

- Remember to write indicators in the density.
- Make use of the fact that $\int f(x)\mathrm{d}x = 1$ if $f(\cdot)$ is a probability density function.