STAT 4010 – Bayesian Learning

Tutorial 2

Spring 2022

Cheuk Hin (Andy) CHENG (Email | Homepage)

Di SU (Email | Homepage)

1 More Examples of Representation

Example 1.1. (Q3 Assignment 1 Spring 2021) Let $n, m \in \mathbb{N}$ be known constants. Assume

$$\begin{aligned}[x_1, \ldots, x_{n+m} \mid \theta] &\overset{\text{IID}}{\sim} |\mathrm{N}(0, 2\theta)|; \\ \theta &\sim 4/\operatorname{Exp}(1).\end{aligned}$$

where $|\mathrm{N}(0, 2\theta)|$ and $4/\operatorname{Exp}(1)$ denote half-normal and inverse-exponential distributions.

1. (10%) Find the posterior distribution $[\theta \mid x_{1:n}]$.

   Solution:

   The posterior is

   $$\begin{aligned} f(\theta \mid x_{1:n}) &\propto f(\theta) f(x_{1:n} \mid \theta) \\ &\propto \frac{4}{\theta^2} \exp\left(-\frac{4}{\theta}\right) \prod_{i=1}^{n} \left(\frac{1}{(2\theta)^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2(2\theta)} x_i^2 \right\} \right) \\ &\propto \frac{1}{\theta^2} \exp\left(-\frac{4}{\theta}\right) \frac{1}{\theta^{\frac{n}{2}}} \exp\left\{ -\frac{1}{4\theta} \sum_{i=1}^{n} x_i^2 \right\} \\ &\propto \theta^{-\left(\frac{n}{2}+1\right)-1} \exp\left\{ -\frac{1}{\theta} \left(\frac{1}{4} \sum_{i=1}^{n} x_i^2 + 4 \right) \right\} \end{aligned}$$

   which follows $\beta_n / \operatorname{Ga}(\alpha_n)$ (inverse gamma distribution) where $\alpha_n = n/2 + 1$ and $\beta_n = 4 + \sum_{i=1}^{n} x_i^2/4$.

2. (10%) Let

   $$\widehat{\sigma}^2_{1:n} := \frac{1}{n} \sum_{i=1}^{n} x_i^2 \quad \text{and} \quad \widehat{\sigma}^2_{n+1:n+m} := \frac{1}{m} \sum_{i=n+1}^{n+m} x_i^2$$

   Find the posterior predictive distribution $\left[\widehat{\sigma}^2_{n+1:n+m} \mid x_{1:n}\right]$.

   Solution:

   For some random variable X, it holds that $|X| = \sqrt{X^2}$. In particular, for $Z \sim \mathrm{N}(0,1)$, $Z^2 \sim \chi^2_1$, and thus $|Z| \sim \chi_1$. Also recall $\operatorname{Ga}(\alpha_n) = \chi^2_{2\alpha_n}/2$.
   Let $V_{n+1}, \ldots, V_{n+m} \sim \chi^2_1$ and $U_n \sim \chi^2_{2\alpha_n}$ be mutually independent. Using representation, we have for $i = n+1, \ldots, n+m$,

   $$\theta \mid x_{1:n} = \frac{2\beta_n}{U_n} \quad \text{and} \quad x_i = \sqrt{2\theta V_i}.$$

Therefore, $x_i \mid x_{1:n} = \sqrt{4\beta_n V_i/U_n}$ for $i = n+1, \ldots, n+m$. We also have, conditioned on $x_{1:n}$,

$$\begin{aligned}\frac{1}{m}\sum_{i=n+1}^{n+m} x_i^2 &= \frac{4\beta_n}{m}\frac{\sum_{i=n+1}^{n+m} V_i}{U_n} \\ &= \frac{2\beta_n}{\alpha_n}\frac{\sum_{i=n+1}^{n+m} V_i/m}{U_n/(2\alpha_n)}. \\ &= \frac{2\beta_n}{\alpha_n}\,\mathrm{F}_{m,2\alpha_n},\end{aligned}$$

Therefore, $\left[\widehat{\sigma}^2_{n+1:n+m} \mid x_{1:n}\right] \sim 2\beta_n \mathrm{F}(m, 2\alpha_n)/\alpha_n$.

3. (10%) Derive

$$p_m = \mathrm{P}\left(\widehat{\sigma}^2_{n+1:n+m} > 1 + \widehat{\sigma}^2_{1:n} \mid x_{1:n}\right), \quad m \in \mathbb{N}$$

Write a R-function `postPr = function (m=1, x)` to compute p_m, where the parameters `m` and `x` denote m and $x_{1:n}$, respectively. Suppose $x_{1:5} = (1.92, 0.59, 0.52, 2.30, 0.39)^{\mathrm{T}}$. Plot $p_1, \ldots, p_{100}$. Interpret. (Use no more than 30 words.)

Solution:

Denote $n^{-1}\sum_{i=1}^n x_i^2$ by $\overline{x^2_{1:n}}$. Let $F \sim \mathrm{F}(m, 2\alpha_n)$. Note that for $m \in \mathbb{N}$,

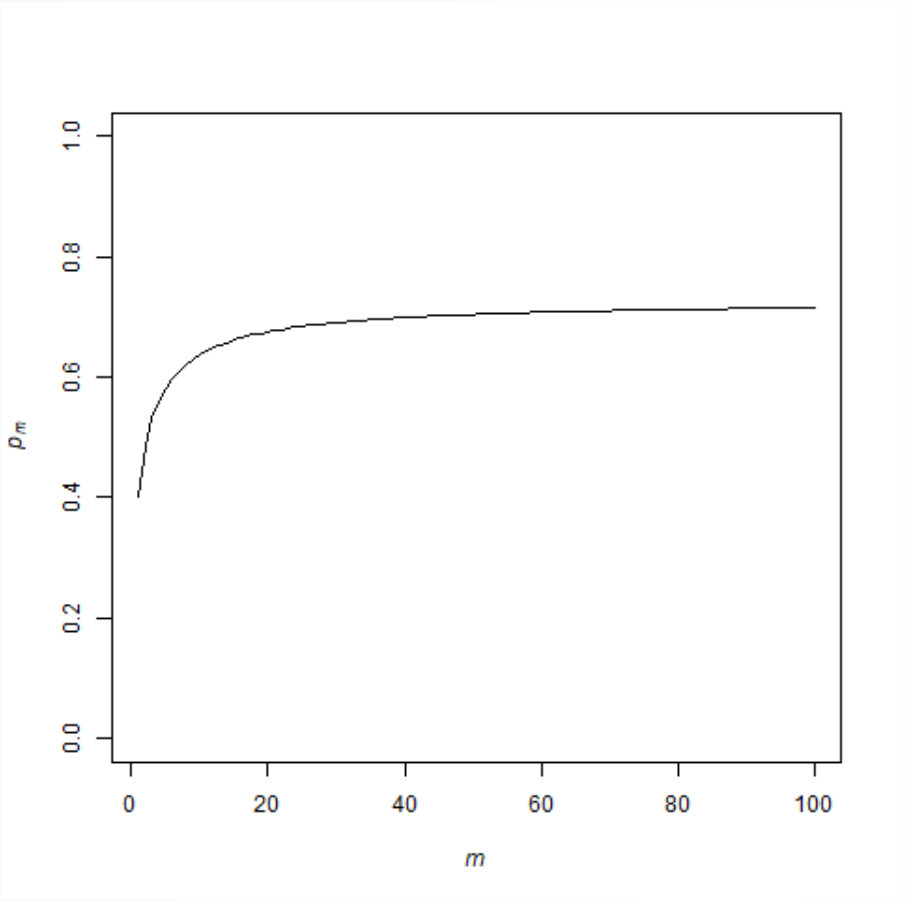
$$\begin{aligned}p_m &= \mathrm{P}\left(\frac{1}{m}\sum_{i=n+1}^{n+m} x_i^2 > 1 + \frac{1}{n}\sum_{i=1}^{n} x_i^2 \mid x_{1:n}\right) \\ &= \mathrm{P}\left(\mathrm{F} > \frac{\alpha_n\left(1+\overline{x^2_{1:n}}\right)}{2\beta_n} \mid x_{1:n}\right) \\ &= 1 - \mathtt{pf}\left(\frac{\alpha_n\left(1+\overline{x^2_{1:n}}\right)}{2\beta_n}, m, 2\alpha_n\right).\end{aligned}$$

We can compute p_m with the following R code:

```r
#data
x = c(1.92, 0.59, 0.52, 2.30, 0.39)

#the main function
pdf.post = function(m = 1,x){
    n = length(x)
    alpha.n = n/2 +1
    beta.n = sum(x^2)/4+4
    p = 1-pf(alpha.n*( mean(x^2)+1 )/2/beta.n, m, 2*alpha.n)
    p
}

#application to data
pm = sapply(1:100, pdf.post, x, simplify=TRUE)
plot = 0
if(plot){
    png('./tut2.png') #save the plot
    plot(1:100, pm, type="l", ylim=c(0,1),
     ylab=bquote(italic(p[m])), xlab=bquote(italic(m)))
    dev.off()
}
```



The graph shows that

- With the sample size of further observations increasing, prediction usually involves larger and larger variability than the variability in observed data
- In the long run, the convergence to a value lower than 1 means exhaustion of knowledge from existing data.

Any other reasonable answers will be accepted.

Remark 1.1. The half normal distribution is also called the Chi distribution (intuitively it is the square root of a Chi-square random variable).

Distribution	Notation	Representation	PDF
Half normal	$x \sim \lvert \mathrm{N}(0, \sigma^2) \rvert$	$x = \sigma \lvert z \rvert, z \sim \mathrm{N}(0,1)$	$f(x) = \frac{\sqrt{2}}{\sqrt{\sigma^2 \pi}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\} \mathbb{1}(x > 0)$

The next example shows how representation can be used to find expectation and simplify simulation.

Example 1.2. Assume $X \sim \mathrm{Ga}(\alpha)/\beta$. Suppose you are new to `R` and can only use the commands `runif`, `mean`. Use simulation to find $\mathsf{E}(X)$ with $(\alpha, \beta) = (4, 2)$ and $(\alpha, \beta) = (5, 10)$.

Solution:

Firstly, we can represent a Gamma-distributed random variable in terms of a function of

uniformly distributed random variables. Secondly, if $X \sim \text{Exp}(1)$, then its c.d.f. is $F(x) = 1 - e^{-x}$ which has a close form. Thirdly, $F(X) \sim \text{Unif}(0,1)$. This is because, the c.d.f of $F(X)$ is the same as of a random variable having distribution Unif(0,1), precisely,

$$\Pr(F(X) \leq z) = \Pr(X \leq F^{-1}(z)) = F(F^{-1}(z)) = z, z \in [0,1].$$

Note also $F^{-1}(y) = -ln(1-y)$. Therefore, letting $U \overset{\text{IID}}{\sim} \text{Unif}(0,1)$, $V := 1 - U \sim \text{Unif}(0,1)$, we can represent

$$X = F^{-1}(U) = -ln(1-U) = -ln(V).$$

Recall from the representation of Gamma distribution and the fact that $\text{Ga}(1) = \text{Exp}(1)$. Therefore,

$$\text{Ga}(\alpha)/\beta = \frac{1}{\beta}\sum_{i=1}^{\alpha} Exp_i(1) = \frac{-1}{\beta}\sum_{i=1}^{\alpha} ln(V_i), \tag{1.1}$$

where $V_{1:\alpha} \overset{\text{IID}}{\sim} \text{Unif}(0,1)$. In R, we can generate a Gamma distribution with (α, β) and simulate the mean of gamma distribution by following Monte Carlo simulation:

Step 1: Generate α IID copies of uniform(0,1) random samples;

Step 2: Obtain a sample of Gamma random variable by (1.1);

Step 3: Repeat step 1 and 2 for n times, so we have n samples from the Gamma distribution;

Step 4: Compute the sample mean of the random gamma samples as an estimator of the population.

```r
##Write step 1 and 2 as a function
sim_gamma <- function(alpha=1,beta=1){
  v = runif(alpha) ##generate alpha uniform(0,1) random samples
  g = -1/beta*sum(log(v))
  return(g)
}

##monte carlo simulation to estimate the population mean
n = 100000
sim = array(NA,n) ##An array to store sample
a=4
b=2
set.seed(4010)
for (i in 1:n) {
  sim[i] = sim_gamma(a,b)
}

##Sample mean vs true mean a/b = 2
mean(sim) ##2.005
```

Remark 1.2. In representation "=" means equal in distribution.

Example 1.3. (Sample variance is independent of location parameter) Consider the following

model,

$$\theta \sim F(\cdot)$$
$$x_{1:n} \mid \theta \overset{\text{IID}}{\sim} N(\theta, 1),$$

where $F(\cdot)$ is some distribution. In this case, θ is the location parameter. Let $z_{1:n} \overset{\text{IID}}{\sim} N(0,1)$ and clearly $z_{1:n}$ are independent to θ. Using representation, we have

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - n^{-1}\sum_{j=1}^{n} x_j)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(\theta + z_i - n^{-1}\sum_{j=1}^{n}\{\theta + z_j\})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(z_i - n^{-1}\sum_{j=1}^{n} z_j)^2. \end{aligned}$$

Therefore, the sample variance depends on $z_{1:n}$ and is free of θ. Thus S^2 is independent of θ.

Remark 1.3. We call a statistic (a function of data), ancillary statistic if its distribution is free of the unknown parameter. Intuitively, ancillary statistic contains no information about the parameter.

2 Invariant Priors

In Bayesian framework, the inference is "subjective" in the sense of prior. One remedy is to adopt non-informative prior. Usually, invariance is used as the guiding principle of designing non-informative prior.

Definition 1. *(Location and Scale Families)*

- *Location family: The family of distributions $\{f(\cdot \mid \theta) : \theta \in \Theta\}$ is said to be a location family with a location parameter $\theta \in \Theta = \mathbb{R}^k$ if there is a f_0 such that*
$$f(x \mid \theta) = f_0(x - \theta). \tag{2.1}$$
- *Scale family:The family of distributions $\{f(\cdot \mid \theta) : \theta \in \Theta\}$ is said to be a (one-dimensional) scale family with a scale parameter $\theta \in \Theta = \mathbb{R}^+$if there is a f_0 such that*
$$f(x \mid \theta) = f_0(x/\theta)/\theta. \tag{2.2}$$

Definition 2. *Let $g : \Theta \to \Phi$ be a bijective function, and $\phi = g(\theta)$. A principle is said to be invariant under reparametrization g if it induces priors $\theta \sim f_\theta(\theta)$ and $\phi \sim f_\phi(\phi)$ such that*

$$\int_A f_\theta(\theta)\mathrm{d}\theta = \int_{g(A)} f_\phi(\phi)\mathrm{d}\phi, \quad \text{for all } A,$$

where $g(A) = \{g(\theta) : \theta \in A\}$. In particular, if g is differentiable, then

$$f_\theta(\theta) = \left|\frac{\mathrm{d}\phi}{\mathrm{d}\theta}\right| f_\phi(\phi).$$

Theorem 2.1. *(Invariant Prior of Location and Scale Families)*
Consider the family of distributions $F = \{f(\cdot \mid \theta) : \theta \in \Theta\}$.

1. *(Location parameter) If* $f(x \mid \theta)$ *is given by (2.1), and* $0 \in \Theta$, *the invariant prior of* θ *is* $f(\theta) \propto 1$.
2. *(Scale parameter) If* $f(x \mid \theta)$ *is given by (2.2), and* $1 \in \Theta$, *the invariant prior of* θ *is* $f(\theta) \propto 1/\theta$.

Remark 2.1.

- If X has a density from a location family, we can represent $X = \theta + Z$ where Z has density f_0. If instead X is from a scale family, we can represent it by $X = \theta Z$.
- Let $c \in \mathbb{R}$. If θ is the parameter of a location family, then $\psi := g(\theta) = \theta + c$ is still a parameter of a location family. Hence the *invariant* prior *for this family* should be *invariant* to the transformation $g(\theta) = \theta + c$. Similarly, for scale family, we want the prior to be invariant to the transformation $g(\theta) = c\theta$.
- The density f_0 is called the standard density.

Example 2.1. (Location and Scale Parameters of Common Distributions)

Distribution (X)	f_0 (Z)	Representation	Location Parameter	Scale Parameter
$\mathrm{N}(\mu, \sigma^2)$	$\mathrm{N}(0,1)$	$\mu + \sigma z$	μ	σ
$\theta\mathrm{Exp}(1)$	$\mathrm{Exp}(1)$	θz	NA	θ
$\mathrm{Ga}(\alpha)/\beta$	$\mathrm{Ga}(\alpha)$	$\frac{1}{\beta} z$	NA	$\frac{1}{\beta}$
$\beta/\mathrm{Ga}(\alpha)$	$1/\mathrm{Ga}(\alpha)$	βz	NA	β
$\mathrm{Unif}(a,b)$	$\mathrm{Unif}(0,1)$	$a + (b-a)z$	a	$b-a$

Table 1: Location and Scale Parameters of Common Distributions

Example 2.2. (Both a location and scale parameter) Consider the model $\{f(\cdot \mid \theta) : \theta \in \Theta\}$ with $\Theta = \mathbb{R}^+$, where θ is both the location and scale parameter, i.e.,

$$f(x \mid \theta) = \frac{1}{\theta} f_0((x-\theta)/\theta),$$

for some f_0. Find an invariant prior for θ.

Solution:

There are many methods. Two of them are illustrated here.

- **Method 1: direct calculation**. Directly copying the proof of Theorem 2.1 (2) with some trivial modification, one can show that an invariant prior for θ is $f(\theta) \propto 1/\theta$.
- **Method 2: representation**. If $x \sim f(\cdot \mid \theta)$, then we can represent x as

$$x = \theta + \theta z,$$

where $z \sim f_0(\cdot)$. Let $\widetilde{z} = z + 1$. We have

$$x = \theta\widetilde{z},$$

which is simply a scale model. By Theorem 2.1, an invariant prior for θ is $f(\theta) \propto 1/\theta$.

Example 2.3. You have recorded the waiting time (in minute) of the N bus at one particular stop, the data are 18, 17, 21, 23, 20, 21. You want to infer what is the average waiting time, and design the following model:

$$\theta \sim F(\cdot), \tag{2.3}$$

$$x_{1:n} \mid \theta \overset{\text{IID}}{\sim} \theta\text{Exp}(1), \tag{2.4}$$

where $F(\cdot)$ is some distribution.

1. Since you do not have any idea about the waiting time, propose an invariant prior, compute the posterior, find the posterior mean and posterior probability that θ is less than 20.
   Solution:
   Note that θ is a scale parameter. By Theorem 2.1, an invariant prior is

   $$f(\theta) \propto \frac{1}{\theta}.$$

   The prior is improper clearly. The posterior is

   $$\begin{aligned} f(\theta \mid x_{1:n}) &\propto f(\theta) f(x_{1:n} \mid \theta) \\ &\propto \frac{1}{\theta} \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{\{\theta>0\}} e^{-x_i/\theta} \mathbb{1}_{\{x_i \in (0,\infty)\}} \\ &\propto \theta^{-n-1} e^{-\sum_{i=1}^{n} x_i/\theta} \mathbb{1}_{\{\theta>0\}}. \end{aligned}$$

   Therefore, $\theta \mid x_{1:n} \sim \beta_n/\text{Ga}(\alpha_n)$, where $\alpha_n = n$ and $\beta_n = \sum_{i=1}^{n} x_i$. The posterior mean is

   $$\frac{\beta_n}{\alpha_n - 1} = \frac{120}{5} = 24.$$

   You may use the R package "`invgamma`" to compute the posterior probability, and $\Pr(\theta < 20 \mid x_{1:n}) = 0.4457$.

2. The university claims that the average waiting time of N bus will be controlled below 20 minutes. Do your data and model support this claim?
   Solution:
   Based on our data, the posterior mean is larger than 20, and the posterior probability of waiting for less than 20 minutes is less than 0.5. Hence our belief on the claim is not strong.

Example 2.4. Consider you and five of your friends have done an IQ test, the scores are 110,108,128,95,117,98. You want to infer what is the maximum of human's IQ, and design the following model:

$$\theta \sim F(\cdot) \tag{2.5}$$

$$x_{1:n} \mid \theta \overset{\text{IID}}{\sim} \text{Unif}(0, \theta), \tag{2.6}$$

where $F(\cdot)$ is some distribution. Since you do not have any idea about the limit, propose an invariant prior, compute the posterior, find the posterior mean and variance. Is your prior and posterior proper?

Solution:

Note that θ is a scale parameter. By Theorem 2.1, an invariant prior for θ is

$$f(\theta) \propto \frac{1}{\theta}.$$

The prior is improper clearly. The posterior is

$$\begin{aligned} f(\theta \mid x_{1:n}) &\propto f(\theta) f(x_{1:n} \mid \theta) \\ &\propto \frac{1}{\theta} \mathbb{1}_{\{\theta > 0\}} \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{\{x_i \in (0,\theta)\}} \\ &\propto \frac{1}{\theta^{n+1}} \mathbb{1}_{\{\theta > \max_i x_i\}}. \end{aligned}$$

Let $x_{(n)} = \max_{i=1,2,\ldots,n} x_i$, we can find the proportionality constant by

$$\frac{1}{c} = \int_{x_{(n)}}^{\infty} \frac{1}{\theta^{n+1}} \mathrm{d}\theta = \frac{1}{n x_{(n)}^n} \Rightarrow c = n x_{(n)}^n.$$

Therefore, the posterior is proper. We can compute the posterior mean and variance,

$$\begin{aligned} \mathsf{E}(\theta \mid x_{1:n}) &= \int_{x_{(n)}}^{\infty} \theta \frac{c}{\theta^{n+1}} \mathrm{d}\theta = c \int_{x_{(n)}}^{\infty} \frac{1}{\theta^n} \mathrm{d}\theta = x_{(n)} \frac{n}{n-1} = 153.6, \\ \mathsf{E}(\theta^2 \mid x_{1:n}) &= \int_{x_{(n)}}^{\infty} \theta^2 \frac{c}{\theta^{n+1}} \mathrm{d}\theta = x_{(n)}^2 \frac{n}{n-2}, \\ \mathsf{Var}(\theta^2 \mid x_{1:n}) &= \mathsf{E}(\theta^2 \mid x_{1:n}) - \mathsf{E}(\theta \mid x_{1:n})^2 = x_{(n)}^2 \frac{n}{(n-1)^2(n-2)} = 983.04. \end{aligned}$$

Note that, the posterior mean and the posterior variance do not exist when we have only one datum.

Takeaway: Improper prior can still lead to a proper posterior. Non-informative prior doesn't contain much prior knowledge, so with a small sample size, the posterior variance is still large.