Application of Monte Carlo Methods: Bayesian inference

Bayes Theorem

- Bayesian inference differs from classical inference in that in Bayesian inference, model parameters are viewed as random variables.
- Recall that probability distributions can be used to express either available information or believes on random or unknown phenomenon.

Bayes Theorem

- In Bayesian inference, unknown parameters are given a distribution (known as prior distribution) that expresses prior information available on them.
- We then combine together the prior distribution, the model and the data to obtain the posterior distribution on which inference is based. We use the Bayes theorem to go from the prior to the posterior distribution.

Bayes Theorem

- Mathematically, the Bayes theorem follows from the definition of conditional probability. Given two events A, B,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A).$$

  Hence

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}. \tag{1}$$

- (1) is known as Bayes Theorem.

Bayes Theorem

Here is a typical application of Bayes theorem:

- We randomly select an individual from a population and consider the following events.
  E: the event "the selected person is a college graduate",
  H_i: "the selected person is in the i-th income quartile",
  $i = 1, \ldots, 4$.
- A survey reveals that $\mathbb{P}(E|H_1) = 0.11$, $\mathbb{P}(E|H_2) = 0.19$, $\mathbb{P}(E|H_3) = 0.31$ and $\mathbb{P}(E|H_4) = 0.53$. What is the probability that a randomly selected individual is in the i-th income quartile if it is observed that she is college educated?

Bayes Theorem

- $$\mathbb{P}(H_i|E) = \frac{\mathbb{P}(E|H_i)\mathbb{P}(H_i)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E|H_i)\mathbb{P}(H_i)}{\sum_{j=1}^{4}\mathbb{P}(E|H_j)P(H_j)}. \qquad (2)$$

  $\mathbb{P}(H_1|E) = 0.09$, $\mathbb{P}(H_4|E) = 0.47$.
- Note how we have used the law of total probability to express $\mathbb{P}(E)$ as

  $$\mathbb{P}(E) = \sum_{j=1}^{4}\mathbb{P}(E|H_j)P(H_j) = \frac{1}{4}(0.11+0.19+0.31+0.53) = 0.285.$$

Bayes Theorem

- Without any additional information, the probability of being in i-th quartile is $1/4$.
- However once we get more information (that is, we know that the selectec person is college educated) we revise our probability using Bayes theorem.

Bayes Theorem

- The same idea extends to random variables.
- If (X, Y) have joint pmf $p_{X,Y}$, then

$$p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y).$$

  So that

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_x p_{Y|X}(y|x)p_X(x)}.$$

- If (X, Y) have joint pdf $f_{X,Y}$, then

$$f_{(X,Y}(x,y) = f_X(x)f_{Y|X}(y|x) = f_Y(y)f_{X|Y}(x|y).$$

  Hence

$$f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(x|y)}{f_Y(y)} = \frac{f_X(x)f_{Y|X}(y|x)}{\int f_X(x)f_{Y|X}(y|x)\mathrm{d}x}.$$

Bayes Theorem

An example:

- Suppose $f_{X|\Lambda}(x|\lambda) = \lambda e^{-\lambda x}$ if $x \geq 0$, (zero otherwise).
- Suppose $f_\Lambda(\lambda) = e^{-\lambda}$, if $\lambda \geq 0$, (and zero otherwise).
- Find $f_{\Lambda|X}$.

$$f_{\Lambda|X}(\lambda|x) = \frac{f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda)}{f_\Lambda(\lambda)} = \frac{\lambda e^{-\lambda(x+1)}}{f_X(x)}$$

- $$f_X(x) = \int_0^\infty f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda) \mathrm{d}\lambda = \frac{1}{(1+x)^2}.$$
- Hence

$$f_{\Lambda|X}(\lambda|x) = \begin{cases} (1+x)^2 \lambda e^{-\lambda(x+1)} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

which is the density of the Gamma distribution $\mathrm{Ga}(2, x+1)$.

Bayes Theorem

- In the example above, the calculation of $f_X(x)$ is unnecessary, because $f_X(x)$ does not depend on λ. We express this by writing

$$f_{\Lambda|X}(\lambda|x) \propto f_{X|\Lambda}(x|\lambda) f_X(x) \propto \lambda e^{-\lambda(x+1)} \quad \text{for} \quad x \geq 0.$$

- Then we "recognize" the form $\lambda e^{-\lambda(x+1)}$ and we conclude that

$$\Lambda|\{X = x\} \sim \mathrm{Ga}(2, x+1).$$

More on the proportionality sign

- If a density has the form

$$f(x) = \frac{h(x)}{C}, \quad \text{or} \quad f(x) = Ch(x),$$

  for some constant C that does not depend on x, we express this by writing

$$f(x) \propto h(x).$$

- The constant C does not matter. We can deduce it from the functional form, since the integral of the density is 1.
- For instance, if a joint density $f_{X,Y}(x,y) = f_1(x)f_2(x,y)$. Then the conditional density of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f_1(x)f_2(x,y)}{\int f_1(x)f_2(x,y)\mathrm{d}y} \propto f_2(x,y).$$

- $f_1(x)$ and the denominator are constants in y.

Bayesian inference

- Bayesian inference is an application of Bayes theorem to statistical problems.
- We view the parameter Θ and the data generating mechanism Y as random variables.
- We formulate a prior distribution $\pi_\Theta(\theta)$ for the parameter.
- We model the data generating mechanism given $\Theta = \theta$ by its conditional distribution $f_{Y|\Theta}(y|\theta)$.
- Hence the joint distribution of the parameter and the data generating process (Θ, Y) has density

$$\pi_{(\Theta,Y)}(\theta, y) = \pi_\Theta(\theta) f_{Y|\theta}(y|\theta) = f_Y(y)\pi_{\Theta|Y}(\theta|y).$$

Bayesian inference

- After observing the data $Y = y$, using Bayes theorem we obtain the posterior distribution

$$\pi_{\Theta|Y}(\theta|y) = \frac{\pi_\Theta(\theta) f_{Y|\Theta}(y|\theta)}{\int \pi_\Theta(\theta) f_{Y|\Theta}(y|\theta) \mathrm{d}\theta} \propto \pi_\Theta(\theta) f_{Y|\Theta}(y|\theta).$$

- The posterior distribution is our inference about Θ given our prior belief π_Θ, and the observed data y.
- We summarize the posterior graphically, or by computing various statistics: mean, mode, median, variance, quantiles.
- We qualify these statistics as "posterior", e.g. posterior mean...

Bayesian inference

- Most quantities of interest in Bayesian data analysis can be written as integrals. For instance the posterior mean

$$\int \theta \pi_{\Theta|Y}(\theta|y) d\theta$$

  is a commonly used summary statistic of the posterior distribution.
- Most of the time these integrals cannot be easily evaluated. Monte Carlo methods are very useful to help approximate these integrals.

Bayesian inference

- Example: we wish to know the proportion of people older than 70 that are overall happy with the life they lived.
- Suppose we ask a group of $n = 82$ from that pop. and $y = 45$ responded 'Yes'.
- In the Bayesian framework, that proportion is a random variable $\Theta \in (0, 1)$. And we need to build a model for both the parameter (the actually proportion), and the data Y.
- For a given value θ of Θ, we assume that the response is generated from

$$Y \sim \mathbf{Bin}(n, \theta),$$

with pmf $f_{Y|\Theta}(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$.

Bayesian inference

Digression: The **Beta** distribution.

- The Beta distribution **Beta**(α, β) for $\alpha > 0$, $\beta > 0$ is a distribution on $(0, 1)$ that is typically used to model proportions and other quantities with values between $(0, 1)$. Its pdf is

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},\ x \in (0, 1).$$

- The mean is $\alpha/(\alpha+\beta)$. Hence if $\alpha = \beta$, the mean is $1/2$ and the distribution is symmetric about 0.5.

Bayesian inference

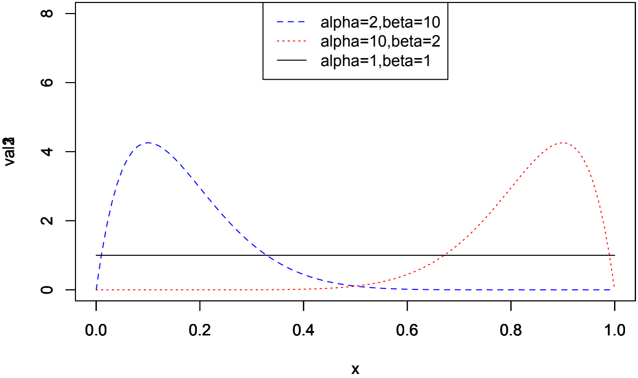

Figure 1: Some pdf shapes of the beta distribution.

Bayesian inference

- So we could also use a prior **Beta**(α, β) for our parameter Θ.
- Recall that the posterior distribution is

$$\pi_{\Theta|Y}(\theta|y) \propto \pi_\Theta(\theta) p_{Y|\Theta}(y|\theta) \propto \pi_\Theta(\theta)\theta^y(1-\theta)^{n-y}.$$

- If we take the prior as a **Beta**(α, β) distribution then:

$$\begin{aligned}\pi_{\Theta|Y}(\theta|y) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^y(1-\theta)^{n-y} \\ &= \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1},\end{aligned}$$

  which is **Beta**$(\alpha + y, \beta + n - y)$.
- In this example the posterior distribution happens to be a familiar distribution. So no Monte Carlo or any other numerical approximation method is needed.

Bayesian inference

- The posterior mean is
$$\frac{y+\alpha}{\alpha+\beta+n} = \frac{y}{n}\frac{n}{\alpha+\beta+n} + \frac{\alpha}{\alpha+\beta}\left(1 - \frac{n}{\alpha+\beta+n}\right).$$
- This formula says that the posterior mean is a weighted average of the prior mean $\alpha/(\alpha+\beta)$ and the maximum likelihood (mle) estimate y/n.
- As $n \to \infty$, the posterior mean moves towards the mle y/n.
- Hence, as we get more and more data, our prior view matters less and less.

Bayesian inference

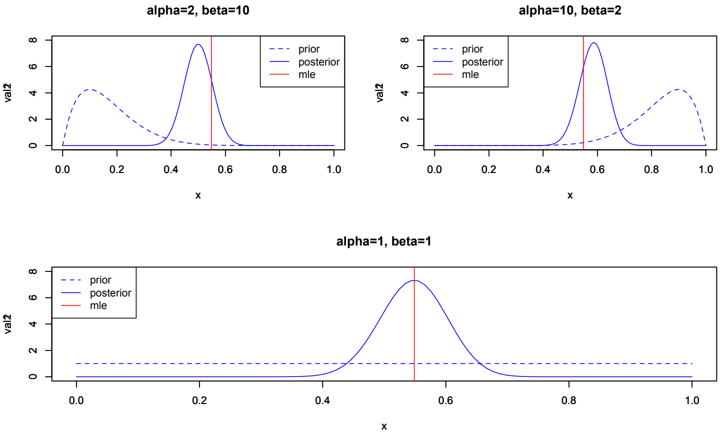

Figure 2: Prior/Posterior plots for the binomial example.

Bayesian inference

- Another example. Suppose we wish to estimate the mean of a random variable X.
- Suppose that we observe $X_1 = x_1, \ldots, X_n = x_n$, where, given $\Theta = \theta$,

$$X_i \overset{i.i.d.}{\sim} \mathbf{N}(\theta, \sigma^2),$$

  with σ^2 known.
- We formulate a prior distribution for Θ. We assume that

$$\Theta \sim \mathbf{N}(0, \tau_0^2),$$

  with $\tau_0 > 0$ known.

Bayesian inference

- Hence the prior is

$$\pi_\Theta(\theta) = \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left(-\frac{\theta^2}{2\tau_0^2}\right).$$

- Writing $X_{1:n} = (X_1, \ldots, X_n)$, the likelihood is

$$f_{X_{1:n}|\Theta}(x_1, \ldots, x_n|\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right).$$

- Then the posterior distribution is

$$\pi_{\Theta|X_{1:n}}(\theta|x_{1:n}) \propto \pi_\Theta(\theta) f_{X_{1:n}|\Theta}(x_1, \ldots, x_n|\theta).$$

Bayesian inference

- We have

$$\pi_{\Theta|X_{1:n}}(\theta|x_{1:n}) \propto \exp\left(-\frac{\theta^2}{2\tau_0^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right).$$

- This posterior can be rearranged to show that

$$\Theta|\{X_{1:n} = x_{1:n}\} \sim \mathbf{N}(\mu_n, \tau_n^2),$$

where, with $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$,

$$\mu_n = \frac{\frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}, \quad \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}.$$

Bayesian inference

- In this example, we see also that the posterior distrbution is a well known distribution.
- We have
$$\mu_n = \frac{\frac{1}{\tau_0^2} 0 + \frac{n}{\sigma^2} \bar{x}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \quad \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}.$$
- The posterior mean μ_n is a weighted average of the prior mean 0 and the data average.
- As we get more and more data ($n \to \infty$), the weight of the data average dominates, and the posterior behaves like $\mathbf{N}(\bar{x}, \sigma^2/n)$.

Bayesian inference

- Another example.
- Suppose we want to estimate the correlation ρ between two variables from measurements $(x_1, y_1), \ldots, (x_N, y_N)$.
- We assume that these measurements are realizations of iid bivariate normal $D = (X_1, Y_1), \ldots, (X_N, Y_N)$, each with distribution

$$\mathbf{N}\left[0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right],$$

  given a correlation parameter ρ.

Bayesian inference

- The conditional distribution of D given ρ is given by

$$
\begin{aligned}
f_{D|\rho}&((X_1, Y_1), \ldots, (X_N, Y_n)|\rho) \\
&= \prod_{j=1}^{N} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x_j^2 + y_j^2 - 2\rho x_j y_j}{2(1-\rho^2)}\right) \\
= &\left(\frac{1}{4\pi^2(1-\rho^2)}\right)^{N/2} \exp\left(-\frac{1}{2(1-\rho^2)}(S_x^2 + S_y^2 - 2\rho S_{xy})\right),
\end{aligned}
$$

where $S_x^2 = \sum_{i=1}^N x_i^2$, $S_y^2 = \sum_{i=1}^N y_i^2$, and $S_{xy} = \sum_{i=1}^N x_i y_i$.

Bayesian inference

- The correlation parameter ρ lies between -1 and 1.
- We assume a uniform prior distribution $\rho \sim \mathcal{U}(-1, 1)$, with density $f_\rho(\rho) = 1/2$ for $\rho \in (-1, 1)$ and $f_\rho(\rho) = 0$ otherwise.
- This leads to a posterior distribution on the interval $(-1, 1)$

$$\pi(\rho|D) = \frac{f_\rho(\rho) f_{D|\rho}((X_1, Y_1), \ldots, (X_N, Y_n)|\rho)}{\int_{-1}^{1} f_\rho(\rho) f_{D|\rho}((X_1, Y_1), \ldots, (X_N, Y_n)|\rho)\, d\rho}$$

$$\propto \left(\frac{1}{1-\rho^2}\right)^{N/2} \exp\left(-\frac{1}{2(1-\rho^2)}(S_x^2 + S_y^2 - 2\rho S_{xy})\right),$$

and $\pi(\rho|D) = 0$ if $\rho \notin (-1, 1)$.

- We want the posterior mean of ρ and a posterior interval for ρ.

Bayesian inference

- This posterior distribution is not a density that we recognize and can easily deal with.
- For this example we need numerical methods or Monte Carlo methods to approximate the calculation of the posterior mean.
- We can approximate the integral $\int_{-1}^{1} \rho\pi(\rho|D)d\rho$ by importance sampling using the density

$$g(\rho) = \begin{cases} \frac{1}{\sqrt{2\pi v}} e^{-\frac{1}{2v}(\rho-\mu)^2}, & \text{if } x \in (-1,1) \\ 0 & \text{otherwise} \end{cases}$$

where μ is the sample correlation between X and Y, and

$$v = \frac{1-|\mu|}{2cn}.$$

Bayesian inference

- Whether g is a good fit for π depends on the constant c. We will need to tune c for that purpose.
- We can tune c by computing the quantity CV in the importance algorithm.
- The main ingredient is the calculation of the importance ratio

$$w(\rho) = \exp\left(\frac{1}{2v}(\rho-\mu)^2 - \frac{N}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}(S_x^2 + S_y^2 - 2\rho S_{xy})\right).$$

Bayesian inference

```
omfun = function(rho,const_c,dt){
   n = length(dt[,1])
   murho = cor(dt[,1],dt[,2]);
   vrho = (1-abs(murho))/(2*const_c*n);
   Sx = sum(dt[,1]^2);
   Sy = sum(dt[,2]^2);
   Sxy = sum(dt[,1]*dt[,2]);
   val = 0.5*n*log(1-rho^2)
        -0.5*(Sx+Sy-2*rho*Sxy)/(1-rho^2);
   val = val +0.5*(rho-murho)*(rho-murho)/vrho;
   return(exp(val))
}
```

Bayesian inference

```
##generate data
n=100;
rho = 0.8;
X = rnorm(n);
Y = rho*X + sqrt(1-rho^2)*rnorm(n);
data = cbind(X,Y);
```

Bayesian inference

```
##Importance sampling
Nmc = 1e4;
c_val = 1;
rhohat = cor(X,Y);
vrho = (1-abs(rhohat))/(2*c_val*n);
R = rnorm(Nmc,rhohat,vrho);
R = R[(R>-1)&(R<1)] #this are draws from g
```

Bayesian inference

```
##Importance sampling
om = sapply(R,omfun,const_c = c_val, dt=data)
CV = sqrt(var(om))/mean(om) #compute the CV
Z = (om/mean(om))*R;
rho_est = mean(Z);
mc_err = sqrt(var(Z)/length(R))
print(c(rho_est,mc_err,CV))
##If CV small then we can trust rho_est
```