MAT 4381

Bayesian Inference

Study Guide

Winter 2024

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Bayesian vs. Frequentist

Frequentist: The parameter is not a random variable

Bayesian: The parameter is a random variable

Bayes Theorem

Ω be the sample space

$E_1, E_2, \dots$ sequence of exhaustive events

$A \subset \Omega$ another event

$$P(E_k|A) = \frac{P(A|E_k)P(E_k)}{\sum_k P(A|E_k)P(E_k)}$$

Bernoulli Trials

X takes value 1 if event occurs and 0 if event do not occur

Bernoulli Distribution: $Ber(\theta)$

$P(X=1) = \theta$

$P(X=0) = 1-\theta$

$P(X=x) = f(x|\theta) = \theta^x(1-\theta)^{1-x}\, I(x\in\{0,1\})$

$E[X] = \sum_{x=0}^{1} x f(x|\theta) = \theta$

$Var(X) = \theta(1-\theta)$

Binomial Distribution

$X_1, \dots, X_n \overset{iid}{\sim} Ber(\theta)$

$Y = \sum_{i=1}^{n} X_i \in \{0,1,2,\dots,n\} \sim Bin(n,\theta)$

$P(Y=y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}\, I(y\in\{0,1,\dots,n\})$

$E[Y] = n\theta \quad Var(Y) = n\theta(1-\theta)$

Bayes Formula for Distributions

$$f(\theta|y) = \frac{f(y|\theta)g(\theta)}{f_1(y)}$$

If θ is discrete with p.m.f. $g(\theta)$

$$f(\theta|y) = \frac{f(y|\theta)g(\theta)}{f_1(y)} = \frac{f(y|\theta)g(\theta)}{\sum_j f(y|\theta_j)g(\theta_j)}$$

If $\theta \sim g(\theta)$ is continuous

$$f(\theta|y) = \frac{f(y|\theta)g(\theta)}{f_1(y)} = \frac{f(y|\theta)g(\theta)}{\int f(y|\theta)g(\theta)d\theta}$$

The denominator is sometimes called the proportionality constant

$f(\theta|y) = C f(y|\theta)g(\theta)$

or $f(\theta|y) \propto f(y|\theta)g(\theta) = \text{likelihood} \times \text{prior}$

C some constant and $\propto$ means "proportional to"

$g(\theta)$: prior distribution for θ

$f(\theta|y)$: posterior distribution

Conjugate Prior

- A conjugate prior is also a convenient prior, since it allows one to easily calculate the posterior distribution

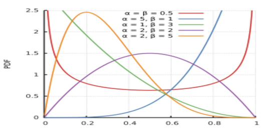
Beta Distribution

$$g(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} I(0<\theta<1)$$

The distribution is the $Beta(\alpha,\beta)$ distribution

We use the fact that

$$\frac{\Gamma(\alpha)}{\beta^\alpha} = \int_0^\infty x^{\alpha-1}e^{-\beta x}dx, \quad \alpha,\beta>0$$



The Beta distribution is a conjugate prior for binomial distribution

If $(Y|\theta=\theta) \sim f(y|\theta)$

$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} I(y\in\{0,1,\dots,n\})$

and $(\theta|\alpha,\beta) \sim Beta(\alpha,\beta)$

we have

$f(\theta|y) \propto f(y|\theta)g(\theta)$

In other words,

$f(\theta|y) \propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

The posterior is:

$$f(\theta|y) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)\Gamma(\beta+n-y)}\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}$$

If the prior is $Beta(\alpha,\beta)$, then the posterior is $Beta(\alpha+y, \beta+n-y)$

If $\theta \sim Beta(\alpha,\beta)$, then

$$E[\theta] = \frac{\alpha}{\alpha+\beta}$$

$$Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$\text{Mode: } \hat{\theta} = \frac{\alpha-1}{\alpha+\beta-2}$$

i) For large values of α and β mode and mean are roughly the same.

ii) Median for $Beta(\alpha,\beta)$ cannot be written in a closed form

iii) Peak of mode of beta density gets sharper as $\alpha+\beta$ increases

iv) We need to use numerical methods to calculate the med for different values of α and β

Credible Region

- We want to find I s.t. $P(\theta\in I) = 1-\alpha$, where $1-\alpha$ is the confidence level
- We can also use simulation method

Choosing Beta function

- First, consider priors that "don't express a strong" opinion.
- Later, consider priors which incorporate a strong prior belief. If there are no preferences on the value for θ, then one might consider using a "flat prior". That is, the uniform distribution between 0 and 1 $(Beta(1,1))$
- In some sense, this implies that each value of θ as equally likely
- Since the prior is flat, the posterior distribution has the same shape as the likelihood multiplied by a constant C

$h(\theta|x) = C f(x|\theta)g(\theta) \propto f(x|\theta)$

- Since the uniform distribution is $Beta(1,1)$, then the posterior is $Beta(y+1, n-y+1)$

Making Bayesian Inference

The posterior contains the results of the analysis

To communicate the result, we can either present the entire posterior distribution, or give summary statistics of the posterior

For credible regions, there are 2 common types:

- Smallest interval
- Equal tail area

Jeffreys Prior

Jeffreys proposed that an acceptable "non-informative prior finding principle" should be invariant under monotone transformations of the parameter

The Jeffrey's prior satisfying this invariance is proportional to $\sqrt{I(\theta)}$, where $I(\theta)$ is the Fisher information on θ. Can be generalized when θ is k-dimensional to $\sqrt{\det(I(\theta))}$

$$I(\theta) = -E\left[\frac{\partial^2 \ln f(x|\theta)}{\partial\theta^2}\right]$$

For k-dimensional, we have

$$I(\theta) = -E\left[\left(\frac{\partial^2 \ln f(x|\theta)}{\partial\theta_i\theta_j}\right)_{(i,j)}\right]$$

Multinomial Distribution

$(X_1,\dots,X_k) \sim Multinomial(n, p_1, \dots, p_k)$

$p_k = 1-p_1-p_2-\dots-p_{k-1}$

Models probability of counts for each side of an experiment with k exhaustive outcomes

$$f(x_1,x_2,\dots,x_k) = \binom{n}{x_1,x_2,\dots,x_{k-1}} p_1^{x_1}p_2^{x_2}\cdots p_{k-1}^{x_{k-1}}(1-p_1-p_2-\dots-p_{k-1})^{x_k}$$

where $x,y = 0,1,2,\dots,n, \; x_1+x_2+\dots+x_{k-1} \le n$

Simple Case: Trinomial Distribution

$(X, Y) \sim Multinomial(n, p_1, p_2)$

$$f(x,y) = \binom{n}{x,y} p_1^x p_2^y (1-p_1-p_2)^{n-x-y}$$

$x,y = 0,1,2,\dots,n, \; x+y\le n$

We can also get that

$X \sim Bin(n, p_1)$

$Y \sim Bin(n, p_2)$

$Y|X=x \sim Bin\left(n-x, \frac{p_2}{1-p_1}\right)$

$E[Y|X=x] = (n-x)\frac{p_2}{1-p_1}$

$Var[Y|X=x] = (n-x)\left(\frac{p_2}{1-p_1}\right)\left(1-\frac{p_2}{1-p_1}\right)$

$E[XY] = \frac{p_2}{1-p_1}E[X(n-X)] = -np_1p_2$

$Cov(X,Y) = -n(n+1)p_1p_2$

$$\rho(X,Y) = \frac{-n(n+1)p_1p_2}{\sqrt{np_1(1-p_1)np_2(1-p_2)}} = \frac{-(n+1)p_1p_2}{\sqrt{p_1p_2(1-p_1)(1-p_2)}}$$

Jeffreys Prior for Trinomial Distribution:

$$g(p_1,p_2) \propto \frac{1}{\sqrt{p_1p_2(1-p_1-p_2)}}$$

Special case of Dirichlet distribution $Dir(\frac{1}{2},\frac{1}{2},\frac{1}{2})$

Dirichlet Distribution

$(X_1,X_2,\dots,X_{k-1}) \sim Dir(\alpha_1,\alpha_2,\dots,\alpha_k)$

$$f(x_1,\dots,x_{k-1}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)}\left(\prod_{i=1}^{k-1} x_i^{\alpha_i-1}\right)\left(1-\sum_{i=1}^{k-1}x_i\right)^{\alpha_k-1}$$

Defined on $S = \{(x_1,\dots,x_{k-1}): \sum_{i=1}^{k-1}x_i<1, \; x_i>0, \; i=1,\dots,k\}$

Posterior for Trinomial

$(\theta_1,\dots,\theta_{k-1} | X_1=x_1,\dots,X_{k-1}=x_{k-1}) \sim Dir(\alpha_1+x_1,\dots,\alpha_{k-1}+x_{k-1}, n-\sum x_i+\alpha_k)$

$$E[\theta|X_1=x_1,\dots,X_{k-1}=x_{k-1}] = \left[\frac{\alpha_1+x_1}{\sum\alpha_i+n}, \dots, \frac{\alpha_k+x_k}{\sum\alpha_i+n}\right]$$

$\theta = [\theta_1 \; \theta_2 \; \cdots \; \theta_{k-1} \; \theta_k]^T$

$\theta_k = 1-\theta_1-\dots-\theta_{k-1} \quad x_k = n-x_1-\dots-x_{k-1}$

Notice: $(\theta_1,\theta_2) \sim Dir(\alpha_1,\alpha_2, \alpha_3+\dots+\alpha_k)$

Basic Properties of Dirichlet Distribution

$(Y_1,\dots,Y_{k-1}) \sim Dir(\alpha_1,\dots,\alpha_k) \quad \alpha_0 = \sum_{i=1}^k \alpha_i$

$$E[Y_i] = \frac{\alpha_i}{\alpha_0} \qquad Var[Y_i] = \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$$

$$Cov(X_i,X_j) = -\frac{\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}$$

Conjugate Prior and Posterior for Poisson

$$X_1,\dots,X_n \overset{iid}{\sim} f(x|\theta) = \frac{\exp(-\theta)\theta^x}{x!} I(x=0,1,\dots)$$

$$\text{Likelihood: } \prod_{i=1}^n f(x_i|\theta) = \prod_{i=1}^n \frac{\exp(-\theta)\theta^{x_i}}{x!} \propto \exp(-n\theta)\theta^y$$

$y = \sum_{i=1}^n x_i$

$$\text{Conjugate Prior: } g(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta) I(\theta>0)$$

$$\text{Posterior: } h(\theta|y) \propto \theta^{y+\alpha+1}\exp(-(\beta+n)\theta)\, I(\theta>0)$$

$$h(\theta|y) = \frac{(\beta+n)^{y+\alpha}}{\Gamma(y+\alpha)}\theta^{y+\alpha+1}\exp(-(\beta+n)\theta)\, I(\theta>0)$$

$$E[\theta|Y=y] = \frac{y+\alpha}{n+\beta}$$

$$Var[\theta|Y=y] = \frac{y+\alpha}{(n+\beta)^2}$$

$$Mode(\theta|Y=y) = \frac{y+\alpha-1}{n+\beta}$$

Jeffrey's Prior for Poisson

$$f(x|\lambda) = \frac{\exp(-n\lambda)\,\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

$$-E\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] = \frac{n}{\lambda}$$

Jeffrey's prior: $g(\lambda) \propto \frac{1}{\sqrt{\lambda}}$

This leads to an improper prior since we have

$$\int_0^\infty \frac{d\lambda}{\sqrt{\lambda}} = \infty$$

No normalizing constant will turn it into a prior p.d.f.

We can still use it to calculate the posterior

$$\lambda|X \propto \exp(-n\lambda)\,\lambda^{y+\frac{1}{2}-1}$$

The use of improper prior leads to Gamma posterior

$$\lambda|X \sim \frac{1}{\Gamma(\sum_{i=1}^n x_i + \frac{1}{2})}\lambda^{\sum_{i=1}^n x_i + \frac{1}{2} - 1}\exp(-n\lambda)$$

$$E[\lambda|x_1,\dots,x_n] = \frac{\sum_{i=1}^n x_i + 0.5}{n}$$

$$Var[\lambda|x_1,\dots,x_n] = \frac{\sum_{i=1}^n x_i + 0.5}{n^2}$$

Normal Distribution

$$(Y_1,\dots,Y_n|\mu) \sim f(y_1,\dots,y_n|\mu) \propto \exp\left(-\sum_{i=1}^n \frac{(y_i-\mu)^2}{2\sigma^2}\right)$$

σ^2 known

$$f(y_1,\dots,y_n|\mu) \propto \exp\left(-\frac{\mu^2 - 2\mu\bar{y}}{2\sigma^2/n}\right)$$

The Jeffrey's prior is

$$g(\mu) \propto 1$$

This constitutes an improper prior

However, we can still calculate the posterior

$$(\mu|Y_1=y_1,\dots,Y_n=y_n) \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

The normal prior

$Y_1,\dots,Y_n \overset{iid}{\sim} N(\theta,\sigma^2)$

σ^2 known

Prior: $g(\theta) \propto \exp(-(\theta-\mu)^2/\tau^2)$

Likelihood: $\prod_{i=1}^n f(y_1,\dots,y_n|\mu) \propto f(\bar{y}|\mu)$

$$\propto \exp\left(-\frac{(\bar{y}-\mu)^2}{2\sigma^2/n}\right)$$

Posterior for θ

$$h(\theta|y_1,y_2,\dots,y_n) \propto \exp(-(\theta-\mu)^2/\tau^2)\cdot\exp\left(-\frac{(\bar{y}-\mu)^2}{2\sigma^2/n}\right)$$

$$\propto \exp\left(\frac{(\bar{y}-\theta)^2}{2\sigma^2/n} - \frac{(\theta-\mu)^2}{2\tau^2}\right)$$

$$\propto \exp\left(-\frac{\theta^2-2\theta\bar{y}}{2\sigma^2/n} - \frac{\theta^2-2\mu\theta}{2\tau^2}\right)$$

$$\propto \exp\left(-\frac{\theta^2(\tau^2+\sigma^2/n) - 2\theta(\tau^2\bar{y}+\mu\sigma^2/n)}{2\tau^2\sigma^2/n}\right)$$

$$\propto \exp\left(-\frac{\theta^2 - 2\theta(\tau^2\bar{y}+\mu\sigma^2/n)/(\tau^2+\sigma^2/n)}{\frac{2\tau^2\sigma^2/n}{\tau^2+\sigma^2/n}}\right)$$

$$(\theta|y_1,\dots,y_n) \sim N\left(\frac{\tau^2\bar{y}+\mu\sigma^2/n}{\tau^2+\sigma^2/n}, \frac{\tau^2\sigma^2/n}{\tau^2+\sigma^2/n}\right)$$

$$E[\theta|Y=y] = \frac{\tau^2\bar{y}+\mu\sigma^2/n}{\tau^2+\sigma^2/n}$$

$$MAP = E[\theta|y_1,\dots,y_n] = Median(\theta|Y=y)$$

Posterior precision: $\frac{\tau^2+\sigma^2/n}{\tau^2\sigma^2/n} = \frac{1}{\sigma^2/n} + \frac{1}{\tau^2}$

Posterior precision is equal to prior precision plus observation precision

Since $E[\theta|y_1,\dots,y_n] = \frac{\sigma^2/n}{\tau^2+\sigma^2/n}\mu + \frac{\tau^2}{\tau^2+\sigma^2/n}\bar{y}$, the posterior mean is the weighted average of the prior mean and the mean of observations, where the weights are the proportions of the precision to the posterior precision.

Credible region for θ:

$$\frac{\tau^2\bar{y}+\mu\sigma^2/n}{\tau^2+\sigma^2/n} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\tau^2\sigma^2/n}{\tau^2+\sigma^2/n}}$$

If σ^2 unknown, we can use

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i-\bar{y})^2$$

We change $Z_{\alpha/2}$ to $t_{\alpha/2}(n-1)$

Credible region becomes

$$\frac{\tau^2\bar{y}+\mu\sigma^2/n}{\tau^2+s^2/n} \pm t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{\tau^2 s^2/n}{\tau^2+s^2/n}}$$

Jeffrey's prior for σ and σ^2

Suppose μ is fixed

Then we get the Jeffrey's prior for σ and σ^2 are

$$g(\sigma) \propto \frac{1}{\sigma} \quad \text{and} \quad g(\sigma^2) \propto \frac{1}{\sigma^2}$$

Bayes Factor and Relative Evidence

In model selection, say 2 models.

The Bayes factor is defined by

$$BF(1,2) = \frac{P(M_1|Data)P(M_2)}{P(M_2|Data)P(M_1)}$$

With a prior on model, the Bayes factor can be calculated. With equal probability on models, we get

$$BF(1,2) = \frac{P(M_1|Data)}{P(M_2|Data)}$$

Inverse χ^2 Distribution

$V \sim \chi^2(r)$. The p.d.f. for V is

$$f(v) = \frac{1}{2^{r/2}\Gamma(r/2)}\exp(-v/2)\,v^{r/2-1}I(v>0)$$

Define $U = \frac{1}{V}$

We say $U \sim Inv\text{-}\chi^2(v)$

The p.d.f for inverse-χ^2 is

$$g(u) = \frac{1}{2^{r/2}\Gamma(r/2)}\exp\left(-\frac{1}{2u}\right)u^{-r/2-1}\,I(u>0)$$

$$E[U] = \frac{1}{r-2} \qquad Mode(U) = \frac{1}{r+2}$$

$$Var(U) = \frac{2}{(r-2)^2(r-4)}$$

Mean exists if $r>2$ and $r>4$

Inverse Gamma Distribution

$X \sim Gamma(\alpha,\beta)$

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}\exp(-\beta x)\,x^{\alpha-1}I(x>0)$$

Define $U = X^{-1}$. The p.d.f for U is

$$g(u) = \frac{\beta^\alpha}{\Gamma(\alpha)}\exp(-\beta/u)\,u^{-(\alpha+1)}\,I(u>0)$$

Denote by $U \sim IG(\alpha,\beta)$

$$E[U] = \frac{\beta}{\alpha-1} \qquad Mode(U) = \frac{\beta}{\alpha+1}$$

$$Var(U) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

Inverse Gamma prior on σ^2, μ fixed

$(y_1,\dots,y_n|\mu,\sigma^2) \overset{iid}{\sim} N(\mu,\sigma^2)$

$$g(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}\exp\left(-\frac{\beta}{\sigma^2}\right)(\sigma^2)^{-(\alpha+1)}$$

$$h(\sigma^2|\mu,y_1,\dots,y_n) \propto (\sigma^2)^{-n/2}\exp\left(-\frac{\Sigma(y_i-\mu)^2}{2\sigma^2}\right)\cdot\exp(-\beta/\sigma^2)\cdot(\sigma^2)^{-(\alpha+1)}$$

$$\propto (\sigma^2)^{-(\alpha+\frac{n}{2}+1)}\exp\left(-\frac{\beta+\sum_{i=1}^n(y_i-\mu)^2/2}{\sigma^2}\right)$$

Therefore

$$(\sigma^2|\mu,y_1,\dots,y_n) \sim IG\left(\alpha+\frac{n}{2}, \beta+\sum_{i=1}^n(y_i-\mu)^2/2\right)$$

In this case

$$E[\sigma^2|\mu,y_1,\dots,y_n] = \frac{\beta+\sum_{i=1}^n(y_i-\mu)^2/2}{\alpha+\frac{n}{2}-1}$$

$$Mode = \frac{\beta+\sum_{i=1}^n(y_i-\mu)^2/2}{\alpha+\frac{n}{2}+1}$$

If $X \sim \chi^2(v)$, then $1/X \sim Inv\text{-}\chi^2(v)$

Inverse chi-squared distribution is a special case of gamma distribution with parameters $\alpha = v/2$, $\beta = 1/2$, or $\alpha = v/2$, $\beta = v\tau^2/2$ for scaled inverse-χ^2 distribution

Posterior Predictive Distribution

Let y_{n+1} be the next observations drawn after the random sample $Y_n = (y_1, y_2, \dots, y_n)$.

The predictive accuracy for $(y_{n+1}|y_1,y_2,\dots,y_n)$ is the conditional density

$$f(y_{n+1}|Y_n) = \int f(y_{n+1}|\mu)\,g(\mu|Y_n)\,d\mu$$

Jeffrey's Prior for normal family with both μ and σ^2 unknown

$$X \sim N(\mu,\sigma^2) \qquad \theta = [\mu \;\; \sigma^2]^T$$

Then we get

$$-E\left[\frac{\partial^2 \ln f}{\partial\mu^2}\right] = \frac{1}{\sigma^2} \qquad -E\left[\frac{\partial^2 \ln f}{\partial\sigma^2\partial\mu}\right] = 0$$

$$-E\left[\frac{\partial^2 \ln f}{\partial(\sigma^2)^2}\right] = \frac{1}{\sigma^4}$$

The Jeffrey's prior is

$$g(\mu,\sigma^2) \propto (\sigma^2)^{-3/2}$$

To calculate posterior, we get

$$h(\mu,\sigma^2|y_1,\dots,y_n) \propto (\sigma^2)^{-(n+3)/2}\exp\left(-\frac{(n-1)S_Y^2 + n(\mu-\bar{y})^2}{2\sigma^2}\right)$$

This shows that

$$(\mu,\sigma^2|y_1,\dots,y_n) \sim N\text{-}\Gamma^{-1}\left(\mu=\bar{y}, \lambda=n, \alpha=n, \beta=\sum_{i=1}^n(y_i-\bar{y})^2/2\right)$$

Normal Inverse Gamma Distribution

Let $\lambda > 0$ and

$$(y|\mu,\sigma^2,\lambda) \sim N(\mu, \sigma^2/\lambda)$$

$$(\sigma^2|\alpha,\beta) \sim IG(\alpha,\beta)$$

Then we say (y,σ^2) has normal inverse gamma distribution with parameters $(\mu,\lambda,\alpha,\beta)$ which is denoted by

$$(y,\sigma^2) \sim NIG(\mu,\lambda,\alpha,\beta)$$

$$(y,\sigma^2) \sim N\text{-}\Gamma^{-1}(\mu,\lambda,\alpha,\beta)$$

We can write the joint p.d.f. for (y,σ^2) as follows

$$f(y,\sigma^2|\mu,\lambda,\alpha,\beta) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}}\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1/2+1)}\exp\left(\frac{2\beta+\lambda(y-\mu)^2}{2\sigma^2}\right)$$

$y \in \mathbb{R}$, $\sigma^2 > 0$

By definition, $(\sigma^2|\alpha,\beta) \sim \Gamma^{-1}(\alpha,\beta)$

We can get that $y \sim t(2\alpha, \mu, \beta/(\lambda\alpha))$

Student t distribution

Let $W \sim N(0,1)$ and $V \sim \chi^2(r)$ and W and V are mutually independent. The joint distribution for W and V is

$$f(v,w) = \frac{1}{\sqrt{2\pi}}\frac{1}{2^{r/2}\Gamma(r/2)}\exp\left(-\frac{w^2}{2}\right)v^{r/2-1}\exp(-v/2)$$

for $w \in \mathbb{R}$, $v > 0$

If we define $T = \frac{W}{\sqrt{\frac{V}{r}}}$ and $U = V$,

we get

$$g(u,t) = \frac{1}{\sqrt{\pi r}}\frac{1}{2^{(r+1)/2}\Gamma(r/2)}u^{(r+1)/2-1}\exp(-u(1+t^2/r)/2)$$

$$g(t) = \frac{1}{\sqrt{\pi r}}\frac{\Gamma((r+1)/2)}{\Gamma(r/2)(1+\frac{t^2}{r})^{(r+1)/2}} \quad \text{when } t>0$$

This is known as the student t distribution

Notes:

- $g(t)$ is a Cauchy distribution if $r=1$
- $g(t) \to N(0,1)$ as $r \to \infty$

If we use the transformation $V = \sigma t + \mu$ when $\sigma > 0$, we get

$$h(v) = \frac{1}{\sigma\sqrt{\pi r}}\frac{\Gamma((r+1)/2)}{\Gamma(r/2)}\frac{1}{\left(1+\frac{1}{r}\frac{(v-\mu)^2}{\sigma^2}\right)^{(r+1)/2}}$$

Denote this distribution by $t(r,\mu,\sigma)$

If we let $U = \sigma\left(\frac{W}{\sqrt{V/r}}\right) + \mu$, we get

$$Median(U) = Mode(U) = E[U] = \sigma E[W]E\left[\frac{1}{\sqrt{V/r}}\right] + \mu$$

If $E\left[\frac{1}{\sqrt{V/r}}\right]$ exists then $E[U] = \mu$

$$Var(U) = \frac{r\sigma^2}{r-2}$$

Properties of Normal Inverse Gamma Distribution

$$(y,\sigma^2) \sim N\text{-}\Gamma^{-1}(\mu,\lambda,\alpha,\beta) \Longrightarrow (\sigma^2|\alpha,\beta) \sim \Gamma^{-1}(\alpha,\beta)$$

$$E[\sigma^2|\alpha,\beta] = \frac{\beta}{\alpha-1}$$

$$Var[\sigma^2|\alpha,\beta] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

$$Var[y] = E\left[\frac{\sigma^2}{\lambda}\right] = \frac{\beta}{\lambda(\alpha-1)}$$

Normal-Inverse Gamma Prior

$$(y_1,\dots,y_n|\mu,\sigma^2) \overset{iid}{\sim} N(\mu,\sigma^2)$$
$$\mu|\sigma^2 \sim N(\mu_0, \sigma^2/n_0)$$
$$\sigma^2 \sim IG(\alpha,\beta)$$

n_0: equivalent sample size of the prior

n: equivalent sample size we would obtain for n observations

The posterior is

$$f(\mu,\sigma^2|y_1,\dots,y_n) \propto (\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{\sum_{i=1}^n(y_i-\mu)^2}{2\sigma^2}\right)$$
$$\times(\sigma^2)^{-\frac{1}{2}}\exp\left(\frac{-n_0(\mu-\mu_0)^2}{2\sigma^2}\right)$$
$$\times(\sigma^2)^{-(\alpha+1)}\exp\left(\frac{\beta}{\sigma^2}\right)$$

Therefore,

$$f(\mu,\sigma^2|y_1,\dots,y_n) \propto (\sigma^2)^{-(n+2\alpha+3)/2}$$
$$\times\exp\left(-\frac{\sum_{i=1}^n(y_i-\mu)^2}{2\sigma^2}-\frac{n_0(\mu-\mu_0)^2}{2\sigma^2}-\frac{\beta}{\sigma^2}\right)$$

Expand to get

$$f(\mu,\sigma^2|y_1,\dots,y_n)\propto\exp\left(-\mu^2\left(\frac{n+n_0}{2\sigma^2}\right)+\left(\frac{n\bar{y}+n_0\mu_0}{\sigma^2}\right)\mu\right)$$
$$\times(\sigma^2)^{-(n+2\alpha+3)/2}\exp\left(-\frac{\sum_{i=1}^n y_i^2+n_0\mu_0^2+2\beta}{2\sigma^2}\right)$$

Therefore

$$f(\mu,\sigma^2|y_1,\dots,y_n)\propto\exp\left(-\frac{\mu^2-2\mu\left(\frac{n\bar{y}+n_0\mu_0}{n+n_0}\right)}{\frac{2\sigma^2}{n+n_0}}\right)$$
$$\times(\sigma^2)^{-(n+2\alpha+3)/2}\exp\left(-\frac{\sum_{i=1}^n y_i^2+n_0\mu_0^2+2\beta}{2\sigma^2}\right)$$

In other words,

$$f(\mu,\sigma^2|y_1,\dots,y_n)\propto\exp\left(-\frac{\left(\mu-\frac{n\bar{y}+n_0\mu_0}{n+n_0}\right)^2}{\frac{2\sigma^2}{n+n_0}}\right)$$
$$\times(\sigma^2)^{-(n+2\alpha+3)/2}\exp\left(-\frac{\sum_{i=1}^n y_i^2+n_0\mu_0^2+2\beta}{2\sigma^2}+\frac{(n\bar{y}+n_0\mu_0)^2}{2\sigma^2(n+n_0)}\right)$$

The posterior for σ^2 is

$$f(\sigma^2|y_1,\dots,y_n)\propto(\sigma^2)^{-(\alpha+\frac{n}{2}+1)}\times\exp\left(-\frac{\sum_{i=1}^n y_i^2+n_0\mu_0^2+2\beta}{2\sigma^2}+\frac{(n\bar{y}+n_0\mu_0)^2}{2\sigma^2(n+n_0)}\right)$$

$$\sigma^2|y_1,\dots,y_n\sim IG\left(\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(y_i-\bar{y})^2}{2}+\frac{nn_0}{n+n_0}\frac{(\bar{y}-\mu_0)^2}{2}\right)$$

For the posterior of μ, we have

$$h(\mu|\sigma^2,\vec{y})=\frac{f(\mu,\sigma^2|\vec{y})}{f(\sigma^2|\vec{y})}$$

$$\mu|\sigma^2,\vec{y}\sim N\left(\frac{n\bar{y}+n_0\mu_0}{n+n_0},\ \frac{\sigma^2}{n+n_0}\right)$$

$$h(\mu|\vec{y})=\int_0^\infty h(\mu|\sigma^2,\vec{y})h(\sigma^2|\vec{y})\,d\sigma^2$$

Simulating Posterior for μ and σ^2 given data

$$\sigma^2|y_1,\dots,y_n\sim IG\left(\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(y_i-\bar{y})^2}{2}+\frac{nn_0}{n+n_0}\frac{(\bar{y}-\mu_0)^2}{2}\right)$$

$$\mu|\sigma^2,y_1,\dots,y_n\sim N\left(\frac{n\bar{y}+n_0\mu_0}{n+n_0},\ \frac{\sigma^2}{n+n_0}\right)$$

We can first simulate σ^2 from the inverse gamma distribution. Then we use it to simulate μ from the normal distribution.

Algorithm:
- Load data $y_1,\dots,y_n$
- Find $\bar{y}$, α, and β parameters
- Simulate σ^2 a number of times
- Use σ^2 to simulate μ

We have μ is symmetrically distributed. We can find a credible region the following way:
- Sort the simulated μ values
- Find 2.5 and 97.5 percentiles
- Then we get the credible region

This method is more tricky with σ^2 since it is not symmetrically distributed.

Acceptance-Rejection Sampling

If we only know the unscaled target distribution, we can use acceptance-rejection sampling to draw a random sample from the target distribution.

This method works by drawing a random sample from an easily sample candidate distribution (g_0) by accepting some which satisfy a criterion. Let M be a positive constant and $g_0(\theta)$ be the easily sampled candidate and

$$g(\theta)f(y|\theta)\le Mg_0(\theta)$$

The goal is to draw a sample from the posterior

$$h(\theta|y)\propto g(\theta)f(y|\theta)$$

The algorithm is as follows:
1. Draw a sample from $g_0(\theta)$
2. Calculate $w(\theta)=\dfrac{g(\theta)f(y|\theta)}{Mg_0(\theta)}$

   Note: $w(\theta)\sim(0,1]$
3. Draw $u\sim Unif(0,1)$
4. If $u<w(\theta)$, accept θ
5. Otherwise, reject θ and iterate step 1 until the first acceptance.

Theorem: Let f and g be two p.d.f.s with c.d.f.s F and G, respectively.

Assume $f(y)\le Mg(y)$ for all y.

Draw $U\sim Unif(0,1)$ and $Y\sim g(y)$ independent from U. If $U\le\frac{f(Y)}{Mg(Y)}$ then accept Y, otherwise, ignore and iterate the previous steps until a sample is accepted. Define N to be the number of repeats until the first sample acceptance. We have:
- $N\sim Geom(p)$

  $P(N=n)=p(1-p)^{n-1}I(n\in\{1,2,\dots\})$

  with $p=\frac{1}{M}$
- $P\left(Y\le y\,|\,U<\frac{f(Y)}{Mg(Y)}\right)=F(y)$

Plug in Principle

Let $X_1,\dots,X_n$ be a sequence of i.i.d. random variables with unknown distribution F. A natural estimate for F is

$$F_n=\frac{1}{n}\sum_{i=1}^n\delta_{x_i}$$

The method that replaces the empirical process, a natural estimate, for the actual distribution, is usually called the plug in principle. In other words the parameter $T(F)$ is estimated by the statistic $T(F_n)$. In general from the plug in principle,

$$T(F)=\int g(x)\,dF(x)$$

should be estimated by

$$T(F_n)=\int g\,dF_n=\frac{g(x_1)+\dots+g(x_n)}{n}$$

This roots from the fact that $F_n\xrightarrow{a.s.}F$ (strong law of large numbers)

Alternative Methodology

The expert makes a prior assumption on F (i.e. F belongs to a family of distributions) and then it will be updated after observing data.

Dirichlet Process

Consider a space $\mathcal{X}$ with a σ-algebra $\mathcal{A}$ of subsets of $\mathcal{X}$. Let H be a fixed probability measure on $(\mathcal{X},\mathcal{A})$ and α be a positive number.

A random probability measure $P=\{P(A)\}_{A\in\mathcal{A}}$ is called a $DP(\alpha H)$, if for any finite measurable partition $\{A_1,\dots,A_k\}$ of $\mathcal{X}$

$$(P(A_1),\dots,P(A_k))\sim Dir(\alpha H(A_1),\dots,\alpha H(A_k))$$

$$P\sim DP(\alpha,H)$$

- For any $A\in\mathcal{A}$, we have
$$P(A)\sim Beta(\alpha H(A),\alpha H(A^c))$$
- For any $A\in\mathcal{A}$, we have
$$E[P(A)]=\frac{\alpha H(A)}{\alpha H(A)+\alpha H(A^c)}=H(A)$$
$$Var[P(A)]=\frac{\alpha^2H(A)H(A^c)}{(\alpha H(A)+\alpha H(A^c))^2(\alpha H(A)+\alpha H(A^c)+1)}$$
$$=\frac{H(A)(1-H(A))}{1+\alpha}$$

Note:
$$P(|P(A)-H(A)|>\varepsilon)\le\frac{Var(P(A))}{\varepsilon^2}=\frac{H(A)(1-H(A))}{\varepsilon^2(1+\alpha)}$$

Therefore,
$$P(A)\xrightarrow{d}H(A)$$

Imposing a large α means that you are working more or less with H. Small α means that, we provide more flexibility to P.

Posterior of Dirichlet Process

If $X_1,\dots,X_n$ is a set of realizations from $P\sim DP(\alpha,H)$, then

$$P|X_1,\dots,X_n\sim DP(\alpha_m^*,H_m^*)$$

where

$$\alpha_m^*=\alpha+m\quad\text{and}\quad H_m^*=\frac{\alpha}{\alpha+m}H+\frac{m}{\alpha+m}\frac{\sum_{i=1}^m\delta_{x_i}}{m}$$

$$\delta_x(A)=\begin{cases}1 & \text{if } x\in A\\ 0 & \text{if } x\notin A\end{cases}$$

Therefore, H^* is a mixture of H and empirical process.

Nonparametric Bayes Estimates

Under the squared error loss, an estimate for P is

$$E[P|X_1,\dots,X_m]=H_m^*$$

As $m\to\infty$, $H_m^*\xrightarrow{a.s.}P$, where P is the actual distribution of X.

We also have

$$\sqrt{m}(P_m^*-P)\xrightarrow{d}B(F)$$

where B is Brownian bridge.

Frullani Integral

Let $a,b>0$. If f' exists and is continuous, $|f(0)|<\infty$ and $|f(\infty)|<\infty$ then

$$\int_0^\infty\frac{f(bx)-f(ax)}{x}dx=(f(\infty)-f(0))\ln\left(\frac{b}{a}\right)$$

Special case of Frullani Integral

$$\int_0^\infty(\exp(\theta u)-1)\frac{\exp(-u)}{u}du=-\ln(1-\theta)$$

Lévy Measure of Gamma$(\alpha,1)$

For $\alpha>0$, let

$$N(x)=\alpha\int_x^\infty\frac{\exp(-t)}{t}dt$$

Gamma Process

Define

$$G_t=\sum_{i=1}^\infty N^{-1}(T_i+t)$$

$$T_i=E_1+\dots+E_i\qquad(E_i)_{i\ge1}\overset{iid}{\sim}Exp(1)$$

The moment generating function for G_t is

$$M(\theta,t)=\exp\left(-\int_t^\infty\left(1-\exp(\theta N^{-1}(v))\right)dv\right)$$

$$M(\theta,0)=(1-\theta)^{-\alpha}$$

Therefore, we get

$$\sum_{i=1}^\infty N^{-1}(T_i)\sim\Gamma(\alpha,1)$$

Let $\theta_i\overset{iid}{\sim}P_0$. Then the set indexed process

$$G(\cdot)=\sum_{i=1}^\infty N^{-1}(T_i)\delta_{\theta_i}(\cdot)$$

is called a Gamma process

Theorem: We have
- $G(A)\sim\Gamma(\alpha P_0(A),1)$
- $G(A)$ is independent of $G(B)$ if $A\cap B=\emptyset$

Dirichlet Process (Formal Definition)

Let $\theta_i\overset{iid}{\sim}P_0$ be independent from $\{T_i\}$ such that $\alpha>0$. The random probability measure

$$P(\cdot)=\sum_{i=1}^\infty\frac{N^{-1}(T_i)}{\sum_{i=1}^\infty N^{-1}(T_i)}\delta_{\theta_i}(\cdot)$$

is called a Dirichlet process with concentration parameter α and concentration measure P_0 and is denoted $P\sim Dir(\alpha,P_0)$

Most of the time P_0 is defined on $\mathbb{R}^d$. In this case:

$$H(x_1,\dots,x_k)=P_0(X_1\le x_1,\dots,X_k\le x_k)$$

We usually write

$$P\sim Dir(\alpha,H)$$

The Dirichlet Process is difficult to use in practice since there's no closed form for $N^{-1}(x)$. There are also infinitely many weights to it.

The above representation of the Dirichlet process is know as the Ferguson's Series Representation, where
- $T_i=E_1+\dots+E_i$, $(E_i)_{i\ge1}\overset{iid}{\sim}Exp(1)$
- $(\theta_i)_{i\ge1}\overset{iid}{\sim}H$
- $(\theta_i)_{i\ge1}$ and $(T_i)_{i\ge1}$ are independent
- $N(x)$ is the Lévy measure of a Gamma$(\alpha,1)$ random variable.

Sethuraman's Representation of DP

$$P^{Seth}(\cdot) = \sum_{i=1}^{\infty} P_i \delta_{\theta_i}(\cdot)$$

where:

- $(\theta_i)_{i\geq 1} \overset{iid}{\sim} H$
- $(P_i)_{i\geq 1}$ is an independent sequence of random variables (called weights) such that

$$P_1 = \beta_1, \quad P_i = \beta_i \prod_{k=1}^{i-1}(1-\beta_k), \ i \geq 2$$

- $(\beta_i)_{i\geq 1} \overset{iid}{\sim} Beta(1, \alpha)$
- $(\theta_i)_{i\geq 1}$ and $(P_i)_{i\geq 1}$ independent

Sethuraman's Truncated Representation

$$P_n^{Seth}(\cdot) = \sum_{i=1}^{n} P_i \delta_{\theta_i}(\cdot)$$

where:

- $(\theta_i)_{i\leq n}$ and $(P_i)_{i\leq n}$ are defined as in Sethuraman's representation with $\beta_n = 1$
- The assumption that $\beta_n = 1$ is necessary to make the weights added to 1 almost surely
- A random stopping rule for choosing $n = n(\varepsilon)$ was proposed by Muliere and Tradella (1998), where, for $\varepsilon \in (0,1)$,

$$n = \inf\{i : P_i = (1-\beta_1)\cdots(1-\beta_{i-1}), \beta_i < \varepsilon\}$$

Markov Process

A stochastic process X_t is a Markov process if for all $n \in \{0,1,2,\dots\}$,

$$P(X_t \in (a,b] \mid X_{t_0} = x_0, X_{t_1} = x_1, \dots, X_{t_n} = x_n)$$
$$= P(X_t \in (a,b] \mid X_{t_n} = x_n)$$

when $t_0 < t_1 < \dots < t_n < t$

For Discrete time Markov Chain we can write

$$P(X_{n+1} = j \mid X_0 = x_{i_0}, \dots, X_n = x_{i_n}) = P(X_{n+1} = j \mid X_n = x_{i_n})$$

Probability Transition Matrix

The probability transition matrix P is a matrix where the (i,j)-th entry is

$$P_{ij} = P(X_2 = j \mid X_1 = i)$$

Notice that $\sum_j P_{ij} = 1$. This means that sum of each row of P is 1