Bayesian Theory
Study Note

Ian S.W. Ma

Winter 2020

Contents

1 Statistics Basics

General Structure: Probability Space/Triple

Probability Space/Triple $(\Omega, \mathcal{F}, \mathbb{P})$

- Ω: **Sample Space**, set of all posible outcomes.
- $\mathcal{F}$: **Set of events** σ, each event is a subset of Ω
- $\mathbb{P}$ Probability of event $\sigma \in \mathcal{F}$

Conditional Probability

Conditional probability of A given B: $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \Leftrightarrow \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B)$

General properties

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

  Events are mutually exclusive $\Rightarrow \mathbb{P}(A \cap B) = 0$
- $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$

  Events are independent $\Rightarrow \mathbb{P}(A|B) = \mathbb{P}(A)$
- Law of Total Probability

  If events A_1, A_2, A_N are mutually exclusive and exhaustive ($\bigcup_{i=1}^{N} A_i = \Omega$) then:

$$\mathbb{P}(B) = \sum_{i=1}^{N} \mathbb{P}(A_i \cap B) = \sum_{i=1}^{N} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

Random Variables

$X \sim$ Distribution	$\mathbb{E}[X]$	$\mathrm{Var}[X]$
Normal(μ, σ^2)	μ	σ^2
Bernoulli(θ)	θ	$\theta(1-\theta)$
Binomial(θ)	$n\theta$	$n\theta(1-\theta)$
Poisson(μ)	μ	μ
Uniform(α, β)	$\frac{\alpha+\beta}{2}$	$\frac{(\alpha+\beta)^2}{12}$
Beta(α, β)	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Gamma(α, β)	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Inv. Gamma(α, β)	$\frac{\beta}{\alpha-1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$
Exponential(λ)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Random Vectors

$\mathbf{X} = X_1, ..., X_k \sim$ Distribution	$\mathbb{E}[X_i]$	$\mathrm{Var}[X_i]$	$\mathrm{Cov}(X_i, X_j)$
Multivariative Normal$(\boldsymbol{\mu}, \Sigma)$	μ_i	$\Sigma_{i,i}$	$\Sigma_{i,j}$
Multinomial$(n, \theta_1, ..., \theta_k)$	$n\theta_i$	$n(1-\theta_i)$	$-n\theta_i\theta_j$

2 Bayes Throrem

Discrete Case

$$
\begin{aligned}
\mathbb{P}(X = x_i | Y = y_j) &= \frac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(Y = y_j)} \\
&= \frac{\mathbb{P}(Y = y_j | X = x_i)\mathbb{P}(X = x_i)}{\mathbb{P}(Y = y_j)} \\
&= \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal}} \\
&= \frac{\mathbb{P}(Y = y_j | X = x_i)\mathbb{P}(X = x_i)}{\sum_{k=1}^{n} \mathbb{P}(X = x_k, Y = y_j)} \\
&= \frac{\mathbb{P}(Y = y_j | X = x_i)\mathbb{P}(X = x_i)}{\sum_{k=1}^{n} \mathbb{P}(Y = y_j | X = x_k) P(X = x_k)} \\
\Rightarrow \mathbb{P}((X|Y)) &= \frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\sum_x \mathbb{P}(X = x, Y)} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal}}
\end{aligned}
$$

Continous Case

$$
f(X|Y) = \frac{f(X, Y)}{g(Y)} = \frac{g(Y|X)f(X)}{\int g(Y|X)f(X)dX}
$$

Bayesian Statistical Inference

- $L(\theta|y) = f(y|\theta)$: Likelihood
- $\pi(\theta)$: Prior
- $m(y) = p(y)$: Marginal distribution/Normalizing constant

$$
\text{Posterior Distribution: } p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int p(\theta, y)d\theta} = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \propto f(y|\theta)\pi(\theta)
$$

Predictive Distrubutions

$$
\begin{aligned}
\text{Prior Predictive Distribution:} \qquad & p(y^{new}) = \int_\theta f(y^{new}|\theta)\pi(\theta)d\theta \\
\text{Posterior Predictive Distribution:} \qquad & p(y^{new}|y^{old}) = \int_\theta f(y^{new}|\theta, y^{old})p(\theta|y^{old})d\theta
\end{aligned}
$$

Bayes Theorem for Multiple Parameters $\Theta = \{\theta_1, ..., \theta_q\}$

$$
\mathbb{P}(\Theta|\boldsymbol{y}) = \frac{\mathbb{P}(\boldsymbol{y}|\Theta)\mathbb{P}(\Theta)}{\mathbb{P}(\boldsymbol{y})} \propto \mathbb{P}(\boldsymbol{y}|\Theta)\mathbb{P}(\Theta)
$$

3 Conjugate Priors

Sample Distribution	Parameter	Prior
Bernoulli(θ)	θ	Beta(α, β)
Binomial(n, θ)	θ	Beta(α, β)
Poisson(θ)	θ	Gamma(α, β)
Exponential(θ)	θ	Gamma(α, β)
Normal(μ, σ_0^2)	μ	Normal(μ_h, σ_h^2)
Normal(μ_0, σ^2)	σ^2	Inverse Gamma(α, β)
Normal($\mu_0, 1/\tau$)	$\tau = 1/\sigma^2$	Gamma(α, β)
Multinomial($n, \theta_1, ..., \theta_k$)	$\theta_1, ..., \theta_k$	Dirichlet($\alpha_1, ..., \alpha_k$)
Uniform($0, \theta$)	θ	Pareto(θ_m, α)

4 Normal Distribution Priors

Unknown μ, known σ^2

- Sample $\boldsymbol{y} \in \mathbb{R}^n$
- Prior $\mu \sim \text{Normal}(\mu_0, \sigma_0^2) = \text{Normal}\left(\mu_0, \frac{\sigma^2}{m}\right) \Rightarrow m = \sigma^2/\sigma_0^2$
- Posterior:

$$\mu|\boldsymbol{y}, \sigma^2 \sim \text{Normal}\left(\frac{m\mu_0 + n\bar{y}}{m+n}, \frac{\sigma^2}{m+n}\right) = \text{Normal}\left((1-w)\mu_0 + w\bar{y}, \frac{\sigma^2}{m+n}\right)$$

$w = \frac{m}{m+n}$

Unknown μ, known τ

- Sample $\boldsymbol{y} \in \mathbb{R}^n$
- Prior $\mu \sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{m}\right) = \text{Normal}\left(\mu_0, \frac{1}{\tau m}\right)$
- Posterior:

$$\mu|\boldsymbol{y}, \sigma^2 \sim \text{Normal}\left(\frac{m\mu_0 + n\bar{y}}{m+n}, \frac{1}{\tau(m+n)}\right) = \text{Normal}\left((1-w)\mu_0 + w\bar{y}, \frac{1}{\tau(m+n)}\right)$$

$w = \frac{m}{m+n}$

Unknown τ, known μ

- Sample $\boldsymbol{y} \in \mathbb{R}^n$
- Prior $\tau \sim \text{Gamma}(\alpha, \beta)$
- Posterior:

$$\tau|\boldsymbol{y}, \mu \sim \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{z^2}{2}\right)$$

$z^2 = \sum_i (y_i - \mu)^2$

Unknown σ^2, known μ

- Sample $\boldsymbol{y} \in \mathbb{R}^n$
- Prior $\mu, \sigma^2 \sim$ Normal $\left(\mu_0, \frac{\sigma^2}{\kappa}\right) \times$ Inverse Gamma(α, β)
- Posterior:

$$\mu, \sigma^2 | \boldsymbol{y} \sim \text{Normal}\left(\frac{\kappa\mu_0 + n\bar{y}}{\kappa + n}, \frac{\sigma^2}{\kappa + n}\right) \times \text{Inverse Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{(n-1)s^2}{2} + \frac{\kappa n(\bar{y} - \mu_0)^2}{2(\kappa + n)}\right)$$

$s^2 = \sum_i (y_i - \bar{y})^2/(n-1)$
$\kappa = \sigma^2(\alpha - 1)/\beta$ or given (seriously it's not written anywhere Ken, I'm just guessing here man)

5 Jeffrey's Prior

Definition

Jeffery Prior: $\pi_{JP}(\theta) \propto \sqrt{I(\theta|y)}$
Fisher Information:

$$I(\theta|y) = \mathbb{E}\left[\left(\frac{d\log f(y|\theta)}{d\theta}\right)^2\right] = -\mathbb{E}\left[\frac{d^2\log f(y|\theta)}{d\theta^2}\right]$$

1:1 transformation of Jeffrey's Prior

$$\pi(\phi) = \pi_{JP}(\theta)\left|\frac{d\theta}{d\phi}\right|$$

Jeffrey's Prior for Multivariate Parameter Vector

Gradient of log likelihood:

$$S(\boldsymbol{\theta}) = \begin{bmatrix} \partial_{\theta_1} \log f(x) \\ \vdots \\ \partial_{\theta_n} \log f(x) \end{bmatrix}$$

Hessian Matrix of log likelihood:

$$H(\boldsymbol{\theta}) = \text{Jacobian}\left[S(\boldsymbol{\theta})\right] = \begin{bmatrix} \partial_{\theta_1}\partial_{\theta_1} \ln f(x) & \cdots & \partial_{\theta_1}\partial_{\theta_n} \ln f(x) \\ \vdots & \ddots & \vdots \\ \partial_{\theta_n}\partial_{\theta_1} \ln f(x) & \cdots & \partial_{\theta_n}\partial_{\theta_n} \ln f(x) \end{bmatrix}$$

Fisher Information: $I(\theta|x) = -\mathbb{E}\left[H(\theta)\right]$
Jeffrey's Prior: $\pi_{JP} \propto \sqrt{\det(I(\theta|x))}$

6 Reference Prior

Kullback-Leibler (KL) divergence

A measure of the difference between two pmfs/pdfs. When $KL(f,g) = 0$, the two distributions are idetical.

Properties

- $KL(f,g) \neq KL(g,f)$
- $KL(fmg) \geq 0$

Continous parameter x

$$KL(f,g) = \int \ln\left[\frac{f(x)}{g(x)}\right] dx = \mathbb{E}_f[\log f(X)] - \mathbb{E}_g[\log f(X)]$$

Discrete parameter x

$$KL(f,g) = \sum_{x \in X} \ln\left[\frac{f(x)}{g(x)}\right] = \mathbb{E}_f[\log f(X)] - \mathbb{E}_g[\log f(X)]$$

Reference Prior

Read week 4 notes (2.2)

7 Point Estimates

Components of Decision Theory with emphasis on parameter estimation

- State Spcae Θ, unknown true value $\theta \in \Theta$
- Action Space $\mathcal{A} \ni a$, sometimes $\mathcal{A} = \Theta$
- Sampling Distribution $f(\boldsymbol{y}|\theta)$
- Loss Function $\mathcal{L}(a|\theta)$
- Risk $R_\theta(a|\boldsymbol{y}) = \mathbb{E}_\theta[L(a|\theta, y)] = \int_{\theta \in \Theta} \mathcal{L}(a|\theta) p(\theta|\boldsymbol{y}) d\theta$
- Bayes Estimator of a parameter $\hat{\theta}_{BE} = \arg\min_{\hat{\theta} \in \Theta} R_\theta(\hat{\theta}|\boldsymbol{y})$

Common Loss Functions and Corresponding Estimators

- **Squared Error Loss**: $\mathcal{L}(\theta|\hat{\theta}) = (\theta - \hat{\theta})^2$, Bayes Estimator $\hat{\theta} = \mathbb{E}\left[\theta|\boldsymbol{y}\right]$
- **Absolute Error Loss**: $\mathcal{L}(\theta|\hat{\theta}) = |\theta - \hat{\theta}|$, Bayes Estimator $\hat{\theta} = \theta_{0.5}$, $(\mathbb{P}(\theta \leq \theta_{0.5}) = 0.5)$
- **0-1 Loss**: $\mathcal{L}(\theta|\hat{\theta}) = I(\theta \neq \hat{\theta})$, Bayes Estimator $\hat{\theta}$ is the posterior mode

8 Interval Estimates

$$P \times 100\% \text{ Bayesian Credible interval} = [LB, UB] \text{ where } \int_{LB}^{UB} p(\theta, \boldsymbol{y}) d\theta = P$$

If loooking for one side-confidence bounds then $LB = -\infty$ or $UB = \infty$

Symmetric Credicble Interval

$$[LB, UB] \text{ where } \mathbb{P}(\theta \leq LB) = \mathbb{P}(\theta \geq UB) = \alpha/2 = (1-P)/2$$

Highest Posterior Density Interval (HPDI)

Credible Interval where all values outside the interval has a density smaller than any of the values in the interval.

Credible Regions

$$\iint p(\theta_1, \theta_2 | \boldsymbol{y}) d\theta_1 d\theta_2 = 1 - \alpha = P$$

9 Classic Hypothesis Testing

1. Assume hypothesis H_0 is true
2. Calculate test statistic $T(\boldsymbol{y}_{obs})$ based on observed sample data regarding to H_0 and H_1
3. Conditional on H_0 being true, p-value $= \mathbb{P}(T(\boldsymbol{y})$ more extreme than $T(\boldsymbol{y}_{obs})|\theta, H_0)$
4. Reject H_0 and accept H_1 for sufficiently small p-values, do not reject H_0 otherwise

10 Bayesian Hypothesis Testing

Suppose there are tow hypotheses about parameter θ:

$$H_0 : \theta \in \Theta_1 \quad H_1 : \theta \in \Theta_1$$

where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

The Bayesian approach specifies prior probabilities on each hypotheses:

$$p_0 = \mathbb{P}(H_0 \text{ is true}) = \mathbb{P}(\theta \in \Theta_0)$$
$$p_1 = \mathbb{P}(H_1 \text{ is true}) = \mathbb{P}(\theta \in \Theta_1)$$

Where the posteriors are as follows:

$$\mathbb{P}(H_0|\boldsymbol{y}) = \mathbb{P}(\theta \in \Theta_0|\boldsymbol{y})$$
$$\mathbb{P}(H_1|\boldsymbol{y}) = \mathbb{P}(\theta \in \Theta_1|\boldsymbol{y})$$

Where $\mathbb{P}(H_0|\boldsymbol{y}) + \mathbb{P}(H_1|\boldsymbol{y}) = 1$

Simple

- Single parameter values $H_0 : \theta = \theta_0 \quad H_1 : \theta = \theta_1$
- Posteriors:

$$
\begin{aligned}
\mathbb{P}(H_0|\boldsymbol{y}) &= \mathbb{P}(\theta = \theta_0|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta_0)p_0}{m(\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\theta_0)p_0}{f(\boldsymbol{y}|\theta_0)p_0 + f(\boldsymbol{y}|\theta_1)p_1} \\
\mathbb{P}(H_1|\boldsymbol{y}) &= \mathbb{P}(\theta = \theta_1|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta_1)p_1}{m(\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\theta_1)p_1}{f(\boldsymbol{y}|\theta_0)p_0 + f(\boldsymbol{y}|\theta_1)p_1} \\
\mathbb{P}(H_0|\boldsymbol{y}) &= 1 - \mathbb{P}(H_1|\boldsymbol{y})
\end{aligned}
$$

- posterior odds of H_0 against H_1:

$$
\frac{\mathbb{P}(H_0|\boldsymbol{y})}{\mathbb{P}(H_1|\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\theta_0)p_0}{f(\boldsymbol{y}|\theta_1)p_1}
$$

Composite

- Single parameter values $H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1$
- Prior: $p_i = \int_{\theta \in \Theta_i} \pi(\theta) d\theta$
- Posteriors:

$$
\begin{aligned}
\mathbb{P}(H_i|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}, H_i)}{m(\boldsymbol{y})} = \frac{p_i \ p(\boldsymbol{y}|H_i)}{m(\boldsymbol{y})} = \frac{p_i \int p(\boldsymbol{y}, \theta|H_i) d\theta}{m(\boldsymbol{y})} = \frac{p_i \int f(\boldsymbol{y}|\theta)\pi(\theta|H_i) d\theta}{m(\boldsymbol{y})} \\
&= \frac{p_i \int_{\theta \in \Theta_i} f(\boldsymbol{y}|\theta) \frac{\pi(\theta)}{p_i} d\theta}{m(\boldsymbol{y})} = \frac{\int_{\theta \in \Theta_i} f(\boldsymbol{y}|\theta)\pi(\theta) d\theta}{m(\boldsymbol{y})} = \int_{\theta \in \Theta_i} p(\theta|\boldsymbol{y}) d\theta = \mathbb{P}(\theta \in \Theta_i|\boldsymbol{y})
\end{aligned}
$$

- posterior odds of H_0 against H_1:

$$
\frac{\mathbb{P}(H_0|\boldsymbol{y})}{\mathbb{P}(H_1|\boldsymbol{y})} = \frac{\int_{\theta \in \Theta_0} f(\boldsymbol{y}|\theta)\pi(\theta) d\theta}{\int_{\theta \in \Theta_1} f(\boldsymbol{y}|\theta)\pi(\theta) d\theta} = \frac{\mathbb{P}(\theta \in \Theta_0|\boldsymbol{y})}{\mathbb{P}(\theta \in \Theta_1|\boldsymbol{y})} = \frac{\mathbb{P}(\theta \in \Theta_0|\boldsymbol{y})}{1 - \mathbb{P}(\theta \in \Theta_0|\boldsymbol{y})}
$$

Multiple models

- Set of models $M_1, ..., M_K$
- H_i: The correct model is M_i

$$
\mathbb{P}(H_i|\boldsymbol{y}) = \frac{\mathbb{P}(H_i, \boldsymbol{y})}{\mathbb{P}(\boldsymbol{y})} = \frac{\mathbb{P}(H_i, \boldsymbol{y})}{\sum_{j=1}^{K} \mathbb{P}(H_j, \boldsymbol{y})}
$$

Bayes Factor

- Prior odds for H_0 against $H_1 = p_0/p_1$
- Posterior odds for H_0 against $H_1 = \mathbb{P}(H_0|\boldsymbol{y})/\mathbb{P}(H_1|\boldsymbol{y})$
- Bayes Factor for H_0 against H_1:

$$BF_{01} = \frac{\text{Posterior odds for } H_0 \text{ against } H_1}{\text{Prior odds for } H_0 \text{ against } H_1} = \frac{\mathbb{P}(H_0|\boldsymbol{y})/\mathbb{P}(H_1|\boldsymbol{y})}{p_0/p_1}$$
$$BF_{10} = \frac{1}{BF_{01}}$$

BF_{ij}	Interpretation
< 3	No edvidence for H_i over H_j
> 3	Positive edvidence for H_i
> 20	Strong edvidence for H_i
> 150	Very strong edvidence for H_i

Simple H_0 vs Simple H_1

$$BF_{01} = \frac{\mathbb{P}(H_0|\boldsymbol{y})/\mathbb{P}(H_1|\boldsymbol{y})}{p_0/p_1} = \frac{(f(\boldsymbol{y}|\theta_0)p_0)/(f(\boldsymbol{y}|\theta_1)p_1)}{p_0/p_1} = \frac{f(\boldsymbol{y}|\theta_0)}{f(\boldsymbol{y}|\theta_1)}$$

Composite H_0 vs Composite H_1

$$BF_{01} = \frac{\mathbb{P}(H_0|\boldsymbol{y})/\mathbb{P}(H_1|\boldsymbol{y})}{p_0/p_1} = \frac{\left[\int_{\theta\in\Theta_0} f(\boldsymbol{y}|\theta)\pi(\theta)d\theta\right] / \left[\int_{\theta\in\Theta_1} f(\boldsymbol{y}|\theta)\pi(\theta)d\theta\right]}{p_0/p_1} = \frac{\mathbb{P}(\theta\in\Theta_0|\boldsymbol{y})/\mathbb{P}(\theta\in\Theta_1|\boldsymbol{y})}{p_0/p_1}$$

Simple H_0 vs Composite H_1

$$BF_{01} = \frac{\mathbb{P}(H_0|\boldsymbol{y})/\mathbb{P}(H_1|\boldsymbol{y})}{p_0/p_1} = \frac{[f(\boldsymbol{y}|\theta_0)p_0] / \left[p_1 \int_{-\infty}^{\infty} f(\boldsymbol{y}|\theta)\pi(\theta)d\theta\right]}{p_0/p_1} = \frac{f(\boldsymbol{y}|\theta_0)}{\int_{-\infty}^{\infty} f(\boldsymbol{y}|\theta)\pi(\theta)d\theta} = \frac{f(\boldsymbol{y}|\theta_0)}{m(\boldsymbol{y})}$$

Multipes Hypotheses

Read 5B.7 cba to type

11 Deterministic Numerical Integration

$$\int_{x_i}^{x_{i+m}} f(x)dx \approx \sum_{j=1}^{m} w_{ij} f(x_{i+j})$$

12 Monte Carlo Integration

$$\text{Expectaqtion } \mathbb{E}[\theta] = \int \theta p(\theta) d\theta$$
$$\text{Samples of } \theta : \theta_1, ..., \theta_N$$
$$\Rightarrow \text{Monte Carlo Estimate } \hat{\mathbb{E}}[\theta] = \frac{1}{N}\sum_{i=1}^{N} \theta_i \approx \mathbb{E}[\theta]$$

By the Law of Large Numbers, $\mathbb{P}(\lim_{N\to\infty} \hat{\mathbb{E}}[\theta] = \mathbb{E}[\theta]) = 1$.

Many Bayesian integrals can be viewed as expectations, such as probabilities and normalising constants.

Direct Sampling

- Target Distribution, often the posterior $p(\theta|\boldsymbol{y})$
- Integrals of interest: $\mathbb{E}[h(\theta|\boldsymbol{y})] = \int h(\theta)p(\theta|\boldsymbol{y})dy$
  - Common choises for $h(\theta)$:
    $h(\theta) = \theta$, the posterior mean
    $h(\theta) = (\theta - \mathbb{E}[\theta])^2$, the posterior variance
    $h(\theta) = I(a \leq \theta \leq b)$, $\mathbb{P}(\theta \in [a, b])$
- Samples $\theta_1, ..., \theta_N$
- Monto Carlo estimation of Integrals of interest $\mathbb{E}[h(\theta|\boldsymbol{y})] \approx \hat{\mathbb{E}}[h(\theta|\boldsymbol{y})] = \frac{1}{N}\sum h(\theta_i)$
- Monte Carlo Error $= \hat{\mathbb{E}}[h(\theta|\boldsymbol{y})] - \mathbb{E}[h(\theta|\boldsymbol{y})]$

Inverse Probability Integral Transform Method (Inverse PIT)

Continous Random Variable

- Let Y be continous random variables with cdf $F_Y(y) \in [0, 1]$
- Define new random variable $U = F_Y(Y) \sim \text{Uniform}(0, 1)$

The Inverse PIT method

- Transform a uniform variable U using the inverse CDF of Y $(g(U) = F_Y^{-1}(U))$. $g \sim Y$
- Therefore if one can evaluate F^{-1} for Y, then sample y can be generated by $y = F^{-1}(u)$, where $u \sim \text{Uniform}(0, 1)$

Discrete Random Variable

The Inverse PIT method

- Same for Continous method where:

$$F^{-1}(u) = \min_y\{y : F_Y(y) \geq u\}$$

Rejection Sampling

Target distribution $p(\theta|\boldsymbol{y})$

- Generates independent samples form envelope distribution $g(\theta)$
- Only some generated values are kept/used, while the rest are discarded
- Often used for generating univariate random variables than a multivariate random vector

General Rejection Algorithm

For target distribution $p()\theta$ and envelope $g(\theta)$, the (upper) bound M pf p/g is defined:

$$M \geq \frac{p(\theta)}{g(\theta)} \quad \forall\theta \Leftrightarrow M(g)\theta \geq p(\theta)$$

The Algorithm to generate a single value form $p(\theta)$ is as follows:

1. Generate θ^* from $g(\theta^*)$
2. Generate u form Uniform$(0, Mg(\theta^*))$
3. If $u \leq p(\theta^*)$ keep θ^* else reject and return step 1

The acceptance rate is $\frac{1}{M}$ which M is defined as above. For optimal acceptance rate, we choose $M_{opt} = \sup\left\{\frac{p(\theta)}{g(\theta)}\right\}$

Importance Sampling

Simlar to Rejection sampling. However, we adjust the values to correspond to the rightr distribution instead of rejecting them. Therefore all generated θ are kept, there's no rejection.

Suppose there is another distribution has the same support as $p(\theta|\boldsymbol{y})$, the integral then can be written as $\mathbb{E}_p[h(\theta|\boldsymbol{y})] = \mathbb{E}_g[h(\theta)w(\theta)]$ where $w(\theta) = p(\theta|\boldsymbol{y})/g(\theta)$ is the **importance ratio**.

The Monte Carlo estimats can be calculated $\hat{\mathbb{E}}_p[h(\theta|\boldsymbol{y})] \equiv \hat{\mathbb{E}}_g[h(\theta)w(\theta)] = \frac{1}{N}\sum h(\theta^i)w(\theta^i)$

Sampling Importance Re-Sampling

Algorithm for generating a samplem sixe of n

1. Choosing $N > n$. generate N samples of $\theta^i, i = 1, ..., N$ form the envolope distribution (importance sampler) $g(\theta)$
2. Calculate the importance ratio $w_i = p(\theta^i|\boldsymbol{y})/g(\theta)$, then scale by the some of weights:

$$w^{*i} = \frac{w^i}{\sum_j^N w^j}$$

3. randomly sample n θ^i's with replacement, with probabilities $w^{*1}, ..., w^{*N}$

13 Markov Chain Monte Carlo (MCMC)

Overwiew

- **Dependent Samples**: $\theta|\boldsymbol{y}$
- **Iterative Conditional Generation**: $g(\theta^i|\theta^{i-1})$
- **Burn-in**: The initial vales $\theta^1, ..., \theta^B$ are not distributed according to th target distribution $p(\theta)$, therefore would be discarded
- **Sample for Inference**: The additional $N - B$ samples $\theta^{B+1}, ..., \theta^N$ are used for inference

Markov Chain

- **State Space** S: is a set of all possible states j
- **Process** $\theta^t = j$: process at time t is in state j

Definition

$$\begin{aligned}\mathbb{P}(\theta^0 = x_0, ..., \theta^T = x_T) &= \mathbb{P}(\theta^T = x_T|\theta^0 = x_0, ..., \theta^{T-1} = x_{T-1}) \\ &\quad \times \mathbb{P}(\theta^{T-1} = x_{T-1}|\theta^0 = x_0, ..., \theta^{T-2} = x_{T-2}) \\ &\quad \times \mathbb{P}(\theta^0 = \theta_0)\end{aligned}$$

Markov Peoperty: if $\theta^t|\theta^{t-1}$ is independent to other values, then
$\mathbb{P}(\theta^T = x_T|\theta^0 = x_0, ..., \theta^{T-1} = x_{T-1}) = \mathbb{P}(\theta^T = x_T|\theta^{T-1} = x_{T-1})$. Therefore:

$$\mathbb{P}(\theta^0 = x_0, ..., \theta^T = x_T) = \prod_{t=1}^{T} \mathbb{P}(\theta^t = x_t|\theta^{t-1} = x_{t-1}) \times \mathbb{P}(\theta^0 = x_0)$$

The sequence of random variables $\theta^1, ..., \theta^K$ with Markov Property is a Markov Chain

Terminology and Notation

- **One-Step Transistion Probability** $p^t_{ij} = \mathbb{P}(\theta^t = j|\theta^{t-1} = i)$
- if $p^t_{ij} = p_{ij}\ \forall t$ then the chain is **time homogeneous** otherwise **time imhomogeneous**
- **Transistion probability matrix** $(P_t)_{ij} = p_{ij}$, where each row sums to one

  for a time homogeneous chain $P_t = P\ \forall t$

Limiting Behaviour of Markov Chain

- **Recurrent State**: state where a Markov Chain returns with probability 1
- **Nonnull State**: state where it will eventually recurrence in a finite ecpected time
- **Irreducible Markov Chain**: $\exists m \in \mathbb{N}\ s.t.\ \mathbb{P}(\theta^{m+n} = j|\theta^n = i) > 0\ \forall i$
- **Periodic Markov Chain**: $\exists m \in \mathbb{N}\ s.t.\ \exists d \in\ \ s.t.\ \mathbb{P}(\theta^{m+n} = j|\theta^n = i) > 0 \cap m/d \in \mathbb{Z}^+\ \forall i$

  otherwise **Aperiodic Markov Chain**
- **Ergodic Markov Chain**: **Irreducible**, **aperiodic**, all states are **nonnull** and **recurrent**

Marginal amd Stationary Distributoin

- **Marginal Probability Distribution** for state at time t is denoted $\boldsymbol{\pi}_t$, which is a vector of probabilities that sum to one
- $(\pi_t)_i = \boldsymbol{\pi}_t(i) = \mathbb{P}(\theta^t = i)$
- $\boldsymbol{\pi}_{t+1}^\top = \boldsymbol{\pi}_t^\top P \Leftrightarrow \boldsymbol{\pi}_{t+1} = P^\top \boldsymbol{\pi}_t$
- If $\boldsymbol{\pi}^\top P = \boldsymbol{\pi}^\top$ then $\boldsymbol{\pi}$ is a **stationa distribution**

Reversible Markov Chain

If a time homogeneous Markov Chain with tpm P and Stationary distribution $\boldsymbol{\pi}$ fufill the **Detailed Balance equation** then it is a reversible Markov Chain.

Detailed Balance equation is written $\boldsymbol{\pi}(i)p_{ij} = \boldsymbol{\pi}(j)p_{ji}$

Key Theoretical Results

If a Markov Chain with tpm P and is irreducible, aperiodic and has a stationary distribution $\boldsymbol{\pi}$, the the followings are true:

- $\boldsymbol{\pi}$ is unique, $\exists! \boldsymbol{\pi}\ s.t.\ \boldsymbol{\pi}^\top P = \boldsymbol{\pi}^\top$
- $\lim_{n\to\infty} \mathbb{P}(\theta^{t+n} = j|\theta^t = i) = \boldsymbol{\pi}(j)$
- $\boldsymbol{\pi}$ is the follows solution to the following program:

$$\sum_{i\in S} \boldsymbol{\pi}(i) = 1$$
$$\boldsymbol{\pi}(j) = \sum_{i\in S} \boldsymbol{\pi}(i)p_{ij}$$
$$\text{Subject to : } \boldsymbol{\pi}(j) \geq 0\ \forall j \in S$$

14 Metropolis-Hasting Algorithm

Steps

Iteration t gievn sample Θ^{t-1}:

1. Generate candidate value Θ^c from proposal distribution $q(\Theta^c|\Theta^{t-1})$
2. Calculate Metropolis Hasting Ratio (MHR) aka. accpetance rate if < 1:

$$\text{MHR}(\Theta^{t-1}, \Theta^C) = \frac{p(\Theta^c)q(\Theta^{t-1}|\Theta^c)}{p(\Theta^{t-1})q(\Theta^c|\Theta^{t-1})}$$

3. Generate $u \sim \text{Uniform}(0, 1)$
4. if $u \leq \min(1, \text{MHR}(\Theta^{t-1}, \Theta^C))$, then $\Theta^t = \Theta^c$ (Keep Θ^c) Otherwise $\Theta^t = \Theta^{t-1}$
5. Set $t = t + 1$, return step 1

Special case proposals

Random walks ($\theta^c = \theta^c + \epsilon$)

Special cases for ϵ: $\epsilon \sim \begin{cases} \text{Normal}(0, \sigma^2) \\ t_{2\ df} \\ \text{Uniform}(-b, b) \end{cases}$ where $q(\theta^c|\theta^o) = q(\theta^o|\theta^c) \Rightarrow \text{MHR}(\theta^o, \theta^c) = \frac{p(\theta^c)}{p(\theta^o)}$

Independence proposals

$$\text{MHR}(\theta^o, \theta^c) = \frac{p(\theta^c)q(\theta^o)}{p(\theta^o)q(\theta^c)}$$

Multidimentional Θ: One-at-a-time updats

let $\Theta_U^{t+1}, \Theta_0^t \subset \Theta$ be the set of updated or yet to be updated $\theta \in \Theta$. Then:

$$\text{MHR}(\theta_i^t, \theta_i^c) = \frac{p(\Theta_U^{t+1}, \theta_i^c, \Theta_0^t \backslash \{\theta_i^c\})q(\theta_i^t|\Theta_U^{t+1}, \theta_i^c, \Theta_0^t \backslash \{\theta_i^c\})}{p(\Theta_U^{t+1}, \Theta_0^t)q(\theta_i^c|\Theta_U^{t+1}, \Theta_0^t)}$$

15 MCMC Diagnostics

Notations

- **Burn-in Length** B
  The idea burn-in length B is the point where the chain is converged (aka. future values will be in the limiting distribution)
- **Trace plots/Sample paths**
  A time series plot of chains of generated values

Multiple Chains

The reason to plot multiple chains for a single random values is that a random variable θ may have multiple modes $\theta_1, ...$, therefore we would like to capture as many as possible. A plot of such chains does not prove convergence, but instead can indicates a lack of convergence.

Let's say variables $\theta_1, ..., \theta_K$ converge at $B = \{B_1, ..., B_K\}$ respectively, then we say any sample length $N \geq \max B$ would indicates a "good mixing"

A more quantitive assesement of a good mixing/lack of convergence is the **Brooks-Gelman-Rubin** (BGR) statistics, and is defined as below:

$$R = \frac{\text{Width of 80\% credible interval of all chains combined}}{\text{Average width of 80\% credible interval for each chain}}$$

If the chain has converged, $R \approx 1$. If the chains have yet been converged then $R > 1$

Variance of Monte Carlo Estimate $\hat{\mathbb{E}}[\theta]$

Notation

- $\gamma_k = \text{Cov}\left(\theta^i, \theta^{i+k}\right)$: autocovariance of lag $k \geq 0$ of the chain
- $\sigma^2 = \gamma_0$: variance if the chain
- $\rho_k = \gamma_k / \sigma^2$: autocorrelation of lag k
- τ_n^2 / n: Variance of $\hat{\mathbb{E}}[\theta]$

$$\begin{aligned}
\text{Var}\left[\hat{\mathbb{E}}[\theta]\right] &= \frac{\tau_n^2}{n} \\
\tau_2^2 &= \sigma^2 \left[1 + 2\sum_{k=1}^{n} \frac{n-k}{n}\rho_k\right] \quad \text{(Ineffiency factor/ Integrated autocorrelation)}
\end{aligned}$$

As $n \to \infty$:

$$\tau_2^2 = \sigma^2 \left[1 + 2\sum_{k=1}^{n} \frac{n-k}{n}\rho_k\right] \approx \sigma^2 \left(1 + 2\sum_{k=1}^{\infty} \rho_k\right) \Rightarrow \text{Var}\left[\hat{\mathbb{E}}[\theta]\right] \approx \frac{\sigma^2\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)}{n}$$

Effective Sample Size, ESS

$$n_{eff} = \frac{n}{1 + 2\sum_{k=1}^{n} \rho_k} \Rightarrow \text{Var}\left[\hat{\mathbb{E}}\left[\theta\right]\right] \approx \frac{\sigma^2}{n_{eff}}$$

Autocorrelation Plots

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\sigma^2}} = \frac{\sum_{i=1}^{n-1}\left(\theta^i - \bar{\theta}\right)\left(\theta^{i+k} - \bar{\theta}\right)}{\sum_{i=1}^{n}\left(\theta^i - \bar{\theta}\right)^2}, \quad \bar{\theta} \text{ is the sample average}$$

Central Limit Theorem for $\hat{\mathbb{E}}\left[\theta\right]$

$$\hat{\mathbb{E}}\left[\theta\right] \sim \text{Asymptotically Normal}\left(\mathbb{E}\left[\theta\right], \frac{\sigma^2}{n_{eff}}\right) \Rightarrow \frac{\hat{\mathbb{E}}\left[\theta\right] - \mathbb{E}\left[\theta\right]}{\sigma/\sqrt{n_{eff}}} \sim \text{Normal}(0, 1)$$

Estimating Var $\left[\hat{\mathbb{E}}\left[\theta\right]\right]$

Batching is a popular method for estimating Var $\left[\hat{\mathbb{E}}\left[\theta\right]\right]$, whe algorithm is as below:

1. Partition chain of length n into m batches of $T \in [10, 30]$ successive values
2. For each batch calculate the mean of the values $\bar{\theta}_i$
3. Estimate of Var $\left[\hat{\mathbb{E}}\left[\theta\right]\right]$ is then calculated:

$$\frac{\hat{\tau^2}}{n} = \frac{\frac{T}{m-1}\sum_{i=1}^{m}\left(\bar{\theta}_i - \bar{\theta}\right)^2}{n}$$

Improve Performance

- Changing proposal distribution
- Reparameterising model
- Blocking

Gibbs Sampler

Can be seen as a special case of a Metropolis-Hasting sampler with two features:

- The proposal distributions are conditional distributions for θ's
- All candidates values are accepted ($\text{MHR}(\theta_i^t, \theta_i^c) = 1$)

The Algorithm

1. Let $\Theta = (\theta_1, ..., \theta_q) \sim p(\Theta)$
2. For $i \in 1, .., q$ denote thr full conditional distribution for θ_i by

$$p(\theta_i|\theta_{-i}) = p(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_q)$$

3. Initialise chain $\Theta^0 = (\theta_1^0, ..., \theta_q^0)$
4. At iteration $t+1$ given Θ^t, Θ^{t+1} is generated as follows:
   - Generate θ_1^{t+1} from $p(\theta_1|\theta_2^t, ..., \theta_q^t)$

     $\vdots$
   - Generate θ_i^{t+1} from $p(\theta_i|\theta_1^{t+1}, ..., \theta_{i-1}^{t+1}, \theta_{i+1}^t, ..., \theta_q^t)$

     $\vdots$
   - Generate θ_q^{t+1} from $p(\theta_i|\theta_1^{t+1}, ..., \theta_{q-1}^{t+1})$
5. The (joint) transistion probability of going from Θ^t to Θ^{t+1} is:

$$\mathcal{K}_{Gibbs}(\Theta^t, \Theta^{t+1}) = \prod_{i=1}^{q} p(\theta_i^{t+1}|\theta_j^{t+1}, j < i \ \cap \ \theta_j^t, j > i)$$

   An the stationary distribution is the target $p(\Theta)$

Gibbs Sampler vs Metropolis-Hasting

Gibbs Sampler

- Strengths
  - Automatically defined proposal dist.
  - Accepts all candidate values
  - views as an "adaptive algorithm"
- Weaknesses
  - Conditional distribution may not be tractable
  - Sampling may be computationally intensive
  - 100% acceptance $\neq$ good mizing

Metropolis Hasting

- Strengths
  - flexible proposal dist., can be fast sample from, fast to evaluate MHR
  - Don't need to know conditional dist.
  - easily block updating
- Weaknesses
  - Hard to find good proposal with good mixing
  - Can take time to tune a proposal

Distrubution Table

Discrete Distributions

θ ~Distrubution	Mass function	Mean	Variance
Binimial(n, p)	$f(\theta) = \binom{b}{\theta} p^{\theta}(1-p)^{n-\theta}$	np	$np(1-p)$
Poisson(λ)	$f(\theta) = \dfrac{\lambda^{\theta} \exp -\lambda}{\theta!}$	λ	λ
Geometric(p)	$f(\theta) = p(1-p)^{\theta-1}$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
Negative Binomial(α, β)	$f(\theta) = \binom{\theta+\alpha-1}{\alpha-1} \left(\dfrac{\beta}{\beta+1}\right)^{\alpha} \left(\dfrac{1}{\beta+1}\right)^{\theta}$	$\dfrac{\alpha}{\beta}$	$\dfrac{\alpha}{\beta^2}(\beta+1)$
Multinomial$(n, \boldsymbol{p})$	$f(\boldsymbol{\theta}) = \dfrac{n!}{\prod_{i=1}^{k} \theta_i!} \prod_{i=1}^{k} p_i^{\theta_i}$	$n\boldsymbol{p}$	$\text{Var}[\theta_i] = np_i(1-p_i)$

Noted that:

- $\binom{n}{k} = \dfrac{n!}{k!(n-k!)}$

Continous Dostribution

θ ∼Distrubution	Density function	Mean	Variance
Uniform$[a,b]$	$f(\theta)=\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal(μ,σ^2)	$f(\theta)=\frac{\exp\left(-\frac{(\theta-\mu)^2}{2\mu^2}\right)}{\sqrt{2\pi\sigma^2}}$	μ	σ^2
log Normal(μ,σ^2)	$f(\theta)=\frac{\exp\left(-\frac{(\log\theta-\mu)^2}{2\mu^2}\right)}{\theta\sqrt{2\pi\sigma^2}}$	$\exp\left(\mu+\frac{\sigma^2}{2}\right)$	$\exp(2\mu+\sigma^2)(\exp(\sigma^2)-1)$
$\beta(\alpha,\beta)$	$f(\theta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Exponential(λ)	$f(\theta)=\lambda\exp(-\lambda\theta)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$\Gamma(\alpha,\beta)$	$f(\theta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta)$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
$\Gamma^{-1}(\alpha,\beta)$	$f(\theta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}\exp\left(-\frac{\beta}{\theta}\right)$	$\frac{\beta}{\alpha-1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$
χ^2_v	$f(\theta)=\frac{2^{-v/2}}{\Gamma(v/2)}\theta^{\frac{v}{2}-1}\exp\left(-\frac{\theta}{2}\right)$	v	$2v$
χ^{-2}_v	$f(\theta)=\frac{2^{-v/2}}{\Gamma(v/2)}\theta^{-\left(\frac{v}{2}+1\right)}\exp\left(-\frac{1}{2\theta}\right)$	$\frac{1}{v-2}$	$\frac{2}{(v-2)^2(v-4)}$
Dirichlet$(\alpha_1,...\alpha_k)$	$f(\theta)=\frac{\Gamma(\alpha_0)}{\prod_{i=1}^k\Gamma(\alpha_i)}\prod_{i=1}^k\theta_i^{\alpha_i-1}$	$\mathbb{E}[\theta_i]=\frac{\alpha_i}{\alpha_0}$	$\mathrm{Var}[\theta_i]=\frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$

Noted that for Dirichlet distribution, $\alpha_0=\sum_{i=1}^k\alpha_i$