Bayesian Data Analysis

by Gelman

Page

Uncertainity

1) Aleatoric Uncertainity : Uncertainity is due to randomness in the outcome.

2) Epistemic Uncertainity : Uncertainity arises due to lack of knowledge.

$P(y|\theta) \rightarrow$ Liklihood
$P(\theta|y) \rightarrow$ Posterior

Book : Bayesian Data Analysis (CRC press)
- Gelman et al

Doing Bayesian Data Analysis
- John Kruske.

Assessment

End Sem - 45%.
Mid Sem - 30%.
Quiz + Assignment - 20% + 5% (Tuning).

Bayesian Inference

- Process of fitting a probability model to a set of data and summarizing the results by a prob. distribution over the parameters of the model and on the unobserved quantities. like prediction outcome.

Date
Page

Process of Bayesian Data Analysis (BDA)

1) Setting up a full model (& its outcome)
   ↳ Joint probability distribution of the observed and the unobserved quantities in the ~~data~~ problem. θ.

2) Conditioning on the observed data $P(\theta|y)$, $P(\tilde{y}|y)$

3) Evaluating the fit of the model and the implications of the posterior distribution.

Notation

$\theta \rightarrow$ Unobservable vector quantities or model parameters.

$y \rightarrow$ Denotes the observed data.

$\tilde{y} \rightarrow$ Unknown but potentially observed quantities

Generally
- Greek letter will denote the model parameter
- Lower case Roman letters will denote the observed or observable scalars or vectors
- Upper case Roman letters will denote the observed or observable matrix
- Vectors would be represented as a column vector, $u^T v$ would be a scalar and uu^T is a matrix

Date
Page

Exchangeability

Ex:- H, T, H, H, H, T ....

Suppose θ is known

$P(H,T,H,H,H,T) = P(H)\,P(T)\cdot P(H)\cdot\ldots$

Exchangable If the sequence is permuted the probability will not change

$= \theta^{\#\text{heads}}(1-\theta)^{\#\text{fails}}$

- Consider the size of tumors observed every 1 year interval, $S_1, S_2, S_3 \ldots$

$P(S_1, S_2, S_3, S_4)$ } Non Exchangable Model
$P(S_1, S_3, S_4, S_2)$ }

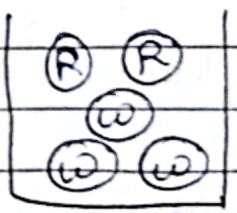
Note Exchangeability is related with iid.

iid $\Rightarrow$ Exchangeability.

BUT

Exchangeability $\not\Rightarrow$ iid.

$\Rightarrow$ Polya's Urn



Taking out a ball without replacement

$P(R,W,R,W,W) =$

$\frac{2}{5}\times\frac{3}{4}\times\frac{1}{3}\times 1\times 1 = \frac{1}{10}$

$P(W,R,W,W,R) = \frac{3}{5}\times\frac{2}{4}\times\frac{2}{3}\times\frac{1}{2}\times 1 = \frac{1}{10}$

This is exchangable but not iid as the next term depends on previous one.

Date
Page

Explanatory Variables

The pair (x, y); is exchangeable.

Bayes Rule of Inference

$$\theta = \frac{\#\text{ Red Marbles}}{\text{Total Marbles}}$$

$$P(y, \theta) = P(\theta|y) \cdot P(y) = P(y|\theta) \cdot P(\theta)$$

$$P(\theta|y) = \frac{P(y|\theta) \cdot P(\theta)}{P(y)}$$

($P(y|\theta)$ → Likelihood; $P(\theta)$ → Prior (Not dependent on past data))

$P(y)$ → Marginal $\sum_{\theta} P(y|\theta) P(\theta)$

$$P(\theta|y) \propto P(y|\theta) P(\theta)$$

Predictive Inference

To make prediction about an unknown observable.

$$p(y) = \int_{\theta} p(y, \theta)\, d\theta = \int_{\theta} p(y|\theta) p(\theta)\, d\theta$$

$p(y)$ → Prior predictive Distribution.

↳ Prior because no past data is used
Predictive because its making a prediction on unknown variable.

Date
Page

$P(\tilde{y}|y)$ → Suppose an object is weighed on a scale n times, $y_1, y_2 \cdots y_n$

Posterior Predictive Distribution.

Let the actual weight be μ and the variance in the scale be σ^2

Let the unknown parameters denoted by θ

$$P(\tilde{y}|y) = \int_\theta P(\tilde{y}, \theta|y)$$

$$= \int_\theta P(\tilde{y}|\theta, y)\, P(\theta|y) \cdot d\theta$$

$$= \int_\theta P(\tilde{y}|\theta)\, P(\theta|y) \cdot d\theta$$

Posterior Odd Ratio

$$\frac{P(\theta_1|y)}{P(\theta_2|y)} = \frac{\frac{P(y|\theta_1) \cdot P(\theta_1)}{P(y)}}{\frac{P(y|\theta_2) P(\theta_2)}{P(y)}} = \frac{P(\theta_1)}{P(\theta_2)} \times \boxed{\frac{P(y|\theta_1)}{P(y|\theta_2)}}$$

$\frac{P(\theta_1)}{P(\theta_2)}$ ↓ Prior Odds

$\frac{P(y|\theta_1)}{P(y|\theta_2)}$ ↓ Likelihood Ratio.

If θ changes a little $P(y|\theta_1) \cdot P(\theta)$ changes largely.

Problems related to these.

1) Prior may not be available all the time
2) Modeling the likelihood with a distribution
3) $P(y)$. θ can have exponential space and could have multiple θs.

Date
Page

Case Study : Autocorrect and Autocompletion

If someone writes "radom"
The possibilities are : → "random"
→ "radon"
→ "radom"

What is the probability that the person has used the word correctly.

Let y be the word the person has written
Here, y = radom

Let θ be the word that the person is actually supposed to write.

θ — θ_1 = "random"
θ — θ_2 = "radon"
θ — θ_3 = "radom"

$$P(\theta|y) \propto P(y|\theta) P(\theta)$$

Prior Information : Finding the frequency of θ in search keywords

θ	$P(\theta)$
random	7.6×10^{-5}
radon	6.5×10^{-6}
radom	3.12×10^{-7}

Date
Page

Liklihood: $P(y|\theta)$ Can be obtained from spell check

θ	$P(y\|\theta) = p(\text{radom}\|\theta)$
random	0.00193
radon	0.000143
radam	0.975

Posterior

θ	$P(y\|\theta)P(\theta)$	$P(\theta\|y)$
random	1.47×10^{-7}	0.325
radon	8.6×10^{-10}	0.002
radam	3.04×10^{-7}	0.674

We can include the context words and model it like $P(\theta|\tilde{y}, y)$
We can also add sentiments to improve results.

Useful results from Probability Theory

$p(u,v)$ → Joint probability distribution
$p(u|v)$ → Conditional probability distribution
$p(u)$ → Marginal probability distribution ($\int_v p(u,v)dv$

$$p(u,v,w) = p(u,v|w)p(w) = p(u|v,w)p(v|w)p(w)$$

↳ Ex:- An urn contains n_r red balls, n_g green balls and n_b blue balls. We pick three balls without replacement. What is the probability that they are of different color.

Date
Page

Mean

$$E(u) = \int_u u\, p(u)\, du \quad \text{or} \quad \sum_u u\, p(u)$$

Variance

$$Var(u) = \int_u (u - E(u))^2 p(u)\, du = E[u - E(u)]$$

Covariance

$$Cov(u) = \int_u (u - E(u))(u - E(u))^T du, \quad u \text{ is a column vector}$$

Modeling using conditional probability

→ Probability models often express the distribution conditionally or hierarchically

→ The height y of a student selected random in a class

Avg. height of boys = 170
Avg. height of girls = 160

If $p(\text{boys}) = p(\text{girls}) = 1/2$

$$p(y) = p(y|\text{boy})\, p(\text{boy}) + p(y|\text{girl})\, p(\text{girl})$$

↑ Marginal probability with bimodal distribution.

Conditional Mean

$$E(u) = \int_u u\, p(u)\, du = \iint_{u,v} u\, p(u,v)\, du\, dv$$

$$E(u) = \iint_{u\,v} u\, f(u|v) f(v)\, du\, dv$$

$$= \int_{v} \int_{u} \underbrace{u\, f(u|v)\, du}_{\text{Conditional mean}}\ f(v)\, dv$$

$$= \int_{v} E(u|v)\, f(v)\, dv$$

$$= E[E(u|v)]$$

Conditional Variance

$$\text{Var}(u) = E(\text{var}(u|v)) + \text{var}(E(u|v))$$

$$E(\text{var}(u|v)) + \text{var}(E(u|v))$$

$$= E(E(u^2|v) - (E(u|v))^2) + E[(E(u|v))^2] - (E(E(u|v)))^2$$

$$= E(E(u^2|v)) - E((E(u|v))^2) + E[(E(u|v))^2] - (E(E(u|v)))^2$$

$$= E(E(u^2|v)) - (E(E(u|v)))^2$$

$$= E(u^2) - (E(u))^2$$

$$= \text{Var}(u)$$

Date
Page

Q Let X be a discrete random variable with support $S = \{0,1\}$ and let Y be a discrete Random variable with support $S_2 = \{0,1,2\}$. The joint distribution $p_{X,Y}(x,y)$ as is given as follows.

$p_{X,Y}(x,y)$ $X \downarrow$ $Y \rightarrow$	0	1	2	$p_X(x)$
0	1/8	2/8	1/8	1/2
1	2/8	1/8	1/8	1/2

1) Find the conditional mean of $Y|X$.

Sol^n ~~$E(Y|X) = (1/8 + 2/8 + 1/8)\cdot\frac{1}{2} + (2/8 + 1/8 + 1/8)\cdot\frac{1}{2}$~~

$$E(Y|X) = \sum_y y\, p(Y|X)$$

$$E(Y|X=0) = \sum_y y\, p(y|x=0)$$

$$= 0\times p(y=0|x=0) + 1\, p(y=1|x=0) + 2\, p(y=2|x=0)$$

$$p_{X,Y}(x,y) = p(x|y)p(y)$$

$$p(y|x)\cdot p(x) \quad \text{or} \quad p(y|x) = \frac{p_{X,Y}(x,y)}{p(x)}$$

$$E(Y|X=0) = 1\times\frac{2/8}{1/2} + 2\,\frac{1/8}{1/2} = 1$$

$$E(Y|X=1) = 0\, p(y=0|x=1) + 1\, p(y=1|x=1) + 2\, p(y=2|x=1)$$

$$\frac{1/8}{1/2} + 2\,\frac{1/8}{1/2} = \frac{3}{4}$$

2) Find the conditional variance of $Y|X$

Sol^n $\text{Var}(Y|X) = \mathbb{E}(Y^2|X) - (\mathbb{E}(Y|X))^2$

$\sum_y y^2 f(Y|X)$

$0\, f(Y|X=0) + 1\, f(Y \neq X=0) + 4\, f(Y \neq X=0)$

$\frac{1}{2} +$

Transformation of Variables

Suppose u is a random variable

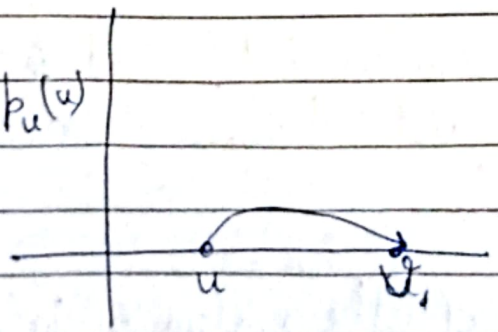
$f_u(u) \rightarrow$ Distribution of the Random Variable u.

$v = f(u)$

$f_v(v) \rightarrow$ Distribution of v we want to find.

Case-1 $f_u(u)$ is discrete and v is a one to one mapping of u, $v = f(u)$

$$f_v(v) = f_u(f^{-1}(v))$$



Q Suppose X is a geometric distribution with $f_x(x)$

$$f_x(x) = \frac{3}{4}\left(\frac{1}{4}\right)^{x-1}, \quad x \in \mathbb{Z}$$

Find the probability distribution of the R.V; $Y = X^2$

Case 1 [illegible]

Date
Page

Sol"

$$Y = X^2$$
$$X = \sqrt{Y}$$

$$f_Y(y) = \begin{cases} \frac{3}{4}\left(\frac{1}{4}\right)^{\sqrt{y}-1} & \text{for } Y = 1, 4, 9 \ldots \\ 0 & \text{otherwise} \end{cases}$$

Case 1 (a) Suppose X_1 and X_2 are 2 discrete random variables with a joint probability distribution $f_{X_1 X_2}(x_1, x_2)$ and Let Y_1 be a function $Y_1 = u_1(x_1, x_2)$ and $Y_2 = u_2(x_1, x_2)$

One to One Transformation

So we can Solve x_1 and x_2 in terms of y_1 and y_2

$$x_1 = w_1(y_1, y_2)$$
$$x_2 = w_2(y_1, y_2)$$

Find distribution $f_{Y_1, Y_2}(y_1, y_2)$

$$f_{Y_1, Y_2}(y_1, y_2) = f_{X_1, X_2}\left(w_1(y_1, y_2), w_2(y_1, y_2)\right)$$

Q Suppose X_1 and X_2 be two random independent random variable with Poisson distibution. with mean μ_1, μ_2

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Find the distribution of the random variable

$$Y = X_1 + X_2$$

Sol"

$$Y = X_1 + X_2$$
$$V = X_1$$

$X_1 = V$ & $X_2 = Y - V$

$$f_{X_1,X_2}(x_1,x_2) = \frac{\lambda^{x_1+x_2} e^{-\lambda}}{x_1!\,x_2!}$$

$$f_{X_1,X_2}(x_1,x_2) = \frac{M_1^{x_1} e^{-M_1}}{x_1!} \cdot \frac{M_2^{x_2} e^{-M_2}}{x_2!}$$

$$f_{Y,V}(y,v) = \frac{M_1^{v} e^{-M_1}}{v!} \frac{M_2^{y-v} e^{-M_2}}{(y-v)!}$$

$$f_Y(y) = \sum_v \frac{M_1^{v}}{v!} \frac{M_2^{y-v}}{(y-v)!} e^{-M_1-M_2}$$

$$\frac{e^{-(M_1+M_2)} M_2^{y}}{y!} \sum_v \left(\frac{M_1}{M_2}\right)^{v} \frac{y!}{v!(y-v)!}$$

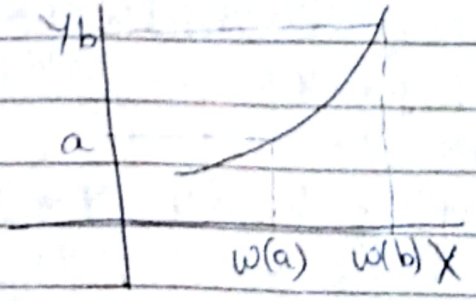
$$\frac{e^{-(M_1+M_2)} M_2^{y}}{y!} \sum_v {}^{y}C_v \left(\frac{M_1}{M_2}\right)^{v}$$

$$\frac{e^{-(M_1+M_2)} M_2^{y}}{y!} \sum {}^{y}C_v M_1^{y-v} M_2^{-v}$$

$$\frac{e^{-(M_1+M_2)} (M_1+M_2)^{y_1}}{y_1!} \sim f(M_1+M_2)$$

Case-2 Suppose X is the continuous random variable, $f_x(x)$ being its distribution and let $Y = u(X)$ define a one to one mapping from X to Y. So $X = w(Y)$. Then the probability distribution $f_Y(y) = f_x(w(y))|w'(y)|$ where $J = w'(y)$, Jacobian.

Proof



$$P(a<Y<b) \int_{w(a)}^{w(b)} f_x(x)\,dx.$$

Date
Page

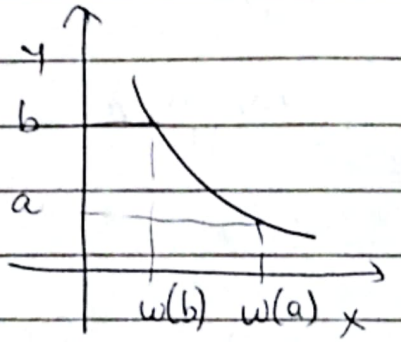
$$= \int_a^b f_x(w(y))$$

$$P_Y(a<Y<b) = \int_a^b f_y(y)\,dy$$

$$= \int_{w(a)}^{w(b)} f_x(w(y))\,dx \qquad \begin{matrix} x = w(y) \\ dx = w'(y)\,dy \end{matrix}$$

$$= \int_a^b f_x(w(y))\,w'(y)\,dy$$

$$f_x(y) = f_x(w(y))\,|w'(y)|.$$



Decreasing function.

$$\int_a^b f_y(y)\,dy$$

$$= -\int_{w(b)}^{w(a)} f_x(w(y))\,dx.$$

$$= -\int_a^b f_x(w(y))\,w'(y)\,dy.$$

($w'(y) < 0$; makes (+)ve.)

$$f_x(y) = f_x(w(y))\,|w'(y)|.$$

Q Let X be a continuous random variable with prob. dist^n $f_x(x) = \begin{cases} x/12 & \text{for } 1<x<5 \\ 0 & \text{otherwise.} \end{cases}$

Find the probability dist^n of the R.V $Y = 2x-3$

Sol^n $f_Y(y) = \begin{cases} \frac{y+3}{24} \cdot \frac{1}{2} & \text{for } 1 < \frac{y+3}{2} < 5 \\ 0 & \text{otherwise} \end{cases}$

Date
Page

$$f_Y(y) = \begin{cases} \frac{y+3}{48} & -1 < y < 7 \\ 0 & \text{Otherwise} \end{cases}$$

Case-3 Suppose X is a vector of continuous random variable and $Y = u(X)$ with one to one mapping $X = w(Y)$

$$f_Y(y) = f_X(w(Y))|J|$$

where J is a matrix

Note $\frac{\partial Y}{\partial X}$, $Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$ $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

$$\frac{\partial Y}{\partial X} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_2} \\ \vdots & & & \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix} = J$$

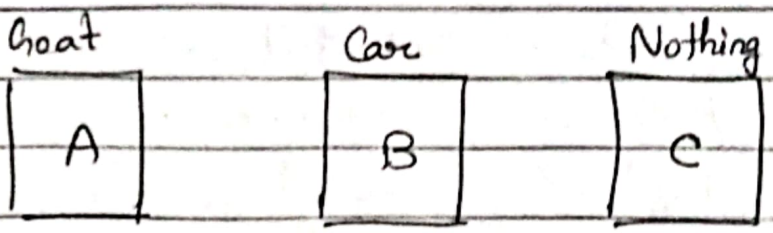
Case-4 If X is a continuous random variable with probability distribution given as $f_X(x)$ and $Y = U(X)$ defines a transformation between the value of X and Y that is not one-to-one. If the interval over which X is defined can be partitioned into k mutually disjoint sets such that each of the inverse functions $x_1 = w_1(y)$, $x_2 = w_2(y) \ldots x_n = w_n(y)$ that defines a one to one correspondence then the $f_Y(y)$ can be defined as

$$f_Y(y) = \sum_{i=1}^{k} f_X(w_i(y))|J_i| \quad \text{where } |J_i| = |w_i'(y)|$$

$i = 1, 2, 3 \ldots k$

Date
Page

Q



"Lets Make a Deal" / "The Monte-Hall" Problem.
When the participant chooses a door some other non-prize door is opened then he is being given a chance to switch his choice.

A is chosen by participant.

$P(A) \xrightarrow{1/3}$ Likelihood if A contains prize host to open B.

$P(B=\text{opened} \mid A) = 1/2$

$P(B) \xrightarrow{1/3} P(B=\text{opened} \mid B) = 0$

$P(C) \xrightarrow{1/3} P(B=\text{opened} \mid C) = 1$

A is chosen and C has the prize so opening of B is certain.

$$P(\text{prize in A} \mid B \text{ opened}) = \frac{P(B=\text{open} \mid A) \cdot P(A)}{\sum_{A,B,C} P(B=\text{open} \mid A) \cdot P(A)}$$

$$= \frac{1/2 \cdot 1/3}{1/2 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{1}{3}$$

$P(B \mid B) = 0$

$P(C \mid B=\text{opened}) = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{2}{3}$

Posterior: Prob. prize is in C / Given B is opened by the host.

Date
Page

Distributions

$$P(y|\theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

↓ Likelihood

↳ Binomial Distribution.

Posterior: $P(\theta|y, n) = \frac{P(y|\theta) \cdot P(\theta)}{\int_{\theta=0}^{1} P(y|\theta) \cdot P(\theta)\, d\theta}$

$\propto P(y|\theta) \cdot P(\theta)$ → Prior (can be anything uniform)

$\propto P(y|\theta)$

$= \binom{n}{y} \theta^y (1-\theta)^{n-y}$

$$\boxed{\beta(\alpha, \beta|\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}}$$

$P(\theta|y, n) \propto \beta(y+1, n-y+1, \theta)$

$$\boxed{\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)!}$$
$$\boxed{\Gamma(1) = 1}$$

$$P(\theta|y, n) = \frac{n!}{y!(n-y)!} \theta^y (1-\theta)^{n-y}$$

$$= \frac{n(n-1)!}{y!(n-y)!} \theta^y (1-\theta)^{n-y}$$

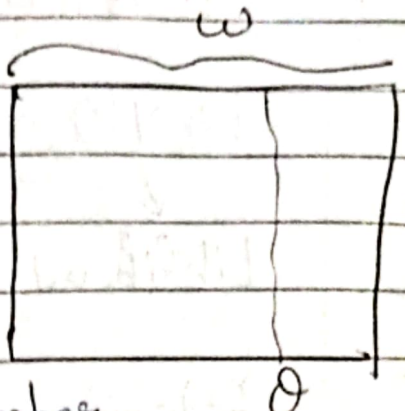
$$= \frac{n\Gamma(n)}{y\Gamma(y)(n-y)\Gamma(n-y)} \theta^y (1-\theta)^{n-y}$$

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

$$\propto \frac{n}{y(n-y)} \beta(\theta, y+1, n-y+1)$$

Date
Page

Bayes Experiment



1) A ball w was thrown randomly θ is the fraction of the width on which it dropped.

2) Another ball O was dropped n number of times. The number of success 'Y' in the count of the times of the ball landed on the right of θ.

$$P(Y|\theta,n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

What he tried to observe $P(\theta \in (\theta_1, \theta_2)) | y)$

$$= \frac{\int_{\theta_1}^{\theta_2} P(Y|\theta)\, P(\theta)\, d\theta}{\int_0^1 P(Y|\theta)\, P(\theta)\, d\theta}$$

Assumptions:

→ θ is uniformly distributed in the range [0,1]

→ The denominator

$$\int_0^1 P(y|\theta)\cdot P(\theta)\, d\theta$$

$$\propto \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y}\, d\theta \quad \{\because P(\theta) \text{ is uniform for all}\}$$

Date
Page

$$= \binom{n}{y} \int_0^1 \theta^y (1-\theta)^{n-y} d\theta = I_n \text{ (say)}$$

$$\int uv\,dx = u\int v\,dx + \int\left[u' \int v\,dx\right] dx$$

$$I_n = \binom{n}{y}\left[\theta^y \int (1-\theta)^{n-y} d\theta + \int \left(y\theta^{y-1} \int (1-\theta)^{n-y} d\theta\right) d\theta\right]$$

$$\binom{n}{y}\left[\frac{-\theta^y (1-\theta)^{n-y+1}}{n-y+1}\right]_0^1 + \int y\theta^{y-1} \frac{(1-\theta)^{n-y+1}}{n-y+1} d\theta$$

Applying limits.

$$\binom{n}{y} \int_0^1 y \frac{\theta^{y-1}(1-\theta)^{n-y+1}}{n-y+1} d\theta$$

$$\binom{n}{y}\left(\frac{y}{n-y+1}\right)\left[\theta^{y-1} \int (1-\theta)^{n-y+1} d\theta + \int \frac{(y-1)\theta^{y-2}(1-\theta)^{n-y+2}}{n-y+2}\right]$$

$$\binom{n}{y}\left(\frac{y}{n-y+1}\right)\left[\frac{-\theta^{y-1}(1-\theta)^{n-y+2}}{n-y+2} + \int \frac{y-1}{n-y+2} \theta^{y-2}(1-\theta)^{n-y+2}\right]$$

$$+\binom{n}{y}\left(\frac{y}{n-y+1}\right)\left(\frac{y-1}{n-y+2}\right) \cdots \cdots \left(\frac{1}{(n+1)}\right) \int_0^1 (1-\theta)^n d\theta$$

$$\binom{n}{y} \frac{y!}{(n-y+1)(n-y+2)\cdots n} \left[\frac{-(1-\theta)^{n+1}}{n+1}\right]_0^1$$

$$\binom{n}{y} \frac{y!}{(n-y+1)(n-y+2)\cdots n(n+1)}$$

$$\frac{n!}{y!(n-y)!} \cdot \frac{y!}{(n-y+1)\cdots n(n+1)} = \frac{n!}{(n+1)!} = \frac{1}{n+1}$$

Date
Page

Q What is the probability of getting a success if he already got y successes.

Sol?

$$P(\tilde{y}=1|y) = \int_0^1 P(\tilde{y}=1, \theta|y)\, d\theta$$

$$\int_0^1 P(\tilde{y}=1|\theta, y) \cdot P(\theta|y)\, d\theta$$

$$\int_0^1 P(\tilde{y}=1|\theta) \cdot P(\theta|y)\, d\theta$$

$$\int_0^1 \theta\, P(\theta|y)\, d\theta = \mathbb{E}[\theta|y]$$

$$\frac{\int_0^1 \theta\, P(y|\theta) \cdot P(\theta)\, d\theta}{\int_0^1 P(y)\, d\theta}$$

Posterior Predective Distribution.

$$P(\tilde{y}=1|y) = \int_0^1 P(\tilde{y}, \theta|y)\, d\theta$$

$$= \int_0^1 P(\tilde{y}|\theta, y) \cdot P(\theta|y) \cdot d\theta$$

$$= \int_0^1 P(\tilde{y}|\theta) \cdot P(\theta|y)\, d\theta$$

$$= \int_0^1 \theta\, P(\theta|y)\, d\theta$$

$$= \int_0^1 \theta\, \frac{P(y|\theta) \cdot P(\theta)}{P(y)}\, d\theta$$

$$\propto \int_0^1 \theta\, \frac{P(y|\theta)\, d\theta}{\frac{1}{n+1}}\, d\theta$$

$$= (n+1)\int_0^1 \theta \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

$$= (n+1)\binom{n}{y}\int_0^1 \theta^{y+1}(1-\theta)^{n-y}$$

$$I_{n-y,y+1} = \int_0^1 (1-\theta)^{n-y}\theta^{y+1} d\theta$$

$$= (1-\theta)^{n-y}\int_0^1 \theta^{y+1} + \int (n-y)(1-\theta)^{n-y-1}\int \theta^{y+1} d\theta$$

$$\left[\frac{(1-\theta)^{n-y}\theta^{y+2}}{y+2} + \int_0^1 \frac{n-y}{y+2}(1-\theta)^{n-y-1}\theta^{y+2} d\theta\right.$$

$$= \frac{n-y}{y+2}\int_0^1 (1-\theta)^{n-y-1}\theta^{y+2} d\theta.$$

$$= \frac{n-y}{y+2} I_{n-y+1,y+2}$$

$$= \left(\frac{n-y}{y+2}\right)\left(\frac{n-y-1}{y+3}\right)\left(\frac{n-y-2}{y+4}\right)\cdots\frac{1}{n+1}\int_0^1 \theta^{n+1} d\theta$$

$$= \frac{(n-y)!\,(y+1)!}{(n+2)!}$$

$$(n+1)\binom{n}{y} I_{n-y,y+1} = \frac{(n+1)\,n!}{y!(n-y)!} \cdot \frac{(n-y)!\,(y+1)!}{(n+2)!}$$

$$= \frac{y+1}{n+2}$$

When $y=0$ in n trials

$$P(\tilde{y}=1|y) = 1/n+2$$

When $y=n$

$$P(\tilde{y}=1/y) = \frac{n+1}{n+2}.$$

Date
Page

The posterior is actually a compromise between the prior and the data information.

$$\mathbb{E}(u) = \int u\, P(u)\, du$$

$$= \iint_{u\,v} u\, P(u,v)\, du\, dv$$

$$\int_u \int_v u\, P(u|v) \cdot P(v)\, du\, dv$$

$$\int_v dv\, P(v) \int_u u\, P(u|v)\, du$$

$$= \int_v \mathbb{E}(u|v) \cdot P(v)\, dv$$

$$= \mathbb{E}\,\mathbb{E}(u|v))$$

Similarly, $\mathbb{E}(\theta) = \mathbb{E}(\mathbb{E}(\theta|y))$

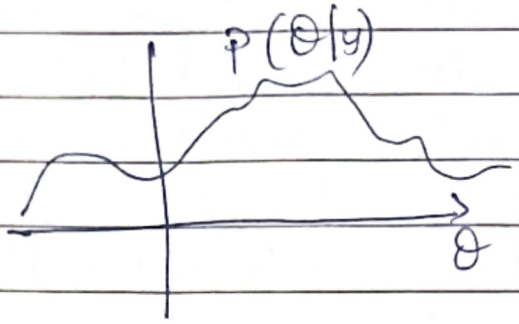
$$\text{Var}(\theta) = \mathbb{E}(\text{Var}(\theta|y)) + \text{Var}(\mathbb{E}(\theta|y))$$

↓ Expected variance of posterior

↓ Variance of the expected posterior

How to summarize the posterior inference

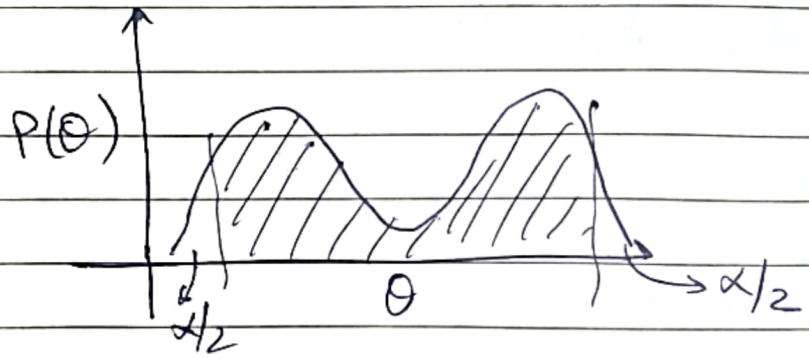
- Posterior probability distribution $P(\theta|y)$
  - Contains all the information about the current parameter θ.
- For Multiple parameter we may use a contour plot.



- Numerical summaries like mean, median and mode

- Variation, Interquartile range using Box plot.

- Posterior quantiles and intervals
  - Central interval of a posterior probability.
    → 100(1−α)%. interval.



- Highest probability density.

The part where the frequency of inner part must be larger than all the outside region.

Priors

2 basic interpretations of prior distribution.

1) Population Interpretation
   - Prior distribution represents a population of possible parameter values from which the current parameter value is drawn

$\frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

$E(\beta) = \frac{a}{a+b}$ $\quad Var(\beta) = \frac{ab}{(a+b)(a+b-1)}$

Date
Page

2) Stage of knowledge representation
- There is an actual prior distribution
- Current value represents a random realization from this distribution.

Problem : The knowledge of entire population from which θ be drawn is not possible. Ex:- Industrial failure.

⇒ Binomial example with different priors

$$p(y|\theta) \propto \theta^{y}(1-\theta)^{n-y}$$

If we assume prior also to be of the same form.

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \propto Beta(\alpha, \beta)$$

This is similar to the probability of $\alpha-1$ successes in $\alpha+\beta-2$ trials.
α, β are hyperparameters that are assumed.

Posterior: $P(\theta|y) \propto P(y|\theta)\cdot P(\theta) = \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$
$= Beta(y+\alpha, n-y+\beta)$

→ When prior & posterior are of the same form they are called Conjugate prior

$$E(\theta|y) = \frac{y+\alpha}{n+\alpha+\beta} \qquad Var(\theta|y) = \frac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n-1)}$$

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

When y, n tends to be larger for fixed α, β

$$E(\theta|y) \to y/n$$

$$Var(\theta|y) \to \frac{1}{n}\left(\frac{y}{n}\right)\left(1-\frac{y}{n}\right) \to 0 \text{ as } n \to \infty$$

Formal distribution of a conjugate prior.

If F is a class of sampling distribution $p(y|\theta)$ and P is a class of prior distribution θ then class P is conjugate for F if

$$p(\theta|y) \in P \;\forall\; p(\cdot|\theta) \in F \text{ and } p(\theta) \in P$$

We want P to be of same form as F Then it is called natural conjugate prior.

Exponential family of distribution

Class F is a family of exponential distribution if all its members have the form

$$\boxed{p(y_i|\theta) = f(y_i)\, g(\theta)\, e^{\phi(\theta) u(y_i)}}$$

- Not necessary that θ is single parameter.

If θ is multiparameter then,

$$\boxed{p(y_i|\theta) = f(y_i)\, g(\theta)\, e^{\phi(\theta)^T u(y_i)}}$$

- $\phi(\theta)$ & $u(y)$ must be of same dimension.
- $\phi(\theta)$ is called natural parameter of F

Suppose a sequence of observation is made $Y = \{y_1, y_2 \ldots y_n\}$

$$L = P(y|\theta) = \left[\prod_{i=1}^{n} f(y_i)\right] (g(\theta))^n \exp\left(\phi(\theta)^T \sum_{i=1}^{n} u(y_i)\right)$$

$$\propto (g(\theta))^n \exp\left(\phi(\theta)^T \underbrace{\sum_{i=1}^{n} u(y_i)}\right)$$

$$T(u) = \sum_{i=1}^{n} u(y_i) \text{ is called "sufficient statistic"}$$

When y_i tends to ∞

If $p(\theta) \propto g(\theta)^\eta e^{\phi^T(\theta)\nu}$

Then the posterior.

$$p(\theta|y) \propto g(\theta)^{\eta+n} e^{\phi^T(\theta)(\nu + T(y))}$$

which is of same form as $p(\theta)$.

Non-Informative Prior

When prior distribution has no population basis. They can be difficult to construct. Here it is necessary to derive prior that play very less role in the posterior distribution.

Such priors are called "Reference prior distributions"

The prior density is non informative / flat or diffuse

Idea is, let the data speak for itself.

Date
Page

Jeffreys Invariance Principle.

→ Used to define non-informative prior distribution
→ Based on one-to-one transformation of the parameters

$$\phi = h(\theta) \Rightarrow \theta = h^{-1}(\phi)$$

→ By transformation the prior density is equivalent in terms of expressing the same beliefs to the prior density on ϕ.

$$p(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right| = p(\theta) \left| h'^{-1}(\phi) \right|$$

→ Any rule for determining the prior density $p(\theta)$ should yield equivalent result if applied to the transformed variable. ie. $p(\phi)$ which is computed by determining $p(\theta)$

$$p(y,\phi) = p(y|\phi)\, p(\phi)$$

→ Considering this principle the non uniform prior density is given as $\boxed{p(\theta) \propto |J(\theta)|^{1/2}}$, where $J(\theta)$ is the fisher information for θ.

$$\boxed{J(\theta) = \mathbb{E}\left(\left(\frac{d \log P(y|\theta)}{d\theta} \right)^2 \middle| \theta \right)}$$

→ Gives a measure of the amount of the information a parameter carries about the likelihood function.

→ Helps quantify the sensitivity of the model parameters to changes in data distribution.

Date
Page

$$J(\theta) = -\mathbb{E}\left(\frac{d^2 \log P(y|\theta)}{d\theta^2} \middle| \theta\right)$$

$$\mathbb{E}\left(\left(\frac{d \log P(y|\theta)}{d\theta} \middle| \theta\right) = \int_y \left[\frac{\partial}{\partial\theta} \log P(y|\theta)\right] f(y|\theta) dy$$

$$\int_y \frac{1}{f(y|\theta)} \cdot \frac{\partial}{\partial\theta} f(y|\theta) \cdot f(y|\theta)\, dy$$

$$\int_y \frac{\partial}{\partial\theta} f(y|\theta)\, dy = \frac{\partial}{\partial\theta} \underbrace{\int_y f(y|\theta) dy}_{1} = 0$$

$$\frac{d^2 \log f(y|\theta)}{d\theta^2} = \frac{d}{d\theta}\left(\frac{1}{f(y|\theta)} \frac{df(y|\theta)}{d\theta}\right) \quad \frac{-1}{(f(y|\theta))^2} \frac{\partial f(y|\theta)}{\partial\theta}$$

$$= \frac{1}{f(y|\theta)} \cdot \frac{d^2 f(y|\theta)}{d\theta^2} - \left(\frac{1}{f(y|\theta)} \cdot \frac{d}{d\theta} f(y|\theta)\right)^2$$

$$= \frac{1}{f(y|\theta)} \frac{d^2 f(y|\theta)}{d\theta^2} - \left(\frac{d}{d\theta} \log f(y|\theta)\right)^2$$

Taking expectation on both side.

$$\mathbb{E}\left(\frac{d^2 \log f(y|\theta)}{d\theta^2}\right) = \underbrace{\mathbb{E}\left(\frac{1}{f(y|\theta)} \frac{d^2 f(y|\theta)}{d\theta^2}\right)}_{0} - \mathbb{E}\left(\frac{d}{d\theta} \log f(y|\theta)\right)^2$$

$$\int_y \frac{1}{f(y|\theta)} \frac{d^2 f(y|\theta)}{d\theta^2} f(y|\theta)\, dy$$

$$\frac{d^2}{d\theta} \int_y f(y|\theta)\, dy$$

$$\frac{\partial^2(1)}{\partial\theta} = 0.$$

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

Q Prove that Jeffereys prior is invariant to parametrization

$$f(\theta) = |J(\theta)|^{1/2}$$

→ $J(\phi)$ at $\theta = h^{-1}(\phi)$

show $f(\phi) = f(\theta)\left|\frac{d\theta}{d\phi}\right|$

$$J(\phi) = -E\left(\frac{d^2}{d\phi^2}\log f(y|\phi) \middle| \phi\right)$$

$$= -E\left(\frac{d^2}{d\theta^2}\log f(y|\theta = h^{-1}(\phi)) \middle| \theta\right)\left(\frac{d\theta}{d\phi}\right)^2$$

$$= J(\theta)\left|\frac{d\theta}{d\phi}\right|^2$$

$$(J(\phi))^{1/2} = [J(\theta)]^{1/2}\left|\frac{d\theta}{d\phi}\right|$$

Q Take a binomial distribution and find Jeffereys prior

$$J(\theta) = -E\left(\frac{d^2}{d\theta^2}\log f(y|\theta) \middle| \theta\right)$$

$$-E\left(\frac{d^2}{d\theta^2}\log\left[\binom{n}{y}\theta^y(1-\theta)^{n-y}\right] \middle| \theta\right)$$

$$-E\left(\binom{n}{y}\frac{d^2}{d\theta^2}\left[\theta^y(1-\theta)^{n-y}\right] \middle| \theta\right)$$

$$\binom{n}{y}\frac{\partial}{\partial\theta}\left(y\theta^{y-1}(1-\theta)^{n-y} - \theta^y(n-y)(1-\theta)^{n-y-1}\right) \middle| \theta$$

Date
Page

$$-E\left(\binom{n}{y}\frac{\partial}{\partial\theta}\left(y\theta^{y-1}(1-\theta)^{n-y}-\theta^{y}(n-y)(1-\theta)^{n-y-1}\right)\Big|\theta\right)$$

$$-E\left(\binom{n}{y}\left[y(y-1)\theta^{y-2}(1-\theta)^{n-y}-y\theta^{y-1}(n-y)(1-\theta)^{n-y-1}\right.\right.$$
$$\left.\left.-\left\{y\theta^{y-1}(n-y)(1-\theta)^{n-y-1}-\theta^{y}(n-y)(n-y-1)(1-\theta)^{n-y-2}\right\}\right]\right)$$

$$J(\theta)=-E\left(\frac{\partial^2}{\partial\theta^2}\log\binom{n}{y}+y\log\theta+(n-y)\log(1-\theta)\Big|\theta\right)$$

$$-E\left(\frac{\partial}{\partial\theta}\left(\frac{y}{\theta}-\frac{(n-y)}{1-\theta}\right)\Big|\theta\right)$$

$$-E\left(\frac{-y}{\theta^2}+\frac{(n-y)}{(1-\theta)^2}\Big|\theta\right)$$

$$-\left(\frac{-1}{\theta^2}E(y)-\frac{1}{(1-\theta)^2}E(n-y)\right)$$

$$-\left(\frac{-n\theta}{\theta^2}-\frac{1}{(1-\theta)^2}(n-n\theta)\right)$$

$$\frac{n}{\theta}+\frac{n}{1-\theta}=\frac{n}{\theta(1-\theta)}$$

$$p(\theta)\propto\sqrt{\frac{n}{\theta(1-\theta)}}=n^{1/2}\theta^{-1/2}(1-\theta)^{-1/2}$$
$$\propto \text{Beta}(1/2,1/2)$$

For two cases all principles (for generating non-informative prior) will give similar results.

a) If $p(y)$ is such that $p(y-\theta|\theta)$ is free of θ & y
Then $y-\theta$ is called as a pivotal quantity
θ is called a location parameter

b) ~~If $p(y)$ is such that~~

$$p(y-\theta|y) \propto p(\theta)\, p(y-\theta|\theta)$$

↳ constant.

b) If $p(y)$ is such that $p(y|\theta)$ is free of θ and y then $\frac{y}{\theta} = u$ is also a pivotal quantity

θ in this case is a scale parameter.

Problem with non informative prior : It is difficult to find some prior which is flat. The cdf may not be equal to 1 in all case.

Weakly informative priors: Proper prior with some general information.

Ex- Money in the wallet → May range from 100 to 3000
So we can consider a normal distribution with mean 500

⇒ CASE STUDY

A study was conducted to know what is the prob. of girl birth and boy birth in case the mother is suffering from Placenta Previa.

Proportion of girl birth in the population is 0.485
For sample of Placenta Previa out of 980 samples, 437 were girls. P(girl) = 0.485

Sol^n $P(\text{girl} \mid \text{Mother has placenta previa}) = \frac{437}{980}$

Date
Page

θ = prob. of a girl birth in patients with PP

To find $\mathbb{E}(\theta)$ and $\text{Var}[\theta]$

$P(\theta | y, n)$ → Posterior (girls birth)

$\hookrightarrow P(\theta) \cdot P(y|\theta, n)$

$P(y|\theta, n) \propto \theta^y (1-\theta)^{n-y}$

$P(\theta) \propto \text{Beta}(1,1)$

$P(\theta) \cdot P(y|\theta, n) \propto \theta^y (1-\theta)^{n-y}$

$= \text{Beta}(y+1, n-y+1)$

$\mathbb{E}(\text{Beta}(y+1, n-y+1)) = \frac{(y+1)}{n+2} = 0.446$

$\text{Var}(\text{Beta}(y+1, n-y+1)) = \frac{(y+1)(n-y+1)}{(n+2)^2 (n+3)}$

$= 251 \times 10^{-6}$

$S.D \simeq 0.016.$

Q Suppose you have Beta(4,4) prior distribution on the probability θ that a coin will yeild a head when spun in a specified manner. The coin is independently spun 10 times and head appears fewer than 3 times. You are not told how many heads were seen only the number is less than 3. Calculate the posterior density (exactly) upto a proportionality constant.

Sol^n

$P(\theta) \sim \text{Beta}(4,4)$

Unobserved parameter → θ

Observation → $y < 3$

↳ no. of heads in 10 trials

Date
Page

$P(\theta|y<3)$ → To find posterior.

$$P(\theta|y<3) \propto P(y<3|\theta) \cdot P(\theta).$$

$$= P(y=0|\theta)P(\theta) + P(y=1|\theta)\cdot P(\theta) + P(y=2|\theta)P(\theta)$$

$$= \left[\binom{n}{0}\theta^0(1-\theta)^n + \binom{n}{1}\theta^1(1-\theta)^{n-1} + \binom{n}{2}\theta^2(1-\theta)^{n-2}\right]P(\theta)$$

$$\left[(1-\theta)^n + n\theta(1-\theta)^{n-1} + \frac{n(n-1)}{2}\theta^2(1-\theta)^{n-2}\right]P(\theta)$$

$$\left((1-\theta)^{10} + 10\theta(1-\theta)^9 + 45\theta^2(1-\theta)^8\right)\theta^3(1-\theta)^3$$

$$= \theta^3(1-\theta)^{11}\left((1-\theta)^2 + 10\theta(1-\theta) + 45\theta^2\right).$$

$$= \theta^3(1-\theta)^{11}(1+\theta^2-2\theta+10\theta-10\theta^2+45\theta^2)$$

$$= \theta^3(1-\theta)^{11}(36\theta^2+8\theta+1)$$

Q Mind Game

One person is guessing random numbers from 1 to 10 20 times and based on that the other person will guess the 10 samples whether they are greater than 5 or less than equal to 5.

Sol' Observation = $Y = \{y_1, y_2 \cdots y_{10}\}$

where $y_i = 0$ if count ≤ 5

1 otherwise.

$$P(\tilde{y}|y) = \int_\theta P(\tilde{y}, \theta|y)$$

$$\int_\theta P(\tilde{y}|\theta)\cdot P(\theta|y)$$

$$\int_\theta \theta\, P(\theta|y)$$

$$\int_\theta \theta \cdot \frac{P(y|\theta)\cdot P(\theta)}{P(y)} = \frac{y+1}{n+2}$$

Date
Page

→ What we did till now.

→ Assumed a likelihood distribution
- Binomial
- Normal
- Poisson
- Exponential

→ Priors : Informative → Conjugate.
Non-Informative
Weak Informative.

→ Posterior
→ Prior predictive distribution
→ Posterior prediction.

Multiparameter Modelling

NORMAL DISTRIBUTION

Likelihood $P(y|N(\mu,\sigma^2)) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$

Assume known variance but unknown mean μ.

Prior (Conjugate) → $e^{A\theta^2+B\theta+C} \sim N(\mu_0, \tau_0^2)$

↳ Since the likelihood is exponent so the prior will also be like this.

(μ_0 & τ_0 are hyperparameters)

Date
Page

$$p(\theta) = \frac{1}{\sqrt{2\pi}\,\tau_0} e^{\frac{-1}{2\tau_0^2}(\theta-\mu_0)^2} \rightarrow \text{Prior dist}^n$$

Posterior

$$p(\theta|y) \propto p(y|\theta)\cdot p(\theta)$$

where $\theta = \mu$

$$= \frac{1}{2\pi\sigma\tau_0} e^{\frac{-1}{2\sigma^2}(y-\theta)^2 - \frac{1}{2\tau_0^2}(\theta-\mu_0)^2}$$

$$\exp\left(\frac{-1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$

$$\frac{-1}{2\sigma^2\tau_0^2}\left[(\tau_0 y - \tau_0\theta)^2 + (\sigma\theta - \sigma\mu_0)^2\right]$$

$$\frac{-1}{2\sigma^2\tau_0^2}\left(\tau_0^2y^2 + \tau_0^2\theta^2 - 2\tau_0^2\theta y + \sigma^2\theta^2 + \sigma^2\tau_0^2 - 2\sigma^2\theta\tau_0\right.$$

$$\left.(\tau_0^2+\sigma^2)\theta^2 - (2\tau_0^2y + 2\sigma^2\tau_0)\theta + \tau_0^2y^2 + \sigma^2\tau_0^2\right)$$

$$\frac{-1}{2\sigma^2\tau_0^2}\left[\tau_0^2y^2 + \tau_0^2\theta^2 - 2\tau_0^2y\theta + \sigma^2\theta^2 + \sigma^2\mu_0^2 - 2\sigma^2\theta\mu_0\right.$$

$$= \frac{-1}{2\sigma^2\tau_0^2}\left((\tau_0^2+\sigma^2)\theta^2 - 2(\tau_0^2y + \sigma^2\mu_0)\theta + (\tau_0^2y^2 + \sigma^2\mu_0^2)\right.$$

$$= \frac{-1}{2\sigma^2\tau_0^2/(\tau_0^2+\sigma^2)}\left(\theta^2 - 2\left(\frac{\tau_0^2y + \sigma^2\mu_0}{\tau_0^2+\sigma^2}\right)\theta + \frac{\tau_0^2y^2 + \sigma^2\mu_0^2}{\tau_0^2+\sigma^2}\right)$$

let $\frac{\tau_0^2y + \sigma^2\theta}{\tau_0^2+\sigma^2}$ as M_1 and $\frac{\tau_0^2y^2 + \sigma^2\mu_0^2}{\tau_0^2+\sigma^2}$ as z

$$\frac{-1}{2\sigma^2\tau_0^2/(\tau_0^2+\sigma^2)}\left(\theta^2 - 2\theta M_1 + M_1^2 - M_1^2 + z\right)$$

let $\frac{\sigma^2\tau_0^2}{\tau_0^2+\sigma^2}$ as τ_1

Date
Page

$$\exp\left(\frac{-1}{2\tau}\left[\frac{(\theta-\mu_1)^2}{\tau^2} + \frac{\mu_1^2 - z}{\tau^2}\right]\right)$$

$$\propto \exp\left(\frac{-1}{2}\frac{(\theta-\mu_1)^2}{\tau^2}\right)$$

Posterior : $\sim N(\mu_1, \tau^2)$

MULTIPARAMETER MODEL

⇒ Concept of "Nuisance parameter"

→ Unknown and unobserved quantities common in practical problem.

→ Out of these one or few of these quantity is relevant.

→ Obtain a marginal probability distribution
  → First obtain the joint probability distribution over all unknowns
  → Integrate the distribution over the unknown that are not of interest.
  → These unknowns are the nuisance parameter

⇒ Averaging over the nuisance parameter

$p(\theta_1, \theta_2 | y)$ → Given

Suppose, We want to find $p(\theta_1 | y)$

$$p(\theta_1|y) = \int_{\theta_2} p(\theta_1, \theta_2 | y)\, d\theta_2$$

$$= \int_{\theta_2} f(y|\theta_1,\theta_2)\cdot f(\theta_1,\theta_2)\,d\theta_2$$

Q Consider 2 coins C_1 and C_2 with the following characteristic $P(\text{Heads}|C_1) = 0.6$ and $P(\text{Heads}|C_2) = 0.4$. Choose one of the coin as random and assume spinning it repeatedly. Now we observe that first two spin results in "Tails". Now what is expectation of the number of additional spin until head shows up?

2.29

Expected no of trial given coin 1 = $1/p$.

$P(C_1|TT)$ $\quad\quad$ $E(E(N|C_1,TT))$

⇒ Normal Data with non-informative priors.

- Consider a vector y of n independent observations from a univariate normal distribution $N(\mu,\sigma^2)$
- Consider a non-informative prior $f(\mu,\sigma^2) \propto (\sigma^2)^{-1}$
- Finding the joint posterior density

$$f(\mu,\sigma^2|y) \propto f(y|\mu,\sigma^2)\cdot f(\mu,\sigma^2)$$

$$\propto \left[\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\,\sigma}\, e^{\frac{-1}{2\sigma^2}(y_i-\mu)^2}\right)\right]\left(\frac{1}{\sigma^2}\right)$$

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

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

$$\propto (\sigma)^{-n-2}\exp\left(\frac{-1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2+(\bar{y}-\mu)^2+2(y_i-\bar{y})(\bar{y}-\mu)\right]\right)$$

$$= (\sigma)^{-n-2}\exp\left(\frac{-1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2+\sum_{i=1}^{n}(\bar{y}-\mu)^2+2\sum y_i\bar{y}-y_i\mu-(\bar{y})^2+\bar{y}\mu\right]\right)$$

$$= (\sigma)^{-n-2}\exp\left(\frac{-1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2+n(\bar{y}-\mu)^2\right]\right)$$

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

$$= (\sigma)^{n-2}\exp\left(\frac{-1}{2\sigma^2}\left((n-1)S^2 + n(\bar{y}-\mu)^2\right)\right) \quad -(1)$$

Sufficient Statistic: The statistic that we need to sufficiently define the distribution in this case S^2 and $\bar{y}$ are sufficient statistic.

$p(\mu|y,\sigma^2)$, assuming σ^2 is known.

$$p(\mu|y,\sigma^2) \propto p(y,\sigma^2|\mu)\,p(\mu)$$
$$\propto p(y|\mu,\sigma^2)\,p(\sigma^2|\mu)\cdot p(\mu).$$

$$p(\mu|y,\sigma^2) \propto p(y|\mu,\sigma^2)\,p(\mu|\sigma^2)$$

$$\propto \exp\left(\frac{-1}{2\sigma^2}(n(\mu-\bar{y})^2\right)$$

$$\propto \exp\left(\frac{-1}{2\sigma^2/n}(\mu-\bar{y})^2\right) \sim N(\bar{y},\sigma^2/n) \quad -(2)$$

$$p(\sigma^2|y) = \int_{\mu} p(\mu|\sigma^2,y)\,d\mu$$

↳ Marginal posterior distribution of the variance.

$$\int_{\mu}(\sigma)^{-n-2}\exp\left(\frac{-1}{2\sigma^2}\left[(n-1)S^2 + n(\bar{y}-\mu)^2\right]\right)d\mu$$

$$(\sigma)^{-n-2}\, e^{-\frac{1}{2\sigma^2}(n-1)S^2}\int_{\mu}\exp\frac{-1}{2\sigma^2}n(\bar{y}-\mu)^2\,d\mu$$

$$\propto \left(\frac{1}{2\sigma^2}\right)\cdot(2(\bar{y}-\mu)).$$

Date
Page

$$\propto \int e^{\frac{n}{2\sigma^2}(\bar{y}^2 - 2\bar{y}\mu + \mu^2)} d\mu$$

$$\propto \int e^{\frac{-n2\bar{y}}{2\sigma^2}\mu} d\mu \int e^{\frac{-n}{2\sigma^2}\mu^2} d\mu$$

$$e^{\frac{-n2\bar{y}}{2\sigma^2}\mu}$$

$$\propto \int_{\mu} \exp\left(\frac{-1}{2\left(\frac{\sigma^2}{n}\right)}(\mu - \bar{y})^2\right) d\mu = \frac{\sqrt{2\pi\sigma^2}}{\sqrt{n}} \int N(\bar{y}, \frac{\sigma^2}{n}) \cdot 1$$

$$f(\sigma^2|y) \propto (\sigma)^{-n-2} \exp\left(\frac{-1}{2\sigma^2}(n-1)S^2\right) \sqrt{\frac{2\pi\sigma^2}{n}}$$

$$\propto \boxed{(\sigma^2)^{-\frac{(n+1)}{2}} \exp\left(\frac{-1}{2\sigma^2}(n-1)S^2\right)} \quad \text{— (3)}$$

↳ Scaled Inverse χ^2 density.

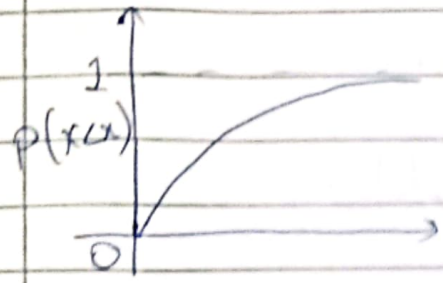
Q Find $f(\mu|y)$.

?! How to sample from joint posterior density?

@ First draw σ^2 from (3)

Note ⇒ Techniques for generating sample that follows a distribution

1) Inverse Transform

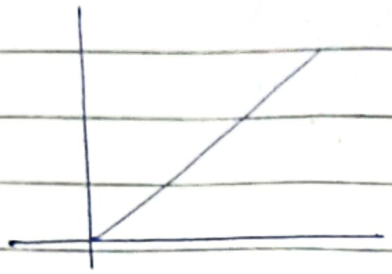


$$P(X \le x) = \sum_{x=-\infty}^{x} P(x)$$

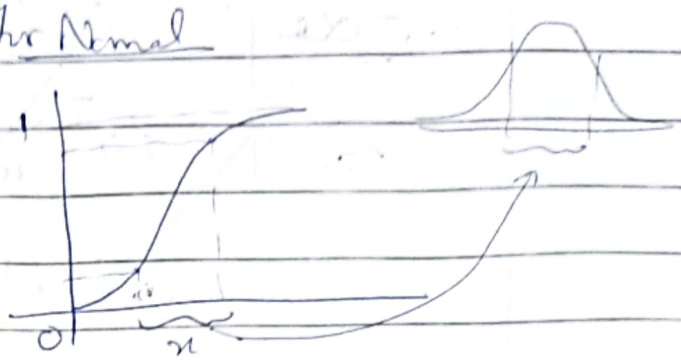
Take a number b/w 0 & 1 and find the corresponding value of x.

Date
Page

For Uniform



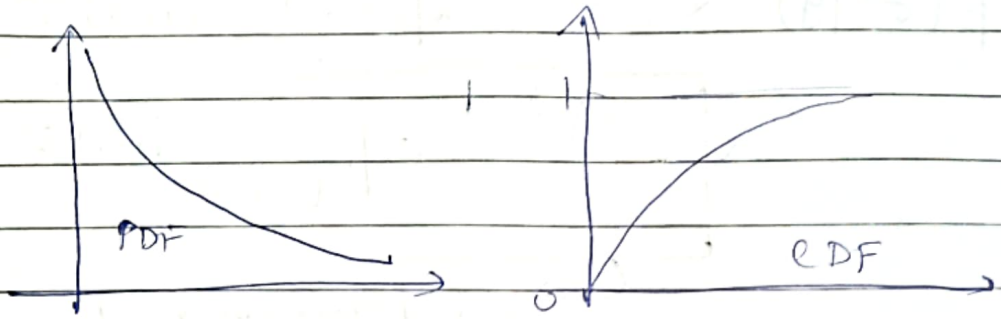
For Normal



For Exponential distribution.

$$P(X = x) = \lambda e^{-\lambda x}$$

$$P(X \leq x) = 1 - e^{-\lambda x}.$$



Generate a random number $U \in [0,1]$

$$U = 1 - e^{-\lambda x}$$

$$e^{-\lambda x} = 1 - U$$

$$-\lambda x = \ln(1-U)$$

$$x = -\frac{1}{\lambda} \ln(1-U)$$

In this technique we will take a random no. and for the CDF we find x corresponding to y. by finding inverse. transform.

2) Acceptance / Rejection Technique.

- Consider an alternate distribution(A) whose inverse CDF is solvable.
- Generate samples x based on this alternate distribution.

Date
Page

$\frac{p_0(x)}{p_A(x)}$, if this ratio is greater than a threshold then accept otherwise reject.

(b) For a given σ^2, draw μ from (2)
Now we have μ & σ^2 which will follow (1)

$$P(\tilde{y}|y) \propto \int_{\sigma^2}\int_{\mu} P(\tilde{y}, \mu, \sigma^2|y)\, d(\sigma^2)\, d\mu$$

$$= \int_{\sigma^2}\int_{\mu} P(\tilde{y}|\mu, \sigma^2, y) \cdot P(\mu, \sigma^2|y)\, d(\sigma^2)\, d\mu.$$

$$\int_{\sigma^2}\int_{\mu} N(\tilde{y}, \mu, \sigma^2) \quad \text{derived using (4)(1)}$$

The integration is difficult to calculate so we will use the Sampling techniques.

- First draw samples from the joint posterior distribution $P(\mu, \sigma^2|y)$
- Given μ and σ^2 ~~draw samples from~~ find $N(\tilde{y}, \mu, \sigma^2)$ ie $\tilde{y} \sim N(\mu, \sigma^2)$

Multinomial model for categorical Data.

↓

~~The number of~~

- Data for which each observation has k possible outcomes.

- If Y is the vector of the counts of number of observations of each outcome

$$p(Y|\theta) \propto \prod_{j=1}^{k} \theta_j^{y_j} \quad ; \sum_{j=1}^{k} \theta_j = 1, \sum_{j=1}^{k} y_j = n$$

Constraints.

$$P(\theta|y) \propto P(y|\theta) \cdot P(\theta)$$

for binomial conjugate prior
$Beta(\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

For Multinomial conjugate prior is Dirichlet Distribution.

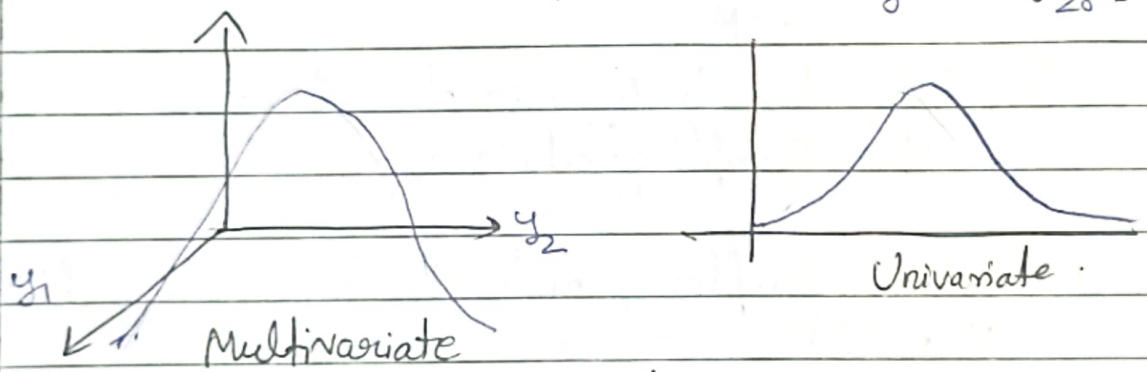
$$D(\alpha_1, \alpha_2, \alpha_3 \dots \alpha_k) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$$

$$P(\theta|y) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j + y_j - 1} \sim D(\alpha_1 + y_1, \alpha_2 + y_2, \dots \alpha_k + y_k)$$

Multivariate Normal Model.

Likelihood

For univariate case: $f(y; \mu, \sigma^2) \propto \frac{1}{\sigma} \exp\left(\frac{-1}{2\sigma^2}(x-\mu)^2\right)$



Suppose y is an observable quantity with d components. $y \in \mathbb{R}^d$

$$y|\mu, \Sigma \sim N(\mu, \Sigma)$$

$$\underset{\in \mathbb{R}^d}{y} \propto |\Sigma|^{1/2} \exp\left(\frac{-1}{2}(Y-\mu)^T \Sigma^{-1} (Y-\mu)\right)$$

$$E((y_i - \mu_i)(y_j - \mu_j))$$

$$\Sigma = \frac{1}{\sigma^2} \begin{bmatrix} Cov(Y_1, Y_1) & Cov(Y_1, Y_2) & \dots & Cov(Y_1, Y_d) \\ Cov(Y_2, Y_1) & Cov(Y_2, Y_2) & \dots & Cov(Y_2, Y_d) \\ & & & \\ Cov(Y_d, Y_1) & Cov(Y_d, Y_2) & \dots & Cov(Y_d, Y_d) \end{bmatrix}$$

↳ Symmetric

↳ Positive definite matrix ($x^T \Sigma x = 0$ → +ive eigen values) → The Δel^n to this is +ive and it becomes quadratic while solving

Suppose you are given n observations, $y_1, y_2 \cdots y_n$

Likelihood: $p(y_1, y_2 \cdots y_n | \mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\left(\frac{-1}{2} \sum_{i=1}^{n} (y_i - \mu_i)^T \Sigma^{-1} (y_i - \mu_i)\right)$

y_i

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_d \end{bmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1d} \\ \vdots & & & \\ \Sigma_{d1} & \Sigma_{d2} & \cdots & \Sigma_{dd} \end{bmatrix} \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_d \end{bmatrix}$$

$$\hat{y}_1 (\hat{y}_1 \Sigma_{11} + \hat{y}_2 \Sigma_{21} \cdots \hat{y}_d \Sigma_{d1})$$
$$+ \hat{y}_2 (\hat{y}_1 \Sigma_{12} + \hat{y}_2 \Sigma_{22} \cdots \hat{y}_d \Sigma_{d2})$$
$$+ \cdots$$
$$+ \hat{y}_d (\hat{y}_1 \Sigma_{1d} + \hat{y}_2 \Sigma_{2d} \cdots \hat{y}_d \Sigma_{dd})$$

$$= \text{tr}(\Sigma^{-1} S_0), \quad \text{where } S_0 = \sum_{i=1}^{n} (y_i - \mu)(y_i - \mu)^T$$

$$\boxed{p(y_1, y_2 \cdots y_n | \mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\left(\frac{-1}{2} \text{tr}(\Sigma^{-1} S_0)\right)}$$

Conjugate Analysis

→ Assume we do not know μ, Σ is known

→ Prior for μ is assumed to be normal $\mu \sim N(\mu_0, \Lambda_0)$, $\mu_0 \in \mathbb{R}^d$, $\Lambda_0 \in \mathbb{R}^{d \times d}$

posterior $p(\mu | y, \Sigma) \propto p(y | \mu, \Sigma) p(\mu | \Sigma)$

$\propto N(y, \mu, \Sigma) N(\mu, \mu_0, \Lambda_0)$

$$\propto \exp\left(\frac{-1}{2}\left[(\mu-\mu_0)^T \Lambda_0^{-1}(\mu-\mu_0) + \sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right]\right)$$

$$(\mu-\mu_0)^T\Lambda_0^{-1}(\mu-\mu_0)$$

$$\mu^T\Lambda_0^{-1}\mu - \mu_0^T\Lambda_0^{-1}\mu - \mu^T\Lambda_0^{-1}\mu_0 + \mu_0^T\Lambda_0^{-1}\mu_0$$

Quadratic — Scalar quantities

$$-\mu^T\Lambda_0^{-1}\mu_0 - \mu^T\Lambda_0^{-1}\mu_0$$

$$= \underbrace{\mu^T\Lambda_0^{-1}\mu}_{\text{Quadratic in terms of } \mu} - \underbrace{2\,\mu_0^T\Lambda_0^{-1}\mu_0}_{\text{Linear in terms of } \mu} + \underbrace{\mu_0^T\Lambda_0^{-1}\mu_0}_{\text{Constant in terms of } \mu} \quad \text{— (I)}$$

$$\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu) =$$

$$= \sum_{i=1}^{n}\left[\mu^T\Sigma^{-1}\mu - 2\mu^T\Sigma^{-1}y_i + y_i^T\Sigma^{-1}y_i\right]$$

$$= n\mu^T\Sigma^{-1}\mu - 2n\mu^T\Sigma^{-1}\bar{y} + \text{constant} \quad \text{— (II)}$$

~~$\propto \exp$~~ from (I) + (II)

$$\mu^T\Lambda_0^{-1}\mu + n\mu^T\Sigma^{-1}\mu - 2\mu^T(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}) + \text{constant}$$

$$\mu^T(\Lambda_0^{-1} + n\Sigma^{-1})\mu - 2\mu^T(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}) + \text{const.}$$

$$\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}$$

$$\Lambda_n^{-1}\mu_n = (\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$$

$$\mu_n = \Lambda_n(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$$

$$= (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$$

$\sim N(\mu; \mu_n, \Lambda_n)$

Posterior distribution of μ in a multivariate normal distribution with known Σ. Assuming prior $\mu \sim N(\mu_0, \Lambda_0)$

$$f(\mu|Y,\Sigma) \sim N(\mu_n, \Lambda_n)$$
$$\mu_n = (\Lambda_0^{-1} + n\Sigma^{-1})(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{Y})$$
$$\Lambda_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}$$

Posterior Conditionals and marginal distribution of a subvectors of μ with known Σ

We know, $y \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$

$$\mu = \begin{bmatrix} \mu_1^k \\ \mu_2^k \\ \mu_3^u \\ \mu_4^u \end{bmatrix} \begin{matrix} \rightarrow \text{known} \\ \\ \rightarrow \text{unknown} \end{matrix} \rightarrow \text{we want to find their distribution}$$

- Marginal posterior of a subset of parameter (say $\mu^{(1)}$)
- Given a known subset ($\mu^{(2)}$)
- A dataset y

Q We want $\mu^{(1)} | \mu^{(2)}, y$

Result is a multivariate normal distribution

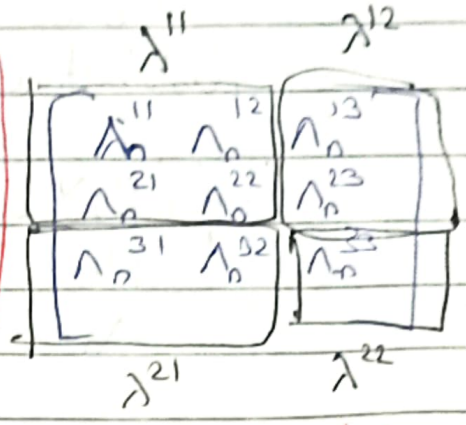
$$\mu^{(1)} | \mu^{(2)}, y \sim N(\mu_n^{(1)} + \beta^{1/2}(\mu^{(2)} - \mu_n^{(2)}), \Lambda^{1/2})$$

$\mu_n \rightarrow$ Posterior mean vector
$\mu_n^{(1)} \rightarrow$ Appropriate subvector of μ_n with similar grouping as $\mu^{(1)}$
$\mu_n^{(2)} \rightarrow$ Appropriate ... of μ_n with ... $\mu^{(2)}$

$\Lambda_n \rightarrow$ Appropriate Submatrix of the given n variance matrix

$$\beta^{1/2} = \Lambda_n^{(1,2)} \left(\Lambda_n^{(2,2)}\right)^{-1}$$

$$\Lambda^{1/2} = \Lambda_n^{(1,1)} - \Lambda_n^{(1,2)} \left(\Lambda_n^{(2,2)}\right)^{-1} \Lambda_n^{(2,1)}$$



Posterior Predictive Distribution of a new data

Suppose we want to find $\tilde{y}|y$ for known variance Σ.

$$p(\tilde{y}|y) = \int_{\mu} p(\tilde{y}, \mu | y, \Sigma)\, d\mu$$

$$= \int_{\mu} \underbrace{p(\tilde{y}|\mu, y, \Sigma)}_{\text{likelihood}} \cdot p(\mu|y, \Sigma)\, d\mu$$

$$= \int_{\mu} p(\tilde{y}|\mu, \Sigma) \cdot \underbrace{p(\mu|y, \Sigma)}_{\text{Posterior of } \mu \text{ with known } \Sigma} d\mu$$

$$\int_{\mu} N(\mu, \Sigma) \cdot N(\mu_n, \Lambda_n)\, d\mu$$

which is normal distribution

Rather than integrating we find the mean and the variance of posterior predictive distribution.

We know $\mathbb{E}(u) = \mathbb{E}(\mathbb{E}(u|v))$

$$\mathbb{E}(\tilde{y}|y) = \mathbb{E}(\mathbb{E}(\tilde{y}|y, \mu)|y)$$

$$= \mathbb{E}(\mu|y) = \mu_n$$

$$\text{Var}(\tilde{y}|y) = \mathbb{E}(\text{var}(\tilde{y}|y,\mu)|y) + \text{Var}(\mathbb{E}(\tilde{y}|y,\mu)|y)$$
$$= \mathbb{E}(\Sigma|y) + \text{Var}(\mu|y)$$
$$= \Sigma + \Lambda_n$$

Non-Informative Prior

$P(\mu) \propto$ constant.

The distribution generated for a non-informative prior is proper only when $n \geq d$

(n: no. of samples; d: dimension of the vector (y) multivariate.)

Multivariate normal with unknown mean and variance

Chi square distribution when extended to multivariate it becomes wishart distribution.

$\Sigma \sim \text{Inv-Wishart}(\Lambda_0^{-1})$

$\mu|\Sigma \sim N(\mu_0, \Sigma/k_0)$; k_0 is a scale parameter

Joint prior $= p(\mu, \Sigma) = \text{Inv-Wishart}$ ↳ # of prior measurements on the Σ scale.

The joint posterior density is also a Inv-Wishart distribution

Case Study – Bioassay Experiment

Scientific Problem:

- Administer various levels of dose of a drug to some batch of test animals.

- The animal response is characterized by a dichotomous outcome. (alive or dead)

- Data is of the following form

(x_i, n_i, y_i)

x_i → i^{th} dose level

n_i → No. of animals to which i^{th} dose level is given

y_i → How many of the n_i animals died.

- Sample data for tests conducted on 20 animals

Dose (x_i in log(gm/ml))	Number of animals (n_i)	Number of deaths (y_i)
0.86	5	0
0.3	5	1
0.05	5	3
0.73	5	5

We want to establish a dose-response relation.

So Modeling the dose-response relation

- Assume that the outcomes of 5 animals within each group is exchangeable.
- Reasonable to model them as independent with equal probabilities.
- Let θ_i be that probability for group i

  So $y_i | \theta_i \sim Bin(n_i, \theta_i)$ — ①

  ↳ prob. of death given dose x_i

- Assume that the outcome probability of each group is independent of each other.
- Simple assumption is $\theta_1, \theta_2, \theta_3, \theta_4$ are exchangeable.

$p(\theta_1, \theta_2, \theta_3, \theta_4) \propto 1$

Assuming $\theta_i \sim$ Beta posterior.

Assuming, $\theta_i = \alpha + \beta x_i$

α, β are parameter of the model

- Problem is with $x_i \to \infty$, $\theta_i \to \infty$; But $\theta_i \in [0, 1]$
- So let $\text{logit}(\theta_i) = \alpha + \beta x_i$

$$\text{logit}(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \beta x_i$$

$$\frac{\theta_i}{1-\theta_i} = e^{\alpha+\beta x_i}$$

$$\frac{1-\theta_i}{\theta_i} = e^{-(\alpha+\beta(x_i))}$$

or $\theta_i = \dfrac{1}{1+e^{-(\alpha+\beta x_i)}}$

Liklihood

For group i

$$p(y_i \mid \alpha, \beta, n_i, x_i) \propto \theta^{y_i} (1-\theta)^{n_i - y_i}$$

Where $\theta_i = \dfrac{1}{1+e^{-(\alpha+\beta x_i)}} = \text{logit}^{-1}(\alpha+\beta x_i)$

$$p(y_i \mid \alpha, \beta, n_i, x_i) \propto \left[\text{logit}^{-1}(\alpha+\beta x_i)\right]^{y_i} \left[1-\text{logit}^{-1}(\alpha+\beta x_i)\right]^{n_i - y_i}$$

Joint posterior of α, β.

$$p(\alpha, \beta \mid y_i, n_i, x_i) \propto p(y_i \mid \alpha, \beta, n_i, x_i) \cdot p(\alpha, \beta)$$

$$p(\alpha, \beta \mid y, n, x) \propto p(\alpha, \beta)\, p(y \mid \alpha, \beta, n, x)$$

$$p(\alpha,\beta \mid y,n,x) \propto p(\alpha,\beta) \prod_{i=1} p(y_i \mid \alpha,\beta,n_i,x_i)$$

As likelihood is complex so we have to rely on computation

Assuming $p(\alpha,\beta) \propto 1$ ie. constant.

For Sampling

We need to have some initial estimate of the parameter

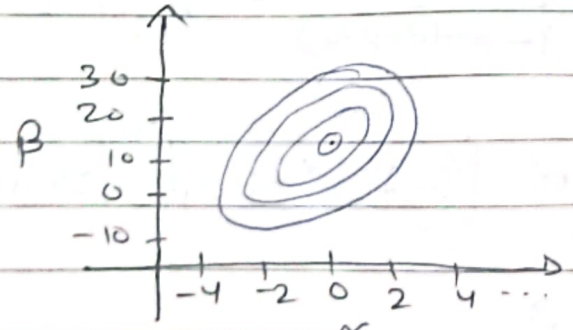
Get an estimate of α and β by fitting a logistic regression using MLE.

Estimated $(\hat{\alpha},\hat{\beta}) = (0.8, 7.7)$

$S.E(\hat{\alpha}) = 1.02$

$SE(\hat{\beta}) = 4.9$

- Consider the range $(\alpha,\beta) \in ([-5,10],[-10,40])$
Take small values for incrementing and get the value of posterior for and get the contour plot with these grid values -



Steps to find $p(\theta)$

-> Start with an initial estimate of $(\hat{\alpha},\hat{\beta})$ by fitting with a LR using MLE
(Logistic Regression.

Date
Page

$$y = \frac{1}{1+e^{\alpha+\beta x}}$$

$$\theta_i = \frac{\text{No. of deaths}(y_i)}{\text{Total number of animals}(n_i)}$$

$$p(\alpha,\beta|x,n,y) = \cancel{p(\alpha,\beta)}^{1} \cdot p(y|\alpha,\beta,n,x)$$

$$\propto \prod_{i=1}^{k} p(y_i|\alpha,\beta,x_i,n_i)$$

$$\propto \prod_{i=1}^{k} \theta_i^{y_i} (1-\theta)^{n_i-y_i}$$

$$\alpha_{new} = \alpha_{old} + \delta\alpha$$

($\delta\alpha$: grid size)

If grid size is very large then we will get fine info. about the distribution.

2) The distribution that we get is unnormalized, so there's a need of normalization. Sampling should be used.

Sampling from the joint posterior

$$p(\alpha|y_i,n_i,x_i)$$

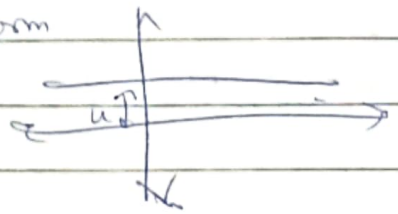
2) Compute the marginal posterior distribution of α by numerically summing over β.
→ in the discrete distribution computed over the grid

$$p(\alpha_i|y,n,x) \approx \sum_{\beta=1}^{S} p(\alpha,\beta|y,n,x)$$

2) For $S = 1$ to 1000
Draw α^S directly from the directly computed $p(\alpha,\beta|y,n,x)$

b) Draw β^s from the discrete conditional distribution.
$P(\beta | \alpha, y)$

c) For each sampled α^s and β^s add a ~~rng~~ random jitter ϵ
$E(\epsilon) = 0$ and it is uniform
$\epsilon \sim U[-u, +u]$



Introduction to Bayesian Computation Techniques

Revolves around 2 steps:

a) Computation of the posterior $p(\theta|y)$
b) Computation of the posterior predictive distribution $p(\tilde{y}|y)$

- Standardized distributions are easier to deal with like normal, poisson, Exponential, Gamma

Normalized & Unnormalized

The distribution (which may be multivariate) that needs to be computed (simulated) is called **Target distribution**
Let it be denoted as $p(\theta|y)$

Assume $p(\theta|y)$ to be easily computed for any value of θ upto a factor only involving y

We assume there is some easily computable function $q(\theta|y)$ which is easily computed and is unnormalized density of $p(\theta|y)$

proposal distribution

Date
Page

$\frac{q(\theta|y)}{p(\theta|y)}$ is a constant that only depends on y and is proportional to the posterior density.

Numerical Integration / Quadrature Finding

- It refers to a method in which the integration over the continuous function is evaluated
- By computing the value of the function at finite set of points.

Two methods for numerical integration

- Simulation based (stochastic)
- Deterministic method.

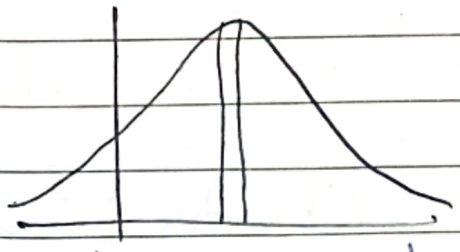
Simulation Based

$$\mathbb{E}(h(\theta|y)) = \int_{\theta} h(\theta|y)\, p(\theta|y)\, d\theta = \int_{\theta} h(\theta)\, p(\theta|y)\, d\theta$$

Draw posterior draws θ^s from $p(\theta|y)$

Estimate the integral

$$\frac{1}{S}\sum_{i=1}^{S} h(\theta^i)$$



When we integrate we need to have the sum the value of the $h(\theta)$ at that i as when we sample the $h(\theta)$ will be higher at the peak and less at the tails.

- Samples drawn are independent of each other

Date
Page

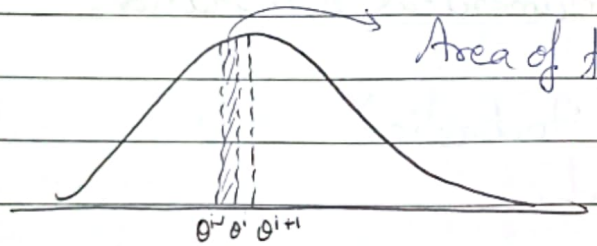
Deterministic Method

Evaluate the integrand $\int h(\theta) p(\theta|y) d\theta$ at selected points θ^s.

$$\mathbb{E}(h(\theta|y)) = \int h(\theta) p(\theta|y) = \frac{1}{S} \sum_{i=1}^{S} \omega^{(i)} h(\theta^{(i)}) p(\theta^{(i)}|y)$$

↳ weight given to a particular integral.

$\omega^{(i)}$ is the weight corresponds to volume of space at θ^i



Area of trapezium $= \frac{1}{2}\left(h(\theta^{(i-1)}) + h(\theta^{(i)})\right) \Delta\theta^{(i)}$

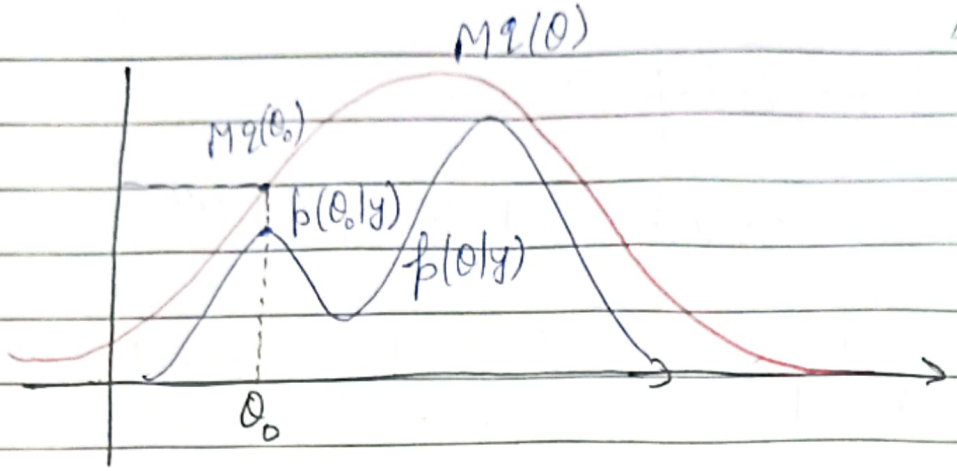
$$\mathbb{E}(h(\theta|y)) = \frac{1}{2} \sum_{i=1}^{S} \left(h(\theta^{(i-1)}) + h(\theta^{(i)})\right) \Delta\theta^{(i)} p(\theta^{(i)}|y)$$

Rejection Sampling

Objective : To estimate $p(\theta|y)$

- Consider a proposal distribution $q(\theta)$ where inverse of $q(\theta)$ is easy to find
- $q(\theta)$ is defined for all θ for which $p(\theta|y)$ is defined.
- $q(\theta)$ when multiplied by a constant M is always greater than $p(\theta|y)$ $\forall \theta$ for which $p(\theta|y)$ is defined

$$\boxed{M q(\theta) \geq P(\theta|y)}$$



- Use $q(\theta)$ to sample a point say θ_0
- Generate a random sample u uniformly over $[0, M]q(\theta_0)$
- Accept θ_0 if $\frac{p(\theta_0|y)}{Mq(\theta_0)} \geq u$

- Here θ_0 gets accepted with a higher prob. if $p(\theta_0|y)$ is closer to $M(q(\theta_0))$

Problem with Rejection Sampling.

→ If there is a multivariate distribution then we need to have exponentially high values for θ

Advantage

→ Generation of normalized distribution.
If we have $\alpha_1, \alpha_2 \ldots \alpha_n$ are generated sample
Normalized procedure: $\frac{p(\alpha_i|y)}{\sum_{j=1}^{n} p(\alpha_i|y)}$

Date
Page

Importance Sampling

- Suppose we are interested in $\mathbb{E}(h(\theta|y))$
- Importance Sampling finds the point estimate of the distribution directly without the use of envelope fn.

$$\mathbb{E}(h(\theta)|y) = \int_{\theta} h(\theta) f(\theta|y) d\theta$$

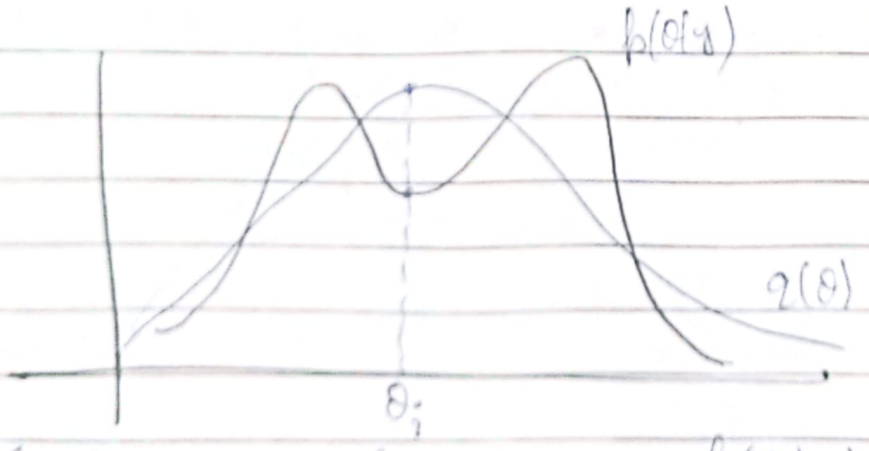
$$= \int_{\theta} h(\theta) \frac{f(\theta|y)}{g(\theta)} . g(\theta) d\theta$$

$g(\theta)$ is a distribution whose samples are easily generated. and given θ it is easy to determine the density $f(\theta|y)$ & $g(\theta|y)$.

- for each sample θ_i generated from $g(\theta)$
  - find the ratio $\frac{f(\theta_i|y)}{g(\theta_i)} = \omega_i$

Generate S sample of θ from $g(\theta)$

$$\mathbb{E}(h(\theta)|y) = \frac{1}{S} \sum_{i=1}^{S} h(\theta_i) \omega_i$$



for θ_i $g(\theta_i)$ is large but $\omega_i = \frac{f(\theta_i|y)}{g(\theta_i)}$ will be < 1 and acts as correction factor.

Date
Page

For the case when $p(\theta|y)$ is unnormalized.

$$p(\theta|y) = \frac{r(\theta|y)}{\int r(\theta|y)\, d\theta}$$

$$\mathbb{E}(h(\theta)|y) = \frac{\int h(\theta)\, r(\theta|y)\, d\theta}{\int_{-\infty}^{\infty} r(\theta|y)\, d\theta}$$

Suppose proposal distribution $q(\theta|y)$

$\frac{r(\theta|y)}{\int_{-\infty}^{\infty} r(\theta|y)\, d\theta}$ is estimated using proposal distribution $\frac{q(\theta|y)}{\int_{-\infty}^{\infty} q(\theta|y)\, d\theta}$

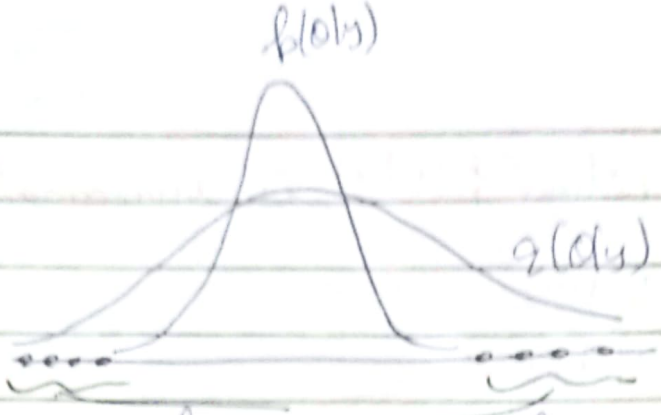
$$\mathbb{E}(h(\theta)|y) = \frac{\int_{\theta} h(\theta) \left(\frac{r(\theta|y)}{q(\theta|y)}\right) q(\theta|y)\, d\theta}{\int_{-\infty}^{\infty} \frac{r(\theta|y)}{q(\theta|y)}\, q(\theta|y)\, d\theta}$$

Let $w_i = \frac{r(\theta|y)}{q(\theta|y)}$

Given S samples from $q(\theta|y)$

$$\frac{\frac{1}{S}\sum_{i=1}^{S} h(\theta_i)\, w_i}{\frac{1}{S}\sum_{i=1}^{S} w_i} = \frac{\sum_{i=1}^{S} h(\theta_i)\, w_i}{\sum_{i=1}^{S} w_i}$$

Problem in determining distribution instead of point estimate. There are many points which are sampled from $q(\theta|y)$ which are not the part of the $p(\theta|y)$ like



These points are not belonging of form $f(\theta|y)$

Sample Importance Resampling

$$r(\theta|y) = \frac{r(\theta|y)}{q(\theta|y)}\, q(\theta|y)$$

Suppose you have generated samples from $q(\theta|y)$

$q(\theta|y) \rightarrow \{\theta^1, \theta^2, \theta^3 \ldots \theta^N\}$ N is quite large.

Generate another set of samples

$\tilde{q} \rightarrow \{\tilde{\theta}^1, \tilde{\theta}^2, \tilde{\theta}^3 \ldots \tilde{\theta}^R\}$

Accept θ^i with probability $\frac{w_i}{\sum_{i=1} w_i}$ to generate $\tilde{\theta}^i$

Only can be used in the cases of finite variance

can be used in case of

Steps

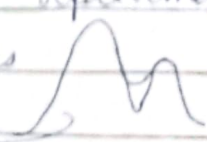
Take sample θ^i from Q

Generate a random number $r \in U[0,1]$

If $r < \frac{w_i}{\sum w_i}$ accept θ^i to $\tilde{\theta}^i$

Else repeat it for θ^i, ∞

In general we have to sample without replacement but analytically it is difficult to track as



Most of the samples are from 1st peak so the data will be biased if we do with replacement.

Date
Page

Markov Chain Monte Carlo (MCMC)

Focus the posterior density evaluations to the part of the parameter space.

$$E_{p(\theta|y)} f(\theta) = \int f(\theta)\, p(\theta|y)\, d\theta$$

We can use the unnormalized posterior $q(\theta|y) = p(y|\theta) \cdot p(\theta)$

Grid $E_{p(\theta|y)} f(\theta) \approx \dfrac{\sum_{s=1}^{S} f(\theta^s)\, q(\theta^{(s)}|y)}{\sum_{s=1}^{S} q(\theta^s|y)}$

Works for few dimensions, As the no. of dimension increases the no. of grids increases exponentially and most of the grid points have posterior as 0. so it is computationally expensive.

Markov Chain

$$p(S_t | S_1, S_2, S_3 \cdots S_{t-1}) = p(S_t | S_{t-1})$$

Stationarity Property: If we run the chain multiple times then the fraction of time we are on state S_i is same for all i.

Date
Page

Gibb's Sampling

If $y \sim N(\mu, \sigma^2)$ and μ & σ^2 are both unknown so we can't apply markov chain directly so we will use gibbs sampling.

We try to generate samples based on 1D conditional distribution $p(\mu|\sigma^2)$ and $p(\sigma^2|\mu)$

Algo

Sample θ_j^t from $p(\theta_j | \theta_{-j}^{t-1}, y)$

where $\theta_{-j}^{t-1} = (\theta_1^t, \ldots \theta_{j-1}^t, \theta_{j+1}^{t-1} \ldots \theta_d^{t-1})$

If μ & σ^2 are highly correlated then the algo will be very slow.

For Joint pdf

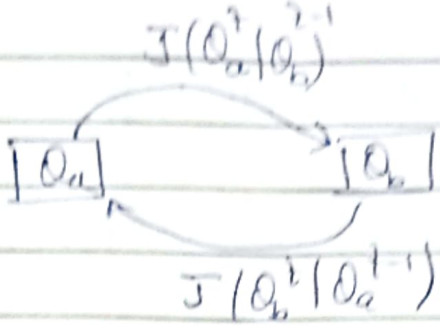
For conditional the correlation creates problem of slow convergence.

Metropolis Algorithm

1) Starting point θ^0
2) $t = 1, 2 \ldots$
   a) pick a proposal θ^* from the proposal dist^n $J_t(\theta^* | \theta^{t-1})$

   Proposal dist^n has to be symmetric ie.

   $J_t(\theta_a | \theta_b) = J_t(\theta_b | \theta_a) \; \forall \; \theta_a \theta_b$



b) Calculate acceptance ratio

$$r = \frac{p(\theta^t|y)}{p(\theta^{t-1}|y)}$$

c) set

$$\theta^t = \begin{cases} \theta^* & \text{with prob } \min(r,1) \\ \theta^{t-1} & \text{otherwise.} \end{cases}$$

→ No need of normalization coeff. because they will cancle out while calculating r.

1 Prove that simulated series is a Markov chain which has unique stationary distribution.

→ a) Irreducible : +ive prob. of eventually reaching any state from any other state.

We have continuous dist^n so from any state we can go to any other state

b) Aperiodic : return times are not periodic
Holds for a random walk on any proper dist^n.

$\pi_k = p(\theta^t | \theta^{t-k}, y)$ → If for any k the prob. is 1
Since every state has some +ive prob. for jump in random walk so the prob is near 1

Date
Page

→ Recurrent/Non transient : Prob to return to a state i is 1.
- holds for a random walk on any proper dist^n

2. Prove that the stationary dist^n is the target dist^n $p(\theta|y)$

Consider $\theta^{t-1} \sim p(\theta|y)$

Consider any two such points θ_a and θ_b drawn from $p(\theta|y)$ and labeled so that $p(\theta_b|y) \geq p(\theta_a|y)$

$$p(\theta^{t-1}=\theta_a, \theta^t=\theta_b) = p(\theta_a|y) \cdot J(\theta_b|\theta_a)$$

→ $p(\theta^{t-1}=\theta_b, \theta^t=\theta_a)$

Due to this the α is not 1

$$= p(\theta_b|y)\, J(\theta_a|\theta_b) \cdot \alpha \quad \left(\alpha \to \frac{p(\theta^t=\theta_a^*)}{p(\theta^{t-1}=\theta_b)}\right)$$

$$= p(\theta_a|y) \cdot J(\theta_a|\theta_b).$$

The marginal to be in θ_a at time t & $t-1$ is same.

Metropolis - Hastings Algorithm

Generalization of metropolis algo with non symmetric proposal dist^n.

$$\alpha = \frac{p(\theta^*|y) / J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y) / J_t(\theta^{t-1}|\theta^*)}$$

Ex:- Using Gibbs Sampler to sample from the posterior density of μ and σ^2 in an 1D normal distribution

Posterior distribution of μ & σ^2 $\quad f(\mu, \sigma^2 | y)$

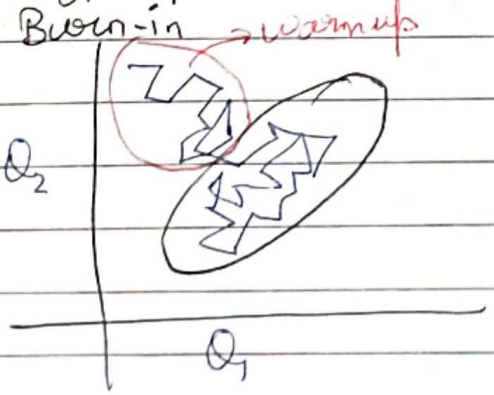
Posterior distribution of μ

Problems with Metropolis Algorithm.

Usually doesn't scale well to high dimension. as there is exponentially growth of search space.

In the higher dimension it is difficult to get the correct proposal distribution due to which a most of the sampled points will be rejected so the efficiency of the algorithm reduces.

(Warm-up / Burn-in : Remov draws from the begining of the chain



Using several chains (To cover the distribution well)

We have to remove warm up and run the chain until they get well mixed and undistinguishable.

Sometimes in visual representation due to shrink in scale we can get false idea about convergence.

Date
Page

$\hat{R}$: Comparision of within and between variances of the chains

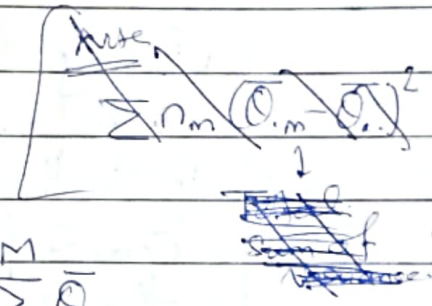
M chain each having N draws

(within chain variance (W)

$$W = \frac{1}{M}\sum_{m=1}^{M} s_m^2 \quad , \quad s_m^2 = \frac{1}{N-1}\sum_{n=1}^{N}(\theta_{nm} - \bar{\theta}_{.m})^2$$

Between chain variance (B)

$$B = \frac{N}{M-1}\sum_{m=1}^{M}(\bar{\theta}_{.m} - \bar{\theta}_{..})^2$$



$$\bar{\theta}_{.m} = \frac{1}{N}\sum_{n=1}^{N}\theta_{nm} \quad , \quad \bar{\theta}_{..} = \frac{1}{M}\sum_{m=1}^{M}\bar{\theta}_{.m}$$

$$\widehat{var}^{+}(\theta|y) = \frac{N-1}{N}W + \frac{1}{N}B$$ → This overestimates marginal posterior variance if the starting points are overdispersed

$$\hat{R} = \sqrt{\frac{\widehat{var}^{+}}{W}}$$

For finite N, W underestimates marginal variance

It should be very near to 1 (~~0.95~~ ~~0.99~~)

if $R > 1.01$, keep sampling.

If $\hat{R}$ close to 1, it is still possible that chain have not converged:

→ If starting points are not overdispersed
→ dist^n far from normal (inf variance)
→ just by chance.

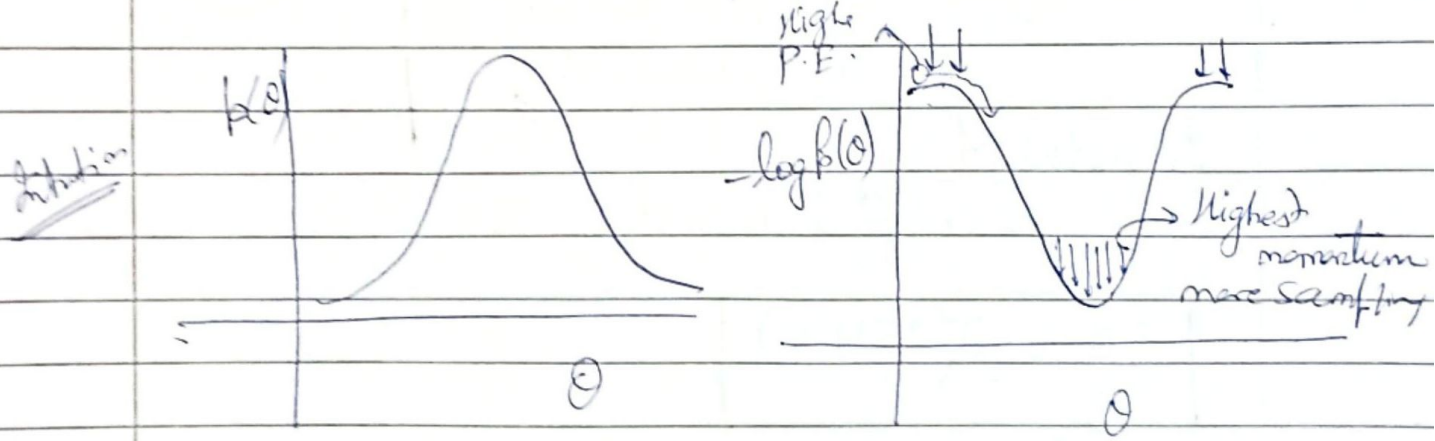
Date
Page

$\hat{R}$ - 1.2

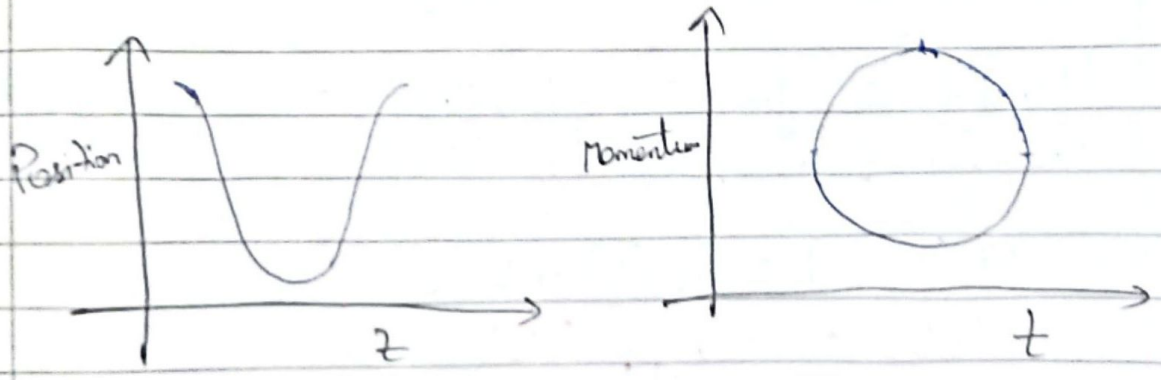
We remove the warm-up then divide the remaing chain into two halves and then we test whether the both the half have similar dist^ (mean & vari-)
This makes the assurance that the series has become Stationary.

Hamiltonian Monte Carlo.

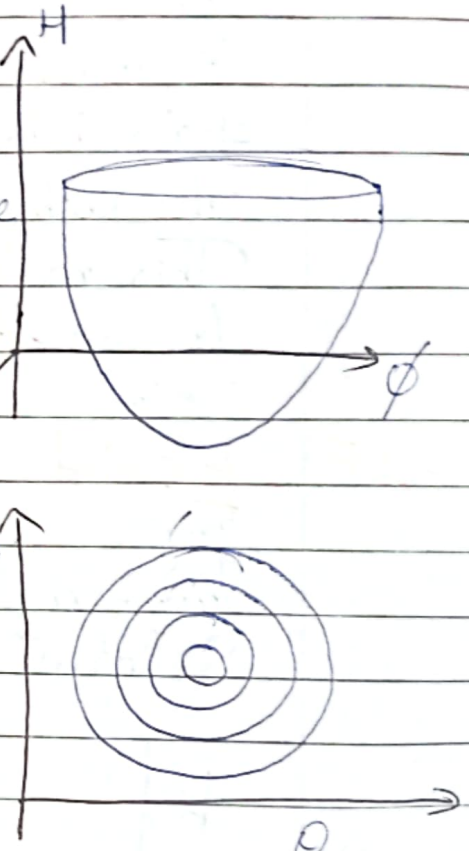
faster and efficient than metropolis algorithm.
Scale well with multiple dimension.



- Hamiltonian Equation
  - operates on a d-dimensional position vector θ and a d-dimensional momentum vector ϕ
    $\theta \in \mathbb{R}^d$ and $\phi \in \mathbb{R}^d$



So the full state given as $(\theta, \phi) \in \mathbb{R}^{2d}$
Described by a function $H(\theta, \phi)$



Equations of the motion
↳ How θ and ϕ changes over time

Given by hamiltonian equations

$$\frac{d\theta}{dt} = \frac{\partial H}{\partial \phi} \qquad \frac{d\phi}{dt} = \frac{-\partial H}{\partial \theta}$$

for d-dimension

$$\frac{d\theta_i}{dt} = \frac{\partial H}{\partial \phi_i} \quad (i) \qquad \frac{d\phi_i}{dt} = \frac{-\partial H}{\partial \theta_i} \quad (ii)$$

for $i = 1, 2, \cdots d$

In matrix form-
Combine θ and ϕ in a vector $z = (\theta, \phi) \in \mathbb{R}^{2d}$

$$\boxed{\frac{\partial z}{\partial t} = J \nabla H(z)}$$

→ $\nabla H(z)$ is the gradient of H i.e.

$$[\nabla H(z)]_k = \frac{\partial H}{\partial z_k}$$

$$J = \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ I_{d\times d} & 0_{d\times d} \end{bmatrix}_{2d\times 2d} \in \mathbb{R}^{2d\times 2d}$$

Hamiltonian function gives the total energy which is the sum of the potential & kinetic energy.

$$H(\theta, \phi) = U(\theta) + K(\phi)$$

$$U(\theta) = -\log p(\theta|y)$$

↳ Potential Energy.

$$K(\phi) = \frac{1}{2}\phi^T M^{-1}\phi$$

[Analogy $KE = \frac{p^2}{2m}$]

- M is the mass matrix
  - Symmetric, Positive Definite
- Typically M is considered to be diagonal with a multiple of the identity matrix I $\begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix}$ this type.

$$\hookrightarrow K(\phi) = -\log p(\phi)$$
where $p(\phi) = \mathcal{N}(0, M)$

Thus equations (I) and (II)

$$\frac{d\theta_i}{dt} = \frac{\partial H}{\partial \phi_i} = \frac{\partial K(\phi)}{\partial \phi} = M^{-1}\phi$$

$$\frac{\partial K(\phi)}{\partial \phi} = \frac{1}{2}\left[(M^{-1}\phi)\frac{\partial \phi}{\partial \phi} + (M^{-1}\phi)\frac{\partial \phi}{\partial \phi}\right]$$

$$= \frac{2}{2}(M^{-1}\phi) = M^{-1}\phi$$

$$\frac{\partial \phi_i}{dt} = \frac{-\partial H}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta} = -\frac{\partial(-\log(p(\theta|y))}{\partial \theta_i}$$

↳ Gradient of the posterior

If we will solve the above differential Eq"s -

$\theta_i(7+s) = f(\theta(7))$

$\phi_i(7+s) = g(\phi(7))$

Now the sampling will become deterministic and that's the benifit of this method.

- For a 1D case:

  θ and ϕ are scalar.

$$H(\theta, \phi) = U(\theta) + K(\phi)$$

Let $U(\theta) = \theta^2/2$

$U(\theta) = -\log(f(\theta/y)) = \theta^2/2$ (Standard normal as posterior)

$K(\phi) = \phi^2/2$ [This is ~~Standard Normal~~ which we generally use]

$$\left.\begin{aligned} \frac{d\theta}{dt} &= \frac{\partial H}{\partial \phi} = \phi \\ \frac{\partial \phi}{\partial t} &= \frac{-\partial H}{\partial \theta} = -\theta \end{aligned}\right\} \text{Equations of SHM.}$$

$$\theta(t) = A\cos(t) + B\sin(t)$$
$$\phi(t) = -A\sin(t) + B\cos(t)$$

[Initial Conditions
At $t=0$, $\theta(0) = \theta_0$
$\phi(0) = \phi_0$]

$\theta_0 = A$; $\phi_0 = B$

$$\boxed{\begin{aligned} \theta(t) &= \theta_0\cos(t) + \phi_0\sin(t) \\ \phi(t) &= -\theta_0\sin(t) + \phi_0\cos(t) \end{aligned}}$$

Let $\theta_0 = r\cos a$ $\phi_0 = -r\sin a$

$$\boxed{\begin{aligned} \theta(t) &= r\cos(a+t) \\ \phi(t) &= -r\sin(a+t) \end{aligned}} \qquad \begin{aligned} r^2 &= \theta_0^2 + \phi_0^2 \\ \tan a &= -\frac{\phi_0}{\theta_0} \end{aligned}$$

Certain properties of Hamiltonian Monte Carlo dynamics.

a) **Reversible**: Knowledge of $\theta(t)$ gives a one-to-one mapping to $\theta(t+s)$. Hence path can be traced.

Date
Page

b) Conservation of Hamiltonian

$H(\theta, \phi)$ we be conserved for a ϕ and θ. It will follow a fixed contour so it will efficiently cover the contour.

c) Volume preservation

If we take a set of θ and ϕ to constitute a region If the θ and ϕ moves then the region will move with constant volume.

d) Symplecticity

Ensure that coverage based on dynamics is good enough to get enough values of θ & ϕ.

Discretizing Hamiltonian Equations.

Let $K(\phi) = \sum_{i=1}^{d} \frac{\phi_i^2}{2m_i}$

- Simple Method
  - Euler method

$$\phi_i(\tau+\epsilon) = \phi_i(\tau) + \epsilon \frac{d\phi_i}{d\tau} \cdots$$

$$= \phi_i(\tau) - \epsilon \frac{\partial H}{\partial \theta_i}\Big|_{\theta_i(\tau)}$$

$$= \phi_i(\tau) - \epsilon \frac{\partial U}{\partial \theta_i}\Big|_{\theta_i(\tau)}$$

$$\theta_i(\tau+\epsilon) = \theta_i(\tau) + \epsilon \frac{\partial \theta_i(\tau)}{\partial \tau}$$

$$= \phi_i(\tau) + \epsilon \frac{\partial H}{\partial \phi}$$

$$= \theta_i(\tau) + \epsilon \frac{\phi_i(\tau)}{m_i}$$

If we take small ϵ then it will converge to the original contour lines but if it is large then it will deviate from the original contours

The spiraling out is the main problem in this approach.

- Modified Euler Method

$$\phi_i(\tau+\epsilon) = \phi_i(\tau) - \epsilon \frac{\partial U}{\partial \theta_i}\Big|_{\theta_i(\tau)}$$

$$\theta_i(\tau+\epsilon) = \theta_i(\tau) + \epsilon \frac{\theta_i(\tau+\epsilon)}{d\tau}$$ → Taking the gradient of the same step.

$$\theta_i(\tau) + \epsilon \frac{\phi_i(\tau+\epsilon)}{m_i}$$

- Leapfrog Integration

$$\phi_i(\tau+\epsilon/2) = \phi_i(\tau) - \epsilon/2 \frac{\partial U}{\partial \theta_i}\Big|_{\theta_i(\tau)}$$

$$\theta_i(\tau+\epsilon) = \theta_i(\tau) + \epsilon \frac{\phi_i(\tau+\epsilon/2)}{m_i}$$

$$\phi_i(\tau+\epsilon) = \phi_i(\tau+\epsilon/2) - \epsilon/2 \frac{\partial U}{\partial \theta_i}\Big|_{\theta_i(\tau+\epsilon)}$$

Problem If there is a peak in a gradient then with longer stepsize

Problem
In case of high divergence leapfrog integrator error increases. If small step size → very huge time req. If very large stepsize then ~~jumps to~~ diverges more.

Sol? Use an adaptive method
- Step size can be high for lower divergence to move faster
- Decrease step size adaptively for high divergence

This is followed in NUTS (No U-Turn Sampling)
→ Sampling after L steps.

After getting ϕ^* & θ^* after leapfrog we can apply metropolis hasting like method to accept it with

$$r = \frac{p(\theta^*|y)\, p(\phi^*)}{p(\theta^{t-1}|y)\, p(\phi^{t-1})}$$

$$\theta^t = \begin{cases} \theta^* & \text{with prob. } \min(r,1) \\ \theta^{t-1} & \text{otherwise} \end{cases}$$

Hierarchial Model

$p(\theta_j\|\tau)$	θ_1	θ_2	θ_3	--- θ_n	parameters
	↓	↓	↓	↓	
$p(y_{ij}\|\theta_j)$	y_{i1}	y_{i2}	y_{i3}	y_{in}	observation

Joint posterior

$$p(\theta, \tau | y) \propto p(y | \theta, \tau) \cdot p(\theta, \tau)$$
$$p(y|\theta) \cdot p(\theta|\tau) \cdot p(\tau)$$

Multiply and Marginalize to get the likelihood.

$$\begin{cases} \theta_j | \alpha, \beta \sim \text{Beta}(\theta_j | \alpha, \beta) \\ y_j | n_j, \theta_j \sim \text{Bin}(y_j | n_j, \theta_j) \end{cases}$$

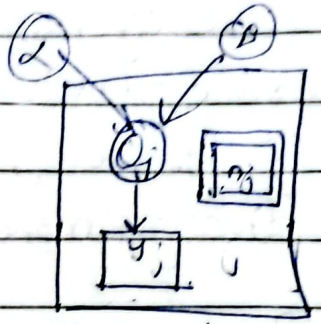


Plate diagram for j^{th} hospital.

Joint posterior

$$p(\theta_1, \theta_2 \ldots \theta_n, \alpha, \beta | y)$$

$$= \prod_{j=1}^{J} p(\theta_j | \alpha, \beta, y)\, p(\alpha, \beta | y)$$

$$p(y|\alpha,\beta)\, p(\alpha,\beta)$$

→ Apply sampling

$$\propto (\alpha+\beta)^{-5/2}$$

Hierarchical model has less variance for smaller value of N.

If ϕ denotes the hyperprior parameters & θ denotes the parameters

- Joint posterior distribution $p(\theta, \phi | y)$

  Joint prior distribution
  $$p(\theta, \phi) = p(\phi)\, p(\theta|\phi)$$

  → $$p(\theta, \phi | y) \propto p(\theta, \phi)\, p(y | \theta, \phi)$$
  $$\propto p(\theta, \phi)\, p(y|\theta)$$

Hierarchichal model works in case of independent systems which are exchangable

Date
Page

- Marginal posterior density $\int_{\theta} p(\theta, \phi | y) d\theta$

$$p(\theta, \phi | y) = p(\theta | \phi, y)\, p(\phi | y)$$

$$\text{or } p(\phi | y) = \frac{p(\phi, \theta | y)}{p(\theta | \phi, y)}$$

The rat tumor model

- Data from 71 experiment
  - Each experiment follows an independent binomial distribution.

$$y_j \sim Bin(\theta_j ; n_j)$$

y_j: No. of rats with (+)ve outputs; n_j: is the total number of rats.

θ_j is assumed to be independent samples from a prior beta distribution.

$$\theta_j \sim Beta(\alpha, \beta)$$

Joint posterior distribution of $p(\theta, \alpha, \beta)$ where $\theta = (\theta_1, \theta_2, \theta_3 \ldots \theta_j)$

$$\propto p(\theta, \alpha, \beta)\, p(y | \theta, \alpha, \beta)$$

$$\propto p(\alpha, \beta)\, p(\theta | \alpha, \beta)\, p(y | \theta, \alpha, \beta)$$

$$p(\theta_1, \theta_2 \ldots \theta_j | \alpha, \beta) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1} (1-\theta_j)^{\beta-1}$$

$$p(y | \theta, \alpha, \beta) \propto \prod_{j=1}^{J} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j}$$

Date
Page

$$\propto p(\alpha,\beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \cdot \prod_{j=1}^{J} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j}$$

$$\propto p(\alpha,\beta) \prod_{j=1}^{J} \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}$$

Conditional Posterior Distribution $[(\theta|\alpha,\beta), y]$

$$p[(\theta|\alpha,\beta), y] = p(\theta|\alpha,\beta)\, p(y|[\theta|\alpha,\beta])$$

$$\propto \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \binom{n_j}{y_j} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j}$$

$$\prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \cdot \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}$$

Marginal posterior distribution $p(\alpha,\beta|y)$

$$p(\alpha,\beta|y) = \frac{p(\alpha,\beta,\theta|y)}{p(\theta|\alpha,\beta,y)}$$

(Dividing the eq^n that we got)

$$\propto p(\alpha,\beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \Big/ \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}$$

$$p(\theta|y) \propto p(y|\theta)\cdot p(\theta)$$

$$p[y|\theta(\alpha,\beta)]\, p(\theta|\alpha,\beta)$$

Model Checking

- Internal Validation (On the Slides)
- External validation

Date
Page

External Validation

Suppose observed data or

$$y = y_1, y_2, \ldots y_n$$

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

Measures

1) Mean Square Error : $\frac{1}{n}\sum (y_i - \mathbb{E}(y_i|\theta))^2$ ← Change the θ and generate diff y_i

2) Weighted MSE : $\frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \mathbb{E}(y_i|\theta))^2}{\text{Var}(y_i|\theta)}$

Advantages of these techniques :

- Easy to compute.

Disadvantage

Doesn't perform well for models other than normal.

2) Log predictive density or Log Liklihood or (LPD)

$\log(p(y|\theta))$ → For normal dist^n the lpd is same as MSE.

- When n is large then the lpd converges towards the a value of $p(y|\theta)$ that follows the highest value of $\log(y|\theta_{MAP})$ that follows the highest posterior density.

Date
Page

Information Criterion Based Measures.

- Akaike Information Criterion:

- Find the Expected log predictive density based on observed data → elpd.
- Find the measure of penality to subtract.
  - k = no. of parameters of the estimated model.

$$\log f(y|\hat{\theta}_{MLE})$$

$$E(lpd) = \int \log f(y|\hat{\theta}_{MLE}) \cdot f(y)\, dy$$

$f(y)$ → Original / True dist° of y (data)
This is unknown.

Estimating elpd

$$\log(f(y|\hat{\theta}_{MLE})) - k$$

$$AIC = -2\log(f_y(\hat{\theta}_{MLE})) + 2k.$$

Disadvantages

→ Not much effective for non-linear model. Only work for linear model with flat prior.

→ For hierarchical model it count the bias so k is too harsh penalty.

Date
Page

Deviance Information Criterion (DIC)

(2 changes than AIC)

1) Log predictive Density, uses $\hat{\theta}_{MAP}$ $\log p(y|\hat{\theta}_{MAP})$

2) Rather than using k as penalty term it uses the variance term of $\log p(y|\theta_j)$ ← Penalty Term

$$DIC = -2 \log p(y|\hat{\theta}_{MAP}) + 2 p_{DIC}$$

$$p_{DIC} = \underset{j=1}{\overset{S}{Var}} \log p(y|\theta_j)$$

More var high uncertainty so more penalty.

Watanbe-Akaike Information Criterion (WAIC)

→ Use the uncertainity of the measure of θ in the lpd term.

→ Uses the variance of each data point.

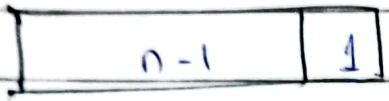
$$WAIC = -2 \sum_{i=1}^{n} \log \frac{1}{S} \sum_{j=1}^{S} p(y_i|\theta_j) + 2 p_{WAIC}$$

$$p_{WAIC} = \sum_{i=1}^{n} \underset{j=1}{\overset{S}{Var}} \log(p(y_i|\theta_j))$$

First we will calculate posterior (we can sample S data points and then calculate the mean of likliehood log(), Sum it over i

Date
Page

Leave one out - Cross Validation (Loo-Cv)



Take n-1 data point for training and use the rest 1 data as out of sample testing

No need for penalty.

Leave 1.

$$\text{Loo-Cv} \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{j=1}^{S} p(y_i | y_{-i}; \theta_j)\right)$$

Which measure is better ??