# Bayesian Computational Statistics

reference book: Bayesian Data Analysis; Gelman et al.

## Bayes' rule and its consequences

conditional probabilities

$P(A|B)$ probability of event A given that event B has occured.

$$P(A|B) = P(A \cap B) / P(B)$$

example: We have a standard 6 sided die

$$A = \{5\}, \quad B = \{1, 3, 5\}$$

$$P(A|B) = P(A \cap B) / P(B) = 1/3$$

$$P(A) = 1/6$$

$$C = \{2, 4, 6\} \quad \text{so } P(A \cap C)/P(C) = 0/3 = 0$$

Note: $P(A|B) \neq P(B|A)$

Bayes' rule relates $P(A|B)$ to $P(B|A)$:

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

It gives us a framework for making and updating estimates of $P(A|B)$ based on evidence.
Updating our beliefs in the face of new information.

<u>example</u>: medical testing

event A: having the disease
event B: testing positive

prior $P(A)$ : 0.01 (1%)

likelihood $P(B|A)$ : 0.99

false positive rate to compute the marginal: 0.05

marginal: $P(B) = P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)$

$= 0.99 \cdot 0.01 + 0.05 \cdot 0.99$

$\approx 0.0594$

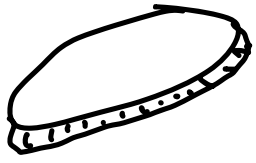
posterior $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{0.99 \cdot 0.01}{0.0594} \approx \frac{0.16}{2}$

a bit surprising.

## Bayesian inference

1. Start with a prior distribution
2. Collect data
3. Compute the likelihood
4. Compute the marginal probability
5. Compute the posterior via Bayes' rule (updated beliefs)

example: Is it a fair coin?



$H_0$: It is fair
$H_1$: It is biased

prior: 0.5

Data: T, T, T

$$P(H_0 \mid TTT) = \frac{P(TTT \mid H_0)\, P(H_0)}{P(TTT)}$$

$$P(TTT) = P(TTT \mid H_0) \cdot P(H_0) + P(TTT \mid H_1) \cdot P(H_1)$$

$$P(TTT \mid H_1) = \int_0^1 p^3 dp = \frac{p^4}{4}\Big|_0^1 = 1/4, \quad \text{so}$$

$$P(H_0 \mid TTT) = 1/3$$

# Fundamentals of Bayesian Inference

Bayes' Rule $\quad P(A|B) = \dfrac{P(B|A) \cdot P(A)}{P(B)}$

Quizz exercise:

| | | prior | data |
|---|---|---|---|
| $H_0$ : | coin is fair | 0.66 | HHHHH |
| $H_1$ : | coin is biased | 0.34 | (5H) |

likelihood:

$$P(5H \mid H_0) = \left(\frac{1}{2}\right)^5 = 1/32 = 0.03125$$

$$P(5H \mid H_1) = \int_0^1 p^5 \, dp = \left[p^6/6\right]_0^1 = 1/6$$

marginal: $\quad P(5H) = P(5H \mid H_0) \cdot P(H_0) + P(5H \mid H_1) \cdot P(H_1)$

$$= 0.03125 \cdot 0.66 + (1/6) \cdot 0.34$$

$$\approx \underline{0.078}$$

posterior:

$$p(H_0 \mid 5H) = \frac{P(5H \mid H_0) \cdot P(H_0)}{P(5H)}$$

$$= \frac{0.03125 \cdot 0.66}{0.078} \approx \underline{0.268}$$

Bayesian Inference:

process of fitting a probability model to a set of data using Bayes' rule

Notation:

$\theta$ : parameter, scalars or vectors
e.g $\theta = (\beta_0, \beta_1)$

$y$ : observed data

$\tilde{y}$ : unknown but potentially observable data

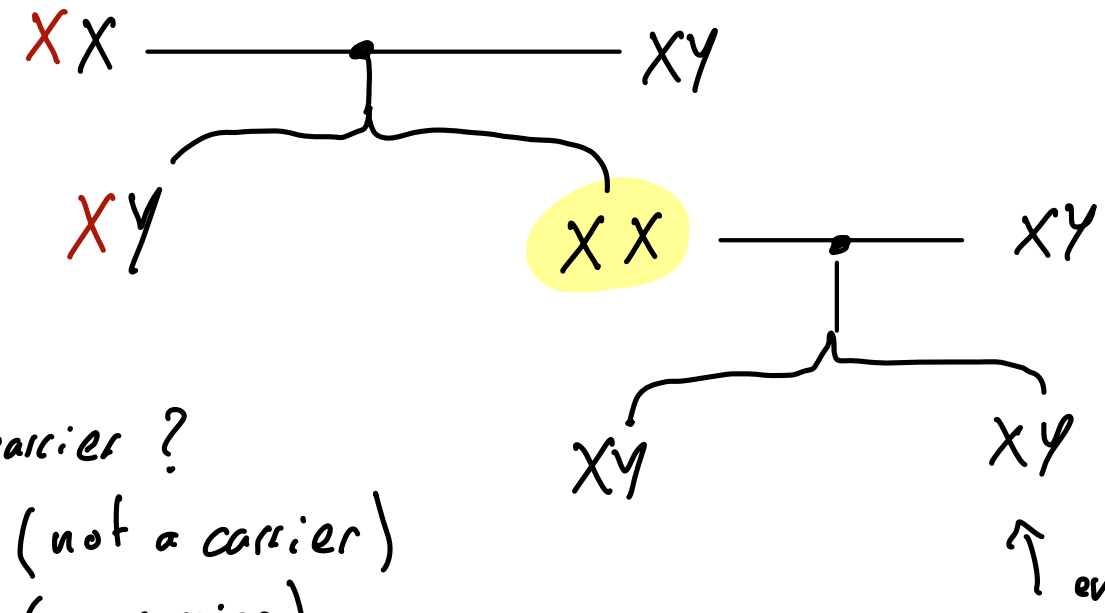
$p(x)$ : pdf of $x$

$p(x, y)$ : joint distribution of $x$ and $y$

Hemophilia example:

from the book

$x$ - linked trait




Yellow is a carrier ?

$H_0$ $\theta = 0$ (not a carrier)

$H_1$ $\theta = 1$ (a carrier)

data : $\bar{y} = (0, 0)$

prior : (50/50) $= 0.5$

likelihood : $p(\bar{y} \mid \theta = 1) = 1/2 \cdot 1/2 = 1/4$

$p(\bar{y} \mid \theta = 0) = 1 \cdot 1 = 1$

marginal (for scaling) : $p(\bar{y}) = p(\bar{y} \mid \theta = 1) \cdot p(\theta = 1) + p(\bar{y} \mid \theta = 0) \cdot p(\theta = 0)$

$= 1/4 \cdot 1/2 + 1 \cdot 1/2 = \underline{5/8}$

posterior :

$$p(\theta = 1 \mid \bar{y}) = \frac{1/8}{5/8} = 1/5 = \underline{0.2}$$

What if there is a third child who is also XY and not afflicted. (update)

$$\bar{y} = (0)$$

$$p(\theta = 1 \mid y) = \frac{p(y \mid \theta = 1) \cdot p(\theta = 1)}{p(y)}$$

$$= \frac{(1/2) \cdot (1/5)}{\frac{1}{2} \cdot \frac{1}{5} + 1 \cdot 4/5} = 1/9 \approx 0{,}11$$

## Exchangeability

order of observations doesn't matter

Subjechivity (vs) objectivity

⇒ prior

## Example from the book

Spelling corrections

'radom' → 'random', 'radon', 'radom'

data : $y$ = radom

$$p(\theta \mid y) \propto p(\theta)\, p(y \mid \theta)$$

scaling can be done easily at the end

prior :

| $\theta$ | rel. freq | prob |
|---|---|---|
| random | $7.6 \cdot 10^{-5}$ | |
| radon | $6.05 \cdot 10^{-6}$ | |
| radom | $3.12 \cdot 10^{-7}$ | |

$\Rightarrow$ rewriting

| $\theta$ | rel. freq. | prob |
|---|---|---|
| random | $760 \cdot 10^{-7}$ | 0.923 |
| radon | $60.5 \cdot 10^{-7}$ | 0.073 |
| radom | $3.12 \cdot 10^{-7}$ | 0.004 |

likelihood

| $\theta$ | $p(\text{'radom'} \mid \theta)$ |
|---|---|
| random | 0.00193 |
| radon | 0.000143 |
| radom | 0.975 |

posterior :

| $\theta$ | $p(\theta)\ p(\text{'radom'} \mid \theta)$ | $p(\text{'radom'} \mid \theta)$ |
|---|---|---|
| random | $1.47 \cdot 10^{-7}$ $(1470 \cdot 10^{-10})$ | $\sim 0.325$ |
| radon | $8.65 \cdot 10^{-10}$ | $\sim 0.002$ |
| radom | $3.04 \cdot 10^{-7}$ $(3040 \cdot 10^{-10})$ | $\sim 0.673$ |

example where we don't need marginals because we can scale the results at the very end.

# Bayesian Computation

Faulty caliper problem from the book

length in millimeters

off by $-1$ mm or $1$ mm

| | | prior |
|---|---|---|
| $\theta = 1$ | $y \sim N(1,1)$ | 0.5 |
| $\theta = -1$ | $y \sim N(-1,1)$ | 0.5 |

likelihood:

$$p(y=0.5 \mid \theta=1) = \frac{1}{2\pi} e^{-\left(\frac{(y-1)^2}{2}\right.} \approx \underline{0.1405}$$

marginal:

$$p(y=0.5) = \begin{matrix} p(y=0.5 \mid \theta=1)\ p(\theta=1) + \\ p(y=0.5 \mid \theta=-1)\ p(\theta=-1) \end{matrix}$$

$$\approx \underline{0.09605}$$

posterior:

$$p(\theta=1 \mid y=0.5) = \frac{0.07022}{0.09605} \approx \underline{0.73}$$

Stan package in R

PyStan in python

example in R: estimation of a distribution

# General approach to Bayesian Computation

## Binomial and Posterior Distributions

Binary data 0, 1 Bernoulli outcomes

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

$y$ : number of success
$\theta$ : proportion of success

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$


Biased $\theta = 0.75$
$H = 1, T = 0$

What is the probability of TTT ?

$$p(y=0, n=3 \mid \theta=0.75) = \binom{3}{0} 0.75^0 0.25^3 \approx 0.016$$

$$p(y=1, n=3 \mid \theta=0.75) \approx 0.14$$

{TTH or THT or HTT}

example 2

$\theta$ : proportion of female birth

$y$ : number of female birth in $n$ recorded births

$\theta \sim U_{[0,1]}$

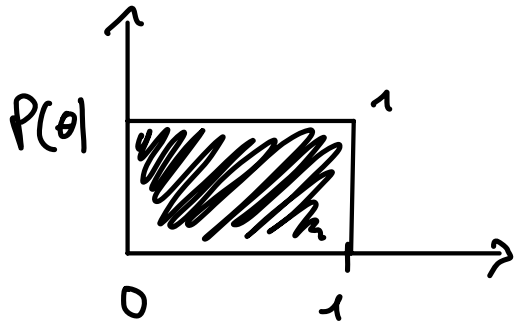
What is the posterior distribution ?

Binomial likelihood.

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

uniform prior :

$p(\theta) = 1$ for $\theta \in [0,1]$




posterior :

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

$$= \binom{n}{y} \theta^y (1-\theta)^{n-y} \qquad \theta \in [0,1]$$

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

once normalized

$$p(\theta|y) = B(\underbrace{y+1}_{\alpha}, \underbrace{n-y+1}_{\beta})$$

Beta distribution

$\tilde{y}$ : predictions with the m next births

$$\tilde{y} \sim Bin(m, \theta)$$

so $\theta \sim$ Beta and no longer Uniform

$$p(\tilde{y} \mid y)$$

So $$p(\tilde{y} \mid y) = \int p(\tilde{y}, \theta \mid y)\, d\theta$$

$$= \int p(\tilde{y} \mid \theta y)\, p(\theta \mid y)\, d\theta$$ chain rule of probability

$$= \int \underbrace{p(\tilde{y} \mid \theta)}_{\text{prediction}} \underbrace{p(\theta \mid y)}_{\text{posterior}}\, d\theta$$ } posterior predictive distribution

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

$$p(\theta \mid y) = Beta(y+1, n-y+1)$$

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

$$p(\tilde{y} \mid y) = \int \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \binom{m}{\tilde{y}} \theta^{y+\tilde{y}} (1-\theta)^{n+m-(\tilde{y}+y)}$$

$$= \binom{m}{\tilde{y}} \frac{\Gamma(n+2)}{\Gamma(y+1)\,\Gamma(n-y+1)} \frac{\Gamma(y+\tilde{y}+1)\,\Gamma(m+n-\tilde{y}-y+1)}{\Gamma(m+n+2)}$$

$$\cdot \int \frac{\Gamma(m+n+2)}{\Gamma(y+\tilde{y}+1)\,\Gamma(m+n-\tilde{y}+y+1)} \theta^{y+\tilde{y}} (1-\theta)^{m+n-(\tilde{y}+y)}\, d\theta$$

$= 1$

Beta distribution

$$= \binom{m}{\tilde{y}} \frac{\Gamma(n+2)}{\Gamma(y+1)\,\Gamma(n-y+1)} \frac{\Gamma(y+\tilde{y}+1)\,\Gamma(m+n-\tilde{y}-y+1)}{\Gamma(m+n+2)}$$

$$= \binom{m}{\tilde{y}} \frac{(n+1)!\,(y+\tilde{y})!\,(m+n-\tilde{y}+y)!}{y!\,(n-y)!\,(m+n+1)!}$$

$$= \binom{m}{\tilde{y}} \frac{n+1}{m+n+1} \frac{n!}{y!\,(n-y)!} \frac{(y+\tilde{y})!\,(m+n-(y+\tilde{y}))!}{(m+n)!}$$

$(m+n+1)(m+n)!$

$$= \binom{m}{\tilde{y}} \binom{n}{y} \begin{bmatrix} m+n \\ y+\tilde{y} \end{bmatrix}^{-1} \frac{n+1}{m+n+1}$$

$n$ = # of data points

$y$ = # of successes

$m$ = # of predicted points

$\tilde{y}$ = # of predicted successes

prediction : next birth is female ?

$$m = \tilde{y} = 1$$

$$p(\tilde{y}=1, m=1 \mid y) = 1 \cdot \binom{n}{y} \left[\binom{n+1}{y+1}\right]^{-1} \frac{n+1}{n+2}$$

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

$$\simeq \boxed{\frac{y+1}{n+2}}$$

$\frac{y}{n}$ : data

$\frac{1}{2}$ : mean of our posterior

$\mathrm{Beta}(\alpha, \beta)$ have a mean of $\frac{\alpha}{\alpha+\beta}$

$\Rightarrow$ What is the mean of the posterior distribution ?

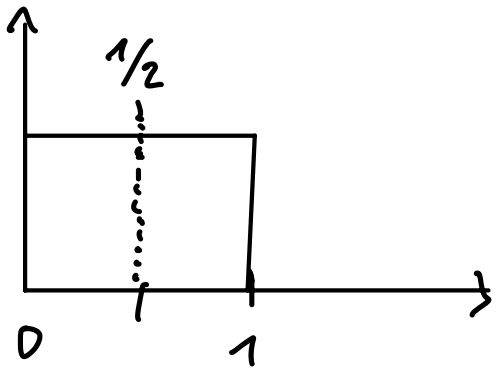
$$p(\theta \mid y) = \mathrm{Beta}(y+1, n-y+1)$$

$$E[\theta \mid y] = \frac{y+1}{n-y+1+y+1} = \boxed{\frac{y+1}{n+2}}$$

If $n$ is close to $0$, then the average is more influenced by the average of the posterior distribution.

As $n$ gets large,

$$E[\theta|y] \sim \frac{y}{n}$$




Given $y$, what is the expected value of $\theta$, $E[\theta|y]$ in this problem?

$$E[\theta|y] = \frac{y+1}{n+2}$$

If data is large

$$E[\theta|y] = \frac{y}{n}$$

$$\text{var}(\theta) = E[\text{var}(\theta|y)] + \text{var}(E[\theta|y])$$

we observe that

$$E[\text{var}(\theta|y)] < \text{var}(\theta)$$

# Priors

Reflect initial information

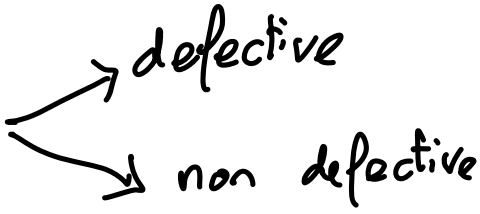
Informative (vs) Noninformative priors

The prior should encompass all possible values

## Example:

defact rate : assumed 5% with variance 0.25%

What could be a good informative prior ?




$\Rightarrow$ likelihood is Binomial

conjugate prior :

$$\theta \sim \text{Beta}(a, b) \qquad \mu = \frac{a}{a+b}$$

$$\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

solution :

$$\frac{a}{a+b} = 0.05 \quad \Rightarrow \quad 19a = b$$

$$0.0025 = \frac{ab}{(a+b)^2(a+b+1)}$$

$$= \frac{19}{400(20a+1)}$$

$$\Rightarrow \quad a = \frac{18}{20} = 0.9$$

$$b = 17.1$$

So the prior is $\theta \sim B(0.9, 17.1)$

Binomial distribution

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

uniform prior $p(\theta) = 1$ for $\theta \in [0,1]$

informative prior $\text{Beta}(a, b)$

$$p(\theta|y) \propto \underbrace{\theta^y (1-\theta)^{n-y}}_{\text{likelihood}} \underbrace{\theta^{a-1} (1-\theta)^{b-1}}_{\text{prior}}$$

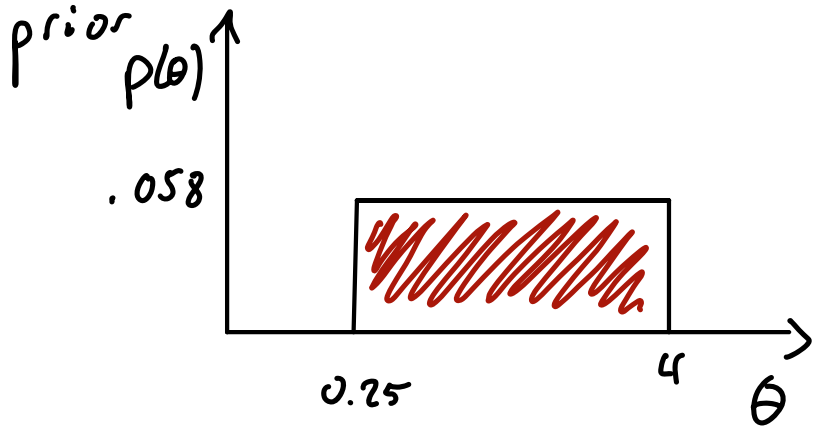
$$= \theta^{y+a-1} (1-\theta)^{n-y+b-1}$$

$$\theta|y \sim \text{Beta}(y+a, n-y+b)$$

$\Rightarrow$ conjugacy

# Nonconjugate prior distributions

## Normal likelihood and Uniform prior




Uniform

$$p(\theta) = \begin{cases} 0.308 & \text{if } \theta \in [0.25, 4] \\ 0 & \end{cases}$$

posterior

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

$$\propto \begin{cases} \exp\left(-\frac{(y-\mu)^2}{2}\right) & \theta \in [0.25, 4] \\ 0 & \text{otherwise} \end{cases}$$

truncated Normal distribution (non conjugate)

# Weakly informative priors

## Quizz

(2) $p(\theta) \sim B(2,2)$

$p(y|\theta) \sim Bin(10, 1/2)$

$y = 3H$

$y = 3 \quad n = 10 \quad a = 2, \; b = 2$

$p(\theta|y) \sim Beta(a+y, n-y+b)$

$\sim Beta(5, 9)$

(1) $p(\theta) \sim B(2,2)$

$y < 3, \; n = 10$

$p(y|\theta) \sim Bin(n, K)$

$p(\theta|y) \sim \theta(1-\theta)^9 (1 + 8\theta + 36\theta^2)$

## Other single-parameter models

example.

likelihood $p(y|\theta) \propto e^{-(y-\theta)^2/2\sigma^2}$

prior: $p(\theta) \propto e^{A\theta^2 + B\theta + C}$

aim: show that this prior is a normal distribution

$$A\theta^2 + B\theta + C = A\left(\theta + B/2A\right)^2 - B^2/4A^2 + C$$

$$p(\theta) \propto e^{A\theta^2 + B\theta + C} \propto e^{A(\theta + B/2A)^2}$$

let $A = -1/2\tau_0^2$ $\quad -B/2A = \mu_0$

$$\boxed{p(\theta) \propto e^{-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2}}$$

$$\boxed{\theta \sim N(\mu_0, \tau_0^2)}$$

prior

aim: show that the posterior is also normal

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

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

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

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

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

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

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

$$\frac{1}{\tau_1^2} = \underbrace{\frac{1}{\sigma^2}}_{\text{likelihood}} + \underbrace{\frac{1}{\tau_0^2}}_{\text{prior}} \qquad \mu_1 = \frac{y/\sigma^2 + \mu_0/\tau_0^2}{1/\sigma^2 + 1/\tau_0^2}$$

$$\propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)$$

$$\theta \mid y \sim N(\mu_1, \tau_1^2) \qquad \text{posterior}$$

The inverse variance is called precision.

Now, for new data points $\tilde{y}$

$$p(\tilde{y}\mid y) = \int p(\tilde{y}\mid\theta)\, p(\theta\mid y)\, d\theta$$

$$\propto \int \exp\left(-\frac{(\tilde{y}-\theta)^2}{2\sigma^2}\right)\exp\left(-\frac{(\theta-\mu_0)^2}{2\tau_1^2}\right) d\theta$$

$$\propto \exp\left(A\tilde{y} + B\tilde{y} + C\right)$$

$$\Rightarrow \tilde{y}|y \sim N(\cdot, \cdot)$$

$$\begin{aligned}
E[\tilde{y}|y] &= E[E[\tilde{y}|\theta, y]|y] \\
&= E[E[\tilde{y}|\theta]|y] \\
&= E[\theta|y] \\
&= \mu_1
\end{aligned}$$

$$\begin{aligned}
\text{var}(\tilde{y}|y) &= E[\text{var}(\tilde{y}|\theta)|y] + \text{var}(E[\tilde{y}|\theta]|y) \\
&= E[\sigma^2|y] + \text{var}(\theta|y) \\
&= \sigma^2 + \tau_1^2
\end{aligned}$$

$$\tilde{y}|y \sim N(\mu_1, \sigma^2 + \tau_1^2)$$

How do we handle multiple data points?

Prior: $p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right)$

likelihood:

$$p(y|\theta) \propto \prod_{i=1}^{n} \exp\left(-\frac{(y_i-\theta)^2}{2\sigma^2}\right)$$

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

posterior:

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

$$\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2 - \frac{1}{\tau_0^2}(\theta-\mu_0)^2\right)$$

$$= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\theta\sum_{i=1}^{n} y_i + n\theta^2\right) - \frac{1}{2\tau_0^2}\left(\theta^2 - 2\theta\mu_0 + \mu_0^2\right)\right)$$

Let $\bar{y}$ be the sample mean, so $n\bar{y} = \sum_{i=1}^{n} y_i$

$$= \exp\left(\theta^2\left(-\frac{n}{2\sigma^2} - \frac{1}{2\tau_0^2}\right) - 2\theta\left(\frac{n\bar{y}}{\sigma^2} + \frac{1}{\tau_0^2}\right)\right)$$

$$\tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)^{-1}$$

$$= \exp\left(-\frac{1}{2\tau_n^2}\left(\theta^2 - 2\tau_n^2\left(\frac{n\bar{y}}{\sigma^2} + \frac{1}{\tau_0^2}\right)\theta\right)\right)$$

$$\mu_n = \tau_n^2\left(\frac{n\bar{y}}{\sigma^2} + \frac{1}{\tau_0^2}\right)$$

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

$$\Rightarrow \quad \theta \mid y \sim N(\mu_n, \tau_n^2)$$

<u>example</u>: screws of 5 mm with variation

mean ? variance: $\sigma^2 = 0.5$

Data: $y = \{5.1, 4.9, 5.2, 5.3, 5\}$

$n = 5$ $\quad \tau_0^2 = 1$

$\bar{y} = 5.1$ $\quad \mu_0 = 5$

$\sigma^2 = 0.5$

$$\tau_n^2 = \left(\frac{1}{1} + \frac{5}{0.5}\right)^{-1} = \frac{1}{11} \approx 0.11$$

$$\mu_n = \frac{1}{11}\left(\frac{25.5}{0.5} + \frac{5}{1}\right) = \frac{56}{11} \approx 5.1$$

$$\theta | y \sim N(5.1, 0.1)$$

easy to compute with conjugacy

## Normal distribution with unknown variance

## But mean is known

Review: scaled inverse $\chi^2$ distribution

$$\theta \sim \text{Inv}\,\chi^2(\nu, s^2)$$

($\nu$: deg. of freedom, $s^2$: scale)

$$p(\theta) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} s^{\nu} \theta^{-(\nu/2+1)} e^{\frac{-\nu s^2}{2\theta}}$$

$$\propto \theta^{-(\nu/2+1)} e^{-\frac{\nu s^2}{2\theta}}$$

likelihood:

$$p(y|\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-\theta)^2/2\sigma^2}$$

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

Define: $v := \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$ sample variance

$$p(y|\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(\frac{nv}{2\sigma^2}\right)$$

prior: $$p(\sigma^2) \propto (\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left(\frac{-\nu_0\sigma_0^2}{2}\cdot\frac{1}{\sigma^2}\right)$$

posterior:

$$\begin{aligned} p(\sigma^2|y) &\propto p(\sigma^2)\, p(y|\sigma^2) \\ &\propto (\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0\sigma_0^2}{2}\cdot\frac{1}{\sigma^2}\right) \\ &\quad \cdot (\sigma^2)^{-n/2} \exp\left(-\frac{nv}{2\sigma^2}\right) \\ &= (\sigma^2)^{-\frac{1}{2}(\nu_0+n+2)} \exp\left(-\frac{1}{2\sigma^2}(\nu_0\sigma_0^2+nv)\right) \end{aligned}$$

$$\sigma^2 \sim \text{Inv}\,\chi^2\left(\nu_0+n, \frac{\nu_0\sigma_0^2+nv}{\nu_0}\right)$$

## Poisson distributions

$\theta$: average number of events

$y$: actual observations

review : $\text{Gamma}(a, b)$ $a$ : shape

$b$ : inverse scale

mean : $a/b$

variance : $a/b^2$

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}$$

$$\propto \theta^{a-1} e^{-b\theta}$$

poisson likelihood

$$p(y|\theta) = \prod_{i=1}^{n} \frac{1}{y_i!} \theta^{y_i} e^{-\theta} \propto \theta^{n\bar{y}} e^{-n\theta}$$

prior :

$$p(\theta) \propto e^{-b_0\theta} \theta^{a_0-1}$$

posterior : $p(\theta|y) \propto p(\theta)\, p(y|\theta)$

$$\propto \theta^{n\bar{y}+a_0-1} e^{-n\theta - b_0\theta}$$

$$\theta|y \sim \text{Gamma}(a_0 + n\bar{y},\ b_0 + n)$$

example: 3/200,000 died of asthma (1 year)

world: 0.6/100,000 prior ↖ data

prior: mean of 0.6 $\Rightarrow \frac{a}{b} = 0.6$

suppose: $a = 3$, $b = 5$

posterior: $\text{Gamma}(a_0 + y, b_0 + x) = \text{Gamma}(6, 7)$

mean: $6/7 \approx 0.86$ variance: $6/7^2$

## Exponential distributions

$$p(y|\theta) = \theta^n e^{-n\bar{y}\theta}$$

waiting time or time between events $\theta$.

conjugate prior: Gamma.

## multiparameter models

# Naisance parameters

example

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

Jeffrey's prior (noninformative) for $\mu$ and $\sigma^2$

Jeffrey's invariance principle

$$p(\theta) \propto [J(\theta)]^{1/2} \quad \text{where}$$

$$J(\theta) = -E\left[\frac{\partial^2 \log p(y|\theta)}{\partial \theta^2} \,\middle|\, \theta\right]$$

$$\theta = (\mu, \sigma^2)$$

$$L := p(y|\theta) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

$$\log(L) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-\mu)^2$$

$$= -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}(y-\mu)^2$$

$$\frac{\partial \log(L)}{\partial \mu} = \frac{1}{\sigma^2}(y-\mu)$$

$$\frac{\partial \log(L)}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{\sigma^3}(y-\mu)^2$$

$$\frac{\partial^2 \log(L)}{\partial \mu^2} = -\frac{1}{\sigma^2}$$

$$\frac{\partial \log(L)}{\partial \mu \, \partial \sigma} = -\frac{2(y-\mu)}{\sigma^3}$$

$$\frac{\partial \log(L)}{(\partial \sigma)^2} = \frac{1}{\sigma^2} - \frac{3}{\sigma^4}(y-\mu)^2$$

$$J(\theta) = -E\left[ \det \begin{vmatrix} -1/\sigma^2 & -\frac{2(y-\mu)}{\sigma^3} \\ -\frac{2(y-\mu)}{\sigma^3} & \frac{1}{\sigma^2} - \frac{3}{\sigma^4}(y-\mu)^2 \end{vmatrix} \right]^{1/2}$$

$$= \left( -E\left[ -\frac{1}{\sigma^4} + \frac{3}{\sigma^6}(y-\mu)^2 - \frac{4(y-\mu)^2}{\sigma^6} \right] \right)^{1/2}$$

remember $E[(y-\mu)^2] = \sigma^2$ so,

$$= \left( \frac{1}{\sigma^4} + \frac{1}{\sigma^4} \right)^{1/2}$$

$$= \left( \frac{2}{\sigma^4} \right)$$

$$= \frac{\sqrt{2}}{\sigma^2} \propto \frac{1}{\sigma^2}$$

$p(\theta) \propto \frac{1}{\sigma^2}$ Noninformative prior

posterior:

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

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

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

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

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

$$= \sum_{i=1}^{n}(y_i-\bar{y})^2 + 2(\bar{y}-\mu)\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})}_{=0} + n(\bar{y}-\mu)^2$$

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

Sample variance

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

joint posterior:

$$p(\mu, \sigma^2 | y) \propto (\sigma^2)^{-n/2-1} \exp\left(-\frac{1}{2\sigma^2}\left(s^2(n-1) + n(\bar{y}-\mu)^2\right)\right)$$

We look at the conditional distribution of $\mu$

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

Normalized:

$$p(\mu | \sigma^2, y) = \sqrt{\frac{n}{2\pi\sigma^2}} \exp\left(-\frac{n}{2\sigma^2}(\mu - \bar{y})^2\right)$$

$$\mu | \sigma^2, y \sim N\left(\bar{y}, \sigma^2/n\right)$$ conditional posterior

$\Rightarrow$ marginal posterior:

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

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

$$\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \cdot \underbrace{\int_{-\infty}^{\infty} \exp\left(-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right) d\mu}_{= \sqrt{\frac{2\pi\sigma^2}{n}}}$$

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

$$\propto \sigma^{-n-1} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \Rightarrow \text{Inverse Gamma distribution}$$

$$a = \frac{n}{2} - 1 \quad , \quad b = \frac{(n-1)s^2}{2}$$

## Normal data summary

- $y_i \sim N(\mu, \sigma^2)$, with $\mu, \sigma^2$ unknown
- Noninformative Jeffrey's prior $p(\mu, \sigma^2) \propto 1/\sigma^2$
- joint posterior:
  $p(\mu, \sigma^2 | y) \propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\bar{y}-\mu)^2\right)\right)$
- conditional posterior:
  $\mu | \sigma^2, y \sim N(\bar{y}, \sigma^2/n)$
- marginal distribution:
  $\sigma^2 | y \sim \text{Inv. Gamma}\left(\frac{n}{2} - 1, \frac{(n-1)s^2}{2}\right)$

# Posterior predictive distribution

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

joint posterior:

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

likelihood:

$$\tilde{y}|\mu,\sigma^2 \sim N(\mu,\sigma^2) \quad \text{so...}$$

$$p(\tilde{y}|y) = \int_0^\infty \int_{-\infty}^\infty \sigma^{-1} \exp\left(-\frac{1}{2\sigma^2}(\tilde{y}-\mu)^2\right) \cdot \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\bar{y}-\mu)^2\right)\right)$$

$$= \int_0^\infty \sigma^{-n-3} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \int_{-\infty}^\infty \exp\left(-\frac{1}{2\sigma^2}\left((\tilde{y}-\mu)^2 + n(\bar{y}-\mu)^2\right)\right) d\mu\, d\sigma^2$$

$$(\tilde{y}-\mu)^2 + n(\bar{y}-\mu)^2 = \tilde{y}^2 - 2\tilde{y}\mu + \mu^2 + n\bar{y}^2 - 2n\bar{y}\mu + n\mu^2$$

$$= (n+1)\mu^2 - 2\mu(\tilde{y}+n\bar{y}) + \tilde{y}^2 + n\bar{y}^2$$

$$= (n+1)\mu^2 - 2\mu(\tilde{y}+n\bar{y}) + \frac{(n+1)\tilde{y}^2}{(n+1)} + \frac{(n+1)n\bar{y}^2}{n+1}$$

$$= (n+1)\mu^2 - 2\mu(\tilde{y} + n\bar{y}) + \frac{(n+1)\tilde{y}^2 + (n+1)n\bar{y}^2 + 2n\bar{y}\tilde{y} - 2n\bar{y}\tilde{y}}{(n+1)}$$

$$= (n+1)\mu^2 - 2\mu(\tilde{y} + n\bar{y}) + \frac{\tilde{y}^2 + 2n\tilde{y}\bar{y} + n^2\bar{y}^2 + n\bar{y}^2 - 2n\bar{y}\tilde{y} + n\tilde{y}^2}{(n+1)}$$

$$= (n+1)\mu^2 - 2\mu(\tilde{y} + n\bar{y}) + \frac{(\tilde{y} + n\bar{y})^2 + n(\bar{y} - \tilde{y})}{(n+1)}$$

$$= (n+1)\left[\mu^2 - \frac{2\mu(\tilde{y} + n\bar{y})}{n+1} + \frac{(\tilde{y} + n\bar{y})^2}{(n+1)^2}\right] + \frac{n(\bar{y} - \tilde{y})^2}{n+1}$$

$$= (n+1)\left(\mu - \frac{(\tilde{y} + n\bar{y})^2}{n+1} + \frac{n(\bar{y} - \tilde{y})^2}{n+1}\right)$$

$$= \int_0^\infty \sigma^{-n-3} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}\right.$$

$$\left.(n+1)\left(\mu - \underbrace{\frac{(\tilde{y} + n\bar{y})^2}{n+1}}_{*} + \frac{n(\bar{y} - \tilde{y})^2}{n+1}\right)\right) d\mu \, d\sigma^2$$

$*$ constant w.r.t $\mu$.

$$= \int_0^\infty \sigma^{-n-3} \exp\left( -\frac{(n-1)S^2}{2\sigma^2} - \left(\frac{n}{n+1}\right) \frac{(\bar{y}-\tilde{y})^2}{2\sigma^2} \right) \cdot$$

$$\underbrace{\int_{-\infty}^{\infty} \exp\left( -\frac{n+1}{2\sigma^2} \left( \mu - * \right) \right) d\mu}_{= \sqrt{\frac{2\pi\sigma^2}{n+1}}} d\sigma^2$$

$$= \int_0^\infty \sigma^{-n-2} \sqrt{\frac{2\pi\sigma^2}{n+1}} \exp\left( -\frac{(n-1)S^2}{2\sigma^2} \right) \exp\left( -\frac{1}{2\sigma^2} \left(\frac{n}{n+1}\right) (\bar{y}-\tilde{y})^2 \right) d\sigma^2$$

$$= \sqrt{\frac{2\pi\sigma^2}{n+1}} \int_0^\infty (\sigma^2)^{-n/2-1} \exp\left( -\frac{1}{2\sigma^2} \left( S^2(n-1) + \frac{n}{n+1} (\bar{y}-\tilde{y})^2 \right) \right) d\sigma^2$$

Gamma distribution

$$\int_0^\infty x^{a-1} e^{-bx} dx = \Gamma(a) / b^a$$

$u = 1/\sigma^2 \Rightarrow \sigma^2 = \frac{1}{u} \Rightarrow d\sigma^2 = -1/u^2 \, du$

$$= \sqrt{\frac{2\pi\sigma^2}{n+1}} \int_0^\infty u^{n/2+1} \exp\left( -\frac{1}{2} u \left( S^2(n-1) + \frac{n}{n+1} (\bar{y}-\tilde{y})^2 \right) \right) \frac{-1}{u^2} du$$

$$= -\sqrt{\frac{2\pi\sigma^2}{n+1}} \int_0^\infty u^{n/2-1} \exp\left(-\frac{1}{2}u\left[s^2(n-1)+\frac{n}{n+1}(\bar{y}-\tilde{y})^2\right]\right) du$$

$$a = n/2 \quad , \quad b = \frac{1}{2}\left(s^2(n-1)+\frac{n}{n+1}(\bar{y}-\tilde{y})^2\right)$$

$$\tilde{y}|y \propto -\sqrt{\frac{2\pi\sigma^2}{n+1}} \frac{\Gamma(n/2)}{\left(\frac{1}{2}\left(s^2(n-1)+\frac{n}{n+1}(\bar{y}-\tilde{y})^2\right)\right)^{n/2}}$$

From here, we can show that $\tilde{y}|y$ follows a t distribution

$$\boxed{\tilde{y}|y \sim t_{n-1}\left(\bar{y}, \left(\frac{n+1}{n}\right)s^2\right)}$$

Reminder t distribution:

$$t \sim \frac{\Gamma(n/2)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\pi\sigma}}\left(1-\frac{1}{n-1}\left(\frac{\bar{y}-\tilde{y}}{s\sqrt{\frac{n+1}{n}}}\right)^2\right)^{-n/2}$$

...

$$\tilde{y}|y \propto -\sqrt{\frac{2\pi}{n+1}}\,\Gamma(n/2)\left(\frac{1}{2}\left(s^2(n-1)+\frac{n}{n+1}(\bar{y}-\tilde{y})^2\right)\right)^{n/2}$$

...

example: Gravity

number of measurements : 70
mean : 6.5
S.d : 2.4

sampling from the posterior.

## Conjugate priors

Normal model with unknown mean and variance.

Multinomial model for categorical data

example: ice cream

categorical

$y = (450, 300, 250)$ data

noninformative prior : dirichelet $\alpha = 1,1,1$

posterior : dirichlet : $\alpha = (451, 301, 251)$

answer : simulation : draws and compute

$\theta_{vanila} - \theta_{chocolate} \sim 100\%$

sampling distribution for multivariate normal

$$y_1, \dots, y_n \sim N(\mu, \Sigma)$$

$$p(y_1, \dots, y_n \mid \mu, \Sigma) \propto \det(\Sigma)^{-n/2} \exp\left(-\frac{1}{2}(y-\mu)\Sigma^{-1}(y-\mu)^T\right)$$

$$= \det(\Sigma)^{n/2} \exp\left(-\frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} S_0\right)\right)$$

where $S_0 = (y-\mu)(y-\mu)^T$

likelihood

<u>example</u>: pen

$$\mu_0 = \begin{pmatrix} 2 \\ 5 \end{pmatrix} \quad \Lambda_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$ no covariance, independence

$$\kappa_0 = 1, \quad \nu_0 = 0$$

$$y = \left\{ \begin{pmatrix} 1.8 \\ 4.8 \end{pmatrix}, \begin{pmatrix} 2.1 \\ 5.1 \end{pmatrix}, \begin{pmatrix} 2.0 \\ 5.3 \end{pmatrix}, \begin{pmatrix} 1.9 \\ 4.9 \end{pmatrix} \right\} \quad \bar{y} = \begin{pmatrix} 1.95 \\ 5.025 \end{pmatrix}$$

$$S = \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^T \approx \begin{pmatrix} 0.0167 & 0.2167 \\ 0.02167 & 0.049167 \end{pmatrix}$$

<u>posterior</u>:

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

$$= \frac{1}{1+4}\begin{pmatrix}2\\5\end{pmatrix} + \left(\frac{4}{1+4}\right)\begin{pmatrix}1.95\\5.025\end{pmatrix}$$

$$= \begin{pmatrix}1.96\\5.02\end{pmatrix}$$ posterior mean

$\kappa_n = 5$, $\nu_n = 4$

$$\Lambda_n = \begin{pmatrix}1&0\\0&1\end{pmatrix} + S + \frac{4}{5+4}\left(\begin{pmatrix}1.95\\5.025\end{pmatrix} - \begin{pmatrix}2\\5\end{pmatrix}\right)\cdot\left(\begin{pmatrix}1.95\\5.025\end{pmatrix} - \begin{pmatrix}2\\5\end{pmatrix}\right)^T$$

$$\approx \begin{pmatrix}1.02 & 0.217\\0.022 & 1.05\end{pmatrix}$$ posterior covariance matrix

## More models and applications

### Bayesian approach to Logistic Regression

likelihood:

$$p(y|\beta) = \exp\left(y^T X\beta - n\sum_{i=1}^{K}\log\left(1+e^{x_i^T\beta}\right)\right)$$

noninformative prior:

$$p(\beta) \propto 1$$

posterior:

$$p(\beta|y) \propto p(\beta)\, p(y|\beta)$$
$$\propto \exp\left( y^T X\beta - n\sum_{i=1}^{K} \log(1+e^{x_i^T\beta}) \right)$$

maximum likelihood estimation

We use MCMC to sample from the posterior

example: Bioassay analysis

comparing two proportions

$$y_1 \sim \text{Bin}(n, \theta_1)$$
$$y_2 \sim \text{Bin}(n, \theta_2)$$

$H_0 : \theta_1 = \theta_2$ (vs) $H_A : \theta_1 > \theta_2$

Define: $z_1 = \log \frac{\theta_1}{1-\theta_1}$ and $z_2 = \log \frac{\theta_2}{1-\theta_2}$

assume

$$z_2|z_1 \sim N(z_1, \tau^2)$$

and $p(z_1) \propto 1$ (noninformative)

let $v = \frac{1}{\sigma}(z_2 - z_1)$, $v \sim N(0,1)$

Improper joint prior:

$$p(z_1, z_2) \propto p(z_2 | z_1)\, p(z_1)$$
$$\propto \frac{1}{\sigma} \exp\left( -\frac{(z_2(\theta_2) - z_1(\theta))^2}{2\sigma^2} \right)$$

variable transformation trick

$$\det\left( \begin{pmatrix} \frac{dz_1}{d\theta_1} & 0 \\ 0 & \frac{dz_2}{d\theta_2} \end{pmatrix} \right) \cdots$$

$$p(\theta_1, \theta_2) \propto \exp\left(-\tfrac{1}{2}v^2\right) \theta_1^{-1}(1-\theta_1)^{-1} \theta_2^{-1}(1-\theta_2)^{-1}$$
$$\propto \exp\left(-\tfrac{1}{2}v^2\right) \theta_1^{\alpha-1}(1-\theta_1)^{\beta-1} \theta_1^{\nu-1}(1-\theta_2)^{\delta-1}$$

Howard prior w/ parameters $\alpha, \beta, \nu, \delta, \sigma$

The posterior distribution is:

$$p(\theta_1, \theta_2 | y) \propto \theta_1^{y_1}(1-\theta_1)^{n_1-y_1} \theta_2^{y_2}(1-\theta_2)^{n_2-y_2}$$
$$\times \exp\left(-\tfrac{1}{2}v^2\right) \theta_1^{\alpha-1}(1-\theta_1)^{\beta-1} \theta_1^{\nu-1}(1-\theta_2)^{\delta-1}$$

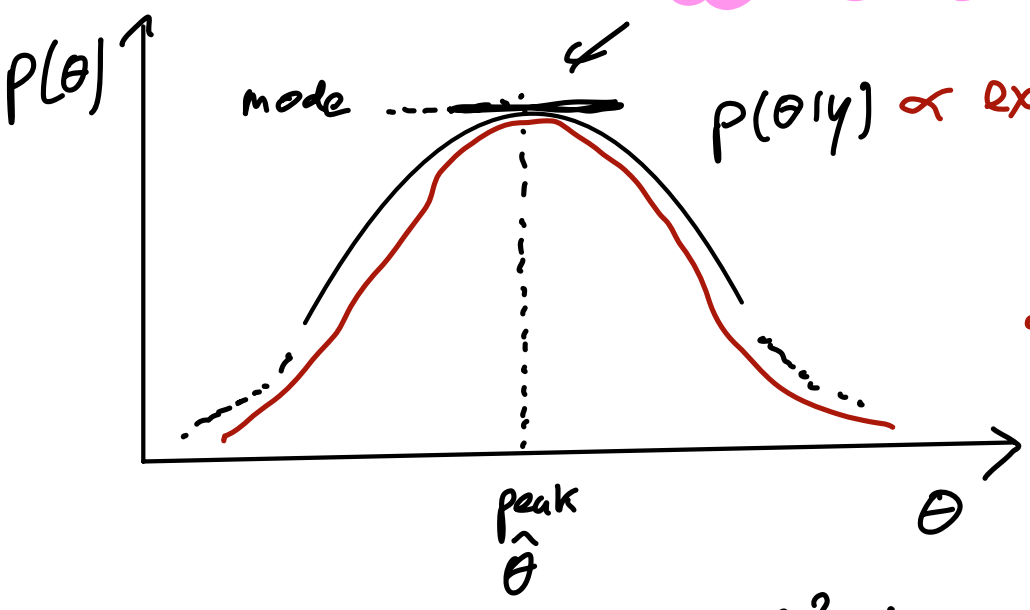
$$\propto \exp\left(-\frac{1}{2}v^2\right)\theta_1^{y_1+\alpha-1}\times(1-\theta_1)^{n_1-y_1+\beta-1}$$
$$\times\,\theta_2^{y_2+\nu-1}\,(1-\theta_2)^{n_2-y_2+\delta-1}$$

$$\text{Howard}\,(y_1+\alpha,\ n_1-y_1+\beta,\ y_2-\nu,\ n_2-y_2+\delta)$$

## Strategy for computation

## Normal approximation

derivative is 0 at peak $\hat{\theta}$




$$p(\theta|y) \propto \exp(a\theta^2+b\theta+c)$$

approximate by a normal distribution

$$\log(p(\theta|y)) \propto a\theta^2+b\theta+c$$

log posterior is roughly quadratic.

$$\frac{\partial}{\partial \theta} \log(p(\theta|y))_{\theta=\hat{\theta}} \simeq 0$$

$\theta$ is likely to be multidimentional

Taylor series

$$f(\theta) \approx f(\theta_0) + \nabla f(\theta_0)^T(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T H f(\theta_0)(\theta - \theta_0)$$

$$\log(p(\theta|y)) \approx \log(p(\hat{\theta}|y)) + 0 + \frac{1}{2}(\theta - \hat{\theta})^T \times \left(\frac{d^2}{(d\theta)^2} \log(p(\hat{\theta}|y))\right)_{\theta=\hat{\theta}} (\theta - \hat{\theta})$$

observed Fisher information

$$I(\theta) = \frac{d^2}{d\theta^2} \log(p(\theta|y))$$

$$\log(p(\theta|y)) \approx \log(p(\hat{\theta}|y)) + 0 + \frac{1}{2}(\theta - \hat{\theta})^T \times \left(\frac{d^2}{(d\theta)^2} \log(p(\hat{\theta}|y))\right)_{\theta=\hat{\theta}} (\theta - \hat{\theta})$$

$$\log(p(\theta|y)) \approx A + \frac{1}{2}(\theta - \hat{\theta})^T I(\theta)(\theta - \hat{\theta})$$

$$p(\theta|y) \approx Be^{\frac{1}{2}(\theta-\hat{\theta})^T I(\theta)(\theta-\hat{\theta})}$$

multivariate Normal distribution

$$p(\theta|y) \approx N(\hat{\theta}, I(\theta)^{-1})$$

<u>example:</u> normal distribution with unknown mean and variance.

<u>prior:</u> noninformative prior

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

show $p(\mu, \log(\sigma)) \propto 1$

$$(\mu, \sigma^2) \mapsto (\mu, \log(\sigma))$$

$$J = \begin{pmatrix} \frac{\partial \mu}{\partial \mu} & \frac{\partial \mu}{\partial \log \sigma^2} \\ \frac{\partial \sigma^2}{\partial \mu} & \frac{\partial \sigma^2}{\partial \log \sigma} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma^2 \end{pmatrix}$$

$$\sigma^2 = e^{2\log\sigma}$$

$$\frac{de^{2\log\sigma}}{d\log\sigma} = 2e^{2\log\sigma} = 2\sigma^2$$

$|J| \approx 2\sigma^2$ so

$$p(\mu \mid \log\sigma) = p(\mu, \sigma^2) \cdot |J| \propto \frac{1}{\sigma^2} \cdot 2\sigma^2$$

$$= 2 \propto 1$$

①

$\theta = (\mu, \log\sigma)$ so our noninformative prior is

$$p(\mu, \log(\sigma)) = 1$$

② find $\hat{\theta}$

So we need to find the peak of $p(\mu, \log\sigma \mid y)$

Recall from module 3:

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

$\log\sigma$ due to $|J|$

$\Rightarrow$

$$\log(p(\mu, \log\sigma \mid y)) \propto -n\log\sigma - \frac{1}{2\sigma^2}\left((n-1)s^2 + n(\bar{y}-\mu)^2\right)$$

This is the function around which we will construct our Taylor series

$\hat{\theta}$ is the point where $\nabla \log(p(\mu, \log\sigma \mid y)) = \vec{0}$

$$\frac{\partial}{\partial \mu} \log(p(\mu, \log\sigma \mid y)) \propto \frac{n(\bar{y}-\mu)}{\sigma^2} = 0$$

$$\frac{\partial}{\partial \log\sigma} \log(p(\mu, \log\sigma \mid y)) \propto -n + \frac{(n-1)s^2 + n(\bar{y}-\mu)^2}{\sigma^2} = 0$$

$$\frac{n(\bar{y}-\mu)}{\sigma^2} = 0 \Rightarrow n(\bar{y}-\mu) = 0$$

$$\Rightarrow \boxed{\bar{y} = \mu} = \hat{\theta}$$

$$-n + \frac{(n-1)s^2 + n(\bar{y}-\mu)^2}{\sigma^2} = 0$$

$$\Rightarrow \frac{-n\sigma^2 + (n-1)s^2}{\sigma^2} = 0$$

$$\Rightarrow (n-1)s^2 = n\sigma^2$$

$$\Rightarrow \frac{(n-1)s^2}{n} = \sigma^2 \neq \hat{\theta}_2$$

$$2\log(\sigma) = \log\left(\frac{(n-1)s^2}{n}\right) \Rightarrow \log\sigma = \frac{1}{2}\log\left(\frac{(n-1)s^2}{n}\right)$$

So

$$\hat{\theta} = \begin{pmatrix} \bar{y} \\ \frac{1}{2}\log\left(\frac{(n-1)s^2}{n}\right) \end{pmatrix}$$

## ③ Find $I(\hat{\theta})$

$$\frac{\partial^2}{(\partial\mu)^2}\log\left(p(\mu, \log\sigma \mid y)\right) = \frac{-n}{\sigma^2}$$

$$\frac{\partial^2}{\partial\mu\,\partial\log\sigma}\log\left(p(\mu, \log\sigma \mid y)\right) = -2n(\bar{y}-\mu)e^{-2\log\sigma}$$

at $\hat{\theta}$ : 0

$$\frac{\partial^2}{(\partial\log\sigma)^2}\log\left(p(\mu, \log\sigma \mid y)\right) = -2\left[(n-1)s^2 + n(\bar{y}-\mu)\right]e^{-2\log\sigma}$$

at $\hat{\theta}$ : $-2n$

$$\hat{\theta} = -n \cdot \frac{n}{(n-1)s^2} = \frac{-n^2}{(n-1)s^2}$$

$$I(\hat{\theta}) = \begin{pmatrix} -\frac{n^2}{(n-1)s^2} & 0 \\ 0 & -2n \end{pmatrix}$$

④ Compute the normal distribution approximation

$$p(\theta|y) \approx N(\hat{\theta}, I(\hat{\theta})^{-1})$$

$$p(\mu, \log\sigma|y) \approx N\left(\begin{pmatrix} \bar{y} \\ \frac{1}{2}\log\left(\frac{(n-1)s^2}{2}\right) \end{pmatrix}, \begin{pmatrix} \frac{n^2}{(n-1)s^2} & 0 \\ 0 & 2n \end{pmatrix}^{-1}\right)$$

## Large-sample theory


$y_1, \dots, y_n$ $\quad y_i \in \{H, T\}$ $\quad H = 1$, $T = 0$

$y$ = # heads
$n$ = # of flips

$\Theta = [0,1]$ parameter space

$$\mathcal{F} := \left\{ \binom{n}{y} \theta^y (1-\theta)^{n-y} \mid \theta \in \Theta \right\}$$

case 1: $f(y) \in \mathcal{F}$ $\exists \theta_0$ s.t. $f(y) = \binom{n}{y} \theta_0^y (1-\theta_0)^{n-y}$

case 2: misspecification of the model

$f(y) \notin \mathcal{F}$ but there is some function $p(\theta_0 | y)$ that's the closest to $f(y)$

# Kullback - Leibler (KL) divergence

## sketch proof 1:

Assumption:

1. $\textcircled{H}$ is finite $\quad \textcircled{H} = \{\theta_0, \theta_1, \ldots, \theta_n\}$
2. $\theta_0$ is the $KL(\theta)$ minimizer
3. $p(\theta = \theta_0) > 0$ it's within our prior

Aim:

show $p(\theta = \theta_0 | y) \longmapsto 1$ as $n \mapsto \infty$

Let $\theta \neq \theta_0$

We pick a value $\theta$ that is not $\theta_0$, so

$$\log\left(\frac{p(\theta|y)}{p(\theta_0|y)}\right) = \log\left(\frac{p(\theta)\prod_{i=1}^{n} p(y_i|\theta)}{p(\theta_0)\prod_{i=1}^{n} p(y_i|\theta_0)}\right)$$

$$= \log\left(\frac{p(\theta)}{p(\theta_0)}\right) + \sum_{i=1}^{n} \log\left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right)$$

observe:

$$\sum_{i=1}^{n} \log\left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right) = \log\left(p(y_i|\theta)\right) - \log\left(p(y_i|\theta_0)\right)$$

$$= \log\left(p(y_i|\theta)\right) - \log\left(f(y_i)\right) + \log\left(f(y_i)\right) - \log\left(p(y_i|\theta_0)\right)$$

$$= \log\left(\frac{f(y_i)}{p(y_i|\theta_0)}\right) - \log\left(\frac{f(y_i)}{p(y_i|\theta)}\right)$$

Recall $y_i$ are iid. So by the law of large numbers

$$\log\left(\frac{p(\theta|y)}{p(\theta_0|y)}\right) = \log\left(\frac{p(\theta)}{p(\theta_0)}\right) + \sum_{i=1}^{n} \log\left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right)$$

$$= \log\left(\frac{p(\theta)}{p(\theta_0)}\right) + \sum_{i=1}^{n}\left[\log\left(\frac{f(y_i)}{p(y_i|\theta_0)}\right) - \log\left(\frac{f(y_i)}{p(y_i|\theta)}\right)\right]$$

$$= \log\left(\frac{p(\theta)}{p(\theta_0)}\right) + n\,\mathbb{E}\left[\log\left(\frac{f(y_i)}{f(y_i|\theta_0)}\right)\right] - n\,\mathbb{E}\left[\log\left(\frac{f(y_i)}{f(y_i|\theta)}\right)\right]$$

$$= \log\left(\frac{p(\theta)}{p(\theta_0)}\right) + n\left[\underbrace{KL(\theta_0)}_{\text{minimizer}} - KL(\theta)\right]$$

Since $\theta \neq \theta_0$ and $\theta_0$ is a minimizer of $KL(\cdot)$

$$\log\left(\frac{p(\theta|y)}{p(\theta_0|y)}\right) \longrightarrow -\infty \quad \text{as} \quad n \to \infty$$

$$\Rightarrow \frac{p(\theta|y)}{p(\theta_0|y)} \to 0 \quad \Rightarrow \quad p(\theta|y) \to 0$$

Because probabilities sum to 1, then

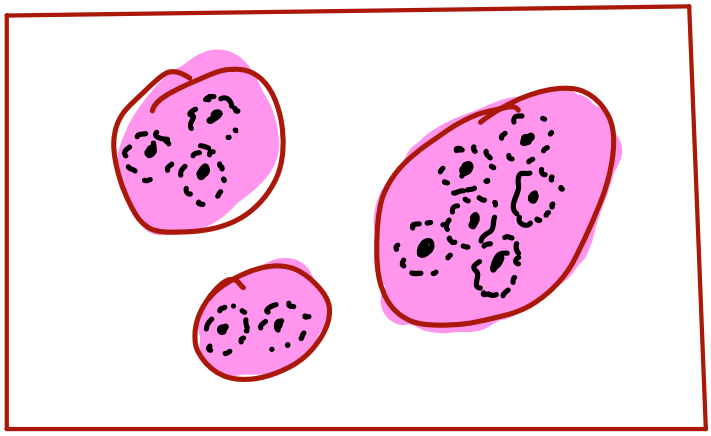
$$p(\theta_0|y) \to 1$$

Hence, the posterior distribution converges

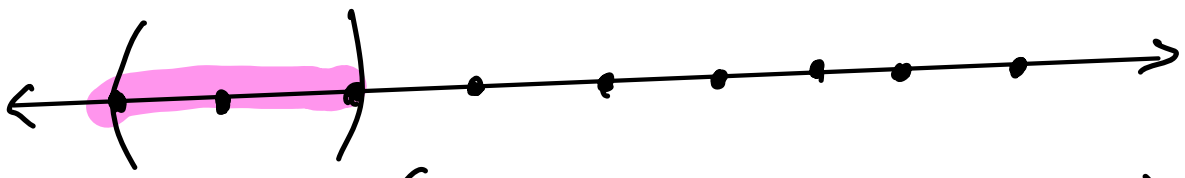
# Compact sets

A set $(H)$ is compact if every open cover contains a finite subcover

Since $(H) \subseteq \mathbb{R}^n$, we picture this as closed and bounded.




Counter example: $\mathbb{Z}$ in $\mathbb{R}$



$$A = \{ z-1, z+1 \mid z \in \mathbb{Z} \}$$

each integer is in exactly one set.
I cannot delete any set in this open cover because only one covers each point.

## Sketch the proof for Theorem 2

Let $(H)$ be a compact set and define $\mathcal{A}$ to be an open cover such that only one set $A_0 \in \mathcal{A}$ contains $\theta_0$.

Since $(H)$ is compact, there exists a finite subcover $\{A_0, A_1, \ldots, A_K\}$ where $A_0$ is the set previously specified.

Use Theorem 1 argument.

Let $A \neq A_0$

$$\log\left(\frac{p(\theta \in A \mid y)}{p(\theta \in A_0 \mid y)}\right) \approx \log\left(\frac{p(\theta \in A)}{p(\theta \in A_0)}\right) +$$

$$n \in \left[\frac{p(y_i \mid \theta \in A)}{p(y_i \mid \theta \in A_0)}\right]$$

$p(\theta \in A \mid y) \to 0$ as $n \to \infty$

$p(\theta \in A_0 \mid y) \to 1$

# Theorem 3

## Asymptotic Normality

## Frequency properties

$$\theta \mid y \sim N\left(\hat{\theta}, [I(\hat{\theta}]^{-1}\right)$$

$$(\theta - \hat{\theta}) \mid y \sim N\left(0, [I(\hat{\theta})]^{-1}\right)$$

Now consider $z = [I(\hat{\theta})]^{1/2} (\theta - \hat{\theta})$

if $x \sim N(\mu, \Sigma)$ and $A$ is a linear transformation, then

$$Ax \sim N(A\mu, A\Sigma A^T)$$

mean: $[I(\hat{\theta})]^{1/2} \cdot 0 = 0$

variance: $[I(\theta)]^{1/2} (I(\hat{\theta}))^{-1} \left([I(\hat{\theta})]^{1/2}\right)^T = I$

hence

$$z|y \sim N(0, I)$$

## hypothesis testing

$\theta$

$H_0 : \theta = 0$

$H_A : \theta > 0$

$y \sim N(\theta, 1)$ uniform prior

data : $y = 1$

| Frequentist approach | Bayesian approach |
|---|---|
| $z = \frac{1-0}{1} = 1$ | $y = 1$ likelihood $p(y\|\theta)$ |
| one-sided p-value of 0.16 | prior $p(\theta)$ |
| two-sided " 0.32 | ⋮ |
| Both "fail to reject" | posterior probability that $\theta > 0$ is 84 %. |
| p-value > 0.05 threshold | updating a belief with uncertainty quantified. |

Bayesian Bootstrapping

Priors and exchangeability

Hierarchical models : normal-normal model

Model Validation

Numerical methods and approximations

Monte carlo simulations (stochastic)

(vs)

Trapezoidal method, ... (deterministic)

Direct approximation using a grid
(brute force)

example grid approximation

$$\theta | y \sim \text{Beta}(y+1, n-y+1) \qquad n = 100,\ y = 48$$

$$\{\theta_1, \theta_2, \dots, \theta_{10}\} \qquad \theta_1 = 0 \text{ and } \theta_{10} = 1$$

$$\| \quad 1/9 \dots$$

Estimate $$\frac{p(\theta_i | y)}{\sum_{j=1}^{10} p(\theta_j | y)}$$

step 1: $u = \text{runif}(1)$ $\quad u_1 = 0.356579$

step 2: use inverse cdf to identify the closest $\theta_i$

cdf: $$F(\omega) = \sum_{v \le \omega} p(v)$$

$$\rightarrow \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}\}$$

$$0 \quad 0 \quad 0 \quad 0.0093 \quad 0.0006 \quad 0 \quad 0 \quad 0 \quad 0$$

$$\downarrow$$

$$0.7024$$

$F(\theta_4) = 0.0093$

$F(\theta_5) = 0.0093 + 0.7024 = 0.7117$

which is closer to $u_1$

$$|u_1 - F(\theta_4)| = 0.347279$$

$$|u_1 - F(\theta_5)| = 0.3551208$$

$\theta_4 \leftarrow$ randomly drew

and so on ...

## Simulating from predictive distributions

## Rejection sampling

Requires that the posterior $p(\theta|y)$ is known.

## Importance sampling

## Sampling - Importance - Ressampling (SIR)

# Markov Chains methods

$$\bar{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \quad \bar{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

$$\bar{\theta} | y \sim N(\bar{y}, \Sigma)$$

$$|\Sigma| = 1 - \rho^2$$

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

The joint probability distribution

$$p(\bar{\theta}|y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}(\bar{\theta}-\bar{y})^T \Sigma^{-1} (\bar{\theta}-\bar{y})\right)$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2(1-\rho^2)} (\theta_1-y_1 \;\; \theta_2-y_2) \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} \theta_1-y_1 \\ \theta_2-y_2 \end{pmatrix}\right)$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[(\theta_1-y_1)^2 - 2\rho(\theta_1-y_1)(\theta_2-y_2) + (\theta_2-y_2)^2\right]\right)$$

$$p(\theta_1|\theta_2) \propto \exp\left(-\frac{1}{2(1-\rho^2)} \left[(\theta_1-y_1)^2 - 2\rho(\theta_2-y_2)(\theta_1-y_1)\right]\right)$$

$$\propto \exp\left(-\frac{1}{2(1-\rho^2)} \left[(\theta_1-y_1) - \rho(\theta_2-y_2)\right]^2\right)$$

$$\left[\theta_1 - \left[y_1 + \rho(\theta_2-y_2)\right]\right]^2$$

$$\theta_1 | \theta_2 \sim N\left( y_1 + \rho(\theta_2 - y_2) \,,\, 1-\rho^2 \right)$$

$\theta_2 | \theta_1$ same calculation

Suppose $\bar{y} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ $\rho = 0.8$

$$\theta_1 | \theta_2, y \sim N\left(0.8\theta_2 \,,\, 0.36\right)$$

$$\theta_2 | \theta_1, y \sim N\left(0.8\theta_1 \,,\, 0.36\right)$$

$\theta_1$ is first and $\theta_2$ is second

$\theta_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$t=1$

$\theta_1^1 \sim N(0, 0.36) \xrightarrow{\text{draw}} \theta_1^1 = 0.12$

$\theta_2^1 \sim N((0.8)(0.12), 0.36) \xrightarrow{\text{draw}} \theta_2^1 = 0.40$

$t=2$

$\theta_1^2 \sim N((0.8)(0.4), 0.36) \xrightarrow{\text{draw}} \theta_1^2 = 0.001$

$\theta_2^2 \sim N((0.8)(0.01), 0.36) \xrightarrow{\text{draw}} \theta_2^2 = 0.148$

and so on ...

That is the process of Gibbs sampling

Metropolis Algorithm

Metropolis - Hastings Algorithm