Lecture 1.pdf

Lecture 2.pdf

Lecture 3.pdf

Lecture 4.pdf

Lecture 5.pdf

Lecture 6.pdf

Lecture 7.pdf

Lecture 8.pdf

Lecture 9.pdf

Lecture 10.pdf

Lecture 11.pdf

Lecture 12.pdf

Lecture 13.pdf

Lecture 14.pdf

Lecture 15.pdf

Lecture 16.pdf

Lecture 17.pdf

Lecture 18.pdf

Lecture 19.pdf

Lecture 20.pdf

Lecture 21.pdf

Lecture 22.pdf

Conditional probabilities

Given two events A and B in a probability space Ω we define:

$P(A \cap B)$ = Probability that A and B happen simultaneously. = joint probability of A and B

$P(A \cap B) = 0$ mutually exclusive events.

$P(A)$ = marginal prob. of A.

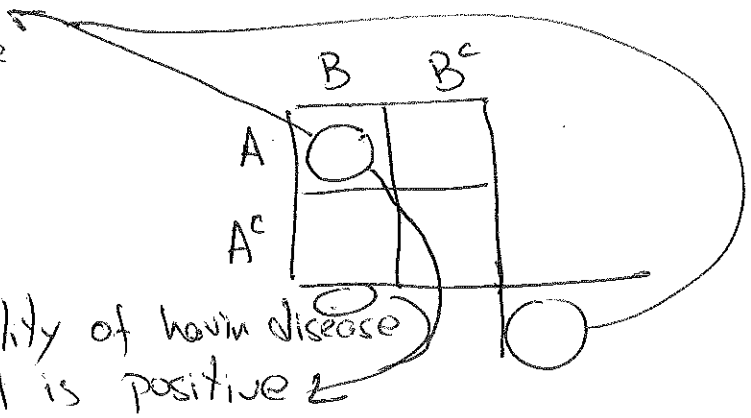
$P(B)$ = marginal prob of B.

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

A = an individual randomly sampled has cancer

B = an individual randomly sample is positive in a test for that cancer.

$P(A \cap B)$ = an individual has cancer and is positive in the disease



$P(A|B)$ = probability of having disease if test is positive

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

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

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ Bayes theorem.

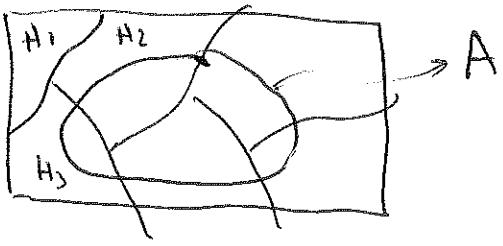
Partitions: $\{H_1, H_2, \dots, H_k\}$ set of events such that

$$H_i \cap H_j = \emptyset$$

$$H_1 \cup H_2 \cdots \cup H_k = \Omega$$

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{\underbrace{\sum_{j=1}^{k} P(A|H_j)P(H_j)}_{P(A)}}$$

Total probability law $P(A) = \sum_{j=1}^{k} P(A|H_j)P(H_j)$



Example: Data on incomes and educational level for a sample of 30 year old males

$\{H_1, H_2, H_3, H_4\}$ events that correspond to a randomly sampled individual being in the 1st, 2nd, 3rd or 4th quantile of the income distribution.

$$P(H_1) = \frac{1}{4}$$

$\{H_1, H_2, H_3, H_4\}$ represents a partition of the space

From the survey data we have

$Pr(E|H_1) = 0.11$
$Pr(E|H_2) = 0.19$
$Pr(E|H_3) = 0.31$
$Pr(E|H_4) = 0.53$

E: an individual has a college education.

(1) $$Pr(H_4|E) = \frac{Pr(E|H_4)P(H_4)}{P(E)}$$

(2) $$P(E) = Pr(E|H_1)P(H_1) + Pr(E|H_2)P(H_2) + Pr(E|H_3)Pr(H_3) + Pr(E|H_4)Pr(H_4)$$

(1) + (2) $\Rightarrow Pr(H_4|E) = 0.47$.

Example: You have two operators for a machine that produces widgets. The number of widgets produced during a day follows a Poisson distribution, with a mean that depends on the operator. The mean for operator 1 is 10.5 and the mean for operator 2 is 7.8.

We collect one day's worth of widgets and want to determine which operator was working that day.

Call X = observation.

$X \sim \text{Poisson}(\lambda)$

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

$$P(X = x \mid \text{operator } 1) = \frac{e^{-10.5}(10.5)^x}{x!}$$

$$P(X = x \mid \text{operator } 2) = \frac{e^{-7.8}(7.8)^x}{x!}$$

$$\Pr(\text{operator } 1 \mid X = x) = \frac{\Pr(X = x \mid Op\,1)\Pr(Op\,1)}{\Pr(X = x \mid Op\,1)\Pr(Op\,1) + \Pr(X = x \mid Op\,2)\Pr(Op\,2)}$$

We need $\Pr(Op\,1)$ ($\Pr(Op\,2) = 1 - \Pr(Op\,1)$)

Option 1: I don't know so set $\Pr(Op\,1) = \frac{1}{2} = \Pr(Op\,2)$

In that case

$$\Pr(Op\,1 \mid \bar{X} = x) = \frac{\Pr(\bar{X} = x \mid Op\,1)}{\Pr(\bar{X} = x \mid Op\,1) + \Pr(\bar{X} = x \mid Op\,2)}$$

$\Rightarrow$ I will say that the operator that was in charge is the one that maximizes

$$\Pr(\bar{X} = x \mid Op)$$

↳ Maximum likelihood estimation (in its simplest form)

Option 2: If you know that one works part time and the other full time then:

$\Pr(Op\,1) = 1/3$ (part time)

$\Pr(Op\,2) = 2/3$ (full time)

$$\Pr(Op\,1 \mid \bar{X} = x) = \frac{\Pr(\bar{X} = x \mid Op\,1)}{\Pr(\bar{X} = x \mid Op\,1) + 2\Pr(\bar{X} = x \mid Op\,2)}$$

Independance

A and B are independent if

$\Pr(A|B) = \Pr(A)$

$\hookrightarrow \Pr(A \cap B) = P(A)P(B)$

Iterating conditionings

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

Random variables:

Unknown numerical quantities about which we make probability statements.

Discrete random variables (Countable outcome spaces)

Continuous random variable (uncountable outcome spaces)

$\mathbb{N}$

$\mathbb{R}$

Discrete random variables

(pmf) probability mass function $f(x) = \Pr(X = x) = F(x) - F(x-1)$

(cdf) cumulative distribution function $\Pr(X \leq x) = F(x)$

Continuous random variables

(pdf) probability density function $f(x) = \frac{d}{dx} F(x)$

(cdf) cumulative distribution function. $F(x) = \Pr(X \leq x)$

Summaries of random variables.

$$E(X) = \begin{cases} \sum_{x \in \Omega} x\, p(x) \\ \int_{\Omega} x\, p(x)\, dx \end{cases}$$

$$\mathrm{Var}(X) = \begin{cases} \sum_{x \in \Omega} (x - E(X))^2 \cdot p(x) \\ \int_{\Omega} (x - E(X))^2\, p(x)\, dx \end{cases}$$

Some random do not have finite means or variances!!

Bernoulli Trials:

Experiments:

1) Two posible outcomes (S/F, 0/1)
2) Repeats are independent
3) Probability of outcomes is constant over time.

Bernoulli: $X_i \sim Ber(\theta)$ $\quad X_i \in \{0,1\}$

$p(x|\theta)$ - $p(x) = \theta^x (1-\theta)^{1-x}$ $\quad \theta$ = success probability

$$= \begin{cases} \theta & \text{if } x=1 \\ 1-\theta & \text{if } x=0 \end{cases}$$

Binomial: $X_1, X_2, \dots X_n \sim Ber(\theta)$ and independent

$Z = X_1 + X_2 + \dots + X_n$ = # of successes in n trials

$Z \sim Bin(n, \theta)$

$$p(z|n,\theta) = \binom{n}{z} \theta^z (1-\theta)^{n-z} \quad z = 0, 1, \dots, n$$

$E(Z) = n\theta$

$Var(Z) = n\theta(1-\theta)$

Geometric:

Y = # of trials before a success is observed in a Bernoulli experiment with success prob θ

$$P(y|\theta) = \theta(1-\theta)^{y-1} \qquad y = 1, 2, 3, \ldots$$

Alternatively, define W = # of failures before the first success

$$Y = W + 1$$

$$p(w|\theta) = \theta(1-\theta)^{w} \qquad w = 0, 1, \ldots$$

$$E(Y) = \frac{1}{\theta}$$

$$Var(Y) = \frac{1-\theta}{\theta^2}$$

Negative Binomial

$$Y_1, Y_2, \ldots Y_K \sim Geo(\theta)$$

$X = Y_1 + Y_2 + \cdots + Y_K$ = number of trials to observe K successes

$$p(x|\theta, K) = \binom{x-1}{K-1} \theta^K (1-\theta)^{x-K}$$

opposite to the Binomial!!!

$$E(X) = \frac{K}{\theta}$$

$$Var(X) = \frac{K(1-\theta)}{\theta^2}$$

Hyper-geometric (sampling uniformly without replacement)

X = # of blue balls in a sample of size n drawn from an urn that contains K blue balls and $N-K$ red balls.

$$p(x \mid n, K, N) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

$E(X) = \frac{Kn}{N} \approx$ # of balls in sample $\times$ prob of blue ball — check!!!

$$\text{Var}(X) = \frac{Kn}{N} \cdot \frac{(N-n)(N-K)}{N(N-1)}$$

Poisson:

X = count of events

$$p(x \mid \lambda) = \frac{\lambda^x \exp\{-\lambda\}}{x!} \qquad x = 0, 1, \ldots$$

Poisson process:

Consider the number of events (phone calls) that you receive during a given hour.

- Disjoint intervals are independent
- Calls arise independently.
- The probability of two or more calls in a very short time period is very small.
- The probability of getting exactly one call in a small interval is constant.

If those 3 assumptions are satisfied then:

X = number of calls in one hours

$X \sim \text{Poisson}(\lambda)$ λ = mean number of calls in one hour.

$E(X) = \lambda$

$\text{Var}(X) = \lambda$

In practice, the fact that $E(X) = \text{Var}(X)$ is quite restrictive. It is often the case that

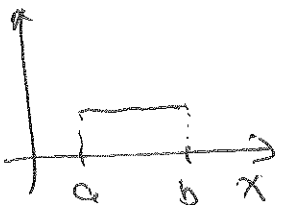
$\text{Var}(X) > E(X)$ overdispersion.

↳ If this is the case the Negative Binomial might be a better option for modeling <u>even if</u> the data generation mechanism has nothing to do with Bernoulli trials!!!

Continuous random variables:

Uniform: $X \sim \text{Uni}[a,b]$

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$



$$E(X) = \frac{a+b}{2}$$

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

"All values have the same likelihood of occurring"

Exponential: interarrival times in a Poisson process

$X \sim \text{Exp}(\mu)$

$$f(x|\mu) = \begin{cases} \frac{1}{\mu} \exp\left\{-\frac{x}{\mu}\right\} & x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

$E(X) = \mu$

$\text{Var}(X) = \mu^2$

$$F(x|\mu) = \int_0^x \frac{1}{\mu} \exp\left\{-\frac{t}{\mu}\right\} dt$$

$$= 1 - \exp\left\{-\frac{x}{\mu}\right\} = \Pr(X \le x)$$

$$\Pr(X > t+s \mid X > s) = \frac{\Pr(X > t+s, X > s)}{\Pr(X > s)} = \frac{\Pr(X > t+s)}{P(X > s)} = \exp\left\{-\frac{t}{\mu}\right\}$$

Gamma: $X_1, X_2, \ldots X_n \sim Exp(\mu)$

$$Y = X_1 + X_2 + \cdots X_n \sim Gamma(n, \mu)$$

$$f(y|n,\mu) = \begin{cases} \frac{1}{\Gamma(n)}\left(\frac{1}{\mu}\right)^n y^{n-1} \exp\left\{-\frac{y}{\mu}\right\} & y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$\Gamma(n)$ = gamma function.

$\Gamma(n) = (n-1)!$ if n is integer

$$\Gamma(n) = \int_0^\infty y^{n-1} \exp\{-y\}\, dy.$$

$E(Y) = n\mu$

$Var(Y) = n\mu^2$.

Beta distribution: $X_1 \sim \text{Gamma}(a_1, b)$

$X_2 \sim \text{Gamma}(a_2, b)$

$$Y = \frac{X_1}{X_1 + X_2} \sim \text{Beta}(a_1, a_2)$$

$$f(y) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1} & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

If $a = b = 1$ then you recover the Uni[0,1] distribution!

$$E(Y) = \int_{-\infty}^{\infty} y f(y)\, dy = \int_0^1 y \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1} dy$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 y^{a}(1-y)^{b-1} dy$$

$\Gamma(a+1) = a\Gamma(a)$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 y^{(a+1)-1}(1-y)^{b-1} dy$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\cancel{\Gamma(b)}} \frac{\Gamma(a+1)\cancel{\Gamma(b)}}{\Gamma(a+b+1)} = \frac{a}{a+b}$$

Normal distribution $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$F(x)$ is not available in closed form!

A common notation is:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} dt.$$

Note that $F(x)$ can be computed in terms of $\Phi(x)$ for any μ and σ.

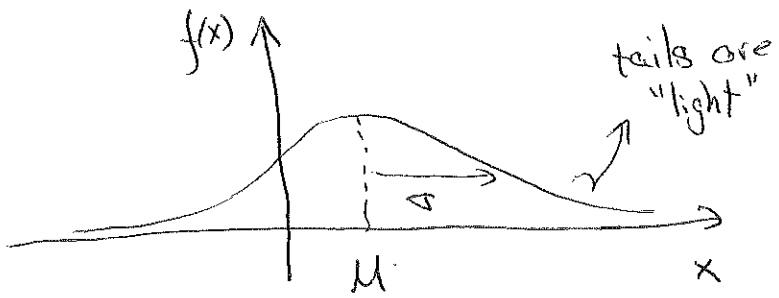
$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right\} dt$$

$$z = \frac{t-\mu}{\sigma} \qquad dz = \frac{dt}{\sigma}$$

$$\Rightarrow F(x) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\} dz = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$E(X) = \mu$

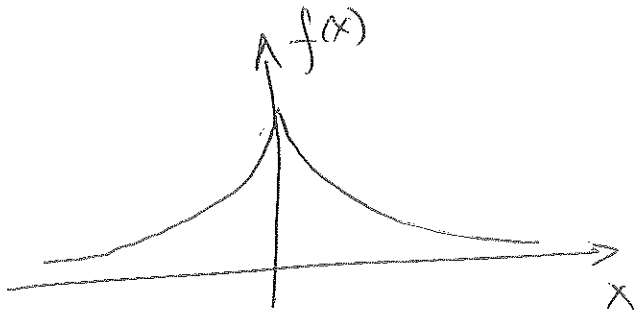
$Var(X) = \sigma^2$.



Double exponential distribution.

$$X \sim DExp(\mu, \lambda)$$

$$f(x) = \frac{1}{2\lambda} \exp\left\{-\frac{1}{\lambda}|x-\mu|\right\}$$



χ^2_n distribution: This is just a special case of the gamma.

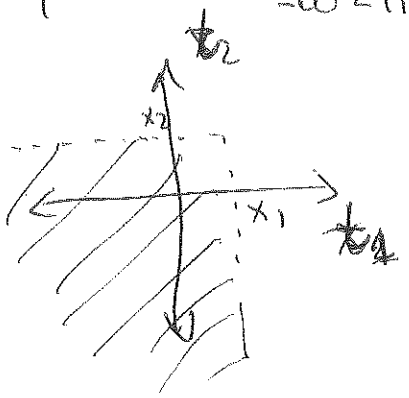
t distribution: We will discuss later when we deal with multiparameter models (inference for normal models).

Multivariate distributions:

Continuous case: Let X_1 and X_2 be two random variables with joint density $f(x_1, x_2)$ such that

$$\Pr(X_1 \le x_1, X_2 \le x_2) = \iint_{\Omega} f(t_1, t_2)\, dt_1 dt_2$$

where $\Omega = \{(t_1, t_2) : -\infty < t_1 \le x_1, -\infty < t_2 \le x_2\}$



You can define the marginal distributions and densities

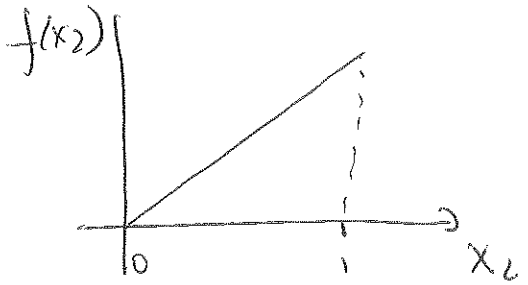
$$f(x_1) = \int f(x_1, x_2)\, dx_2 \quad \text{and}$$

$$f(x_2) = \int f(x_1, x_2)\, dx_1$$

The conditional densities

$$f(x_1 | x_2) = \frac{f(x_1, x_2)}{f(x_2)} = \frac{f(x_2 | x_1) p(x_1)}{\int f(x_2 | x_1) p(x_1) dx_1}$$

$$f(x_2) = \int_{-x_2}^{x_2} 1 \, dx_1 = 2x_2 \qquad x_2 \in [0,1]$$

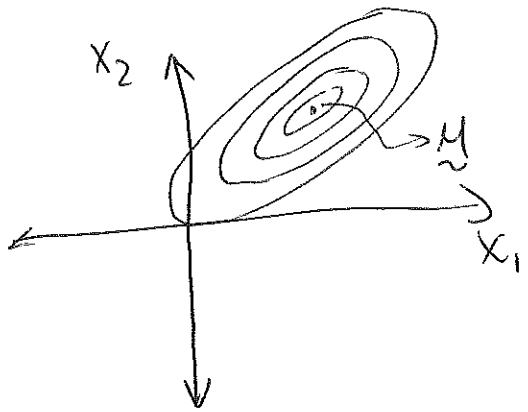


Multivariate normal distribution:

$\underset{\sim}{x} = (x_1, \dots x_n) \sim N_n(\underset{\sim}{\mu}, \Sigma)$ → positive definite matrix

$$f(\underset{\sim}{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x}-\underset{\sim}{\mu})\right\}$$

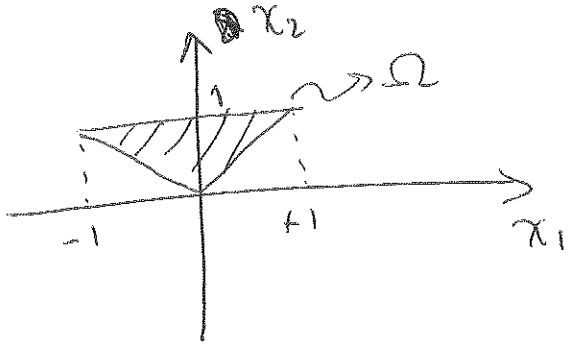
In the bivariate case.



Σ gives you the "shape" ⇒ orientation and axes length.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

ρ = correlation coefficient
σ_i^2 = marginal variances
μ_i = marginal means



$f(x_2)$

$f(x_2 \mid x_1 = a)$

$$f(x_1, x_2) = \begin{cases} 1 & \text{if } (x_1, x_2) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

$$f(x_2 \mid x_1) = \frac{f(x_1, x_2)}{\int f(x_1, x_2)\, dx_2}$$

$$f(x_1) = \int f(x_1, x_2)\, dx_2 = \int_{|x_1|}^{1} 1\, dx_2 = 1 - |x_1|$$

$$f(x_2 \mid x_1) = \frac{1}{1 - |x_1|} \quad \begin{matrix} -1 \le x_1 \le 1 \\ |x_1| \le x_2 \le 1 \end{matrix}$$

Difference between independence and lack of correlation.

If $(X_1, X_2) \sim N(\mu, \Sigma)$ and Σ is diagonal (meaning that the variables are uncorrelated) then they are also independent, and viceversa.

In general independence $\Rightarrow$ lack of correlation. but not the other way around.

$$\rho = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)\,Var(X_2)}} \quad \text{where} \quad Cov(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2)$$

$$E(X_1 X_2) = \iint x_1 x_2 f(x_1, x_2)\, dx_1 dx_2$$

$$\overset{\text{if independent}}{=} \iint x_1 x_2 f(x_1) f(x_2)\, dx_1 dx_2$$

$$= \int x_1 f(x_1)\, dx_1 \times \int x_2 f(x_2)\, dx_2$$

$$= E(X_1)\, E(X_2)$$

Conditional expectations and variances

$$E(X) = E[E(X|Y)]$$

$$\mathrm{Var}(X) = E(\mathrm{Var}(X|Y)) + \mathrm{Var}(E(X|Y))$$

$$E(X) = \iint x f(x,y)\,dx\,dy = \iint x f(x|y) f(y)\,dx\,dy$$

$$= \int \left[\int x f(x|y)\,dx\right] f(y)\,dy$$

$$g(y) = E(X|Y)$$

$$= \int g(y)\,dy = E($$

$$= \int g(y) f(y)\,dy = E(g(Y))$$

$$= E[E(X|Y)]$$

A machine makes a random number N of widgets in a day. Assuming that each is defective independently of the rest with probability θ, what is the expected number of defective widgets in a day?

If you know N, Z: # of defectives.

$$Z \mid N \sim \text{Bin}(N, \theta)$$

$$E(Z) = E[E(Z \mid N)] = E[N\theta] = \theta E(N).$$

Bernoulli Trials:

Experiments:

1) Two posible outcomes (S/F, 0/1)
2) Repeats are independent
3) Probability of outcomes is constant over time.

Bernoulli: $X_i \sim Ber(\theta)$ $\quad X_i \in \{0,1\}$

$p(x|\theta)$ - $p(x) = \theta^x (1-\theta)^{1-x}$ $\quad \theta$ = success probability

$$= \begin{cases} \theta & \text{if } x=1 \\ 1-\theta & \text{if } x=0 \end{cases}$$

Binomial: $X_1, X_2, \dots X_n \sim Ber(\theta)$ and independent

$Z = X_1 + X_2 + \dots + X_n$ = # of successes in n trials

$Z \sim Bin(n, \theta)$

$$p(z|n,\theta) = \binom{n}{z} \theta^z (1-\theta)^{n-z} \quad z = 0, 1, \dots, n$$

$E(Z) = n\theta$

$Var(Z) = n\theta(1-\theta)$

Geometric:

Y = # of trials before a success is observed in a Bernoulli experiment with success prob θ

$$P(y|\theta) = \theta(1-\theta)^{y-1} \qquad y = 1, 2, 3, \ldots$$

Alternatively, define W = # of failures before the first success

$$Y = W + 1$$

$$p(w|\theta) = \theta(1-\theta)^{w} \qquad w = 0, 1, \ldots$$

$$E(Y) = \frac{1}{\theta}$$

$$Var(Y) = \frac{1-\theta}{\theta^2}$$

Negative Binomial

$$Y_1, Y_2, \ldots Y_K \sim Geo(\theta)$$

$X = Y_1 + Y_2 + \cdots + Y_K$ = number of trials to observe K successes

$$p(x|\theta, K) = \binom{x-1}{K-1} \theta^K (1-\theta)^{x-K}$$

opposite to the Binomial!!!

$$E(X) = \frac{K}{\theta}$$

$$Var(X) = \frac{K(1-\theta)}{\theta^2}$$

Hyper-geometric (sampling uniformly without replacement)

X = # of blue balls in a sample of size n drawn from an urn that contains K blue balls and $N-K$ red balls.

$$p(x \mid n, K, N) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

$E(X) = \frac{Kn}{N} \approx$ # of balls in sample $\times$ prob of blue ball — check!!!

$$\text{Var}(X) = \frac{Kn}{N} \cdot \frac{(N-n)(N-K)}{N(N-1)}$$

Poisson:

X = count of events

$$p(x \mid \lambda) = \frac{\lambda^x \exp\{-\lambda\}}{x!} \qquad x = 0, 1, \ldots$$

Poisson process:

Consider the number of events (phone calls) that you receive during a given hour.

- Disjoint intervals are independent
- Calls arise independently.
- The probability of two or more calls in a very short time period is very small.
- The probability of getting exactly one call in a small interval is constant.

If those 3 assumptions are satisfied then:

X = number of calls in one hours

$X \sim \text{Poisson}(\lambda)$ λ = mean number of calls in one hour.

$E(X) = \lambda$

$\text{Var}(X) = \lambda$

In practice, the fact that $E(X) = \text{Var}(X)$ is quite restrictive. It is often the case that

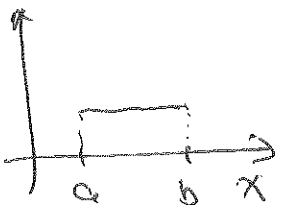
$\text{Var}(X) > E(X)$ overdispersion.

↳ If this is the case the Negative Binomial might be a better option for modeling <u>even if</u> the data generation mechanism has nothing to do with Bernoulli trials!!!

Continuous random variables:

Uniform: $X \sim \text{Uni}[a,b]$

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$



$$E(X) = \frac{a+b}{2}$$

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

"All values have the same likelihood of occurring"

Exponential: interarrival times in a Poisson process

$X \sim \text{Exp}(\mu)$

$$f(x|\mu) = \begin{cases} \frac{1}{\mu} \exp\left\{-\frac{x}{\mu}\right\} & x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

$E(X) = \mu$

$\text{Var}(X) = \mu^2$

$$F(x|\mu) = \int_0^x \frac{1}{\mu} \exp\left\{-\frac{t}{\mu}\right\} dt$$

$$= 1 - \exp\left\{-\frac{x}{\mu}\right\} = \Pr(X \le x)$$

$$\Pr(X > t+s \mid X > s) = \frac{\Pr(X > t+s, X > s)}{\Pr(X > s)} = \frac{\Pr(X > t+s)}{P(X > s)} = \exp\left\{-\frac{t}{\mu}\right\}$$

Gamma: $X_1, X_2, \ldots X_n \sim Exp(\mu)$

$$Y = X_1 + X_2 + \cdots X_n \sim Gamma(n, \mu)$$

$$f(y|n,\mu) = \begin{cases} \frac{1}{\Gamma(n)}\left(\frac{1}{\mu}\right)^n y^{n-1} \exp\left\{-\frac{y}{\mu}\right\} & y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$\Gamma(n)$ = gamma function.

$\Gamma(n) = (n-1)!$ if n is integer

$$\Gamma(n) = \int_0^\infty y^{n-1} \exp\{-y\}\, dy.$$

$E(Y) = n\mu$

$Var(Y) = n\mu^2$.

Happiness data:

Each female of age 65 or over in the 1998 General Social Survey was asked whether or not they were generally happy. Let $Y_i = 1$ if the i-th woman said "happy" and $Y_i = 0$ if the i-th woman said "unhappy". The survey includes $n = 129$ women over 65 who provided answers.

We want to estimate the "prevalence" of happiness in women over 65 in the US.

Likelihood:

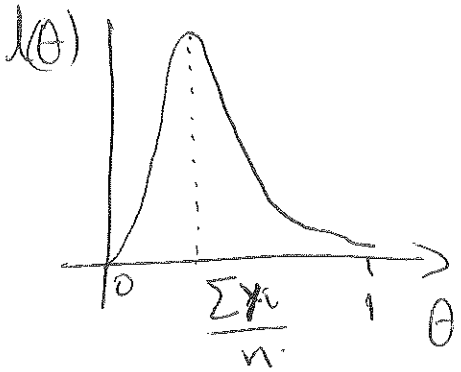
Assuming happiness is constant (and does not on covariates) and independence across observations.

$$f(y_1, \cdots y_n | n, \theta) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} = \theta^{\sum y_i} (1-\theta)^{n - \sum_{i=1}^{n} y_i}$$

$$z = \sum_{i=1}^{n} Y_i$$

$$f(z | n, \theta) = \binom{n}{z} \theta^{z} (1-\theta)^{n-z} = \binom{n}{\sum y_i} \theta^{\sum y_i} (1-\theta)^{n - \sum y_i}$$

$$l(\theta; x) = \theta^{\sum y_i} (1-\theta)^{n - \sum y_i} \quad 0 \leq \theta \leq 1$$



Classical statistics would use the MLE as a "good" estimate of the unknown θ

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ L(\theta)$$

$$\Rightarrow \hat{\theta} = \frac{\Sigma x_i}{n} \quad \text{in this case}$$

$l(\theta) = p(x|\theta)$ I want to combine that with a prior distribution $p(\theta)$ and use Bayes theorem.

What does it mean to elicit a prior? What is a prior?

We treat θ as random, but think of it as your knowledge about θ rather than θ being intrinsically random.

If you have no idea, you might want to say that all possible values have the same probability a priori

$$p(\theta) = \begin{cases} 1 & \text{if } 0 \le \theta \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{\int p(y|\theta)\,p(\theta)\,d\theta} = \begin{cases} \dfrac{\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}}{\int \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}\,d\theta} & 0 \le \theta \le 1 \quad (3) \\ \text{kernel of a } \text{Beta}(\sum y_i + 1,\ n - \sum y_i + 1) & \\ 0 & \text{otherwise} \end{cases}$$

$$= \frac{\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}}{\dfrac{\Gamma(\sum y_i + 1)\,\Gamma(n - \sum y_i + 1)}{\Gamma(n+2)}} = \frac{\Gamma(n+2)}{\Gamma(\sum y_i + 1)\,\Gamma(n - \sum y_i + 1)}\,\theta^{\sum y_i + 1 - 1}(1-\theta)^{n - \sum y_i + 1 - 1}$$

Assume that we have observed in the past values between 0.1 and 0.4 for θ (but never anything smaller or bigger) How do we incorporate this into the analysis.

For convenience, start with $\theta \sim \text{Beta}(a, b)$ and let's tweak a, and b to reflect our prior knoledge.

$$E(\theta) = \frac{a}{a+b} = 0.25$$

Consider a few cases

a	b
0.1	0.3
1	3
10	30

My prior is going to be $a = 10$ $b = 30$

$$p(\theta|y) = \frac{\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}}{\int \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\,d\theta}$$

$$= \frac{\theta^{\sum y_i + a - 1}(1-\theta)^{n-\sum y_i + b - 1}}{\int \theta^{\sum y_i + a - 1}(1-\theta)^{n-\sum y_i + b - 1}\,d\theta}$$

$$= \frac{\Gamma(n+a+b)}{\Gamma(\sum y_i + a)\Gamma(n - \sum y_i + b)}\theta^{\sum y_i + a - 1}(1-\theta)^{n-\sum y_i + b - 1}$$

If you recognize the numerator as the kernel of a distribution you know, then you do not need to solve the integral explicitly because it has to be the reciprocal of the normalizing constant of that distribution

$$p(\theta|y) \propto p(y|\theta)p(\theta) \propto \theta^{\sum y_i + a - 1}(1-\theta)^{n-\sum y_i + b - 1}$$

$$\Rightarrow \theta|y \sim \text{Beta}(a_n, b_n)$$

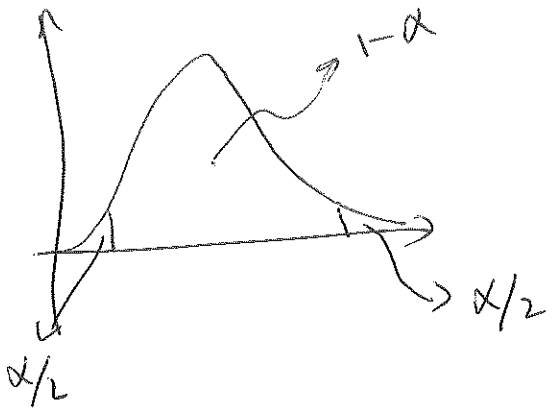
$$a_n = \sum y_i + a \qquad b_n = n - \sum y_i + b$$

To provide a point estimate we would use the posterior mean: $a = b = 1$

$$E(\theta|Y) = \frac{a_n}{a_n + b_n} = \frac{\sum y_i + a}{n + a + b} = \text{point estimator}$$

$$= 0.1984$$

Interval estimator: Credible interval: an interval that receives high probability under the posterior.



An easy interval is found by computing the $\alpha/2$ and $1-\alpha/2$ quantile of the posterior for small α.

(0.1350, 0.2706)

Poisson model

$X_i | \theta \sim$ Poisson (θ)

$\theta \sim$ Gamma (a, b)

$$f(x_1, \dots x_n | \theta) = \prod_{i=1}^{n} \frac{e^{-\theta} \theta^{x_i}}{x_i!} = \left(\prod_{i=1}^{n} \frac{1}{x_i!} \right) e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i}$$

$$f(\theta) = \begin{cases} \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp\{-b\theta\} & \theta \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

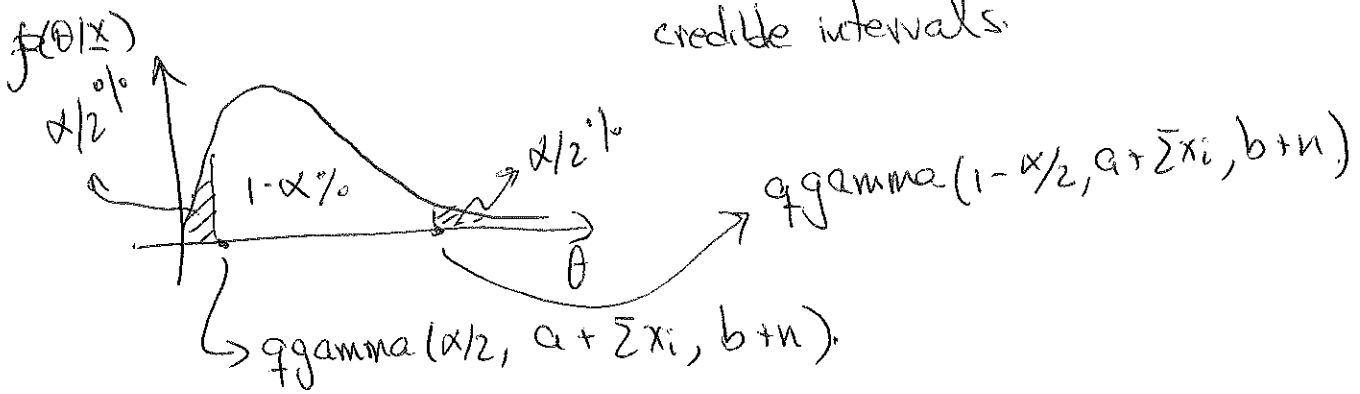
$$f(\theta | x_1, \dots x_n) \propto \theta^{\sum x_i} e^{-n\theta} \theta^{a-1} \exp\{-b\theta\}$$

$$= \theta^{\sum x_i + a - 1} \exp\{-\theta(b+n)\}$$

$$\Rightarrow \theta | x_1, \dots x_n \sim \text{Gamma}(a + \Sigma x_i, b + n).$$

$$E(\theta | x_1, \dots x_n) = \frac{a + \Sigma x_i}{b + n}$$

credible intervals.



Hypothesis testing:

$$H_0: \theta = \theta_0 \quad \text{vs} \quad H_a: \theta \neq \theta_0$$

$$Pr(H_0) = 1 - Pr(H_a) = 1/2$$

$$p(x_i|\theta, H_0) = \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) \theta_0^{\Sigma x_i} \exp\{-n\theta_0\}$$

$$p(x_i|\theta, H_a) = \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) \theta^{\Sigma x_i} \exp\{-n\theta\}$$

$$p(\theta|H_0) = \delta_{\theta_0}(\theta)$$

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

$$\frac{a}{b} = \theta_0 \qquad \nu = a \Rightarrow b = \frac{\nu}{\theta_0}$$

$$= \left(\frac{\nu}{\theta_0}\right)^{\nu} \frac{1}{\Gamma(\nu)} \theta^{\nu-1} \exp\left\{-\frac{\nu}{\theta_0}\theta\right\}$$

$$B_{0a} = \frac{\left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) \theta_0^{\Sigma x_i} \exp\{-n\theta_0\}}{\int \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) \theta^{\Sigma x_i} \exp\{-n\theta\} \left(\frac{\nu}{\theta_0}\right)^{\nu} \frac{1}{\Gamma(\nu)} \theta^{\nu-1} \exp\left\{-\frac{\nu}{\theta_0}\theta\right\} d\theta}$$

$$B_{0a} = \frac{\theta_0^{\Sigma x_i} \exp\{-n\theta_0\}}{\int \left(\frac{\nu}{\theta_0}\right)^{\nu} \frac{1}{\Gamma(\nu)} \theta^{\Sigma x_i + \nu - 1} \exp\{-(n + \frac{\nu}{\theta_0})\theta\} d\theta}$$

$$= \frac{\theta_0^{\Sigma x_i} \exp\{-n\theta_0\}}{\left(\frac{\nu}{\theta_0}\right)^{\nu} \frac{1}{\Gamma(\nu)} \Gamma(\Sigma x_i + \nu) \left(\frac{\Sigma x_i + \nu}{n + \frac{\nu}{\theta_0}}\right)^{(\Sigma x_i + \nu)}}$$

$$= \frac{\Gamma(\nu)}{\Gamma(\nu + \Sigma x_i)} \left(\nu + \Sigma x_i\right) \frac{\theta_0^{\Sigma x_i} \exp\{-n\theta_0\}}{\left(\frac{\nu}{\theta_0}\right)^{\nu} \left(\frac{n + \nu/\theta_0}{\nu + \Sigma x_i}\right)^{\nu + \Sigma x_i}}$$

$$\Pr(H_0 | \underline{x}) = \frac{1}{1 + B_{a0} \times \frac{\Pr(H_a)}{\Pr(H_0)}} = \frac{1}{1 + \frac{1}{B_{0a}} \frac{\Pr(H_a)}{\Pr(H_0)}}$$

How should Bayes Factor be interpreted?

Kass & Raftery 1995

Gaussian data:

$x_1, \dots x_n \sim \theta, \phi \sim N\left(\theta, \frac{1}{\phi}\right)$ so that $\sigma^2 = 1/\phi$

σ^2 = variance

ϕ = precision.

Assume ϕ is known.

$$f(x_1, \dots x_n | \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \phi^{1/2} \exp\left\{ -\frac{\phi}{2} (x_i - \theta)^2 \right\}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \phi^{n/2} \exp\left\{ -\frac{\phi}{2} \sum_{i=1}^{n} (x_i - \theta)^2 \right\}$$

$$\sum_{i=1}^{n} (x_i - \theta)^2 = \sum_{i=1}^{n} \left(x_i^2 - 2x_i\theta + \theta^2\right) = n\theta^2 - 2\theta \bar{x} n + \sum_{i=1}^{n} x_i^2$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$f(x_1, \dots x_n | \theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \phi^{n/2} \exp\left\{ -\frac{\phi}{2} \left(n\theta^2 - 2\theta n \bar{x} + \sum x_i^2\right) \right\}$$

$$\theta \sim N(\mu, \tau^2) \Rightarrow p(\theta) = \frac{1}{\sqrt{2\pi}} \frac{1}{\tau} \exp\left\{ -\frac{1}{2\tau^2} (\theta - \mu)^2 \right\}$$

$$p(\theta|x_1,\dots x_n) \propto \exp\left\{-\frac{\phi}{2}(n\theta^2 - 2\theta n\bar{x}) - \frac{1}{2\tau^2}(\theta^2 - 2\mu\theta)\right\} \quad (5)$$

$$= \exp\left\{-\frac{1}{2}\left[\left(\phi n + \frac{1}{\tau^2}\right)\theta^2 - 2\theta\left(n\phi\bar{x} + \frac{\mu}{\tau^2}\right)\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\phi n + \frac{1}{\tau^2}\right)\left[\theta - \left(\phi n + \frac{1}{\tau^2}\right)^{-1}\left(n\phi\bar{x} + \frac{\mu}{\tau^2}\right)\right]^2\right\}$$

$$\theta|x_1,\dots x_n \sim N\left(\frac{n\phi\bar{x} + \frac{\mu}{\tau^2}}{\left(\phi n + \frac{1}{\tau^2}\right)}, \left(\phi n + \frac{1}{\tau^2}\right)^{-1}\right)$$

Hypothesis testing: two proportions

Data: x = number of suc in pop 1
n = number of trials in pop 1

y = number of suc in pop 2
m = number of trials in pop 2

Likelihood $x \sim Bin(n, \theta_1)$ $y \sim Bin(m, \theta_2)$ independent

Goal $H_0: \theta_1 = \theta_2$ vs $H_a: \theta_1 \neq \theta_2$

$$Pr(H_0 \mid \overbrace{x, y}^{\text{data}}) = 1 - Pr(H_a \mid x, y)$$

We are going to use Bayes theorem.

$$Pr(H_0 \mid x, y) = \frac{\overbrace{Pr(x, y \mid H_0)}^{\text{density}} Pr(H_0)}{\underbrace{Pr(x, y \mid H_0)}_{\text{density}} Pr(H_0) + \underbrace{Pr(x, y \mid H_a)}_{\text{density}} Pr(H_a)}$$

$$= \frac{f(x, y \mid H_0) Pr(H_0)}{f(x, y \mid H_0) Pr(H_0) + f(x, y \mid H_a) Pr(H_a)}$$

In the absence of other information, we might want to use

$$Pr(H_0) = Pr(H_a) = 1/2$$

$$f(x, y \mid H_0) = \int f(x, y \mid \theta_1, \theta_2, H_0)\, p(\theta_1, \theta_2 \mid H_0)\, d\theta_1 d\theta_2 \quad (2)$$

$$f(x, y \mid \theta_1, \theta_2, H_0) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \binom{m}{y} \theta^y (1-\theta)^{m-y}$$

$$f(x, y \mid \theta, H_0) = \binom{n}{x}\binom{m}{y} \theta^{x+y} (1-\theta)^{n+m-x-y}$$

$$p(\theta_1, \theta_2 \mid H_0) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}$$

$$p(\theta \mid H_0)$$

$$f(x, y \mid H_0) = \binom{n}{x}\binom{m}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{x+y+a-1} (1-\theta)^{n+m-x-y+b-1} d\theta$$

$$= \binom{n}{x}\binom{m}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(x+y+a)\,\Gamma(n+m-x-y+b)}{\Gamma(n+m+a+b)}$$

→ marginal likelihood (under model H_0)
prior predictive distribution (under model H_0)

$$f(\underline{x}, \underline{y} \mid H_a) = \iint f(\underline{x}, \underline{y} \mid \theta_1, \theta_2, H_a)\, p(\theta_1, \theta_2 \mid H_a)\, d\theta_1 d\theta_2$$

$$f(\underline{x}, \underline{y} \mid \theta_1, \theta_2, H_a) = \binom{n}{x} \theta_1^x (1-\theta_1)^{n-x} \binom{M}{y} \theta_2^y (1-\theta_2)^{m-y}$$

$$p(\theta_1, \theta_2 \mid H_a) = p(\theta_1 \mid H_a)\, p(\theta_2 \mid H_a)$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta_1^{a-1} (1-\theta_1)^{b-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta_2^{a-1} (1-\theta_2)^{b-1}$$

↳ Same a and b as under H_0!!

$$f(\underline{x}, \underline{y} \mid H_a) = \binom{n}{x}\binom{M}{y} \left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right]^2 \times \int \theta_1^{x+a-1} (1-\theta_1)^{n-x+b-1} d\theta_1$$

$$\times \int \theta_2^{y+a-1} (1-\theta_2)^{m-y+b-1} d\theta_2$$

$$= \binom{n}{x}\binom{M}{y} \left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right]^2 \frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)} \cdot \frac{\Gamma(a+y)\Gamma(b+m-y)}{\Gamma(a+b+m)}$$

$$\Pr(H_0 \mid \underline{x}, \underline{y}) = \frac{\frac{1}{2}\binom{n}{x}\binom{m}{y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(x+y+a)\Gamma(n+m-x-y+b)}{\Gamma(n+m+a+b)}}{\frac{1}{2}\binom{n}{x}\binom{m}{y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(x+y+a)\Gamma(n+m-x-y+b)}{\Gamma(n+m+a+b)} + \text{(4)}}$$

$$\rightarrow \frac{1}{2}\binom{n}{x}\binom{m}{y}\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right]^2 \frac{\Gamma(a+x)\Gamma(b+n-x)\Gamma(a+y)\Gamma(b+m-y)}{\Gamma(a+b+n)\Gamma(a+b+m)}$$

$$\frac{\Gamma(x+y+a)\Gamma(n+m-x-y+b)}{\Gamma(n+m+a+b)}$$

$$= \frac{\frac{\Gamma(x+y+a)\Gamma(n+m-x-y+b)}{\Gamma(n+m+a+b)}}{\frac{\Gamma(x+y+a)\Gamma(n+m-x-y+b)}{\Gamma(n+m+a+b)} + \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+x)\Gamma(b+n-x)\Gamma(a+y)\Gamma(b+m-y)}{\Gamma(a+b+n)\Gamma(a+b+m)}}$$

Other ways to write $\Pr(H_0 \mid \underline{x}, \underline{y})$.

$$\Pr(H_0 \mid \underline{x}, \underline{y}) = \frac{1}{1 + \underbrace{\frac{\Pr(H_a)}{\Pr(H_0)} \frac{\Pr(\underline{x}, \underline{y} \mid H_a)}{\Pr(\underline{x}, \underline{y} \mid H_0)}}_{\text{Posterior odds}}}$$

$$\Rightarrow \underbrace{\frac{\Pr(H_0 \mid \underline{x}, \underline{y})}{\Pr(H_a \mid \underline{x}, \underline{y})}}_{\text{Posterior odds}} = \underbrace{\frac{\Pr(H_0)}{\Pr(H_a)}}_{\text{Prior odds}} \underbrace{\frac{\Pr(\underline{x}, \underline{y} \mid H_0)}{\Pr(x, y \mid H_a)}}_{\text{Bayes factor}}$$

Count data and Poisson likelihoods:

Public health: estimating prevalence.

Data from different counties.

S_i = total population in the county.

X_i = # of recorded cases in the county.

You want to estimate the infection rate θ

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} S_i}$$ = maximum likelihood estimator under a Poisson model

$X_i | \theta \sim$ Poisson $(S_i \theta)$ independent $i = 1, \ldots, n$

$$p(x_1, x_2, \ldots x_n | \theta) = \prod_{i=1}^{n} \frac{(S_i\theta)^{x_i} \exp\{-S_i\theta\}}{x_i!}$$

$$= \frac{\prod_{i=1}^{n} (S_i)^{x_i}}{\prod_{i=1}^{n} x_i!} \theta^{\sum x_i} \exp\{-\theta \sum S_i\}$$

Bayesian equivalent?

Prior? How about $\theta \sim$ Gamma (a, b) a priori

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

$$p(\theta|x) = \frac{\cancel{\frac{\Pi (s_i)^{x_i}}{\Pi x_i!}} \theta^{\Sigma x_i} \exp\{-\theta \Sigma s_i\} \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp\{-b\theta\}}{\int \cancel{\frac{\Pi s_i^{x_i}}{\Pi x_i!}} \theta^{\Sigma x_i} \exp\{-\theta \Sigma s_i\} \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp\{-b\theta\} d\theta} \quad (6)$$

$$= \frac{\theta^{\Sigma x_i + a - 1} \exp\{-(b + \Sigma s_i)\theta\}}{\int \theta^{\Sigma x_i + a - 1} \exp\{-(b + \Sigma s_i)\theta\} d\theta}$$

$$\Rightarrow \theta|\underline{x} \sim \text{Gamma}(a + \Sigma x_i, b + \Sigma s_i)$$

$\phi \sim \text{Gamma}(a, b)$ $\theta|\phi \sim N(\mu, K\phi^{-1})$

$p(\theta, \phi | \underset{\sim}{x})$ is such that

$$\phi|\underset{\sim}{x} \sim \text{Gamma}(a_n, b_n)$$

where $a_n = a + \frac{n}{2}$

$$b_n = b + \frac{1}{2}\sum_{i=1}^{n} x_i^2 + \frac{1}{2}\frac{\mu^2}{K} - \frac{1}{2}\left[\frac{n\bar{x} + \frac{\mu}{K}}{n + \frac{1}{K}}\right]^2$$

$$\theta|\phi, \underset{\sim}{x} \sim N(\mu_n, K_n\phi^{-1})$$

$$\mu_n = \frac{n\bar{x} + \frac{\mu}{K}}{n + \frac{1}{K}}$$

$$K_n = n + \frac{1}{K}.$$

$$p(\theta, \phi | \underset{\sim}{x}) = \frac{b_n^{a_n}}{\Gamma(a_n)}\phi^{a_n - 1}\exp\{-\phi b_n\}$$

$$\times \frac{1}{\sqrt{2\pi}}\frac{\phi^{1/2}}{K_n^{1/2}}\exp\left\{-\frac{\phi}{2K_n}(\theta - \mu_n)^2\right\}$$

$$= \frac{1}{\sqrt{2\pi}}\frac{b_n^{a_n}}{\Gamma(a_n)}\frac{1}{K_n^{1/2}}\phi^{a_n + 1/2 - 1}\exp\left\{-\phi\left[b_n + \frac{(\theta - \mu_n)^2}{2K_n}\right]\right\}$$

$$p(\theta|\underset{\sim}{x}) = \int p(\theta, \phi|\underset{\sim}{x})\,d\phi = \int \frac{1}{\sqrt{2\pi}}\frac{b_n^{a_n}}{\Gamma(a_n)}\frac{1}{K_n^{1/2}}\phi^{a_n + 1/2 - 1}\exp\left\{-\phi\left[b_n + \frac{(\theta - \mu_n)^2}{2K_n}\right]\right\}d\phi$$

$$= \frac{1}{\sqrt{2\pi}}\frac{b_n^{a_n}}{\Gamma(a_n)}\frac{1}{K_n^{1/2}}\frac{\Gamma(a_n + 1/2)}{\left[b_n + \frac{(\theta - \mu_n)^2}{2K_n}\right]^{a_n + 1/2}}$$

$\Rightarrow$ This a (scaled and shifted t distribution)

$$p(\theta|\underline{x}) = \frac{1}{\sqrt{2\pi}} \frac{\Gamma(a_n + 1/2)}{\Gamma(a_n)} \frac{b_n^{a_n}}{K_n^{1/2}} \frac{1}{b_n^{a_n+1/2}}$$

$$\left[1 + \frac{(\theta - \mu_n)^2}{2 b_n K_n}\right]^{-(a_n+1/2)}$$

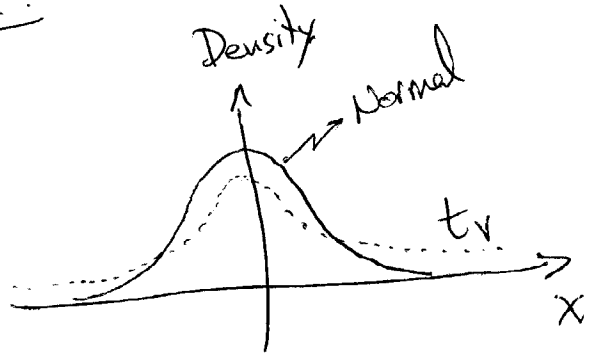
$$= \frac{1}{\sqrt{2\pi}} \frac{\Gamma(a_n+1/2)}{\Gamma(a_n)} \left(\frac{1}{b_n K_n}\right)^{1/2} \left[1 + \frac{(\theta-\mu_n)^2}{2 b_n K_n}\right]^{-(a_n+1/2)}$$

$$z = (\theta - \mu_n) \Big/ \sqrt{\frac{b_n K_n}{a_n}}$$

$$\Rightarrow p(z|\underline{x}) = \frac{1}{\sqrt{2\pi}} \frac{\Gamma(a_n + 1/2)}{\Gamma(a_n)} a_n^{-1/2} \left[1 + \frac{z^2}{2a_n}\right]^{-(a_n+1/2)}$$

$z \sim t_{2a_n}$ degrees of freedom

t distribution:



Point estimation.

$$\tilde{\theta} = E(\theta|\underline{x}) = \mu_n.$$

$$E(\theta|\underline{x}) = E\Big[E(\theta|\phi,\underline{x})\Big] = E\Big[\mu_n\Big] = \mu_n.$$

$$Var(\theta|\underline{x}) = Var\left(E(\theta|\phi,\underline{x})\right) + E\left(Var(\theta|\phi,\underline{x})\right)$$

$$= Var\left(\mu_n\right) + E\left[k_n\phi^{-1}\right]$$

$$= 0 + k_n \cdot \frac{b_n}{a_n - 1} = \frac{k_n b_n}{a_n - 1}$$

$$E(\underbrace{\phi^{-1}}_{\sigma^2}|\underline{x}) = \frac{b_n}{a_n - 1}$$

Credible intervals

Because the distribution is symmetric, the HPD credible interval and the symmetric interval match!!

Hypothesis testing:

$x_1, \dots x_n \sim N(\mu_1, \phi_1^{-1})$ $\quad y_1, \dots y_m \sim N(\mu_2, \phi_2^{-1})$

with ϕ_1 and ϕ_2 known.

$H_0: \mu_1 = \mu_2$ $\quad$ vs $\quad$ $H_a: \mu_1 \neq \mu_2$

$$Pr(H_0) = 1 - Pr(H_a) = 1/2$$

$$p(\underline{x}, \underline{y} \mid H_a, \mu_1, \mu_2) = \prod_{i=1}^{n} p(x_i \mid \mu_1) \prod_{j=1}^{m} p(y_j \mid \mu_2)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \phi_1^{1/2} \exp\left\{-\frac{\phi_1}{2} \sum_{i=1}^{n} (x_i - \mu_1)^2\right\} \times$$

$$\left(\frac{1}{\sqrt{2\pi}}\right)^m \phi_2^{1/2} \exp\left\{-\frac{\phi_2}{2} \sum_{j=1}^{m} (y_j - \mu_2)^2\right\}$$

$$p(\underline{x}, \underline{y} \mid H_0, \mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n+m} (\phi_1 \phi_2)^{1/2} \exp\left\{-\frac{\phi_1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{\phi_2}{2} \sum_{j=1}^{m} (y_j - \mu)^2\right\}$$

$\mu \mid H_0 \sim N(\eta, \tau^2)$

$\mu_1 \mid H_a \sim N(\eta, \tau^2)$ independent

$\mu_2 \mid H_a \sim N(\eta, \tau^2)$

$$p(\underline{x}, \underline{y} \mid H_a) = \iint p(\underline{x}, \underline{y} \mid H_a, \mu_1, \mu_2)\, \underbrace{p(\mu_1, \mu_2 \mid H_a)}_{p(\mu_1 \mid H_a)\, p(\mu_2 \mid H_a)}\, d\mu_1\, d\mu_2 \quad (5)$$

$$= \underbrace{\int \left(\frac{1}{\sqrt{2\pi}}\right)^n \phi_1^{1/2} \exp\left\{-\frac{\phi_1}{2}\sum_{i=1}^{n}(x_i-\mu_1)^2\right\} \frac{1}{\sqrt{2\pi}\tau}\exp\left\{-\frac{1}{2\tau^2}(\mu_1-\eta)^2\right\} d\mu_1}_{(1)}$$

$$\times \underbrace{\int \left(\frac{1}{\sqrt{2\pi}}\right)^m \phi_2^{1/2} \exp\left\{-\frac{\phi_2}{2}\sum_{j=1}^{m}(y_j-\mu_2)^2\right\} \frac{1}{\sqrt{2\pi}\tau}\exp\left\{-\frac{1}{2\tau^2}(\mu_2-\eta)^2\right\} d\mu_2}_{(2)}.$$

$$(1) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} (\phi_1\phi_2)^{1/2} \int \exp\left\{-\frac{\phi_1}{2}\sum_{i=1}^{n}(x_i^2 - 2x_i\mu_1 + \mu_1^2) - \frac{1}{2\tau^2}(\mu_1^2 - 2\mu_1\eta + \eta^2)\right\} d\mu_1$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} (\phi_1\phi_2)^{1/2} \exp\left\{-\frac{\phi_1}{2}\sum_{i=1}^{n}x_i^2 - \frac{\eta^2}{2\tau^2}\right\} \times$$

$$\underbrace{\int \exp\left\{-\frac{1}{2}\left(n\mu_1^2\phi_1 + \frac{\mu_1^2}{\tau^2} - 2\mu_1 n\bar{x}\phi_1 - \frac{2\mu_1\eta}{\tau^2}\right)\right\} d\mu_1}_{(3)}$$

$$(3) = \int \exp\left\{-\frac{1}{2}\left(n\phi_1 + \frac{1}{\tau^2}\right)\left(\mu_1^2 - 2\mu_1\left(\frac{n\phi_1\bar{x} + \frac{\eta}{\tau^2}}{n\phi_1 + \frac{1}{\tau^2}}\right)\right)\right\} d\mu_1$$

$$= \int \exp\left\{-\frac{1}{2}\left(n\phi_1 + \frac{1}{\tau^2}\right)\left[\left(\mu_1 - \frac{n\phi_1\bar{x} + \frac{\eta}{\tau^2}}{n\phi + \frac{1}{\tau^2}}\right)^2 - \left(\frac{n\phi_1\bar{x} + \frac{\eta}{\tau^2}}{n\phi + \frac{1}{\tau^2}}\right)^2\right]\right\} d\phi$$

$$= \exp\left\{\frac{1}{2}\left[\frac{\left(n\phi_1\bar{x} + \frac{\eta}{\tau^2}\right)^2}{n\phi + 1/\tau^2}\right]\right\} \sqrt{2\pi}\left(n\phi_1 + \frac{1}{\tau^2}\right)^{-1/2}.$$

2 → Complete!!!

Recap

$x_1, \dots x_n | \theta \overset{iid}{\sim} N(\theta, \phi^{-1})$ ϕ known, θ unknown

$\theta \sim N(\mu, \tau^2)$

$$\Rightarrow \theta | x_1, \dots x_n \sim N\left(\frac{n\phi\bar{x} + \frac{\mu}{\tau^2}}{n\phi + \frac{1}{\tau^2}}, \frac{1}{n\phi + \frac{1}{\tau^2}} \right)$$

$$E(\theta | x_1, \dots x_n) = \frac{n\phi\bar{x} + \frac{\mu}{\tau^2}}{n\phi + \frac{1}{\tau^2}} = w\bar{x} + (1-w)\mu$$

$$w = \frac{n\phi}{n\phi + \frac{1}{\tau^2}}$$

What is the posterior median in this case?

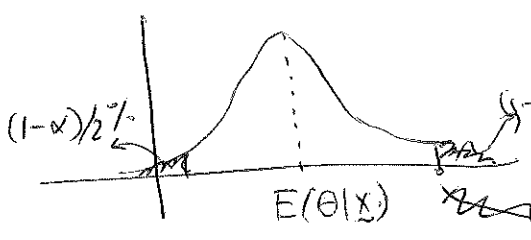
Same as before

What about posterior mode?

Same

What is the α% posterior (symmetric) credible interval?

$$\frac{n\phi\bar{x} + \frac{\mu}{\tau^2}}{n\phi + \frac{1}{\tau^2}} \pm z_{\alpha/2} \sqrt{\frac{1}{n\phi + 1/\tau^2}}$$



$$z_{\alpha/2} = \Phi\left(1 - \frac{(1-\alpha)}{2}\right)$$

If $\alpha = 95\%$ then

Credible interval $$\frac{n\phi\bar{x} + \frac{\mu}{\tau^2}}{n\phi + \frac{1}{\tau^2}} \pm 1.96 \sqrt{\frac{1}{n\phi + \frac{1}{\tau^2}}}$$

If $\tau^2 \to \infty$ this becomes

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} \qquad \text{(remember } \phi = \frac{1}{\sigma^2}\text{)}$$

Consider now the case θ is known and ϕ is unknown.

Remember the likelihood

$$p(x_1, \dots x_n | \phi) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \phi^{n/2} \exp\left\{-\frac{\phi}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right\}$$

$\phi \sim \text{Gamma}(a, b)$

$$\Rightarrow p(\phi) = \frac{b^a}{\Gamma(a)} \phi^{a-1} \exp\{-b\phi\}$$

$$p(\phi | \underset{\sim}{x}) \propto \phi^{n/2} \exp\left\{-\phi \frac{\sum_{i=1}^{n}(x_i - \theta)^2}{2}\right\} \phi^{a-1} \exp\{-b\phi\}$$

$$= \phi^{n/2 + a - 1} \exp\left\{-\phi\left[b + \frac{\sum_{i=1}^{n}(x_i - \theta)^2}{2}\right]\right\}$$

$$\Rightarrow \phi | \underset{\sim}{x} \sim \text{Gamma}\left(\frac{n}{2} + a,\ b + \frac{\sum_{i=1}^{n}(x_i - \theta)^2}{2}\right)$$

$$E(\phi | \underset{\sim}{x}) = \frac{\frac{n}{2} + a}{b + \frac{\sum(x_i - \theta)^2}{2}}$$

$\Gamma(x) = (x-1)\,\Gamma(x-1)$

[redacted]

$$E(\sigma^2 \mid x_1, \dots x_n) = E\left(\frac{1}{\phi} \mid x_1, \dots x_n\right)$$

$$= \int \frac{1}{\phi} \frac{\left(b + \frac{\sum(x_i-\theta)^2}{2}\right)^{a+\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+a\right)} \phi^{a+\frac{n}{2}-1} \exp\left\{-\phi\left(b + \frac{\sum(x_i-\theta)^2}{2}\right)\right\} d\phi$$

$$= \frac{\left(b + \frac{\sum(x_i-\theta)^2}{2}\right)^{a+\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+a\right)} \frac{\Gamma\left(\frac{n}{2}+a-1\right)}{\left(b + \frac{\sum(x_i-\theta)^2}{2}\right)^{a+\frac{n}{2}-1}}$$

$$= \frac{b + \frac{\sum(x_i-\theta)^2}{2}}{a + \frac{n}{2} - 1}$$

Let's derive the distribution of τ^2.

$\eta = \tau^2 = 1/\phi$ $\quad d\eta = \frac{d\phi}{\phi^2} \Rightarrow \phi = \frac{1}{\eta} \quad d\phi = \frac{d\eta}{\eta^2}$

$$P(\eta) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{\eta}\right)^{a-1} \exp\left\{-\frac{b}{\eta}\right\} \frac{1}{\eta^2}$$

$$= \frac{b^a}{\Gamma(a)} \left(\frac{1}{\eta}\right)^{a+1} \exp\left\{-\frac{b}{\eta}\right\}$$

↳ The distribution of η is called the inverse-gamma distribution.

We are calling $\gamma = \sigma^2$ so

$$p(\sigma^2) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left\{-\frac{b}{\sigma^2}\right\} \quad IG(a, b)$$

Now with this prior

$$\sigma^2 | \underline{x} \sim IG\left(a + \frac{n}{2}, \; b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right)$$

Now, what if both θ and ϕ are unknown?

$$p(x_1, \dots x_n | \theta, \phi) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \phi^{n/2} \exp\left\{-\frac{\phi}{2}\sum(x_i - \theta)^2\right\}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \phi^{n/2} \exp\left\{-\frac{\phi}{2}\left[\sum x_i^2 - 2n\bar{x}\theta + \theta^2\right]\right\}$$

$\theta | \phi \sim N\left(\mu, \frac{K}{\phi}\right)$

$\phi \sim \text{Gamma}(a, b)$

$\Rightarrow$ This is different from

$\theta \sim N(\mu, \tau^2)$

$\phi \sim \text{Gamma}(a, b)$!!

$$p(\theta,\phi\mid \underset{\sim}{x}) \propto \phi^{n/2}\exp\left\{-\frac{\phi}{2}\left[\sum x_i^2 - 2n\bar{x}\theta + n\theta^2\right]\right\}\left(\frac{\phi}{\kappa}\right)^{1/2}\exp\left\{-\frac{\phi}{2\kappa}(\theta-\mu)^2\right\}$$

$$\phi^{a-1}\exp\{-b\phi\}$$

$$\propto \phi^{n/2+a+\frac{1}{2}-1}\exp\left\{-\frac{\phi}{2}\left[2b+\sum x_i^2 - 2n\bar{x}\theta + n\theta^2 + \frac{\theta^2}{\kappa} - \frac{2\theta\mu}{\kappa} + \frac{\mu^2}{\kappa}\right]\right\}$$

$$p(\phi\mid \underset{\sim}{x}) = \int p(\theta,\phi\mid \underset{\sim}{x})\,d\theta$$

$$\propto \int \phi^{n/2+a+\frac{1}{2}-1}\exp\left\{-\frac{\phi}{2}\left[2b + \sum x_i^2 - 2n\bar{x}\theta + n\theta^2 + \frac{\theta^2}{\kappa} - \frac{2\theta\mu}{\kappa} + \frac{\mu^2}{\kappa}\right]\right\}d\theta$$

$$= \phi^{\frac{n}{2}+a+\frac{1}{2}-1}\exp\left\{-\frac{\phi}{2}\left[2b + \sum x_i^2 + \frac{\mu^2}{\kappa}\right]\right\}$$

$$\int \exp\left\{-\frac{\phi}{2}\left[\theta^2\left(n+\frac{1}{\kappa^2}\right) - 2\theta\left[n\bar{x} + \frac{\mu}{\kappa}\right]\right]\right\}d\theta$$

$$\left[2\pi\right]^{-1/2}\left[\phi\left(1+\frac{1}{\kappa^2}\right)\right]^{-1/2}.$$

$$\propto \phi^{\frac{n}{2}+a-1}\exp\left\{-\frac{\phi}{2}\left[2b + \sum x_i^2 + \frac{\mu^2}{\kappa}\right]\right\}.$$

$$\Rightarrow \phi\mid \underset{\sim}{x} \sim \text{Gamma}\left(\frac{n}{2}+a,\ b + \frac{\sum x_i^2}{2} + \frac{\mu^2}{2\kappa}\right).$$

$$\int \exp\left\{-\frac{\phi}{2}\left[\theta^2\left(n+\frac{1}{K}\right) - 2\theta\left[n\bar{x} + \frac{\mu}{K}\right]\right]\right\} d\theta$$

$$= \int \exp\left\{-\frac{\phi}{2}\left(n+\frac{1}{K}\right)\left[\left(\theta - \frac{n\bar{x}+\frac{\mu}{K}}{n+\frac{1}{K}}\right)^2 - \left(\frac{n\bar{x}+\frac{\mu}{K}}{n+\frac{1}{K}}\right)^2\right]\right\} d\theta$$

$$= \exp\left\{\frac{\phi}{2}\,\frac{\left(n\bar{x}+\frac{\mu}{K}\right)^2}{\left(n+\frac{1}{K}\right)}\right\} \underbrace{\int \exp\left\{-\frac{\phi}{2}\left(n+1/K\right)\left[\theta - \frac{n\bar{x}+\mu/K}{n+1/K}\right]^2\right\} d\theta}_{\sqrt{2\pi}\left[\phi\left(1+1/K\right)\right]^{-1/2}}$$

$$\Rightarrow p(\phi|\underline{x}) \propto \phi^{\frac{n}{2}+a-1} \exp\left\{-\frac{\phi}{2}\left[2b + \sum_{i=1}^{n} x_i^2 + \frac{\mu^2}{K} - \frac{\left(n\bar{x}+\mu/K\right)^2}{\left(n+1/K\right)}\right]\right\}$$

$$\Rightarrow \phi|\underline{x} \sim \text{Gamma}\left(\frac{n}{2}+a,\ 2b + \frac{1}{2}\sum_{i=1}^{n} x_i^2 + \frac{1}{2}\frac{\mu^2}{K} - \frac{1}{2}\frac{\left(n\bar{x}+\mu/K\right)^2}{\left(n+1/K\right)}\right)$$

$$p(\theta|\phi,\underline{x}) = \frac{p(\theta,\phi|\underline{x})}{p(\phi|\underline{x})}$$

$$p(\theta|\phi,\underline{x}) \propto \frac{\phi^{\frac{n}{2}+a+\frac{1}{2}-1}\exp\left\{-\frac{\phi}{2}\left[2b+\sum x_i^2 - 2n\bar{x}\theta + n\theta^2 + \frac{\theta^2}{\kappa} - \frac{2\theta\mu}{\kappa} + \frac{\mu^2}{\kappa}\right]\right\}}{\phi^{n/2+a-1}\exp\left\{-\frac{\phi}{2}\left[2b+\sum x_i + \frac{\mu^2}{\kappa}\right]\right\}}$$

$$= \phi^{1/2}\exp\left\{-\frac{\phi}{2}\left[\theta^2\left(n+\frac{1}{\kappa}\right) - 2\theta\left(n\bar{x}+\frac{\mu}{\kappa}\right)\right]\right\}$$

$$\propto \exp\left\{-\frac{\phi}{2}\left[\theta^2\left(n+\frac{1}{\kappa}\right) - 2\theta\left(n\bar{x}+\frac{\mu}{\kappa}\right)\right]\right\}$$

$$\theta|\phi,\underline{x} \sim N\left(\frac{n\bar{x}+\frac{\mu}{\kappa}}{\phi\left(n+\frac{1}{\kappa}\right)}, \left[\phi\left(n+\frac{1}{\kappa}\right)\right]^{-1}\right)$$

↳ check!!!

webcast.ucsc.edu

Back on our survey example:

How would we obtain point and interval estimates from a classical perspective for this same problem.

Point estimator: $\frac{\sum_{i=1}^{n} y_i}{n} = \hat{\theta}$ (Maximum likelihood est).

Interval estimator: $\hat{\theta} \pm 2\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \Rightarrow$ For a $\simeq 95\%$ confidence interval.

The classical results are not too far away from the Bayesian. In fact, as n increases the differences diminish.

More generally, under some mild assumptions, point and interval estimates constructed using Bayesian methods behave like their frequentist counterparts.

Batch vs joint update

Suppose you got an additional sample of $m = 63$ women with $\sum x_i = 17$ positive answers.

Batch update.

$\theta \sim \text{Beta}(a=10, b=30)$

$\Downarrow$ $n = 129$, $\sum y_i = 25$

$\theta \sim \text{Beta}(a = \underbrace{10+25}_{35}, b = \underbrace{30+129-25}_{134})$

Because of conjugacy you stay in the same family. ⟹

$\Downarrow$ $m = 63$, $\sum x_i = 17$

$\theta \sim \text{Beta}(a = 52, b = 180)$

$$P_1(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} = \frac{P(y|\theta)P(\theta)}{\int P(y|\theta)P(\theta)\,d\theta}$$

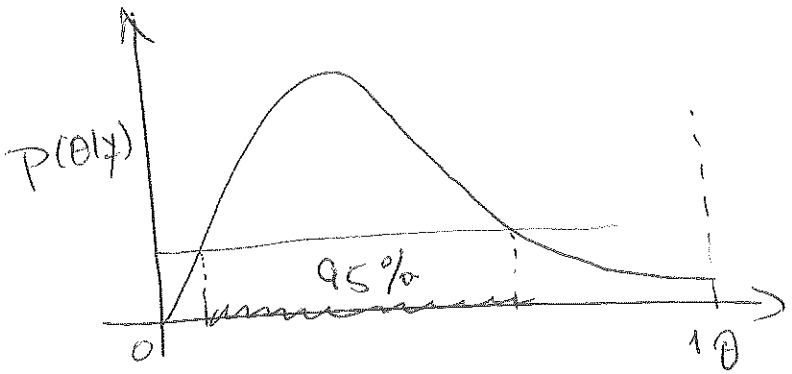
$$P_2(\theta|y, x) = \frac{P(x|\theta, y)P_1(\theta|y)}{\int P(x|\theta, y)P_1(\theta|y)\,d\theta}$$

Note that $P(x|\theta, y) = P(x|\theta)$ in this case.

$$= \frac{P(x|\theta)P(y|\theta)P(\theta)/P(y)}{\int P(x|\theta)P(y|\theta)P(\theta)/P(y)\,d\theta}$$

$$= \frac{p(y|\theta)\,p(x|\theta)\,p(\theta)}{\int p(y|\theta)\,p(x|\theta)\,p(\theta)\,d\theta} = \frac{p(\underset{\sim}{x}, y|\theta)\,p(\theta)}{\int p(\underset{\sim}{x}, y|\theta)\,p(\theta)\,d\theta}$$

HPD vs symmetric intervals:



HPD interval = shorter interval that contains a pre-specified amount of probability (Highest Posterior Density)

They satisfy $p(l|y) = p(u|y)$ (in unimodal settings)

$$\int_l^u p(\theta|y) = 0.95$$

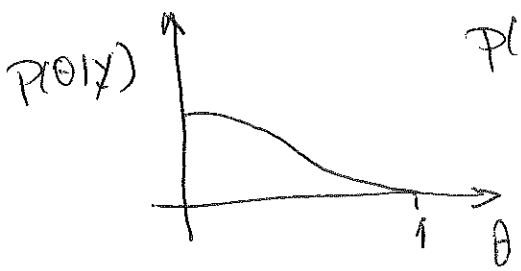
In symmetric posteriors, HPD and symmetric intervals are the same.

$\theta \sim \text{Beta}(1,1) = \text{Uniform}[0,1]$

$n = 129$

$\sum y_i = 0$

$\Rightarrow \theta|y \sim \text{Beta}(1, 130)$



$$p(\theta|y) = \frac{\Gamma(131)}{\Gamma(130)}(1-\theta)^{129}$$

$$= 130(1-\theta)^{129}$$

Mean / median / mode?

↳ utility functions / decision theory

Post Means are not invariant to transformation.

$$E(g(\theta)) \neq g(E(\theta))$$

However the median is invariant to monotone transformations.

Thinking About (joint) means of multiple parameters is easy; defining a median for various parameters is tricky.

In practice I will tend to use the posterior mean.

Hypothesis testing in two populations.

Why is ADHD diagnosis going up for younger kids.

Two populations (before and after cutoff date)

$x = 15$ $\quad$ $y = 12$

$n = 87$ $\quad$ $m = 63$

Before $\quad$ After

$X \sim Bin(n, \theta_1)$ $\quad$ $Y \sim Bin(n, \theta_2)$

Contrast $\quad H_0: \theta_1 = \theta_2 \quad H_a: \theta_1 \neq \theta_2$

From a Bayesian perspective we are going to treat the hypotheses as random quantities.

$$P(H_0) = 1 - P(H_a)$$

$P(x, y \mid H_0, \theta_1, \theta_2)$ $\quad$ $P(x, y \mid H_a, \theta_1, \theta_2)$

$P(\theta_1, \theta_2 \mid H_0)$ $\quad$ $P(\theta_1, \theta_2 \mid H_a)$.

$$P(H_0 \mid \underset{\sim}{x}, \underset{\sim}{y}) = \frac{P(\underset{\sim}{x}, \underset{\sim}{y} \mid H_0) P(H_0)}{P(\underset{\sim}{x}, \underset{\sim}{y} \mid H_0) P(H_0) + P(x, y \mid H_a) P(H_a)}$$

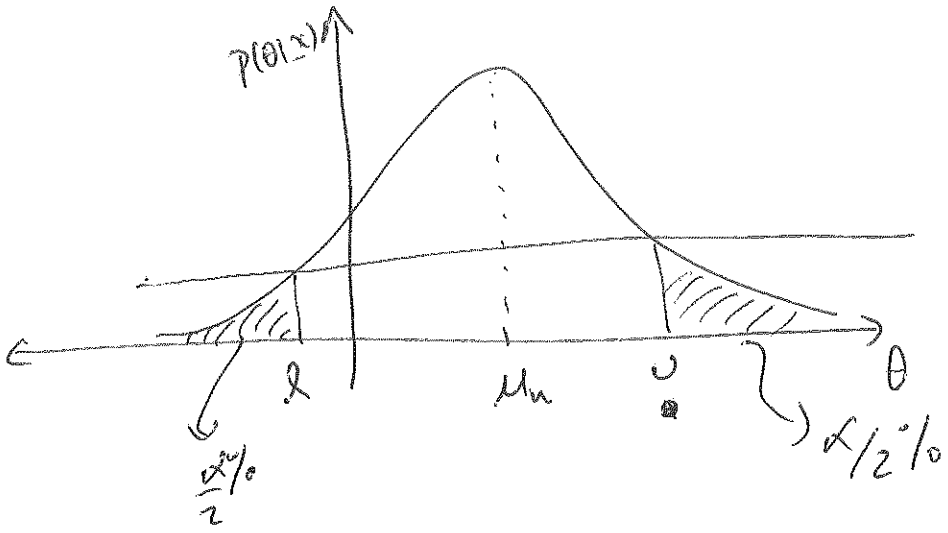
$$P(\underset{\sim}{x}, \underset{\sim}{y} \mid H_0) = \int P(\underset{\sim}{x}, \underset{\sim}{y} \mid \theta_1, \theta_2 H_0) P(\theta_1, \theta_2 \mid H_0) d\theta_1 d\theta_2$$

Credible interval

$$X_i \mid \theta \sim N(\theta, \sigma^2) \quad i = 1, \ldots, n$$

$$\theta \sim N(\mu, \tau^2)$$

$$\theta \mid \underline{x} \sim N\left(\underbrace{\frac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}}_{\mu_n}, \underbrace{\frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}}_{\sigma_n^2} \right)$$



For a $(1-\alpha)\%$ HPD credible interval

$$\frac{\theta - \mu_n}{\sigma_n} \mid \underline{x} \sim N(0,1)$$

$$Pr\left(\frac{\theta - \mu_n}{\sigma_n} > z_{\alpha/2} \right) = \frac{\alpha}{2}$$

$$\Rightarrow 1 - Pr\left(\frac{\theta - \mu_n}{\sigma_n} > z_{\alpha/2} \right) = 1 - \frac{\alpha}{2}$$

$$Pr\left(\frac{\theta - \mu_n}{\sigma_n} < z_{\alpha/2} \right) = 1 - \alpha/2$$

$$\Rightarrow z_{\alpha/2} = qnorm\left(1 - \frac{\alpha}{2}\right)$$

For $\alpha = 5\%$ then

$$z_{\alpha/2} = q_{norm}(0.975) \cong 1.964$$

$$\begin{cases} U = \mu_n + 1.964\,\sigma_n \\ l = \mu_n - 1.964\,\sigma_n \end{cases}$$

Exponential families and conjugacy:

Let $f(\underset{\sim}{x}|\underset{\sim}{\theta})$ be of the form:

$$f(\underset{\sim}{x}|\underset{\sim}{\theta}) = C(\underset{\sim}{\theta})h(\underset{\sim}{x})\exp\left\{\underset{\sim}{R}^T(\underset{\sim}{\theta})\cdot\underset{\sim}{T}(\underset{\sim}{x})\right\}$$

where $\underset{\sim}{x}\in\mathcal{X}$ and $\underset{\sim}{\theta}\in\Theta\in\mathbb{R}^k$ such that the form of $\mathcal{X}$ is independent of Θ

f is said to belong to the exponential family.

A particularly interesting case arises when $\underset{\sim}{R}(\underset{\sim}{\theta})\in\mathbb{R}^k$ and $\underset{\sim}{T}(\underset{\sim}{x})\in\mathbb{R}^k$ in which case we can rewrite.

$$f(\underset{\sim}{x}|\underset{\sim}{\theta}) = C^*(\underset{\sim}{\eta})h(\underset{\sim}{x})\exp\left\{\underset{\sim}{\eta}^T\cdot T(\underset{\sim}{x})\right\}$$

where $\underset{\sim}{\eta} = R(\underset{\sim}{\theta})$

$\underset{\sim}{\eta}$ the natural parameter of the distribution.

Examples:

$x_i|\theta \sim Ber(\theta) \quad i=1,\dots,n$

$$\Rightarrow f(\underset{\sim}{x}|\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} = (1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^{\sum x_i}$$

$$= (1-\theta)^n\exp\left\{\left(\sum_{i=1}^n x_i\right)\cdot\log\frac{\theta}{1-\theta}\right\}$$

$$\eta = \log\frac{\theta}{1-\theta} \quad T(\underset{\sim}{x}) = \sum_{i=1}^n x_i \quad C(\theta) = (1-\theta)^n \quad h(\underset{\sim}{x}) = 1$$

~~[redacted]~~

$X_i | \theta, \sigma^2 \sim N(\theta, \sigma^2)$ θ, σ^2 ~~[redacted]~~ Unknown

$$f(x_1, \dots x_n | \theta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2\right\}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2} \left(\sum x_i^2 - 2n\bar{x}\theta + \theta^2\right)\right\}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{\theta^2}{2\sigma^2}\right\}$$

$$\exp\left\{-\frac{\sum x_i^2}{2\sigma^2} + \frac{n\bar{x}\theta}{\sigma^2}\right\}$$

$K = 2$

$$T(\underline{x}) = \left(\sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i\right) \qquad R(\theta, \sigma^2) = \left(\frac{1}{2\sigma^2}, \frac{\theta}{\sigma^2}\right)$$

[redacted] prior

Note that if $x_1, \dots x_n$ are iid from $f(x_i|\underline{\theta})$ and $f(x_i|\underline{\theta})$ belong to the exponential family, then $f(\underline{x}|\underline{\theta})$ also belongs to the exponential family.

$$f(\underline{x}|\underline{\theta}) = \prod_{i=1}^{n} c(\theta)\, h(x_i) \exp\left\{ \cdot T^T(x_i)\, R(\theta) \right\}$$

$$= [c(\theta)]^n \left[\prod_{i=1}^{n} h(x_i)\right] \exp\left\{ \left(\sum_{i=1}^{n} T(x_i)\right)^T R(\theta) \right\}$$

Now consider a prior of the form.

[redacted]

$$p(\theta|\lambda, \mu) = [c(\theta)]^{\lambda}\, h^*(\mu, \lambda) \exp\{ \mu^T \theta \}$$

Is this prior in the exponential family?

Consider the likelihood

$$f(\underline{x}|\eta) = C(\eta)\, h(\underline{x}) \exp\left\{ T^T(\underline{x})\, \eta \right\}$$

$$p(\eta) = [C^*(\eta)]^{\lambda} h^*(\mu, \lambda) \exp\left\{ \mu^T \eta \right\}$$

$$p(\eta|\underline{x}) \propto [C^*(\eta)]^{\lambda+1} \exp\left\{ [T(\underline{x}) + \mu]^T \eta \right\}$$

$$p(\eta|\underline{x}) = [C^*(\eta)]^{\lambda+1} h^*(\mu + T(\underline{x}), \lambda+1) \exp\left\{ [T(\underline{x}) + \mu]^T \eta \right\}$$

↳ conjugate prior!!!

$x_i | \lambda \sim \text{Poisson}(\lambda) \quad i = 1, \dots n$

$\lambda \sim \text{Gamma}(a, b)$

$$p(\underline{x}|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \left(\prod_{i=1}^{n} \frac{1}{x_i!} \right) e^{-n\lambda} \underbrace{\lambda^{\sum x_i}}_{\exp\{\sum x_i \log \lambda\}}$$

$\eta = \log \lambda$

$T(\underline{x}) = \sum x_i$

$c^*(\eta) = \exp\{-n \exp\{\eta\}\}$

$$p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp\{-b\lambda\} \qquad \lambda = \exp\{\eta\}, \quad d\lambda = \exp\{\eta\} d\eta.$$

$$p(\eta) = \frac{b^a}{\Gamma(a)} \exp\{(a-1)\eta\} \exp\{-\exp\{\eta\} b\}.$$

Complete the matching of terms.

Conjugate priors:

$$f(\underset{\sim}{x}|\underset{\sim}{\eta}) = C(\underset{\sim}{\eta})\, h(\underset{\sim}{x}) \exp\left\{ \underset{\sim}{\eta}^T \underset{\sim}{T}(\underset{\sim}{x}) \right\} \Rightarrow$$ Natural parameter exponential family

↳ Conjugate prior

$$p(\underset{\sim}{\eta}|\lambda, \underset{\sim}{x}_0) \propto \exp\left\{ \underset{\sim}{\eta}^T \underset{\sim}{x}_0 + \lambda \log c(\eta) \right\}$$

⇒ Posterior

$$p(\underset{\sim}{\eta}|\underset{\sim}{x}) \propto \exp\left\{ \underset{\sim}{\eta}^T (\underset{\sim}{x}_0 + T(\underset{\sim}{x})) + (\lambda + 1) \log c(\eta) \right\}$$

Updated parameters are

$$\lambda_n = \lambda + 1$$

$$\underset{\sim}{x}_n = \underset{\sim}{x}_0 + T(\underset{\sim}{x})$$

For the Poisson likelihood:

$$p(x_1, \dots x_n|\lambda) = \frac{e^{-\lambda n} \lambda^{\sum x_i}}{\prod_{i=1} x_i!} \propto \lambda^{\sum x_i} \exp\{-\lambda n\} = \exp\left\{ \sum x_i \overbrace{\log \lambda}^{\eta} - n\lambda \right\}$$

$$\eta = \log \lambda$$

$$p(x_1, \dots x_n|\eta) \propto \exp\left\{ \eta \sum x_i - n \exp\{\eta\} \right\}$$

$$p(\eta) \propto \exp\left\{ \eta x_0 - \lambda \exp\{\eta\} \right\}$$

$$p(\eta|\underset{\sim}{x}) \propto \exp\{\eta(\gamma_0 + \textstyle\sum x_i) - (\gamma + n)\exp\{\eta\}\}$$

$$\Downarrow$$

$$\exp\left\{\eta\left(\gamma\left(\frac{\gamma_0}{\gamma}\right) + \textstyle\sum x_i\right) - (\gamma+n)\exp\{\eta\}\right\}$$

Back transform:

$$\eta = \log\lambda \qquad d\eta = \frac{d\lambda}{\lambda}$$

$$\Rightarrow p(\lambda|\underset{\sim}{x}) \propto \exp\{\log\lambda(\overbrace{\gamma_0 + \textstyle\sum x_i}^{x_n}) - \overbrace{(\gamma+n)}^{\gamma_n}\lambda\}\frac{1}{\lambda}$$

$$\propto \lambda^{\gamma_0 + \sum x_i - 1}\exp\{-(\gamma+n)\lambda\}$$

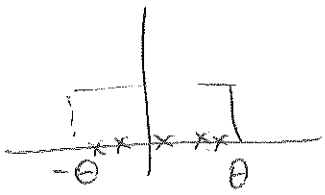
$$\Rightarrow \lambda|\underset{\sim}{x} \sim \text{Gamma}(\gamma_0 + \textstyle\sum x_i, \gamma + n)$$

Let $x_1, \dots x_n \overset{iid}{\sim} f(x_i|\theta)$ and $E(x) = \xi(\theta)$ with f in the exponential family with conjugate prior $\pi(\theta) \propto \exp\{\theta\gamma_0 - \lambda\psi_0\}$ then

$$E(\xi(\theta)) = \frac{\gamma_0}{\lambda} \quad \text{and} \quad E(\xi(\theta)|\underset{\sim}{x}) = \frac{\lambda(\gamma_0/\lambda) + n\bar{x}}{\lambda + n}$$

Conjugacy outside the exponential family

$X_1, \dots X_n \overset{iid}{\sim} U_n[-\theta, \theta]$



$$f(x_1, \dots x_n | \theta) = \begin{cases} \left(\frac{1}{2\theta}\right)^n & \text{if } -\theta \le x_1 \le \theta \\ & \vdots \\ & -\theta \le x_n \le \theta \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} \left(\frac{1}{2\theta}\right)^n & \text{if } \theta > \max_i |x_i| \\ 0 & \text{otherwise} \end{cases}$$

2) Not in the exponential family.

$$f(\underset{\sim}{x} | \theta) = \begin{cases} \left(\frac{1}{2\theta}\right)^n & \theta > \max_i \{|x_i|\} \\ 0 & \text{otherwise} \end{cases}$$

$$p(\theta) = \begin{cases} \frac{\beta \alpha^\beta}{\theta^{\beta+1}} & \theta > \alpha \Rightarrow \text{Pareto}(\alpha, \beta) \\ 0 & \text{otherwise} \end{cases}$$

$$p(\theta | \underset{\sim}{x}) \propto \left(\frac{1}{\theta}\right)^n \cdot \left(\frac{1}{\theta}\right)^{\beta+1} \quad \text{if } \theta > \max\left\{\alpha, \max_i |x_i|\right\}$$

$\alpha_n = \max\{\alpha, \max|x_i|\}$ $\beta_n = \beta + n$

$\theta|\underline{x} \sim \text{Pareto}\left(\max\{\alpha, \max\{x_i\}\}, \beta+n\right)$

↳ Pareto is not in the exponential family either (if you consider both α and β as parameters)

Closed form expression outside conjugacy.

$x_1, \dots x_n|\theta \overset{iid}{\sim} N(\theta, \sigma^2)$

$$p(\theta) = \frac{1}{2\lambda}\exp\left\{-\frac{1}{\lambda}|\theta-\mu|\right\}$$

$$p(\underline{x}|\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum x_i^2 - 2\theta n\bar{x} + n\theta^2\right]\right\}$$

$$p(\theta|\underline{x}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[n\theta^2 - 2n\theta\bar{x}\right] - \frac{1}{\lambda}|\theta-\mu|\right\}$$

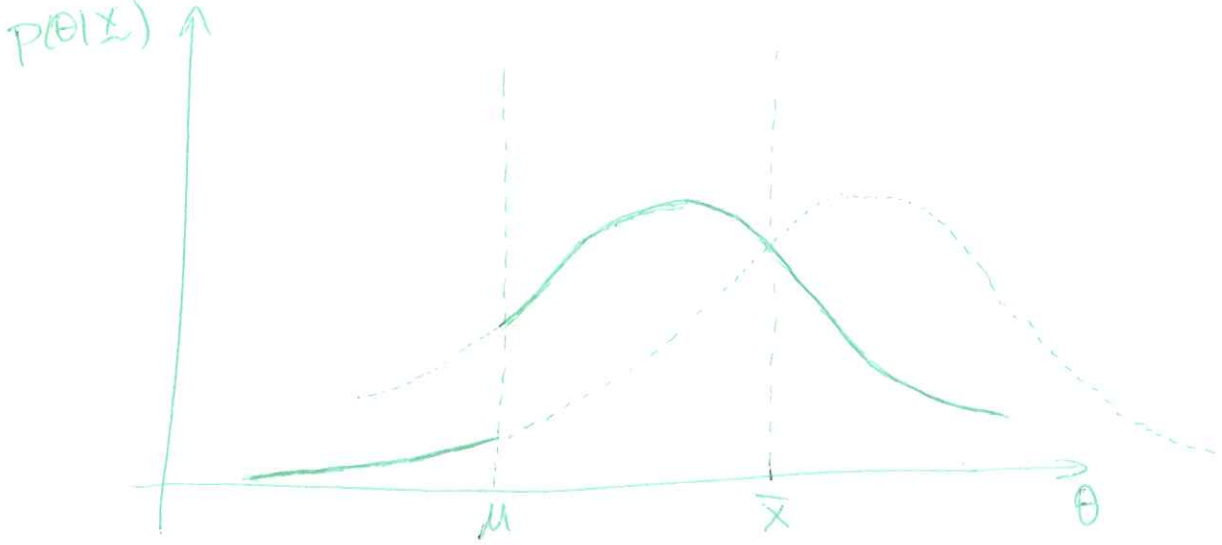
For $\theta > \mu$ $\quad |\theta-\mu| = \theta-\mu$

$$\Rightarrow p(\theta|\underline{x}) \propto \exp\left\{-\frac{1}{2}\left[\frac{n\theta^2}{\sigma^2} - \frac{2n\theta\bar{x}}{\sigma^2} - \frac{2}{\lambda}(\theta-\mu)\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\frac{n\theta^2}{\sigma^2} - 2\theta\left[\frac{n\bar{x}}{\sigma^2} + \frac{1}{\lambda}\right]\right]\right\}$$

For $\theta < \mu$

$$p(\theta|\underline{x}) \propto \exp\left\{-\frac{1}{2}\left[\frac{n\theta^2}{\sigma^2} - 2\theta\left[\frac{n\bar{x}}{\sigma^2} - \frac{1}{\lambda}\right]\right]\right\}$$


p(θ|x)
μ
x̄
θ

$$\int_{-\infty}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\frac{x^2}{\sigma^2}\right\} dx = \frac{1}{2}$$

(1)

$$E\left[\mathbb{1}_{(-\infty, m)}(x)\right] \qquad X \sim N(0, \sigma^2)$$

$$\mathbb{1}_{\Omega}(x) = \begin{cases} 1 \text{ if } x \in \Omega \\ 0 \text{ if } x \notin \Omega \end{cases}$$

How do we make it easier to compute expectations?

Use the CLT (Central Limit Theorem) and the LLN (Law of Large Numbers)

LLN:

$$E(f(x)) \simeq \int f(x) p(x) dx \simeq \frac{1}{N}\sum_{i=1}^{n} f(x^{(i)}) \quad x^{(i)} \sim p.$$

~~pdf~~ pmf for X

Computing integrals can be reduced to a problem of sampling from a probability distribution.

The power of simulation Example:

		Test: Diseased	Test: Healthy
Truth	Diseased	✓	Error
	Healthy	Error	✓

False Positive rate: $\Pr(+ \mid \bar{D})$

True positive rate: $\Pr(+ \mid D)$

} Easy to get from data but not what you really care about

We care about $\Pr(D \mid +)$

$\Pr(\bar{D} \mid -)$

$$\Pr(D \mid +) = \frac{\Pr(+ \mid D)\Pr(D)}{\Pr(+ \mid D)\Pr(D) + \Pr(+ \mid \bar{D})\Pr(\bar{D})}$$

We will also need data on $\Pr(D)$ = incidence of the disease

Data:

x = # of positives in the test in a population of n_1 diseased individuals

y = # of positives in the test in a population of n_2 healthy individuals

z = # of diseased individuals in a sample of n subjects

independent
$$\begin{cases} X \sim \text{Bin}(n_1, \theta_1) \\ Y \sim \text{Bin}(n_2, \theta_2) \\ Z \sim \text{Bin}(m, \phi) \end{cases}$$

θ_1 = true positive rate = $\Pr(+|D)$

θ_2 = false positive rate = $\Pr(+|\bar{D})$

ϕ = incidence = $\Pr(D)$

η = True diagnostic probability = $\Pr(D|+)$.
is my quantity of interest

$$\eta = \frac{\theta_1 \phi}{\theta_1 \phi + \theta_2 (1-\phi)}$$

Let's start with the interpretable parameters

$$\left.\begin{array}{l} \theta_1 \sim \text{Beta}(a_1, b_1) \\ \theta_2 \sim \text{Beta}(a_2, b_2) \\ \phi \sim \text{Beta}(a_3, b_3) \end{array}\right\} \text{independent}$$

A posteriori: they are independent

$$\left.\begin{array}{l} \theta_1|x \sim \text{Beta}(\overbrace{a_1 + x}^{a_{1,n}}, \overbrace{b_1 + n_1 - x}^{b_{1,n}}) \\ \theta_2|y \sim \text{Beta}(a_2 + y, b_2 + n_2 - y) \\ \phi|z \sim \text{Beta}(a_3 + z, b_3 + m - z) \end{array}\right\} \text{independent}$$

$$\begin{aligned} p(\theta_1, \theta_2, \phi | x, y, z) &\propto p(x|\theta_1) p(y|\theta_2) p(z|\phi) p(\theta_1) p(\theta_2) p(\phi) \\ &= \left[p(x|\theta_1) p(\theta_1)\right]\left[p(y|\theta_2) p(\theta_2)\right]\left[p(z|\phi) p(\phi)\right] \\ &= p(\theta_1|x)\, p(\theta_2|y)\, p(\phi|z) \end{aligned}$$

How would you proceed to obtain, for example, the posterior mean for η

$$E(\eta \mid X,Y,Z) = \iiint \frac{\theta_1 \phi}{\theta_1 \phi + \theta_2 (1-\phi)} \frac{\Gamma(a_{1,n}+b_{1,n})}{\Gamma(a_{1,n})\Gamma(b_{1,n})} \theta_1^{a_{1,n}-1} (1-\theta_1)^{b_{1,n}-1}$$

$$\frac{\Gamma(a_{2,n}+b_{2,n})}{\Gamma(a_{2,n})\Gamma(b_{2,n})} \theta_2^{a_{2,n}-1} (1-\theta_2)^{b_{2,n}-1}$$

$$\frac{\Gamma(a_{3,n}+b_{3,n})}{\Gamma(a_{3,n})\Gamma(b_{3,n})} \phi^{a_{3,n}-1} (1-\phi)^{b_{3,n}-1} \, d\theta_1 \, d\theta_2 \, d\phi$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \frac{\theta_1^{(i)} \phi^{(i)}}{\theta_1^{(i)} \phi^{(i)} + \theta_2^{(i)} (1-\phi^{(i)})}$$

$$\theta_1^{(i)}, \theta_2^{(i)}, \phi^{(i)} \overset{iid}{\sim} p(\theta_1, \theta_2, \phi \mid X, Y, Z)$$

$$\Downarrow$$

$$\text{ind.} \begin{cases} \theta_1^{(i)} \sim \text{Beta}(a_{1,n}, b_{1,n}) \\ \theta_2^{(i)} \sim \text{Beta}(a_{2,n}, b_{2,n}) \\ \phi^{(i)} \sim \text{Beta}(a_{3,n}, b_{3,n}) \end{cases}$$

The posterior of interest was "simple" in this case, so we can directly simulate from the posterior.

For most problems, however, the posterior might not have a simple form:

$$x_1, \dots x_n \overset{\theta,\phi}{\sim} N(\theta, \phi^{-1}) \qquad \theta, \phi \text{ unknown.}$$

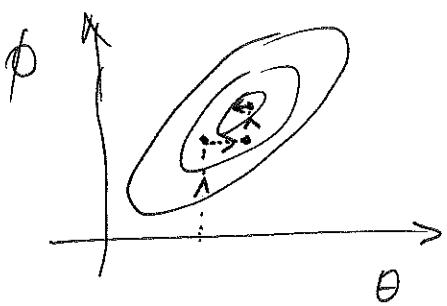
$$\left.\begin{aligned} \theta &\sim N(\mu, \tau^2) \\ \phi &\sim \text{Gamma}(a, b) \end{aligned}\right\} \Rightarrow \text{Not conjugate!!}$$

Posterior:

$$p(\theta, \phi \mid \underline{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \phi^{-n/2} \exp\left\{-\frac{\phi}{2}\left(\sum x_i^2 - 2\theta n\bar{x} + n\theta^2\right)\right\}$$
$$\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\tau^2}\left(\theta^2 - 2\theta\mu + \mu^2\right)\right\} \frac{b^a}{\Gamma(a)} \phi^{a-1} \exp\{-b\phi\}$$

Let's go for a simpler task: find posterior mode.

$$\underset{(\phi,\theta)}{\arg\max}\ p(\theta, \phi \mid \underline{x}) = \underset{(\phi,\theta)}{\arg\max}\ \log p(\theta, \phi \mid \underline{x})$$



ACM = Alternate conditional modes.

Diagnostics for MCMC algorithms

- Recall that, although Metropolis-Hastings algorithm work if we run them long enough, the fact that samples are dependent presents challenges:
  - Convergence.
  - Mixing.

Diagnostics for MCMC algorithms

- Mixing: As with every Monte Carlo algorithm an important question is how many samples should be generated in order to have estimates of the parameters with a reasonable level of accuracy.
  - However, because samples are (most usually positively) correlated, standard theory (standard central limit theorems, Chebyshev's inequality, etc.) does not apply here!
  - Mixing is also related to how fast the high probability areas of the posterior distribution are explored.

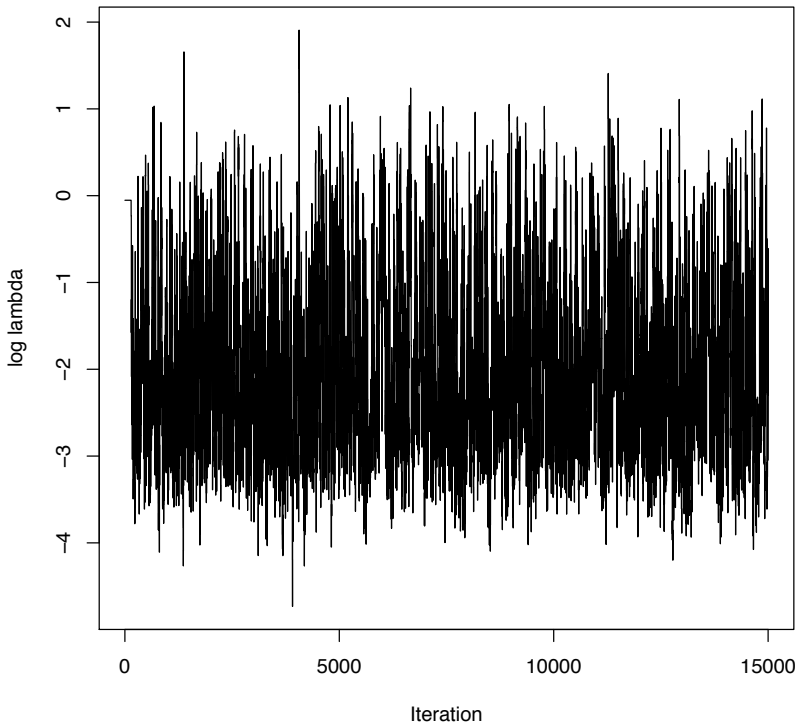
Diagnostics for MCMC algorithms

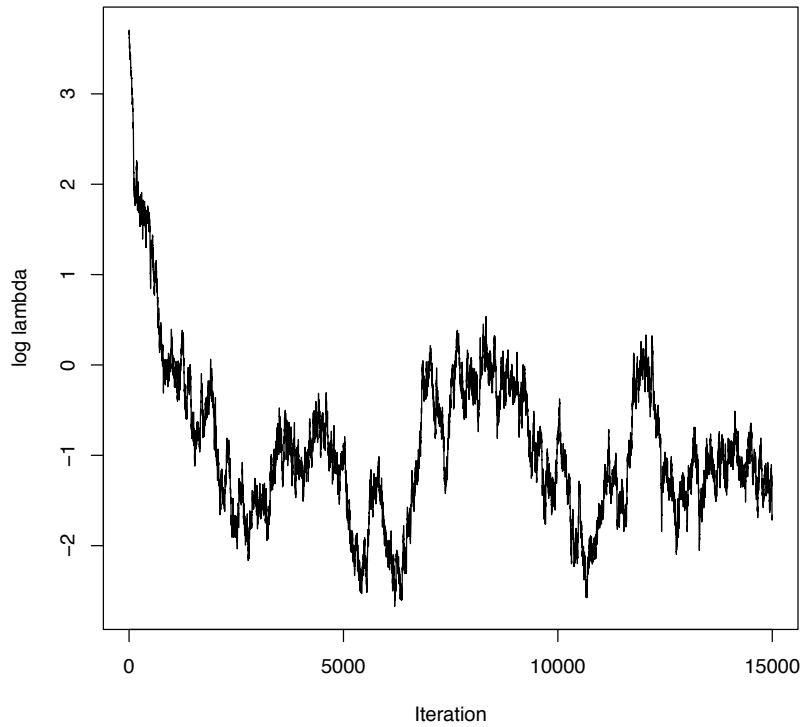
- Convergence: An additional question when using MCMC algorithms that is not a concern in standard Monte Carlo is how many of the initial observations need to be discarded before we can perform any inference.
  - Remember that, unless your initial state is sampled from the stationary distribution, the samples of your chain are only distributed as the stationary distribution when the number of iterations goes to infinity.
  - In practice, if your posterior is unimodal, you just want your initial state to be close to the high-probability areas.
  - When posterior are multimodal convergence can be a huge issue.

Diagnostics for MCMC algorithms

- These two questions are interrelated, but are not identical. For example, you can accelerate convergence to the posterior by carefully selecting the initial state of your chain even if the chain mixes slowly. On the other hand, even if your initial state is really distributed from the stationary distribution (e.g., by using perfect sampling), you might need to run a very long chain to have an acceptable level of uncertainty in your estimates because the autocorrelation in the parameters of interest can be very high.
- Time series plots of the realized chains (the so-called *trace* plots) provide visual intuition about these two concepts, and can be used as an informal check (which should be supplemented with some of the formal techniques discussed later).

Diagnostics for MCMC algorithms





Checking for convergence

- Generally speaking, determining *a priori* how many iterations are needed for burn-in in not possible. Indeed, except for some special cases, general theory about the behavior of the algorithm (e.g., the value of the largest eigenvalue of the kernel function) is not easy to derive.
- For this reason, most practical algorithms to monitor convergence look at the time-series behavior of a small number of selected parameters (or functions of parameters).
  - Selecting which parameters need to be monitored is tricky and model dependent. In high-dimensional models I usually monitor either the likelihood or the posterior, a small (random) subset of the main first-stage parameters, and a couple of important hyperparameters.
  - All criteria here are implemented in the `R` package `CODA`.
- We typically cannot ensure convergence, just argue that there is no evidence of lack of convergence.

Checking for convergence (Multi-chain methods)

- As the name suggests multi-chain methods use $M > 1$ realizations of the algorithm (each of length B, typically started from "over dispersed" values) and compare the output of the chains. Once the statistical properties of the realizations appear to be similar, the chains are deemed to have converged.
- Once example of this type of approach was introduced by Gelman & Rubin (1992) who propose to monitor a statistic that compares the average within-chain variability against the between-chains variability:

$$R = \left(\frac{B-1}{B} + \frac{M+1}{M} \frac{Z_B}{W_B} \right) \frac{\nu_B}{\nu_B - 2},$$

where Z_B is an estimate of between-chain variability, W_B is an estimate of within-chain variability and ν_B is the approximate number of degrees of freedom.

Checking for convergence (Multi-chain methods)

If $\zeta = h(\theta)$ is the quantity of interest being monitored, define

$$\bar{\xi}_m = \frac{1}{B}\sum_{b=1}^{B} \xi_m^{(b)} \qquad \bar{\xi} = \frac{1}{M}\sum_{m=1}^{M} \bar{\xi}_m \qquad s_m^2 = \frac{1}{B}\sum_{b=1}^{B} \left(\xi_m^{(b)} - \bar{\xi}\right)^2$$

Then

$$Z_B = \frac{1}{M}\sum_{m=1}^{M} \left(\bar{\xi}_m - \bar{\xi}\right)^2 \qquad W_B = \frac{1}{M}\sum_{m=1}^{M} s_m^2$$

and

$$s_m^2 = \frac{1}{B}\sum_{b=1}^{B} \left(\xi_m^{(b)} - \bar{\xi}\right)^2 \qquad \nu_B = \frac{2\left(\frac{B-1}{B} W_B + \frac{M+1}{M} Z_B\right)^2}{W_B}$$

Checking for convergence (Multi-chain methods)

- We have $BZ_B/W_B \sim F_{M-1,2W_B^2/\varpi_B}$ approximately, where

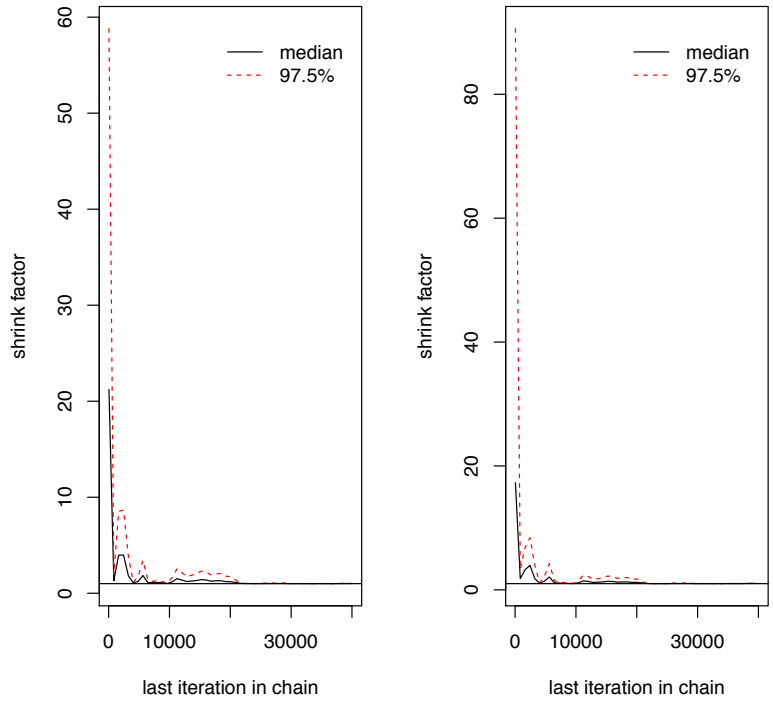
$$\varpi_B = \frac{1}{M^2}\left[\sum_{m=1}^{M} s_m^4 - \frac{1}{M}\left(\sum_{m=1}^{M} s_m^2\right)^2\right]$$

- Once the chains have converged we expect $R \to 1$, so the approximate distribution can be used to test convergence at different values of B.
- Samples from different chains are then combined together for computing any expectation of interest.

Checking for convergence (Multi-chain methods)

```
#...  Run algotithm with the first set of initial
values
chain1.mcmc = as.mcmc(cbind(log(lambda.out),
log(kappa2.out)))


#...  Run algotithm with the second set of
initial values
chain2.mcmc = as.mcmc(cbind(log(lambda.out),
log(kappa2.out)))
x = as.mcmc.list(list(chain1.mcmc, chain2.mcmc))
x.gr = gelman.diag(x, autoburnin=F)
gelman.plot(x)
```



Checking for convergence (Multi-chain methods)

- I tend to prefer multi-chain methods. One advantage is that they can be naively parallelized.
- However, finding over-dispersed initial points can be hard. For example, if the posterior is multimodal and you do not start the chain at least once on each mode you might never find out.
- For slow mixing chains, inferences based on a single long chain might be more accurate than the inferences based on many small chains put together. (However, this is an issue mostly if computational resources are limited.)

Checking for convergence (Single-chain methods)

- An alternative to multi-chain methods is to use a single long chain and see if the properties of the chain have "stabilized".
- More specifically, we can divide the original chain of length B into M sections of approximately equal length and compare their spectral densities.
- In particular Geweke (1992) proposes to compare the first and second half of the chain (i.e., use $M = 2$). The algorithm is implemented in the `R` function `geweke.diag` and `geweke.plot`.

Mixing

- Again, I want to emphasize that mixing is a subtly different problem from convergence.
- Note that a chain that mixes slowly, if started on a low-probability region of the space, might take a long time to converge. However, if we start it in a high probability region convergence is much less of an issue.
- A simple way to decide how long the algorithm must be run *after* the burn-in period is to evaluate the variance of the estimators.

Mixing

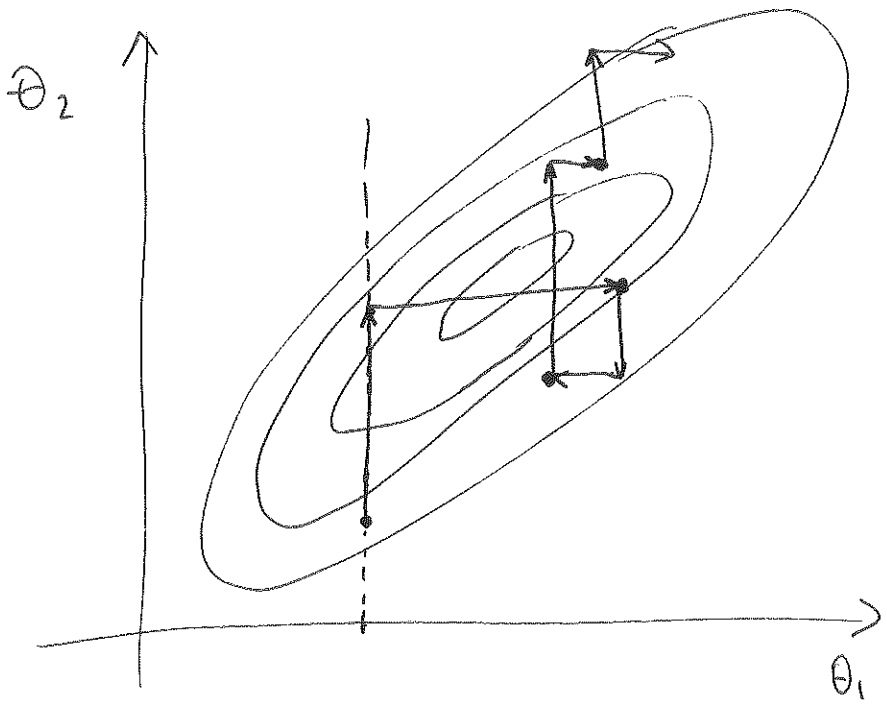
- Let h be an integrable function. Recall that we aim at approximating $\mathrm{E}_{\boldsymbol{\theta}|\mathbf{y}}\left\{h(\boldsymbol{\theta})\right\} \approx \frac{1}{B}\sum_{b=1}^{B} h\left(\boldsymbol{\theta}^{(b)}\right)$
- If the samples $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(B)}$ were independent, the variance of the estimator would be equal to $\frac{\mathrm{Var}\{h(\boldsymbol{\theta})\}}{B}$. In practice a rough estimate of $\mathrm{Var}\left\{h(\boldsymbol{\theta})\right\}$ can also be obtained by Monte Carlo.
- From Markov chain theory, it is easy to show that for an MCMC that has converged the variance of the same estimator is $\tau_h \frac{\mathrm{Var} h(\boldsymbol{\theta})}{B}$ where τ_h is the integrated autocorrelation:

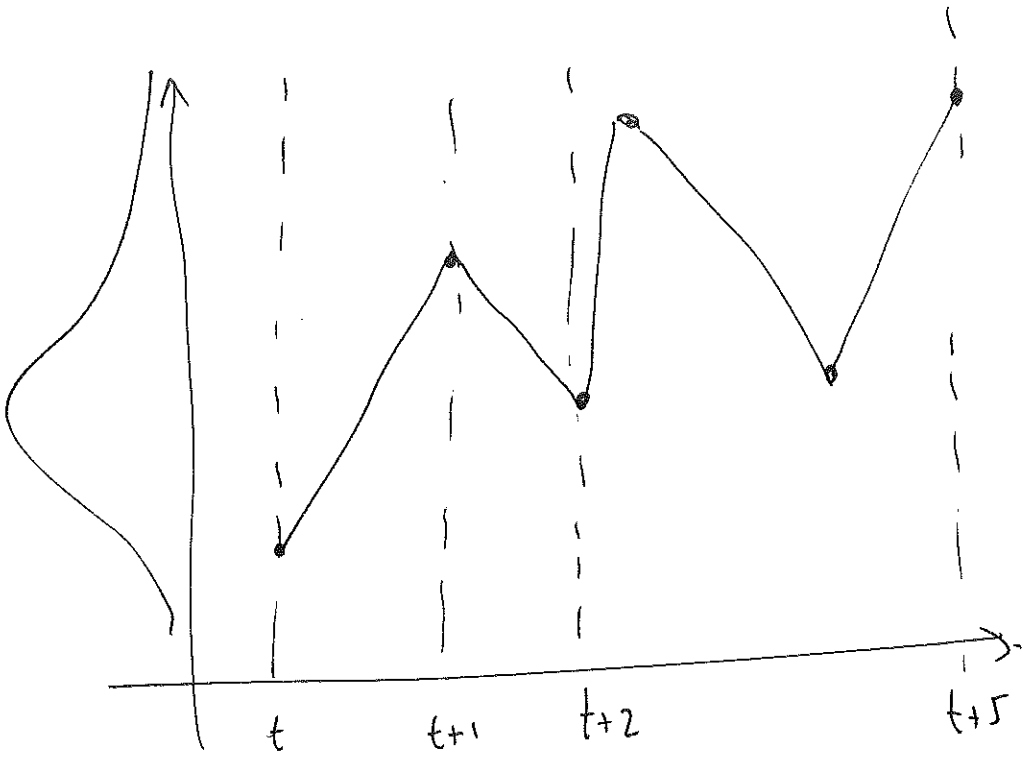
$$\tau_h = 1 + 2\sum_{b=1}^{\infty} \mathrm{Cor}\left\{h\left(\boldsymbol{\theta}^{(1)}\right), h\left(\boldsymbol{\theta}^{(b)}\right)\right\}$$

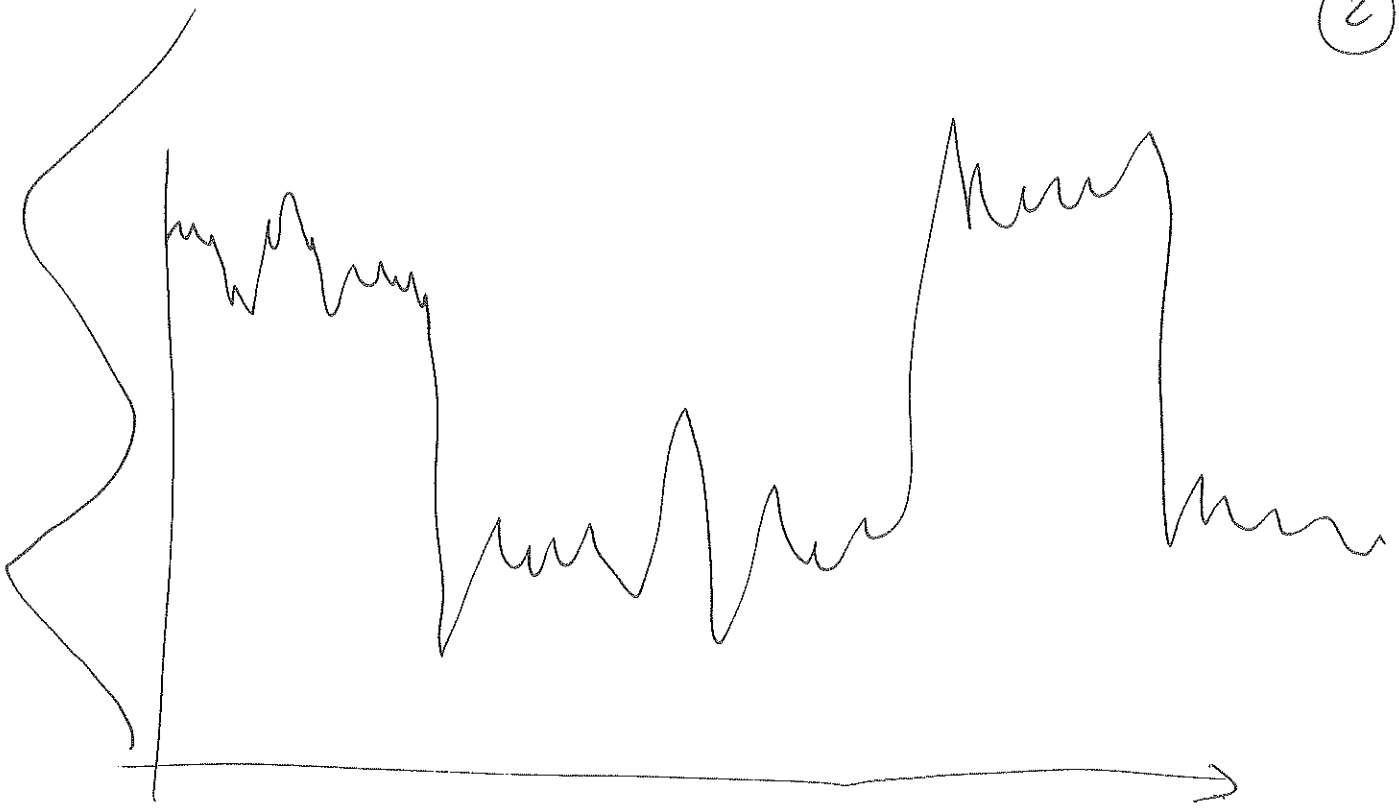
- We call $\frac{B}{\tau_h}$ the equivalent sample size.

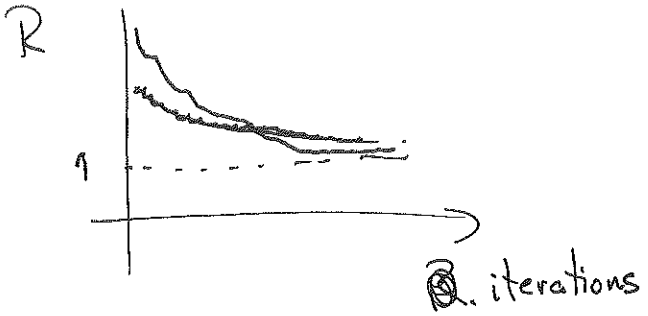
Mixing

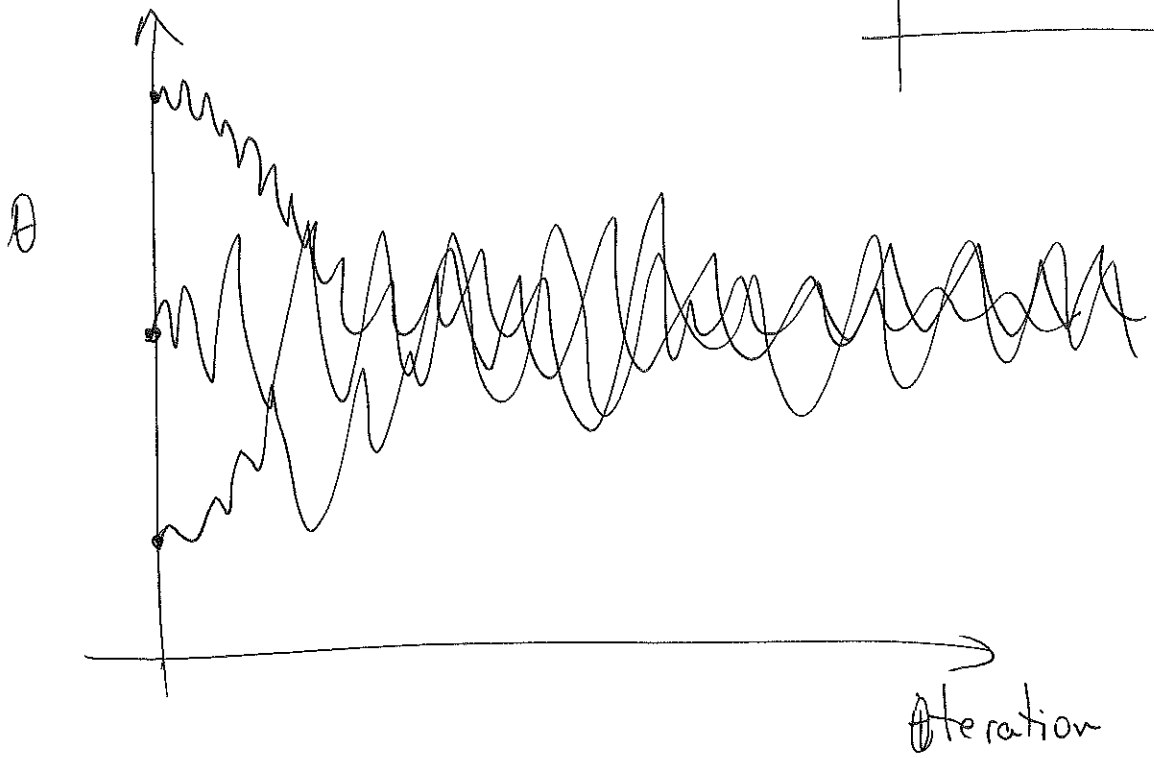
- The equivalent sample size calculation is implemented in `CODA` through the function `effectiveSize`.
- Note that the equivalent sample size will be different for different parameters. That means that a given number of iterations might be enough for one parameter, but not for another.
- Some people like to "thin" their chain (retain every X number of iterations and drop the rest) as a way to decrease τ_g. However, this also reduces B!
  - In general, the tradeoff is not favorable.
  - In my opinion, unless storage is an issue, thinning is not a good idea.


θ_2
θ_1


t
t+1
t+2
t+5




R
1
iterations


θ
Iteration

X, Y be two random variables

$E(X+Y) = E(X) + E(Y)$ no matter whether X, Y are independent or not

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

↳ If $X_1, .. X_n$ independent and identically distributed!

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{n\,Var(X_i)}{n^2} = \frac{Var(X_i)}{n}.$$

Prior selection:

1) Subjective Bayes.

2) Objective Bayes
   - Invariance.
   - "Minimal" information.

3) Robust Bayes.
   - Bounded influence.

$y_1, y_2, \cdots y_n$

$\Downarrow$

data.

μ

ρ

σ^2

$p(\mu, \rho, \sigma^2 | y_1, \cdots y_n)$ $\Leftarrow$ ultimate goal

$\propto p(y_2, \cdots y_n | y_1, \mu, \rho, \sigma^2)\, p(\mu, \rho, \sigma^2)$

$\Downarrow$

$p(\mu)\, p(\rho)\, p(\sigma^2)$

$\swarrow$ $\Downarrow$ $\searrow$

$p(y_t | \mu, \rho, \sigma^2, y_{t-1})$ Gaussian $\quad$ Uni$[-1, 1]$ $\quad$ Inverse Gamma

$$p(y_n, \cdots y_2 | y_1, \mu, \rho, \sigma^2) = p(y_n | y_{n-1}, \cancel{\cdots y_1}, \mu, \rho, \sigma^2) \times$$

$$p(y_{n-1} | y_{n-2}, \cancel{\cdots y_1}, \mu, \rho, \sigma^2) \times$$

$$\cdots\cdots$$

$$p(y_3 | y_2, \cancel{y_1}, \mu, \rho, \sigma^2) \times$$

$$p(y_2 | y_1, \mu, \rho, \sigma^2)$$

$$= \prod_{t=2}^{n} p(y_t | y_{t-1}, \mu, \rho, \sigma^2) = \prod_{t=2}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left\{ -\frac{1}{2\sigma^2} (y_t - \mu - \rho(y_{t-1} - \mu))^2 \right\}$$

$$y_t = \mu + \rho(y_{t-1} - \mu) + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2)$$

$$\Updownarrow$$

$$y_t \mid \mu, \rho, y_{t-1}, \sigma^2 \sim N(\mu + \rho(y_{t-1} - \mu), \sigma^2)$$

Priors:
- Subjective priors
- "Objective" priors
  - Invariance
  - Minimal information
- Robust priors

Invariant priors: Jeffreys' priors

(Consistency across parameterizations)

Expected Fisher information:

Likelihood for your problem is $p(\underline{x}|\theta)$ then

$$I(\theta) = -E_{x|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log p(\underline{x}|\theta)\right]$$

Example: $X \sim Bin(n, \theta)$

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

$$\Rightarrow l = \log p(x|\theta) = \log\binom{n}{x} + x\log\theta + (n-x)\log(1-\theta)$$

$$\frac{\partial l}{\partial\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta} \Rightarrow E_{x|\theta}\left[-\frac{x}{\theta^2} - \frac{(n-x)}{(1-\theta)^2}\right] =$$

$$\frac{\partial^2 l}{\partial\theta^2} = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2} = -\left[\frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2}\right] = \frac{n(1-\theta+\theta)}{\theta(1-\theta)}$$

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

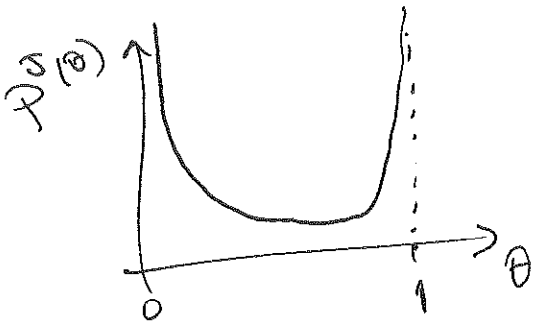
The Jeffreys' prior is proportional to the squared root of the information number

$$p^J(\theta) \propto \sqrt{I(\theta)}$$

For the binomial:

$$p^J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

$\Rightarrow$ The Jeffreys' prior for θ and a Binomial likelihood is a Beta(1/2, 1/2)



$$p^J(\theta) = \frac{\Gamma(1)}{\Gamma(1/2)\Gamma(1/2)}\theta^{-1/2}(1-\theta)^{-1/2}$$

$$\eta = \log\frac{\theta}{1-\theta}$$

What is the Jeffreys' prior for η?

Procedure 1:

Change of variables on $\pi^J(\theta)$.

$$\eta = \log \frac{\theta}{1-\theta} \Rightarrow \exp\{\eta\} = \frac{\theta}{1-\theta} \Rightarrow \exp\{\eta\} - \theta\exp\{\eta\} = \theta$$

$$\Rightarrow \theta = \frac{\exp\{\eta\}}{1+\exp\{\eta\}}$$

$$\Rightarrow \theta = \frac{1}{1+\exp\{-\eta\}} = (1+\exp\{-\eta\})^{-1}$$

$$d\theta = (+1)(1+\exp\{-\eta\})^{-2}\exp\{-\eta\}$$

$$= \frac{\exp\{-\eta\}}{(1+\exp\{-\eta\})^2}$$

$$\pi^J(\eta) = \frac{1}{[\Gamma(1/2)]^2}\left[\frac{1}{1+\exp\{-\eta\}}\right]^{-1/2}\left[\frac{\exp\{-\eta\}}{1+\exp\{-\eta\}}\right]^{-1/2}\frac{\exp\{-\eta\}}{[1+\exp\{-\eta\}]^2}$$

$$= \frac{1}{[\Gamma(1/2)]^2}\frac{\exp\{-\frac{1}{2}\eta\}}{1+\exp\{-\eta\}}$$

Homework: check that procedure 2 gives the same prior

$$p(x|\eta) = \binom{n}{x}\left(\frac{1}{1+\exp\{-\eta\}}\right)^x\left(\frac{\exp\{-\eta\}}{1+\exp\{-\eta\}}\right)^{n-x}$$

$$\pi^J(\eta) \propto \left|-E_{x|\eta}\left[\frac{\partial^2}{\partial\eta^2}\log p(x|\eta)\right]\right|^{1/2}$$

Consider now $X \sim \text{Neg Bin}(K, \theta)$ ($K \to$ number of successes; $X \to$ # of trials)

$$p(x|\theta) = \binom{x-1}{k-1} \theta^K (1-\theta)^{x-K}$$

$$l = \log p(x|\theta) = \log \binom{x-1}{k-1} + K \log \theta + (x-K)\log(1-\theta)$$

$$\frac{\partial l}{\partial \theta} = \frac{K}{\theta} - \frac{x-K}{1-\theta}$$

$$\frac{\partial^2 l}{\partial \theta^2} = -\frac{K}{\theta^2} - \frac{(x-K)}{(1-\theta)^2}$$

$$E_{x|\theta}\left(\frac{\partial^2 l}{\partial \theta^2}\right) = -E_{x|\theta}\left[\frac{K}{\theta^2} + \frac{(x-K)}{(1-\theta)^2}\right] = -\left[\frac{K}{\theta^2} + \frac{E(x)-K}{(1-\theta)^2}\right]$$

$$E(x) = \frac{K}{\theta} \Rightarrow I(\theta) = \frac{K}{\theta^2} + \frac{K\left(\frac{1}{\theta}-1\right)}{(1-\theta)^2}$$

$$= \frac{K\left((1-\theta)^2 + \theta^2\left(\frac{1}{\theta}-1\right)\right)}{\theta^2(1-\theta)^2} = K\,\frac{1-2\theta+\theta^2+\theta-\theta^2}{\theta^2(1-\theta)^2}$$

$$= K\,\frac{(1-\theta)}{\theta^2(1-\theta)^2} = K\,\frac{1}{\theta^2(1-\theta)}$$

$$p^J(\theta) \propto \frac{1}{\theta(1-\theta)^{1/2}} = \theta^{-1}(1-\theta)^{-1/2}$$

↳ Improper prior!! (Does not integrate to anything finite!!! Not a density!!!

Even though the prior is improper;

$$p(\theta|x) \propto \theta^{k}(1-\theta)^{x-k}\ \theta^{-1}(1-\theta)^{-1/2}$$

$$\Rightarrow p(\theta|x) \propto \theta^{k-1}(1-\theta)^{x-k-1/2}$$

↳ Kernel of a Beta$(k, x-k+1/2)$

General result:

* For location parameters the invariant prior is uniform.
* For scale parameters the invariant prior is uniform on the log scale.

Example: $X_1, \dots X_n \sim N(\mu, \sigma^2)$

* If μ unknown but σ^2 known.

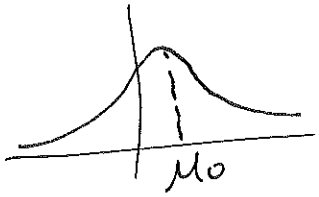
$$p^J(\mu) \propto 1$$

* If σ^2 is unknown but μ is known

$$p^J(\sigma) \propto \frac{1}{\sigma}.$$

For many models in the exponential family you can get the Jeffreys' prior as a limit of conjugate priors

$$p(\mu|\mu_0, \tau^2) = \frac{1}{\sqrt{2\pi}\tau} \exp\left\{-\frac{1}{2\tau^2}(\mu-\mu_0)^2\right\}$$



Objective Bayes: Jeffreys prior

- For location families where $p(y \mid \theta) = f(y - \theta)$ for some density f, the Jeffreys prior is simply

$$\pi^{J,L}(\theta) \propto 1.$$

- For scale families where $p(y \mid \theta) = \frac{1}{\theta} f\left(\frac{y}{\theta}\right)$ for some density f, the Jeffreys prior is

$$\pi^{J,S}(\theta) \propto \frac{1}{\theta}.$$

  This is equivalent to placing a uniform prior on the log scale.

- In both cases the Jeffreys prior is improper, and we need to carefully check that the corresponding posterior is proper!

Objective Bayes: Jeffreys prior

- The Jeffreys priors for location and scale families can be obtained as the limit of proper priors:
  - For the location family, consider a Gaussian prior

$$\pi^L(\theta) = \frac{1}{\sqrt{2\pi}\tau} \exp\left\{-\frac{1}{2\tau^2}(\theta-\mu)^2\right\}$$

    Note that $\lim_{\tau\to\infty} \pi^L(\theta) = \pi^{J,L}(\theta)$.

  - For the scale family, consider a (slightly reparameterized) inverse Gamma family

$$\pi^S(\theta) = \frac{(\alpha\beta)^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)} \exp\left\{-\frac{\alpha\beta}{\theta}\right\}$$

    where $E\{\theta\} = \beta$. Here $\lim_{\alpha\to 0} \pi^S(\theta) = \pi^{J,S}(\theta)$.

Objective Bayes: Jeffreys prior

- **Example (Normal distribution with unknown mean and variance):** Let $y \mid \mu, \phi \sim \mathrm{N}_p(\mu, \phi)$, so that $\phi = \sigma^2$ is the variance. The Hessian in this case is

$$\begin{vmatrix} -\frac{1}{\phi} & -\frac{1}{\phi^2}\sum_{i=1}^n (y_i - \mu) \\ \frac{1}{\phi^2}\sum_{i=1}^n (y_i - \mu) & \frac{n}{2\phi^2} - \frac{1}{\phi^3}\sum_{i=1}^n (y_i - \mu)^2 \end{vmatrix}.$$

Since $\mathrm{E}\left\{\sum_{i=1}^n (y_i - \mu)\right\} = 0$ and $\mathrm{E}\left\{\sum_{i=1}^n (y_i - \mu)^2\right\} = n\phi$,

$$|\mathcal{I}(\mu, \phi)| = \frac{1}{\phi^3}$$

and

$$\pi^J(\mu, \phi) \propto \phi^{-3/2}.$$

Objective Bayes: Jeffreys prior

- Multivariate Jeffreys priors sometimes have undesirable behaviors.
- An alternative is the so-called independence Jeffreys prior (originally proposed by Jeffreys himself) is very common.
- The independence Jeffreys $\pi^{IJ}(\mu, \phi)$ assumes prior independence and uses the corresponding univariate Jeffreys priors, $\pi^{J}(\mu, \phi) = \pi^{J}(\mu)\pi^{J}(\phi)$.
- In the case of the normal model with unknown mean and variance this reduces to

$$\pi^{IJ}(\mu, \phi) \propto \frac{1}{\phi}.$$

- Independence Jeffreys priors are not invariant to transformations that involve more than one parameter.

Objective Bayes: Reference priors

- Another (similar) solution is to construct the prior sequentially.
- Let $y \mid \theta \sim p(y \mid \theta_1, \theta_2)$ where θ_1 is the parameter of interest and θ_2 is the nuisance parameter.
  - Use $p(y \mid \theta_1, \theta_2)$ to construct the Jeffreys prior for θ_2 assuming a fixed value of θ_1 to obtain $\pi^J(\theta_2 \mid \theta_1)$.
  - Compute $\tilde{p}(y \mid \theta_1) = \int p(y \mid \theta_1, \theta_2)\pi^J(\theta_2 \mid \theta_1)\mathrm{d}\theta_2$.
  - Construct the Jeffreys prior for θ_1 using $\tilde{p}(y \mid \theta_1)$.
- The order of the parameters matters, and you might get a different prior depending on what is the main parameter of interest.
- Can be generalized to more than two groups of parameters.
- Can be justified as maximizing the posterior information.

Bayesian hypothesis testing

- Improper priors such as those that often arise from non-informative approaches are often appropriate for both point and interval estimation.
- However, their use in problems that involve models of different dimensions (as is the case in hypothesis testing) is extremely problematic.
- To illustrate this, let's consider Barlett's paradox.

Barlett's paradox

- **Example:** Let $y_1, \ldots, y_n$ be such that $y_i \mid \theta \sim \text{N}(\theta, 1)$ and consider testing the hypotheses $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ under the prior $\theta \mid H_1 \sim \text{N}(0, \tau^2)$.

  The corresponding Bayes factor (which we already calculated in our introductory example) is

$$B_{10}(\tau^2, \bar{y}) = (1 + n\tau^2)^{-1/2} \exp\left\{ \frac{1}{2} \frac{n^2 \bar{y}^2}{n + \frac{1}{\tau^2}} \right\},$$

  and it is easy to verify that $\lim_{\tau^2 \to \infty} B_{10}(\tau^2, \bar{y}) = 0$ and therefore $\lim_{\tau^2 \to \infty} p(H_0 \mid y_1, \ldots, y_n) = 1$ for any values of n and $\bar{y}$.

Barlett's paradox

- This appears to be paradoxical because we would expect the Bayes factor to prefer H_1 when $\bar{y}$ is large (and indeed $\lim_{\bar{y}\to\infty} B_{10}(\tau^2, \bar{y}) = \infty$). Also, we had shown that Bayesian model selection is consistent as $n \to \infty$.
- In conclusion, no matter how strong the information we have in our sample, the Bayes factor constructed using a flat prior will always favor the null model.
  - The same phenomenon appears in most comparisons that involve testing point vs. composite hypotheses using improper priors on the parameter being tested.
- More generally, this result highlights that priors can have a big impact for model selection, even for large sample sizes.

Barlett's paradox

- The key to understand Barlett's paradox is to notice that the order in which you take the limits matter! If you first take $\tau^2 \to \infty$ first, the Bayes factor is ill behaved.
- Another way to think about Barlett's paradox is that it arises because improper priors are defined up to an arbitrary normalizing constant, the Bayes factor depends on such constant and their value is therefore undefined.
  - This does not happen in estimation problems because the arbitrary constant appears both in the numerator and the denominator of the posterior distribution, canceling out.
- Another way to think about it is that the improper prior essentially puts probability 0 on H_1 a priori.
- Barlett's paradox does not arise when testing composite vs. composite hypotheses where the same improper prior is used under both hypotheses

Barlett's paradox

- **Example:** Again, let $y_1, \ldots, y_n$ be such that $y_i \mid \theta \sim \text{N}(\theta, 1)$ and consider testing the hypotheses $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ under the prior $\theta \sim \text{N}(0, \tau^2)$. The posterior odds in this case reduce to

$$\frac{p(H_0 \mid y_1, \ldots, y_n)}{p(H_1 \mid y_1, \ldots, y_n)} = \frac{\int_{-\infty}^{0} \exp\left\{ -\frac{1}{2}\left(n + \frac{1}{\tau^2}\right)\left(\theta - \frac{n\bar{y} + \frac{\mu}{\tau^2}}{n + \frac{1}{\tau^2}}\right)^2 \right\}}{\int_{0}^{\infty} \exp\left\{ -\frac{1}{2}\left(n + \frac{1}{\tau^2}\right)\left(\theta - \frac{n\bar{y} + \frac{\mu}{\tau^2}}{n + \frac{1}{\tau^2}}\right)^2 \right\}}.$$

so that

$$0 < \lim_{\tau^2 \to \infty} \frac{p(H_0 \mid y_1, \ldots, y_n)}{p(H_1 \mid y_1, \ldots, y_n)} < \infty$$

for any finite values of $\bar{y}$ and n. Taking a limit is not a problem here!

Barlett's paradox

- **Example (cont):** Intuitively, the improper prior is not a problem here because the (undefined!) normalizing constant of the prior appears in both the numerator and the denominator, so it cancels out!

  More generally, traditional improper priors used in estimation will work well for testing composite vs. composite hypotheses, but NOT for testing point vs. composite hypotheses. Different ("well calibrated") default priors are required in that setting.

Beyond conjugate models

- When you do not have conjugacy the process is the same, but you often cannot make much progress.
- There are a few exceptions. A well known one is the use of double exponential priors for the location parameter of a Gaussian likelihood,

$$p(y_1, \ldots, y_n \mid \theta) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta)^2\right\},$$

$$\pi(\theta) = \frac{1}{2\tau}\exp\left\{-\frac{1}{\tau}\left|\theta - \mu\right|\right\}.$$

Beyond conjugate models

- The posterior is in this case:

$$p(\theta \mid y_1, \ldots, y_n) = \frac{a(n, \bar{y})}{a(n, \bar{y}) + b(n, \bar{y})} \text{TN}\left(\lambda_1(n, \bar{y}), \frac{1}{n}, -\infty, \mu\right) \mathbb{I}(\theta < \mu) + \frac{b(n, \bar{y})}{a(n, \bar{y}) + b(n, \bar{y})} \text{TN}\left(\lambda_2(n, \bar{y}), \frac{1}{n}, \mu, \infty\right) \mathbb{I}(\theta \geq \mu)$$

where

$$\lambda_1(n, \bar{y}) = \bar{y} + \frac{1}{n\tau} \qquad \lambda_2(n, \bar{y}) = \bar{y} - \frac{1}{n\tau}$$

and

$$a(n, \bar{y}) = \exp\left\{-\frac{\mu}{\tau} + \frac{n}{2}\lambda_1^2(n, \bar{y})\right\} \Phi\left(\{\mu - \lambda_1(n, \bar{y})\}\sqrt{n}\right)$$

$$b(n, \bar{y}) = \exp\left\{\frac{\mu}{\tau} + \frac{n}{2}\lambda_2^2(n, \bar{y})\right\} \left\{1 - \Phi\left(\{\mu - \lambda_2(n, \bar{y})\}\sqrt{n}\right)\right\}$$

Beyond conjugate models

- The posterior mean is

$$\frac{a(n,\bar{y})}{a(n,\bar{y})+b(n,\bar{y})}\left[\lambda_1(n,\bar{y})-\frac{1}{\sqrt{n}}\frac{\phi\left(\sqrt{n}\{\mu-\lambda_1(n,\bar{y})\}\right)}{\Phi\left(\sqrt{n}\{\mu-\lambda_1(n,\bar{y})\}\right)}\right]+$$
$$\frac{b(n,\bar{y})}{a(n,\bar{y})+b(n,\bar{y})}\left[\lambda_2(n,\bar{y})+\frac{1}{\sqrt{n}}\frac{\phi\left(\sqrt{n}\{\mu-\lambda_2(n,\bar{y})\}\right)}{1-\Phi\left(\sqrt{n}\{\mu-\lambda_2(n,\bar{y})\}\right)}\right],$$

  where ϕ and Φ are the density and cdf of the standard normal distribution. Note that this is highly non-linear!

- One key feature is that the prior has bounded influence! $\Rightarrow$ Posterior distribution is robust to misspecification of the prior.

$$\lim_{\mu\to\infty} = \bar{y}+\frac{1}{n\tau} \qquad \lim_{\mu\to-\infty} = \bar{y}-\frac{1}{n\tau}$$

  (Note the difference with a Gaussian prior.)

Bayesian Assymptotics

- If θ is continuous and the the observations are iid, then under mild regularity conditions the posterior distribution is approximately normal

$$\theta \mid y \sim \mathrm{N}\left(\hat{\theta}(y), H^{-1}(\hat{\theta}(y))\right),$$

  where $\hat{\theta}(y)$ represents the maximum likelihood estimator of θ and $H(\hat{\theta}(y))$ is the observed information matrix.
- Sometimes called the "Bayesian CLT".
- Other variants are possible, all of them are based on Laplace approximations to the posterior

Bayesian Assymptotics - Laplace expansions

- Consider an iid univariate sequence $y_1, \ldots, y_n$ where $y_i \mid \theta \sim p(y \mid \theta)$ and do a Taylor expansion of the log-likelihood:

$$\log\{\ell(\theta)\} \approx \log \ell\left(\hat{\theta}\right) + \frac{1}{2} \frac{\partial^2}{\partial \theta^2} \log \ell\left(\hat{\theta}\right)\Big|_{\theta=\hat{\theta}} \left(\theta - \hat{\theta}\right)^2$$

  where $\frac{\partial^2}{\partial \theta^2} \log \ell\left(\hat{\theta}\right)\Big|_{\theta=\hat{\theta}}$ is the observed information matrix.

- For the prior, you have two options:
  - You can either do a similar Taylor expansion of the prior.
  - Under regularity conditions, the likelihood concentrates as n grows, so the prior can be discarded.

Bayesian Assymptotics - Laplace expansions

- This result has a few important consequences:
  - For large sample sizes, Bayesian and frequentist point and interval estimation procedures yield essentially the same results! (This is not true for hypothesis testing!)
  - For large sample sizes, the form of the prior does not matter much! (Again, this is not true for hypothesis testing!)
  - Most "reasonable" Bayesian point estimators are consistent (in a classical, frequentist sense).
  - For large sample sizes, the approximation can be used to easily construct point and interval estimates.
- The quality of the approximation might depend on the parameterization.

The Metropolis-Hastings algorithm

- We move on now to discuss specific MCMC algorithms used in Bayesian computation. We start with the Metropolis-Hastings (MH) algorithm, which encompasses many others as special cases.
- The idea behind the MH algorithm is similar to that behind the rejection algorithm:
  - Pick a proposal distribution $q(\vartheta \mid \theta)$ (think of it as *almost* the transition kernel of your Markov chain) that will be used to generate potentially new states.
  - The chain either stays on the old value or moves to this new proposed values according to a certain probability, that is chosen to ensure that the chain is reversible!

The Metropolis-Hastings algorithm

Given $\theta^{(t)}$

1. Generate $\vartheta^{(t+1)} \sim q\left(\vartheta \mid \theta^{(t)}\right)$.
2. Take

$$\theta^{(t+1)} = \begin{cases} \vartheta^{(t+1)} & \text{with probability } \rho\left(\vartheta^{(t+1)}, \theta^{(t)}\right) \\ \theta^{(t)} & \text{with probability } 1 - \rho\left(\vartheta^{(t+1)}, \theta^{(t)}\right) \end{cases},$$

where

$$\rho\left(\vartheta^{(t+1)}, \theta^{(t)}\right) = \min\left\{1, \frac{p\left(\vartheta^{(t+1)} \mid \mathbf{y}\right)}{p\left(\theta^{(t)} \mid \mathbf{y}\right)} \frac{q\left(\theta^{(t)} \mid \vartheta^{(t+1)}\right)}{q\left(\vartheta^{(t+1)} \mid \theta^{(t)}\right)}\right\}.$$

The Metropolis-Hastings algorithm

- $\rho\left(\vartheta^{(t+1)}, \theta^{(t)}\right)$ is called the Metropolis-Hastings acceptance probability.
- Note that the algorithm depends on the ratio $\frac{p\left(\vartheta^{(t+1)}|\mathbf{y}\right)}{p\left(\theta^{(t)}|\mathbf{y}\right)}$, so it works even if $p\left(\theta \mid \mathbf{y}\right)$ is known only up to a normalizing constant.
- There are a number of variants of the algorithm!

The Metropolis-Hastings algorithm

In the random-walk Metropolis-Hastings, given $\theta^{(t)}$:

1. Generate $\vartheta^{(t+1)} \sim g\left(\left|\vartheta - \theta^{(t)}\right|\right)$, where g is a density.
2. Take

$$\theta^{(t+1)} = \begin{cases} \vartheta^{(t+1)} & \text{with probability } \rho\left(\vartheta^{(t+1)}, \theta^{(t)}\right) \\ \theta^{(t)} & \text{with probability } 1 - \rho\left(\vartheta^{(t+1)}, \theta^{(t)}\right) \end{cases},$$

where

$$\rho\left(\vartheta^{(t+1)}, \theta^{(t)}\right) = \min\left\{1, \frac{p\left(\vartheta^{(t+1)} \mid \mathbf{y}\right)}{p\left(\theta^{(t)} \mid \mathbf{y}\right)}\right\}.$$

Common choices for g include zero-mean Gaussian (my favorite) or uniform distributions!

The Metropolis-Hastings algorithm

- **Example (Gumble likelihood):** Consider a setting where $y_1, \ldots, y_n$ corresponds to a random sample from a Gumble distribution with location parameter θ, i.e.,

$$p(y_i \mid \theta) = \exp\{-(y_i - \theta) - \exp\{-(y_i - \theta)\}\} \qquad y_i \in \mathbb{R}$$

and assume that we let $\theta \sim \mathrm{N}(\xi, \kappa^2)$ a priori.

The posterior distribution associated with this model is intractable. However, creating a RWMH algorithm to sample from it is straightforward. Because θ can in principle take any real value, a Gaussian random walk seems appropriate,

$$q(\vartheta \mid \theta) = \frac{1}{\sqrt{2\pi}\tau} \exp\left\{-\frac{1}{2\tau^2}(\vartheta - \theta)^2\right\}$$

The Metropolis-Hastings algorithm

- **Example (Gumble likelihood, cont):** The corresponding acceptance probability becomes

$$\rho(\vartheta, \theta) = \min\left\{1, \frac{\exp\left\{-\sum_{i=1}^{n}(y_i - \vartheta) - \sum_{i=1}^{n}\exp\{-(y_i - \vartheta)\}\right\}}{\exp\left\{-\sum_{i=1}^{n}(y_i - \theta) - \sum_{i=1}^{n}\exp\{-(y_i - \theta)\}\right\}} \frac{\exp\left\{-\frac{(\vartheta-\xi)^2}{2\kappa^2}\right\}}{\exp\left\{-\frac{(\theta-\xi)^2}{2\kappa^2}\right\}}\right\}$$

(Note that, as we discussed before, the ratio of the proposals cancels out because of the symmetry!)

This algorithm is implemented in the file `gumble.R`. Compare the results using $\tau^2 = 0.001$ (too small!), $\tau^2 = 0.07$ (about right, roughly 40% acceptance rate), and $\tau^2 = 5$ (too large!)

The Metropolis-Hastings algorithm

- **Example (Gaussian process):** Consider now a model where a vector $\mathbf{y} = (y_1, \ldots, y_n)$ follows a joint multivariate distribution with mean 0 and variance-covariance matrix $\mathbf{\Sigma}$ with

$$[\mathbf{\Sigma}(\lambda, \kappa)]_{i,j} = \kappa \exp\left\{-\frac{|x_i - x_j|}{\lambda}\right\}, \qquad \lambda, \kappa > 0.$$

  In the previous expression, λ and κ are unknown *positive* parameters, and the values $x_1, \ldots, x_n$ associated with each of the entries of $\mathbf{y}$ are assumed to be known.

This problem is more complicated than the previous one in two ways: Now $\boldsymbol{\theta} = (\lambda, \kappa)$ is two-dimensional, and the parameters are restricted to be positive!

The Metropolis-Hastings algorithm

- **Example (Gaussian process, cont):** Because the parameters must be positive, we use Gamma priors for both:

$$p(\lambda) = \frac{b_1^{a_1} \lambda^{a_1 - 1}}{\Gamma(a_1)} \exp\{-b_1 \lambda\} \quad p(\kappa) = \frac{b_2^{a_2} \kappa^{a_2 - 1}}{\Gamma(a_2)} \exp\{-b_2 \kappa\}$$

For the same reason a Gaussian random walk directly on (λ, κ) is not a great idea (you would be automatically rejecting every time you generate negative values).

However, a (bivariate!) Gaussian random walk for $(\log \lambda, \log \kappa)$ is feasible and does not lead to automatic rejections.

(Another alternative is to use a reflecting Gaussian random walk, but I will not pursue that idea further.)

The Metropolis-Hastings algorithm

- **Example (Gaussian process, cont):** Since we work with a transformation of the parameters, there are two ways to derive the algorithm (they lead to different formal expressions, but they are equivalent!):
  1. Do a transformation so that the problem is reformulated as sampling from $p(z_1, z_2 \mid \mathbf{y})$ where $z_1 = \log \lambda$ and $z_2 = \log \kappa$ instead of $p(\lambda, \kappa \mid \mathbf{y})$. Once samples $z_1^{(1)}, \ldots, z_1^{(B)}$ and $z_2^{(1)}, \ldots, z_2^{(B)}$ have been generated, samples $\lambda^{(1)}, \ldots, \lambda^{(B)}$ and $\kappa^{(1)}, \ldots, \kappa^{(B)}$ can be constructed by letting $\lambda^{(b)} = \exp\{z_1^{(b)}\}$ and $\kappa^{(b)} = \exp\{z_2^{(b)}\}$
  2. Keep the original target $p(\lambda, \kappa \mid \mathbf{y})$ but think of your proposal as being a bivariate log-Gaussian distribution rather than a Gaussian (and recognize that your proposal is almost symmetric, but not quite!).

  Note that, in both cases, a Jacobian is going to be involved.

The Metropolis-Hastings algorithm

- **Example (Gaussian process, cont):** Taking the second route, the proposal distribution is

$$p(\vartheta_1, \vartheta_2 \mid \lambda, \kappa) = \frac{1}{2\pi} \frac{1}{\vartheta_1 \vartheta_2} |\mathbf{\Omega}|^{-1/2} \exp\left\{ -\frac{1}{2} \begin{pmatrix} \log \vartheta_1 - \log \lambda \\ \log \vartheta_2 - \log \kappa \end{pmatrix}^T \mathbf{\Omega}^{-1} \begin{pmatrix} \log \vartheta_1 - \log \lambda \\ \log \vartheta_2 - \log \kappa \end{pmatrix} \right\}$$

Note that $\mathbf{\Omega}$ does not have to be diagonal, so the moves could (and, turns out, should!) be correlated.

The Metropolis-Hastings algorithm

- **Example (Gaussian process, cont):** The corresponding acceptance probability for the algorithm becomes

$$\rho(\vartheta_1, \vartheta_2, \lambda, \kappa) = \min\left\{1, \frac{|\mathbf{\Sigma}(\vartheta_1, \vartheta_2)|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{y}^T \left[\mathbf{\Sigma}(\vartheta_1, \vartheta_2)\right]^{-1} \mathbf{y}\right\}}{|\mathbf{\Sigma}(\lambda, \kappa)|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{y}^T \left[\mathbf{\Sigma}(\lambda, \kappa)\right]^{-1} \mathbf{y}\right\}} \frac{\vartheta_1^{a_1} \exp\{b_1\vartheta_1\} \vartheta_2^{a_2} \exp\{b_2\vartheta_2\}}{\lambda^{a_1} \exp\{b_1\lambda\} \kappa^{a_2} \exp\{b_2\kappa\}} \frac{\vartheta_1\vartheta_2}{\lambda\kappa}\right\}$$

(Note that the symmetric part of the proposal cancels out but the Jacobians are not symmetric, and they remain ...)

This algorithm is implemented in the file `GP-MH.R`. To tune the variance, run the algorithm once with $\mathbf{\Omega} = \text{diag}\{0.3, 0.3\}$, compute $\text{Var}\left\{(\log\lambda, \log\kappa)^T \mid \mathbf{y}\right\}$, and then rerun it with $\mathbf{\Omega} = \frac{2.38^2}{d}\text{Var}\left\{(\log\lambda, \log\kappa)^T \mid \mathbf{y}\right\}$. (In this case $d = 2$.)

The Metropolis-Hastings algorithm

Suggested exercise: Show that both approaches to constructing the random-walk Metropolis for the Gaussian process example lead to the same algorithm!

The Metropolis-Hastings algorithm

- The magic numbers $\frac{2.38^2}{d}\text{Var}(\boldsymbol{\theta}|\mathbf{y})$ for proposals and 44% acceptance rates for unidimensional problems and 23% for multivariate problems comes from Gelman et al. (1996) Roberts et al. (1997) and Roberts et al. (2001), which derived general theoretical results and performed experiments on a Gaussian.
- Since, with enough data, most posteriors are approximately Gaussian, these are reasonable rules of thumb!
- Note that, because we are making two independent runs, the fact that we change the variance of the chain is not (conceptually) an issue as it does not invalidate the Markov property.
- If a good approximation for $\text{Var}(\boldsymbol{\theta}|\mathbf{y})$ is available (rarely the case) then the preliminary run can be skipped.

Multivariate normal model

Multivariate normal distribution.

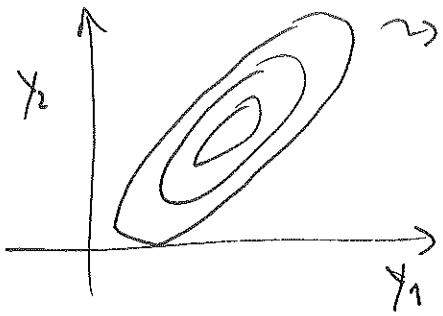
$\underset{\sim}{y} = (y_1, \dots, y_p)^T \qquad \underset{\sim}{y} \sim N_p(\underset{\sim}{\theta}, \Sigma) \qquad \underset{\sim}{\theta} = (\theta_1, \dots \theta_p)^T$

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

if

$$p(\underset{\sim}{y} \mid \underset{\sim}{\theta}, \Sigma) = \left(\frac{1}{\sqrt{2\pi}}\right)^p |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\underset{\sim}{y} - \underset{\sim}{\theta})^T \Sigma^{-1} (\underset{\sim}{y} - \underset{\sim}{\theta})\right\} \qquad \Sigma = \Sigma^T$$

For $p = 2$



$\rightsquigarrow$ contour plots for $p(\underset{\sim}{y} \mid \underset{\sim}{\theta}, \Sigma)$

How do we learn the parameter $\underset{\sim}{\theta}$ and Σ from a sample $\underset{\sim}{y}_1, \dots \underset{\sim}{y}_n$.

First, assume Σ is known but $\underset{\sim}{\theta}$ unknown.

$$p(\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^p |D|^{-1/2} \exp\left\{-\frac{1}{2}(\underset{\sim}{\theta} - \underset{\sim}{d}) D^{-1} (\underset{\sim}{\theta} - \underset{\sim}{d})\right\}$$ is a conjugate prior

$\theta \sim N_p(d, D)$

$$p(\theta \mid \underline{y}_1, \dots \underline{y}_n) \propto \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}}\right)^p |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\underline{y}_i - \underline{\theta})^T \Sigma^{-1} (\underline{y}_i - \underline{\theta})\right\} \quad (2) \times$$

$$\left(\frac{1}{\sqrt{2\pi}}\right)^p |D|^{-1/2} \exp\left\{-\frac{1}{2}(\underline{\theta} - \underline{d})^T D^{-1} (\underline{\theta} - \underline{d})\right\}$$

$$\propto \exp\left\{-\frac{1}{2}(\underline{\theta} - \underline{d})^T D^{-1} (\underline{\theta} - \underline{d}) - \frac{1}{2}\sum_{i=1}^{n} (\underline{y}_i - \underline{\theta})^T \Sigma^{-1} (\underline{y}_i - \underline{\theta})\right\}$$

$$= \exp\left\{-\frac{1}{2}\left[\underline{\theta}^T D^{-1}\underline{\theta} - \overbrace{\underline{\theta}^T D^{-1}\underline{d}}^{\theta^T D^{-1} d} - \overbrace{\underline{d}^T D^{-1}\underline{\theta}}^{d^T D^{-1}\theta} + \underline{d}^T D^{-1}\underline{d} + \sum_{i=1}^{n} \underline{y}_i^T \Sigma^{-1} \underline{y}_i \right.\right.$$
$$\left.\left. - \underline{\theta}^T \Sigma^{-1} \sum_{i=1}^{n} \underline{y}_i - \left(\sum_{i=1}^{n} \underline{y}_i^T\right) \Sigma^{-1} \underline{\theta} + n\underline{\theta}^T \Sigma^{-1} \underline{\theta}\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\underline{\theta}^T D^{-1}\underline{\theta} + n\underline{\theta}^T \Sigma^{-1}\underline{\theta} - 2\underline{\theta}^T D^{-1}\underline{d} - 2\underline{\theta}^T \Sigma^{-1} n \bar{\underline{y}}\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\underline{\theta}^T (D^{-1} + n\Sigma^{-1})\underline{\theta} - 2\underline{\theta}^T \left[D^{-1}\underline{d} + n\Sigma^{-1}\bar{\underline{y}}\right]\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\underline{\theta} - (D^{-1} + n\Sigma^{-1})^{-1}(D^{-1}\underline{d} + n\Sigma^{-1}\bar{\underline{y}})\right]^T (D^{-1} + n\Sigma^{-1})\right.$$
$$\left.\left[\underline{\theta} - (D^{-1} + n\Sigma^{-1})^{-1}(D^{-1}\underline{d} + n\Sigma^{-1}\bar{\underline{y}})\right]\right\}$$

$$\underline{\theta} \mid \text{Data} \sim N\left((D^{-1} + n\Sigma^{-1})^{-1}(D^{-1}\underline{d} + n\Sigma^{-1}\bar{\underline{y}}),\ (D^{-1} + n\Sigma^{-1})^{-1}\right)$$

We say that $\Omega\,(=\Sigma^{-1})$ has a Wishart distribution ③

if:

$$p(\Omega) = \frac{1}{2^{\nu p/2}\,|A|^{-\nu/2}\,\Gamma_p(\nu/2)}\,|\Omega|^{(\nu-p-1)/2}\,\exp\{-\mathrm{tr}(A\Omega)\}$$

where $\mathrm{tr}(B)$: trace of the matrix

$$= \sum_{i=1}^{p} b_{ii} = \text{sum of diagonal elements}$$

$$\Gamma_p(\nu/2) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(\frac{\nu}{2} + \frac{1-j}{2}\right)$$

↓ "multivariate" gamma function

↳ regular gamma function

$\Omega \in \mathbb{R}^{p\times p}$ and symmetric + positive definite.

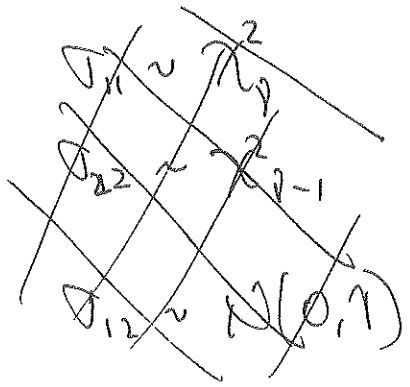
Two parameter ν, A

$$E(\Omega) = A^{-1}\nu$$

$$E(\Omega^{-1}) = E(\Sigma) = \frac{A}{\nu-p-1}$$

How to sample a Wishart (ν, I)

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix}$$



Check!!!

$$\Omega = BB^T$$

$$B = \begin{pmatrix} b_{11} & 0 & 0 & \cdots \\ b_{21} & b_{22} & 0 & \cdots \\ b_{31} & b_{32} & b_{33} & \cdots \end{pmatrix}$$

$b_{11} \sim \sqrt{\chi^2_\nu}$ $\quad \omega_{11} \sim \chi^2_\nu$

$b_{22} \sim \sqrt{\chi^2_{\nu-1}}$ $\Rightarrow$ $\omega_{22} \sim \chi^2_\nu$

$\vdots$ $\quad \vdots$

$b_{21} \sim N(0,1)$

The Wishart is a conjugate prior for the precision matrix $\Omega = \Sigma^{-1}$ Assume θ Known.

$$p(y_1, \cdots y_n \mid \theta, \Omega) \propto |\Omega|^{+n/2} \exp\left\{-\frac{1}{2}\sum (\underline{y}_i - \underline{\theta})^T \Omega (\underline{y}_i - \underline{\theta})\right\}$$

$$p(\Omega \mid y_1, \cdots y_n) \propto |\Omega|^{(\nu-p-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(A\Omega)\right\} |\Omega|^{n/2}$$

$$\exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}(y_i - \underline{\theta})^T \Omega (y_i - \underline{\theta})\right\}$$

If the dimensions allow it, then

$$\text{trace}(ABC) = \text{trace}(BCA) = \text{trace}(CAB)$$

$$\Rightarrow p(\Omega \mid \text{data}) \propto |\Omega|^{(\nu+n-p-1)/2} \exp\left\{-\frac{1}{2}\left[\text{tr}(A\Omega) + \text{tr}\left(\sum_{i=1}^{n}(y_i-\theta)(y_i-\theta)^T \Omega\right)\right]\right\}$$

$$\Rightarrow \Omega \mid \text{data} \sim \text{Wishart}\left(\nu+n,\ A + \underbrace{\sum_{i=1}^{n}(y_i-\theta)(y_i-\theta)^T}_{p \times p \text{ matrix}}\right)$$

Inference for multivariate normal distributions

$\underset{\sim}{Y}_1, \dots \underset{\sim}{Y}_n \overset{iid}{\sim} N_p(\underset{\sim}{\theta}, \Sigma)$ $\quad \Omega = \Sigma^{-1}$ both $\underset{\sim}{\theta}$ and Σ are unknown

$\underset{\sim}{\theta}, \Sigma$

Two options:

* Fully conjugate Normal-Wishart prior

$$\underset{\sim}{\theta} | \Sigma \sim N_p(\underset{\sim}{d}, \tfrac{1}{k}\Sigma)$$

$$\Omega = \Sigma^{-1} \sim \text{Wishart}(\nu, A)$$

* Conditionally conjugate prior

$$\underset{\sim}{\theta} \sim N_p(\underset{\sim}{d}, D)$$

$$\Omega = \Sigma^{-1} \sim \text{Wishart}(\nu, A)$$

We will focus on the second!!

In that case computation proceed using "Gibbs sampling"

Posterior:

$$p(\theta, \Omega | \underset{\sim}{y}_1, \dots \underset{\sim}{y}_n) \propto \left(\frac{1}{\sqrt{2\pi}}\right)^{np} |\Omega|^{n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} (\underset{\sim}{y}_i - \underset{\sim}{\theta})^T \Omega (\underset{\sim}{y}_i - \underset{\sim}{\theta})\right\}$$

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{p} |D|^{-1/2} \exp\left\{-\frac{1}{2}(\underset{\sim}{\theta} - \underset{\sim}{d})^T D^{-1} (\underset{\sim}{\theta} - \underset{\sim}{d})\right\}$$

$$|\Omega|^{\frac{\nu - p - 1}{2}} \exp\left\{-\frac{1}{2}\text{trace}\{A\Omega\}\right\}$$

$$p(\underline{\theta} \mid \Omega, \underline{y}) = p(\underline{\theta} \mid \cdots)$$

$$\propto \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\underline{y}_i - \underline{\theta})^T \Omega (\underline{y}_i - \underline{\theta}) \right\}$$

$$\exp\left\{ -\frac{1}{2} (\underline{\theta} - \underline{d})^T D^{-1} (\underline{\theta} - \underline{d}) \right\}$$

$$\theta \mid \Omega, \underline{y} \sim N\left((n\Omega + D^{-1})^{-1} (n\Omega \bar{y} + D^{-1}\underline{d}), (n\Omega + D^{-1})^{-1} \right)$$

$$p(\Omega \mid \theta, \underline{y}) \propto |\Omega|^{n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\underline{y}_i - \theta)^T \Omega (\underline{y}_i - \theta) \right\}$$

$$|\Omega|^{\frac{\nu - p - 1}{2}} \exp\left\{ -\frac{1}{2} \mathrm{tr}(A\Omega) \right\}$$

$$\Omega \mid \theta, \underline{y} \sim \mathrm{Wis}\left(\nu + n, A + \sum_{i=1}^{n} (\underline{y}_i - \theta)(\underline{y}_i - \theta)^T \right)$$

A slightly fancier version deals with some observations being missing.

$$Y_i = \begin{pmatrix} 2.03 \\ 1.82 \\ -1.07 \\ \vdots \\ .053 \end{pmatrix} \qquad Y_{i'} = \begin{pmatrix} 2.17 \\ NA \\ 0.17 \\ \vdots \\ -0.16 \end{pmatrix}$$

$\Downarrow$ fully observed

$\Downarrow$ partially observed.

Y = all observations.

$Y = (Y_{obs}, Y_{mis})$ → the value of the observations that I did not get

$P(Y_{obs})$

The previous model can still be used for Y.

$$Y_{obs}, Y_{mis} \mid \theta, \Omega \sim \prod_{i=1}^{n} N(Y_i \mid \theta, \Omega)$$

$\Downarrow$

$$p(Y_{obs}, Y_{mis}, \theta, \Omega) = p(Y_{obs}, Y_{mis} \mid \theta, \Omega)\, p(\theta, \Omega)$$

We are interest in

$$p(\theta, \Omega \mid Y_{obs}) = \int p(\theta, \Omega, Y_{mis} \mid Y_{obs})\, dY_{mis}.$$

We can sample from $P(Y_{mis}, \theta, \Omega \mid Y_{obs})$ using Gibbs sampling.

$P(\theta \mid \Omega, Y_{mis}, Y_{obs})$

$P(\Omega \mid \theta, Y_{mis}, Y_{obs})$ Three full conditional distributions

$P(Y_{mis} \mid \theta, \Omega, Y_{obs})$

* The first two reduce to the case where there is no missing values.

* For the third, each missing value comes from a normal distribution that is the conditional distribution of the missing entry given the observed ones.

$$Y_i = \begin{pmatrix} Y_{obs,i} \\ Y_{mis,i} \end{pmatrix} \Bigg| \begin{pmatrix} \underline{\theta}_{obs} \\ \underline{\theta}_{mis} \end{pmatrix}, \begin{pmatrix} \Sigma_{oo} & \Sigma_{om} \\ \Sigma_{mo} & \Sigma_{mm} \end{pmatrix} \sim N(Y_i \mid \underline{\theta}, \Sigma)$$

$\underline{\theta}$ $\quad$ Σ

$$Y_{mis,i} \mid Y_{obs,i}, \underline{\theta}, \Sigma \sim N\left(\underline{\theta}_{mis} + \Sigma_{mo}\Sigma_{oo}^{-1}(Y_{obs,i} - \theta_{obs}),\ \Sigma_{m,m} - \Sigma_{mo}^{*}\Sigma_{oo}^{-1}\Sigma_{om}\right)$$

Hierarchical modeling

Y_{ij} = SAT of the j-th student in the i-th school of a given school district.

Simplest model:

$$Y_{ij} | \theta, \sigma^2 \overset{\text{ind}}{\sim} N(\theta, \sigma^2) \quad i = 1, \dots I \text{ and } j = 1, \dots J_i$$

$$\theta \sim N(\mu, \tau^2)$$

$$\sigma^2 \sim \text{IGam}(a, b)$$

Does not account for the nested structure in the data

A better model:

$$Y_{ij} | \theta_i, \sigma^2 \sim N(\theta_i, \sigma^2) \qquad \sigma^2 \sim IG(a, b)$$

$$\theta_i \sim N(\mu, \tau^2)$$

If μ and τ^2 are fixed, then this is almost equivalent to fitting a separate model for each school.

Instead use priors for μ, τ^2.

$$\Rightarrow \theta_i | \mu, \tau^2 \sim N(\mu, \tau^2)$$

$$\mu \sim N(\eta, K^2) \qquad \tau^2 \sim \text{IGam}(c, d)$$

- Repeated measurements model
- Random effects model
- Hierarchical model

How do we perform computation for this model?

$$p(\theta_1, \ldots \theta_I, \sigma^2, \mu, \tau^2 \mid \{y_{ij}\})$$

$$= \left[\prod_{i=1}^{I} \prod_{j=1}^{J_i} p(y_{ij} \mid \theta_i, \sigma^2)\right]\left[\prod_{i=1}^{I} p(\theta_i \mid \mu, \tau^2)\right] p(\mu)\, p(\tau^2)\, p(\sigma^2)$$

Full conditionals:

$\theta_i \mid \cdots$ $\quad$ $\sigma^2 \mid \cdots$

$\mu \mid \cdots$ $\quad$ $\tau^2 \mid \cdots$

$$\Rightarrow p(\theta_i \mid \cdots) \propto \left[\prod_{j=1}^{J_i} p(y_{ij} \mid \theta_i, \sigma^2)\right] p(\theta_i \mid \mu, \tau^2)$$

$$\propto \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{J_i} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{J_i}(y_{ij} - \theta_i)^2\right\} \frac{1}{\sqrt{2\pi}\tau} \exp\left\{-\frac{1}{2\tau^2}(\theta_i - \mu)^2\right\}$$

$$\Rightarrow \theta_i \mid \cdots \sim N\left(\frac{\frac{J_i \bar{y}_{i\cdot}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{J_i}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{J_i}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

$$p(\mu \mid \cdots) \propto \left[\prod_{i=1}^{I} p(\theta_i \mid \mu, \tau^2)\right] p(\mu)$$

$$\propto \left(\frac{1}{\sqrt{2\pi}\,\tau}\right)^{I} \exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^{I}(\theta_i - \mu)^2\right\} \frac{1}{\sqrt{2\pi}\,K} \exp\left\{-\frac{1}{2K^2}(\mu - \eta)^2\right\}$$

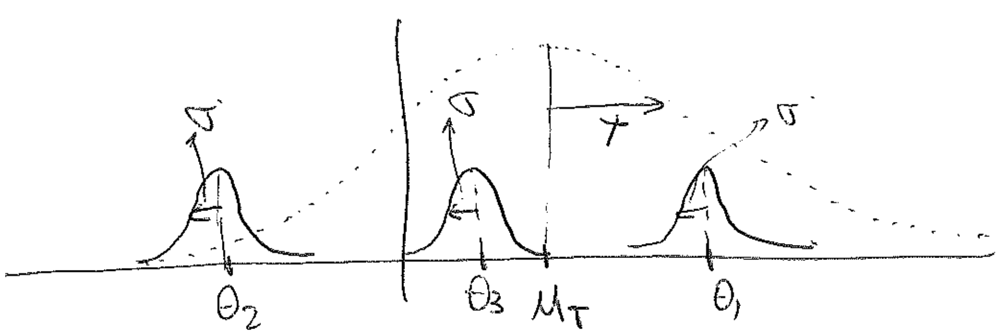
$$\mu \mid \cdots \sim N\left(\frac{\frac{I\bar{\theta}}{\tau^2} + \frac{\eta}{K^2}}{\frac{I}{\tau^2} + \frac{1}{K^2}} , \frac{1}{\frac{I}{\tau^2} + \frac{1}{K^2}} \right)$$

$$p(\sigma^2 \mid \cdots) \propto \prod_{i=1}^{I} \left(\frac{1}{\sigma^2}\right)^{J_i/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{J_i}(y_{ij} - \theta_i)^2\right\} \left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left\{-\frac{b}{\sigma^2}\right\}$$

$$\left(\frac{1}{\sigma^2}\right)^{a + \frac{\sum_{i=1}^{I} J_i}{2} + 1} \exp\left\{-\frac{1}{\sigma^2}\left(b + \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \theta_i)^2\right)\right\}$$

$$\sigma^2 \mid \cdots \sim IGam\left(a + \frac{\sum_{i=1}^{I} J_i}{2}, \; b + \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \theta_i)^2\right)$$

$$\tau^2 \mid \cdots \sim IGam\left(c + \frac{I}{2}, \; d + \frac{1}{2}\sum_{i=1}^{I}(\theta_i - \mu)^2\right)$$



Alternative model

$$Y_{ij} \mid \theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2)$$

$$\theta_i \mid \mu, \tau^2 \sim N(\mu, \tau^2)$$

$$\sigma_i^2 \sim IGam(a, b)$$

$$\mu_0 \sim N(\gamma, K^2)$$

$$b \sim Gam(\xi_1, \xi_2)$$

$$\tau^2 \sim IGam(c, d)$$

NIMBLE

$$x_i | \lambda_i \overset{iid}{\sim} \text{Poisson}(\lambda_i)$$

$$\lambda_i = \theta_i t_i \qquad \text{where } t_i \text{ is known}$$

$$\theta_i | \alpha, \beta \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$$

$$\alpha \sim \text{Gamma}(1, 1)$$

$$\beta \sim \text{Gamma}(0.1, 1)$$

Full conditionals:

$\theta_i \mid \text{data}, \alpha, \beta$

$\alpha \mid \text{data}, \{\theta_i\}$

$\beta \mid \text{data}, \{\theta_i\}$

$$p(\theta_i | \cdots) \propto \frac{e^{-\theta_i t_i} (\theta_i t_i)^{x_i}}{x_i!} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_i^{\alpha-1} \exp\{-\beta \theta_i\}$$

$$\propto \exp\{-\theta_i (\beta + t_i)\} \; \theta_i^{x_i + \alpha - 1}$$

$$\theta_i | \cdots \sim \text{Gamma}(\alpha + x_i, \beta + t_i)$$

$$p(\beta|\cdots) \propto \left[\prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta_i^{\alpha-1} \exp\{-\beta\theta_i\}\right] \beta^{0.1-1} \exp\{-\beta\}$$

$$\propto \beta^{n\alpha} \exp\{-\beta \sum_{i=1}^{n} \theta_i\} \beta^{-0.9} \exp\{-\beta\}$$

$$= \beta^{n\alpha-0.9} \exp\{-\beta\left[1 + \sum_{i=1}^{n} \theta_i\right]\}$$

$$\beta|\cdots \sim \text{Gamma}\left(n\alpha + 0.1, \; 1 + \sum_{i=1}^{n} \theta_i\right)$$

$$p(\alpha|\cdots) \propto \left[\prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta_i^{\alpha-1} \exp\{-\beta\theta_i\}\right] \exp\{-\alpha\}$$

$$\propto \frac{\beta^{n\alpha}}{[\Gamma(\alpha)]^n} \left[\prod_{i=1}^{n} \theta_i\right]^{\alpha-1} \exp\{-\alpha\}$$

Needs a random walk.

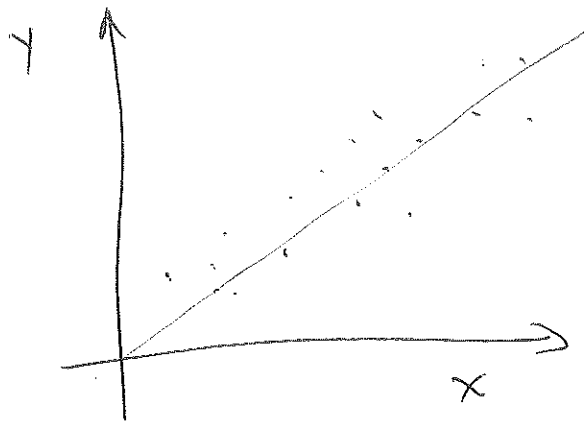
Linear models:

$$Y_i = \beta_0 + \sum_{K=1}^{P} X_{iK}\beta_K + \varepsilon_i$$

- Y_i: observed values
- β_0, β_K: unknown parameters
- X_{iK}: known predictor
- ε_i: random error

Simple linear regression:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (p=1)$$



Vector form:

$$Y_i = \underset{\sim}{X}_i^T \underset{\sim}{\beta} + \varepsilon_i$$

$$\underset{\sim}{X}_i^T = (1, X_{i1}, X_{i2}, \dots X_{ip})$$

$$\beta^T = (\beta_0, \beta_1, \beta_2, \dots \beta_p)$$

Matrix form:

$$\underset{\sim}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

$$\underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X}_1^T \\ \underset{\sim}{X}_2^T \\ \vdots \\ \underset{\sim}{X}_3^T \end{pmatrix} \qquad \underset{\sim}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

First you need a likelihood if you are going to do any kind of (Bayesian) inference

$\underset{\sim}{\varepsilon} \sim N_n(0, \Sigma)$ simplest case $\Sigma = \sigma^2 I$

but you could have more sophisticated scenarios

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & 0 & \cdots & 0 \\ \rho & 1 & \rho & \cdots & 0 \\ 0 & \rho & 1 & \ddots & \\ \vdots & & & & 1 \\ 0 & 0 & 0 & & \end{pmatrix}^{-1}$$ ⇒ AR

Yet another way to write the model is

$$\underset{\sim}{y} \sim N_n(X\underset{\sim}{\beta}, \Sigma)$$

How do we do Bayesian inference for this model.

Consider first

$\underset{\sim}{y} \sim N_n(X\underset{\sim}{\beta}, \sigma^2 D)$ with D known and $\underset{\sim}{\beta}, \sigma^2$ unknown.

Priors:

$\underset{\sim}{\beta} \sim N(\eta, C)$

$\sigma^2 \sim IGam(a, b)$

Computation with a Gibbs sampler:

$$p(\underline{\beta}, \sigma^2 | \text{Data}) \propto \cancel{\left(\frac{1}{\sqrt{2\pi}}\right)^n} \left(\frac{1}{\sigma^2}\right)^{n/2} \cancel{|D|^{-1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\underline{y} - X\underline{\beta})^T D^{-1}(\underline{y} - X\underline{\beta})\right\}$$

$$\cancel{\left(\frac{1}{\sqrt{2\pi}}\right)^p} |C|^{-1/2} \exp\left\{-\frac{1}{2}(\underline{\beta} - \underline{\eta})^T C^{-1}(\underline{\beta} - \underline{\eta})\right\}$$

$$\left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left\{-\frac{b}{\sigma^2}\right\}$$

$$p(\underline{\beta} | \sigma^2, \text{data}) \propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\underline{\beta}^T X^T D^{-1} X\underline{\beta} - \frac{2}{\sigma^2}\underline{\beta}^T X^T D^{-1} y + \underline{\beta}^T C^{-1}\underline{\beta} - 2\underline{\beta}^T C^{-1}\underline{\eta}\right]\right\}$$

$$= \exp\left\{-\frac{1}{2}\left[\underline{\beta}^T\left[\frac{1}{\sigma^2}X^T D^{-1} X + C^{-1}\right]\underline{\beta} - 2\underline{\beta}^T\left(\frac{1}{\sigma^2}X^T D^{-1} y + C^{-1}\underline{\eta}\right)\right]\right\}$$

$$\underline{\beta} | \sigma^2 \sim N\left(\left(\frac{1}{\sigma^2}X^T D^{-1} X + C^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}X^T D^{-1} y + C^{-1}\underline{\eta}\right), \left(\frac{1}{\sigma^2}X^T D^{-1} X + C^{-1}\right)^{-1}\right)$$

$$p(\sigma^2 | \underline{\beta}, \text{data}) \propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\underline{\beta})^T D^{-1}(y - X\underline{\beta})\right\}$$

$$\left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left\{-\frac{b}{\sigma^2}\right\}$$

$$p(\sigma^2 \mid \beta, \text{data}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+a+1} \exp\left\{-\frac{1}{\sigma^2}\left(b + (\underline{y} - X\underline{\beta})^T D^{-1} (\underline{y} - X\underline{\beta})\right)\right\} \quad (4)$$

$$\sigma^2 \sim \text{IGam}\left(a + \frac{n}{2},\ b + \frac{1}{2}(\underline{y} - X\underline{\beta})^T D^{-1} (\underline{y} - X\underline{\beta})\right)$$

If $D = I \Rightarrow$ independence of the observations

and $X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ and $\underline{\beta} = \beta_0$

Then

$$(\underline{y} - X\beta)^T D^{-1} (\underline{y} - X\underline{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0)^2.$$

Gibbs sampler is relatively easy.

(More) Special cases of the linear model:

ANOVA model (one-way)

$$y_{ij} = \theta_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2) \qquad \begin{matrix} i = 1, \dots I \\ j = 1, \dots J_i \end{matrix}$$

$$\theta_i \sim N(\mu, \tau^2)$$

This can be rewritten in matrix form:

$$y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1,J_1} \\ y_{2,1} \\ \vdots \\ y_{2,J_2} \\ \vdots \\ y_{I,1} \\ \vdots \\ y_{I,J_I} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & & & & \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \\ 0 & 1 & 0 & \cdots & 0 \\ & & & & \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}}_{X} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_I \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1,J_1} \\ \vdots \\ \varepsilon_{I1} \\ \vdots \\ \varepsilon_{I,J_I} \end{pmatrix}$$

Fourier Regression:

$$y_i = f(x_i) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

$$f(x) = \sum_{k=1}^{K} b_k(x)\, \theta_k \qquad \Rightarrow \quad y_i = \sum_{k=1}^{K} b_k(x)\, \theta_k + \varepsilon_i$$

$$\Downarrow$$

$$\cos\left(\frac{\omega x}{K}\right)$$

$$y = \begin{pmatrix} b_1(x_1) & \cdots & b_K(x_1) \\ b_1(x_2) & \cdots & b_K(x_2) \\ \vdots & & \\ b_1(x_N) & \cdots & b_K(x_N) \end{pmatrix} \underbrace{\begin{pmatrix} \theta_1 \\ \vdots \\ \theta_K \end{pmatrix}}_{\theta} + \varepsilon$$