Spatio-temporal Methods in Environmental Epidemiology

Alexandra M. Schmidt

Session 2 - An introduction to Bayesian Inference

Department of Epidemiology, Biostatistics and Occupational Health
McGill University

alexandra.schmidt@mcgill.ca

Applied Bayesian Statistics School 2025

University of Genova, Department of Architecture and Design
03-06, June 2025

Part I, Section 1: Introduction to Bayesian inference

Bayesian inference

From http://www.bayesian.org/

”Scientific inquiry is an iterative process of integrating accumulating information. Investigators assess the current state of knowledge regarding the issue of interest, gather new data to address remaining questions, and then update and refine their understanding to incorporate both new and old data. Bayesian inference provides a logical, quantitative framework for this process. It has been applied in a multitude of scientific, technological, and policy settings.”

Bayes' Theorem and methodology

Summarized presentation of methodology

Bayes Theorem

Observations y: described by density $f(y|\theta)$
Likelihood: $l(\theta) = f(y|\theta)$
θ: index of f (parameter)
Canonical situation: random sample
$y = (y_1, ..., y_n)$ taken from $f(y|\theta)$.

Example

Measurements of a physical quantity θ with errors $e_i \sim N(0, \sigma^2)$, σ^2 known.
$y_i = \theta + e_i$, $i = 1, ..., n$ and $f(y|\theta) =$

$$\prod_{i=1}^{n} f_N(y_i; \theta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(y_i - \theta)^2}{\sigma^2} \right\}$$

θ is more than simple index

- Situation repeats in more general cases
- Very likely that researcher has prior information about θ
- This may be modelled through density $p(\theta)$
- Lots of controversy in the past

Bayes' Theorem and methodology

Inference process based on distribution of θ after observing $y \rightarrow$ posterior distribution (as opposed to prior)
Obtained through Bayes theorem as

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta)f(y|\theta)}{f(y)} \qquad \text{or} \\ \pi(\theta) &\propto p(\theta)\, l(\theta) \end{aligned}$$

$$\text{posterior dist.} \quad \propto \quad \text{prior dist.} \times \text{likelihood}$$

And

$$f(y) = \int f(y|\theta)p(\theta)d\theta$$

It is not generally necessary to compute the denominator.

Bayes' Theorem and methodology

Predictive Distribution

Prediction (or forecast) of a future observation y^* after observing y based on the distribution of $(y^*|y)$

$$f(y^*|y) = \int f(y^*, \theta|y)d\theta = \int f(y^*|\theta)\, p(\theta \mid y)\, d\theta$$

if y and y^* are conditionally independent given $\theta \rightarrow$ eg. random sample

Part I, Section 2: Conjugate Families

The main conjugate families

Theorem 1

Let $\theta \sim N(\mu, \tau^2)$ and $X \mid \theta \sim N(\theta, \sigma^2)$, with σ^2 known.
The posterior distribution of θ is $(\theta \mid X = x) \sim N(\mu_1, \tau_1^2)$ where

$$\mu_1 = \frac{\tau^{-2}\mu + \sigma^{-2}x}{\tau^{-2} + \sigma^{-2}} \quad \text{e} \quad \tau_1^{-2} = \tau^{-2} + \sigma^{-2} \tag{1}$$

- The precision is the inverse of the variance
- From the result above, the posterior precision of μ is the sum between the prior precision and the likelihood, and it does not depend on x

Sampling from the normal distribution with known variance

- Looking at the precision as a measure on the amount of information, and defining $w = \tau^{-2}/(\tau^{-2} + \sigma^{-2}) \in (0, 1)$
- w measures the relative information contained in the prior distribution with respect to the total information (prior + likelihood) Then we can write

$$\mu_1 = w\mu + (1 - w)x$$

which is a weighted average between the prior mean and the likelihood mean

Proving the previous theorem

Proof : From Bayes theorem

$$\begin{aligned} p(\theta \mid x) &\propto l(\theta; x)p(\theta) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2 - \frac{1}{2\tau^2}(\theta-\mu)^2\right\} \\ &\propto \exp\left\{-\frac{\theta^2}{2\sigma^2} - \frac{\theta^2}{2\tau^2} + \frac{x\theta}{\sigma^2} + \frac{\mu\theta}{\tau^2}\right\} \\ &= \exp\left\{-\frac{\theta^2}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right) + \theta\left(\frac{x}{\sigma^2}+\frac{\mu}{\tau^2}\right)\right\} \end{aligned}$$

where, in the first step, all the constants were included in the proportionality constant

Now let $\tau_1^2 = (\tau^{-2} + \sigma^{-2})^{-1}$ and $\mu_1 = (\sigma^{-2}x + \mu\tau^{-2})\tau_1^2$

Proving the previous theorem (cont.)

Substituting the expressions in the previous slide, we have

$$\begin{aligned} p(\theta \mid x) &\propto \exp\left\{ -\frac{\theta^2}{2\tau_1^2} + \frac{\theta\mu_1}{\tau_1^2} \right\} \\ &\propto \exp\left\{ -\frac{1}{2\tau_1^2}(\theta - \mu_1)^2 \right\} \\ &\propto \frac{1}{\sqrt{2\pi\tau_1^2}} \exp\left\{ -\frac{1}{2\tau_1^2}(\theta - \mu_1)^2 \right\} \end{aligned}$$

Note that the last term in the expression above corresponds to that of a normal density
Therefore, the constant of proportionality is equal to 1 and $(\theta \mid x) \sim N(\mu_1, \tau_1^2)$.

The main conjugate families

Binomial Distribution

The family of beta distributions is conjugate to the binomial (or Bernoulli) model.

Normal with known variance

Theorem 1 stated that the normal family is conjugate to the normal model. For a sample of size n we have

$$l(\theta; \mathbf{x}) \propto \exp\left\{-\frac{n}{2\sigma^2}(\overline{x} - \theta)^2\right\}$$

Note that $p(\mathbf{x} \mid \theta) \propto p(\overline{X} \mid \theta)$. Assuming $\theta \sim N(\mu, \tau^2)$ we have that $\theta \mid \mathbf{x} \sim N(\mu_1, \tau_1^2)$ where

$$\mu_1 = \frac{n\sigma^{-2}\overline{x} + \tau^{-2}\mu}{n\sigma^{-2} + \tau^{-2}} \text{ and } \tau_1^{-2} = n\sigma^{-2} + \tau^{-2}.$$

The main conjugate families

Poisson distribution

Suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is a random sample from the Poisson distribution with parameter θ, denoted $Pois(\theta)$. Its joint probability function is

$$p(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!}.$$

The Gamma family of distributions is conjugate to the Poisson model.

Exponential distribution

Suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is a random sample from the Exponential distribution with parameter θ, denoted $Exp(\theta)$. Its joint probability function is

$$p(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i}.$$

The Gamma family of distributions is conjugate to the Exponential model.

The main conjugate families

Multinomial distribution

Let $\mathbf{X} = (X_1, \cdots, X_p)$ and $\theta = (\theta_1, \cdots, \theta_p)$ be, respectively, the number of observed cases and the probabilities associated with each of p categories in a sample of size n. Assume that $\sum_{i=1}^{n} X_i = n$ and $\sum_{i=1}^{p} \theta_i = 1$. $\mathbf{X}$ is said to have a multinomial distribution with parameters n and $(\theta_1, \cdots, \theta_p)$. And

$$p(\mathbf{x} \mid \theta) = \frac{n!}{\prod_{i=1}^{p} x_i!} \prod_{i=1}^{p} \theta_i^{x_i}.$$

It also belongs to the exponential family and $l(\theta) \propto \prod \theta_i^{x_i}$. Its kernel is the same as the kernel of the density of a Dirichlet distribution. The Dirichlet family with integer parameters $a_1, \cdots, a_p$ is natural conjugate with respect to the multinomial sampling distribution. Again, little is lost by extending natural conjugacy over all Dirichlet distributions.

The main conjugate families

The posterior will then be

$$p(\theta \mid \mathbf{x}) \propto \left[\prod_{i=1}^{p} \theta_i^{x_i}\right]\left[\prod_{i=1}^{p} \theta_i^{a_i-1}\right] = \prod_{i=1}^{p} \theta_i^{x_i+a_i-1},$$

which is a Dirichlet with parameters
$a_1 + x_1, \cdots, a_p + x_p$ and is denoted by
$(\theta \mid \mathbf{x}) \sim D(a_1 + x_1, \cdots, a_p + x_p)$.
The proportionality constant is given by

$$\frac{\Gamma(a+n)}{\prod_{i=1}^{p} \Gamma(a_i + x_i)}$$

This conjugate analysis generalizes the analysis for binomial samples with beta priors.

The main conjugate families

Normal with known mean and unknown variance

Let $\mathbf{X} = (X_1, \cdots, X_p)$ be a random sample from the $N(\theta, \sigma^2)$, θ known, and $\phi = \sigma^{-2}$. In this case we have that

$$l(\phi; \mathbf{x}) = p(\mathbf{x} \mid \theta, \phi) \propto \phi^{n/2} \exp\left\{ -\frac{\phi}{2} n s_0^2 \right\} \text{ where}$$

$s_0^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \theta)^2$.

The conjugate prior may have the kernel of $l(\phi; \mathbf{x})$, which is in the gamma distribution form. As the gamma family is closed under sampling we can consider $\phi \sim Ga(n_0/2, n_0\sigma_0^2/2)$ or, equivalently, $n_0\sigma_0^2\phi \sim \chi^2_{n_0}$. Then

$$\phi \mid \mathbf{x} \sim Ga\left(\frac{n_0 + n}{2}, \frac{n_0\sigma_0^2 + n s_0^2}{2} \right),$$

or, equivalently, $(n_0\sigma_0^2 + n s_0^2)\phi \mid \mathbf{x} \sim \chi^2_{n_0+n}$.

The main conjugate families

Normal with unknown mean and variance

Assume that (θ, ϕ) be such that $X_i \sim N(\theta, \sigma^2)$, where $\phi = \sigma^{-2}$. We will consider the prior for (θ, ϕ) in two stages:

$$(\theta \mid \phi) \sim N(\mu_0, (c_0\phi)^{-1}) \text{ and}$$
$$n_0\sigma_0^2\phi \sim \chi^2_{n_0} \text{ or } \phi \sim Ga\left(n_0/2, n_0\sigma_0^2/2\right),$$

where $(\mu_0, c_0, n_0, \sigma_0^2)$ are obtained from the initial information H. This distribution is usually called normal-gamma or normal-χ^2 with parameters $(\mu_0, c_0, n_0, \sigma_0^2)$ and joint density given by:

$$p(\theta, \phi) \propto \phi^{(n_0+1)/2-1} \exp\left\{-\frac{\phi}{2}\left[n_0\sigma_0^2 + c_0(\theta-\mu_0)^2\right]\right\}$$

Note that

$$\begin{aligned} p(\theta) &= \int_0^\infty \phi^{a-1} e^{-b\phi} d\phi = \frac{\Gamma(a)}{b^a} \\ &\propto [n_0\sigma_0^2 + c_0(\theta-\mu_0)^2]^{-(n_0+1)/2} \end{aligned}$$

The main conjugate families

Rearranging the terms of the last equation gives

$$p(\theta) \quad \propto \quad \left[1 + \frac{(\theta - \mu_0)^2}{n_0(\sigma_0^2/c_0)}\right]^{-(n_0+1)/2}$$

implying that $\theta \sim t_{n_0}(\mu_0, \sigma_0^2/c_0)$.
The conditional distribution of $\phi \mid \theta$ can be obtained from the joint distribution of (θ, ϕ) and

$$\phi \mid \theta \sim Ga\left(\frac{(n_0+1)}{2}, \frac{[n_0\sigma_0^2 + c_0(\theta - \mu_0)^2]}{2}\right)$$

The joint distribution of a random sample $\mathbf{X} = (X_1, \cdots, X_n)$ is

$$p(\mathbf{x} \mid \theta, \phi) = \prod_{i=1}^{n} \phi^{1/2} \exp\left\{-\frac{\phi}{2}(x_i - \theta)^2\right\}$$

$$\propto \phi^{n/2} \exp\left\{-\frac{\phi}{2}\left[ns^2 + n(\overline{x} - \theta)^2\right]\right\}$$

where $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.
It has the same kernel as the normal-gama density for (θ, ϕ). The normal-gamma is closed under sampling.

The main conjugate families

The posterior distribution will then be

$$\begin{aligned} p(\theta, \phi \mid \mathbf{x}) &\propto p(\mathbf{x} \mid \theta, \phi) p(\theta, \phi) \\ &\propto \phi^{[(n+n_0+1)/2]-1} \\ &\quad \times \exp\left\{ -\frac{\phi}{2} [n_0 \sigma_0^2 + n s^2 + c_0(\theta - \mu_0)^2 + n(\overline{x} - \theta)^2 \right\}. \end{aligned}$$

But

$$c_0(\theta - \mu_0)^2 + n(\theta - \overline{x})^2 = (c_0 + n)(\theta - \mu_1)^2 + \frac{c_0 n}{c_0 + n}(\mu_0 - \overline{x})^2$$

where $\mu_1 = (c_0 \mu + n\overline{x})/(c_0 + n)$. Then we have that

$$\begin{aligned} p(\theta, \phi) &\propto \phi^{[(n+n_0+1)/2]-1} \\ &\exp\left\{ -\frac{\phi}{2} \left[n_0 \sigma_0^2 + n s^2 + \frac{c_0 n}{c_0 + n}(\mu_0 - \overline{x})^2 + (c_0 + n)(\theta - \mu_1)^2 \right] \right\} \end{aligned}$$

The main conjugate families

The joint posterior for $(\theta, \phi \mid \mathbf{x})$ is normal-gamma with parameters $(\mu_1, c_1, n_1, \sigma_1^2)$ given by

$$\mu_1 = \frac{c_0\mu_0 + n\bar{x}}{c_0 + n} \qquad c_1 = c_0 + n$$

$$n_1 = n_0 + n \qquad n_1\sigma_1^2 = n_0\sigma_0^2 + ns^2 + \frac{c_0 n}{c_0 + n}(\mu_0 - \bar{x})^2.$$

The normal-gamma family is conjugate with respect to the normal sampling model when θ and σ^2 are both unknown.

The main conjugate families

Summary of the Distributions in the normal case

	Prior	Posterior
$\theta \mid \phi$	$N(\mu_0, (c_0\phi)^{-1})$	$N(\mu_1, (c_1\phi)^{-1})$
ϕ	$n_0\sigma_0^2\phi \sim \chi^2_{n_0}$	$n_1\sigma_1^2\phi \sim \chi^2_{n_1}$
θ	$t_{n_0}(\mu_0, \sigma_0^2/c_0)$	$t_{n_1}(\mu_1, \sigma_1^2/c_1)$
$\phi \mid \theta$	$[n_0\sigma_0^2 + c_0(\theta-\mu_0)^2]\phi \sim \chi^2_{n_0+1}$	$[n_1\sigma_1^2 + c_1(\theta-\mu_1)^2]\phi \sim \chi^2_{n_1+1}$

Example (1)

Normal data, $Y_i \sim N(\theta, \sigma^2)$, σ^2 known, then

$$\begin{aligned} l(\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_i-\theta)^2}{\sigma^2}\right\} \\ &\propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\right\} \end{aligned}$$

where $\bar{y}$ is the average of the y_i's.
Model completed with prior for θ,
$p(\theta) = N(\mu, \tau^2)$, μ and τ^2 known.
Then, we have that

$$\pi(\theta) \sim N(\mu_1, \tau_1^2)$$

where
$\tau_1^{-2} = n\sigma^{-2} + \tau^{-2}$ and $\mu_1 = \tau_1^2(n\sigma^{-2}\bar{y} + \tau^{-2}\mu)$
$\tau^2 \to \infty$: non-informative prior $p(\theta) \propto c$ and
$\pi(\theta) = N(\bar{y}, \sigma^2/n)$.
See `ConjugatePriors.r`

Part I, Section 3: Bayes Estimators and Linear Models

Bayes Estimators

A Bayes estimator is an estimator that is chosen to minimize the posterior mean of some measure of how far the estimator is from the parameter.

Loss Function A loss function is a real-valued function of two variables $L(\theta, a)$, where $\theta \in \Omega$ and a is a real number. The interpretation is that the statistician loses $L(\theta, a)$ if the parameter equals θ and the estimate equals a

Definition of a Bayes Estimator

Suppose that one can observe the value $\mathbf{y}$ of random vector $\mathbf{Y}$ before estimating θ, and let $p(\theta \mid \mathbf{x})$ denote the posterior pdf of θ on Ω. For each estimate a that the statistician might use, her expected loss in this case will be

$$E[L(\theta, a) \mid \mathbf{x}] = \int_{\Omega} L(\theta, a) p(\theta \mid \mathbf{x}) d\theta \tag{2}$$

You can choose an estimate a for which the expectation above is a minimum.

Bayes Estimators

Definition Let $L(\theta, a)$ be a loss function. For each possible value $\mathbf{x}$ of $\mathbf{X}$, let $\delta^*(\mathbf{x})$ be a value of a such that $E[L(\theta, a) \mid \mathbf{x}]$ is minimized. Then δ^* is called a *Bayes estimator* of θ. Once $\mathbf{X} = \mathbf{x}$ is observed, $\delta^*(\mathbf{x})$ is called a *Bayes estimate* of θ.
Then, for each possible value of $\mathbf{x}$ of $\mathbf{X}$, the value $\delta^*(\mathbf{x})$ is chosen so that

$$E[L(\theta, \delta^*(\mathbf{x})) \mid \mathbf{x}] = \min_{\text{All a}} E[L(\theta, a) \mid x]$$

Corollary Let θ be a real-valued parameter. Suppose that the squared error loss function, $L(\theta, a) = (\theta - a)^2$ is used, and that $E(\theta \mid \mathbf{x}) < \infty$. Then, a Bayes estimator of θ is $\delta^*(\mathbf{x}) = E(\theta \mid \mathbf{X})$.

Corollary Let θ be a real-valued parameter. Suppose that the absolute error loss function, $L(\theta, a) = |\theta - a|$ is used, then a Bayes estimator of θ is equal to the median of the posterior distribution of θ.

Bayes Estimators

Another form to reduce the effect of large estimation errors is to consider loss functions that remain constant whenever $|\delta - \theta| > k$ for some k arbitrary.
$\Rightarrow$ The most common choice is the limiting value as $k \to 0$.
This loss function associates a fixed loss when an error is committed, irrespective of its magnitude. This loss is usually known as the 0-1 loss.

Lemma: Let $L_3(\delta, \theta) = \lim_{\varepsilon \to 0} I_{|\theta - \delta|}([\varepsilon, \infty))$. The estimator of θ is $\delta_3 =$ mode(θ), the mode of the posterior distribution of θ.
$\to$ also known as the generalized maximum likelihood estimator (GMLE).

Consistency of Bayes Estimators

Under fairly general conditions, and for a wide class of loss functions, the Bayes estimators of some parameters θ will form a consistent sequence of estimators as the sample size $n \to \infty$.

Linear Models

Model specification

- $\mathbf{Y} \mid \beta, \sigma^2 \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$
  - $\mathbf{Y}_{n\times 1}$ vector of observations
  - $\mathbf{X}_{n\times p}$ design matrix
  - $\beta_{p\times 1}$ parameter vector
- $E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$.
- Kernel of the likelihood function

$$\begin{aligned} l(\beta, \sigma^2; \mathbf{y}) &\propto \sigma^{-n} \exp\left\{-\frac{S(\beta)}{2\sigma^2}\right\}, \\ S(\beta) &= (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta). \end{aligned}$$

MLE : $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Bayesian linear models

Bayesian inference

- Model: $\mathbf{y} \mid \beta, \sigma^2 \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$.
- Prior specification for β and $\phi = \sigma^{-2}$.

Conjugate Prior

- $$\begin{cases} \beta|\phi \sim N_p(\mu_0, \phi^{-1}\mathbf{C}_0^{-1}),\ \phi\mathbf{C_0} \text{ precision} \\ \phi \sim Ga\left(\frac{n_0}{2}; \frac{n_0\sigma_0^2}{2}\right) \end{cases}$$
- Joint prior: $p(\beta, \phi) \sim$ Normal $\times$ Gamma;
- Marginal prior for β: $p(\beta) = \frac{p(\beta,\phi)}{p(\phi|\beta)}(*)$
- From the joint prior $p(\beta, \phi)$ it follows that
  - $\phi \mid \beta \sim Ga\left(\frac{n_0+p}{2}, \frac{(n_0+p)[n_0\sigma_0^2+(\beta-\mu_0)'\mathbf{C}_0(\beta-\mu_0)]}{2}\right)$
  - Using $(*)$ or integrating $p(\beta, \phi)$ we have that

$$p(\beta) \propto [n_0\sigma_0^2 + (\beta - \mu_0)\mathbf{C}_0(\beta - \mu_0)]^{-(n_0+p)/2},$$

$$\Rightarrow \beta \sim tM_{n_0}(\mu_0, \sigma_0^2 \mathbf{C}_0{}^{-1}).$$

Bayesian linear models

- On the other hand, the kernel of $l(\beta, \phi; \mathbf{y})$, evaluated at the MLE $\hat{\beta}$, given by

$$\phi^{n/2} \exp\left\{-\frac{\phi}{2}\left[S_e + (\beta - \hat{\beta})'\mathbf{X}'\mathbf{X}(\beta - \hat{\beta})\right]\right\}$$

  has the same form of the prior distribution

- It can be shown that

$$\begin{aligned}(\beta - \mu_0)'\mathbf{C}_0(\beta - \mu_0) + (\beta - \hat{\beta})'\mathbf{X}'\mathbf{X}(\beta - \hat{\beta}) &= \\ (\beta - \mu_1)'\mathbf{C}_1(\beta - \mu_1) + \mu_0'\mathbf{C}_0\mu_0 + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} + \mu_1'\mathbf{C}_1\mu_1 &\end{aligned}$$

  $\mu_1 = \mathbf{C}_1^{-1}(\mathbf{C}_0\mu_0 + \mathbf{X}'\mathbf{y})$ e $\mathbf{C}_1 = \mathbf{C}_0 + \mathbf{X}'\mathbf{X}$.

- Also

$$\begin{aligned} S_e + \mu_0'\mathbf{C}_0\mu_0 + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} + \mu_1'\mathbf{C}_1\mu_1 &= \\ \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} + \mu_0'\mathbf{C}_0\mu_0 + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} + \mu_1'(\mathbf{C}_0\mu_0 + \mathbf{X}'\mathbf{y}) &= \\ (\mathbf{y} - \mathbf{X}\mu_1)'\mathbf{y} + (\mu_0 - \mu_1)'\mathbf{C}_0\mu_0. &\end{aligned}$$

Bayesian linear models

Conjugate posterior distribution

- $p(\beta, \phi \mid \mathbf{y}) \propto \phi^{p/2} \exp\left\{-\frac{\phi}{2}(\beta-\mu_1)'\mathbf{C}_1(\beta-\mu_1)\right\}$

$$\times \phi^{(n_1/2)-1} \exp\left\{-\frac{\phi}{2} n_1 \sigma_1^2\right\},$$

$n_1 = n + n_0$
$n_1\sigma_1^2 = n_0\sigma_0^2 + (\mathbf{y} - \mathbf{X}\mu_1)'\mathbf{y} - (\mu_1 - \mu_0)'\mathbf{C}_0\mu_0.$

- Then $p(\beta, \phi \mid \mathbf{y}) \sim$ Normal $\times$ Gamma

$$\begin{cases} \beta|\phi, \mathbf{y} \sim N_p(\mu_1, (\phi\mathbf{C}_1)^{-1}), \\ \phi|\mathbf{y} \sim Ga\left(\frac{n_1}{2}; \frac{n_1\sigma_1^2}{2}\right) \end{cases}$$

- Marginal posterior: $\beta \mid \mathbf{y} \sim tM_{n_1}(\mu_1, \sigma_1^2\mathbf{C}_1^{-1})$,

$$\begin{cases} E(\beta \mid \mathbf{y}) = \mu_1 = \mathbf{C}_1^{-1}(\mathbf{C}_0\mu_0 + \mathbf{X}'\mathbf{y}), \\ Var(\beta \mid \mathbf{y}) = \frac{n_1}{n_1-2}\sigma_1^2\mathbf{C}_1^{-1}, n_1 > 2. \end{cases}$$

Bayesian linear models

- Marginal Posterior : $\phi|\mathbf{y} \sim Ga\left(\frac{n_1}{2}, \frac{n_1\sigma_1^2}{2}\right)$,

$$\begin{cases} E(\phi|\mathbf{y}) = \phi_1, \\ Var(\phi|\mathbf{y}) = \frac{2\phi_1^2}{n_1} \end{cases}$$

- Credible intervals for β_j and ϕ: are obtained through the percentiles of the t_{n_1} and $Ga\left(\frac{n_1}{2}, \frac{n_1\sigma_1^2}{2}\right)$, respectively.

- Inference about β:

$$(\beta - \mu_1)'\mathbf{C}_1(\beta - \mu_1)\sigma_1^2 \mid \mathbf{y} \sim F(p, n_1 - p).$$

Bayesian linear models

Jeffreys prior:

$$p(\beta,\phi) \propto \phi^{-1} \leftrightarrows \mu_0 \equiv 0, \sigma_0 \to 0 \text{ e } C_0 \to 0.$$

- Posterior density Normal $\times$ Gamma: $p(\beta,\phi \mid \mathbf{y})$

$$\begin{aligned} &\propto \quad \phi^{(n/2)-1} \exp\left\{-\frac{\phi}{2}[S_e + (\beta-\hat{\beta})'\mathbf{X}'\mathbf{X}(\beta-\hat{\beta})]\right\} \\ &\propto \quad \phi^{p/2} \exp\left\{-\frac{\phi}{2}(\beta-\hat{\beta})\mathbf{X}'\mathbf{X}(\beta-\hat{\beta})\right\} \\ &\qquad \times \phi^{((n-p)/2)-1} \exp\left\{-\frac{\phi}{2}(n-p)s^2\right\}. \end{aligned}$$

- Then, $\beta \mid \mathbf{y} \sim tM_{n-p}(\hat{\beta}, s^2(\mathbf{X}'\mathbf{X})^{-1})$,

$$\phi \mid \mathbf{y} \sim Ga\left(\frac{n-p}{2}, \frac{(n-p)s^2}{2}\right) \text{ and}$$

$$(\beta-\hat{\beta})'\mathbf{X}'\mathbf{X}(\beta-\hat{\beta})/s^2 \sim F(p, n-p).$$

- Bayesian and classical results are similar

Bayesian linear models - Example

Multiple linear regression: analysis of heart data

See files `Heart_data.pdf` and `HeartBayesian.r`

Part I, Section 4: Approximating methods

Bayesian Estimation Methods

- We can usually write down the functional form of the posterior distributions required for Bayesian inference.
- For most realistic models, these functions are complex and high dimensional: Difficult to evaluate them analytically.
- Instead, it is often straightforward to *simulate* realisations from the required posterior distributions.
- Posterior summaries (e.g. posterior mean) are easily obtained by simple data summaries of the simulated values.
- Markov Chain Monte Carlo (MCMC) methods are a convenient class of simulation algorithms for this purpose.

Heuristic View of Simulation Methods for Bayesian Inference

- Imagine generating a random sample of values from a probability distribution (e.g.: normal);
- Construct a histogram from the sample;
- If the sample is large enough, histogram can provide virtually complete information about the distribution from which these samples were drawn:
  - Mean, variance, percentiles of sample
    $\approx$ mean, variance, percentiles original distribution.
- MCMC methods enable us to generate large samples from the posterior distributions of model parameters:
  - These samples can be summarised to estimate properties (e.g. mean, variance, percentiles) of the posterior distribution.

Markov chain Monte Carlo Methods

A Markov chain Monte Carlo algorithm to simulate from $\pi(.)$ is any method which produces a homogeneous, ergodic and irreducible Markov chain, whose stationary distribution is $\pi(.)$;

A chain is **ergodic** if it is (i) aperiodic and (ii) positively recurrent;

Periodicity: a chain is aperiodic if none of its states is visited after d steps with probability one, for any d integer and $d > 0$.

Positive Recurrence: a chain is positively recurrent when the mean number of steps for the chain to return to any state is finite.

Markov chain Monte Carlo Methods

A chain is irreducible if, with positive probability, it moves from one (any) point to another in a finite number of iterations.
Results:

- If the Markov chain is homogeneous, irreducible, positively recurrent and aperiodic, then the limit distribution exists and the states of this chain are, approximately, realizations of this stationary distribution.
- Ergodic Means:

$$\bar{h}_m = \frac{1}{m} \sum_{i=1}^{m} h(\theta_i) \to E_\pi[h(\theta)]$$

for $m \to \infty$.

This result is similar to Monte Carlo integration.

Gibbs Sampling

Geman and Geman (1984), Gelfand and Smith (1990)
It is an algorithm that generates a sequence

$$\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots\},$$

from a Markov chain, whose limit/equilibrium distribution is $\pi(\theta)$ and whose transition kernel is given by the product of the full conditional distributions.
Algorithm:

1. $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)})$
2. $\theta^{(j)}$ obtained from $\theta^{(j-1)}$

$$\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1 | \theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}) \\
\theta_2^{(j)} &\sim \pi(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}) \\
\theta_3^{(j)} &\sim \pi(\theta_3 | \theta_1^{(j)}, \theta_2^{(j)}, \theta_4^{(j-1)}, \ldots, \theta_p^{(j-1)}) \\
&\vdots \\
\theta_p^{(j)} &\sim \pi(\theta_p | \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_{p-1}^{(j)})
\end{aligned}$$

Example

Assume that $\mathbf{X} \sim N_2(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

It is also known that for $i \neq j = 1, 2$,

$$\begin{aligned} (X_i | X_j = x_j) &\sim N(\mu_{i|j}, \sigma_{i|j}^2) \\ \mu_{i|j} &= \mu_i + \sigma_{ij}\sigma_j^{-2}(x_j - \mu_j) \\ \sigma_{i|j} &= \sigma_i^2 - \sigma_{ij}^2\sigma_j^{-2} \end{aligned}$$

Let

$$\mu = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \Sigma = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$$

See file `GibbsExample.r`

Change of mean in the Poisson

Assume that

$$y_i|\lambda,\phi,k \sim \begin{cases} Po(\lambda) & \text{para } i = 1,2,\ldots,k \\ Po(\phi) & \text{para } i = k+1,k+2,\ldots,n \end{cases}$$

such that,

$$p(\mathbf{y}|\lambda,\phi,k) = \left[\prod_{i=1}^{k} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}\right]\left[\prod_{i=k+1}^{n} \frac{\phi^{y_i} e^{-\phi}}{y_i!}\right]$$

and likelihood function given by:

$$l(\lambda,\phi,k|\mathbf{y}) \propto \lambda^{t_1(\mathbf{y},k)} e^{-k\lambda} \phi^{t_2(\mathbf{y},k)} e^{-(n-k)\phi}$$

where

$$t_1(\mathbf{y},k) = \sum_{i=1}^{k} y_i \qquad t_2(\mathbf{y},k) = \sum_{i=k+1}^{n} y_i$$

Carlin, Gelfand and Smith (1992) apply this model to the British coalmining disaster data.

Prior Distributions:

$$\begin{aligned} \pi(\lambda|\alpha,\beta) &\sim Ga(\alpha,\beta) \\ \pi(\phi|\gamma,\delta) &\sim Ga(\gamma,\delta) \\ Pr(k=i) &= 1/n \end{aligned}$$

Posterior Full Conditional Distributions

$$\begin{aligned} \pi(\lambda|\mathbf{y},k,\phi) &\sim Ga(\alpha+t_1(\mathbf{y},k),\beta+k) \\ \pi(\phi|\mathbf{y},k,\lambda) &\sim Ga(\gamma+t_2(\mathbf{y},k),\delta+n-k) \\ Pr(k=i|\mathbf{y},\phi,\lambda) &\propto \lambda^{t_1(\mathbf{y},i)}\phi^{t_2(\mathbf{y},i)}e^{-(i\lambda+(n-i)\phi)} \end{aligned}$$

Metropolis-Hastings

Metropolis et. al. (1953), Hastings (1970)
Like the Gibbs sampling, it is an algorithm that generates a sequence

$$\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots\},$$

from a Markov chain whose stationary distribution is $\pi(\theta)$.

1. initial value $\theta^{(0)}$.
2. proposed value $\xi \sim q(\xi|\theta^{(i-1)})$
3. accepted value:

$$\theta^{(i)} = \begin{cases} \xi & \text{with probability } \alpha \\ \theta^{(i-1)} & \text{with probability } 1-\alpha \end{cases}$$

where

$$\alpha = \min \left\{ 1, \frac{\frac{\pi(\xi)}{q(\xi|\theta^{(i-1)})}}{\frac{\pi(\theta^{(i-1)})}{q(\theta^{(i-1)}|\xi)}} \right\}$$

Special Cases

1. Symmetric Chains: $q(\theta|\xi) = q(\xi|\theta)$

$$\alpha = \min\left\{1, \frac{\pi(\xi)}{\pi(\theta)}\right\}$$

2. Random Walk: $q(\theta|\xi) = q(|\theta - \xi|)$

$$\alpha = \min\left\{1, \frac{\pi(\xi)}{\pi(\theta)}\right\}$$

3. Independent Chains: $q(\theta|\xi) = q(\theta)$

$$\alpha = \min\left\{1, \frac{\omega(\xi)}{\omega(\theta)}\right\}$$

where $\omega(\xi) = \pi(\xi)/q(\xi)$.

Gibbs $\subset$ Metropolis

The Gibbs sampling is equivalent to a composition of p Metropolis-Hastings algorithms, whose acceptance probabilities are always equals 1.

$$\frac{\pi(\theta_1^{(j)}, \ldots, \theta_{i-1}^{(j)}, \xi, \theta_{i+1}^{(j-1)}, \ldots, \theta_p^{(j-1)})}{\pi(\xi | \theta_1^{(j)}, \ldots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j-1)}, \ldots, \theta_p^{(j-1)})}$$

equals

$$\pi(\theta_1^{(j)}, \ldots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j-1)}, \ldots, \theta_p^{(j-1)})$$

Example: Mixture of Normals

Let,

$$\mathbf{X} \sim 0.7N(\mu_1, \Sigma_1) + 0.3N(\mu_2, \Sigma_2)$$

where

$$\mu_1 = \begin{pmatrix} 4 \\ 5 \end{pmatrix} \; \mu_2 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$

and

$$\Sigma_1 = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix} \; \Sigma_2 = \begin{pmatrix} 1 & -0.7 \\ -0.7 & 1 \end{pmatrix}$$

Assume one wants to use

$$q(\mathbf{X}^{(i)}|\mathbf{X}^{(i-1)}) \sim N(\mathbf{X}^{(i-1)}, \mathbf{V})$$

as the proposal of the Metropolis-Hastings algorithm.

JAGS

Just Another Gibbs Sampler

`http://mcmc-jags.sourceforge.net/`

It is a program for analysis of Bayesian hierarchical models using Markov Chain Monte Carlo (MCMC) simulation not wholly unlike BUGS. JAGS was written with three aims in mind:

- To have a cross-platform engine for the BUGS language
- To be extensible, allowing users to write their own functions, distributions and samplers
- To be a plaftorm for experimentation with ideas in Bayesian modelling

NIMBLE

Numerical Inference for statistical Models for Bayesian and Likelihood Estimation

`https://r-nimble.org/`

NIMBLE is a system for building and sharing analysis methods for statistical models, especially for hierarchical models and computationally-intensive methods. NIMBLE is built in R but compiles your models and algorithms using C++ for speed. It includes three components:

- A system for using models written in the BUGS model language as programmable objects in R.
- An initial library of algorithms for models written in BUGS, including basic MCMC, which can be used directly or can be customized from R before being compiled and run.
- A language embedded in R for programming algorithms for models, both of which are compiled through C++ code and loaded into R.

NIMBLE can also be used without BUGS models as a way to compile simple R-like code into C++, which is then compiled and loaded into R with an interface function or object.

Stan[1]

`http://mc-stan.org/`

`Stan` is a state-of-the-art platform for statistical modeling and high-performance statistical computation.
Users specify log density functions in Stan's probabilistic programming language and get:

- full Bayesian statistical inference with MCMC sampling (NUTS, HMC)
- approximate Bayesian inference with variational inference (ADVI)
- penalized maximum likelihood estimation with optimization (L-BFGS)
- Stan's math library provides differentiable probability functions & linear algebra (C++ autodiff)
- Additional R packages provide expression-based linear modeling, posterior visualization, and leave-one-out cross-validation
- It does not allow inference for discrete parameters

[1] Named after Stanislaw Ulam, one of the developers of Monte Carlo methods in the 1940s

INLA

`http://www.r-inla.org/`

Based on the Integrated Nested Laplace Approach (Rue et al., 2009)

- Makes use of an integrated nested Laplace approximation and its simplified version
- Provides accurate approximations to the posterior *marginals*
- The main benefit of these approximations is computational: where Markov chain Monte Carlo algorithms need hours or days to run, INLA provides more precise estimates in seconds or minutes

INLA

- Consider a model with parameters θ_1 that are assigned normal priors, with the remaining parameters being denoted θ_2 with $G = dim(\theta_1)$ and $V = dim(\theta_2)$
- For ease of explanation, assume $\theta_1 \sim N_G(\mathbf{0}, \Sigma)$ where Σ depends on elements in θ_2
- The posterior is proportional to

$$\begin{aligned}
\pi(\theta_1, \theta_2 \mid \mathbf{y}) &\propto \pi(\theta_1 \mid \theta_2)\pi(\theta_2)\prod_{i=1}^{n} p(\mathbf{y}_i, \mid \theta_1, \theta_2) \\
&\propto \pi(\theta_2) \mid \Sigma(\theta_2) \mid^{-1/2} \exp\left\{ -\frac{1}{2}\theta_1^T \Sigma(\theta_2)^{-1}\theta_1 \right. \\
&\qquad \left. + \sum_{i=1}^{n} \log p(\mathbf{y}_i \mid \theta_1, \theta_2) \right\} \qquad (3)
\end{aligned}$$

- Of particular interest are the posterior univariate marginal distributions $\pi(\theta_{1g} \mid \mathbf{y})$, $g = 1, \cdots, G$, and $\pi(\theta_{2v} \mid \mathbf{y})$, $v = 1, 2, \cdots, V$

INLA

- The normal parameters θ_1 are dealt with by analytical approximations (as applied to the term in the exponent of (3), conditional on specific values of θ_2)
- Numerical integration techniques are applied to θ_2, so that V should not be too large for accurate inference
- For elements of θ_1 we write

$$\pi(\theta_{1g} \mid \mathbf{y}) = \int \pi(\theta_1 \mid \theta_2, \mathbf{y}) \times \pi(\theta_2 \mid \mathbf{y}) d\theta_2$$

which may be evaluated via the approximation

$$\begin{aligned} \tilde{\pi}(\theta_{1g} \mid \mathbf{y}) &= \int \tilde{\pi}(\theta_{1g} \mid \theta_2, \mathbf{y}) \times \tilde{\pi}(\theta_2 \mid \mathbf{y}) d\theta_2 \\ &\approx \sum_{k=1}^{K} \tilde{\pi}(\theta_{1g} \mid \theta_2, \mathbf{y}) \times \tilde{\pi}(\theta_2 \mid \mathbf{y}) \Delta_k, \end{aligned}$$

for a set of weights Δ_k, $k = 1, 2 \cdots, K$. Laplace or related analytical approximations are applied to carry out the integration (over $\theta_{1g'}$, $g' \neq g$) required for evaluation of $\tilde{\pi}(\theta_{1g} \mid \theta_2, \mathbf{y})$.

INLA

- To produce the grid of points $\{\theta_2^{(k)}, k = 1, 2, \cdots, K\}$ which numerical integration is performed, the mode of $\tilde{\pi}(\theta_2 \mid \mathbf{y})$ is located and the Hessian is approximated, from which the grid of points $\{\theta_2^{(k)}, k = 1, 2, \cdots, K\}$, with associated weights Δ_k, is created and used in the approximation above

- The output of INLA consists of posterior *marginal* distributions, which can be summarized via means, variances, and quantiles

Budworms example

- The larvae of the tobacco budworm causes much damages to crops in the US and Central and South America. (Bad for the farmers, good for public health?!)
- A study is conducted to investigate the dose of drug needed to kill the adult moths. Six different doses were applied to 20 male and 20 female moths. The number knocked down (uncoordinated) or dead 72 hours after exposure was recorded.
- We fit a Bayesian logistic regression on dose.

Budworms example

$$
\begin{aligned}
Y_i &\sim Binom(n_i, p_i) \\
\log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 \text{dose}_i + \beta_2 \text{male}_i
\end{aligned}
$$

- Bayesian model is complete after assigning a prior distribution for $\beta = (\beta_0, \beta_1, \beta_2)$
- Regardless of the prior distribution you assign, the posterior does not have a closed analytical form
- if one assumes that components of β are independent, each normally distributed with some known mean c and variance C, we have that

$$
\begin{aligned}
p(\beta_0, \beta_1, \beta_2 \mid \mathbf{y}) &\propto l(\mathbf{y} \mid \beta) p(\beta) \\
&\propto \prod_{j=1}^{k} p_j^{y_j} (1-p_j)^{(n_j - y_j)} \prod_{i=0}^{2} \exp\left\{ -\frac{1}{2C} (\beta_i - c)^2 \right\}
\end{aligned}
$$

See file `BudwormsEg-bino-Bayesian.r`

Part II, Section 5: Hierarchical Models

Hierarchical Priors

A good strategy to specify the prior distribution or to describe better the experimental situation, is to divide it into stages or into a hierarchy (Lindley & Smith, 1972). The prior specification is made in two phases:

1. structural, for the division into stages;
2. subjective, for quantitative specification at each stage.

Example: Let $\mathbf{Y} = (Y_1, \cdots, Y_n)$ with $Y_i \sim N(\theta_i, \sigma^2)$, with σ^2 known. There are many choices for specification of the prior for $\theta = (\theta_1, \cdots, \theta_n)$. Some options:

Hierarchical Priors

- θ_i's are independent $\Rightarrow p(\theta) = \prod_i p(\theta_i)$.
- θ_i's are a sample from a population with $p(\theta|\lambda)$ where λ contains the parameters describing the population
  $\Rightarrow p(\theta|\lambda) = \prod_{i=1}^{n} p(\theta_i \mid \lambda)$
  This specification corresponds to the first stage. To complete the prior setting, it is necessary to specify the second stage: the distribution of λ, $p(\lambda)$.
  Note that $p(\lambda)$ corresponds to the second stage and **does not** depend on the first stage. Recall that

$$p(\theta) = \int p(\theta, \lambda) d\lambda = \int p(\theta \mid \lambda) p(\lambda) d\lambda$$

$$= \int \prod_{i=1}^{n} p(\theta_i \mid \lambda) p(\lambda) d\lambda.$$

  Note that θ_i's are supposed exchangeable.

Hierarchical Priors

Since the distribution of λ is independent of the first stage, it can be stated as:

1. Concentrated: $p(\lambda = \lambda) = 1$
2. Discrete: $p(\lambda = \lambda_j) = p_j$, $j = 1, \cdots, k$, with $\sum_j p_j = 1$. In this case the distribution of θ will be a finite mixture of the densities $p(\theta \mid \lambda_j)$ with weights p_j, $j = 1, \cdots, k$.
3. Continuous: as before, the distribution of θ will be a continuous mixture of $p(\theta \mid \lambda)$ with weights given by $p(\lambda)$.

In the example we can assume: $\theta_i \sim N(\mu, \tau^2), i = 1, 2, \cdots, n$ then $\lambda = (\mu, \tau^2)$. Assuming $p(\tau^2 = \tau_0) = 1$ and μ normally distributed then θ has a multivariate normal distribution.

OR assuming $p(\mu = \mu_0) = 1$ and τ^{-2} with a gamma prior distribution implies that θ has a multivariate Student t distribution.

Hierarchical Priors

Nothing prevents these ideas from going further into the hierarchy. For example, the distribution of λ can depend on ϕ, in this case,

$$p(\theta) = \int_{\Phi} \int_{\Lambda} p(\theta|\lambda)\, p(\lambda \mid \phi) p(\phi) d\lambda\, d\phi.$$

The parameters λ and ϕ are called hyperparameters and are introduced to ease the prior specification. In practice, it is very hard to interpret the parameters of third or higher stages, so it is common practice to use a non-informative prior for these levels.

A three-stage hierarchical model

Consider

$$\mathbf{y}_i = \mathbf{x}_i\beta + \mathbf{z}_i\mathbf{b}_i + \varepsilon_i$$

with $\mathbf{b}_i$ and ε_i independent and distributed as $\mathbf{b}_i \mid \mathbf{D} \sim N_{q+1}(\mathbf{0}, \mathbf{D})$, and $\varepsilon_i \mid \sigma_\varepsilon^2 \sim N_{n_i}(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$, $i = 1, 2, \cdots, m$.

- The second stage assumption for $\mathbf{b}_i$ can be motivated using the concept of exchangeability. Then it can be shown that

$$p(\mathbf{b}_1, \cdots, \mathbf{b}_m) = \int \prod_{i=1}^{m} p(\mathbf{b}_i \mid \phi)\pi(\phi)d\phi$$

  so that $[\mathbf{b}_1, \cdots, \mathbf{b}_m]$ are *conditionally* independent, given *hyperparameters* ϕ, with the hyperparameters having a distribution known as *hyperprior*.

- Hence, we have a two-stage hierarchical model

$$\begin{aligned} \mathbf{b}_i \mid \phi &\sim_{\text{i.i.d}} p(\cdot \mid \phi) \\ \phi &\sim_{\text{i.i.d}} \pi(\cdot) \end{aligned}$$

A three-stage hierarchical model

- Parametric choices for $p(\cdot \mid \phi)$ and $\pi(\cdot)$ are based on the application, though computational convenience may also be a consideration. In general, if collections of units cluster due to an observed covariate that we believe will influence $\mathbf{b}_i$, then our prior should reflect this.
- This framework contrasts with the sampling theory approach in which the random effects are assumed to be a random sample from a hypothetical *infinite* population.

The three-stage model is:

- Stage one: Likelihood:

$$p(\mathbf{y}_i \mid \beta, \mathbf{b}_i, \sigma^2_\varepsilon) \qquad i = 1, 2, \cdots, m$$

- Stage two: Random Effects prior:

$$\mathbf{b}_i \mid \phi \sim_{\text{i.i.d}} p(\cdot \mid \phi)$$

- Stage three: Hyperprior

$$p(\beta, \mathbf{D}, \sigma^2_\varepsilon).$$

Hyperpriors : It is common to assume independent priors:

$$p(\beta, \mathbf{D}, \sigma^2_\varepsilon) = \pi(\beta)\pi(\mathbf{D})\pi(\sigma^2_\varepsilon).$$

A three-stage hierarchical model

Hyperpriors :

It is common to assume independent priors:

$$p(\beta, \mathbf{D}, \sigma_\varepsilon^2) = \pi(\beta)\pi(\mathbf{D})\pi(\sigma_\varepsilon^2).$$

- A multivariate normal distribution for β and an inverse gamma distribution for σ_ε^2 are often reasonable choices, since they are flexible
- These choices also lead to conditional distributions that have convenient forms for Gibbs sampling
- The prior specification of $\mathbf{D}$ is less straightforward
  - If $\mathbf{D}$ is a diagonal matrix with elements σ_k^2, $k = 0, 1, \cdots, q$ then an obvious choice is

$$\pi(\sigma_0^2, \cdots, \sigma_q^2) = \prod_{k=0}^{q} \text{IGa}(a_k, b_k)$$

    IGa(a_k, b_k) inverse gamma distribution with known parameters a_k and b_k
  - A prior for non-diagonal $\mathbf{D}$ is more troublesome; there are $(q+2)(q+1)/2$ elements, with the restriction that the matrix is positive definite $\rightarrow$ natural choice is an Inverse Wishart distribution

Inverse Wishart Distribution

Suppose $\mathbf{Z}_1, \cdots, \mathbf{Z}_r \sim_{i.i.d} N_p(\mathbf{0}, \mathbf{S})$, with $\mathbf{S}$ a non-singular variance-covariance matrix, and let

$$\mathbf{W} = \sum_{j=1}^{r} \mathbf{Z}_j \mathbf{Z}_j^T.$$

Then $\mathbf{W}$ follows a Wishart distribution, denoted $W(r, S)$, with probability density function

$$p(\mathbf{w}) = c^{-1} |\mathbf{w}|^{(r-p-1)/2} \exp\left\{ -\frac{1}{2} tr(\mathbf{w}\mathbf{S}^{-1}) \right\}$$

where $c = 2^{rp/2} \Gamma_p(r/2) |\mathbf{S}|^{r/2}$ with $\Gamma_p(r/2) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma[(r+1-j)/2]$, the generalized gamma function. We require $r > p-1$ for a proper density. The mean is

$$E[\mathbf{W}] = r\mathbf{S}$$

Taking $p = 1$ yields a Gamma distribution with parameters $r/2$ and $1/(2S)$. Further, taking $S = 1$ gives a χ^2_r r.v.

Inverse Wishart Distribution

- If $\mathbf{W} \sim W(r, \mathbf{S})$, the distribution of $\mathbf{D} = \mathbf{W}^{-1}$ is known as the inverse Wishart distribution, denoted $InvW(r, \mathbf{S})$, with density

$$p(\mathbf{d}) = c^{-1} |\mathbf{d}|^{-(r+p+1)/2} \exp\left\{ -\frac{1}{2} tr(\mathbf{d}^{-1}\mathbf{S}) \right\}$$

  where c is like before.
- $E(\mathbf{D}) = \frac{\mathbf{S}^{-1}}{r-p-1}$ and is defined for $r > p+1$. If $p = 1$ we recover the inverse gamma distribution, $IGa(r/2, 1/2S)$, with $E(D) = \frac{1}{S(r-2)}$, and $Var(D) = \frac{1}{S^2(r-2)(r-4)}$, so that small values of r gives a more dispersed distribution (which is true for general p).
- One way of thinking about prior specification is to imagine that the prior data for the precision consists of observing r multivariate normal r.v. with empirical covariance matrices $\mathbf{R} = \mathbf{S}^{-1}$.
- We summarize samples from the Wishart via marginal distributions for σ_0, σ_1, and ρ since these are more interpretable.
- Example $\mathbf{D}^{-1} \sim W_2(r, \mathbf{R}^{-1})$, $r = 4$, $E[\mathbf{D}] = \frac{\mathbf{R}}{4-1-2} = \mathbf{R}$ with $\mathbf{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

A three-stage hierarchical model - Implementation

For simplicity, assume $\mathbf{x}_i = \mathbf{z}_i$. It is convenient to reparametrize in terms of $[\beta_1, \beta_2, \cdots, \beta_m, \tau, \beta, \mathbf{W}]$, where $\beta_i = \beta + \mathbf{b}_i$, $\tau = \sigma_\varepsilon^2$, and $\mathbf{W} = \mathbf{D}^{-1}$. The joint posterior is proportional to

$$p(\beta_1, \beta_2, \cdots, \beta_m, \tau, \beta, \mathbf{W} \mid \mathbf{y}) \propto \prod_{i=1}^{m} [p(\mathbf{y_i} \mid \beta_i, \tau) p(\beta_i \mid \beta, \mathbf{W})] \pi(\beta)\pi(\tau)\pi(\mathbf{W}) \quad (4)$$

with priors

$$\beta \sim N_{q+1}(\beta_0, \mathbf{V}_0),\ \tau \sim Ga(a_0, b_0),\ \mathbf{W} \sim W_{q+1}(r, \mathbf{R}^{-1})$$

- Marginal distributions, and summaries of these distributions are not available in closed form.
- Possibilities:
  - INLA (`http://www.r-inla.org/`) is ideally suited to the LMM
  - MCMC is an alternative, which we describe next.

A three-stage hierarchical model - MCMC Implementation

Posterior full conditional distributions

- $\beta \mid \beta_1, \cdots \beta_m, \mathbf{W} \sim$
  $N_{q+1}\left[(m\mathbf{W}+\mathbf{V}_0^{-1})\left(\mathbf{W}\sum_{i=1}^m \beta_i + \mathbf{V}_0^{-1}\beta_0\right), \left(m\mathbf{W}+\mathbf{V}_0^{-1}\right)^{-1}\right]$
- $\tau \mid \beta_i, \mathbf{y} \sim Ga\left[a_0 + \frac{\sum_{i=1}^m n_i}{2}, b_0 + \frac{1}{2}\sum_{i=1}^m (\mathbf{y}_i - \mathbf{x}_i\beta_i)^T(\mathbf{y}_i - \mathbf{x}_i\beta_i)\right]$
- $\beta_i \mid \tau, \mathbf{W}, \mathbf{y} \sim N_{q+1}\left[(\tau\mathbf{x}_i^T\mathbf{x}_i + \mathbf{W})^{-1}(\tau\mathbf{x}_i^T\mathbf{y}_i + \mathbf{W}\beta), (\tau\mathbf{x}_i^T\mathbf{x}_i + \mathbf{W})^{-1}\right]$
- $\mathbf{W} \mid \beta_1, \cdots, \beta_m, \beta \sim W_{q+1}\left[r+m, \left(\mathbf{R} + \sum_{i=1}^m (\beta_i - \beta)(\beta_i - \beta)^T\right)^{-1}\right]$

  Note that $E[\mathbf{D} \mid \beta_1, \cdots, \beta_m, \beta] = \frac{\mathbf{R} + \sum_{i=1}^m \left(\beta_i - \beta)(\beta_i - \beta\right)^T}{r+m-q-2}$, suggesting that it is better to pick a small $\mathbf{R}$, since a large $\mathbf{R}$ will always dominate the sum of squares. If m is small the prior is always influential.

Example - dental growth data

- Table 1 records dental measurements of the distance in millimeters from the center of the pituitary gland to the pteryo-maxillary fissure in 11 girls and 16 boys at the ages of 8, 10, 12 and 14 years.
- Here we have an example of repeated measures or longitudinal data.
- Figure 1 plots these data and we see that dental growth for each child increases in an approximately linear fashion.
- One common aim of such studies is to identify the within-individual and between-individual sources of variability.

Example - dental growth data

Girls	8	10	12	14
1	21	20	21.5	23
2	21	21.5	24	25.5
3	20.5	24	24.5	26
4	23.5	24.5	25	26.5
5	21.5	23	22.5	23.5
6	20	21	21	22.5
7	21.5	22.5	23	25
8	23	23	23.5	24
9	20	21	22	21.5
10	16.5	19	19	19.5
11	24.5	25	28	28
Boys	8	10	12	14
1	26	25	29	31
2	21.5	22.5	23	26.5
3	23	22.5	24	27.5
4	25.5	27.5	26.5	27
5	20	23.5	22.5	26
6	24.5	25.5	27	28.5
7	22	22	24.5	26.5
8	24	21.5	24.5	25.5
9	23	20.5	31	26
10	27.5	28	31	31.5
11	23	23	23.5	25
12	21.5	23.5	24	28
13	17	24.5	26	29.5
14	22.5	25.5	25.5	26
15	23	24.5	26	30
16	22	21.5	23.5	25

Table: Dental growth data for girls and boys.

Example - dental growth data

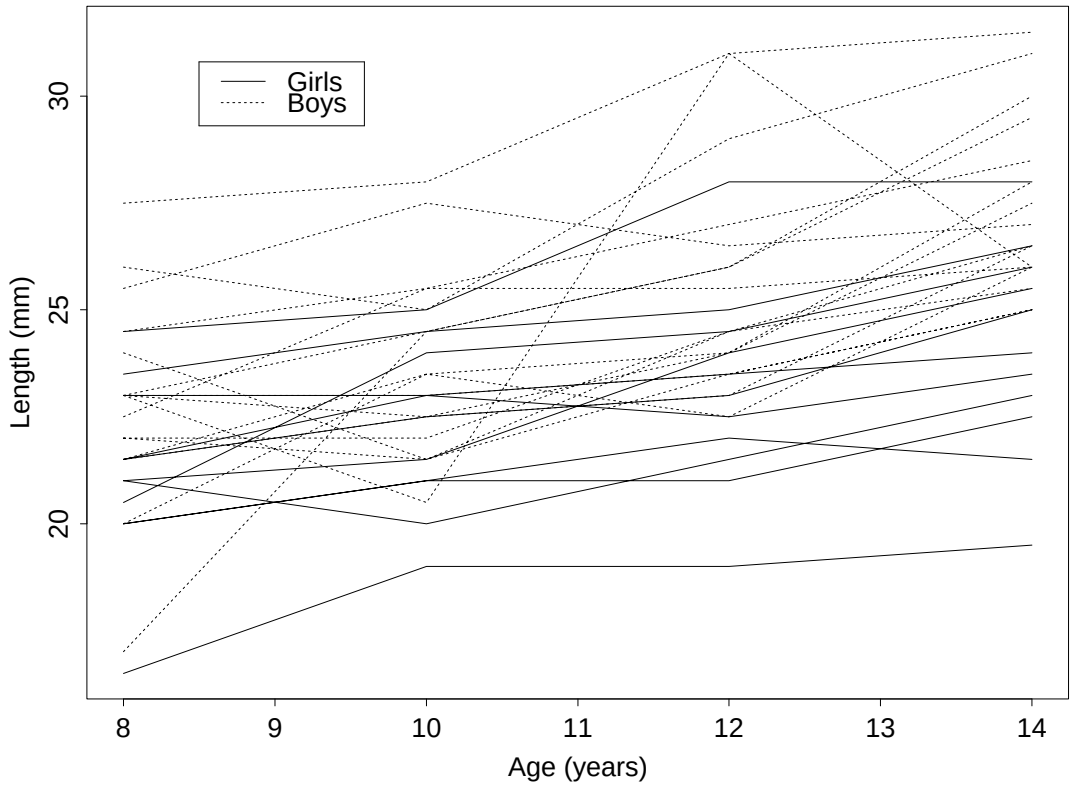

Figure: Dental growth data for girls and boys.

A three-stage hierarchical model - Dental Growth Example

The three-stage model is:

- Stage one: Likelihood:

$$y_{ij} = \beta_{i0} + \beta_{1i} t_j + \varepsilon_{ij}, \quad \text{with } \varepsilon_{ij} \mid \tau \sim_{iid} N(0, \tau^{-1})\, i = 1, 2, \cdots, m$$

- Stage two: Random Effects prior:

$$\beta_i \mid \beta, \mathbf{D} \sim_{\text{i.i.d}} N_2(\beta, \mathbf{D})$$

$$\beta_i = \begin{pmatrix} \beta_{i0} \\ \beta_{i1} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \mathbf{D} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix}$$

- Stage three: Hyperprior

$$p(\beta, \mathbf{D}, \tau) = \pi(\tau) \times \pi(\beta) \times \pi(\mathbf{D})$$

If we assume improper priors for β and τ, we have that

$$p(\beta, \mathbf{D}, \tau) \propto \tau^{-1} \pi(\mathbf{D})$$

with $D^{-1} \sim W_2(r, \mathbf{R}^{-1})$

A three-stage hierarchical model - Dental Growth Example

See file `OrthogirlsRandomInterceptRandomSlopeModel.r`