Bayesian Inference for the LMEM

Consider the model

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

with $\boldsymbol{b}_i \sim_{iid} \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$, $\boldsymbol{\epsilon}_i \sim_{ind} \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_{n_i}\sigma^2_\epsilon)$, with $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ independent.

The form of the posterior follows from exploiting conditional independencies:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b} \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b})\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}) = \prod_{i=1}^m p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}_i)\pi(\boldsymbol{b} \mid \boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\alpha}) \\ &= \prod_{i=1}^m \left\{p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}_i)\pi(\boldsymbol{b}_i \mid \boldsymbol{\alpha})\right\} \pi(\boldsymbol{\beta})\pi(\boldsymbol{\alpha}) \qquad (24) \end{aligned}$$

Alternatively, we can derive the posterior for $\boldsymbol{\beta}, \boldsymbol{\alpha}$ directly:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\alpha} \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha})\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{i=1}^m p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha})\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &= \prod_{i=1}^m \int p(\boldsymbol{y}_i, \boldsymbol{b}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha})\, d\boldsymbol{b}_i \pi(\boldsymbol{\beta}, \boldsymbol{\alpha}) \end{aligned}$$

where the integrand is giving by the term in curly brackets in (24).

The prior on $\boldsymbol{b}_i$ is justified by the context, formally via **exchangeability**.

Exchangeability

Definition: A finite set $Y_1, ..., Y_n$ of random variables is said to be *exchangeable* if every permutation $(Y_1, ..., Y_n)$ has the same joint distribution as every other permutation. An infinite collection is exchangeable if every finite subcollection is exchangeable.

Every collection of independent and identically distributed random variables is exchangeable.

Theorem: *De Finetti's representation Theorem for 0/1 random variables.*

If $Y_1, Y_2, ...$ is an infinitely exchangeable sequence of 0/1 random variables, there exists a distribution $\pi(\cdot)$ such that the joint mass function $\Pr(y_1, ..., y_n)$ has the form

$$\Pr(y_1, ..., y_n) = \int_0^1 \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}\pi(\theta)\, \mathrm{d}\theta,$$

where

$$\int_0^\theta \pi(u)\, \mathrm{d}u = \lim_{n\to\infty} \Pr\left(\frac{Z_n}{n} \leq \theta\right),$$

with $Z_n = Y_1 + ... + Y_n$, and $\theta = \lim_{n\to\infty} Z_n/n$.

Proof: See Bernardo and Smith (1994) for more details.

Let $z_n = y_1 + ... + y_n$ be the number of 1's (which we label "successes") in the first n observations. Then, due to exchangeability,

$$\Pr(y_1 + ... + y_n = z_n) = \left(\begin{array}{c} n \\ z_n \end{array} \right) \Pr(Y_{\pi(1)}, ..., Y_{\pi(n)}),$$

for all permutations π of $\{1, ..., n\}$ such that $y_{\pi(1)} + ... + y_{\pi(n)} = z_n$. Then we can embed the event $y_1 + ... + y_n = z_n$ within a sequence, $y_1, ..., y_N$, $N \geq n$, and

$$\begin{aligned} \Pr\left(\sum_{i=1}^{n} y_i = z_n\right) &= \sum_{Z_N = z_n}^{N-(n-z_n)} \Pr(y_1 + ... + y_n = z_n, y_1 + ... + y_N = z_N) \\ &= \sum_{z_N = z_n}^{N-(n-z_n)} \Pr(y_1 + ... + y_n = z_n \mid y_1 + ... + y_N = z_N) \\ &\times \Pr(y_1 + ... + y_N = z_N). \end{aligned}$$

To obtain the conditional probability we observe that it is as if we have a population of N people of which z_N are successes, and $N - z_N$ failures, from which we draw n people, the probability of z_n successes is then hypergeometric.

Hence

$$\Pr(y_1 + ... + y_n = z_n) = \sum_{z_N = z_n}^{N-(n-z_n)} \frac{\left(\begin{array}{c} z_N \\ z_n \end{array} \right) \left(\begin{array}{c} N - z_N \\ n - z_n \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} \Pr(z_N)$$

Here $\Pr(z_N)$ is the "prior" belief in the number of successes out of N.

Let $N \to \infty$ and by the strong law of law numbers $\theta = \lim_{N \to \infty} z_N/N$.

The hypergeometric tends to a binomial with parameters n and θ, and the prior $\Pr(z_N)$ is translated into a prior for θ, $\pi(\theta)$. Hence we have

$$\Pr(y_1 + ... + y_n = z_n) \to \left(\begin{array}{c} n \\ z_n \end{array} \right) \int \theta^{z_n} (1 - \theta)^{n - z_n} \pi(\theta) \, \mathrm{d}\theta,$$

as $N \to \infty$.

Implications

The interpretation of this theorem is of great significance:

- We may view the Y_i to be independent, Bernoulli random variables, conditional on a random variable θ.
- θ is itself assigned a probability distribution $\pi()$.
- π may be interpreted as "beliefs about the limiting relative frequency of 1's".

In conventional language, we have the *likelihood function*

$$p(Y_1, ..., Y_n|\theta) = \prod_{i=1}^{n} p(Y_i|\theta) = \prod_{i=1}^{n} \theta^{Y_i}(1-\theta)^{1-Y_i},$$

where the *parameter* θ is assigned a *prior distribution* $\pi(\theta)$.

Further results

General Representation Theorem:

If $Y_1, Y_2, ...$ is an infinitely exchangeable sequence of random variables with probability measure P, there exists a distribution function Q such that the joint mass function $p(Y_1, ..., Y_n)$ has the form

$$p(Y_1, ..., Y_n) = \int \prod_{i=1}^{n} p(Y_i|\boldsymbol{\theta})\pi(\theta)\mathrm{d}\boldsymbol{\theta},$$

with $p(\cdot|\boldsymbol{\theta})$ denoting the density function corresponding to the 'unknown parameter' $\boldsymbol{\theta}$.

Further assumptions on $Y_1, Y_2, ...$ are required to identify $p(\cdot|\boldsymbol{\theta})$.

Relevance of Exchangeability

If we believe *a priori* that $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_m$ are exchangeable (and are considered within a hypothetical infinite sequence of such random variables), then it can be shown using representation theorems that the prior can be written in the form

$$p(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_m) = \int \prod_{i=1}^{m} p(\boldsymbol{\theta}_i|\boldsymbol{\phi})\pi(\boldsymbol{\phi}) \, \mathrm{d}\boldsymbol{\phi},$$

that is, they are conditionally independent, given *hyperparameters* $\boldsymbol{\phi}$, with the hyperparameters having a *hyperprior* distribution.

Hence we have a two-stage (hierarchical) prior:

Stage A: $\boldsymbol{\theta}_i|\boldsymbol{\phi} \sim_{iid} p(\cdot|\boldsymbol{\phi})$, $i = 1, ..., m$.

Stage B: $\boldsymbol{\phi} \sim_{iid} \pi(\cdot)$.

Parametric choices for $p(\cdot|\boldsymbol{\phi})$ and $\pi(\cdot)$ are usually made to balance flexibility and computational convenience.

Contrast with the sampling theory approach in which the random effects are assumed to be a random sample from a hypothetical infinite population.

Bayesian Computation

We have seen that to summarize posterior distributions integration is required and, in all but the simplest (conjugate) models, these integrals are not analytically tractable.

Integration is also required to integrate out the random effects in nonlinear mixed effects models, to obtain the likelihood, and later we will review a number of analytical and numerical approaches, for now we concentrate on Markov chain Monte Carlo (MCMC).

The first key idea is the duality between densities and samples from that density: given a density we can always generate samples, and given samples we can reconstruct the density.

Simulation-based techniques have revolutionized Bayesian statistics, by allowing the fitting of very complex models.

Example: Binomial Likelihood with Weird Functions of Interest

Suppose we have

$$Y_j \mid p_j \sim \text{Binomial}(n_j, p_j)$$

$j = 1, 2$, with independent priors

$$p_j \sim U(0, 1)$$

The posteriors are available analytically as

$$p_j \mid y_j \sim \text{Beta}(y_j + 1, n_j - y_j + 1)$$

but suppose we are interested in inference for the odds ratio

$$\phi = \frac{p_1}{1 - p_1} / \frac{p_2}{1 - p_2}$$

and for the relative risk

$$\theta = \frac{p_1}{p_2}$$

The following is `R code` to simulate from

$$p_1 \mid y_1, y_2 \text{ and } p_2 \mid y_1, y_2$$

and

$$\phi \mid y_1, y_2 \text{ and } \theta \mid y_1, y_2$$

when

$$n_1 = 35, n_2 = 45, y_1 = 30, y_2 = 10$$

```
> n1 <- 35; n2 <- 45; y1 <- 30; y2 <- 10
> nsamp <- 1000
> p1 <- rbeta(nsamp,y1+1,n1-y1+1); p2 <- rbeta(nsamp,y2+1,n2-y2+1)
> odds <- (p1/(1-p1))/(p2/(1-p2));  rr <- p1/p2
> par(mfrow=c(2,2))
> hist(p1,xlim=c(0,1))
> hist(p2,xlim=c(0,1))
> hist(odds)
> hist(rr)
> sum(odds[odds>10])/sum(odds) # Posterior prob that odds ratio is > than 10
[1] 0.945683
```

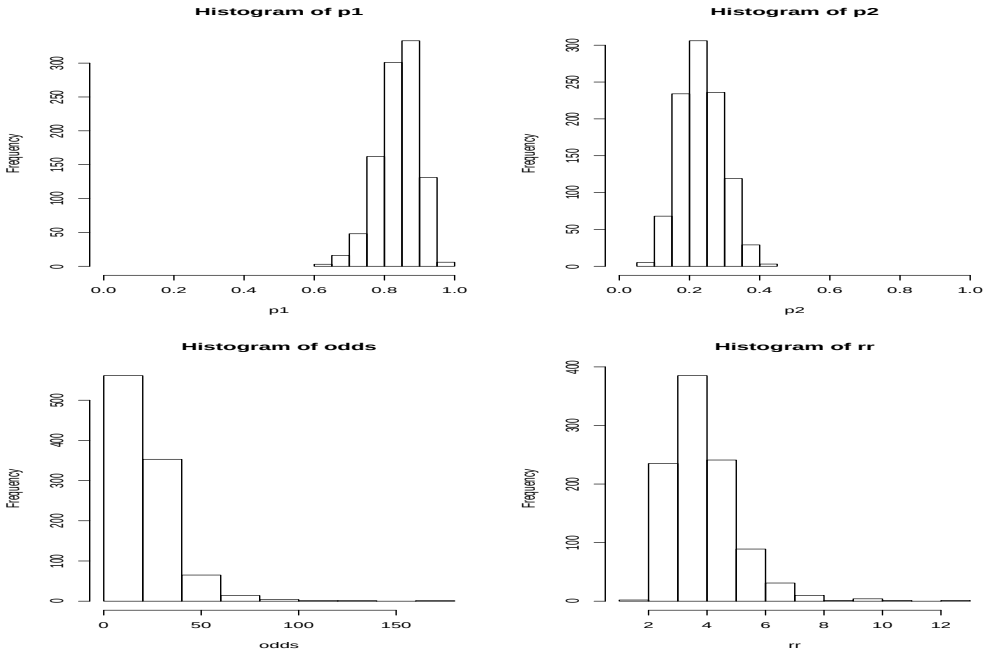

Figure 7: Posterior distributions for p_1, p_2, the odds ratio $\frac{p_1}{1-p_1}/\frac{p_2}{1-p_2}$ and for the relative risk $\theta = \frac{p_1}{p_2}$.

The Composition Method

A useful technical for simulating from joint posterior distributions is the following.

Write the joint posterior distribution for θ_1, θ_2 as

$$p(\theta_1, \theta_2 \mid \boldsymbol{y}) = p(\theta_1 \mid \boldsymbol{y}) p(\theta_2 \mid \theta_1, \boldsymbol{y})$$

Then a simulating algorithm to produce independent samples from $p(\theta_1, \theta_2 \mid \boldsymbol{y})$ is, for $s = 1, ..., S$:

1. Simulate $\theta_1^{(s)} \sim_{ind} p(\theta_1 \mid \boldsymbol{y})$.
2. Simulate $\theta_2^{(s)} \sim_{ind} p(\theta_2 \mid \theta_1^{(s)}, \boldsymbol{y})$.

Markov chain Monte Carlo

MCMC is a very general technique that has revolutionized practical Bayesian statistics.

In the usual derivation of Markov chains over a discrete sample space we are given a transition matrix and the aim is to find the stationary distribution (if it exists). Probabilities of movement depend on the current state only, hence the name.

In the context of sampling from a distribution $\pi(\cdot)$, the aim is to construct a Markov chain whose stationary distribution is π.

Samples $\boldsymbol{\theta}^{(s)}$, $s = 1, ..., S$, produced by a Markov chain "look" more and more like *dependent* samples from π as $S \rightarrow \infty$. The dependency does not cause a problem in terms of estimation since

$$\frac{1}{S}\sum_{s=1}^{S} f(\boldsymbol{\theta}^{(s)}) \rightarrow \mathrm{E}\{f(\boldsymbol{\theta})\},$$

as $S \rightarrow \infty$ (provided the expectation exists).

The only difficulty with the dependency is establishing an appropriates Monte Carlo error on the resultant estimator. The Gibbs sampler, and the Metropolis-Hastings algorithm are common strategies.

Gibbs Sampling

Consider a two-parameter problem in which the (intractable) posterior is:

$$\pi(\theta_1, \theta_2|\boldsymbol{y}) \propto l(\theta_1, \theta_2) \times \pi(\theta_1, \theta_2).$$

We have

$$\pi(\theta_1, \theta_2|\boldsymbol{y}) = p(\theta_1|\boldsymbol{y}) \times p(\theta_2|\theta_2, \boldsymbol{y}),$$

but $p(\theta_1|\boldsymbol{y})$ will typically be unavailable.

Gibbs sampling proceeds by iterating between the steps:

$$\theta_1^{(s)} \sim p(\theta_1|\theta_2^{(s-1)}, \boldsymbol{y}),$$

and

$$\theta_2^{(s)} \sim p(\theta_2|\theta_1^{(s)}, \boldsymbol{y}),$$

to produce the sequence

$$(\theta_1^{(0)}, \theta_2^{(0)}), (\theta_1^{(1)}, \theta_2^{(1)}), ..., (\theta_1^{(s)}, \theta_2^{(s)}), ...$$

which may be viewed as a draw from $\pi(\theta_1, \theta_2|\boldsymbol{y})$

Example: Normal likelihood, unknown mean and variance

Likelihood:

$$Y_i|\boldsymbol{\beta}, \sigma^2 \sim N(\boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2), i = 1, ..., n.$$

Prior:

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{V}), \sigma^{-2} \sim \mathrm{Ga}(a, b).$$

Posterior

$$\pi(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto l(\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta})\pi(\sigma^2),$$

is intractable unless $p(\boldsymbol{\beta})$ is improper uniform and the prior for σ^2 is inverse gamma.

Gibbs sampling iterates between $\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2$ and $\sigma^{-2}|\boldsymbol{y}, \boldsymbol{\beta}$ where

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2) &\propto l(\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}) \\ &\sim N(\boldsymbol{\mu}^*, \boldsymbol{V}^*), \\ p(\sigma^{-2}|\boldsymbol{y}, \boldsymbol{\beta}) &\propto l(\boldsymbol{\beta}, \sigma^2)\pi(\sigma^{-2}) \\ &\sim \mathrm{Ga}\left(a + \frac{n}{2}, b + \frac{(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})}{2}\right). \end{aligned}$$

where

$$\boldsymbol{\mu}^* = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\sigma^{-2} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{V}^{-1})^{-1}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\hat{\boldsymbol{\beta}}\sigma^{-2} + \boldsymbol{\mu}\boldsymbol{V}^{-1}),$$

and

$$\boldsymbol{V}^* = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\sigma^{-2} + \boldsymbol{V}^{-1})^{-1}.$$

Metropolis-Hastings Algorithm

Generalizes the Metropolis algorithm to allow a non-symmetric proposal density.

Suppose $\boldsymbol{\theta}^{(0)}$ denotes the initial point. The Metropolis-Hastings algorithm then consists of, at iteration s:

- Sample $\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s-1)} \sim g(\cdot|\boldsymbol{\theta}^{(s-1)})$.
- Calculate
$$r = \frac{\pi(\boldsymbol{\theta}^*)/g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s-1)})}{\pi(\boldsymbol{\theta}^{(s-1)})/g(\boldsymbol{\theta}^{(s-1)}|\boldsymbol{\theta}^*)}.$$
- Set
$$\boldsymbol{\theta}^{(s)} = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \min(r,1), \\ \boldsymbol{\theta}^{(s-1)} & \text{otherwise.} \end{cases}$$

Important point: the calculation of r does not depend on the normalizing constant of the target density π.

Issues:

- Convergence of the Markov chain?
- Parameterization.

Convergence

- Early iterations $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(m)}$ reflect the (arbitrary) starting value $\theta^{(0)}$.
- These iterations are called the *burn-in*.
- Chain will gradually 'forget' its initial state and converge to the unique stationary distribution which is independent of $\theta^{(0)}$.
- Burn-in samples should be ignored when summarizing the samples for posterior inference via Monte Carlo integration, i.e.
$$E[g(\theta)] \quad \approx \quad \frac{1}{n-m} \sum_{s=m+1}^{n} g(\theta^{(s)})$$

Convergence Diagnosis

- Strictly speaking, convergence is only achieved for $n = \infty$.
- But we only need Markov chain to be 'approaching' convergence for Monte Carlo integration to yield a consistent estimate of the true expectation.
- How do we determine m, the number of 'burn-in' iterations?
- Informal examination of time series plots and running of multiple chains is a must.
- Two issues: have we 'found' the posterior? Do we have enough samples to answer the inferential questions? Some chains may be very slow mixing (examination of autocorrelation is important).

Parameterization

The Markov chain will display better mixing properties if the parameters are approximately independent in the posterior.

In an extreme case, if we have independence then

$$p(\theta_1, ..., \theta_k|\boldsymbol{y}) = \prod_{i=1}^{k} p(\theta_i|\boldsymbol{y}),$$

and Gibbs sampling via the conditional distributions $p(\theta_i|\boldsymbol{y}), i = 1, ..., n$, is equivalent to direct sampling from the posterior.

In general it is better to sample 'blocks' of parameters that are approximately independent.

Hyperpriors

Consider the LMEM

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

with $\boldsymbol{b}_i \sim \mathrm{N}_{q+1}(\boldsymbol{0}, \boldsymbol{D})$, and $\boldsymbol{\epsilon}_i \sim \mathrm{N}_{n_i}(\boldsymbol{0}, \sigma^2_\epsilon \boldsymbol{I}_{n_i})$, $i = 1, ..., m$. A Bayesian analysis requires prior distributions on $\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2_\epsilon$; it is common to assume independent priors

$$\pi(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2_\epsilon) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{D})\pi(\sigma^2_\epsilon).$$

For $\boldsymbol{\beta}$ a multivariate normal distribution and for σ^2_ϵ an inverse gamma distribution are often specified since they lead to conditional distributions of convenient form for Gibbs sampling, but other choices are possible.

If $\boldsymbol{D}$ is a diagonal matrix with elements σ^2_k, $k = 0, 1, ..., q$, then a prior that leads to conjugate conditional distributions in a Gibbs sampling algorithm is

$$\pi(\sigma^2_0, ..., \sigma^2_q) = \prod_{k=0}^{q} \mathrm{IGa}(a_k, b_k),$$

where $\mathrm{IGa}(a_k, b_k)$ denotes the inverse gamma distribution with pre-specified parameters a_k, b_k, $k = 0, ..., q$.

The Wishart Distribution

A prior for a non-diagonal $\boldsymbol{D}$ is more troublesome; there are $(q+2)(q+1)/2$ elements, with the restriction that the resultant matrix is positive definite.

The inverse Wishart distribution is the conjugate choice, and is the only distribution for which any great practical experience has been gained.

Suppose $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_r \sim_{iid} \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{S})$, with $\boldsymbol{S}$ a non-singular variance-covariance matrix, and let

$$\boldsymbol{W} = \sum_{j=1}^{r} \boldsymbol{Z}_j \boldsymbol{Z}_j^{\mathrm{T}}. \tag{25}$$

Then $\boldsymbol{W}$ follows a Wishart distribution, denoted $\mathrm{W}_p(r, \boldsymbol{S})$, and

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

where

$$c = 2^{rp/2}\Gamma_p(r/2) \mid \boldsymbol{S} \mid^{r/2}, \tag{26}$$

with

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

the generalized gamma function, and $r \geq p$ for a proper density.

The mean is given by

$$\mathrm{E}[\boldsymbol{W}] = r\boldsymbol{S}.$$

The Wishart distribution is a multivariate version of the gamma distribution. Taking $p = 1$ yields

$$p(w) = \frac{(2S)^{-r/2}}{\Gamma(r/2)} w^{r/2-1} \exp(-w/2S),$$

for $w > 0$, the gamma distribution $\mathrm{Ga}(r/2, 1/(2S))$. Further, taking $S = 1$ gives a χ^2_r random variable, which is clear from (25).

The Inverse Wishart Distribution

If $\boldsymbol{W} \sim \mathrm{W}_p(r, \boldsymbol{S})$, the distribution of $\boldsymbol{D} = \boldsymbol{W}^{-1}$ is known as the inverse Wishart distribution, and is given by

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

where c is again given by (26). The mean is given by

$$\mathrm{E}[\boldsymbol{D}] = \frac{\boldsymbol{S}^{-1}}{r-p-1}$$

and is defined for $r > p + 1$. If $p = 1$ we recover the inverse gamma distribution $\mathrm{IGa}(r/2, 1/2S)$ with $\mathrm{E}[D] = 1/[s(r-2)]$ and $\mathrm{var}(D) = 1/[S^2(r-2)(r-4)]$ (so that small r gives a larger spread).

Thinking ahead to application in the LMEM if $\boldsymbol{W} \sim \mathrm{W}_{q+1}(r, \boldsymbol{R}^{-1})$, then

$$\mathrm{E}[\boldsymbol{W}] = r\boldsymbol{R}^{-1},$$

and

$$\mathrm{E}[\boldsymbol{D}] = \boldsymbol{R}/(r-q-1-1),$$

so that $\boldsymbol{R}$, may be scaled to be a prior estimate of $\boldsymbol{D}$, with r acting as a strength of belief in the prior.

Issues with the Wishart Prior

- A problem with the Wishart distribution is that it is deficient in second moment parameters since there is only a single degrees of freedom parameter r. So, for example, it is not possible to have differing levels of certainty in the tightness of the prior distribution for different elements of $\boldsymbol{D}$. With diagonal $\boldsymbol{D}$ and independent inverse gamma priors we have a precision parameter for each variance.
- The form of the conditional distribution suggests that it may be better to err on the side of picking $\boldsymbol{R}$ too small (if m small, prior always influential).
- Intuition: as if our prior data for the precision consists of observing r normal random variables with variance-covariance matrices $\boldsymbol{R}$.
- We need to take $r \geq q+1$ for a proper prior, with the flattest prior corresponding to $r = q+1$. A proper prior is required to ensure propriety of the posterior distribution.
- Figure 8 displays samples from the Wishart distribution $W_2\{20, (20\boldsymbol{S})^{-1}\}$ where $\boldsymbol{S} = \begin{bmatrix} 0.4 & 0 \\ 0 & 1.0 \end{bmatrix}$. The mean is $\mathrm{E}[\boldsymbol{W}] = \boldsymbol{S}^{-1} = \begin{bmatrix} 2.5 & 0 \\ 0 & 1.0 \end{bmatrix}$.

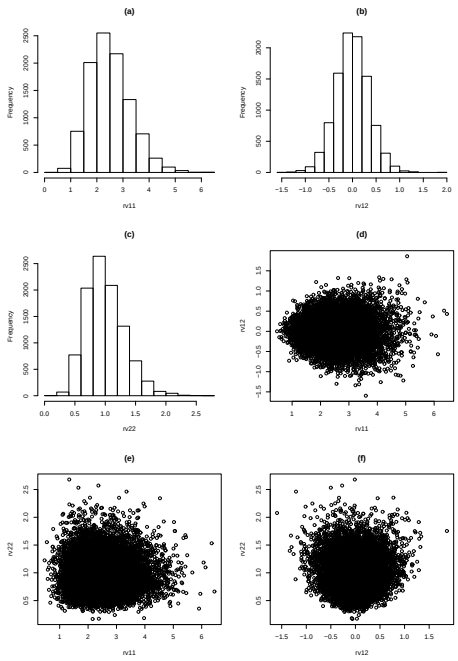

Figure 8: Histograms of (a) w_{11}, (b) w_{12}, (c) w_{22}, scatterplots of (d) w_{11}, w_{12}, (e) w_{11}, w_{22}, w_{12}, w_{22}

Conditional Conjugacy

We now consider a Gibbs sampling scheme and assume for simplicity that $\boldsymbol{x}_i = \boldsymbol{z}_i$. It is computational more convenient to reparameterize in terms of the set $\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \tau, \boldsymbol{\beta}, \boldsymbol{W}\}$ where $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{b}_i$, $\tau = \sigma_\epsilon^{-2}$, $\boldsymbol{W} = \boldsymbol{D}^{-1}$.

The joint posterior is

$$p(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \tau, \boldsymbol{\beta}, \boldsymbol{W} \mid \boldsymbol{y}) \propto \prod_{i=1}^{m} \{p(\boldsymbol{y}_i \mid \boldsymbol{\beta}_i, \tau) p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{W})\} \pi(\boldsymbol{\beta}) \pi(\tau) \pi(\boldsymbol{W}),$$

with priors:

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathrm{N}_{q+1}(\boldsymbol{\beta}_0, \boldsymbol{V}_0) \\ \tau &\sim \mathrm{Ga}(a_0, b_0) \\ \boldsymbol{W} &\sim \mathrm{W}_{q+1}(r, \boldsymbol{R}^{-1}) \end{aligned}$$

and derive the required conditional distributions:

- $p(\boldsymbol{\beta} \mid \tau, \boldsymbol{W}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{y})$
- $p(\tau \mid \boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{y})$
- $p(\boldsymbol{W} \mid \boldsymbol{\beta}, \tau, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{y})$
- $p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \tau, \boldsymbol{W}, \boldsymbol{y})$, $i = 1, ..., m$.

Conditional for $\boldsymbol{\beta}$

$$\boldsymbol{\beta} \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{W} \sim \mathrm{N}_{q+1}\left\{ \left(m\boldsymbol{W} + \boldsymbol{V}_0^{-1}\right)^{-1} \left(\boldsymbol{W} \sum_{i=1}^{m} \boldsymbol{\beta}_i + \boldsymbol{V}_0^{-1} \boldsymbol{\beta}_0 \right), \left(m\boldsymbol{W} + \boldsymbol{V}_0^{-1}\right)^{-1} \right\}$$

Conditional for τ

$$\tau \mid \boldsymbol{\beta}_i, \boldsymbol{y} \sim \mathrm{Ga}\left(a_0 + \frac{\sum_{i=1}^{m} n_i}{2}, b_0 + \frac{1}{2} \sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta}_i)^{\mathrm{T}} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta}_i) \right)$$

Conditional for $\boldsymbol{\beta}_i$

$$\boldsymbol{\beta}_i \mid \tau, \boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{y} \sim \mathrm{N}_{q+1}\left\{ (\tau \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i + \boldsymbol{W})^{-1} (\tau \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{y}_i + \boldsymbol{W}\boldsymbol{\beta}), (\tau \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i + \boldsymbol{W})^{-1} \right\}$$

Note the way that the conditional independencies have been exploited so that in each case we condition on only a subset of the parameters.

Conditional for $\boldsymbol{W}$

First note that

$$(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{\beta}_i - \boldsymbol{\beta}) = \mathrm{tr}((\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{\beta}_i - \boldsymbol{\beta})) = \mathrm{tr}(\boldsymbol{W} (\boldsymbol{\beta}_i - \boldsymbol{\beta})(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}}).$$

Then

$$\begin{aligned}
\boldsymbol{W} \mid \boldsymbol{y}, \boldsymbol{\beta}_i, \boldsymbol{\beta} &\propto \prod_{i=1}^{m} p(\boldsymbol{\beta}_i \mid \boldsymbol{W}) \times \pi(\boldsymbol{W}) \\
&\propto |\boldsymbol{W}|^{(m+r-q-1-1)/2} \exp\left\{ -\frac{1}{2} \left[\sum_{i=1}^{m} (\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{\beta}_i - \boldsymbol{\beta}) + \mathrm{tr}(\boldsymbol{W}\boldsymbol{R}) \right] \right\} \\
&= |\boldsymbol{W}|^{(m+r-q-1-1)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr} \left(\boldsymbol{W} \left[\sum_{i=1}^{m} (\boldsymbol{\beta}_i - \boldsymbol{\beta})(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}} + \boldsymbol{R} \right] \right) \right\}
\end{aligned}$$

Hence the conditional distribution is

$$\boldsymbol{W} \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{\beta}, \boldsymbol{y} \sim \mathrm{W}_{q+1} \left\{ r + m, \left(\boldsymbol{R} + \sum_{i=1}^{m} (\boldsymbol{\beta}_i - \boldsymbol{\beta})(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}} \right)^{-1} \right\}.$$

Example: Dental Data for Girls

Three-Stage Hierarchical Model:

First Stage:

$$y_{ij} = \beta_{0i} + \beta_{1i}(t_j - 11) + \epsilon_{ij},$$

with $\epsilon_{iid} \sim \mathrm{N}(0, \tau^{-1})$, $j = 1, ..., 4$, $i = 1, ..., 11$.

Second Stage: Let

$$\boldsymbol{\beta}_i = \begin{bmatrix} \beta_{0i} \\ \beta_{1i} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \boldsymbol{D} = \begin{bmatrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{bmatrix},$$

and then

$$\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D} \sim \mathrm{N}_2(\boldsymbol{\beta}, \boldsymbol{D}),$$

$i = 1, ..., m$.

Third Stage:

$$\pi(\tau, \boldsymbol{\beta}, \boldsymbol{D}^{-1}) \propto \mathrm{Ga}(0, 0) \times \mathrm{N}_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10^6 & 0 \\ 0 & 10^6 \end{bmatrix} \right) \times W_2(r, \boldsymbol{R}^{-1}).$$

Results below are for priors, with prior mean

$$\mathrm{E}[\boldsymbol{D}] = \frac{1}{r-q-2}\boldsymbol{R} = \frac{1}{r-3}\boldsymbol{R} = \begin{bmatrix} 1.0 & 0 \\ 0 & 0.1 \end{bmatrix}$$

(since $q = 1$) and different degrees of freedom r.

We see sensitivity to the prior in inference for $\boldsymbol{D}$, but not for $\boldsymbol{\beta}$.

Note the greater shrinkage to the prior mean for the second and third priors.

r	$\boldsymbol{R}$	β_0	β_1
4	1.0 0 0 0.1	22.6 (21.4,23.8)	0.48 (0.33,0.63)
7	4.0 0 0 0.4	22.6 (21.5,23.7)	0.48 (0.31,0.65)
28	25 0 0 2.5	22.6 (21.8,23.5)	0.48 (0.28,0.67)

Table 4: Posterior medians and 95% intervals for population means, under three priors.

r	Diag $\boldsymbol{R}$	D_{00}	D_{01}	D_{11}
4	1.0 0.1	3.48 (1.66, 8.75)	0.13 (-0.10,0.54)	0.03 (0.01,0.10)
7	4.0 0.4	2.97 (1.51, 6.63)	0.10 (-0.14,0.46)	0.05 (0.02,0.12)
28	25 2.5	1.78 (1.14, 2.97)	0.04 (-0.10,0.20)	0.08 (0.05,0.14)

Table 5: Posterior medians and 95% intervals for population variances, under two priors.

The code below is for the analysis with $r = 4$, BUGS parametrizes the Wishart in terms of $\boldsymbol{R}^{-1}$ and r.

```
model
{
for( i in 1 : N ) {
    for( j in 1 : T ) {
        Y[i , j] ~ dnorm(mu[i , j],eps.tau)
        mu[i , j] <- beta[i,1] + beta[i,2] * (x[j]-11)
    }
    beta[i,1:2] ~ dmnorm(beta.mu[1:2],iSigma[1:2,1:2])
}
beta.mu[1:2] ~ dmnorm(mean[1:2], prec[1:2, 1:2])
iSigma[1:2, 1:2] ~ dwish(R[1:2, 1:2], r)
Sigma[1:2, 1:2] <- inverse(iSigma[1:2, 1:2])
eps.tau <- exp(logtau)
logtau ~ dflat()
sigma <- 1 / sqrt(eps.tau)
}
```



```
list(x = c(8,10,12,14), N = 11, T = 4,
Y = structure(
.Data =   c(21,20,21.5,23,
21,21.5,24,25.5,
20.5,24,24.5,26,
23.5,24.5,25,26.5,
21.5,23,22.5,23.5,
20,21,21,22.5,
21.5,22.5,23,25,
23,23,23.5,24,
20,21,22,21.5,
16.5,19,19,19.5,
24.5,25,28,28),
.Dim = c(11,4)),mean = c(0, 0),r=4,
R = structure(.Data = c(1, 0, 0, 0.1),
.Dim = c(2, 2)),
prec = structure(.Data = c(1.0E-6, 0,0,1.0E-6),
.Dim = c(2, 2))))
list(beta = structure(.Data = c(18,.5,18,.5,18,.5,18,.5,18,.5,18,.5,18,.5,18,
.5,18,.5,18,.5,18,.5), .Dim=c(11,2)), beta.mu = c(18,.5),
     iSigma = structure(.Data = c(1, 0, 0, 0.1), .Dim = c(2, 2)), logtau = 0)
```