SDS 383D: Modeling II **Spring 2017**

Section 2: Bayesian inference in Gaussian models

Student: Natalia Zuniga-Garcia

2.1 Bayesian inference in a simple Gaussian model

Let's start with a simple, one-dimensional Gaussian example, where

$$y_i|\mu, \sigma^2 \sim \mathrm{N}(\mu, \sigma^2).$$

We will assume that μ and σ are unknown, and will put conjugate priors on them both, so that

$$\begin{aligned}\sigma^2 &\sim \text{Inv-Gamma}(\alpha_0, \beta_0)\\ \mu|\sigma^2 &\sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)\end{aligned}$$

or, equivalently,

$$\begin{aligned}y_i|\mu, \omega &\sim \mathrm{N}(\mu, 1/\omega)\\ \omega &\sim \text{Gamma}(\alpha_0, \beta_0)\\ \mu|\omega &\sim \text{Normal}\left(\mu_0, \frac{1}{\omega\kappa_0}\right)\end{aligned}$$

We refer to this as a normal/inverse gamma prior on μ and σ^2 (or a normal/gamma prior on μ and ω). We will now explore the posterior distributions on μ and $\omega(/\sigma^2)$ – much of this will involve similar results to those obtained in the first set of exercises.

Exercise 2.1 *Derive the conditional posterior distributions $p(\mu, \omega|y_1, \ldots, y_n)$ (or $p(\mu, \sigma^2|y_1, \ldots, y_n)$) and show that it is in the same family as $p(\mu, \omega)$. What are the updated parameters α_n, β_n, μ_n and κ_n?*

Solution

We know that,

$$p(\mu, \omega|y_i) \propto p(y_i|\mu, \omega)p(\mu, \omega)$$

First, we find the likelihood $p(y_i|\mu, \omega)$,

$$p(y_i|\mu, \omega) = \left(\frac{\omega}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{\omega}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}$$

Now we find the prior $p(\mu, \omega)$,

$$p(\mu, \omega) = p(\mu|\omega)p(\omega)$$

Where,

$$p(\mu|\omega) = \left(\frac{\omega\kappa_0}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\omega\kappa_0}{2}(\mu - \mu_0)^2\right\}$$

$$p(\omega) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \omega^{\alpha_0 - 1} exp\{-\beta_0 \omega\}$$

Therefore we have,

$$p(\mu, \omega) = \left(\frac{\omega \kappa_0}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\omega \kappa_0}{2}(\mu - \mu_0)^2\right\} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \omega^{\alpha_0 - 1} exp\{-\beta_0 \omega\}$$

$$= \left(\frac{\kappa_0}{2\pi}\right)^{\frac{1}{2}} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \omega^{\alpha_0 - \frac{1}{2}} \exp\left\{-\frac{\omega}{2}[\kappa_0(\mu - \mu_0)^2 + 2\beta_0]\right\}$$

$$\propto \mathrm{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0)$$

Now we can estimate the posterior,

$$p(\mu, \omega | y_i) \propto p(y_i | \mu, \omega) p(\mu, \omega)$$

$$\propto \omega^{\frac{n}{2}} \exp\left\{-\frac{\omega}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right\} \omega^{\alpha_0 - \frac{1}{2}} \exp\left\{-\frac{\omega}{2}[\kappa_0(\mu - \mu_0)^2 + 2\beta_0]\right\}$$

$$\propto \omega^{\frac{1}{2}} \omega^{\alpha_0 + \frac{n}{2} - 1} \exp\{-\beta_0 \gamma\} \exp\left\{-\frac{\omega}{2}[\kappa_0(\mu - \mu_0)^2 + \sum_{i=1}^{n}(y_i - \mu)^2]\right\}$$

Where,

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

and,

$$\kappa_0(\mu - \mu_0)^2 + n(\mu - \overline{y})^2 = (\kappa_0 + n)\left(\mu - \frac{\kappa_0 \mu_0 + n\overline{y}}{\kappa_0 + n}\right)^2 + \frac{\kappa_0 n(\overline{y} - \mu_0)^2}{\kappa_0 + n}$$

Then we have that,

$$\kappa_0(\mu - \mu_0)^2 + \sum_{i=1}^{n}(y_i - \mu)^2 = (\kappa_0 + n)\left(\mu - \frac{\kappa_0 \mu_0 + n\overline{y}}{\kappa_0 + n}\right)^2 + \frac{\kappa_0 n(\overline{y} - \mu_0)^2}{\kappa_0 + n} + \sum_{i=1}^{n}(y_i - \overline{y})^2$$

Now going back to the posterior,

$$p(\mu, \omega | y_i) \propto \omega^{\frac{1}{2}} \omega^{\alpha_0 + \frac{n}{2} - 1} \exp\{-\beta_0 \omega\} \exp\left\{-\frac{1}{2}[\kappa_0(\mu - \mu_0)^2 + \sum_{i=1}^{n}(y_i - \mu)^2]\right\}$$

$$\propto \omega^{\frac{1}{2}} \omega^{\alpha_0 + \frac{n}{2} - 1} \exp\{-\beta_0 \omega\} \exp\left\{-\frac{\omega}{2}\left[(\kappa_0 + n)\left(\mu - \frac{\kappa_0 \mu_0 + n\overline{y}}{\kappa_0 + n}\right)^2 + \frac{\kappa_0 n(\overline{y} - \mu_0)^2}{\kappa_0 + n} + \sum_{i=1}^{n}(y_i - \overline{y})^2\right]\right\}$$

$$\propto \omega^{\frac{1}{2}} \exp\left\{-\frac{\omega}{2}(\kappa_0 + n)\left(\mu - \frac{\kappa_0 \mu_0 + n\overline{y}}{\kappa_0 + n}\right)^2\right\} \omega^{\alpha_0 + \frac{n}{2} - 1} \exp\left\{-\beta_0 \omega - \frac{\omega}{2}\left[\frac{\kappa_0 n(\overline{y} - \mu_0)^2}{\kappa_0 + n} + \sum_{i=1}^{n}(y_i - \overline{y})^2\right]\right\}$$

$$\propto \text{Normal}\left(\frac{\kappa_0 \mu_0 + n\overline{y}}{\kappa_0 + n}, \frac{1}{\omega(\kappa_0 + n)}\right) \text{Gamma}\left(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{1}{2}\left[\frac{\kappa_0 n(\overline{y} - \mu_0)^2}{\kappa_0 + n} + \sum_{i=1}^{n}(y_i - \overline{y})^2\right]\right)$$

$$\propto \mathrm{NG}(\mu_n, \kappa_n, \alpha_n, \beta_n)$$

Thus, the updated parameters are:

- $\alpha_n = \alpha_0 + \frac{n}{2}$
- $\beta_n = \beta_0 + \frac{1}{2}\left[\frac{\kappa_0 n(\overline{y}-\mu_0)^2}{\kappa_0+n} + \sum_{i=1}^n (y_i - \overline{y})^2\right]$
- $\mu_n = \frac{\kappa_0\mu_0+n\overline{y}}{\kappa_0+n}$
- $\kappa_n = \kappa_0 + n$

Exercise 2.2 *Derive the conditional posterior distribution $p(\mu|\omega, y_1, \ldots, y_n)$ and $p(\omega|y_1, \ldots, y_n)$ (or if you'd prefer, $p(\mu|\sigma^2, y_1, \ldots, y_n)$ and $p(\sigma^2|y_1, \ldots, y_n)$). Based on this and the previous exercise, what are reasonable interpretations for the parameters $\mu_0, \kappa_0, \alpha_0$ and β_0?*

Solution

We know that,

$$p(\mu|\omega, y_i) \propto p(y_i|\mu, \omega)p(\mu|\omega)$$

Where,

$$p(\mu|\omega) = \left(\frac{\omega\kappa_0}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\omega\kappa_0}{2}(\mu-\mu_0)^2\right\}$$

$$p(y_i|\mu, \omega) = \left(\frac{\omega}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{\omega}{2}\sum_{i=1}^n (y_i-\mu)^2\right\}$$

Then,

$$p(\mu|\omega, y_i) \propto \left(\frac{\omega}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{\omega}{2}\sum_{i=1}^n (y_i-\mu)^2\right\} \left(\frac{\omega\kappa_0}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\omega\kappa_0}{2}(\mu-\mu_0)^2\right\}$$

$$\propto \exp\left\{-\frac{\omega}{2}[\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^n (y_i-\mu)^2]\right\} \propto \exp\left\{-\frac{\omega}{2}(\kappa_0+n)\left(\mu - \frac{\kappa_0\mu_0+n\overline{y}}{\kappa_0+n}\right)^2\right\}$$

$$\propto \text{Normal}\left(\frac{\kappa_0\mu_0+n\overline{y}}{\kappa_0+n}, \frac{1}{\omega(\kappa_0+n)}\right)$$

Now for ω we have,

$$p(\omega|y_i) \propto p(y_i|\mu, \omega)p(\omega)$$

Where,

$$p(\omega) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\omega^{\alpha_0-1}exp\{-\beta_0\omega\}$$

Then,

$$p(\omega|y_i) \propto \left(\frac{\omega}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{\omega}{2}\sum_{i=1}^n (y_i-\mu)^2\right\} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\omega^{\alpha_0-1}exp\{-\beta_0\omega\}$$

$$\propto \exp\left\{-\beta_0\omega - \frac{\omega}{2}\left[\frac{\kappa_0 n(\overline{y}-\mu_0)^2}{\kappa_0+n} + \sum_{i=1}^n (y_i-\overline{y})^2\right]\right\}$$

$$\propto \text{Gamma}\left(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{1}{2}\left[\frac{\kappa_0 n(\overline{y}-\mu_0)^2}{\kappa_0+n} + \sum_{i=1}^n (y_i-\overline{y})^2\right]\right)$$

What are reasonable interpretations for the parameters $\mu_0, \kappa_0, \alpha_0$ and β_0?

- κ_0 is the "prior sample size" for μ.
- μ_0 is the prior guest on μ and μ_n is the posterior of μ that we can see is a weighted average based on the sample size n and the "prior sample size" κ_0.
- α_0 is the "prior sample size" for the error variance σ^2.
- β_0 is the "prior sum of squares" for the error variance σ^2. For the posterior, we have β_n which combines the "prior sum of squares" β_0, the sample sum of squares $\sum_{i=1}^{n}(y_i - \overline{y})^2$, and a term due to the discrepancy between the prior mean and the sample mean.

Exercise 2.3 *Show that the marginal distribution over μ is a centered, scaled t-distribution (note we showed something very similar in the last set of exercises!), i.e.*

$$p(\mu) \propto \left(1 + \frac{1}{\nu}\frac{(\mu - m)^2}{s^2}\right)^{-\frac{\nu+1}{2}}$$

What are the location parameter m, scale parameter s, and degree of freedom ν?

Solution

We can compute the marginal as follow,

$$p(\mu) \propto \int_0^\infty p(\mu, \omega) d\omega$$

Where,

$$p(\mu, \omega) = \left(\frac{\kappa_0}{2\pi}\right)^{\frac{1}{2}} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \, \omega^{\alpha_0 - \frac{1}{2}} \exp\left\{-\frac{\omega}{2}[\kappa_0(\mu - \mu_0)^2 + 2\beta_0]\right\}$$

Then,

$$p(\mu) \propto \int_0^\infty \omega^{\alpha_0 + \frac{1}{2} - 1} \exp\left\{-\frac{\omega}{2}[\kappa_0(\mu - \mu_0)^2 + 2\beta_0]\right\} d\omega$$

Which is the kernel of a Gamma$\left(\alpha_0 + \frac{1}{2}, \beta_0 + \frac{\kappa_0(\mu-\mu_0)^2}{2}\right)$. Then, we have that the normalization constant of the distribution is,

$$\begin{aligned}
p(\mu) &\propto \frac{\Gamma(\alpha_0 + \frac{1}{2})}{[\beta_0 + \frac{\kappa_0(\mu-\mu_0)^2}{2}]^{\alpha_0 + \frac{1}{2}}} \\
&\propto \left[\beta_0 + \frac{\kappa_0(\mu - \mu_0)^2}{2}\right]^{-\alpha_0 - \frac{1}{2}} \\
&\propto \left[1 + \frac{1}{2\alpha_0}\frac{\alpha_0\kappa_0(\mu - \mu_0)^2}{\beta_0}\right]^{-\frac{(2\alpha_0+1)}{2}}
\end{aligned}$$

Where,

- $m = \mu_0$
- $\nu = 2\alpha_0$

- $s = \sqrt{\frac{\beta_0}{\alpha_0\kappa_0}}$

Exercise 2.4 *The marginal posterior $p(\mu|y_1,\ldots,y_n)$ is also a centered, scaled t-distribution. Find the updated location, scale and degrees of freedom.*

Solution

We have that,

$$p(\mu|y_i) \propto \int_0^\infty p(\mu,\omega|y_i)d\omega$$

$$\propto \int_0^\infty \omega^{\frac{1}{2}} \exp\left\{-\frac{\omega}{2}(\kappa_0+n)\left(\mu - \frac{\kappa_0\mu_0+n\overline{y}}{\kappa_0+n}\right)^2\right\}\omega^{\alpha_0+\frac{n}{2}-1}\exp\left\{-\beta_0\omega - \frac{\omega}{2}\left[\frac{\kappa_0 n(\overline{y}-\mu_0)^2}{\kappa_0+n} + \sum_{i=1}^n (y_i-\overline{y})^2\right]\right\}d\omega$$

Similarly to the previous exercise, we find that,

$$p(\mu|y_i) \propto \left[1 + \frac{1}{2(\alpha_0+\frac{n}{2})}\frac{(\alpha_0+\frac{n}{2})(\kappa_0+n)(\mu-\frac{\kappa_0\mu_0+n\overline{y}}{\kappa_0+n})^2}{\beta_0+\frac{1}{2}\left[\frac{\kappa_0 n(\overline{y}-\mu_0)^2}{\kappa_0+n}+\sum_{i=1}^n(y_i-\overline{y})^2\right]}\right]^{-\frac{(2(\alpha_0+\frac{n}{2})+1)}{2}}$$

Where,

- $m = \frac{\kappa_0\mu_0+n\overline{y}}{\kappa_0+n} = \mu_n$
- $\nu = 2(\alpha_0+\frac{n}{2}) = 2\alpha_n$
- $s = \sqrt{\frac{\beta_0+\frac{1}{2}\left[\frac{\kappa_0 n(\overline{y}-\mu_0)^2}{\kappa_0+n}+\sum_{i=1}^n(y_i-\overline{y})^2\right]}{(\alpha_0+\frac{n}{2})(\kappa_0+n)}} = \sqrt{\frac{\beta_n}{\alpha_n\kappa_n}}$

Exercise 2.5 *Derive the posterior predictive distribution $p(y_{n+1},\ldots,y_{n+m}|y_1,\ldots,y_m)$.*

Solution

The posterior predictive is,

$$p(y_{new_i}|y_i) = \frac{p(y_{new_i},y_i)}{p(y_i)}$$

First we need to find $p(y_i)$, the marginal distribution of y_i. We know that,

$$p(\mu,\omega|y_i) = \frac{p(y_i|\mu,\omega)p(\mu,\omega)}{p(y_i)}$$

$$p(y_i) = \frac{p(y_i|\mu,\omega)p(\mu,\omega)}{p(\mu,\omega|y_i)}$$

Let's denote our prior as $p(\mu,\omega) = \frac{1}{Z_0}\mathrm{NG}(\mu_0,\kappa_0,\alpha_0,\beta_0)$. Where Z_0 is the normalizing constant $Z_0 = \frac{\Gamma(\alpha_0)}{\beta_0^{\alpha_0}}(\frac{2\pi}{\kappa_0})^{\frac{1}{2}}$. Similarly, our posterior is $p(\mu,\omega|y_i) = \frac{1}{Z_n}\mathrm{NG}(\mu_n,\kappa_n,\alpha_n,\beta_n)$. Also, let's our likelihood be represented as $p(y_i|\mu,\omega) = (\frac{1}{2\pi})^{\frac{n}{2}}\prod_i^n \mathrm{N}(y_i|\mu,\omega)$. Then,

$$p(y_i) = \frac{p(y_i|\mu,\omega)p(\mu,\omega)}{p(\mu,\omega|y_i)} = p(y_i|\mu,\omega) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\frac{Z_n\mathrm{NG}(\mu_0,\kappa_0,\alpha_0,\beta_0)\prod_i^n \mathrm{N}(y_i|\mu,\omega)}{Z_0\mathrm{NG}(\mu_n,\kappa_n,\alpha_n,\beta_n)}$$

Previously, we found that $\mathrm{NG}(\mu_n, \kappa_n, \alpha_n, \beta_n) = \mathrm{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0) \prod_i^n \mathrm{N}(y_i|\mu, \omega)$, so:

$$p(y_i) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{Z_n}{Z_0} = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{\Gamma(\alpha_n)\beta_0^{\alpha_0}}{\Gamma(\alpha_0)\beta_n^{\alpha_n}} \left(\frac{\kappa_0}{\kappa_n}\right)^{\frac{1}{2}}$$

Now, we can go back to our posterior predictive,

$$p(y_{new_i}|y_i) = \frac{p(y_{new_i}, y_i)}{p(y_i)} = \frac{Z_{n+m}}{Z_0}\left(\frac{1}{2\pi}\right)^{\frac{n+m}{2}} \frac{Z_0}{Z_n}(2\pi)^{\frac{n}{2}} = \frac{Z_{n+m}}{Z_n}\left(\frac{1}{2\pi}\right)^{\frac{m}{2}}$$

$$p(y_{new_i}|y_i) = \frac{\Gamma(\alpha_{n+m})}{\Gamma(\alpha_n)} \frac{\beta_n^{\alpha_n}}{\beta_{n+m}^{\alpha_{n+m}}} \left(\frac{\kappa_n}{\kappa_{n+m}}\right)^{\frac{1}{2}} \left(\frac{1}{2\pi}\right)^{\frac{m}{2}}$$

Now, for simplicity, let's use the case where $m = 1$. Then we have that $y_1 = \overline{y}$ and $\sum_{i=1}^{1}(y_i - \overline{y})^2 = 0$,

- $\alpha_{n+1} = \alpha_n + \frac{1}{2}$
- $\beta_{n+1} = \beta_n + \frac{1}{2}\frac{\kappa_n(y-\mu_n)^2}{\kappa_n+1}$
- $\kappa_{n+1} = \kappa_n + 1$

$$p(y_{new_i}|y_i) = \frac{\Gamma(\alpha_{n+1})}{\Gamma(\alpha_n)} \frac{\beta_n^{\alpha_n}}{\beta_{n+1}^{\alpha_{n+1}}} \left(\frac{\kappa_n}{\kappa_{n+1}}\right)^{\frac{1}{2}} \left(\frac{1}{2\pi}\right)^{\frac{1}{2}}$$

$$= \frac{\Gamma(\alpha_n + \frac{1}{2})}{\Gamma(\alpha_n)} \frac{\beta_n^{\alpha_n}}{(\beta_n + \frac{1}{2}\frac{\kappa_n(y-\mu_n)^2}{\kappa_n+1})^{\alpha_n+\frac{1}{2}}} \left(\frac{\kappa_n}{\kappa_n+1}\right)^{\frac{1}{2}} \left(\frac{1}{2\pi}\right)^{\frac{1}{2}}$$

$$= \left(\frac{1}{\pi}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{(2\alpha_n+1)}{2})}{\Gamma(\frac{2\alpha_n}{2})} \left(\frac{\alpha_n\kappa_n}{2\alpha_n\beta_n(\kappa_n+1)}\right)^{\frac{1}{2}} \left(1 + \frac{\alpha_n\kappa_n(y-\mu_n)}{2\alpha_n\beta_n(\kappa_n+1)}\right)^{-\frac{(2\alpha_n+1)}{2}}$$

Which is a T-distribution where,

- $m = \mu_n$
- $\nu = 2\alpha_n$
- $s = \sqrt{\frac{\beta_n(\kappa_n+1)}{\alpha_n\kappa_n}}$

Exercise 2.6 *Derive the marginal distribution over $y_1, \ldots, y_n$.*

Solution

We solved this in the previous Exercise,

$$p(y_i) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{\Gamma(\alpha_n)\beta_0^{\alpha_0}}{\Gamma(\alpha_0)\beta_n^{\alpha_n}} \left(\frac{\kappa_0}{\kappa_n}\right)^{\frac{1}{2}}$$

2.2 Bayesian inference in a multivariate Gaussian model

Let's now assume that each y_i is a d-dimensional vector, such that

$$y_i \sim \mathrm{N}(\mu, \Sigma)$$

for d-dimensional mean vector μ and $d \times d$ covariance matrix Σ.

We will put an *inverse Wishart* prior on Σ. The inverse Wishart distribution is a distribution over positive-definite matrices parametrized by $\nu_0 > d-1$ degrees of freedom and positive definite matrix Λ_0^{-1}, with pdf

$$p(\Sigma|\nu_0, \Lambda_0^{-1}) = \frac{|\Lambda|^{\nu_0/2}}{2^{(\nu_0 d)/2}\Gamma_d(\nu_0/2)}|\Sigma|^{-\frac{\nu_0+d+1}{2}}e^{-\frac{1}{2}\mathrm{tr}(\Lambda\Sigma^{-1})}$$

where $\Gamma_d(x) = \pi^{d(d-1)/4}\prod_{i=1}^d \Gamma\left(x - \frac{j-1}{2}\right)$.

Exercise 2.7 *Show that in the univariate case, the inverse Wishart distribution reduces to the inverse gamma distribution.*

Solution

The univariate case has $d = 1$,

$$p(\Sigma|\nu_0, \Lambda_0^{-1}) = \frac{|\Lambda|^{\nu_0/2}}{2^{\nu_0/2}\Gamma_d(\nu_0/2)}|\Sigma|^{-\frac{\nu_0+2}{2}}e^{-\frac{1}{2}\Lambda\Sigma^{-1}}$$

$$= \frac{\left(\frac{|\Lambda|}{2}\right)^{\nu_0/2}}{\Gamma_d(\nu_0/2)}|\Sigma|^{-\frac{\nu_0+2}{2}}e^{-\frac{1}{2}\Lambda\Sigma^{-1}}$$

Which is the inverse Gamma$(\frac{\nu_0}{2}, \frac{\Lambda}{2})$

Exercise 2.8 *Let* $\Sigma \sim$ *Inv-Wishart*(ν_0, Λ_0^{-1}) *and* $\mu|\Sigma \sim N(\mu_0, \Sigma/\kappa_0)$*, so that*

$$p(\mu, \Sigma) \propto |\Sigma|^{-\frac{\nu_0+d+1}{2}}e^{-\frac{1}{2}tr(\Lambda_0\Sigma^{-1})-\frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)}$$

Where,

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

$$p(\mu|\Sigma) = \frac{1}{(2\pi)^{1/2}}|\Sigma/\kappa_0|^{-1/2}\exp\left\{-\frac{\kappa_0}{2}(\mu-\mu_0)^T(\Sigma)^{-1}(\mu-\mu_0)\right\}$$

and let

$$y_i \sim N(\mu, \Sigma)$$

Show that $p(\mu, \Sigma|y_1, \ldots, y_n)$ *is also normal-inverse Wishart distributed, and give the form of the updated parameters* μ_n, κ_n, ν_n *and* Λ_n*. It will be helpful to note that*

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu) =& \sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d}(x_{ij}-\mu_j)(\Sigma^{-1})_{jk}(x_{ik}-\mu_k)\\ =& \sum_{j=1}^{d}\sum_{k=1}^{d}(\Sigma^{-1})_{ab}\sum_{i=1}^{n}(x_{ij}-\mu_j)(x_{ik}-\mu_k)\\ =& tr\left(\Sigma^{-1}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)\end{aligned}$$

Based on this, give interpretations for the prior parameters.

Solution

We know that,

$$p(\mu,\Sigma|y_i) = \frac{p(y_i|\mu,\Sigma)p(\mu,\Sigma)}{p(y_i)} \propto p(y_i|\mu,\Sigma)p(\mu,\Sigma)$$

First, we estimate the likelihood

$$p(y_i|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}}|\Sigma|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T(\Sigma)^{-1}(y_i-\mu)\right\}$$

Previously we had that our prior is a normal-inverse Wishart$(\mu_0,\kappa_0,\nu_0,\Lambda_0)$,

$$p(\mu,\Sigma) \propto |\Sigma|^{-\frac{\nu_0+d+1}{2}}\exp\left\{-\frac{1}{2}\text{tr}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0))\right\}$$

So we can estimate the posterior as,

$$p(\mu,\Sigma|y_i) \propto p(y_i|\mu,\Sigma)p(\mu,\Sigma) \propto$$

$$|\Sigma|^{-\frac{\nu_0+d+1}{2}}\exp\left\{-\frac{1}{2}\text{tr}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0))\right\}|\Sigma|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T(\Sigma)^{-1}(y_i-\mu)\right\}$$

$$\propto |\Sigma|^{-\frac{\nu_0+d+1+n}{2}}\exp\left\{-\frac{1}{2}\text{tr}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0) - \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T(\Sigma)^{-1}(y_i-\mu)\right\}$$

Using,

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

We have,

$$p(\mu,\Sigma|y_i) \propto |\Sigma|^{-\frac{\nu_0+d+1+n}{2}}\exp\left\{-\frac{1}{2}\text{tr}(\Lambda_0\Sigma^{-1}) - tr\left(\Sigma^{-1}\frac{\kappa_0}{2}(\mu-\mu_0)^T(\mu-\mu_0)\right) - \frac{1}{2}tr\left(\Sigma^{-1}\sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)^T\right)\right\}$$

$$p(\mu,\Sigma|y_i) \propto |\Sigma|^{-\frac{\nu_0+d+1+n}{2}}\exp\left\{-\frac{1}{2}\text{tr}\left(\Lambda_0 + \kappa_0(\mu-\mu_0)^T(\mu-\mu_0) + \sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)^T\Sigma^{-1}\right)\right\}$$

$$\propto |\Sigma|^{-\frac{\nu_0+d+1+n}{2}}\exp\left\{-\frac{1}{2}\text{tr}\left(\Lambda_0 + \kappa_0(\mu-\mu_0)^T(\mu-\mu_0) + \sum_{i=1}^{n}(y_i-\overline{y})^T(y_i-\overline{y}) + n(\mu-\overline{y})^T(\mu-\overline{y})\Sigma^{-1}\right)\right\}$$

$$\propto |\Sigma|^{-\frac{\nu_0+n+d+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\Lambda_0 + S + (\kappa_0+n)\,\mu^2 - 2\mu(\kappa_0\mu_0 + n\overline{y}) + \kappa_0\mu_0^2 + n\overline{y}^2\right)\Sigma^{-1}\right\}$$

Where, $S = \sum_{i=1}^{n}(y_i - \overline{y})(y_i - \overline{y})^T$

$$\propto |\Sigma|^{-\frac{\nu_0+n+d+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\Lambda_0 + S + \left(\mu - \frac{\kappa_0\mu_0 + n\overline{y}}{\kappa_0+n}\right)^T\left(\mu - \frac{\kappa_0\mu_0 + n\overline{y}}{\kappa_0+n}\right) + \frac{\kappa_0 n}{\kappa_0+n}(\overline{y}-\mu_0)^T(\overline{y}-\mu_0)\Sigma^{-1}\right)\right\}$$

$$\propto |\Sigma|^{-\frac{\nu_0+n+d+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\overline{y}-\mu_0)^T(\overline{y}-\mu_0)\Sigma^{-1}\right) + \left(\mu - \frac{\kappa_0\mu_0 + n\overline{y}}{\kappa_0+n}\right)^T\Sigma^{-1}\left(\mu - \frac{\kappa_0\mu_0 + n\overline{y}}{\kappa_0+n}\right)\right\}$$

Which is also a normal-inverse Wishart$(\mu_n, \kappa_n, \nu_n, \Lambda_n)$ with,

- $\mu_n = \frac{\kappa_0\mu_0 + n\overline{y}}{\kappa_0+n}$
- $\kappa_n = \kappa_0 + n$
- $\nu_n = \nu_0 + n$
- $\Lambda_n = \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\overline{y}-\mu_0)(\overline{y}-\mu_0)^T$

Give interpretations for the prior parameters.

We can think on the interpretation for the univariate normal we did on Question 2.2:

- κ_0 is the "prior sample size" for μ.
- μ_0 is the prior guest on μ and μ_n is the posterior of μ that we can see is a weighted average based on the sample size n and the "prior sample size" κ_0.
- ν_0 is the "prior sample size" for the covariance Σ.
- Λ_0 is the "prior sum of squares" for the covariance Σ. For the posterior, we have Λ_n which combines the "prior sum of squares" Λ_0, the sample sum of squares $S = \sum_{i=1}^{n}(y_i - \overline{y})^2$, and a term due to the discrepancy between the prior mean and the sample mean.

2.3 A Gaussian linear model

Lets now add in covariates, so that

$$\mathbf{y}|\beta, X \sim \text{Normal}(X\beta, (\omega\Lambda)^{-1})$$

where $\mathbf{y}$ is a vector of n responses; X is a $n \times d$ matrix of covariates; and Λ is a known positive definite matrix. Let's assume $\beta \sim \text{Normal}(\mu, (\omega K)^{-1})$ and $\omega \sim \text{Gamma}(a, b)$, where K is assumed fixed.

Exercise 2.9 *Derive the conditional posterior* $p(\beta|\omega, y_1, \ldots, y_n)$

Solution

We know that,

$$\begin{gathered}
p(\beta|\omega, y_i) \propto p(y_i|\beta, \omega)p(\beta|\omega) \\
p(\omega|y_i) \propto \text{Normal}(X\beta, (\omega\Lambda)^{-1})\text{Normal}(\mu, (\omega K)^{-1}) \\
\propto \exp\left\{-\frac{1}{2}\left((\omega\Lambda)(Y - X\beta)^T(Y - X\beta) + (\omega K)(\beta - \mu)^T(\beta - \mu)\right)\right\} \\
\propto \exp\left\{-\frac{\omega}{2}\left(\Lambda(Y - X\beta)^T(Y - X\beta) + K(\beta - \mu)^T(\beta - \mu)\right)\right\} \\
\propto \exp\left\{-\frac{\omega}{2}\left(Y^T\Lambda Y - 2Y^T\Lambda X\beta + (X\beta)^T X\beta + \beta^T K\beta - 2\right)\right\} \\
\propto \exp\left\{-\frac{1}{2}\left(\omega(K + X^T\Lambda X)\left(\beta - \frac{X^T\Lambda Y + K\mu}{K + X^T\Lambda X}\right)\right)\right\}
\end{gathered}$$

Which is a $\text{Normal}(\mu_n, (\omega K_n)^{-1})$

- $K_n = K + X^T\Lambda X$
- $\mu_n = K^{-1}(X^T\Lambda Y + K\mu)$

Exercise 2.10 *Derive the marginal posterior* $p(\omega|y_1, \ldots, y_n)$

Solution

We know that,

$$\begin{gathered}
p(\omega|y_i) = \int p(\beta|\omega)p(\omega)p(y_i|\beta, \omega)d\beta \\
= p(\omega)\int p(\beta|\omega)p(y_i|\beta, \omega)d\beta
\end{gathered}$$

$$= \omega^{\alpha-1}\exp\{-b\omega\}\int\left(\frac{\omega K}{2\pi}\right)^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\omega(\beta - \mu)^T K(\beta - \mu)\right\}\left(\frac{\omega\Lambda}{2\pi}\right)^{\frac{n}{2}}\exp\left\{-\frac{1}{2}\omega(Y - X\beta)^T\Lambda(Y - X\beta)\right\}d\beta$$

$$= \omega^{\alpha+\frac{1}{2}+\frac{n}{2}-1}\exp\{-b\omega\}\int\exp\left\{-\frac{1}{2}\omega(\beta - \mu)^T K(\beta - \mu) + (Y - X\beta)^T\Lambda(Y - X\beta)\right\}d\beta$$

$$= \omega^{\alpha+\frac{1}{2}+\frac{n}{2}-1}\exp\{-b\omega\}\exp\left\{-\frac{1}{2}\omega(\mu^T K\mu + Y^T\Lambda Y)\right\}\int \exp\left\{-\frac{1}{2}\omega(\beta^T K\beta - 2\mu^T K\beta - 2Y^T\Lambda X\beta + (X\beta)^T\Lambda X\beta)\right\}d\beta$$

$$= \omega^{\alpha+\frac{1}{2}+\frac{n}{2}-1}\exp\left\{-\omega\left(b + -\frac{1}{2}(\mu^T K\mu + Y^T\Lambda Y)\right)\right\}$$

$$\int \exp\left\{-\frac{1}{2}\omega(K + X^T\Lambda X)\left[\left(\beta - \frac{\mu K + Y^T\Lambda X}{K + X^T\Lambda X}\right)^2 - \left(\frac{\mu K + Y^T\Lambda X}{K + X^T\Lambda X}\right)^2\right]\right\}d\beta$$

$$= \omega^{\alpha+\frac{1}{2}+\frac{n}{2}-1}\exp\left\{-\omega\left(b + \frac{1}{2}(\mu^T K\mu + Y^T\Lambda Y) - \frac{1}{2}\mu_n^T(K + X^T\Lambda X)\mu_n\right)\right\}$$

Where, $\mu_n = \frac{\mu K + Y^T\Lambda X}{K + X^T\Lambda X}$

Therefore, we have a Gamma(a_n, b_n) where,

- $a_n = a + \frac{n+1}{2}$
- $b_n = b + \frac{1}{2}\left[(\mu^T K\mu + Y^T\Lambda Y) - \mu_n^T(K + X^T\Lambda X)\mu_n\right]$

Exercise 2.11 *Derive the marginal posterior,* $p(\beta|y_1, \ldots, y_n)$

Solution

$$p(\beta|Y) = \int p(\beta|\omega, y)p(\omega|y)d\omega$$

$$= \int \left(\frac{1}{2\pi}\omega(K + X^T\Lambda X)\right)^{\frac{1}{2}}\exp\left\{-\frac{1}{2}(\beta - \mu_n)^T\omega(K + X^T\Lambda X)(\beta - \mu_n)\right\}\frac{b_n^{a_n}}{\Gamma(a_n)}\omega^{a_n-1}\exp\{-\omega b_n\}d\omega$$

$$= \left(\frac{1}{2\pi}(K + X^T\Lambda X)\right)^{\frac{1}{2}}\frac{b_n^{a_n}}{\Gamma(a_n)}\int \omega^{a_n+\frac{1}{2}-1}\exp\left\{-\omega\left[b + \frac{1}{2}(\beta - \mu_n)^T\omega(K + X^T\Lambda X)(\beta - \mu_n)\right]\right\}d\omega$$

Where the integral is a kernel of the Gamma distribution, then

$$p(\beta|Y) = \left(\frac{1}{2\pi}(K + X^T\Lambda X)\right)^{\frac{1}{2}}\frac{b_n^{a_n}\Gamma(a_n + \frac{1}{2})}{\Gamma(a_n)}\left[b_n + \frac{1}{2}\left((\beta - \mu_n)^T(K + X^T\Lambda X)(\beta - \mu_n)\right)\right]^{-\frac{2a_n+1}{2}}$$

Which is a t-distribution with:

- $m = \mu_n$
- $\nu = 2a_n$
- $s^2 = \frac{b_n}{a_n}K_n^{-1}$

Exercise 2.12 *Download the dataset dental.csv from Github. This dataset measures a dental distance (specifically, the distance between the center of the pituitary to the pterygomaxillary fissure) in 27 children. Add a column of ones to correspond to the intercept. Fit the above Bayesian model to the dataset, using $\Lambda = I$ and $K = I$, and picking vague priors for the hyperparameters, and plot the resulting fit. How does it compare to the frequentist LS and ridge regression results?*

Solution

Using the script *Section2_12.R*,

Table 2.1: Results

	Least Squares	**Ridge $\lambda = 1$**	**Bayesian Linear Model**
Intercept	15.3857	13.9440	15.5579
Age	0.6602	0.7656	0.6497
Sex (Male=1)	2.3210	2.5793	2.3220

```r
# Bayesian Model Description:
# - Lambda = I and K = I
# - Vague prior for the hyperparameters
# - (Y|beta,omega) ~ N(X %*% beta, (omega * Lamda)^{-1})
# - (beta|omega) ~ N(mu, (omega * K)^{-1})
# - omega ~ Gamma(a, b)

# Hyperparameters we need E[omega]=1 and Var[omega]=inf (vague)
Lambda <- diag(rep(1, n)) # matrix nxn
K <- diag(rep(0.01, p)) # precision matriz pxp
mu <- matrix(0, p) # prior guess on mu
a <- 0.01 # prior sample size for the error variance
b <- 0.01 # prior sum of square errors for the error variance

# Updated parameters
# (beta|y,omega) ~ N(mu.new, (omega * K.new)^{-1})
# (omega|y) ~ Gamma(a.new, b.new)

XtLambdaX <- t(X) %*% Lambda %*% X
beta.hat <- solve(XtLambdaX) %*% crossprod(X, Lambda) %*% y
y.hat <- X %*% beta.hat

K.new <- K + XtLambdaX
mu.new <- solve(K.new) %*% (XtLambdaX %*% beta.hat + K %*% mu)
a.new <- a + (n + 1) / 2
s <- t(mu) %*% K %*% mu + t(y.hat) %*% Lambda %*% y.hat
r <- t(mu) %*% (K + XtLambdaX) %*% mu
b.new <- as.numeric(b + 1 / 2 * (s - r))

# Sampling updated parameters
n.iter <- 5000
beta <- matrix(0, n.iter, p)
omega <- rep(0, n.iter)
for (i in 1:n.iter) {
  # omega
  omega.zero <- rgamma(1, a.new, b.new)
  omega[i] <- omega.zero
  # beta
  beta.zero <- mvrnorm(1, mu.new, solve(omega.zero * K.new))
  beta[i,1] <- beta.zero[1]
  beta[i,2] <- beta.zero[2]
  beta[i,3] <- beta.zero[3]
}
```

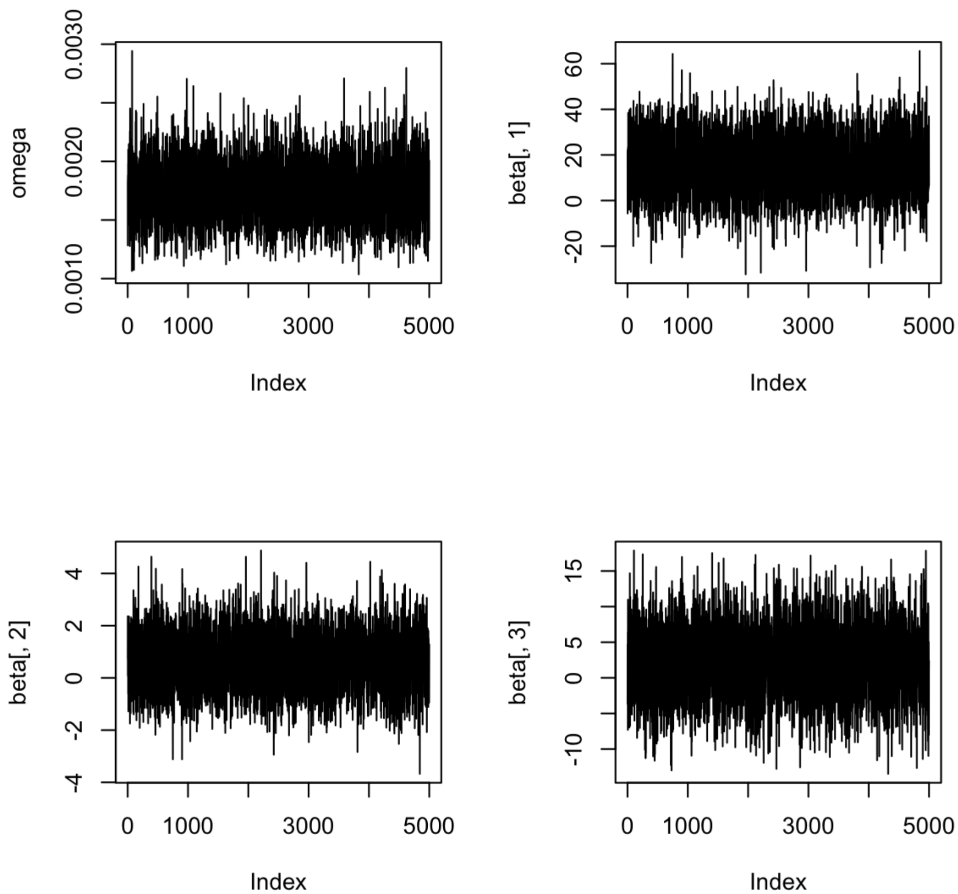
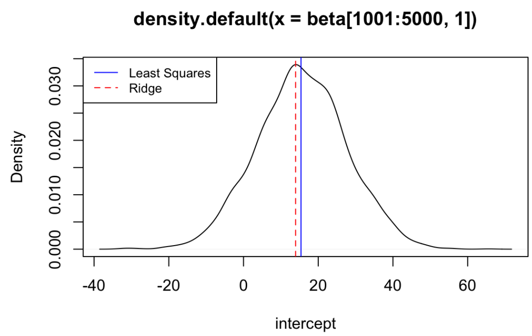
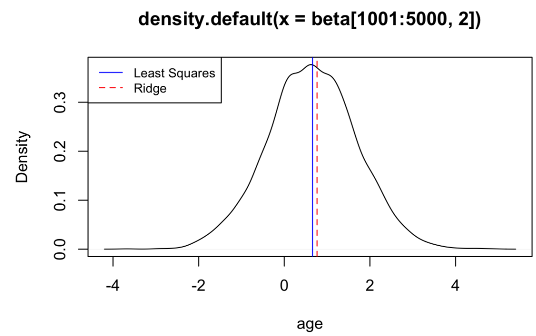


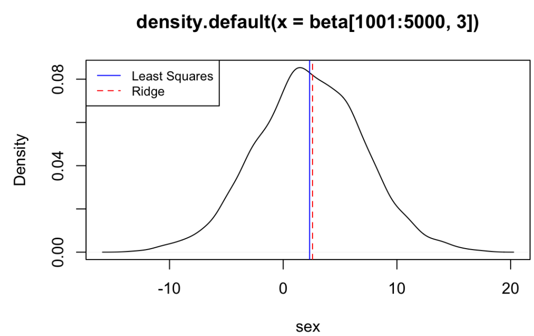
Figure 2.1: Trace plots



(a) Beta Intercept



(b) Beta Age



(c) Beta Sex

Figure 2.2: Resulting fit comparison

2.4 A hierarchical Gaussian linear model

The dental dataset has heavier tailed residuals than we would expect under a Gaussian model. We've seen previously that we can model a scaled t-distribution using a scale mixture of Gaussians; let's put that into effect here. Concretely, let

$$\begin{aligned}
\mathbf{y}|\beta,\omega,\Lambda \sim & \mathrm{N}(X\beta,(\omega\Lambda)^{-1}) \\
\Lambda = & \mathrm{diag}(\lambda_1,\dots,\lambda_n) \\
\lambda_i \overset{iid}{\sim} & \mathrm{Gamma}(\tau,\tau) \\
\beta|\omega \sim & \mathrm{N}(\mu,(\omega K)^{-1}) \\
\omega \sim & \mathrm{Gamma}(a,b)
\end{aligned}$$

Exercise 2.13 *What is the conditional posterior, $p(\lambda_i|\mathbf{y},\beta,\omega)$?*

Solution

$$p(\lambda_i|\mathbf{y},\beta,\omega) \propto p(y|\omega,\beta,\Lambda) \propto p(y|\omega,\beta,\Lambda)p(\Lambda)p(\omega,\beta)$$

Λ is independent of ω and β

$$p(\lambda_i|\mathbf{y},\beta,\omega) \propto \lambda_i^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\omega\lambda_i(y_i - x_i\beta)^2\right\} \lambda_i^{\tau-1} \exp\{-\tau\lambda_i\} \propto \lambda_i^{\tau+\frac{1}{2}-1} \exp\left\{-\lambda_i\left(\frac{1}{2}\omega(y_i - x_i\beta)^2 + \tau\right)\right\}$$

Which is a Gamma$(\tau + \frac{1}{2}, \frac{\omega}{2}(y_i - x_i\beta)^2 + \tau)$

Exercise 2.14 *Write a Gibbs sampler that alternates between sampling from the conditional posteriors of λ_i, β and ω, and run it for a couple of thousand samplers to fit the model to the dental dataset.*

Solution

Using the script *Section2_14.R*,

```
# GIBB SAMPLER
# Hyperparameters we need E[omega]=1 and Var[omega]=inf (vague)
K <- diag(rep(0.1, p)) # precision matriz pxp
mu <- matrix(0, p) # prior guess on mu
a <- 1 # prior sample size for the error variance
b <- 1 # prior sum of square errors for the error variance
tau <- 1

# Sampling updated parameters
n.iter <- 5000
beta <- matrix(NA, n.iter, p)
omega <- matrix(NA, n.iter)
lambda <- matrix(NA, n.iter, n)

beta[1,] <- rep(0, p)
lambda[1,] <- rep(1, n)
omega[1] <- 1

for (i in 2:n.iter) {
  Lambda <- diag(lambda[i-1,])
  XtLambdaX <- t(X) %*% Lambda %*% X
  beta.hat <- solve(XtLambdaX) %*% crossprod(X, Lambda) %*% y
  y.hat <- X %*% beta.hat
  K.new <- K + XtLambdaX
  mu.new <- solve(K.new) %*% (XtLambdaX %*% beta.hat + K %*% mu)

  # Betas
  beta[i,] <- mvrnorm(1, mu.new, solve(omega[i-1] * K.new))

  # omega
  a.new <- a + (n + 1) / 2
  s <- t(mu) %*% K %*% mu + t(y.hat) %*% Lambda %*% y.hat
  r <- t(mu) %*% (K + XtLambdaX) %*% mu
  b.new <- as.numeric(b + 1 / 2 * (s - r))
  omega[i] <- rgamma(1, a.new, b.new)

  #lambda
  an <- tau + 1/2
  bn <- (1/2) * (omega[i] * (y - X %*% beta[i,])^2 + tau)
  lambda[i,] <- rgamma(1, an, bn)
}
```

Table 2.2: Results

	Least Squares	**Ridge $\lambda = 1$**	**Hierarchical Linear Model**
Intercept	15.3857	13.9440	15.2847
Age	0.6602	0.7656	0.6775
Sex (Male=1)	2.3210	2.5793	2.2454

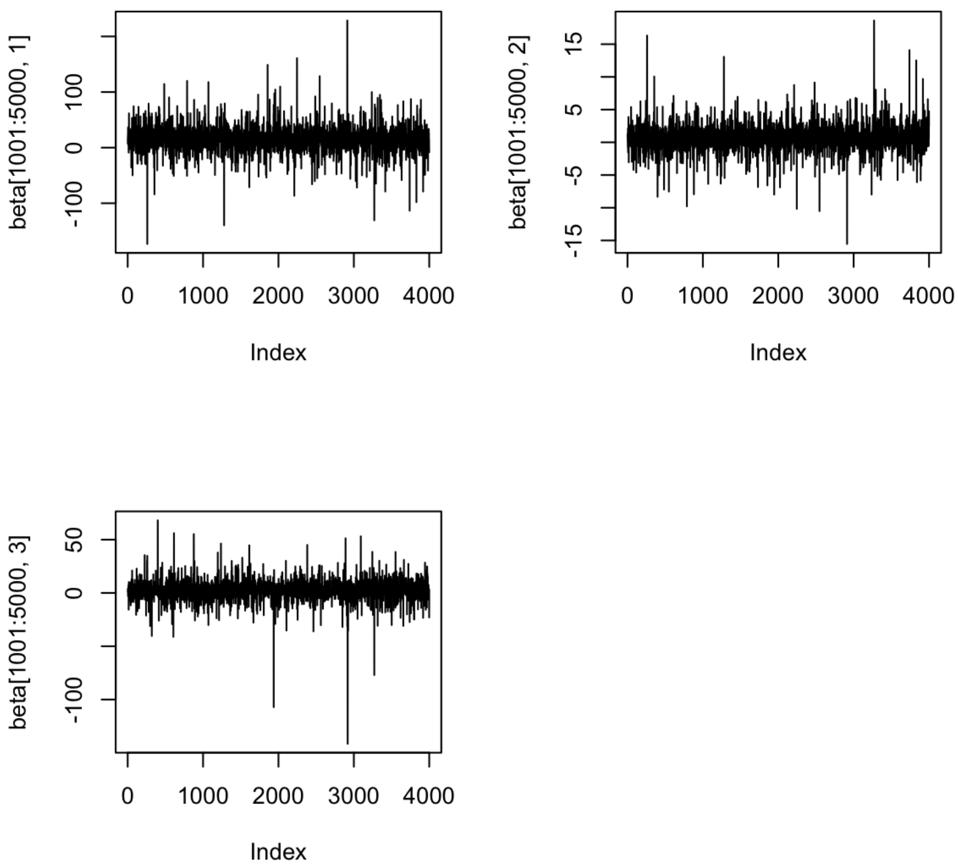
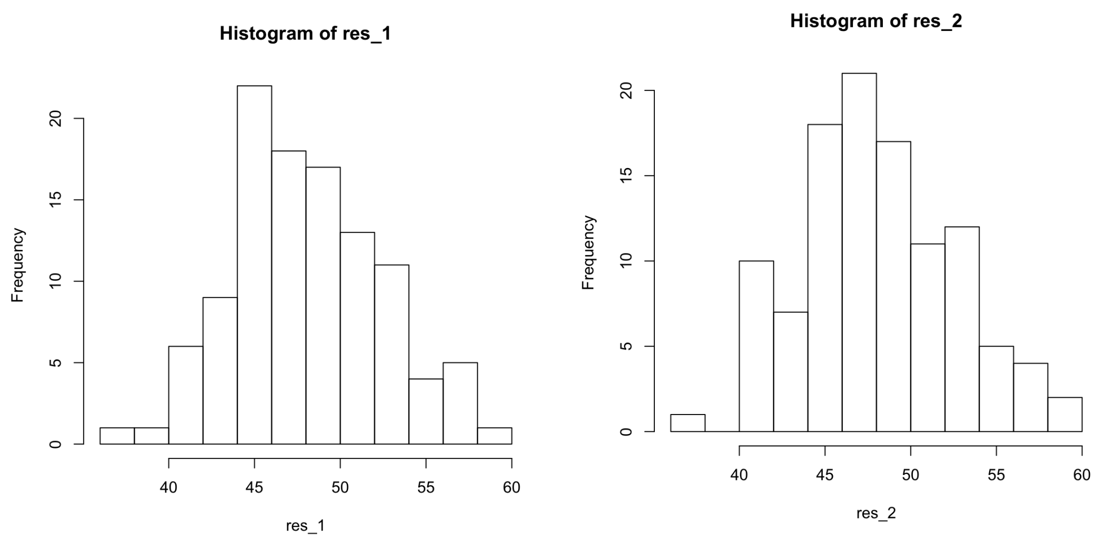


Figure 2.3: Trace plots

Exercise 2.15 *Compare the two fits. Does the new fit capture everything we would like? What assumptions is it making? In particular, look at the fit for just male and just female subjects. Suggest ways in which we could modify the model, and for at least one of the suggestions, write an updated Gibbs sampler and run it on your model.*

Solution

The plot of the residuals of the two models is shown in Figure 2.4. We can observe that the hierarchical model present ”fatter tails”, this is because the model is more robust with extreme values.



(a) Residuals from the Gaussian Linear Model

(b) Residuals from the Hierarchical Gaussian Linear Model

Figure 2.4: Residuals plot

We can modify the model by adding interaction terms, since the effect on women and on men is different, this can help to improve the model.