Prior Elicitation and Relative Belief Inferences for Linear Models

April 10, 2024

1 Introduction

This paper describes how to elicit a prior and how to implement Monte Carlo algorithms for the posterior calculations for a number of different linear models where the underlying error distributions are assumed to be normally distributed. Elicitation algorithms are provided for the parameters associated with means and variances of response variables while a uniform prior is always used for the correlation matrix when there several response variables. A website containing an implementation of this approach can be found at the website:

`?????`

The code can be downloaded at:

`https://?????`

Section 2 discusses the simplest case where there is a single response varaible and no predictors. Section 3 describes the generalization to the case of $p > 1$ response variables and no predictors. Section 4 considers the linear regression model and Section 5 consders the multivariate regression model.

2 Location-Scale Normal Model

Suppose the sampling model is $y \sim N(\mu, \sigma^2)$ where (μ, σ^2) is completely unknown. A conjugate prior is then given by

$$\begin{aligned} 1/\sigma^2 &\sim \text{gamma}(\alpha_{01}, \alpha_{02}) \\ \mu \,|\, \sigma^2 &\sim N(\mu_0, \lambda_0^2 \sigma^2) \end{aligned} \tag{1}$$

where $(\mu_0, \lambda_0, \alpha_{01}, \alpha_{02})$ are hyperparameters that need to be specified via an elicitation based upon what the practitioner knows about the measurement process in the application that produces y.

Consider now the elicitation of these hyperparameters and for this we need to specify a probability γ that represents virtual certainty. For example, $\gamma = 0.99$ is a reasonable choice and it will be used extensively in the elicitation. We break the elicitation into two parts. Note that, given the structure of the prior, namely, the prior on μ is conditional on a given value of σ^2, it makes sense to first elicit the values $(\alpha_{01}, \alpha_{02})$. Accordingly, we break the elicitation into two parts.

2.1 The Prior on σ

For this we note that the interval $\mu_{true} \pm \sigma_{true} z_{(1+\gamma)/2}$ will contain a measurement y with virtual certainty. Of course, we don't know $(\mu_{true}, \sigma_{true})$ and so we will place a prior on these quantities as specified in (1) and drop the "true" subscript. We suppose first that the practitioner's knowledge is such that they can specify some limits on the half-length of this interval as in specifying constants (s_1, s_2) where

$$s_1 \leq \sigma z_{(1+\gamma)/2} \leq s_2$$

such that these inequalities hold with virtual certainty. This implies that

$$\frac{z^2_{(1+\gamma)/2}}{s_2^2} \leq \frac{1}{\sigma^2} \leq \frac{z^2_{(1+\gamma)/2}}{s_1^2} \tag{2}$$

holds with virtual certainty. Note that it is possible to take $s_1^2 = 0$ but the bigger we can take it the better and we will suppose in the following that $s_1^2 > 0$ but it might be small.

Now let $G(\alpha_{01}, \alpha_{02}, \cdot)$ denote the gamma$(\alpha_{01}, \alpha_{02})$ cdf, using the rate parameterization, so that $G(\alpha_{01}, \alpha_{02}, x) = G(\alpha_{01}, 1, \alpha_{02} x)$. Therefore, with the quantile function of the gamma$(\alpha_{01}, \alpha_{02})$ denoted by $G^{-1}(\alpha_{01}, \alpha_{02}, \cdot)$, the interval given by (2) contains $1/\sigma^2$ with virtual certainty, when α_{01}, α_{02} satisfy,

$$\begin{aligned} G^{-1}(\alpha_{01}, \alpha_{02}, (1+\gamma)/2) &= z^2_{(1+p)/2}, /s_1^2, \\ G^{-1}(\alpha_{01}, \alpha_{02}, (1-\gamma)/2) &= z^2_{(1+p)/2}/s_2^2, \end{aligned}$$

or equivalently

$$\begin{aligned} G(\alpha_{01}, \alpha_{02}, z^2_{(1+p)/2}/s_1^2) &= (1+\gamma)/2, \\ G(\alpha_{01}, \alpha_{02}, z^2_{(1+p)/2}/s_2^2) &= (1-\gamma)/2, \end{aligned}$$

or equivalently

$$G(\alpha_{01}, 1, \alpha_{02} z^2_{(1+p)/2}/s_1^2) = (1+\gamma)/2, \tag{3}$$

$$G(\alpha_{01}, 1, \alpha_{02} z^2_{(1+p)/2}/s_2^2) = (1-\gamma)/2. \tag{4}$$

It is a simple matter to solve these equations for $(\alpha_{01}, \alpha_{02})$. For this choose an initial value for α_{01} and, using (3), find x such that $G(\alpha_{01}, 1, x) = (1+\gamma)/2$, which implies $\alpha_{02} = x s_{1i}^2 / z^2_{(1+p)/2}$. If the left-side of (4) is less (greater) than $(1-\gamma)/2$, then decrease (increase) the value of α_{01} and repeat step 1. Continue iterating this process until satisfactory convergence is attained.

it	α_{01}	α_{02}	prob. content of (s_1, s_2)
1	25	23.96118	0.9950000
2	12.5	14.14578	0.9950000
3	6.25	8.760395	0.9949823
4	3.125	5.723011	0.9898152
5	4.6875	7.292580	0.9946849
6	3.90625	6.524291	0.9937067
7	3.515625	6.128473	0.9924021
8	3.320312	5.927058	0.9913268
9	3.222656	5.8253790	0.99063500
10	3.173828	5.774283	0.9902425
11	3.149414	5.748669	0.9900334

Table 1: Output from the iterative procedure to determine (a_{01}, a_{02}) based on $(s_1, s_2) = (2, 10)$ where the starting value of a_{01} is 25 and error criiterion $\epsilon = 0.0001$.

If the initial value chosen for α_{01} is not chosen large enough, then the iteration will not converge and in any case one should specify a maximum number of iterations. If the iteration fails, then just choose a larger value of α_{01} and it is worth noting that the algorithm is robust to this choice as in choosing a very large value simply leads to more iterations with virtually the same values of $(\alpha_{01}, \alpha_{02})$ chosen for the hyperparameters. Table 1 contains the output of the iteration when $(s_1, s_2) = (2, 10)$ and the starting value of α_{01} is 25 and the algorithm stops when the content of (s_1, s_2) differs from 0.99 by less than $\epsilon = 0.0001$. So the elicited values are $(\alpha_{01}, \alpha_{02}) = (3.149414, 5.748669)$. It is a good idea to plot the elicited prior of $\sigma z_{(1+\gamma)/2}$ to make sure this makes sense and this is provided in Figure 1.

2.2 The Prior on μ

We now consider eliciting the hyperparameters (μ_0, λ_0). To start we specify an interval (m_1, m_2) such that we are virtually certain this contains the true value of μ. A natural choice for μ_0 is then $\mu_0 = (m_1 + m_2)/2$ but other choices could be made in (m_1, m_2). Now, given the prior determined for $1/\sigma^2$ in part 1 we have that the prior on μ, as determined by (1) is given by

$$\mu \sim \mu_0 + \sqrt{\frac{\alpha_{02}}{\alpha_{01}}} \lambda_0 t_{2\alpha_{01}}$$

where $t_{2\alpha_{01}}$ is distributed according to a t distribution with $2\alpha_{01}$ degrees of freedom. Let $H_{2\alpha_{01}}$ denote the cdf of $t_{2\alpha_{01}}$ so λ_0 must satisfy, using the symmetry of t distributions about 0 for the final equality and we have substituted

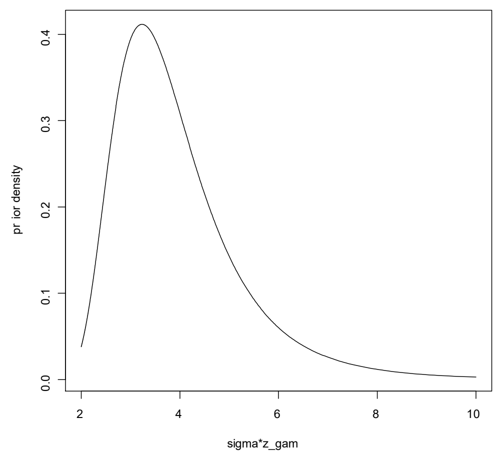

Figure 1: A plot of the elicited prior on $\sigma z_{(1+\gamma)/2}$.

$\mu_0 = (m_1 + m_2)/2$,

$$\begin{aligned}\gamma &= H_{2\alpha_{01}}\left(\frac{m_2-\mu_0}{\sqrt{\alpha_{02}/\alpha_{01}}\lambda_0}\right) - H_{2\alpha_{01}}\left(\frac{m_1-\mu_0}{\sqrt{\alpha_{02}/\alpha_{01}}\lambda_0}\right)\\ &= 2H_{2\alpha_{01}}\left(\frac{m_2-m_1}{\sqrt{\alpha_{02}/\alpha_{01}}\lambda_0}\right) - 1.\end{aligned}$$

Then denoting the quantile function of the $t_{2\alpha_{01}}$ distribution by $H^{-1}_{2\alpha_{01}}$

$$\frac{m_2-m_1}{2\sqrt{\alpha_{02}/\alpha_{01}}\lambda_0} = H^{-1}_{2\alpha_{01}}\left(\frac{1+\gamma}{2}\right)$$

or

$$\lambda_0 = \frac{m_2-m_1}{2\sqrt{\alpha_{02}/\alpha_{01}}H^{-1}_{2\alpha_{01}}\left(\frac{1+\gamma}{2}\right)}$$

and this completes the elicitation.

Consider the previous example where we determined

$$(\alpha_{01}, \alpha_{02}) = (3.149414, 5.748669)$$

and now suppose that $(m_1, m_2) = (-2, 10)$. So,

$$2\alpha_{01} = 2(3.149414) = 6.298828$$

is the degrees of freedom for the t distribution and $H^{-1}_{2\alpha_{01}}((1+\gamma)/2) = 3.636274$ so the interval $(-3.636274, 3.636274)$ contains γ of the probability for the $t_{2\alpha_{01}}$

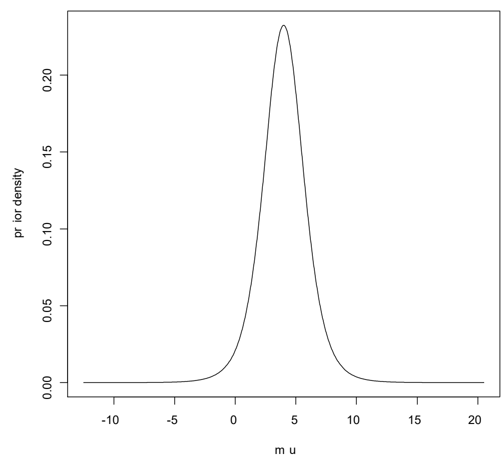

Figure 2: A plot of the elicited prior on μ.

distribution. For the density of μ this is recentered to $\mu_0 = (m_1 + m_2)/2 = 4$ and rescaled by the factor $\sigma_0 = \sqrt{\alpha_{02}/\alpha_{01}}\lambda_0$ where

$$\begin{aligned}\lambda_0 &= \frac{m_2 - m_1}{2\sqrt{\alpha_{02}/\alpha_{01}}H^{-1}_{2\alpha_{01}}\left(\frac{1+\gamma}{2}\right)} = \frac{12}{2(1.351042)(3.636274)} \\ &= 1.22131.\end{aligned}$$

Figure 2 is a plot of the prior on μ.

2.3 The Posterior

Suppose now that $\mathbf{y} = (y_1, \ldots, y_n)'$ is observed with mss $(\bar{y}, s^2)$ where $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$. The prior specified in (1) is a conjugate prior and leads to the following posterior.

$$\begin{aligned}1/\sigma^2 \,|\, \mathbf{y} &\sim \text{gamma}(\alpha_{01} + n/2, \alpha_{02}(\mathbf{y})) \\ \mu \,|\, \sigma^2, \mathbf{y} &\sim N(\mu(\mathbf{y}), (1/\lambda_0^2 + n)^{-1}\sigma^2)\end{aligned} \tag{5}$$

where

$$\begin{aligned}\alpha_{02}(\mathbf{y}) &= \alpha_{02} + s^2/2 + n(\bar{y} - \mu_0)^2/2(1 + n\lambda_0^2), \\ \mu(\mathbf{y}) &= (1/\lambda_0^2 + n)^{-1}(\mu_0/\lambda_0^2 + n\bar{y}).\end{aligned}$$

As such, it is possible to directly sample from the posterior to determine posterior expectations.

3 Multivariate Normal

3.1 The Prior

Suppose the sampling model is $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \Sigma)$ where $(\boldsymbol{\mu}, \Sigma)$ is completely unknown. We write $\Sigma = DRD$ where $D = \text{diag}(\sigma_{11}^{1/2}, \ldots, \sigma_{pp}^{1/2})$ and R is the correlation matrix. Rather than a conjugate prior we consider the prior

$$\begin{aligned} 1/\sigma_{ii} &\sim \text{gamma}(\alpha_{01i}, \alpha_{02i}) \text{ for } i = 1, \ldots, p \\ R &\sim \text{uniform}(C_p) \text{ where } C_p = \text{ the set of } p \times p \text{ correlation matrices} \\ \boldsymbol{\mu} \,|\, \Sigma &\sim N_p(\boldsymbol{\mu}_0, \Lambda_0 \Sigma \Lambda_0) \text{ where } \Lambda_0 = \text{diag}(\lambda_{01}, \ldots, \lambda_{0p}). \end{aligned} \tag{6}$$

So, $(\boldsymbol{\mu}_0, \Lambda_0, \boldsymbol{\alpha}_{01}, \boldsymbol{\alpha}_{02})$ are the $4p$ hyperparameters that need to be specified via an elicitation based upon what the practitioner knows about the measurement process in the application that produces y.

Here the values of $(\boldsymbol{\alpha}_{01}, \boldsymbol{\alpha}_{02})$ are elicited exactly as specified in Section 2 and this is done variable by variable. Now note that $\Lambda_0 \Sigma \Lambda_0 = \Lambda_0 DRD \Lambda_0$ and $\Lambda_0 D = \text{diag}(\lambda_{01}\sigma_{11}^{1/2}, \ldots, \lambda_{0p}\sigma_{pp}^{1/2})$ and so $\boldsymbol{\mu} \,|\, \Sigma \sim N_p(\boldsymbol{\mu}_0, \Lambda_0 \Sigma \Lambda_0)$. Therefore, we have that the

$$\mu_i \,|\, \Sigma \sim N(\mu_{0i}, \lambda_{0i}^2 \sigma_{ii})$$

and so the marginal conditional prior on μ_i only depends on σ_{ii} and the hyperparameters (μ_{0i}, λ_{0i}). Then, since $1/\sigma_{ii} \sim \text{gamma}(\alpha_{i01}, \alpha_{i02})$ this implies that the unconditional prior on μ_i is given by

$$\mu_i \sim \mu_{0i} + \sqrt{\frac{\alpha_{02i}}{\alpha_{01}i}} \lambda_{0i} t_{2\alpha_{01i}}.$$

In other words, the values (μ_{0i}, λ_{0i}) can be elicited variable by variable exactly as we have done in Section 2 but with possibly different limits (m_{1i}, m_{2i}) holding with virtual certainty for μ_i.

3.2 The Posterior

The prior (4) leads to some computational issues for the posterior that need to be addressed. First we consider a conjugate prior given by

$$\begin{aligned} \Sigma^{-1} &\sim W_p(\Sigma_0, f_0) \\ \boldsymbol{\mu} \,|\, \Sigma &\sim N_p(\boldsymbol{\mu}_0, \lambda_0^2 \Sigma). \end{aligned} \tag{7}$$

The prior (6) differs from (7) in several significant ways. For (7) we need to specify Σ_0 which means we need to elicit both variances and covariances with the latter being much harder to do. Also we need to specify the degrees of freedom parameter. More seriously, while eliciting $\boldsymbol{\mu}_0$ can be done similar to what we have described already, the conjugacy of this prior demands that $\lambda_{01} = \cdots = \lambda_{0p} = \lambda_0$ and this results in a considerable loss of control over the priors on the individual means μ_i and this is not satisfactory. Still, as shall be demonstrated, the posterior resulting from (7) has value when evaluating expectations with respect to the posterior induced by (6) and so is stated here.

The following result is well-known and is stated here as it will be used when considering the posterior associated with (6). Here $Y = (\mathbf{y}_1 \cdots \mathbf{y}_n)' \in \mathbb{R}^{n\times p}$ is the observed sample, $\bar{\mathbf{y}}' = \mathbf{1}'Y/n$ and $S = (Y - \mathbf{1}\bar{\mathbf{y}}')'(Y - \mathbf{1}\bar{\mathbf{y}}')$.

Theorem 1. With the prior given by (7), the posterior of $(\boldsymbol{\mu}, \Sigma^{-1})$ given Y, is

$$\begin{aligned} \Sigma^{-1} \,|\, Y &\sim W_p(\Sigma(Y), f_0 + n), \\ \boldsymbol{\mu} \,|\, \Sigma, Y &\sim N_p(\boldsymbol{\mu}(Y), (n + 1/\lambda_0^2)^{-1}\Sigma), \end{aligned} \tag{8}$$

where

$$\begin{aligned} \Sigma(Y) &= \left(\Sigma_0^{-1} + S + \frac{n}{1 + n\lambda_0^2}(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)'\right)^{-1}, \\ \boldsymbol{\mu}(Y) &= (n + 1/\lambda_0^2)^{-1}(\boldsymbol{\mu}_0/\lambda_0^2 + n\bar{\mathbf{y}}). \end{aligned}$$

Note that when $p = 1$, then $\text{gamma}(\alpha_{01}, \alpha_{02}) = W_1(1/2\alpha_{02}, 2\alpha_{01})$ which explains the apparent difference in appearance between (8) and (5).

Now consider the prior given by (6) and note that for matrix A, $\text{etr}(A) = \exp(\text{tr}(A))$. For this let $w_p(\cdot; V, f)$ denote the density of a $\text{Wishart}_p(V, f)$ density and $\varphi_p(\cdot; \boldsymbol{\mu}, \Sigma)$ denote the density of a $N_p(\boldsymbol{\mu}, \Sigma)$ distribution.

Theorem 2. With the prior given by (6), the posterior of $(\boldsymbol{\mu}, \Xi) = (\boldsymbol{\mu}, \Sigma^{-1})$ is proportional to

$$w_p(\Xi; S^{-1}, n - p - 1)\varphi_p(\boldsymbol{\mu}; \bar{\mathbf{y}}, \Xi^{-1}/n)k(\boldsymbol{\mu}, \Xi) \tag{9}$$

where

$$\begin{aligned} k(\boldsymbol{\mu}, \Xi) &= \exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\Lambda_0^{-1}\Xi\Lambda_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right\} \times \\ &\quad \prod_{i=1}^{k} (1/\sigma_{ii})^{\alpha_{01i} + (p+1)/2} \exp\left(-\alpha_{02i}/\sigma_{ii}\right). \end{aligned}$$

Proof: We need to show that the joint posterior of $(\boldsymbol{\mu}, \Xi)$ is proportional to

$$|\Xi|^{(n-2p-2)/2}\text{etr}\left\{-\frac{1}{2}S\Xi\right\}|\Xi|^{1/2}\exp\left\{-\frac{n}{2}(\boldsymbol{\mu} - \bar{\mathbf{y}})'\Xi(\boldsymbol{\mu} - \bar{\mathbf{y}})\right\}k(\boldsymbol{\mu}, \Xi).$$

With $D = \text{diag}(\sigma_{11}^{1/2}, \ldots, \sigma_{pp}^{1/2})$ put $U = D^{-2} = \text{diag}(1/\sigma_{11}, \ldots, 1/\sigma_{pp})$, then the posterior of $(\boldsymbol{\mu}, U, R)$ is proportional to,

$$\begin{aligned} &L(\boldsymbol{\mu}, U, R \,|\, Y)\pi(\boldsymbol{\mu} \,|\, U, R)\pi(U) \\ \propto\; &|\Sigma|^{-n/2}\exp\left\{-\frac{n}{2}(\bar{\mathbf{y}} - \boldsymbol{\mu})'\Sigma^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})\right\}\text{etr}\left\{-\frac{1}{2}S\Sigma^{-1}\right\} \times \\ &|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\Lambda_0^{-1}\Sigma^{-1}\Lambda_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right\}\prod_{i=1}^{p} u_i^{\alpha_{01i} - 1}\exp\left(-\alpha_{02i}u_i\right). \end{aligned}$$

Now put $V = \text{diag}(v_1, \ldots, v_p) = U^{-1} = \text{diag}(1/u_1, \ldots, 1/u_p)$ and note that the transformation $U \to V$ has Jacobian $v_1^{-2} \cdots v_p^{-2}$. Therefore, the posterior of $(\boldsymbol{\mu}, V, R)$, where $\Sigma = V^{1/2} R V^{1/2}$, is proportional to

$$|\Sigma|^{-n/2} \text{etr}\left\{-\frac{1}{2} S \Sigma^{-1}\right\} |\Sigma|^{-1/2} \exp\left\{-\frac{n}{2}\left((\bar{\mathbf{y}} - \boldsymbol{\mu})' \Sigma^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu})\right)\right\} \times$$
$$\exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)' \Lambda_0^{-1} \Sigma^{-1} \Lambda_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right\} \prod_{i=1}^{p} v_i^{-\alpha_{01i} - 1} \exp\left(-\frac{\alpha_{02i}}{v_i}\right).$$

Now make the transformation $(\boldsymbol{\mu}, V, R) \to (\boldsymbol{\mu}, \Sigma)$. We need to calculate the Jacobian of the transformation $(V, R) \to \Sigma$. For this we have $\sigma_{ij} = v_i^{1/2} v_j^{1/2} r_{ij}$ when $i \neq j$ and $\sigma_{ii} = v_i$. Therefore, when $i \neq j$

$$\frac{\partial \sigma_{ij}}{\partial v_i} = \frac{1}{2} v_i^{-1/2} v_j^{1/2} r_{ij}, \; \frac{\partial \sigma_{ij}}{\partial v_j} = \frac{1}{2} v_i^{1/2} v_j^{-1/2} r_{ij}, \; \frac{\partial \sigma_{ij}}{\partial r_{ij}} = v_i^{1/2} v_j^{1/2}$$

and $\partial \sigma_{ii} / \partial v_i = 1$ with all other partial derivatives equal to 0. Now in the matrix of partial derivatives order the rows according to

$$(\sigma_{11}, \ldots, \sigma_{pp}, \sigma_{12}, \ldots, \sigma_{1p}, \sigma_{23}, \ldots, \sigma_{p-1p})$$

and order the columns according to

$$(v_1, \ldots, v_p, r_{12}, \ldots, r_{1p}, r_{23}, \ldots, r_{p-1p}).$$

It is then easy to see that this matrix is lower triangular with diagonal given by

$$(1, \ldots, 1, v_1^{1/2} v_2^{1/2}, \ldots, v_1^{1/2} v_p^{1/2}, v_2^{1/2} v_3^{1/2}, \ldots, v_{p-1}^{1/2} v_p^{1/2})$$

with determinant equal to

$$v_1^{(p-1)/2} v_2^{(p-1)/2} \cdots v_p^{(p-1)/2} = |\text{diag}(\Sigma)|^{(p-1)/2}$$

so the Jacobian of the transformation $(V, R) \to \Sigma$ is $|\text{diag}(\Sigma)|^{-(p-1)/2}$. This implies that the posterior of $(\boldsymbol{\mu}, \Sigma)$ is proportional to

$$\exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)' \Lambda_0^{-1} \Sigma^{-1} \Lambda_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right\} \times$$
$$\prod_{i=1}^{p} \left(\frac{1}{\sigma_{ii}}\right)^{\alpha_{01i} + (p+1)/2} \exp\left(-\frac{\alpha_{02i}}{\sigma_{ii}}\right) \times$$
$$|\Sigma|^{-n/2} \text{etr}\left\{-\frac{1}{2} S \Sigma^{-1}\right\} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\bar{\mathbf{y}} - \boldsymbol{\mu})' (\Sigma/n)^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu})\right\}.$$

Finally, make the transformation $\Sigma \to \Xi = \Sigma^{-1}$ which has Jacobian $|\Xi|^{-(p+1)}$ (see Muirhead(1982), Theorem 2.1.8). With this, the joint posterior of $(\boldsymbol{\mu}, \Xi)$ is proportional to (9). ■

As will be discussed, one of the virtues of (9) is that we can sample from the joint density $w_p(\Xi; S^{-1}, n-p-1)\varphi_p(\boldsymbol{\mu}; \bar{\mathbf{y}}, \Xi^{-1})$ and what is left over, namely the factor $k(\boldsymbol{\mu}, \Xi)$, does not depend on the data and generally can be considered to have much less of an influence on the posterior, particularly as the sample size grows. It is of some relevance, however, that another form of this result allows direct sampling from the conditional posterior $\boldsymbol{\mu} \,|\, \Sigma, Y$.

Corollary 2.1. With the prior given by (6), the posterior of $(\boldsymbol{\mu}, \Xi) = (\boldsymbol{\mu}, \Sigma^{-1})$ is proportional to

$$\begin{aligned} & w_p(\Xi; S^{-1}, n-p)\varphi_p(\boldsymbol{\mu}; \boldsymbol{\mu}(Y), (n\Xi + \Lambda_0^{-1}\Xi\Lambda_0^{-1})^{-1}) \times \\ & \exp\left\{-\frac{1}{2}(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)'\Lambda_0^{-1}(\Xi - \Xi(n\Lambda_0^{-1}\Xi\Lambda_0^{-1} + \Xi)^{-1}\Xi)\Lambda_0^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)\right\} k(\Xi) \end{aligned} \tag{10}$$

where

$$\begin{aligned} \boldsymbol{\mu}(Y) &= (\Xi + \Lambda_0^{-1}\Xi\Lambda_0^{-1}/n)^{-1}(\Xi\bar{\mathbf{y}} + \Lambda_0^{-1}\Xi\Lambda_0^{-1}\boldsymbol{\mu}_0/n) \\ k(\Xi) &= |n\Xi + \Lambda_0^{-1}\Xi\Lambda_0^{-1}|^{-1}\prod_{i=1}^{k}(1/\sigma_{ii})^{\alpha_{01i}+(p+1)/2}\exp\left(-\alpha_{02i}/\sigma_{ii}\right), \end{aligned}$$

so $\boldsymbol{\mu} \,|\, \Sigma, Y \sim N_p(\boldsymbol{\mu}(Y), (n\Xi + \Lambda_0^{-1}\Xi\Lambda_0^{-1})^{-1})$.

3.3 Importance Sampling

Theorem 2 suggests a good importance sampler for computing integrals with respect to the posterior.

Theorem 3. To approximate the expectation $E_{\Pi(\cdot\,|\,Y)}(h(\boldsymbol{\mu}, \Xi))$ generate a sample $(\boldsymbol{\mu}_i, \Xi_i)$ for $i = 1, \ldots, N$ where

$$\begin{aligned} \Xi_i &\sim W_p(S^{-1}, n-p-1), \\ \boldsymbol{\mu}_i \,|\, \Xi_i &\sim N_p(\bar{\mathbf{y}}, \Xi_i^{-1}/n) \end{aligned} \tag{11}$$

and evaluate the estimate

$$\hat{E}_{\Pi(\cdot\,|\,Y)}(h(\boldsymbol{\mu}, \Xi)) = \frac{\sum_{i=1}^{N} h(\boldsymbol{\mu}_i, \Xi_i)k(\boldsymbol{\mu}_i, \Xi_i)}{\sum_{i=1}^{N} k(\boldsymbol{\mu}_i, \Xi_i)}. \tag{12}$$

For an estimate of its standard error. See Evans and Swartz (2000), Theorem 6.4 for a formula for the standard error. Note that the merit in this importance sampler lies in the fact that $k(\boldsymbol{\mu}, \Xi)$ does not depend on the data Y and it is the data that primarily determines where the posterior concentrates and this is the main determinant of the success of any integration algorithm in this context. The effect of the prior on the posterior generally decreases is as sample size increases as it concentrates about the true values and so the function k becomes effectively constant in the region where the bulk of the posterior probability lies .

In fact, a whole class of importance samplers suggests itself. For we can write,

$$
\begin{aligned}
& |\Xi|^{(n-2p-2)/2}\mathrm{etr}\left\{-\frac{1}{2}S\Xi\right\}|\Xi|^{1/2}\exp\left\{-\frac{n}{2}(\boldsymbol{\mu}-\bar{\mathbf{y}})'\Xi(\boldsymbol{\mu}-\bar{\mathbf{y}})\right\}k(\boldsymbol{\mu},\Xi) \\
= \; & |\Xi|^{(n-2p-2)/2}\mathrm{etr}\left\{-\frac{1}{2}S\Xi\right\}\times \\
& |\Xi|^{1/2}\exp\left\{-\frac{n}{2}(\boldsymbol{\mu}-\bar{\mathbf{y}})'\Xi(\boldsymbol{\mu}-\bar{\mathbf{y}})-\frac{1}{2\lambda_0^2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\Xi(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right\}\times \\
& k(\boldsymbol{\mu},\Xi)\exp\left\{\frac{1}{2\lambda_0^2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\Xi(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right\}.
\end{aligned}
$$

Note that, excepting $k(\boldsymbol{\mu},\Xi)$, the first line is similar to the situation that obtains when obtaining the posterior with part of the conjugate prior on $\boldsymbol{\mu}$ but with no prior on Ξ. So, this now becomes proportional to, using the notation, $\mathbf{x}'A(\cdot)=\mathbf{x}'A\mathbf{x}$ for a quadratic form,

$$
\begin{aligned}
& |\Xi|^{1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\mu}-(1/\lambda_0^2+n)^{-1}(\boldsymbol{\mu}_0/\lambda_0^2+n\bar{\mathbf{y}})'(1/\lambda_0^2+n)\Xi(\cdot)\right\}\times \\
& |\Xi|^{(n-2p-2)/2}\mathrm{etr}\left\{-\frac{1}{2}\left[S+\frac{n}{1+n\lambda_0^2}(\bar{\mathbf{y}}-\boldsymbol{\mu}_0)(\bar{\mathbf{y}}-\boldsymbol{\mu}_0)'\right]\Xi\right\}k_{\lambda_0}(\boldsymbol{\mu},\Xi)
\end{aligned}
$$

where

$$
\begin{aligned}
k_{\lambda_0}(\boldsymbol{\mu},\Xi) \;=\; & \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'(\Lambda_0^{-1}\Xi\Lambda_0^{-1}-\lambda_0^{-2}\Xi)(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right\}\times \\
& \prod_{i=1}^{k}(1/\sigma_{ii})^{\alpha_{01i}+(p+1)/2}\exp\left(-\alpha_{02i}/\sigma_{ii}\right).
\end{aligned}
$$

This establishes the following result.

Theorem 4. To approximate the expectation $E_{\Pi(\cdot\,|\,Y)}(h(\boldsymbol{\mu},\Xi))$ generate a sample $(\boldsymbol{\mu}_i,\Xi_i)$ for $i=1,\ldots,N$ where

$$
\begin{aligned}
\Xi &\sim W_p(\Sigma(Y),n-p-1), \\
\boldsymbol{\mu}\,|\,\Xi &\sim N_p(\boldsymbol{\mu}(Y),(n+1/\lambda_0^2)^{-1}\Xi^{-1}),
\end{aligned}
\tag{13}
$$

where

$$
\begin{aligned}
\Sigma(Y) &= \left(S+\frac{n}{1+n\lambda_0^2}(\bar{\mathbf{y}}-\boldsymbol{\mu}_0)(\bar{\mathbf{y}}-\boldsymbol{\mu}_0)'\right)^{-1}, \\
\boldsymbol{\mu}(Y) &= (n+1/\lambda_0^2)^{-1}(\boldsymbol{\mu}_0/\lambda_0^2+n\bar{\mathbf{y}})
\end{aligned}
$$

and evaluate the estimate

$$
\hat{E}_{\Pi(\cdot\,|\,Y)}(h(\boldsymbol{\mu},\Xi))=\frac{\sum_{i=1}^{N}h(\boldsymbol{\mu}_i,\Xi_i)k_{\lambda_0}(\boldsymbol{\mu}_i,\Xi_i)}{\sum_{i=1}^{N}k_{\lambda_0}(\boldsymbol{\mu}_i,\Xi_i)}. \tag{14}
$$

We are free to choose λ_0^2 to improve the efficiency of the sampler and a natural choice is $\lambda_0^2 = \max(\lambda_{01}^2, \ldots, \lambda_{0p}^2)$ as this guarantees that the importance sampler is based on the most pessimistic choice of this hyperparameter. Note that taking $\lambda_0^2 = \infty$ gives the importance sampler (11).

3.4 Sampling from the Posterior

It is also of interest to be able to at least approximately generate a sample of n from the posterior. For this we can use the SIR algortithm of Rubin (1988). After generating the sample from the importance sampler and computing the weights. The cumulative cdf of the weights is then evaluated as $(i, \sum_{j=1}^{i} w_j)$, a value $U \sim U(0,1)$ is then generated and the value i, satisfying $\sum_{j=1}^{i} w_j \leq U < \sum_{j=1}^{i+1} w_j$ is computed. The value $(\boldsymbol{\mu}_i, \Xi_i)$ is then returned as a value approximately generated from the posterior. This is repeated n independent times to generate the sample $(\boldsymbol{\mu}_{i_1}, \Xi_{i_1}), \ldots, (\boldsymbol{\mu}_{i_n}, \Xi_{i_n})$. This approximation is justified by the fact that a generated value converges weakly to the posterior distribution as $N \to \infty$.

4 Linear Regression

Suppose the sampling model is $y \sim N(\mathbf{x}'\boldsymbol{\beta}, \sigma^2)$ for a set of k predictors $\mathbf{x} = (x_1, \ldots, x_k)'$ and where $(\boldsymbol{\beta}, \sigma^2)$ is completely unknown. A possible prior is then given by

$$\begin{aligned} 1/\sigma^2 &\sim \text{gamma}(\alpha_{01}, \alpha_{02}) \\ \boldsymbol{\beta} \,|\, \sigma^2 &\sim N_k(\boldsymbol{\beta}_0, \sigma^2 \Lambda_0^2) \end{aligned} \tag{15}$$

where $\Lambda_0 = \text{diag}(\lambda_{01}, \ldots, \lambda_{0k})$ and so $(\boldsymbol{\beta}_0, \Lambda_0, \alpha_{01}, \alpha_{02})$ are hyperparameters that need to be specified via an elicitation. The elicitation of $(\alpha_{01}, \alpha_{02})$ for σ proceeds as in Section 2 so we focus on the prior for $\boldsymbol{\beta}$. We consider several different scenarios for this.

4.1 Prior information about each β_i

This proceeds just as in Section 2 for the mean of a multivariate normal. So, we have that the marginal prior on β_i is

$$\beta_i \sim \beta_{0i} + \sqrt{\frac{\alpha_{02i}}{\alpha_{01i}}} \lambda_{0i} t_{2\alpha_{01i}}.$$

So after specifying the limits (m_{1i}, m_{2i}), such that we are virtually certain this contains the true value of β_i, and putting $\beta_{0i} = (m_{1i} + m_{2i})/2$, then

$$\lambda_{0i} = \frac{m_{2i} - m_{1i}}{2\sqrt{\alpha_{02i}/\alpha_{01i}} H_{2\alpha_{01i}}^{-1}\left(\frac{1+\gamma}{2}\right)}.$$

The likelihood for the observed $\mathbf{y} \sim N_n(X\boldsymbol{\beta}, \sigma^2)$ at predictor values $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)' \in \mathbb{R}^{n\times k}$ is given by

$$
\begin{aligned}
L(\boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{y}) &\propto (\sigma^{-2})^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta})\right\} \\
= \quad &\left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - X\mathbf{b})'(\mathbf{y} - X\mathbf{b})\right\} \times \\
&\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{b} - \boldsymbol{\beta})'X'X(\mathbf{b} - \boldsymbol{\beta})\right\}
\end{aligned}
$$

where $\mathbf{b} = (X'X)^{-1}X'\mathbf{y}$ and it is assumed that X is of rank k. It is clear that the prior (15) is conjugate and this leads to the following well-known result for the posterior.

Theorem 4. With the prior given by (15), the posterior of $(\boldsymbol{\beta}, \sigma^2)$ is

$$
\begin{aligned}
1/\sigma^2 \,|\, \mathbf{y} &\sim \text{gamma}(\alpha_{01} + n/2, \alpha_{02}(\mathbf{y})) \\
\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y} &\sim N(\boldsymbol{\beta}(\mathbf{y}), (X'X + \Lambda_0^{-2})^{-1}\sigma^2)
\end{aligned}
\tag{16}
$$

where

$$
\begin{aligned}
\alpha_{02}(\mathbf{y}) &= \alpha_{02} + s^2/2 + (\mathbf{b} - \boldsymbol{\beta}_0)'(\Lambda_0^2 + (X'X)^{-1})^{-1}(\mathbf{b} - \boldsymbol{\beta}_0)/2, \\
\boldsymbol{\beta}(\mathbf{y}) &= (X'X + \Lambda_0^{-2})^{-1}((X'X)\mathbf{b} + \Lambda_0^{-2}\boldsymbol{\beta}_0).,
\end{aligned}
$$

where $s^2 = (\mathbf{y}-X\mathbf{b})'(\mathbf{y}-X\mathbf{b})$.

Therefore it is possible to sample directly from the posterior in this context.

4.2 Prior information about $\mathbf{x}'\boldsymbol{\beta}$ for some set of predictors

Suppose that there is little knowledge about the individual β_i. It is supposed instead that there is a set of k linearly independent predictor vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ such that bounds (m_{i1}, m_{i2}) can be stated so that $\mathbf{v}_i'\boldsymbol{\beta} \in (m_{i1}, m_{i2})$ with virtual certainty. Note that the $\mathbf{v}_i$ could be observed or specified values of the predictors. Putting

$$
V = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_k \end{pmatrix}'
$$

this implies $\boldsymbol{\theta} = V\boldsymbol{\beta} \in (m_{11}, m_{12}) \times \cdots \times (m_{k1}, m_{k2})$ with virtual certainty. Now suppose that the prior for $(\boldsymbol{\alpha},\sigma^2)$ is given by

$$
\begin{aligned}
1/\sigma^2 &\sim \text{gamma}(\alpha_{01}, \alpha_{02}) \\
\boldsymbol{\theta} \,|\, \sigma^2 &\sim N_k(\boldsymbol{\theta}_0, \sigma^2\Lambda_0^2)
\end{aligned}
\tag{17}
$$

and we determine the hyperparameters just as was done previously for $(\boldsymbol{\beta},\sigma^2)$. This implies that the prior on $(\boldsymbol{\beta},\sigma^2)$ is now

$$
\begin{aligned}
1/\sigma^2 &\sim \text{gamma}(\alpha_{01}, \alpha_{02}) \\
\boldsymbol{\beta} \,|\, \sigma^2 &\sim N_k(\boldsymbol{\beta}_0, \sigma^2 V_0)
\end{aligned}
\tag{18}
$$

where $\boldsymbol{\beta}_0 = V_0^{-1}\boldsymbol{\theta}, V_0 = V^{-1}\Lambda_0^2(V^{-1})'$.

The posterior. is then given by the following result.

Theorem 5. With the prior given by (15), the posterior of $(\boldsymbol{\beta}, \sigma^2)$ is

$$\begin{aligned} 1/\sigma^2 \,|\, \mathbf{y} &\sim \text{gamma}(\alpha_{01} + n/2, \alpha_{02}(\mathbf{y})) \\ \boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y} &\sim N(\boldsymbol{\beta}(\mathbf{y}), \sigma^2(X'X + V_0^{-1})^{-1}) \end{aligned} \tag{19}$$

where

$$\begin{aligned} \alpha_{02}(\mathbf{y}) &= \alpha_{02} + s^2/2 + (\mathbf{b} - \boldsymbol{\beta}_0)'((V_0 + (X'X)^{-1})^{-1}(\mathbf{b} - \boldsymbol{\beta}_0)/2, \\ \boldsymbol{\beta}(\mathbf{y}) &= (X'X + V_0^{-1})^{-1}((X'X)\mathbf{b} + V_0^{-1}\boldsymbol{\beta}_0). \end{aligned}$$

Again, it is possible to sample directly from the posterior.

Note that the elicitation methodology of Section 4.1 is really encompassed by that of Section 4.2. For example, the method of Section 4.1 is just V with $\mathbf{v}_i = \mathbf{e}_i$ the i-th standard basis vector.

5 Multivariate Linear Regression

Now suppose we have a p-dimensional response vector $\mathbf{y}$ and predictor variables $\mathbf{x} \in \mathbb{R}^k$ such that

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \sim N_p \left(\begin{pmatrix} \mathbf{x}'\boldsymbol{\beta}_1 \\ \vdots \\ \mathbf{x}'\boldsymbol{\beta}_p \end{pmatrix}, \Sigma \right) = N_p \left(\mathcal{B}'\mathbf{x}, \Sigma \right)$$

where

$$\mathcal{B} = \begin{pmatrix} \boldsymbol{\beta}_1 & \cdots & \boldsymbol{\beta}_p \end{pmatrix} \in \mathbb{R}^{k \times p}.$$

So, $\boldsymbol{\beta}_i = (\beta_{1i}, \ldots, \beta_{ki})'$ contains the regression coefficients relating the i-th response variable y_i to the k predictor variables. Therefore, with $\mathbf{y}_1, \ldots, \mathbf{y}_n$ independent, $\mathbf{y}_i \sim N_p(\mathcal{B}\mathbf{x}_i, \Sigma)$ and

$$X = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{pmatrix}', Y = \begin{pmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_n \end{pmatrix}',$$

the likelihood function is

$$
\begin{aligned}
& L(\mathcal{B},\Sigma\,|\,Y) \\
\propto\ & |\Sigma|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i-\mathcal{B}'\mathbf{x}_i\right)'\Sigma^{-1}\left(\mathbf{y}_i-\mathcal{B}'\mathbf{x}_i\right)\right\} \\
=\ & |\Sigma|^{-n/2}\mathrm{etr}\left\{-\frac{1}{2}\left(\sum_{i=1}^{n}\left(\mathbf{y}_i-\mathcal{B}'\mathbf{x}_i\right)\left(\mathbf{y}_i-\mathcal{B}'\mathbf{x}_i\right)'\right)\Sigma^{-1}\right\} \\
=\ & |\Sigma|^{-n/2}\mathrm{etr}\left\{-\frac{1}{2}\left(Y-X\mathcal{B}\right)'\left(Y-X\mathcal{B}\right)\Sigma^{-1}\right\} \\
=\ & |\Sigma|^{-n/2}\mathrm{etr}\left\{-\frac{1}{2}(Y-XB)'(Y-XB)\Sigma^{-1}+(B-\mathcal{B})'X'X(B-\mathcal{B})\Sigma^{-1}\right\} \\
=\ & |\Sigma|^{-n/2}\mathrm{etr}\left\{-\frac{1}{2}S\Sigma^{-1}\right\}\mathrm{etr}\left\{-\frac{1}{2}(B-\mathcal{B})'X'X(B-\mathcal{B})\Sigma^{-1}\right\}
\end{aligned}
$$

and where $S=(Y-XB)'(Y-XB), B=(X'X)^{-1}X'Y$ since

$$\left(Y-X\mathcal{B}\right)'\left(Y-X\mathcal{B}\right)=(Y-XB)'(Y-XB)+(B-\mathcal{B})'X'X(B-\mathcal{B}).$$

For the prior on Σ we proceed as in Section 2.1 and put

$$
\begin{aligned}
1/\sigma_{ii} &\sim \text{gamma}(\alpha_{01i},\alpha_{02i}) \text{ for } i=1,\ldots,p \\
R &\sim \text{Uniform}(C_p) \text{ where } C_p= \text{ the set of } p\times p \text{ correlation matrices.}
\end{aligned}
$$

For the conditional prior on $\mathcal{B}\,|\,\Sigma$ we will follow Section 4.2 and recall that this just generalizes the approach of Section 4.1. So, we suppose there are design matrices $V_1,\ldots,V_p$ and bounds $(m_{11i},m_{12i})\times\cdots\times(m_{k1i},m_{k2i})$ such that

$$\boldsymbol{\theta}_i=V_i\boldsymbol{\beta}_i\in(m_{11i},m_{12i})\times\cdots\times(m_{k1i},m_{k2i})$$

holds with virtual certainty. This determines $\boldsymbol{\theta}_{i0}$ and diagonal matrix $\Lambda_{0i}\in\mathbb{R}^{p\times p}$, using

$$\lambda_{0ij}=\frac{m_{2ij}-m_{1ij}}{2\sqrt{\alpha_{02j}/\alpha_{01j}}H_{2\alpha_{01j}}^{-1}\left(\frac{1+\gamma}{2}\right)},$$

so

$$\boldsymbol{\theta}_i\,|\,\Sigma\sim N_k(\boldsymbol{\theta}_{i0},\Lambda_{0i}\Sigma\Lambda_{0i}).$$

which implies

$$\boldsymbol{\beta}_i\,|\,\Sigma\sim N_k(\boldsymbol{\beta}_{0i},V_{0i})$$

where $\boldsymbol{\beta}_{0i}=V_i^{-1}\boldsymbol{\theta}_{i0}$ and $V_{0i}=V_i^{-1}\Lambda_{0i}\Sigma\Lambda_{0i}(V_i^{-1})'$. Also, it is assumed that the $\boldsymbol{\beta}_i$ are conditionally independent given Σ. Therefore, the prior on $\mathcal{B}\,|\,\Sigma$ is proportional to

$$|\Sigma|^{-p/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{p}\left(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{0i}\right)'V_{0i}^{-1}\left(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{0i}\right)\right\}.$$

Combining the prior and the likelihood leads to the following analog of Theorem 3 where $\otimes$ denotes the Kronecker product.

Theorem 6. To approximate the expectation $E_{\Pi(\cdot\,|\,Y)}(h(\mathcal{B},\Xi))$ generate a sample $(\mathcal{B}_i,\Xi_i)$ for $i=1,\ldots,N$ where

$$\begin{aligned}\Xi_i &\sim W_p(S^{-1},n-p-1),\\ \mathcal{B}_i\,|\,\Xi_i &\sim N_{k\times p}(B,(X'X)^{-1}\otimes\Sigma)\end{aligned}\tag{20}$$

and evaluate the estimate

$$\hat{E}_{\Pi(\cdot\,|\,Y)}(h(\boldsymbol{\mu},\Xi))=\frac{\sum_{i=1}^N h(\mathcal{B}_i,\Xi_i)k(\mathcal{B}_i,\Xi_i)}{\sum_{i=1}^N k(\mathcal{B}_i,\Xi_i)}\tag{21}$$

where

$$\begin{aligned}k(\boldsymbol{\mu},\Xi) &= \exp\left\{-\frac{1}{2}\sum_{i=1}^p(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{0i})'\,V_{0i}^{-1}\,(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{0i})\right\}\times\\ &\quad\prod_{i=1}^k\left(\frac{1}{\sigma_{ii}}\right)^{\alpha_{01i}+(p+1)/2}\exp\left(-\frac{\alpha_{02i}}{\sigma_{ii}}\right).\end{aligned}$$

Appendix

Some preliminary results are required before proving Theorem 1.

Lemma 1. For sample $Y=(\mathbf{y}_1\cdots\mathbf{y}_n)'\in\mathbb{R}^{n\times p}$, $\mathbf{a}\in\mathbb{R}^p$ and $A\in\mathbb{R}^{p\times p}$ then

$$\begin{aligned}\sum_{i=-1}^n(\mathbf{y}_i-\mathbf{a})'A(\mathbf{y}_i-a) &= n(\bar{\mathbf{y}}-\mathbf{a})'A(\bar{\mathbf{y}}-a)+\sum_{i=-1}^n(\mathbf{y}_i-\bar{\mathbf{y}})'A(\mathbf{y}_i-\bar{\mathbf{y}})\\ &= n(\bar{\mathbf{y}}-\mathbf{a})'A(\bar{\mathbf{y}}-a)+\operatorname{tr}\{(Y-\mathbf{1}\bar{\mathbf{y}}')'A(Y-\mathbf{1}\bar{\mathbf{y}}')\}\\ &= n(\bar{\mathbf{y}}-\mathbf{a})'A(\bar{\mathbf{y}}-a)+\operatorname{tr}\{(Y-\mathbf{1}\bar{\mathbf{y}}')(Y-\mathbf{1}\bar{\mathbf{y}}')'A\}\\ &= n(\bar{\mathbf{y}}-\mathbf{a})'A(\bar{\mathbf{y}}-a)+\operatorname{tr}\{SA\}.\end{aligned}$$

Then by Lemma 1 the likelihood for $(\boldsymbol{\mu},\Sigma)$ based on data Y is given by

$$L(\boldsymbol{\mu},\Sigma\,|\,Y)\propto|\Sigma|^{-n/2}\exp\left\{-\frac{n}{2}(\bar{\mathbf{y}}-\boldsymbol{\mu})'\Sigma^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu})\right\}\operatorname{etr}\left\{-\frac{1}{2}S\Sigma^{-1}\right\}.$$

We make use of the following well-known result.

Lemma 2. If $A_1,A_2\in\mathbb{R}^{p\times p}$ are symmetric with A_1+A_2 invertible and $\mathbf{a}_1,\mathbf{a}_2\in\mathbb{R}^p$, then

$$\begin{aligned}&(\mathbf{x}-\mathbf{a}_1)'A_1(\mathbf{x}-\mathbf{a}_1)+(\mathbf{x}-\mathbf{a}_2)'A_2(\mathbf{x}-\mathbf{a}_2)\\ =\;&(\mathbf{x}-(A_1+A_2)^{-1}(A_1\mathbf{a}_1+A_2\mathbf{a}_2))'(A_1+A_2)(\cdot)+\\ &(\mathbf{a}_1-\mathbf{a}_2)'A_1(A_1+A_2)^{-1}A_2(\mathbf{a}_1-\mathbf{a}_2).\end{aligned}$$

Proof of Theorem 1

The posterior is proportional to

$$
\begin{aligned}
& L(\boldsymbol{\mu}, \Sigma \,|\, Y)\pi(\boldsymbol{\mu} \,|\, \Sigma)\pi(\Sigma^{-1}) \\
\propto\ & |\Sigma|^{-n/2} \exp\left\{-\frac{n}{2}(\bar{\mathbf{y}}-\boldsymbol{\mu})'\Sigma^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu})\right\} \\
& \times|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2\lambda_0^2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right\} \times \\
& \operatorname{etr}\left\{-\frac{1}{2}S\Sigma^{-1}\right\} |\Sigma^{-1}|^{(f_0-p-1)/2}\operatorname{etr}\left\{-\frac{1}{2}\Sigma_0^{-1}\Sigma^{-1}\right\} \\
=\ & |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2\lambda_0^2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \frac{n}{2}(\bar{\mathbf{y}}-\boldsymbol{\mu})'\Sigma^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu})\right\} \times \\
& |\Sigma^{-1}|^{(n+f_0-p-1)/2}\operatorname{etr}\left\{-\frac{1}{2}(\Sigma_0^{-1}+S)\Sigma^{-1}\right\}.
\end{aligned}
$$

By Lemma 2, with $A_1 = (\lambda_0^2\Sigma)^{-1}, A_2 = (\Sigma/n)^{-1}, A_1 + A_2 = (1/\lambda_0^2 + n)\Sigma^{-1}, \mathbf{a}_1 = \bar{\mathbf{y}}, \mathbf{a}_2 = \boldsymbol{\mu}_0$

$$
\begin{aligned}
& \frac{1}{\lambda_0^2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) + n(\bar{\mathbf{y}}-\boldsymbol{\mu})'\Sigma^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu}) \\
=\ & (\boldsymbol{\mu} - (1/\lambda_0^2+n)^{-1}\Sigma(\Sigma^{-1}\boldsymbol{\mu}_0/\lambda_0^2 + n\Sigma^{-1}\bar{\mathbf{y}}))'(1/\lambda_0^2+n)\Sigma^{-1}(\cdot) + \\
& (\bar{\mathbf{y}}-\boldsymbol{\mu}_0)'(\lambda_0^2\Sigma)^{-1}(1/\lambda_0^2+n)^{-1}\Sigma(\Sigma/n)^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu}_0) \\
=\ & (\boldsymbol{\mu} - (1/\lambda_0^2+n)^{-1}(\boldsymbol{\mu}_0/\lambda_0^2 + n\bar{\mathbf{y}}))'(1/\lambda_0^2+n)\Sigma^{-1}(\cdot) + \\
& (n/\lambda_0^2)(1/\lambda_0^2+n)^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu}_0)'\Sigma^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu}_0) \\
=\ & (\boldsymbol{\mu} - (1/\lambda_0^2+n)^{-1}(\boldsymbol{\mu}_0/\lambda_0^2 + n\bar{\mathbf{y}}))'(1/\lambda_0^2+n)\Sigma^{-1}(\cdot) + \\
& (n/(1+n\lambda_0^2))\operatorname{tr}(\bar{\mathbf{y}}-\boldsymbol{\mu}_0)(\bar{\mathbf{y}}-\boldsymbol{\mu}_0)'\Sigma^{-1}.
\end{aligned}
$$

Therefore, the posterior is proportional to

$$
\begin{aligned}
& |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - (1/\lambda_0^2+n)^{-1}(\boldsymbol{\mu}_0/\lambda_0^2 + n\bar{\mathbf{y}}))'(1/\lambda_0^2+n)\Sigma^{-1}(\cdot)\right\} \times \\
& |\Sigma^{-1}|^{(n+f_0-p-1)/2}\operatorname{etr}\left\{-\frac{1}{2}\left[\Sigma_0^{-1} + S + \frac{n}{1+n\lambda_0^2}(\bar{\mathbf{y}}-\boldsymbol{\mu}_0)(\bar{\mathbf{y}}-\boldsymbol{\mu}_0)'\right]\Sigma^{-1}\right\}
\end{aligned}
$$

which gives the result. ■

References

Evans, M. and Swartz, T. (2000) Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford University Press.

Muirhead, R. (1982) Aspects of Multivariate Statistical Theory. John Wiley.

Rubin,D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In Bayesian Statistics 3 (J.M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 395-402. Oxford Univ. Press.