Bayesian Linear Model: Gory Details
Adding more details

Irving Gómez Méndez

These notes are based on the original article of Banerjee (2008), I just made some corrections and extended the explanation in some parts. As some of my students have pointed it out to me, some results can be obtained faster with clever observations (clever observations made by my students).

1 Previous results

Identity of Sherman-Woodbury-Morrison

$$(A+BDC)^{-1} = A^{-1} - A^{-1}B(D^{-1}+CA^{-1}B)^{-1}CA^{-1},$$

A and D square invertible matrices, B and C can be rectangular matrices.

Proof: Multiplying the right side by $A+BDC$, we get:

$$\begin{aligned}
&(A+BDC)\left[A^{-1} - A^{-1}B(D^{-1}+CA^{-1}B)^{-1}CA^{-1}\right]\\
=&(I+BDCA^{-1}) - B(D^{-1}+CA^{-1}B)^{-1}CA^{-1} - BDCA^{-1}B(D^{-1}+CA^{-1}B)^{-1}CA^{-1}\\
=&(I+BDCA^{-1}) - (B+BDCA^{-1}B)(D^{-1}+CA^{-1}B)^{-1}CA^{-1}\\
=&(I+BDCA^{-1}) - BD(D^{-1}+CA^{-1}B)(D^{-1}+CA^{-1}B)^{-1}CA^{-1}\\
=&I+BDCA^{-1} - BDCA^{-1}\\
=&I
\end{aligned}$$

Identity of Sherman-Woodbury-Morrison for determinants

$$|A+BDC| = |A||D||D^{-1}+CA^{-1}B|$$

Proof: See theorem 10.11 of Banerjee and Roy (2014).

Ellipsoidal rectification or multivariate completing squares

If A is a symmetric positive definite matrix (and hence invertible), then

$$u^TAu - 2\alpha^Tu = (u-A^{-1}\alpha)^TA(u-A^{-1}\alpha) - \alpha^TA^{-1}\alpha$$

Proof: It is immediate expanding $(u - A^{-1}\alpha)^T A(u - A^{-1}\alpha)$.

2 Model and the prior distributions

Consider the usual model of regression

$$Y = X^T\beta + \varepsilon,$$

with $\varepsilon|X \sim \mathcal{N}(0, \sigma^2)$ and $\varepsilon \perp\!\!\!\perp \beta|X$.

Assume that we have a sample $\mathcal{D}_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ with $(X_i, Y_i) \overset{iid}{\sim} (X, Y)$. Thus, the model might be written as

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n).$$

The parameters of the model are β and σ. Consider the following prior distributions

$$\sigma^2 \sim IG(a, b), \quad a, b > 0,$$

i.e.

$$p(\sigma^2) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left\{-\frac{b}{\sigma^2}\right\} \mathbb{1}_{(0,\infty)}(\sigma^2).$$

On the other hand, consider

$$\beta|\sigma^2 \sim \mathcal{N}_p(\mu_\beta, \sigma^2 V_\beta),$$

$\mu_\beta \in \mathbb{R}^p$ and V_β symmetric and positive definite matrix.

In this case, we say that the joint distribution of β and σ^2 is Normal-Inverse Gamma with parameters μ_β, V_β, a and b, and it's denoted as

$$\beta, \sigma^2 \sim NIG(\mu_\beta, V_\beta, a, b).$$

To get the density of a $NIG(\mu_\beta, V_\beta, a, b)$ distribution we express the joint distribution in the following way:

$$p(\beta, \sigma^2) = p(\beta|\sigma^2)p(\sigma^2)$$

$$= \frac{b^a}{(2\pi)^{p/2}|V_\beta|^{1/2}\Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a+\frac{p}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left[b + \frac{1}{2}(\beta - \mu_\beta)^T V_\beta^{-1}(\beta - \mu_\beta)\right]\right\} \mathbb{1}_{(0,\infty)}(\sigma^2)\mathbb{1}_{\mathbb{R}^p}(\beta)$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{a+\frac{p}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left[b + \frac{1}{2}(\beta - \mu_\beta)^T V_\beta^{-1}(\beta - \mu_\beta)\right]\right\} \mathbb{1}_{(0,\infty)}(\sigma^2)\mathbb{1}_{\mathbb{R}^p}(\beta) \qquad (1)$$

First, we deduce the non-conditional prior distribution of β. To do so, note that from the definition of the gamma function, we have

$$\Gamma(z) = \int_0^\infty x^{z-1} \exp\{-x\} dx,$$

let be $x = \frac{1}{\sigma^2} w$ and $w = b + \frac{1}{2}(\beta - \mu_\beta)^T V_\beta^{-1}(\beta - \mu_\beta)$. Note that

$$\frac{dx}{d\sigma^2} = -\frac{1}{(\sigma^2)^2} w.$$

moreover, when $x \to 0$, $\sigma^2 \to \infty$ and when $x \to \infty$, $\sigma^2 \to 0$. Thus, with this change of variable, we get

$$\Gamma(z) = \int_0^\infty \left(\frac{w}{\sigma^2}\right)^{z-1} \exp\left\{-\frac{w}{\sigma^2}\right\} \frac{w}{(\sigma^2)^2} d\sigma^2$$

and

$$\Gamma(a+p/2) = \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{a+\frac{p}{2}+1} w^{a+\frac{p}{2}} \exp\left\{-\frac{w}{\sigma^2}\right\} d\sigma^2.$$

Now, marginalizing the joint density of β and σ^2 with respect to σ^2, we have

$$\begin{aligned} p(\beta) &= \frac{b^a}{(2\pi)^{p/2}|V_\beta|^{1/2}\Gamma(a)} \mathbb{1}_{\mathbb{R}^p}(\beta) \int\limits_0^\infty \left(\frac{1}{\sigma^2}\right)^{a+\frac{p}{2}+1} \exp\left\{-\frac{w}{\sigma^2}\right\} d\sigma^2 \\ &= \frac{b^a}{(2\pi)^{p/2}|V_\beta|^{1/2}\Gamma(a)} w^{-(a+\frac{p}{2})}\Gamma(a+p/2)\mathbb{1}_{\mathbb{R}^p}(\beta) \end{aligned}$$

Multivariate t distribution: We say that a random vector X has a multivariate t distribution with $\nu > 0$ degrees of freedom, localization vector $\mu \in \mathbb{R}^p$ and scale matrix Σ (positive definite matrix of dimension $p \times p$), which is denoted as $X \sim t_\nu(\mu, \Sigma)$, if its density is given by

$$\frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{p/2}\pi^{p/2}|\Sigma|^{1/2}} \left[1 + \frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]^{-(\nu+p)/2} \mathbb{1}_{\mathbb{R}^p}(x)$$

Note that the prior density of β can be expressed as

$$\begin{aligned} p(\beta) &= \frac{\Gamma\left(\frac{2a+p}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)(2a)^{p/2}\pi^{p/2}|\frac{1}{a}V_\beta|^{1/2}} b^a \left[b + \frac{1}{2}(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)\right]^{-\frac{2a+p}{2}} \mathbb{1}_{\mathbb{R}^p}(\beta) \\ &= \frac{\Gamma\left(\frac{2a+p}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)(2a)^{p/2}\pi^{p/2}|\frac{b}{a}V_\beta|^{1/2}} \left[1 + \frac{1}{2a}(\beta-\mu_\beta)^T \left(\frac{b}{a}V_\beta\right)^{-1}(\beta-\mu_\beta)\right]^{-\frac{2a+p}{2}} \mathbb{1}_{\mathbb{R}^p}(\beta), \end{aligned}$$

where we can recognize $\beta \sim t_{2a}\left(\mu_\beta, \frac{b}{a}V_\beta\right)$.

3 Likelihood

From the model, we have that $\mathbf{Y}|\mathbf{X}, \beta, \sigma^2 \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 I)$. Thus, the likelihood of the sample is given by

$$p(\mathbf{Y}|\mathbf{X}, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)\right\} \mathbb{1}_{\mathbb{R}^p}(\mathbf{Y})$$

4 Posterior distributions

Remember that the kernel of the posterior distribution might be identify from

$$p(\beta,\sigma^2|\mathcal{D}_n) \propto p(\mathbf{Y}|\mathbf{X},\beta,\sigma^2)p(\beta,\sigma^2)$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{a+\frac{n}{2}+\frac{p}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left[b+\frac{1}{2}(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)+\frac{1}{2}(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)\right]\right\} \mathbb{1}_{\mathbb{R}^p}(\beta)\mathbb{1}_{(0,\infty)}(\sigma^2) \tag{2}$$

On the other hand,

$$(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)+(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)$$

$$=\mu_\beta^T V_\beta^{-1}\mu_\beta+\mathbf{Y}^T\mathbf{Y}+\beta^T V_\beta^{-1}\beta+\beta^T\mathbf{X}^T\mathbf{X}\beta-2\mu_\beta^T V_\beta^{-1}\beta-2\mathbf{Y}^T\mathbf{X}\beta$$

$$=\mu_\beta^T V_\beta^{-1}\mu_\beta+\mathbf{Y}^T\mathbf{Y}+\beta^T\left(V_\beta^{-1}+\mathbf{X}^T\mathbf{X}\right)\beta-2\left(\mu_\beta^T V_\beta^{-1}+\mathbf{Y}^T\mathbf{X}\right)\beta$$

Completing the multivariate square, we have

$$(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)+(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)$$

$$=\mu_\beta^T V_\beta^{-1}\mu_\beta+\mathbf{Y}^T\mathbf{Y}+\left[\beta-\left(V_\beta^{-1}+\mathbf{X}^T\mathbf{X}\right)^{-1}\left(V_\beta^{-1}\mu_\beta+\mathbf{X}^T\mathbf{Y}\right)\right]^T\left(V_\beta^{-1}+\mathbf{X}^T\mathbf{X}\right)\times$$
$$\left[\beta-\left(V_\beta^{-1}+\mathbf{X}^T\mathbf{X}\right)^{-1}\left(V_\beta^{-1}\mu_\beta+\mathbf{X}^T\mathbf{Y}\right)\right]-\left(V_\beta^{-1}\mu_\beta+\mathbf{X}^T\mathbf{Y}\right)^T\left(V_\beta^{-1}+\mathbf{X}^T\mathbf{X}\right)^{-1}\left(V_\beta^{-1}\mu_\beta+\mathbf{X}^T\mathbf{Y}\right)$$

Let be

$$\mu^\star=\left(V_\beta^{-1}+\mathbf{X}^T\mathbf{X}\right)^{-1}\left(V_\beta^{-1}\mu_\beta+\mathbf{X}^T\mathbf{Y}\right)$$

y

$$V^\star=\left(V_\beta^{-1}+\mathbf{X}^T\mathbf{X}\right)^{-1},$$

then

$$(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)+(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)=\mu_\beta^T V_\beta^{-1}\mu_\beta+\mathbf{Y}^T\mathbf{Y}-\mu^{\star T}V^{\star-1}\mu^\star+(\beta-\mu^\star)^T V^{\star-1}(\beta-\mu^\star) \tag{3}$$

Substituting the Equation (3) in the Equation (2), we get

$$p(\beta,\sigma^2|\mathcal{D}_n)$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{a+\frac{n}{2}+\frac{p}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left[b+\frac{1}{2}(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)+\frac{1}{2}(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)\right]\right\} \mathbb{1}_{\mathbb{R}^p}(\beta)\mathbb{1}_{(0,\infty)}(\sigma^2)$$

$$=\left(\frac{1}{\sigma^2}\right)^{a^\star+\frac{p}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left[b^\star+\frac{1}{2}(\beta-\mu^\star)^T V^{\star-1}(\beta-\mu^\star)\right]\right\} \mathbb{1}_{\mathbb{R}^p}(\beta)\mathbb{1}_{(0,\infty)}(\sigma^2),$$

where

$$a^\star = a + \frac{n}{2}$$

and

$$b^\star = b + \frac{1}{2}\left[\mu_\beta^T V_\beta^{-1}\mu_\beta + \mathbf{Y}^T\mathbf{Y} - \mu^{\star T} V^{\star -1}\mu^\star\right]. \tag{4}$$

Comparing this last expression with the kernel of the NIG distribution in the expression (1), we deduce that

$$\beta, \sigma^2 | \mathcal{D}_n \sim NIG(\mu^\star, V^\star, a^\star, d^\star).$$

Furthermore, we can observe that

$$p(\beta|\sigma^2, \mathcal{D}_n) \propto \exp\left\{-\frac{1}{2\sigma^2}(\beta - \mu^\star)^T V^{\star -1}(\beta - \mu^\star)\right\} \mathbb{1}_{\mathbb{R}^p}(\beta),$$

i.e.

$$\beta|\sigma^2, \mathcal{D}_n \sim \mathcal{N}(\mu^\star, \sigma^2 V^\star).$$

Where we can conclude that

$$\sigma^2|\mathcal{D}_n \sim IG(a^\star, b^\star)$$

and

$$\beta|\mathcal{D}_n \sim t_{2a^\star}\left(\mu^\star, \frac{b^\star}{a^\star}V^\star\right)$$

5 A useful expression from $b^\star$

To get a useful expression for $b^\star$, first consider the expression between square brackets in the Equation (4), where we observe that

$$\begin{aligned}
&\mathbf{Y}^T\mathbf{Y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \mu^{\star T} V^{\star -1}\mu^\star \\
=&\mathbf{Y}^T\mathbf{Y} + \mu_\beta^T V_\beta^{-1}\mu_\beta - \left(V_\beta^{-1}\mu_\beta + \mathbf{X}^T\mathbf{Y}\right)^T V^\star \left(V_\beta^{-1}\mu_\beta + \mathbf{X}^T\mathbf{Y}\right) \\
=&\mathbf{Y}^T\left(I - \mathbf{X}V^\star\mathbf{X}^T\right)\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}V^\star V_\beta^{-1}\mu_\beta + \mu_\beta^T\left(V_\beta^{-1} - V_\beta^{-1}V^\star V_\beta^{-1}\right)\mu_\beta
\end{aligned} \tag{5}$$

On the other hand,

$$V_\beta^{-1} - V_\beta^{-1}V^\star V_\beta^{-1} = V_\beta^{-1} - V_\beta^{-1}\left(V_\beta^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1} V_\beta^{-1},$$

using the identity of Sherman-Woodbury-Morrison with $A = V_\beta$, $B = C = I$ y $D = (\mathbf{X}^T\mathbf{X})^{-1}$, we get

$$V_\beta^{-1} - V_\beta^{-1}V^\star V_\beta^{-1} = \left(V_\beta + (\mathbf{X}^T\mathbf{X})^{-1}\right)^{-1}$$

and applying again the identity of Sherman-Woodbury-Morrison with $A = (\mathbf{X}^T\mathbf{X})^{-1}$, $B = C = I$ and $D = V_\beta$, we get

$$\begin{aligned}
V_\beta^{-1} - V_\beta^{-1}V^\star V_\beta^{-1} &= (\mathbf{X}^T\mathbf{X}) - (\mathbf{X}^T\mathbf{X})\left(\mathbf{X}^T\mathbf{X} + V_\beta^{-1}\right)^{-1}(\mathbf{X}^T\mathbf{X}) \\
&= \mathbf{X}^T\left(I - \mathbf{X}V^\star\mathbf{X}^T\right)\mathbf{X}
\end{aligned} \tag{6}$$

Moreover, from the definition of $V^\star$, we observe that

$$V^\star\left(V_\beta^{-1}+\mathbf{X}^T\mathbf{X}\right)=I\Rightarrow V^\star V_\beta^{-1}=I-V^\star\mathbf{X}^T\mathbf{X}$$

y

$$\begin{aligned}\mathbf{X}V^\star V_\beta^{-1}&=\mathbf{X}-\mathbf{X}V^\star\mathbf{X}^T\mathbf{X}\\&=\left(I-\mathbf{X}V^\star\mathbf{X}^T\right)\mathbf{X}\end{aligned}\tag{7}$$

Thus, substituting (6) and (7) in (5), we get

$$\begin{aligned}&\mathbf{Y}^T\mathbf{Y}+\mu_\beta^TV_\beta^{-1}\mu_\beta-\mu^{\star T}V^{\star-1}\mu^\star\\=&\mathbf{Y}^T\left(I-\mathbf{X}V^\star\mathbf{X}^T\right)\mathbf{Y}-2\mathbf{Y}^T\left(I-\mathbf{X}V^\star\mathbf{X}^T\right)\mathbf{X}\mu_\beta+(\mathbf{X}\mu_\beta)^T\left(I-\mathbf{X}V^\star\mathbf{X}^T\right)(\mathbf{X}\mu_\beta)\\=&(\mathbf{Y}-\mathbf{X}\mu_\beta)^T\left(I-\mathbf{X}V^\star\mathbf{X}^T\right)(\mathbf{Y}-\mathbf{X}\mu_\beta),\end{aligned}$$

applying again the identity of Sherman-Woodbury-Morrison, with $A=I$, $B=\mathbf{X}$, $C=\mathbf{X}^T$ and $D=V_\beta$, then

$$\begin{aligned}\left(I+\mathbf{X}V_\beta\mathbf{X}^T\right)^{-1}&=I-\mathbf{X}\left(\mathbf{X}^T\mathbf{X}+V_\beta^{-1}\right)^{-1}\mathbf{X}^T\\&=I-\mathbf{X}V^\star\mathbf{X}^T.\end{aligned}$$

That is,

$$\mathbf{Y}^T\mathbf{Y}+\mu_\beta^TV_\beta^{-1}\mu_\beta-\mu^{\star T}V^{\star-1}\mu^\star=(\mathbf{Y}-\mathbf{X}\mu_\beta)^T\left(I+\mathbf{X}V_\beta\mathbf{X}^T\right)^{-1}(\mathbf{Y}-\mathbf{X}\mu_\beta).$$

Consider again $b^\star$, defined in the expression(4), we have that

$$\begin{aligned}b^\star&=b+\frac{1}{2}\left[\mu_\beta^TV_\beta^{-1}\mu_\beta+\mathbf{Y}^T\mathbf{Y}-\mu^{\star T}V^{\star-1}\mu^\star\right]\\&=b+\frac{1}{2}(\mathbf{Y}-\mathbf{X}\mu_\beta)^T\left(I+\mathbf{X}V_\beta\mathbf{X}^T\right)^{-1}(\mathbf{Y}-\mathbf{X}\mu_\beta).\end{aligned}\tag{8}$$

Furthermore, from the expression (4), we get

$$\begin{aligned}&b^\star+\frac{1}{2}(\beta-\mu^\star)^TV^{\star-1}(\beta-\mu^\star)\\=&b+\frac{1}{2}\left[\mu_\beta^TV_\beta^{-1}\mu_\beta+\mathbf{Y}^T\mathbf{Y}-\mu^{\star T}V^{\star-1}\mu^\star+(\beta-\mu^\star)^TV^{\star-1}(\beta-\mu^\star)\right]\end{aligned}$$

and using Equation (3), we have that

$$b^\star+\frac{1}{2}(\beta-\mu^\star)^TV^{\star-1}(\beta-\mu^\star)=b+\frac{1}{2}\left[(\beta-\mu_\beta)^TV_\beta^{-1}(\beta-\mu_\beta)+(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)\right].\tag{9}$$

On the other hand, adding $\frac{1}{2}(\beta-\mu^\star)^TV^{\star-1}(\beta-\mu^\star)$ in Equation (8), we get

$$b^\star+\frac{1}{2}(\beta-\mu^\star)^TV^{\star-1}(\beta-\mu^\star)=b+\frac{1}{2}\left[(\mathbf{Y}-\mathbf{X}\mu_\beta)^T\left(I+\mathbf{X}V_\beta\mathbf{X}^T\right)^{-1}(\mathbf{Y}-\mathbf{X}\mu_\beta)+(\beta-\mu^\star)^TV^{\star-1}(\beta-\mu^\star)\right].\tag{10}$$

Finally, comparing the right side of Equations (9) and (10), we conclude that

$$\begin{aligned}&(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)+(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)\\ =&(\mathbf{Y}-\mathbf{X}\mu_\beta)^T\left(I+\mathbf{X}V_\beta\mathbf{X}^T\right)^{-1}(\mathbf{Y}-\mathbf{X}\mu_\beta)+(\beta-\mu^\star)^T V^{\star-1}(\beta-\mu^\star)\end{aligned} \tag{11}$$

The relevance of Equation (11) lies in observing that while the left side of the equality has teh sum of two terms in which appears β, there is only one term with β in the right side of the equality. This observations is going to be exploited in the next section where we find the prior predictive distribution.

6 Prior predictive distribution

To find the prior predictive distribution, that is the distribution of $\mathbf{Y}|\mathbf{X}$, we first deduce the distribution of $\mathbf{Y}|\mathbf{X},\sigma^2$, which can be obtained marginalizing the joint distribution of $\mathbf{Y},\beta|\mathbf{X},\sigma^2$. That is, solving the next integral:

$$p(\mathbf{Y}|\mathbf{X},\sigma^2)=\int\limits_{\mathbb{R}^p} p(\mathbf{Y}|\mathbf{X},\beta,\sigma^2)p(\beta|\sigma^2)d\beta.$$

We know from the likelihood that

$$\mathbf{Y}|\mathbf{X},\beta,\sigma^2\sim\mathcal{N}_n(\mathbf{X}\beta,\sigma^2 I)$$

and from the prior conditional distribution of $\beta|\sigma^2$, we have that

$$\beta|\sigma^2\sim\mathcal{N}_p(\mu_\beta,\sigma^2V_\beta).$$

Therefore,

$$p(\mathbf{Y}|\mathbf{X},\sigma^2)=\frac{1}{(2\pi\sigma^2)^{(n+p)/2}|V_\beta|^{1/2}}\int\limits_{\mathbb{R}^p}\exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)+(\beta-\mu_\beta)^TV_\beta^{-1}(\beta-\mu_\beta)\right]\right\}d\beta,$$

using Equation (11),

$$\begin{aligned}&p(\mathbf{Y}|\mathbf{X},\sigma^2)\\ =&\frac{1}{(2\pi\sigma^2)^{(n+p)/2}|V_\beta|^{1/2}}\int\limits_{\mathbb{R}^p}\exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{Y}-\mathbf{X}\mu_\beta)^T\left(I+\mathbf{X}V_\beta\mathbf{X}^T\right)^{-1}(\mathbf{Y}-\mathbf{X}\mu_\beta)+(\beta-\mu^\star)^TV^{\star-1}(\beta-\mu^\star)\right]\right\}d\beta\\ =&\frac{1}{(2\pi\sigma^2)^{n/2}|V_\beta|^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{Y}-\mathbf{X}\mu_\beta)^T\left(I+\mathbf{X}V_\beta\mathbf{X}^T\right)^{-1}(\mathbf{Y}-\mathbf{X}\mu_\beta)\right]\right\}\times\\ &\qquad\int\limits_{\mathbb{R}^p}\frac{1}{(2\pi\sigma^2)^{p/2}}\exp\left\{-\frac{1}{2\sigma^2}\left[(\beta-\mu^\star)^TV^{\star-1}(\beta-\mu^\star)\right]\right\}d\beta\\ =&\frac{|V^\star|^{1/2}}{(2\pi\sigma^2)^{n/2}|V_\beta|^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{Y}-\mathbf{X}\mu_\beta)^T\left(I+\mathbf{X}V_\beta\mathbf{X}^T\right)^{-1}(\mathbf{Y}-\mathbf{X}\mu_\beta)\right]\right\},\end{aligned}$$

note that

$$\frac{|V^\star|}{|V_\beta|} = \frac{|(V_\beta^{-1} + \mathbf{X}^T\mathbf{X})^{-1}|}{|V_\beta|} = \frac{1}{|V_\beta||(V_\beta^{-1} + \mathbf{X}^T\mathbf{X})|},$$

on the other hand, using the identity of Sherman-Woodbury-Morrison for determinants, we have that

$$|I + \mathbf{X}V_\beta\mathbf{X}^T| = |V_\beta||V_\beta^{-1} + \mathbf{X}^T\mathbf{X}|,$$

making $A = I$, $B = \mathbf{X}$, $D = V_\beta$ and $C = \mathbf{X}^T$.

Thus,

$$p(\mathbf{Y}|\mathbf{X}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}|I + \mathbf{X}V_\beta\mathbf{X}^T|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{Y} - \mathbf{X}\mu_\beta)^T\left(I + \mathbf{X}V_\beta\mathbf{X}^T\right)^{-1}(\mathbf{Y} - \mathbf{X}\mu_\beta)\right]\right\},$$

that is

$$\mathbf{Y}|\mathbf{X}, \sigma^2 \sim \mathcal{N}_n\left(\mathbf{X}\mu_\beta, \sigma^2(I + \mathbf{X}V_\beta\mathbf{X}^T)\right).$$

To get the prior predictive distribution, that is the distribution of $\mathbf{Y}|\mathbf{X}$, we marginalize the joint distribution of $\mathbf{Y}, \sigma^2|\mathbf{X}$, that is, we solve the next integral

$$p(\mathbf{Y}|\mathbf{X}) = \int_0^\infty p(\mathbf{Y}|\mathbf{X}, \sigma^2)p(\sigma^2)d\sigma^2.$$

Here, it is convenient to remember that when we deduced the prior distribution of β, we marginalized the joint distribution of β and σ^2, that is, to find its distribution we solve the following integral

$$p(\beta) = \int_0^\infty p(\beta|\sigma^2)p(\sigma^2)d\sigma^2$$

considering that

$$\sigma^2 \sim IG(a, b),$$
$$\beta|\sigma^2 \sim \mathcal{N}_p(\mu_\beta, \sigma^2V_\beta).$$

Where we deduced that $\beta \sim t_{2a}\left(\mu_\beta, \frac{b}{a}V_\beta\right)$. Therefore, by analogy we deduce that

$$\mathbf{Y}|\mathbf{X} \sim t_{2a}\left(\mathbf{X}\mu_\beta, \frac{b}{a}(I + \mathbf{X}V_\beta\mathbf{X}^T)\right).$$

7 Prior predictive distribution, the easy way

Note that the model can be written in the following way

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon_1}, \quad \boldsymbol{\varepsilon_1} \sim \mathcal{N}(0, \sigma^2I) \text{ y } \boldsymbol{\varepsilon_1} \perp\!\!\!\perp \beta,$$
$$\beta = \mu_\beta + \boldsymbol{\varepsilon_2}, \quad \boldsymbol{\varepsilon_2} \sim \mathcal{N}(0, \sigma^2V_\beta) \text{ y } \boldsymbol{\varepsilon_2} \perp\!\!\!\perp \boldsymbol{\varepsilon_1},$$

that is,

$$\mathbf{Y} = \mathbf{X}\mu_\beta + \mathbf{X}\boldsymbol{\varepsilon_2} + \boldsymbol{\varepsilon_1}.$$

This means that $\mathbf{Y}|\sigma^2$ can be expressed as the sum of two independent normal distributions plus a constant, where we can conclude easily that

$$\mathbf{Y}|\sigma^2 \sim \mathcal{N}\left(\mathbf{X}\mu_\beta, \sigma^2(I + \mathbf{X}V_\beta\mathbf{X}^T)\right)$$

8 Posterior predictive distribution

To deduce the posterior predictive distribution, we can marginalized the joint distribution of $\mathbf{Y}|\mathbf{X},\beta,\sigma,\mathcal{D}_n$, which means to solve the next multiple integral

$$p(\mathbf{Y}|\mathbf{X},\mathcal{D}_n)=\int\limits_0^\infty\int\limits_{\mathbb{R}^p}p(\mathbf{Y}|\mathbf{X},\beta,\sigma^2)p(\beta,\sigma^2|\mathcal{D}_n)d\beta d\sigma^2,$$

where

$$\begin{aligned}\beta,\sigma^2|\mathcal{D}_n&\sim NIG\left(\mu^\star,V^\star,a^\star,b^\star\right)\\ \mathbf{Y}|\mathbf{X},\beta,\sigma^2&\sim\mathcal{N}_n\left(\mathbf{X}\beta,\sigma^2I\right).\end{aligned}$$

Analogously, we got the prior predictive distribution, solving

$$p(\mathbf{Y}|\mathbf{X})=\int\limits_0^\infty\int\limits_{\mathbb{R}^p}p(\mathbf{Y}|\mathbf{X},\beta,\sigma^2)p(\beta,\sigma^2)d\beta d\sigma^2,$$

where

$$\begin{aligned}\beta,\sigma^2|\mathcal{D}_n&\sim NIG\left(\mu_\beta,V_\beta,a,b\right)\\ \mathbf{Y}|\mathbf{X},\beta,\sigma^2&\sim\mathcal{N}_n\left(\mathbf{X}\beta,\sigma^2I\right).\end{aligned}$$

We concluded that

$$\mathbf{Y}|\mathbf{X}\sim t_{2a}\left(\mathbf{X}\mu_\beta,\frac{b}{a}(I+\mathbf{X}V_\beta\mathbf{X}^T)\right),$$

by analogy, we deduce that

$$\mathbf{Y}|\mathbf{X},\mathcal{D}_n\sim t_{2a^\star}\left(\mathbf{X}\mu^\star,\frac{b^\star}{a^\star}(I+\mathbf{X}V^\star\mathbf{X}^T)\right),$$

from the same analogy, we deduce that

$$\mathbf{Y}|\mathbf{X},\sigma^2,\mathcal{D}_n\sim\mathcal{N}_n\left(\mathbf{X}\mu^\star,\sigma^2(I+\mathbf{X}V^\star\mathbf{X}^T)\right)$$

9 Summary

In summary, we have the next results.

9.1 Prior distributions

$$\begin{aligned}\sigma^2&\sim IG(a,b)\\ \beta|\sigma^2&\sim\mathcal{N}(\mu_\beta,\sigma^2V_\beta)\\ \sigma^2|\beta&\sim IG\left(a+\frac{p}{2},b+\frac{1}{2}(\beta-\mu_\beta)^TV_\beta^{-1}(\beta-\mu_\beta)\right)\\ \beta,\sigma^2&\sim NIG(\mu_\beta,V_\beta,a,b)\\ \beta&\sim t_{2a}\left(\mu_\beta,\frac{b}{a}V_\beta\right)\end{aligned}$$

9.2 Posterior distributions

$$\begin{aligned}
\sigma^2|\mathcal{D}_n &\sim IG(a^\star, b^\star) \\
\beta|\sigma^2, \mathcal{D}_n &\sim \mathcal{N}(\mu^\star, \sigma^2 V^\star) \\
\sigma^2|\beta, \mathcal{D}_n &\sim IG\left(a^\star + \frac{p}{2}, b^\star + \frac{1}{2}(\beta - \mu^\star)^T V^{\star -1}(\beta - \mu^\star)\right) \\
\beta, \sigma^2|\mathcal{D}_n &\sim NIG(\mu^\star, V^\star, a^\star, b^\star) \\
\beta|\mathcal{D}_n &\sim t_{2a^\star}\left(\mu^\star, \frac{b^\star}{a^\star} V^\star\right)
\end{aligned}$$

9.3 Prior predictive distributions

$$\begin{aligned}
\mathbf{Y}|\mathbf{X}_0, \beta, \sigma^2 &\sim \mathcal{N}_n(\mathbf{X}_0\beta, \sigma^2 I) \\
\mathbf{Y}|\mathbf{X}_0, \sigma^2 &\sim \mathcal{N}_n\left(\mathbf{X}_0\mu_\beta, \sigma^2(I + \mathbf{X}_0 V_\beta \mathbf{X}_0^T)\right) \\
\mathbf{Y}|\mathbf{X}_0 &\sim t_{2a}\left(\mathbf{X}_0\mu_\beta, \frac{b}{a}(I + \mathbf{X}_0 V_\beta \mathbf{X}_0^T)\right)
\end{aligned}$$

9.4 Posterior predictive distributions

$$\begin{aligned}
\mathbf{Y}|\mathbf{X}_0, \beta, \sigma^2, \mathcal{D}_n &\sim \mathcal{N}_n(\mathbf{X}_0\beta, \sigma^2 I) \\
\mathbf{Y}|\mathbf{X}_0, \sigma^2, \mathcal{D}_n &\sim \mathcal{N}_n\left(\mathbf{X}_0\mu^\star, \sigma^2(I + \mathbf{X}_0 V^\star \mathbf{X}_0^T)\right) \\
\mathbf{Y}|\mathbf{X}_0, \mathcal{D}_n &\sim t_{2a^\star}\left(\mathbf{X}_0\mu^\star, \frac{b^\star}{a^\star}(I + \mathbf{X}_0 V^\star \mathbf{X}_0^T)\right)
\end{aligned}$$

where

$$\begin{aligned}
V^\star &= \left(V_\beta^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1} \\
\mu^\star &= V^\star\left(V_\beta^{-1}\mu_\beta + \mathbf{X}^T\mathbf{Y}\right) \\
a^\star &= a + \frac{n}{2} \\
b^\star &= b + \frac{1}{2}\left[\mu_\beta^T V_\beta^{-1}\mu_\beta + \mathbf{Y}^T\mathbf{Y} - \mu^{\star T} V^{\star -1}\mu^\star\right]
\end{aligned}$$

References

Banerjee, Sudipto (2008). "Bayesian linear model: Gory details". In: *URL http://www. biostat. umn. edu/˜ ph7440/pubh7440/BayesianLinearModelGoryDetails. pdf.*

Banerjee, Sudipto and Anindya Roy (2014). *Linear algebra and matrix analysis for statistics.* Crc Press.