Bayesian Linear Model: Gory Details

By Sudipto Banerjee

Let $\mathbf{y} = [y_i]_{i=1}^n$ be an $n \times 1$ vector of independent observations on a dependent variable (or response) from n experimental units. Associated with the y_i, is a $p \times 1$ vector of regressors, say $\mathbf{x}_i$, and lead to the linear regression model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $X = [\mathbf{x}_i^T]_{i=1}^n$ is the $n \times p$ matrix of regressors with i-th row being $\mathbf{x}_i^T$ and is assumed fixed, $\boldsymbol{\beta}$ is the slope vector of regression coefficients and $\boldsymbol{\epsilon} = [\epsilon_i]_{i=1}^n$ is the vector of random variables representing "pure error" or measurement error in the dependent variable. For independent observations, we assume $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2 I_n)$, viz. that each component $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Furthermore, we will assume that the columns of the matrix X are linearly independent so that the rank of X is p.

1 The NIG conjugate prior family

A popular Bayesian model builds upon the linear regression of $\mathbf{y}$ using *conjugate priors* by specifying

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2) &= p(\boldsymbol{\beta} \,|\, \sigma^2)p(\sigma^2) = N(\boldsymbol{\mu}_\beta, \sigma^2 V_\beta) \times IG(a, b) = NIG(\boldsymbol{\mu}_\beta, V_\beta, a, b) \\ &= \frac{b^a}{(2\pi)^{p/2}|V_\beta|^{1/2}\Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a+p/2+1} \times \exp\left[-\frac{1}{\sigma^2}\left\{b + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^T V_\beta^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)\right\}\right] \\ &\propto \left(\frac{1}{\sigma^2}\right)^{a+p/2+1} \times \exp\left[-\frac{1}{\sigma^2}\left\{b + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^T V_\beta^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)\right\}\right], \end{aligned} \tag{2}$$

where $\Gamma(\cdot)$ represents the Gamma function and the $IG(a, b)$ prior density for σ^2 is given by

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

where a, $b > 0$. We call this the Normal-Inverse-Gamma (NIG) prior and denote it as $NIG(\boldsymbol{\mu}_\beta, V_\beta, a, b)$.

The NIG probability distribution is a joint probability distribution of a vector $\boldsymbol{\beta}$ and a scalar σ^2. If $(\boldsymbol{\beta}, \sigma^2) \sim NIG(\boldsymbol{\mu}, V, a, b)$, then an interesting analytic form results from integrating out σ^2

from the joint density:

$$\begin{aligned}
\int NIG(\boldsymbol{\mu},\, V,\, a,\, b)d\sigma^2 &= \frac{b^a}{(2\pi)^{p/2}|V|^{1/2}\Gamma(a)} \int \left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left\{-\frac{1}{\sigma^2}\left[b + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T V^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right]\right\} d\sigma^2 \\
&= \frac{b^a}{(2\pi)^{p/2}|V|^{1/2}\Gamma(a)} \int \exp\left\{-\frac{1}{\sigma^2}\left(b + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T V^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right)\right\} d\sigma^2 \\
&= \frac{b^a \Gamma\left(a+\frac{p}{2}\right)}{(2\pi)^{p/2}|V|^{1/2}\Gamma(a)} \left[b + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T V^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right]^{-\left(a+\frac{p}{2}\right)} \\
&= \frac{\Gamma\left(a+\frac{p}{2}\right)}{\pi^{p/2}|(2a)\frac{b}{a}V|^{1/2}\Gamma(a)} \left[1 + \frac{(\boldsymbol{\beta}-\boldsymbol{\mu})^T\left[\frac{b}{a}V\right]^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})}{2a}\right]^{-\left(\frac{2a+p}{2}\right)}.
\end{aligned}$$

This is a *multivariate t* density:

$$MVSt_\nu(\boldsymbol{\mu},\, \Sigma) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\pi^{p/2}|\nu\Sigma|^{1/2}} \left[1 + \frac{(\boldsymbol{\beta}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})}{\nu}\right]^{-\frac{\nu+p}{2}}, \tag{3}$$

with $\nu = 2a$ and $\Sigma = \left(\frac{b}{a}\right)V$.

2 The likelihood

The likelihood for the model is defined, up to proportionality, as the joint probability of observing the data given the parameters. Since X is fixed, the likelihood is given by

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) = N(X\boldsymbol{\beta}, \sigma^2 I) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})\right\}. \tag{4}$$

3 The posterior distribution from the NIG prior

Inference will proceed from the posterior distribution

$$p(\boldsymbol{\beta},\, \sigma^2 \mid \mathbf{y}) = \frac{p(\boldsymbol{\beta}, \sigma^2)p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)}{p(\mathbf{y})},$$

where $p(\mathbf{y}) = \int p(\boldsymbol{\beta}, \sigma^2)p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)d\boldsymbol{\beta}d\sigma^2$ is the marginal distribution of the data. The key to deriving the joint posterior distribution is the following easily verified *multivariate completion of squares* or *ellipsoidal rectification* identity:

$$\mathbf{u}^T A\mathbf{u} - 2\boldsymbol{\alpha}^T\mathbf{u} = (\mathbf{u} - A^{-1}\boldsymbol{\alpha})^T A(\mathbf{u} - A^{-1}\boldsymbol{\alpha}) - \boldsymbol{\alpha}^T A^{-1}\boldsymbol{\alpha}, \tag{5}$$

where A is a symmetric positive definite (hence invertible) matrix. An application of this identity immediately reveals,

$$\frac{1}{\sigma^2}\left[b+\frac{1}{2}\left\{(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)^T V_\beta(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)+(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})\right\}\right]=\frac{1}{\sigma^2}\left[b^*+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}^*)^T V^{*-1}(\boldsymbol{\beta}-\boldsymbol{\mu}^*)\right],$$

using which we can write the posterior as

$$p(\boldsymbol{\beta},\,\sigma^2\mid\mathbf{y})\propto\left(\frac{1}{\sigma^2}\right)^{a+(n+p)/2+1}\times\exp\left\{-\frac{1}{\sigma^2}\left[b^*+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}^*)^T V^{*-1}(\boldsymbol{\beta}-\boldsymbol{\mu}^*)\right]\right\},\tag{6}$$

where

$$\begin{aligned}\boldsymbol{\mu}^* &= (V_\beta^{-1}+X^TX)^{-1}(V_\beta^{-1}\boldsymbol{\mu}_\beta+X^T\mathbf{y}),\\ V^* &= (V^{-1}+X^TX)^{-1},\\ a^* &= a+n/2,\\ b^* &= b+\frac{1}{2}[\boldsymbol{\mu}_\beta^T V_\beta^{-1}\boldsymbol{\mu}_\beta+\mathbf{y}^T\mathbf{y}-\boldsymbol{\mu}^{*T}V^{*-1}\boldsymbol{\mu}^*].\end{aligned}$$

This posterior distribution is easily identified as a $NIG(\boldsymbol{\mu}^*, V^*, a^*, b^*)$ proving it to be a conjugate family for the linear regression model.

Note that the marginal posterior distribution of σ^2 is immediately seen to be an $IG(a^*, b^*)$ whose density is given by:

$$p(\sigma^2\mid\mathbf{y})=\frac{b^{*a^*}}{\Gamma(a^*)}\left(\frac{1}{\sigma^2}\right)^{a^*+1}\exp\left(-\frac{b^*}{\sigma^2}\right).\tag{7}$$

The marginal posterior distribution of $\boldsymbol{\beta}$ is obtained by integrating out σ^2 from the NIG joint posterior as follows:

$$\begin{aligned}p(\boldsymbol{\beta}|\mathbf{y}) &= \int p(\boldsymbol{\beta},\sigma^2\mid\mathbf{y})d\boldsymbol{\sigma}^2=\int NIG(\boldsymbol{\mu}^*,\,V^*,\,a^*,\,b^*)d\sigma^2\\ &\propto\int\left(\frac{1}{\sigma^2}\right)^{a^*+1}\exp\left\{-\frac{1}{\sigma^2}\left[b^*+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}^*)^T V^{*-1}(\boldsymbol{\beta}-\boldsymbol{\mu}^*)\right]\right\}d\sigma^2\\ &\propto\left[1+\frac{(\boldsymbol{\beta}-\boldsymbol{\mu}^*)^T V^{*-1}(\boldsymbol{\beta}-\boldsymbol{\mu}^*)}{2b^*}\right]^{-(a^*+p/2)}.\end{aligned}$$

This is a *multivariate t* density:

$$MVSt_{\nu^*}(\boldsymbol{\mu}^*,\,\Sigma^*)=\frac{\Gamma\left(\frac{\nu^*+p}{2}\right)}{\Gamma\left(\frac{\nu^*}{2}\right)\pi^{p/2}|\nu^*\Sigma^*|^{1/2}}\left[1+\frac{(\boldsymbol{\beta}-\boldsymbol{\mu}^*)^T\Sigma^{*-1}(\boldsymbol{\beta}-\boldsymbol{\mu}^*)}{\nu^*}\right]^{-\frac{\nu^*+p}{2}},\tag{8}$$

with $\nu^*=2a^*$ and $\Sigma^*=\left(\frac{b^*}{a^*}\right)V^*$.

4 A useful expression for the NIG scale parameter

Here we will prove:

$$b^* = b + \frac{1}{2}\left(\mathbf{y} - X\boldsymbol{\mu}_\beta\right)^T \left(I + XV_\beta X^T\right)^{-1}\left(\mathbf{y} - X\boldsymbol{\mu}_\beta\right) \tag{9}$$

On account of the expression for b^* derived in the preceding section, it suffices to prove that

$$\mathbf{y}^T\mathbf{y} + \boldsymbol{\mu}_\beta^T V^{-1}\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^* V^{*-1}\boldsymbol{\mu}^* = \left(\mathbf{y} - X\boldsymbol{\mu}_\beta\right)^T \left(I + XV_\beta X^T\right)^{-1}\left(\mathbf{y} - X\boldsymbol{\mu}_\beta\right)$$

Substituting $\boldsymbol{\mu}^* = V^*(V^{-1}\boldsymbol{\mu}_\beta + X^T\mathbf{y})$ in the left hand side above we obtain:

$$\begin{aligned}\mathbf{y}^T\mathbf{y} + \boldsymbol{\mu}_\beta^T V_\beta^{-1}\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^* V^{*-1}\boldsymbol{\mu}^* &= \mathbf{y}^T\mathbf{y} + \boldsymbol{\mu}_\beta^T V_\beta^{-1}\boldsymbol{\mu}_\beta - (V_\beta^{-1}\boldsymbol{\mu}_\beta + X^T\mathbf{y})V^*(V_\beta^{-1}\boldsymbol{\mu}_\beta + X^T\mathbf{y}) \\ &= \mathbf{y}^T(I - XV_\beta^* X^T)\mathbf{y} - 2\mathbf{y}^T XV^* V^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^T(V_\beta^{-1} - V_\beta^{-1}V^* V_\beta^{-1})\boldsymbol{\mu}.\end{aligned} \tag{10}$$

Further development of the proof will employ two tricky identities. The first is the well-known Sherman-Woodbury-Morrison identity in matrix algebra:

$$(A + BDC)^{-1} = A^{-1} - A^{-1}B\left(D^{-1} + CA^{-1}B\right)^{-1}CA^{-1}, \tag{11}$$

where A and D are square matrices that are invertible and B and C are rectangular (square if A and D have the same dimensions) matrices such that the multiplications are well-defined. This identity is easily verified by multiplying the right hand side with $A + BDC$ and simplifying to reduce it to the identity matrix.

Applying (11) twice, once with $A = V_\beta$ and $D = (X^T X)^{-1}$ to get the second equality and then with $A = (X^T X)^{-1}$ and $D = V_\beta$ to get the third equality, we have

$$\begin{aligned}V_\beta^{-1} - V_\beta^{-1}V^* V_\beta^{-1} &= V_\beta^{-1} - V_\beta^{-1}(V_\beta^{-1} + XX^T)^{-1}V_\beta^{-1} \\ &= [V_\beta + (X^T X)^{-1}]^{-1} \\ &= X^T X - X^T X(X^T X + V^{-1})^{-1}X^T X \\ &= X^T(I_n - XV^* X^T)X.\end{aligned} \tag{12}$$

The next identity notes that since $V^*(V^{-1} + X^T X) = I_p$, we have $V^* V^{-1} = I_p - V^* X^T X$, so that

$$XV^* V^{-1} = X - XV^* X^T X = (I_n - XV^* X^T)X. \tag{13}$$

Substituting (12) and (13) in (10) we obtain

$$\begin{aligned}\mathbf{y}^T(I_n - XV^*X^T)\mathbf{y} - 2\mathbf{y}^T(I_n - XV^*X^T)\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^T(I_n - XV^*X^T)\boldsymbol{\mu}_\beta \\ = (\mathbf{y} - X\boldsymbol{\mu}_\beta)^T(I_n - XV^*X^T)(\mathbf{y} - X\boldsymbol{\mu}_\beta) \\ = (\mathbf{y} - X\boldsymbol{\mu}_\beta)^T(I_n + XVX^T)^{-1}(\mathbf{y} - X\boldsymbol{\mu}_\beta), \qquad (14)\end{aligned}$$

where the last step is again a consequence of (11):

$$(I_n + XVX^T)^{-1} = I_n - X(V^{-1} + X^TX)^{-1}X^T = I_n - XV^*X^T.$$

5 Marginal distributions – the hard way

To obtain the marginal distribution of $\mathbf{y}$, we first compute the distribution $p(\mathbf{y} \,|\, \sigma^2)$ by integrating out $\boldsymbol{\beta}$ and subsequently integrate out σ^2 to obtain $p(\mathbf{y})$. To be precise, we use the expression for b^* derived in the preceding section, proceeding as below:

$$\begin{aligned}p(\mathbf{y} \,|\, \sigma^2) &= \int p(\mathbf{y} \,|\, \boldsymbol{\beta},\, \sigma^2)p(\boldsymbol{\beta} \,|\, \sigma^2)d\boldsymbol{\beta} = \int N(X\boldsymbol{\beta}, \sigma^2 I_n) \times N(\boldsymbol{\mu}_\beta, \sigma^2 V_\beta)d\boldsymbol{\beta} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n+p}{2}}\,|V_\beta|^{1/2}} \int \exp\left[-\frac{1}{2\sigma^2}\left\{(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)^T V_\beta^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)\right\}\right] d\boldsymbol{\beta} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n+p}{2}}\,|V_\beta|^{1/2}} \\ &\quad \times \int \exp\left[-\frac{1}{2\sigma^2}\left\{(\mathbf{y} - X\boldsymbol{\mu}_\beta)^T(I + XVX^T)^{-1}(\mathbf{y} - X\boldsymbol{\mu}_\beta) + (\boldsymbol{\beta} - \boldsymbol{\mu}^*)^T V^{*-1}(\boldsymbol{\beta} - \boldsymbol{\mu}^*)\right\}\right] d\boldsymbol{\beta} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n+p}{2}}\,|V_\beta|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\mu}_\beta)^T(I + XV_\beta X^T)^{-1}(\mathbf{y} - X\boldsymbol{\mu}_\beta)\right\} \\ &\quad \times \int \exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{\beta} - \boldsymbol{\mu}^*)^T V^{*-1}(\boldsymbol{\beta} - \boldsymbol{\mu}^*)\right\}\right] d\boldsymbol{\beta} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\left(\frac{|V^*|}{|V_\beta|}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\mu}_\beta)^T(I + XV_\beta X^T)^{-1}(\mathbf{y} - X\boldsymbol{\mu}_\beta)\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}\,|I + XV_\beta X^T|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\mu}_\beta)^T(I + XV_\beta X^T)^{-1}(\mathbf{y} - X\boldsymbol{\mu}_\beta)\right\} \\ &= N(X\boldsymbol{\mu}_\beta, \sigma^2(I + XV_\beta X^T)). \qquad (15)\end{aligned}$$

Here we have applied the matrix identity

$$|A + BDC| = |A||D||D^{-1} + CA^{-1}B| \qquad (16)$$

to obtain

$$|I_n + XV_\beta X^T| = |V_\beta||V_\beta^{-1} + X^T X| = \left(\frac{|V_\beta|}{|V^*|}\right).$$

Now, the marginal distribution of $p(\mathbf{y})$ is obtained by integrating a NIG density as follows:

$$\begin{aligned} p(\mathbf{y}) &= \int p(\mathbf{y} \,|\, \sigma^2) p(\sigma^2) d\sigma^2 = \int N(X\boldsymbol{\mu}_\beta, \sigma^2(I + XVX^T)) IG(a,b) d\sigma^2 \\ &= \int NIG(X\boldsymbol{\mu}_\beta, (I + XVX^T), a, b) d\sigma^2 = MVSt_{2a}\left(X\boldsymbol{\mu}, \frac{b}{a}(I + XVX^T)\right). \end{aligned} \tag{17}$$

Rewriting our result slightly differently reveals another useful property of the NIG density:

$$\begin{aligned} p(\mathbf{y}) &= \int p(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2 \\ &= \int N(X\boldsymbol{\beta}, \sigma^2 I_n) \times NIG(\boldsymbol{\mu}_\beta, V_\beta, a, b) d\boldsymbol{\beta} d\sigma^2 = MVSt_{2a}\left(X\boldsymbol{\mu}, \frac{b}{a}(I + XVX^T)\right). \end{aligned} \tag{18}$$

Of course, the computation of $p(\mathbf{y})$ could also be carried out in terms of the NIG distribution parameters more directly as

$$\begin{aligned} p(\mathbf{y}) &= \int p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2 = \int N(X\boldsymbol{\beta}, \sigma^2 I_n) \times NIG(\boldsymbol{\mu}_\beta, V_\beta, a, b) d\boldsymbol{\beta} d\sigma^2 \\ &= \frac{b^a}{(2\pi)^{p/2}|V_\beta|^{1/2}\Gamma(a)} \int \left(\frac{1}{\sigma^2}\right)^{a^*+p/2+1} \times \exp\left\{-\frac{1}{\sigma^2}\left[b^* + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}^*)^T V^{*-1}(\boldsymbol{\beta} - \boldsymbol{\mu}^*)\right]\right\} \\ &= \frac{b^a}{\Gamma(a)(2\pi)^{(n+p)/2}\sqrt{|V_\beta|}} \times \frac{\Gamma(a^*)(2\pi)^{p/2}\sqrt{|V^*|}}{(b^*)^{a^*}} \\ &= \frac{b^a \Gamma\left(a + \frac{n}{2}\right)\sqrt{|V^*|}}{(2\pi)^{n/2}\Gamma(a)\sqrt{|V_\beta|}} \times \left[b + \frac{1}{2}\left\{\boldsymbol{\mu}_\beta^T V_\beta^{-1}\boldsymbol{\mu}_\beta + \mathbf{y}^T\mathbf{y} - \boldsymbol{\mu}^* V^{*-1}\boldsymbol{\mu}^*\right\}\right]^{-(a+n/2)}. \end{aligned} \tag{19}$$

6 Marginal distribution: the easy way

An alternative and much easier way to derive $p(\mathbf{y} \,|\, \sigma^2)$, avoiding any integration at all, is to note that we can write the above model as:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}_1, \text{ where } \boldsymbol{\epsilon}_1 \sim N(\mathbf{0}, \sigma^2 I);$$

$$\boldsymbol{\beta} = \boldsymbol{\mu}_\beta + \boldsymbol{\epsilon}_2, \text{ where } \boldsymbol{\epsilon}_2 \sim N(\mathbf{0}, \sigma^2 V_\beta),$$

where $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are independent of each other. It then follows that

$$\mathbf{y} = X\boldsymbol{\mu}_\beta + X\boldsymbol{\epsilon}_2 + \boldsymbol{\epsilon}_1 \sim N(X\boldsymbol{\mu}_\beta, \sigma^2(I + XV_\beta X^T)).$$

This gives $p(\mathbf{y} \mid \sigma^2)$. Next we integrate out σ^2 to obtain $p(\mathbf{y})$ as in the preceding section to obtain

In fact, the entire distribution theory for the Bayesian regression with NIG priors could proceed by completely avoiding any integration. To be precise, we obtain this marginal distribution first and derive the posterior distribution:

$$p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) = \frac{p(\boldsymbol{\beta}, \sigma^2) \times p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)}{p(\mathbf{y})} = \frac{NIG(\boldsymbol{\mu}_\beta, V_\beta, a, b) \times N(X\boldsymbol{\beta}, \sigma^2 I)}{MVSt_{2a}(X\boldsymbol{\mu}, \frac{b}{a}(I + XV_\beta X^T))},$$

which indeed reduces (after some algebraic manipulation) to the $NIG(\boldsymbol{\mu}^*, V^*, a^*, b^*)$ density.

7 Bayesian Predictions

Next consider Bayesian prediction in the context of the linear regression model. Suppose we now want to apply our regression analysis to a new set of data, where we have observed a new $m \times p$ matrix of regressors $\tilde{X}$, and we wish to predict the corresponding outcome $\tilde{\mathbf{y}}$. Observe that if $\boldsymbol{\beta}$ and σ^2 were known, then the probability law for the predicted outcomes would be described as $\tilde{\mathbf{y}} \sim N(\tilde{X}\boldsymbol{\beta}, \sigma^2 I_m)$ and would be independent of $\mathbf{y}$. However, these parameters are not known; instead they are summarized through their posterior samples. Therefore, all predictions for the data must follow from the *posterior predictive* distribution:

$$\begin{aligned} p(\tilde{\mathbf{y}} \mid \mathbf{y}) &= \int p(\tilde{\mathbf{y}} \mid \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta},\, \sigma^2 \mid \mathbf{y}) d\boldsymbol{\beta} d\sigma^2 \\ &= \int N(\tilde{X}\boldsymbol{\beta},\, \sigma^2 I_m) \times NIG(\boldsymbol{\mu}^*, V^*, a^*, b^*) d\boldsymbol{\beta} d\sigma^2 \\ &= MVSt_{2a^*}\left(\tilde{X}\boldsymbol{\mu}^*, \frac{b^*}{a^*}(I + \tilde{X}V^*\tilde{X}^T)\right), \end{aligned} \tag{20}$$

where the last step follows from (18). There are two sources of uncertainty in the posterior predictive distribution: (1) the fundamental source of variability in the model due to σ^2, unaccounted for by $\tilde{X}\beta$, and (2) the posterior uncertainty in $\boldsymbol{\beta}$ and σ^2 as a result of their estimation from a finite sample $\mathbf{y}$. As the sample size $n \to \infty$ the variance due to posterior uncertainty disappears, but the predictive uncertainty remains.

8 Posterior and posterior predictive sampling

Sampling from the NIG posterior distribution is straightforward: for each $l = 1, \ldots, L$, we sample $\sigma^{2(l)} \sim IG(a + n/2, b^*)$ and $\boldsymbol{\beta}^{(l)} \sim MVN(\boldsymbol{\mu}^*, \sigma^{2(l)} V^*)$. The resulting $\left\{\boldsymbol{\beta}^{(l)}, \sigma^{2(l)}\right\}_{l=1}^{L}$ provide samples from the joint distribution $p(\boldsymbol{\beta},\, \sigma^2 \mid \mathbf{y})$ while $\{\boldsymbol{\beta}^{(l)}\}_{l=1}^{L}$ and $\{\sigma^{2(l)}\}_{l=1}^{L}$ provide samples from the marginal posterior distributions $p(\boldsymbol{\beta} \mid \mathbf{y})$ and $p(\sigma^2 \mid \mathbf{y})$ respectively.

Predictions are carried out by sampling from the posterior predictive density (20). Sampling from this is easy – for each posterior sample $(\boldsymbol{\beta}^{(l)}, \sigma^{2(l)})$, we draw $\tilde{\mathbf{y}}^{(l)} \sim N(\tilde{X}\boldsymbol{\beta}^{(l)}, \sigma^{2(l)} I_m)$. The resulting $\{\tilde{\mathbf{y}}^{(l)}\}_{l=1}^{L}$ are samples from the desired posterior predictive distribution in (20); the mean and variance of this sample provide estimates of the predictive mean and variance respectively.

9 The posterior distribution from improper priors

Taking $V_\beta^{-1} \to 0$ (i.e. the null matrix) and $a \to -p/2$ and $b \to 0$ leads to the improper prior $p(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$. The posterior distribution is $NIG\,(\boldsymbol{\mu}^*, V^*, a^*, b^*)$ with

$$\boldsymbol{\mu}^* = \hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y},$$

$$V^* = (X^T X)^{-1},$$

$$a^* = \frac{n-p}{2},$$

$$b^* = \frac{(n-p)s^2}{2} \quad \text{where} \quad s^2 = \frac{1}{n-p}(\mathbf{y} - X\hat{\boldsymbol{\beta}})^T(\mathbf{y} - X\hat{\boldsymbol{\beta}}) = \frac{1}{n-p}\mathbf{y}^T(I - P_X)\mathbf{y}, \quad \text{where} \quad P_X = X(X^T X)^{-1}X^T.$$

Here $\hat{\boldsymbol{\beta}}$ is the classical least squares estimates (also the maximum likelihood estimate) of $\boldsymbol{\beta}$, s^2 is the classical unbiased estimate of σ^2 and P_X is the projection matrix onto the column space of X.

Plugging in the above values implied by the improper priors into the more general $NIG(\boldsymbol{\mu}^*, V^*, a^*, b^*)$ density, we find the marginal posterior distribution of σ^2 is an $IG\left(\frac{n-p}{2}, \frac{(n-p)s^2}{2}\right)$ (equivalently the posterior distribution of $(n-p)s^2/\sigma^2$ is a χ^2_{n-p} distribution) and the marginal posterior distribution of $\boldsymbol{\beta}$ is a $MVSt_{n-p}(\hat{\boldsymbol{\beta}}, s^2 X^T X)$ with density:

$$MVSt_{n-p}(\boldsymbol{\mu}^*,\, s^2 X^T X) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-p}{2}\right)\pi^{p/2}|(n-p)s^2(X^T X)|^{-1/2}}\left[1 + \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T X^T X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{(n-p)s^2}\right]^{-\frac{n}{2}}.$$

Predictions with non-informative priors again follow by sampling from the posterior predictive distribution as earlier, but some additional insight is gained by considering analytical expressions

for the expectation and variance of the posterior predictive distribution. Again, plugging in the parameter values implied by the improper priors into (20), we obtain the posterior predictive density as a $MVSt_{n-p}\left(\tilde{X}\hat{\boldsymbol{\beta}}, s^2(I + \tilde{X}(X^TX)^{-1}\tilde{X}^T)\right)$.

Note that

$$\begin{aligned} E(\tilde{\mathbf{y}}|\sigma^2, y) &= E[E(\tilde{\mathbf{y}} \mid \boldsymbol{\beta},\, \sigma^2,\, \mathbf{y}) \mid \sigma^2, \mathbf{y}] \\ &= E[\tilde{X}\boldsymbol{\beta} \mid \sigma^2,\, \mathbf{y}] \\ &= \tilde{X}\hat{\boldsymbol{\beta}} = \tilde{X}(X^TX)^{-1}X^T\mathbf{y}, \end{aligned}$$

where the inner expectation averages over $p(\tilde{\mathbf{y}} \mid \boldsymbol{\beta},\, \sigma^2)$ and the outer expectation averages with respect to $p(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y})$. Note that given σ^2, the future observations have a mean which does not depend on σ^2. In analogous fashion,

$$\begin{aligned} var(\tilde{\mathbf{y}} \mid \sigma^2,\, \mathbf{y}) &= E[var(\tilde{\mathbf{y}} \mid \beta,\, \sigma^2,\, \mathbf{y}) \mid \sigma^2, \mathbf{y}] + var[E(\tilde{\mathbf{y}}|\boldsymbol{\beta}, \sigma^2, \mathbf{y})|\sigma^2, \mathbf{y}] \\ &= E[\sigma^2 I_m] + var[\tilde{X}\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}] \\ &= (I_m + \tilde{X}(X^TX)^{-1}\tilde{X}^T)\sigma^2. \end{aligned}$$

Thus, conditional on σ^2, the posterior predictive variance has two components: $\sigma^2 I_m$, representing sampling variation, and $\tilde{X}(X^TX)^{-1}\tilde{X}^T\sigma^2$, due to uncertainty about β.