The Bayesian Linear Model: A more advanced look

By
Sudipto Banerjee

1 The Standard Linear Model

Here we will study the standard linear model in a Bayesian context. Let us suppose that $Y$ is a $N \times 1$ response vector, $X$ is a $N \times p$ vector of regressors, $\beta$ is the corresponding vector of slope parameters and $\epsilon$ is a vector of random errors following a multivariate normal distribution. In general, these errors may be correlated, defined by a covariance matrix $\Sigma_\epsilon$ so that:

$$Y = X\beta + \epsilon; \quad \epsilon \sim MVN(0, \Sigma_\epsilon). \quad (1)$$

1.1 Bayesian inference when $\Sigma_\epsilon$ is known

Let us first suppose that $\Sigma_\epsilon$ is a known $N \times N$ matrix, so that only $\beta$ is unknown. Further suppose that $\beta \sim MVN(\mu_\beta, \Sigma_\beta)$. Then, the posterior distribution of $\beta$ is given by:

$$P(\beta | Y) \propto P(\beta) \times P(Y | \beta, \Sigma_\epsilon)$$

$$\propto \exp \left( -\frac{1}{2} (\beta - \mu_\beta)^T \Sigma_\beta^{-1} (\beta - \mu_\beta) \right) \times \exp \left( -\frac{1}{2} (Y - X\beta)^T \Sigma_\epsilon^{-1} (Y - X\beta) \right)$$

$$\propto \exp \left( -\frac{1}{2} \left\{ (\beta - \mu_\beta)^T \Sigma_\beta^{-1} (\beta - \mu_\beta) + (Y - X\beta)^T \Sigma_\epsilon^{-1} (Y - X\beta) \right\} \right).$$

Even though the computation for the posterior distribution might appear monstrous from the above expression, a neat little observation greatly helps: we will only need to keep track of algebraic expressions involving $\beta$. Any terms not involving $\beta$ can be safely ignored as they are absorbed into the proportionality constant involved in $P(\beta | Y)$. Therefore, expanding the two quadratic forms in the exponents and
collecting the terms involving $\beta$ we have

\[
(\beta - \mu_\beta)^T \Sigma^{-1}_\beta (\beta - \mu_\beta) + (Y - X\beta)^T \Sigma^{-1}_\epsilon (Y - X\beta)
= \beta^T \left( \Sigma^{-1}_\beta + X^T \Sigma^{-1}_\epsilon X \right) \beta - 2\beta^T \left( \Sigma^{-1}_\epsilon \mu_\beta + X^T \Sigma^{-1}_\epsilon Y \right) + \text{terms not involving } \beta
\]

Now we can use a multivariate version of “completing the square” which is essentially the following easily verified result on ellipses:

\[
x^T Ax - 2x^T u = (x - A^{-1}u)^T A(x - A^{-1}u) - u^T A^{-1}u.
\]

So, applying this result to the quadratic form in the exponent of $P(\beta | Y)$ we obtain:

\[
\beta^T \Sigma^{-1}_\beta \beta - 2\beta^T \Sigma^{-1}_\beta \mu_\beta + \text{terms not involving } \beta
= \left( \beta - \mu_\beta \right)^T \Sigma^{-1}_\beta \left( \beta - \mu_\beta \right) + \text{terms not involving } \beta
\]

where $\Sigma_\beta = \left( \Sigma^{-1}_\beta + X^T \Sigma^{-1}_\epsilon X \right)^{-1}$ and $\mu_\beta = \Sigma_\beta \left( \Sigma^{-1}_\epsilon \mu_\beta + X^T \Sigma^{-1}_\epsilon Y \right)$. This identifies the posterior distribution $P(\beta | Y)$ as:

\[
P(\beta | Y) \propto \exp \left\{ -\frac{1}{2} \left( \beta - \mu_\beta \right)^T \Sigma^{-1}_\beta \left( \beta - \mu_\beta \right) \right\}
\]

which is $MVN(\mu_\beta, \Sigma_\beta)$. We summarize this result in “hierarchical language”:

\[
Y | X, \beta, \Sigma_\epsilon \sim MVN (X\beta, \Sigma_\epsilon)
\]

\[
\beta | \mu_\beta, \Sigma_\beta \sim MVN (\mu_\beta, \Sigma_\beta)
\]

\[
\Rightarrow \beta | Y, X, \Sigma_\epsilon, \mu_\beta, \Sigma_\beta \sim MVN \left( \mu_\beta, \Sigma_\beta \right)
\]

where $\Sigma_\beta = \left( \Sigma^{-1}_\beta + X^T \Sigma^{-1}_\epsilon X \right)^{-1}$ and $\mu_\beta = \Sigma_\beta \left( \Sigma^{-1}_\epsilon \mu_\beta + X^T \Sigma^{-1}_\epsilon Y \right)$.

### 1.2 Bayesian inference when $\Sigma_\epsilon$ is unknown

Let us now turn to the setting where $\Sigma_\epsilon$ is unknown. Suppose we use a completely non-informative prior (Jeffrey’s prior) on $\beta$ and $\Sigma_\epsilon$. Then, we have:

\[
P(\beta, \Sigma_\epsilon) \propto \frac{1}{|\Sigma_\epsilon|^{(n+1)/2}}
\]
Then, the posterior becomes:

\[ P(\beta, \Sigma \epsilon | Y) \propto \frac{1}{|\Sigma \epsilon|^{n/2+1}} \exp \left\{ -\frac{1}{2} (Y - X\beta)^T \Sigma^{-1} \epsilon (Y - X\beta) \right\} . \]

In order to obtain the marginal posterior distribution \( P(\Sigma \epsilon | Y) \), we need to integrate out \( \beta \) from the above expression. To facilitate this calculation, we can again use a completing the squares argument to observe that:

\[ (Y - X\beta)^T \Sigma^{-1} (Y - X\beta) = (\hat{\beta} - \beta)^T X^T \Sigma^{-1} X (\hat{\beta} - \beta) + (Y - X\hat{\beta})^T \Sigma^{-1} (Y - X\hat{\beta}) , \]

where \( \hat{\beta} = (X^T X)^{-1} X^T Y \) is the least-squares estimate.

Assigning the non-informative Jeffrey’s prior \( P(\tau^2) \propto 1/\tau^2 \), we can perform posterior sampling from \( \theta = (\beta, \tau^2) \) without resorting to MCMC. To see how, observe that the conditional posterior distribution of \( P(\beta | \tau^2, Y) \) and the marginal distribution of \( P(\tau^2 | Y) \) are given as:

\[
[\beta | \tau^2, Y] \sim MVN(\mu_{\beta|}, \Sigma_{\beta|}), \\
\text{where } \Sigma_{\beta|} = \left( \Sigma^{-1} + \frac{1}{\tau^2} (X^T X)^{-1} \right)^{-1} \text{ and } \mu_{\beta|} = \Sigma_{\beta|} \left( \Sigma^{-1} \mu_{\beta} + \frac{1}{\tau^2} X^T Y \right) 
\]

2 The Bayesian Mixed Model and the Big-N problem

Let us consider the following setup:

\[ Y = X\beta + Zu + \epsilon, \epsilon \sim MVN(0, \Sigma \epsilon) \tag{2} \]

where \( Y \) is an \( N \times 1 \) vector of observations, \( X \) is a \( N \times p \) matrix of regressors, \( \beta \) is a \( p \times 1 \) vector of regression coefficients and \( Z \) is a \( N \times q \) matrix of coefficients/weights for the \( q \times 1 \) vector of “random effects” \( u \). Often the error is taken as i.i.d. so that \( \Sigma \epsilon = \tau^2 I \). In a Bayesian hierarchical setting, we will have priors in the coefficients such as:

\[
\beta \sim MVN(\mu_{\beta}, \Sigma_{\beta}); \\
u \sim MVN(\mu_{u}, \Sigma_{u}); \\
\tau^2 \sim P(\tau^2)
\]
In spatial contexts, the random effect dispersion matrix will be further parametrized as $\Sigma_u = \sigma^2 R(\phi)$, whence further priors must be imposed on $\sigma^2$ and $\phi$, say $P(\sigma^2)$ and $P(\phi)$ respectively.

The optimal strategy for estimating such models depends upon the sizes of the vectors and matrices above. Whenever possible it is better to reduce the dimension of the model (i.e. the number of “unknowns”). For instance, we can marginalize over the random effects $u$ by writing (2) as:

$$Y = X\beta + \epsilon^*; \quad \epsilon^* \sim \text{MVN}(Z\mu_u, Z\Sigma_u Z^T + \tau^2 I). \quad (1')$$

Several different options are worth pointing out. First consider the setting where $u = 0$ with probability 1. This amounts to $\mu_u = 0$ and $\Sigma_u^{-1} = 0$ (the null matrix) whence (2) and (1’) both coincide with the standard Bayesian linear model.