Ch 2 Conjugate Gaussian linear model

Monday, February 18, 2019 12:49 PM

Our model for the following problems is

$$(y \mid \beta, \sigma^2) \sim N(X\beta, (\omega \Lambda)^{-1})$$

where $\omega = \frac{1}{\sigma^2}$ is the precision. The priors are

$$(\beta \mid \omega) \sim N(m, (\omega K)^{-1})$$
$$\omega \sim \text{Gamma}\left(\frac{d}{2}, \frac{\eta}{2}\right)$$

where K is a $p \times p$ precision matrix in the MVN prior for β which we assume to be known.

A) Derive the conditional posterior $p(\beta \mid y, \omega)$.

The joint posterior $p(\beta, \omega \mid y) \propto p(y \mid \beta, \omega)\, p(\beta \mid \omega)\, p(\omega)$

For a linear model, the likelihood is

$$p(y \mid \beta, \omega) = \frac{1}{|2\pi (\omega\Lambda)^{-1}|^{1/2}} \exp\left(-\frac{1}{2}(y - X\beta)^T \omega \Lambda (y - X\beta)\right)$$

The priors $p(\beta \mid \omega)$ and $p(\omega)$

$$p(\beta \mid \omega) = \frac{1}{|2\pi (\omega K)^{-1}|^{1/2}} \exp\left(-\frac{1}{2}(\beta - m)^T \omega K (\beta - m)\right)$$

$$p(\omega) = \frac{(\eta/2)^{d/2}}{\Gamma(\frac{d}{2})\Gamma(\frac{2}{2})} \omega^{\frac{d}{2}-1} \exp\left(-\frac{\eta \omega}{2}\right)$$

so
$$p(\beta, \omega \mid y) \propto \frac{1}{|(\omega\Lambda)^{-1}|^{1/2} |(\omega K)^{-1}|^{1/2}} \exp\Big(\underbrace{-\frac{1}{2}(y - X\beta)^T \omega \Lambda (y - X\beta) - \frac{1}{2}(\beta - m)^T \omega K (\beta - m)}_{\text{complete the square in this term}}\Big) \quad (1)$$
$$\cdot\, \omega^{\frac{d}{2}-1} \exp\left(-\frac{\eta\omega}{2}\right)$$

Completing the square:

$$-\frac{1}{2}(y - X\beta)^T \omega \Lambda (y - X\beta) - \frac{1}{2}(\beta - m)^T \omega K (\beta - m)$$

$$= -\frac{1}{2}\omega\left[(y - X\beta)^T \Lambda (y - X\beta) + (\beta - m)^T K (\beta - m)\right]$$

$$= -\frac{1}{2}\omega\left[(y^T\Lambda - \beta^T X^T \Lambda)(y - X\beta) + (\beta^T K - m^T K)(\beta - m)\right]$$

$$= -\frac{1}{2}\omega\left[y^T \Lambda y - 2y^T \Lambda X\beta + \beta^T X^T \Lambda X \beta + (\beta^T K \beta - 2m^T K\beta + m^T K m)\right]$$

$$= -\frac{\omega}{2}\Big[\beta^T \underbrace{(X^T\Lambda X + K)}_{\text{call this } A}\beta - 2\underbrace{(y^T \Lambda X + m^T K)}_{\text{call this } b}\beta + y^T\Lambda y + m^T K m\Big]$$

$$= -\frac{\omega}{2}\left[\beta^T A \beta - 2b\beta + b^T A^{-1} b - b^T A^{-1} b + y^T \Lambda y + m^T K m\right]$$

$$= -\frac{\omega}{2}\left[(\beta - A^{-1}b)^T A (\beta - A^{-1} b) - b^T A^{-1} b + y^T \Lambda y + m^T K m\right]$$

The parameters of interest are β and ω. Therefore, we can rearrange terms in the exponentials that do not contain β. Hence, we can plug our completed square back into (1), and reorganize to get

$$p(\beta, \omega \mid y) \propto \frac{\omega^{\frac{d}{2}-1}}{|(\omega\Lambda)^{-1}|^{1/2} |(\omega K)^{-1}|^{1/2}} \exp\left(-\frac{\omega}{2}(\beta - A^{-1}b)^T A (\beta - A^{-1}b)\right) \exp\left(-\frac{\omega}{2}(\eta - b^T A^{-1} b + y^T \Lambda y + m^T K m)\right)$$

To find the marginal posterior of β given ω, we can simply ignore all the parts of the joint posterior that do not include β.

$$p(\beta \mid \omega, y) \propto \exp\left(-\frac{\omega}{2}(\beta - A^{-1}b)^T A (\beta - A^{-1}b)\right)$$

$$\sim MVN\left((X^T X + K)^{-1}(y \Lambda X + m^T K),\ (\omega (X^T \Lambda X + K))^{-1}\right)$$ by plugging in the values for A and b.

Like in the univariate case, we can see that the precision adds.

B) Derive the marginal posterior $p(\omega \mid y)$.

To find the marginal posterior of ω, we need to integrate β out of the joint posterior:

$$p(\beta, \omega \mid y) \propto \frac{\omega^{\frac{d}{2}-1}}{|(\omega\Lambda)^{-1}|^{1/2}\,|(\omega K)^{-1}|^{1/2}} \exp\left(-\frac{\omega}{2}(\beta - A^{-1}b)^T A(\beta - A^{-1}b)\right) \exp\left(-\frac{\omega}{2}(\eta - b^T A^{-1} b + y^T \Lambda y + m^T K m)\right)$$

$$p(\omega \mid y) \propto \frac{\omega^{\frac{d}{2}-1}}{|(\omega\Lambda)^{-1}|^{1/2}\,|(\omega K)^{-1}|^{1/2}} \exp\left(-\frac{\omega}{2}(\eta - b^T A^{-1} b + y^T \Lambda y + m^T K)\right) \int_{-\infty}^{\infty} \underbrace{\exp\left(-\frac{\omega}{2}(\beta - A^{-1}b)^T A(\beta - A^{-1}b)\right) d\beta}_{\text{This is the kernel of a MVN}(A^{-1}b, (\omega A)^{-1})}$$

$$= \frac{\omega^{\frac{d}{2}-1}}{|(\omega\Lambda)^{-1}|^{1/2}\,|(\omega K)^{-1}|^{1/2}} \exp\left(-\frac{\omega}{2}(\eta - b^T A^{-1} b + y^T \Lambda y + m^T K)\right) \left|2\pi\,(\omega(X^T\Lambda X + K))^{-1}\right|^{1/2}$$

This is some sort of gamma distribution. To find the α parameter, we need to use the determinant property $\det(cA) = c^n A$ where n is the dimension of A. We have

$$|(\omega\Lambda)^{-1}|^{1/2} = |\omega^{-1}\Lambda^{-1}|^{1/2} \propto \omega^{-\frac{n}{2}}$$

$$|(\omega K)^{-1}|^{1/2} = |\omega^{-1} K^{-1}|^{1/2} \propto \omega^{-\frac{p}{2}}$$

$$|(\omega(X^T\Lambda X + K))^{-1}|^{1/2} = |\omega^{-1}(X^T \Lambda X + K)^{-1}|^{1/2} \propto \omega^{-p/2}$$

Combining these with the $\omega^{\frac{d}{2}-1}$ from the prior gives us

$$\frac{\omega^{\frac{d}{2}-1}}{|(\omega\Lambda)^{-1}|^{1/2}\,|(\omega K)^{-1}|^{1/2}} \left|(\omega(X^T\Lambda X + K))^{-1}\right|^{1/2}$$

$$\propto \omega^{\frac{d}{2}-1} (\omega^{\frac{n}{2}})(\omega^{\frac{p}{2}})(\omega^{-p/2})$$

$$= \omega^{\frac{d+n}{2}-1}$$

Thus, $p(\omega \mid y) \sim \text{Gamma}\left(\frac{d+n}{2}, \frac{1}{2}(\eta - b^T A^{-1} b + y^T \Lambda y + m^T K)\right)$

where $A = (X^T X + K)$

$b = (y \Lambda X + m^T K)$

C) Putting these together, what is the marginal posterior of $p(\beta \mid y)$?

We can get $p(\beta \mid y) = \int p(\beta, \omega \mid y)\, d\omega$

$$= \int_0^{\infty} \frac{\omega^{\frac{d}{2}-1}}{|(\omega\Lambda)^{-1}|^{1/2}\,|(\omega K)^{-1}|^{1/2}} \exp\Big(-\frac{\omega}{2}\underbrace{(\beta - A^{-1}b)^T A(\beta - A^{-1}b)}_{\text{call this } B}\Big) \exp\Big(-\frac{\omega}{2}\underbrace{(\eta - b^T A^{-1} b + y^T \Lambda y + m^T K m)}_{\text{call this } \eta^*}\Big) d\omega$$

$$= \int_0^{\infty} \underbrace{\omega^{\left(\frac{d+n+p}{2}\right)-1} \exp\left(-\frac{\omega}{2}(B + \eta^*)\right)}_{\text{this is the Kernel of the gamma}\left(\frac{d+n+p}{2}, \frac{B+\eta^*}{2}\right)} d\omega$$

$$= \frac{\Gamma\left(\frac{d+n+p}{2}\right)}{\left(\frac{B+\eta^*}{2}\right)^{\frac{d+n+p}{2}}}$$

$$\propto \left(\frac{\eta^*}{2} + \frac{1}{2}(\beta - A^{-1}b)^T A(\beta - A^{-1}b)\right)^{-\frac{(d+n+p)}{2}}$$

$$\propto \left(\eta^* + (\beta - A^{-1}b)^T A(\beta - A^{-1}b)\right)^{-\frac{(d+n+p)}{2}}$$

$$= \left(1 + (\beta - A^{-1}b)^T \frac{A}{\eta^*}(\beta - A^{-1}b)\right)^{-\frac{(d+n+p)}{2}}$$

$$= \left(1 + \frac{1}{\quad} (\beta - A^{-1}b)^T \frac{A(d+n)}{\quad} (\beta - A^{-1}b)\right)^{-\frac{(d+n+p)}{2}}$$

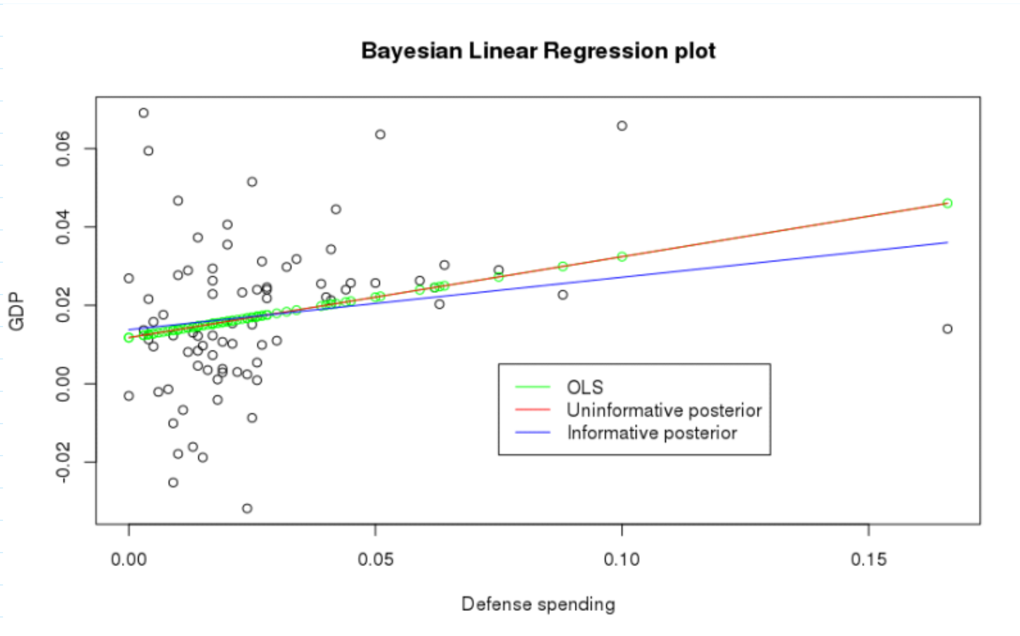
$$= \left(1 + \frac{1}{d+n}(\beta - A^{-1}b)^T \frac{A(d+n)}{\eta^*}(\beta - A^{-1}b)\right)^{-\frac{(d+n+p)}{2}}$$

$$\sim \text{multivariate } T\left(A^{-1}b, \frac{A(d+n)}{\eta^*}\right)$$

where

$$A = (X^TX + K)$$
$$b = (y/^T\Lambda X + m^TK)$$
$$\eta^* = \eta - b^TA^{-1}b + y/^T\Lambda y/ + m^TKm$$

D) Fit the Bayesian linear model of GR6096 vs DEF60 from the "gdpgrowth.csv" file. Use $\Lambda = I$ and something diagonal and vague for the prior precision matrix. Are you happy with the fit of the line?



The figure above features 3 fitted models: the ordinary least squares operator, the Bayesian Linear model with an "uninformative" prior precision matrix, and a Bayesian linear model with an "informative" prior precision matrix. In this case, the uninformative prior is

$$K = \text{diag}(0.00001, 0.00001).$$

The interpretation of this prior is that we have very little information to add to the data we have collected. In particular, we can see that the new mean of Beta can be written as

$$(K + X^TX)^{-1}(Km + X^Ty) \quad \text{when } \Lambda = I$$

$$= (K + X^TX)^{-1}Km + \underbrace{(K + X^TX)^{-1}X^Ty}_{\text{looks like the OLS estimator of } \hat{\beta}}$$

It is possible to see that if K is almost 0, the above expression boils down to the OLS estimate of $\hat{\beta}$. In general, the posterior mean of β is a weight linear combination of our prior beliefs about β and the OLS estimate of β, with the precision matrix K as a weight. The graph verifies this interpretation. The fitted line for the uninformative prior sits exactly on the OLS fitted line.

In fact, when I added in an informative prior,

$$K = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}, \quad m = (0.4, 0.4)$$

In fact, when I added in an informative prior,

$$K = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}, \quad m = (0.4, 0.4)$$

the fit of the model became slightly worse since we effectively communicated that the expected mean of β should be closer to zero. Thus, this fit is unsatisfactory since we do not know what the appropriate priors should be and we also cannot "beat" the OLS estimate unless we make unfounded guesses as to the prior mean of β.