Multiplying the likelihood w/ the prior:

$$P(\beta, \Sigma | y) \propto |\Sigma|^{-\underbrace{\frac{(\nu_0+N+2)}{2}}_{\nu_x = \nu_0 + 1}} \exp\left(-\frac{1}{2} tr\left[\left(\underbrace{S_0^{-1} + \frac{1}{2}(y - X\beta)(y - X\beta)^T}_{S_x}\right)\Sigma^{-1}\right]\right)$$

$$\times \sqrt{k_0} \, |(X^T \Sigma^{-1} X)^{-1}|^{-\frac{1}{2}} \times \exp\left(-\frac{1}{2}(2+k_0)^{2}\right)$$

This has some errors so I am rewriting this below

$$\exp\left(-\frac{1}{2}\left[\left(\beta - \left(\frac{\beta_0 + 2\hat{\beta}}{2+k_0}\right)\right)^T X^T \Sigma^{-1} X \left(\beta - \frac{\beta_0 k_0 \hat{\beta}}{2+k_0}\right)\right]\right)$$

Corrected Bayesian multivariate regression:

$$y_i = T_0 + P_0 E_i + P' E_i^2 + \varepsilon_i$$

$$= (1, \; E_i, \; E_i^2) \begin{pmatrix} T_0 \\ P_0 \\ P' \end{pmatrix} + \varepsilon_i$$

$$= x_i^T \beta + \varepsilon_i \qquad \text{where } \varepsilon_i \sim N(0, \sigma_i^2)$$

not necessarily i.i.d., the observations can have covariance

~~[illegible]~~

$$\Leftrightarrow \quad y = X\beta + \varepsilon \qquad \text{where } \varepsilon \sim N(0, \Sigma)$$

Likelihood:

$$P(y | X, \beta, \Sigma) = N(X\beta, \Sigma)$$
$$\propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y - X\beta)^T \Sigma^{-1} (y - X\beta)\right)$$

~~Note that there are two ways of decomposing this likelihood:~~

Note the decompositions:

① $|\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} tr\left((y - X\hat{\beta})^T \Sigma^{-1} (y - X\hat{\beta})\right)\right)$

$$= |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} tr\left((y - X\hat{\beta})(y - X\hat{\beta})^T \Sigma^{-1}\right)\right)$$

which looks like an Inverse Wishart distribution on Σ of the form:

$$\Sigma \sim IW(S_\varepsilon, V_\varepsilon)$$

where $S_\varepsilon = (y - X\hat{\beta})(y - X\hat{\beta})^T$

② $\cancel{P(y|X,\beta,\Sigma)} \propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} tr\left((y - X\beta)^T \Sigma^{-1} (y - X\beta)\right)\right)$

Noting the linear least squares solution:

$$\cancel{\hat{\beta} = (X^T X)^{-1} X^T y}$$
$$\cancel{X^T X \hat{\beta} = X^T y}$$
$$\cancel{X \hat{\beta} = y}$$

$\hat{\beta} = X^{-1} y$ provided X is invertible

$\Rightarrow y = X\hat{\beta}$

$$\Rightarrow \cancel{P(y|X,\beta,\Sigma)} \propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} tr\left((X(\hat{\beta}-\beta))^T \Sigma^{-1} X(\hat{\beta}-\beta)\right)\right)$$
$$\propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} tr\left((\hat{\beta}-\beta) X^T \Sigma^{-1} X(\hat{\beta}-\beta)\right)\right)$$
$$\propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} tr\left((\beta-\hat{\beta}) X^T \Sigma^{-1} X(\beta-\hat{\beta})\right)\right)$$

which looks like a Normal distribution on β of the form:

$$\beta \sim N(\hat{\beta}, (X^T \Sigma^{-1} X)^{-1})$$

Recalling these decompositions, we can factorize the likelihood into two terms: one term expressing the scatter of y about the least squares solution, and the other expressing the variance of β about the least squares solution, using the sum of squares trick:

$$P(y|X,\beta,\Sigma) \propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} tr\left((y - X\hat{\beta} + X\hat{\beta} - X\beta)(y - X\hat{\beta} + X\hat{\beta} - X\beta)^T \Sigma^{-1}\right)\right)$$

Substituting in $y = X\beta + \Sigma$ and $\hat{\beta} = X^{-1} y$ we see that the cross terms cancel, leading to the factorisation:

$$P(y|X,\beta,\Sigma) \propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} tr\left((y - X\hat{\beta})^T \Sigma^{-1} (y - X\hat{\beta})\right)\right)$$
$$\cdot \exp\left(-\frac{1}{2} tr\left((X\hat{\beta} - X\beta)^T \Sigma^{-1} (X\hat{\beta} - X\beta)\right)\right)$$
$$\propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} tr\left((y - X\hat{\beta})(y - X\hat{\beta})^T \Sigma^{-1}\right)\right)$$
$$\cdot \exp\left(-\frac{1}{2} (\beta - \hat{\beta})^T X^T \Sigma^{-1} X (\beta - \hat{\beta})\right)$$

Prior:

The natural conjugate prior for this likelihood ~~takes~~ the form:

controls our belief in β_0

$$P(\beta, \Sigma) = P(\Sigma)\, P(\beta \mid \Sigma)$$
$$= IW(S_0, v_0)\, N(\beta_0, \tfrac{1}{k_0}(X^T \Sigma^{-1} X)^{-1})$$

i.e. multiplying the likelihood by this prior will produce a posterior of the same distribution due to the properties of exponents and trace.

If the parameters were directly observed then $\frac{1}{k_0}(X^T\Sigma^{-1}X)^{-1}$ would simply be $\frac{1}{k_0}\Sigma^{-1}$. Instead this is communicating the possible variance in β that results from the variance of the observations.

$$P(\beta, \Sigma) = |\Sigma|^{-\left(\frac{v_0+N+1}{2}\right)} \exp(-\tfrac{1}{2} tr(S_0 \Sigma^{-1})) \cdot \sqrt{k_0}\, |(X^T\Sigma^{-1}X)^{-1}|^{\frac{1}{2}} \exp(-\tfrac{k_0}{2}(\beta-\beta_0)^T X^T \Sigma^{-1} X(\beta-\beta_0))$$

Posterior:

The posterior is also NIW:

$$P(\beta, \Sigma \mid y, X) = IW(S_y, v_y)\, N(\beta_y, \tfrac{1}{k_y}(X^T\Sigma^{-1}X)^{-1})$$

Multiplying the likelihood by the prior:

IW part: $IW(S_y, V_y) = IW(S_0, V_0)\, IW(S_\Sigma, V_\Sigma)$

$$= \exp(-\tfrac{1}{2} tr(S_0\Sigma^{-1}) \quad \exp(-\tfrac{1}{2} tr(S_\Sigma \Sigma^{-1}))$$
$$= \exp(-\tfrac{1}{2} tr((S_0 + S_\Sigma)\Sigma^{-1})$$

Normal part: $\propto N\Big(\beta_y, \underbrace{\frac{1}{k_y}\overbrace{(X^T\Sigma^{-1}X)^{-1}}^{\Sigma^*}}_{\Sigma_y}\Big) = N\big(\beta_0, \tfrac{1}{k_0}(X^T\Sigma^{-1}X)^{-1}\big)\, N\big(\hat{\beta}, (X^T\Sigma^{-1}X)^{-1}\big)$

Using the Matrix Cookbook 8.1.8, product of Gaussian densities:

- $$\beta_y = (k_0\Sigma^{*-1} + \Sigma^{*-1})^{-1}(k_0\Sigma^{*-1}\beta_0 + \Sigma^{*-1}\hat{\beta})$$
$$= (1+k_0)^{-1}\cancel{\Sigma^*\Sigma^{*-1}}(k_0\beta_0 + \hat{\beta})$$
$$= \frac{k_0\beta_0 + \hat{\beta}}{1+k_0}$$

This covariance matrix on our parameter Σ_y changes as Σ changes due to the IW portion of the joint distribution.

- $$\Sigma_y = (k_0\Sigma^{*-1} + \Sigma^{*-1})^{-1}$$
$$= [(k_0+1)\Sigma^{*-1}]^{-1}$$
$$= \frac{\Sigma^*}{k_0+1} = \frac{(X^T\Sigma^{-1}X)^{-1}}{k_0+1} \Rightarrow \bullet\ k_y = k_0 + 1$$

$$(c) = \frac{1}{\sqrt{\det(2\pi(\frac{1}{k_0}\Sigma^* + \Sigma^*))}} \exp\Big(-\tfrac{1}{2}(\hat{\beta}-\beta_0)\big(\tfrac{1}{k_0}\Sigma^* + \Sigma^*\big)^{-1}(\hat{\beta}-\beta_0)\Big)$$

$$\propto \sqrt{\frac{k_0}{1+k_0}}\,|\Sigma^*|^{-\frac{1}{2}} \exp\Big(-\tfrac{1}{2}(\hat{\beta}-\beta_0)^T\frac{k_0}{1+k_0}\Sigma^{*-1}(\hat{\beta}-\beta_0)\Big)$$

Combining (c) with the IW part gives:

$$IW(S_y, \nu_y) = |\Sigma|^{-\frac{\nu_0+N+1}{2}} \times \sqrt{\frac{k_0}{1+k_0}}\,|(X^T\Sigma^{-1}X)^{-1}|^{-\frac{1}{2}} \times \exp\Big(-\tfrac{1}{2}\mathrm{tr}\big(S_0\Sigma^{-1} + S_\varepsilon\Sigma^{-1}\big) + \frac{k_0}{1+k_0}(\hat{\beta}-\beta_0)^T X^T\Sigma^{-1}X(\hat{\beta}-\beta_0)\Big)\Big)$$

$$= \sqrt{\frac{k_0}{1+k_0}}\,|\Sigma|^{-\frac{\nu_0+N+1}{2}} \times |X^{-1}\Sigma(X^{-1})^T|^{-\frac{1}{2}} \times \exp\Big(-\tfrac{1}{2}\mathrm{tr}\Big(\big(S_0 + S_\varepsilon + \frac{k_0}{1+k_0}X(\hat{\beta}-\beta_0)(\hat{\beta}-\beta_0)^T X^T\big)\Sigma^{-1}\Big)\Big)$$

$$\propto |\Sigma|^{-\frac{\nu_0+N+1+1}{2}} \exp\Big(-\tfrac{1}{2}\mathrm{tr}\Big(\big(S_0 + (y-X\hat{\beta})(y-X\hat{\beta})^T + \frac{k_0}{1+k_0}X(\hat{\beta}-\beta_0)(\hat{\beta}-\beta_0)^T X^T\big)\Sigma^{-1}\Big)\Big)$$

$$\Rightarrow \bullet\ S_y = S_0 + \underbrace{(y-X\hat{\beta})(y-X\hat{\beta})^T}_{\text{empirical scatter matrix}} + \underbrace{\frac{k_0}{1+k_0}X(\hat{\beta}-\beta_0)(\hat{\beta}-\beta_0)^T X^T}$$

initial scatter matrix due to ~~our belief in~~ our belief in the prior β_0 i.e. if we strongly believe this prior then we should lean towards assuming the observations are wrong.

- $$\nu_y = \nu_0 + 1$$

Posterior marginals:

$$p(\Sigma \mid y, X) = \int p(\beta, \Sigma \mid y, X)\, d\beta$$
$$= IW(S_y, \nu_y)$$

$$\hat{\Sigma}_{map} = \frac{S_y}{\nu_y + N + 1} \qquad E[\Sigma] = \frac{S_y}{\nu_y - N - 1}$$

$$p(\beta \mid y, X) = \int p(\beta, \Sigma \mid y; X)\, d\Sigma$$
$$= T\left(\beta_y, \frac{S_y}{k_y(\nu_y - N + 1)}, \nu_y - N + 1\right)$$

This S_y causes our variance to change depending on the scatter of the deviation and any outliers.