Missing data lecture 5: Bayes

Bayes recap

MLE invariance

If $\widehat{\theta}$ is the MLE then the MLE for a function of θ, say $g(\theta)$, is just $g(\widehat{\theta})$.

Bayesian (in)variance:

Let $\eta = g(\theta)$, and assume for simplicity's sake that g is one-to-one. Then $\theta = g^{-1}(\eta)$. If θ has posterior $p(\theta \mid y)$, the posterior for $g(\theta)$ is:

$$p(g^{-1}(\eta) \mid y) \det \nabla_{\eta} g^{-1}(\eta)$$

This can lead to contradictions under "ignorance".

This presentation follows Gelman et al. (2013) somewhat.

There are priors called Jeffreys' priors (for Harold Jeffreys) that are invariant to reparameterizations. Remember that the Fisher information, or:

$$\mathcal{I}(\theta) = \mathbb{E}\left[\nabla_{\theta} \ell_Y(\theta \mid y) \nabla_{\theta} \ell_Y(\theta \mid y)^T\right]$$

under a reparameterization $\eta = g(\theta)$ with Jacobian $(J_{\eta,\theta})_{ij} = \frac{\partial \eta_i}{\partial \theta_j}$ is:

$$\mathcal{I}(\theta(\eta)) = J_{\eta,\theta}^T \mathbb{E}\left[\left(\nabla_{\theta} \ell_Y(\theta \mid y)\right) |_{\theta = g^{-1}(\eta)} \nabla_{\theta}(\ell_Y(\theta \mid y) \, |_{\theta = g^{-1}(\eta))}^T\right] J_{\eta,\theta}$$

For $\eta = g(\theta)$, assume for simplicity that g is one-to-one, then a prior for θ that is proportional to the square root of the determinant of the Fisher information will be invariant to reparameterization:

$$p(\theta) \propto \det(\mathcal{I}(\theta))^{1/2}$$

Why is this the case? Because under the change of measure formula above the prior for η is:

$$p(\eta) \propto p(g^{-1}(\eta)) \det J_{\eta,\theta}$$

which is

$$\begin{aligned} p(\eta) &\propto \det \mathcal{I}(g^{-1}(\eta))^{1/2} \det J_{\eta,\theta} \\ &\propto \det(J_{\eta,\theta})^{1/2} \det \mathcal{I}(g^{-1}(\eta))^{1/2} \det(J_{\eta,\theta})^{1/2} \\ &\propto \det(J_{\eta,\theta}^T)^{1/2} \det \mathcal{I}(g^{-1}(\eta))^{1/2} \det(J_{\eta,\theta})^{1/2} \\ &\propto \det(J_{\eta,\theta}^T \mathcal{I}(g^{-1}(\eta))^{1/2} J_{\eta,\theta})^{1/2} \\ &\propto \det(\mathcal{I}(\eta))^{1/2} \end{aligned}$$

Thus giving some sense of invariance under a coordinate change. As stated in Gelman et al. (2013), more or less:

> Any rule for determining the prior density $p(\theta)$ should yield an equivalent result if aplied to the transformed parameter; that is, $p(\eta)$ generated using $p(\theta)$ using the change of measure formula should yield the same prior as would have been obtained directly from the model $p(\eta)p(y \mid \eta)$

One issue with Jeffreys' prior is that it is dependent on a likelihood, which can be controversial.

For the Bernoulli trial example from last class, the Jeffreys prior is $\text{Beta}(1/2, 1/2)$.

Frequentist coverage?

Posterior probabilities are strictly "right" under our prior assumption because of the math of Bayes' theorem. However, if we take a Frequentist view of probability, namely that probabilities are defined as limiting proportions of events, we'll need to think about alternative draws of our prior and of our data.

The coverage of our posterior credible intervals will only match the nominal probabilities if the prior we use for our analysis matches that which generated the data. We can show this as computing the marginal posterior $p(\theta \mid y)$ under repeated draws from the prior and data distribution $p(y \mid \theta)$, which is the distribution associated with the density $f_Y(y \mid \theta)$ we'll use in our posterior:

$$\begin{aligned} \theta' &\sim p(\theta) \\ y &\sim p(y \mid \theta') \\ \theta &\sim p(\theta \mid y) \end{aligned}$$

Another way to represent this sampling diagram is through integrals:

$$\begin{aligned}
\int_{\Omega_\theta}\int_{\mathcal{Y}} \frac{p(\theta)f_Y(y \mid \theta)}{\int_{\Omega_\theta} p(\theta)f_Y(y \mid \theta)d\theta} f_Y(y \mid \theta')p(\theta')dy\, d\theta' &= \int_{\mathcal{Y}}\int_{\Omega_\theta} \frac{p(\theta)f_Y(y \mid \theta)}{\int_{\Omega_\theta} p(\theta)f_Y(y \mid \theta)d\theta} f_Y(y \mid \theta')p(\theta')d\theta'\, dy \\
&= \int_{\mathcal{Y}} \frac{p(\theta)f_Y(y \mid \theta)}{\int_{\Omega_\theta} p(\theta)f_Y(y \mid \theta)d\theta} \int_{\Omega_\theta} f_Y(y \mid \theta')p(\theta')d\theta'\, dy \\
&= \int_{\mathcal{Y}} p(\theta)f_Y(y \mid \theta) \\
&= p(\theta)
\end{aligned}$$

See Talts et al. (2018) for more info about how we can use this identity to test whether our algorithms are working correctly.

Bayes computation

Like Frequentist confidence intervals, we can only compute $p(\theta \mid y)$ exactly under special circumstances, like conjugate priors. The reason for this is that the integral in the denominator is usually intractable.

We will usually have to do approximate inference on Bayesian models by using Markov Chain Monte Carlo samplers, which iteratively generate samples that converge in distribution to the true posterior distribution. Bayesian approximate methods instead operate on an expression that is proportional to the posterior:

$$p(\theta \mid y) \propto f_Y(y \mid \theta)p(\theta)$$

One way to think about the MLE is that it is the posterior mode under a prior of $p(\theta) \propto 1$:

$$p(\theta \mid y) \propto f_Y(y \mid \theta)$$

The difference between the likelihood $L_Y(\theta \mid y)$ and the posterior $p(\theta \mid y)$ lies in how we treat the expression. In MLE we're going to maximize the likelihood. In Bayesian inference we care about the full distribution of θ.

This gives some intuition about Bayesian inference. We can think of doing MLE and penalizing certain values of θ:

$$\ell_Y(\theta \mid y) + \text{penalty}(\theta)$$

that will allow the maximizer to favor certain values of θ over others.

If we look at the implied log-posterior ignoring the constant that doesn't depend on θ:

$$\log p(\theta \mid y) = \log f_Y(y \mid \theta) + \log p(\theta)$$

If we maximize this expression we can rewrite this as

$$\log p(\theta \mid y) = \ell_Y(\theta \mid y) + \log p(\theta)$$

and we get the penalized likelihood expression where the penalty is a probability density.

One question might be: ok, we have a full distribution for θ. What do we do with it? While the MLE is a single choice, we now have myriad choices for point estimates derived from Bayesian models. We could use the posterior mean:

$$\mathbb{E}\left[\theta \mid y\right]$$

We could use the posterior median, θ_m:

$$P(\theta > \theta_m \mid y) = P(\theta \leq \theta_m \mid y) = 1/2.$$

We could use another posterior quantile. We could use the mode of the posterior as well.

Asymptotically, one might hope that the Bayesian estimates converge to the Frequentist estimates, and this is true, though one needs to be careful in scenarios where the dimensionality of the parameter space increases with sample size and about how one uses priors.

In Frequentist inference, the only limits on the parameter space come from the likelihood; the normal density requires that $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$. In Bayesian inference, the prior can also restrict the parameter space. For example, in the normal example, one could use a prior for μ that enforced $\mu > 0$. The posterior would then only be able to represent $\mu > 0$. If the true μ were negative, a Bayesian point-estimator wouldn't converge to the true μ.

While the prior adds an extra degree of freedom which seems dangerous, it can yield better estimates when there are small datasets, because there isn't as much information in the data. An example of this would be a simple regression model:

$$y_i \sim \text{Normal}(X_i^T \beta, \sigma^2)$$

We might have some good information that we don't expect β to be nearly infinite, and in fact we expect it to be pretty well concentrated to $[-10, 10]$. Then we could use independent $\text{Normal}(0, 5^2)$ priors for the regression coefficients.

Linear regression with conjugate priors

This and the following section follow Chapter 2 in Rossi, Allenby, and Misra (2024) quite closely.

Let's look at the linear regression model with conjugate priors.

$$y_i = x_i^T \beta + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2)$$

where $x_i \in \mathbb{R}^p$. A full model would imply a model for x_i as well:

$$f_{X,Y}((x_1, y_1), \ldots, (x_n, y_n) \mid \beta, \psi) = \prod_i f_X(x_i \mid \psi) f_Y(y_i \mid x_i, \beta, \sigma^2)$$

If we have a prior for ψ, β, σ^2 that is independent, $p(\psi, \beta, \sigma^2) = p(\psi)p(\beta, \sigma^2)$, then the posterior will factorize into independent distributions as well:

$$\begin{aligned} p_(\beta, \psi, \sigma^2 \mid (x_1, y_1), \ldots, (x_n, y_n)) &\propto \prod_i f_X(x_i \mid \psi) p(\psi) f_Y(y_i \mid x_i, \beta, \sigma^2) p(\beta, \sigma^2) \\ &\propto (\prod_i f_X(x_i \mid \psi) p(\psi)) \prod_i f_Y(y_i \mid x_i, \beta, \sigma^2) p(\beta) \\ &\propto p(\psi \mid x_1, \ldots, x_n) p(\beta, \sigma^2 \mid (x_1, y_1), \ldots, (x_n, y_n)) \end{aligned}$$

Remember from last class how we could intuit the form of the joint prior if we examined the likelihood for θ and chose a prior with the same functional form as that of the likelihood.

In the Bernoulli example, we had a likelihood of the form: $L_Y(\theta \mid y) = \theta^k (1-\theta)^{n-k}$, where $k = \sum_i = y$, which suggested a prior of the form $\theta^a (1-\theta)^b$, which we could regonize as a Beta distribution.

We'll do the same for the regression example. The likelihood for the linear model is:

$$(2\pi\sigma^2)^{-n/2} \exp\left(\frac{1}{2\sigma^2} \sum_i (y_i - x_i^T \beta)^2 \right)$$

which can be simplified somewhat by writing the sum as a dot product between the vector of errors, $e = y - X\beta$ where $y = (y_1, \ldots, y_n)$ and $X^T = (x_1, \ldots, x_n)$.

$$(2\pi\sigma^2)^{-n/2} \exp\left(\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) \right)$$

We can rewrite the term $(y - X\beta)^T (y - X\beta)$ in terms of the least-squares estimator for β, $\hat{\beta} = (X^T X)^{-1} X^T y$ by decomposing y as $y = X\hat{\beta} + y - X\hat{\beta}$:

$$\begin{aligned} (y - X\beta)^T (y - X\beta) &= (X\hat{\beta} + y - X\hat{\beta} - X\beta)^T (X\hat{\beta} + y - X\hat{\beta} - X\beta) \\ &= (y - X\hat{\beta})^T (y - X\hat{\beta}) + (X\beta - X\hat{\beta})^T (X\beta - X\hat{\beta}) - 2(X\beta - X\hat{\beta})^T (y - X\hat{\beta}) \\ &= (y - X\hat{\beta})^T (y - X\hat{\beta}) + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \end{aligned}$$

Let $s^2 = \frac{1}{n-p}(y - X\hat{\beta})^T (y - X\hat{\beta})$, and $\nu = n - p$, so we can rewrite the sum more compactly as:

$$(y - X\beta)^T (y - X\beta) = \nu s^2 + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta})$$

This leads to a likelihood:

$$L_Y(\beta, \sigma^2 \mid y, X) \propto (\sigma^2)^{-\nu/2} \exp\left(\frac{\nu s^2}{2\sigma^2}\right) (\sigma^2)^{-(n-\nu)/2} \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta})\right)$$

Before we derive conjugate priors from this likelihood, we can see that the posterior under flat priors for β and a prior for σ^2, σ^{-2}, leads to a posterior:

$$p(\beta, \sigma^2 \mid y, X) \propto (\sigma^2)^{-(\nu/2+1)} \exp\left(\frac{\nu s^2}{2\sigma^2}\right) (\sigma^2)^{-(n-\nu)/2} \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta})\right)$$

which is a conditional normal posterior for β with a scaled inverse chi-squared posterior for σ^2.

This suggests a conjugate prior of the form $p(\beta, \sigma^2) = p(\sigma^2)p(\beta \mid \sigma^2)$:

$$p(\sigma^2) \propto (\sigma^2)^{-(\nu_0/2+1)} \exp\left(\frac{\nu_0 s_0}{2\sigma^2}\right)$$

and a conditional normal prior for β:

$$p(\beta \mid \sigma^2) \propto (\sigma^2)^{-p/2} \exp\left(-\frac{1}{2\sigma^2}(\beta - \mu_0)^T \Sigma_0^{-1} (\beta - \mu_0)\right)$$

This can be seen as the posterior from a regression run with a prior of $p(\sigma^2) \propto \sigma^{-2}$ and a flat prior on β.

Then the posterior for σ^2, β is simply the product of the priors and the likelihood, which we write as above:

$$p(\beta, \sigma^2 \mid (x_1, y_1), \ldots, (x_n, y_n)) \propto (\sigma^2)^{-(\nu_0/2+1)} \exp\left(\frac{\nu_0 s_0}{2\sigma^2}\right) (\sigma^2)^{-p/2} \exp\left(-\frac{1}{2\sigma^2}(\beta - \mu_0)^T \Sigma_0^{-1} (\beta - \mu_0)\right)$$
$$(2\pi\sigma^2)^{-n/2} \exp\left(\frac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta)\right)$$

This is definitley formidable, but we can simplify things a bit by collecting the terms with β:

$$(y - X\beta)^T (y - X\beta) + (\mu_0 - \beta)^T \Sigma_0^{-1} (\mu_0 - \beta)$$

and decomposing $\Sigma_0^{-1} = L^T L$, and noting that we can write the sum as the following inner product:

$$\begin{bmatrix} (y - X\beta)^T & (L\mu_0 - L\beta)^T \end{bmatrix} \begin{bmatrix} (y - X\beta) \\ L\mu_0 - L\beta) \end{bmatrix}$$

This can be further simplified by constructing a vector

$$u = \begin{bmatrix} y \\ L\mu_0 \end{bmatrix}$$

and a matrix W

$$W = \begin{bmatrix} X \\ L \end{bmatrix}$$

and writing the expresssion as $(u - W\beta)^T(u - W\beta)$. We can then use the same trick as above, by representing u as the projection into the column space of W and the residual:

$$(W\bar{\beta} + u - W\bar{\beta} - W\beta)^T(W\bar{\beta} + u - W\bar{\beta} - W\beta)$$

The expression for $\bar{\beta}$ is:

$$\begin{aligned} \bar{\beta} &= (X^TX + L^TL)^{-1}(X^Ty + L^TL\mu_0) \\ &= (X^TX + \Sigma_0^{-1})^{-1}(X^Ty + \Sigma_0^{-1}\mu_0) \end{aligned}$$

Which simplifies as

$$(u - W\bar{\beta})^T(u - W\bar{\beta}) + (\beta - \bar{\beta})^TW^TW(\beta - \bar{\beta})$$

and after some algebra comes to

$$(y - X\bar{\beta})^T(y - X\bar{\beta}) + (\mu_0 - \bar{\beta})^T\Sigma_0^{-1}(\mu_0 - \bar{\beta}) + (\beta - \bar{\beta})^T(X^TX + \Sigma_0^{-1})(\beta - \bar{\beta})$$

In the following, let $ns^2 = (y - X\bar{\beta})^T(y - X\bar{\beta}) + (\mu_0 - \bar{\beta})^T\Sigma_0^{-1}(\mu_0 - \bar{\beta})$. The posterior is:

$$\begin{aligned} p(\beta, \sigma^2 \mid y, X) \propto &(\sigma^2)^{-(n+\nu_0)/2+1} \exp\left(\frac{(n+\nu_0)(ns^2 + \nu_0 s_0^2)/(n+\nu_0)}{2\sigma^2}\right) \times (\sigma^2)^{-p/2} \\ &\exp\left(-\frac{1}{2\sigma^2}(\beta - \bar{\beta})^T(X^TX + \Sigma_{\bar{\beta}}^{-1})(\beta - \bar{\beta})\right) \\ &\bar{\beta} = (X^TX + \Sigma_{\bar{\beta}}^{-1})^{-1}(\Sigma_{\bar{\beta}}^{-1}\mu_{\beta} + X^TX\hat{\beta}) \end{aligned}$$

Like the Bernoulli problem, the posterior mean for β is a weighted average between the prior mean and the information from the likelihood, which in this case is the least-squared estimator for β. This is often a consequence of using conjugate priors, that the posterior is a compromise between the prior and the likelihood.

This implies the following distributions for σ^2 and $\beta \mid \sigma^2$:

$$\begin{aligned} \sigma^2 &\sim \text{Inv-}\chi^2\left(n + \nu_0, \frac{ns^2 + \nu_0 s_0^2}{n + \nu_0}\right) \\ \beta \mid \sigma^2 &\sim \text{Normal}(\bar{\beta}, \sigma^2\left(X^TX + \Sigma_0^{-1}\right)^{-1}) \end{aligned}$$

The posterior mean for σ^2 is:

$$\mathbb{E}\left[\sigma^2 \mid y, X\right] = \frac{n + \nu_0}{n + \nu_0 - 2}\frac{ns^2 + \nu_0 s_0^2}{n + \nu_0}$$

The expression for ns^2 is interesting because it involves the squared error of the posterior linear predictor for y:

$$\begin{aligned}(y - \mathbb{E}\left[X\beta \mid y, X\right])^T(y - \mathbb{E}\left[X\beta \mid y, X\right])^T &= (y - X\mathbb{E}\left[\beta \mid y, X\right])^T(y - X\mathbb{E}\left[\beta \mid y, X\right]) \\ &= (y - X\bar{\beta})^T(y - X\bar{\beta})\end{aligned}$$

but it also involves the error in the prior mean with respect to the prior covariance matrix:

$$(\mu_0 - \bar{\beta})^T\Sigma_0^{-1}(\mu_0 - \bar{\beta})$$

The effect of this term will decrease as the number of observations increases, but it elucidates how the posterior mean of the error variance is decomposed into several pieces depending on different aspects of the prior and the data.

Bayesian inference in repeated measure models

Two lectures ago we went through how to compute the MLE from this regression model:

$$\begin{aligned}y_i \mid X_i &= X_i\beta + \epsilon_i \\ \epsilon_i &\sim \text{Normal}(0, \Sigma) \\ \epsilon_i &\perp\!\!\!\perp \epsilon_j \forall i \neq j.\end{aligned}$$

This required sequentially computing the MLE for β given an estimate for Σ and computing $\hat{\Sigma}$ given the last estimate for $\hat{\beta}$.

Let's write down the likelihood for this model to see if we can come up with a conjugate prior for the problem.

$$L_Y(\beta, \Sigma \mid y, X) \propto \det(I_n \otimes \Sigma)^{-1/2} \exp\left(-\frac{1}{2}(y - X\beta)^T(I_n \otimes \Sigma)^{-1}(y - X\beta)\right)$$

If we start with the prior for $\beta \mid \Sigma$ we can ignore the determinant and focus on the term in the exponential:

$$-\frac{1}{2}(y - X\beta)^T(I_n \otimes \Sigma)^{-1}(y - X\beta)$$

Let's try a multivariate normal prior:

$$\beta \sim \text{Normal}(\mu_0, \Sigma_0)$$

so we can multiply the likelihood by the prior to get

$$-\frac{1}{2}\left((y - X\beta)^T(I_n \otimes \Sigma)^{-1}(y - X\beta) + (\beta - \mu_0)^T\Sigma_0^{-1}(\beta - \mu_0)\right)$$

which we'll rewrite for convenience as

$$-\frac{1}{2}\left((A(y-X\beta))^TA(y-X\beta)+(L(\beta-\mu_0))^TL(\beta-\mu_0)\right)$$

where $A^TA=(I_n\otimes\Sigma)^{-1}$ and $L^TL=\Sigma_0^{-1}$.

This looks familiar! We can use the same trick as we did above: create a new vector u and matrix W:

$$u=\begin{bmatrix}Ay\\L\mu_0\end{bmatrix},\quad W=\begin{bmatrix}AX\\L\end{bmatrix}$$

and can write

$$(u-W\beta)^T(u-W\beta)=\left((A(y-X\beta))^TA(y-X\beta)+(L(\beta-\mu_0))^TL(\beta-\mu_0)\right).$$

Furthermore, write $u=W\bar{\beta}+u-W\bar{\beta}$, where $\bar{\beta}$ is the least-squares coeffients of the regression of u on W:

$$\bar{\beta}=(W^TW+L^TL)^{-1}W^Tu=(X^T(I_n\otimes\Sigma)^{-1}X+\Sigma_0^{-1})^{-1}(X^T(I_n\otimes\Sigma)^{-1}y+\Sigma_0^{-1}\mu_0)$$

This leads to $(u-W\bar{\beta})^TW=0$, which allows us to cleanly partition $(u-W\beta)^T(u-W\beta)$ into two pieces: $u-W\bar{\beta}$ and $W\beta$:

$$\begin{aligned}(W\bar{\beta}+u-W\bar{\beta}-W\beta)^T&(W\bar{\beta}+u-W\bar{\beta}-W\beta)\\&=(u-W\bar{\beta}+W\bar{\beta}-W\beta)^T(u-W\bar{\beta}+W\bar{\beta}-W\beta)\\&=(u-W\bar{\beta})^T(u-W\bar{\beta})+(W\bar{\beta}-W\beta)^T(W\bar{\beta}-W\beta)+2(u-W\bar{\beta})^T(W\bar{\beta}-W\beta)\\&=(u-W\bar{\beta})^T(u-W\bar{\beta})+(W\bar{\beta}-W\beta)^T(W\bar{\beta}-W\beta)\end{aligned}$$

where the last line follows because $(u-W\bar{\beta})^TW=0$. Because we're focusing only on the posterior, which is a function of β and not data, we can ignore the $(u-W\bar{\beta})^T(u-W\bar{\beta})$ term because it does not involve β and involves only functions of X,y,A,L, which are fixed with respect to β.

We rewrite

$$(W\bar{\beta}-W\beta)^T(W\bar{\beta}-W\beta)$$

as

$$(\beta-\bar{\beta})^TW^TW(\beta-\bar{\beta})=(\beta-\bar{\beta})^T(X^T(I_n\otimes\Sigma)^{-1}X+\Sigma^{-1})(\beta-\bar{\beta})$$

This shows that $\beta\mid\Sigma$ is multivariate normal with:

$$\beta\sim\text{Normal}(\bar{\beta},(X^T(I_n\otimes\Sigma)^{-1}X+\Sigma^{-1})^{-1})$$

Now let's focus on the conditional distribution of $\Sigma\mid\beta$. We'll start with the likelihood written in simpler terms:

$$L_Y(\beta,\Sigma\mid y,X)\propto\det(\Sigma)^{-n/2}\exp\left(-\frac{1}{2}\textstyle\sum_i(y_i-X_i\beta)^T\Sigma^{-1}(y_i-X_i\beta)\right)$$

We can use the trace trick to rearrange things:

$$\begin{aligned}
L_Y(\beta, \Sigma \mid y, X) &\propto \det(\Sigma)^{-n/2} \exp\left(-\frac{1}{2}\textstyle\sum_i (y_i - X_i\beta)^T\Sigma^{-1}(y_i - X_i\beta)\right) \\
&\propto \det(\Sigma)^{-n/2} \exp\left(-\frac{1}{2}\textstyle\sum_i \operatorname{tr}((y_i - X_i\beta)^T\Sigma^{-1}(y_i - X_i\beta))\right) \\
&\propto \det(\Sigma)^{-n/2} \exp\left(-\frac{1}{2}\textstyle\sum_i \operatorname{tr}((y_i - X_i\beta)(y_i - X_i\beta)^T\Sigma^{-1})\right) \\
&\propto \det(\Sigma)^{-n/2} \exp\left(-\frac{1}{2}\operatorname{tr}((\textstyle\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T)\Sigma^{-1})\right)
\end{aligned}$$

This suggests that a conjugate prior for Σ has the form:

$$p(\Sigma) \propto \det(\Sigma)^{-a/2} \exp\left(-\frac{1}{2}\operatorname{tr}(V_0\Sigma^{-1})\right)$$

Fortunately, we're in luck! The Inverse Wishart distribution has the density:

$$p(\Sigma) \propto \det(\Sigma)^{-(\nu_0+p+1)/2} \exp\left(-\frac{1}{2}\operatorname{tr}(V_0\Sigma^{-1})\right)$$

Combining the likelihood with the prior we get something proportional to the conditional posterior for Σ:

$$p(\Sigma \mid y, X, \beta) \propto \det(\Sigma)^{-(n+\nu_0+p+1)/2} \exp\left(-\frac{1}{2}\operatorname{tr}((V_0 + \textstyle\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T)\Sigma^{-1})\right)$$

Putting this together we get the following two conditional posteriors:

$$\begin{aligned}
\beta \mid \Sigma, y, X &\sim \text{Normal}(\bar{\beta}, (X^T(I_n \otimes \Sigma)^{-1}X + \Sigma^{-1})^{-1}) \\
\Sigma \mid \beta, y, X &\sim \text{Inverse-Wishart}(n + \nu_0, V_0 + \textstyle\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T)
\end{aligned}$$

We can use the theory of integral operators to show that given intial conditions Σ^0 and β^0 the following algorithm for $t = 1, \ldots, S$:

$$\begin{aligned}
\beta^{t+1} \mid \Sigma^t, y, X &\sim \text{Normal}(\bar{\beta}, (X^T(I_n \otimes \Sigma^t)^{-1}X + (\Sigma^t)^{-1})^{-1}) \\
\Sigma^{t+1} \mid \beta^t, y, X &\sim \text{Inverse-Wishart}(n + \nu_0, V_0 + \textstyle\sum_i (y_i - X_i\beta^t)(y_i - X_i\beta^t)^T)
\end{aligned}$$

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis*. CRC Press.

Rossi, Peter, Greg Allenby, and Sanjog Misra. 2024. *Bayesian Statistics and Marketing*. John Wiley & Sons.

Talts, Sean, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. 2018. "Validating Bayesian Inference Algorithms with Simulation-Based Calibration." *arXiv Preprint arXiv:1804.06788*.