Wishart Priors

Patrick Breheny

March 28
When more than two coefficients vary, it becomes difficult to directly model each element of the correlation matrix.

For the sake of easily generalizing to larger number of coefficients, let’s rewrite model #3 from the previous lecture using matrix notation:

\[ Y_{ij} \sim N(x_{ij}^T \beta_j, \sigma_y^2) \]

\[ \beta_j \sim N(\mu, \Sigma) \]

The complication, of course, is that now we have to specify a prior for \( \Sigma \), a variance-covariance matrix.
Recall that the semi-conjugate prior for the variance of a univariate normal distribution could be expressed as a scaled $\chi^2$ distribution:

$$c\tau \sim \chi^2(\nu)$$

$$\sigma^2 = \tau^{-1}$$

The same approach can be extended to the multivariate normal case using a multivariate extension of the $\chi^2$ distribution known as the *Wishart distribution*.
The Wishart distribution

Suppose $\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$; the Wishart distribution with $n$ degrees of freedom is defined as the distribution of

$$\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T;$$

we will denote this $\mathbf{S} \sim \text{Wishart}(\Sigma, n)$.

Alternatively, one could parameterize the Wishart distribution in terms of the precision matrix, $\Omega = \Sigma^{-1}$; this is the parameterization used by BUGS and JAGS (note the distinction, though, because most other sources, including our textbook, calls this an “inverse Wishart” distribution).
Interpreting the Wishart

- The big advantage of the Wishart distribution is that it is guaranteed to produce positive definite draws, provided that \( n \geq p \); this is difficult to enforce otherwise.

- The fewer the degrees of freedom \( n \) in the distribution, the larger the variability; thus, \( n = p \) is the least informative choice possible.

- Note that the expected value of the Wishart distribution is \( n \Sigma \); this is helpful if providing an informative prior, where you can think of the prior as equivalent to seeing \( n \) observations, for which the observed variance-covariance matrix is \( n \Sigma \) (again, these would have to be converted to precision matrices in the BUGS/JAGS formulation).

- (See R code for some examples of drawing from the Wishart distribution)
Introduction

- When more than two coefficients vary, it becomes difficult to directly model each element of the correlation matrix.
- For the sake of easily generalizing to larger number of coefficients, let’s rewrite model #3 from the previous lecture using matrix notation:

  \[ Y_{ij} \sim N(x^T_{ij}\beta_j, \sigma^2_y) \]
  \[ \beta_j \sim N(\mu, \Sigma) \]

- The complication, of course, is that now we have to specify a prior for \( \Sigma \), a variance-covariance matrix.
This model is similar to Model #3 from the previous lecture, but is not identical – a Wishart prior is not the same as placing uniform priors on the elements of $\Sigma$ directly – however, for the most part the inferences we obtain are very similar.

The most noticeable difference is that the MCMC sampler runs quite a bit faster and mixes better – this are the usual advantages of semi-conjugacy.

However, another important difference concerns $\rho$, which has a posterior median of -0.1 and a 95% posterior interval of (-0.5, 0.3), which is quite a bit different than the result from the previous model.
The Wishart distribution has a single parameter that determines how informative/restrictive it is.

Often in modeling, one would rather have a prior that is, relatively speaking, more informative/restrictive with respect to the correlation structure than it is with respect to the variances — i.e., we would like to decompose the prior on $\Sigma$ into separate priors on (a) the diagonal elements and (b) the correlation structure.

An interesting approach for doing this is proposed by our authors, which they call a scaled Wishart or scaled inverse-Wishart.
The idea is as follows:

\[ Q \sim \text{Wishart}(I, n) \]
\[ \Sigma = \Xi Q \Xi, \]

where \( \Xi \) is a diagonal matrix with elements \( \{\xi_j\} \), which are typically given a disperse prior such as a uniform distribution over a wide range.

Strictly speaking, this model is not identifiable, in the sense that the parameters \( \{\xi_j\} \) and \( Q \) cannot be interpreted separately.
However, the model is still identifiable in terms of $\Sigma$, which is what we care about:

\[
\begin{align*}
\sigma_j &= \xi_j \sqrt{Q_{jj}} \\
\rho_{jk} &= \frac{Q_{jk}}{\sqrt{Q_{jj}Q_{kk}}} 
\end{align*}
\]
Results

- Again, for this data set, most of the inferences regarding \( \{\alpha_j\}, \{\beta_j\}, \) and the \( \gamma \) parameters are fairly robust to whether we directly specify the prior for all the elements of \( \Sigma \), use a Wishart prior, or a scaled Wishart prior.

- However, the posterior we obtain for \( \rho \), the correlation between \( \alpha \) and \( \beta \), is more similar to our original result using the scaled Wishart than the Wishart: median 0.2, 95% interval: (-0.5, 0.7).

- This is an important observation to be aware of as we move forward: the “least informative” Wishart prior is still fairly informative.