Bayesian Inference for the Hierarchical Normal Model

Reading

This lecture is based on Chapter 4 of Hoff.

Hierarchical Normal Model

Recall our data model:

$$y_{ij} \quad = \quad \mu_j + \varepsilon_{ij} = \mu_j + \varepsilon_{ij}$$

where $\mu_j = \mu + \alpha_j$, $\alpha_j \overset{iid}{\sim} N\left(0, \tau^2\right) \perp \varepsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right)$

In addition to the data model we specified previously, we will specify a prior distribution for (μ, τ^2, σ^2), denoted $p(\theta) = p(\mu, \tau^2, \sigma^2)$.

Bayesian Specification of the Model

We will start with a "default" prior specification given by

$$p(\mu, \tau^2, \sigma^2) = p(\mu|\tau^2)p(\tau^2)p(\sigma^2),$$

where

- $p(\mu|\tau^2)$ is $N\left(\mu_0, \frac{\tau^2}{m_0}\right)$
- $p(\tau^2)$ is inverse-gamma$\left(\frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right)$
- $p(\sigma^2)$ is inverse-gamma$\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$

With this default prior specification, we have nice interpretations of the prior parameters.

- For $p(\mu|\tau^2) = N\left(\mu_0, \frac{\tau^2}{m_0}\right)$, μ_0 is a prior guess at μ, and m_0 describes our certainty in this guess (you can think of it as a prior sample size)
- For $p(\tau^2) = IG\left(\frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right)$, τ_0^2 is a prior guess at the across-group variance τ^2, and η_0 describes our confidence in this guess
- For $p(\sigma^2) = IG\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$, σ_0^2 is a prior guess at the within-group variance σ^2, and ν_0 describes our certainty in this guess

Fully-Specified Model

We have now fully-specified our model with the following components.

1. Unknown parameters $(\mu, \tau^2, \sigma^2, \alpha_1, \cdots, \alpha_J)$
2. Prior distributions, specified in terms of prior guesses $(\mu_0, \tau_0^2, \sigma_0^2)$ and certainty/prior sample sizes (m_0, η_0, ν_0)
3. Don't forget the data y, x from our groups!

Posterior distributions

We can interrogate the posterior distribution of the parameters using Gibbs sampling, as the full conditional distributions have closed forms. We'll next examine the full conditional distributions of each of our parameters $(\mu_0, \tau_0^2, \sigma_0^2, \mu_1, \cdots, \mu_J)$ in turn.

$$
\begin{aligned}
p(\mu \mid \tau^2, \sigma^2, Y, \mu_1, \cdots, \mu_J) &\propto p(\mu, \tau^2, \sigma^2, Y, \mu_1, \cdots, \mu_J) \\
&\propto p(\mu \mid \tau^2) p(\tau^2) p(\sigma^2) \prod_{j=1}^{J} p(\mu_j \mid \mu, \tau^2) \prod_{i=1}^{n_j} p(y_{ij} \mid \mu_j, \sigma^2) \\
&\propto p(\mu \mid \tau^2) \prod_{j=1}^{J} p(\mu_j \mid \mu, \tau^2) \quad \text{(dropping stuff not involving } \mu) \\
&\propto \exp\left\{ -\frac{m_0(\mu - \mu_0)^2}{2\tau^2} \right\} \exp\left\{ -\frac{\sum(\mu - \mu_j)^2}{2\tau^2} \right\}
\end{aligned}
$$

You can show then that $(\mu \mid \tau^2, \mu_1, \cdots, \mu_J) = N(e, v)$ where $e = \frac{J}{m_0+J}\frac{1}{J}\sum \mu_j + \frac{m_0}{J+m_0}\mu_0$ and $v = \frac{\tau^2}{J+m_0}$, and here we can think of m_0 as the number of "prior observations" and μ_0 as the sample mean of those prior observations. In practice, people take μ_0 to be a reasonable prior guess, and if they're really unsure they take $m_0 = 1$, implying that this is equivalent roughly to the certainty you would have from only one sampled value from the population (so not very much information).

How do you show that?

It helps if you remember how to complete the square.

$$
\begin{aligned}
& \exp\left\{\frac{-m_0(\mu-\mu_0)^2}{2\tau^2} - \frac{\sum(\mu-\mu_j)^2}{2\tau^2}\right\} \\
= \quad & \exp\left\{\frac{-m_0\mu_0^2 + 2\mu m_0\mu_0 - m_0\mu_0^2 - J\mu^2 + 2\mu\sum\mu_j - \sum\mu_j^2}{2\tau^2}\right\} \\
\propto \quad & \exp\left\{\frac{-m_0\mu^2 + 2\mu m_0\mu_0 - J\mu^2 + 2\mu\sum\mu_j}{2\tau^2}\right\} \\
= \quad & \exp\left\{\frac{\mu^2(-m_0-J) + \mu(2m_0\mu_0 + 2\sum\mu_j)}{2\tau^2}\right\} \\
= \quad & \exp\left\{\frac{-m_0-J}{2\tau^2}\left[\mu^2 + \mu\left(\frac{2m_0\mu_0 + 2\sum\mu_j}{-m_0-J}\right)\right]\right\} \\
\propto \quad & \exp\left\{\left(\frac{-(m_0+J)}{2\tau^2}\right)\left[\mu - \frac{m_0\mu_0 + \sum\mu_j}{m_0+J}\right]^2\right\}
\end{aligned}
$$

Ahh, this is the kernel of a normal distribution for μ!

We recognize this conditional posterior distribution of $\{\mu \mid \tau^2, \mu_1, \cdots, \mu_J\}$ as the kernel of a normal distribution with mean $\frac{m_0\mu_0+\sum \mu_j}{\mu_0+J}$ and variance $\frac{\tau^2}{m_0+J}$.

To motivate interpretation of the prior parameters, express the mean $\frac{m_0\mu_0+\sum \mu_j}{\mu_0+J} = \frac{J}{m_0+J}\frac{1}{J}\sum \mu_j + \frac{m_0}{J+m_0}\mu_0$.

We can think of

- m_0 as the number of prior observations with sample mean μ_0
- the conditional mean as the average of all the observations (prior and current)
- the conditional variance as the population variance divided by the "total" sample size $J + m_0$

Using similar algebra, we can show the full conditional distribution of μ_j is given by

$$
\begin{aligned}
p(\mu_j \mid \mu, \tau^2, \sigma^2, Y, \mu_1, \cdots, \mu_J) &\propto p(\mu, \tau^2, \sigma^2, Y, \mu_1, \cdots, \mu_J) \\
&\propto p(\mu \mid \tau^2) p(\tau^2) p(\sigma^2) \prod_{j=1}^{J} p(\mu_j \mid \mu, \tau^2) \prod_{i=1}^{n_j} p(y_{ij} \mid \mu_j, \sigma^2) \\
&\propto p(\mu_j \mid \mu, \tau^2) \prod_{j=1}^{n_j} p(y_{ij} \mid \mu_j, \sigma^2) \quad \propto \quad \exp\left\{ -\frac{(\mu_j - \mu)^2}{2\tau^2} \right\} \exp
\end{aligned}
$$

For homework you will show this is a normal distribution with variance $v_j = \frac{1}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$ and mean $v_j \left[\frac{n_j}{\sigma^2} \overline{y}_j + \frac{1}{\tau^2} \mu \right]$.

The full conditional posterior distribution of μ_j has a nice form, and we can see that the conditional mean is a compromise between the sample mean (weighted by the data precision) and the overall mean (weighted by the prior across-group precision), while the conditional precision (inverse variance) can be thought of as the sum of data precision and prior across-group precision.

Inverse Gamma Distribution

The inverse gamma (IG) distribution is a popular choice for modeling variance components given its support on the positive real numbers. If the random variable $\frac{1}{X}$ has a gamma distribution, we say X has an inverse gamma distribution, given by

$$p(x \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}} \text{ for } x > 0.$$

This distribution has mean $\frac{\beta}{\alpha-1}$, mode $\frac{\beta}{\alpha+1}$, and variance $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$.

Using an $\text{IG}\left(\frac{m_0}{2}, \frac{m_0\tau_0^2}{2}\right)$ distribution for τ^2, we can now see that τ_0^2 is somewhere in the "center" of the distribution (between the mode $\frac{m_0\tau_0^2}{m_0+2}$ and the mean $\frac{m_0\tau_0^2}{m_0-2}$). As the "prior sample size" m_0 increases, the difference between these quantities goes to 0.

The full conditional posterior for τ^2 in our hierarchical normal model is given by

$$
\begin{aligned}
p(\tau^2 \quad | \quad & \mu, \sigma^2, Y, \mu_1, \cdots, \mu_J) \propto p(\mu, \tau^2, \sigma^2, Y, \mu_1, \cdots, \mu_J) \\
\propto \quad & p(\tau^2) p(\mu \mid \tau^2) \prod p(\mu_1, \cdots, \mu_J \mid \mu, \tau^2) \\
\propto \quad & \left(\tau^2\right)^{-\frac{m_0}{2}-1} e^{-\frac{m_0 \tau_0^2}{2\tau^2}} \left(\tau^2\right)^{-\frac{1}{2}} e^{-\frac{m_0(\mu-\mu_0)^2}{2\tau^2}} \prod_{j=1}^{J} \left(\tau^2\right)^{-\frac{1}{2}} e^{-\frac{(\mu_j-\mu)^2}{2\tau^2}} \\
\propto \quad & (\tau^2)^{-\frac{m_0+J+1}{2}-1} e^{-\frac{m_0\tau_0^2+m_0(\mu-\mu_0)^2+\sum(\mu_j-\mu)^2}{2\tau^2}}
\end{aligned}
$$

so that the full conditional posterior is also IG.

You can also show the full conditional posterior of σ^2 is $IG\left(\frac{\nu_0+\sum n_j}{2}, \frac{\nu_0\sigma_0^2+\sum\sum(y_{ij}-\mu_j)^2}{2}\right)$ under our prior specification of $\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$ (homework). Again, σ_0^2 and ν_0 can be interpreted as a prior guess at σ^2 associated with a given prior sample size (level of confidence).

Posterior Inference

We can use Gibbs sampling to make inference about the posterior distribution.

Biking Data

Here we fit a hierarchical model for the association between curb and passing distance in the bike data. NEED TO FIT THIS MODEL WITH THE DEFAULT PRIORS INSTEAD OF WITH THE STAN VS USED HERE.

```
## Loading required package: Rcpp

## rstanarm (Version 2.18.2, packaged: 2018-11-08 22:19:38

## - Do not expect the default priors to remain the same in

## Thus, R scripts should specify priors explicitly, even i

## - For execution on a local, multicore CPU with excess RA

## options(mc.cores = parallel::detectCores())

## - Plotting theme set to bayesplot::theme_default().

##
```

print output

```
prior_summary(object=M1_stanlmer)
```

```
## Priors for model 'M1_stanlmer'
## ------
## Intercept (after predictors centered)
##  ~ normal(location = 0, scale = 10)
##      **adjusted scale = 3.83
##
## Auxiliary (sigma)
##  ~ exponential(rate = 1)
##      **adjusted scale = 0.38 (adjusted rate = 1/adjuste
##
## Covariance
##  ~ decov(reg. = 1, conc. = 1, shape = 1, scale = 1)
## ------
## See help('prior_summary.stanreg') for more details
```

print output

```
print(M1_stanlmer,digits=2)
```

```
## stan_lmer
##  family:       gaussian [identity]
##  formula:      `passing distance` ~ (1 | kerb)
##  observations: 2355
## ------
##             Median MAD_SD
## (Intercept) 1.54   0.06
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.37   0.01
##
## Error terms:
##  Groups   Name        Std.Dev.
##  kerb     (Intercept) 0.162
##  Residual             0.371
```

need to do more on example! code, interp, etc

Maybe do simple hospital ranking (or hw?). Then pose question – is hospital mortality high because they're bad, or because sicker people are going there, or both?

Use that old example from Kristian and James? I think I used it at UNC in my bayes class there last.