Lecture 8 - Key

Hierarchical Modeling

This chapter focuses on comparison of means across groups and more generally Bayesian hierarchical modeling. Hierarchical modeling is defined by datasets with a multilevel structure, such as:

- patients within hospitals or
- students within school.

The most basic form of this type of data consists of two-levels, groups and individuals within groups.

Recall, observations are exchangeable if $p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n})$. Consider where $Y_1, \ldots, Y_n$ are test scores from randomly selected students from a given STAT 216 instructor/course. If exchangeability holds for these values, then:

$$\begin{aligned} \phi &\sim p(\phi), \\ Y_1, \ldots, Y_n|\phi &\sim \text{i.i.d. } p(y|\phi). \end{aligned}$$

The exchangeability can be interpreted that the random variables are independent samples from a population with a parameter, ϕ. For instance in a normal model, $\phi = \{\theta, \sigma^2\}$ and the data are conditionally independent from a normal distribution $N(\theta, \sigma^2)$.

In a hierarchical framework this can be extended to include the group number:

$$Y_{1,j}, \ldots, Y_{n_j,j}|\phi_j \sim \text{i.i.d. } p(y|\phi_j).$$

The question now is how to we characterize the information between $\phi_1, \ldots, \phi_m$?

Is it reasonable to assume that the values are independent, that is does the information from ϕ_i tell you anything about ϕ_j?

Now consider the groups as samples from a larger population, then using the idea of exchangeability with group-specific parameters gives:

$$\phi_1, \ldots, \phi_m | \psi \sim \text{i.i.d. } p(\phi|\psi).$$

This is similar to the idea of a random effect model and gives the following hierarchical probability model:

$$\begin{array}{rcll} y_{1,j}, \ldots, y_{n_j,j} | \phi_j & \sim & p(y|\phi_j) & \text{(within-group sampling variability)} \\ \phi_1, \ldots, \phi_m | \psi & \sim & p(\phi|\psi) & \text{(between-group sampling variability)} \\ \psi & \sim & p(\psi) & \text{(prior distribution)} \end{array}$$

The distributions $p(y|\phi)$ and $p(\phi|\psi)$ represent sampling variability:

- $p(y|\phi)$ represents variability among measurements within a group and
- $p(\phi|\psi)$ represents sampling variability across groups.

Hierarchical normal model

The hierarchical normal model is often used for modeling differing means across a population.

$$\begin{array}{rcl} \phi_j = \{\theta_j, \sigma^2\}, p(y|\phi_j) & = & normal(\theta_j, \sigma^2) \text{ within-group model} \\ \psi = \{\mu, \tau\}, p(\theta_j|\psi) & = & normal(\mu, \tau^2) \text{ between-group model} \end{array}$$

Note this model specification assumes constant variance for each within-group model, but this assumption can be relaxed.

This model contains three unknown parameters that need priors, we will use the standard semi-conjugate forms:

$$\begin{array}{rcl} \sigma^2 & \sim & InvGamma(\nu_0/2, \nu_0\sigma_0^2/2) \\ \tau^2 & \sim & InvGamma(\eta_0/2, \eta_0\tau_0^2/2) \\ \mu & \sim & normal(\mu_0, \gamma_0^2) \end{array}$$

Given these priors, we need to derive the full conditional distributions in order to make draws from the posterior distribution. Note the joint posterior distribution, can be expressed as:

$$\begin{array}{rcl} p(\tilde{\theta}, \mu, \tau^2, \sigma^2 | \tilde{y}_1, \ldots, \tilde{y}_n) & \propto & p(\mu, \tau^2, \sigma^2) \times p(\tilde{\theta}|\mu, \tau^2, \sigma^2) \times p(\tilde{y}_1, \ldots, \tilde{y}_m | \tilde{\theta}, \mu, \tau^2, \sigma^2) \\ & \propto & p(\mu)p(\sigma^2)p(\tau^2) \times \left(\prod_{j=1}^{m} p(\theta_j|\mu, \tau^2) \right) \times \left(\prod_{j=1}^{m} \prod_{i=1}^{n_j} p(y_{i,j}|\theta_j, \sigma^2) \right). \end{array}$$

Posterior Samples

1. Sampling μ: $p(\mu|-) \propto p(\mu) \prod_{j=1}^{m} p(\theta_j|\mu, \tau^2)$. This is a familiar setting with two normal models, hence, the posterior is also a normal distribution.

$$\mu|- \sim normal\left(\frac{m\bar{\theta}/\tau^2+\mu_0/\gamma_0^2}{m/\tau^2+1/\gamma_0^2}, \left[m/\tau^2+1/\gamma_0^2\right]^{-1}\right)$$

2. Sampling τ^2: $p(\tau^2|-) \propto p(\tau^2) \prod_{j=1}^{m} p(\theta_j|\mu, \tau^2)$. Again this is similar to what we have seen before.

$$\tau^2|- \sim InvGamma\left(\frac{\eta_0+m}{2}, \frac{\eta_0\tau_0^2+\sum_j(\theta_j-\mu)^2}{2}\right)$$

Now what about $\theta_1, \ldots, \theta_m$?

3. Sampling $\theta_1, \ldots, \theta_m\}$. Consider a single θ_j, then $\theta_j|- \propto p(\theta_j|\mu, \tau^2) \prod_{i=1}^{n_j} p(y_{i,j}|\theta_j, \sigma^2)$. Again this is the case where we have two normal distributions.

$$\theta_j|- \sim normal\left(\frac{n_j,\bar{y}_j/\sigma^2+1/\tau^2}{n_j/\sigma^2+1/\tau^2}, \left[n_j/\sigma^2+1/\tau^2\right]^{-1}\right)$$

4. Sampling $\sigma^2\}$: $p(\sigma^2|-) \propto p(\sigma^2) \prod_{j=1}^{m} \prod_{i=1}^{n_j} p(y_{i,j}|\theta_j, \sigma)$.

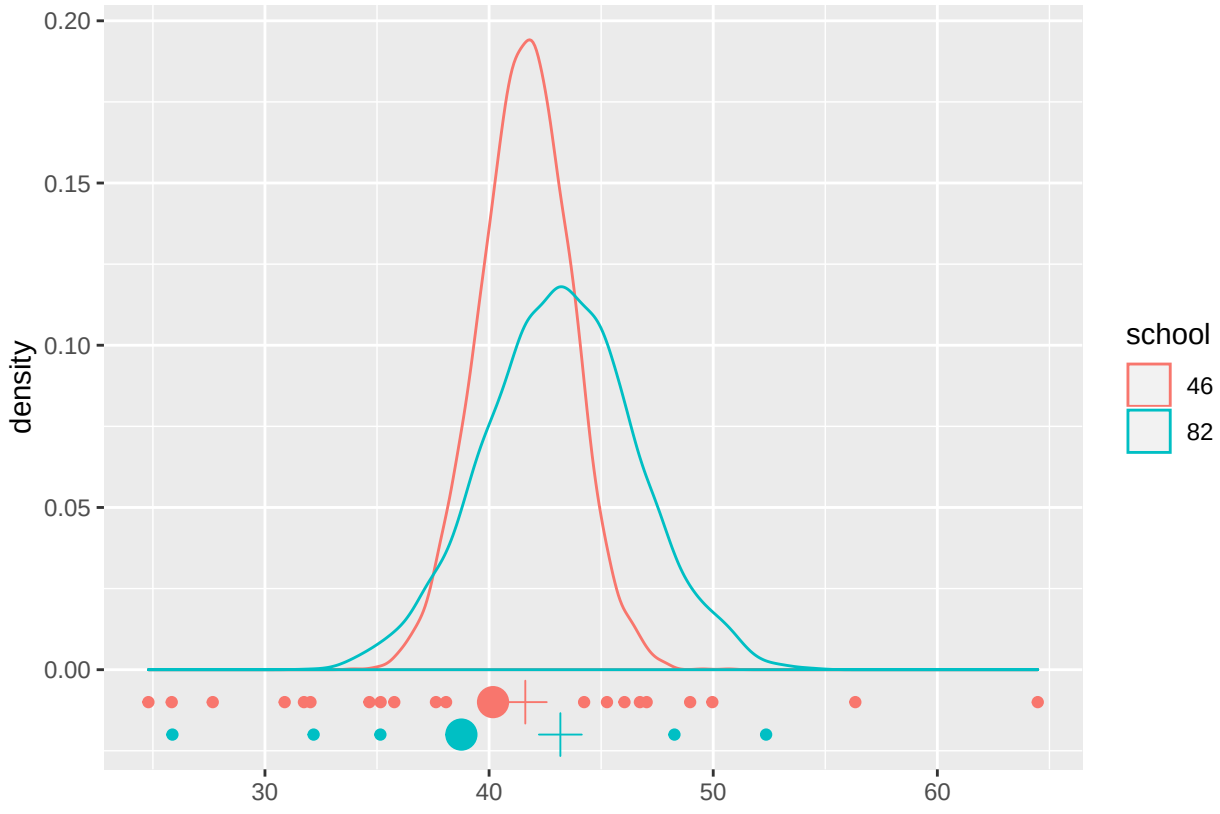
$$\sigma^2|- \sim InvGamma\left(\frac{1}{2}\left[\nu_0+\sum_{j=1}^{m} n_j\right], \frac{1}{2}\left[\nu_0\sigma_0^2+\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j}-\theta_j)^2\right]\right).$$

Data Example

Consider the dataset outline in Chapter 8, that focuses on math tests scores for students spread across 100 schools. Using the Gibbs sampling procedure described above we can fit this model, code courtesy of textbook.

```
Y <- dget("http://www2.stat.duke.edu/~pdh10/FCBS/Inline/Y.school.mathscore")

### weakly informative priors
nu0 <- 1
sigmasq.0 <-100
eta.0<-1
tausq.0<-100
mu.0<-50
gammasq.0<-25
###

### starting values
m <- length(unique(Y[,1])) # number of schools
n<-sv<-ybar<-rep(NA,m)
for(j in 1:m) {
  ybar[j]<-mean(Y[Y[,1]==j,2])
  sv[j]<-var(Y[Y[,1]==j,2])
  n[j]<-sum(Y[,1] ==j)
}

theta<-ybar
sigma2<-mean(sv)
mu<-mean(theta)
tau2<-var(theta)
###

### setup MCMC
set.seed(1)
S<-5000
THETA<-matrix( nrow=S,ncol=m)
MST<-matrix( nrow=S,ncol=3)
###

### MCMC algorithm
for (s in 1:S){
  # sample new values of the thetas
  for (j in 1:m){
    vtheta<-1/(n[j]/sigma2+1/tau2)
    etheta<-vtheta*(ybar[j]*n[j]/sigma2+mu/tau2)
    theta[j]<-rnorm(1,etheta,sqrt(vtheta))
  }

  #sample new value of sigma2
  nun<-nu0+sum(n)
  ss<-nu0*sigmasq.0;
  for(j in 1:m){
    ss<-ss+sum((Y[[j]]-theta[j])^2)
  }
  sigma2<-1/rgamma(1,nun/2,ss/2)
```

```r
  #sample a new value of mu
  vmu<- 1/(m/tau2+1/gammasq.0)
  emu<- vmu*(m*mean(theta)/tau2 + mu.0/gammasq.0)
  mu<-rnorm(1,emu,sqrt(vmu))

  # sample a new value of tau2
  etam<-eta.0+m
  ss<- eta.0*tausq.0 + sum( (theta-mu)^2 )
  tau2<-1/rgamma(1,etam/2,ss/2)

  #store results
  THETA[s,]<-theta
  MST[s,]<-c(mu,sigma2,tau2)
}
```



Interpret the figure created above. What are the small circle, large circles, and the plus signs?

Why does this happen and is it a good thing?

Shrinkage

Recall the posterior mean can be represented as a weighted average, specifically:

$$E[\theta_j|\tilde{y}_j, \mu, \tau^2, \sigma^2] = \frac{\bar{y}_j n_j/\sigma^2 + \mu/\tau^2}{n_j/\sigma^2 + 1/\tau^2}. \tag{1}$$

In this case μ and τ^2 are not chosen parameters from prior distributions, but rather they come from the between group model. So the posterior means for test scores at each school are pulled from the sample mean toward the overall group mean across all of the schools. This phenemenon is known as *bf shrinkage*.

Schools with more students taking the exam see less shrinkage, as there is more weight on the data given more observations. So the figure we discussed before, shows more shrinkage for school 82 as there were fewer observations.

So what about shrinkage, does it make sense? Is it a good thing?

We will soon see that it is an extremely powerful tool and actually dominates the unbiased estimator (the MLE for each distribution). This suprising result is commonly known as Stein's Paradox.

Hierarchical Modeling of Means and Variances

The model we just described and fit was somewhat restrictive in that each school was known to have a common variance. It is likely that schools with a more heterogenous mix of students would have greater variance in the test scores. There are a couple of solutions, the first involves a set of i.i.d. priors on each σ_j^2

$$\sigma_1^2, \ldots, \sigma_m^2 \sim \text{i.i.d. } gamma(\nu_0/2, \nu_0\sigma_0^2/2), \tag{2}$$

however, this results in a full conditional distribution for σ_j that only takes advantage of data from school j. In other words no information from the other schools is used to estimate that variance.

Another option is to consider ν_0 and σ_0^2 as parameters to be estimated in the hierarchical model. A common prior for σ_0^2 would be $p(\sigma_0^2) \sim Gamma(a, b)$. Unfortunately, there is not a semi-conjugate prior distribution for ν_0. The textbook suggests a geometric distribution, where $p(\nu_0) \propto \exp(-\alpha\nu_0)$. Then the full conditional distribution allows a sampling procedure that enumerates over the domain of possible values. This procedure allows shrinkage for the variance terms as well. It is worth noting, that pooling variances is a common way to tackle this particular problem.