Hierarchical models I

Dr. Olanrewaju Michael Akande

Feb 28, 2020

Announcements

- No HW today.
- Next HW immediately after spring break on Monday, March 16.
- Midterm exam next Friday, March 6.
- Practice questions on Sakai later today or tomorrow.
- Review session next Wednesday, March 4.

Outline

- Introduction to hierarchical models
- Shrinkage
- Comparing two groups
- BMI example
- Comparing multiple groups with same variance

Motivation

- Sometimes, we may have a natural grouping in our data, for example
  - students within schools,
  - patients within hospitals,
  - voters within counties or states,
  - biology data, where animals are followed within natural populations organized geographically and, in some cases, socially.
- For such grouped data, we may want to do inference across all the groups, for example, comparison of the group means.
- Ideally, we should do so in a way that takes advantage of the relationship between observations in the same group, but we should also look to borrow information across groups when possible.
- Hierarchical modeling provides a principled way to do so.

Bayes estimators and bias

- Recall the normal model:

$$y_i|\mu, \sigma^2 \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right).$$

- The MLE for the population mean μ is just the sample mean $\bar{y}$.
- $\bar{y}$ is unbiased for μ. That is, for any data $y_i \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, $\mathbb{E}[\bar{y}] = \mu$.
- However, recall that in the conjugate Normal-Gamma normal for example, the posterior expectation is a **weighted average** of the prior mean and the sample mean.
- That is, it is actually biased!

Shrinkage

- Usually through the weighting of the sample data and prior, the Bayes procedure has the tendency to pull the estimate of μ toward the prior mean.
- Of course, the magnitude of the pull depends on the sample size.
- This "pulling" phenomenon is referred to as shrinkage.
- Why would we ever want to do this? Why not just stick with the MLE?
- Well, in part, because shrinkage estimators are often "more accurate" in prediction problems – i.e. they tend to do a better job of predicting a future outcome or of recovering the actual parameter values. Remember variance-bias trade off!
- The fact that a biased estimator would do a better job in many prediction problems can be proven rigorously, and is referred to as Stein's paradox.

Modern relevance

- Stein's result implies, in particular, that the sample mean is an *inadmissible* estimator of the mean of a multivariate normal distribution in more than two dimensions – i.e. there are other estimators that will come closer to the true value in expectation.
- In fact, these are Bayes point estimators (the posterior expectation of the parameter μ).
- Most of what we do now in high-dimensional statistics is develop biased estimators that perform better than unbiased ones.
- Examples: lasso regression, ridge regression, various kinds of hierarchical Bayesian models, etc.
- Today we will get a very basic introduction to Bayesian hierarchical models, which provide a formal and coherent framework for constructing shrinkage estimators.

Why hierarchical models?

- **Bayesian hierarchical models** is a sort of catch-all phrase for a large class of models that have several levels of conditional distributions making up the prior.
- Like simpler one-level priors, they also accomplish shrinkage. However, they are much more flexible.
- Why use them? Several reasons:
  - We may want to exploit more complex dependence structures.
  - We may have many parameters relative to the amount of data that we have, and want to borrow information in estimating them.
  - We may want to shrink toward something other than a simple prior mean/hyper-parameter.

Comparing two groups

- Suppose we want to do inference on mean body mass index (BMI) for two groups (male or female).
- BMI is known to often follow a normal distribution, so let's assume the same here.
- We should expect some relationship between the mean BMI for the two groups.
- We may also think the shape of the two distributions would be relatively the same (at least as a simplifying assumption for now).
- Thus, a reasonable model might be

$$
\begin{aligned}
y_{i,male} &\overset{iid}{\sim} \mathcal{N}\left(\theta_m, \sigma^2\right); \;\; i = 1, \ldots, n_m; \\
y_{i,female} &\overset{iid}{\sim} \mathcal{N}\left(\theta_f, \sigma^2\right); \;\; i = 1, \ldots, n_f.
\end{aligned}
$$

but with some relationship between θ_m and θ_f.

Application

- First, let's do classical inference on such data. The data we will use in the R package `rethinking`.

```r
#install.packages(c("coda","mvtnorm","devtools","loo","dagitty"))
#library(devtools)
#devtools::install_github("rmcelreath/rethinking",ref="Experimental")
#library(rethinking)
data(Howell1)

Howell1[1:15,]
```

```
##     height   weight  age male
## 1  151.765 47.82561 63.0    1
## 2  139.700 36.48581 63.0    0
## 3  136.525 31.86484 65.0    0
## 4  156.845 53.04191 41.0    1
## 5  145.415 41.27687 51.0    0
## 6  163.830 62.99259 35.0    1
## 7  149.225 38.24348 32.0    0
## 8  168.910 55.47997 27.0    1
## 9  147.955 34.86988 19.0    0
## 10 165.100 54.48774 54.0    1
## 11 154.305 49.89512 47.0    0
## 12 151.130 41.22017 66.0    1
## 13 144.780 36.03221 73.0    0
## 14 149.900 47.70000 20.0    0
## 15 150.495 33.84930 65.3    0
```

Data

- For now, focus on data for individuals under age 15.

```
htm <- Howell1$height/100
bmi <- Howell1$weight/(htm^2)
y_male <- bmi[Howell1$age<15 & Howell1$male==1]
y_female <- bmi[Howell1$age<15 & Howell1$male==0]
n_m <- length(y_male)
n_f <- length(y_female)

n_f
```

```
## [1] 84
```

```
n_m
```

```
## [1] 77
```

```
summary(y_male)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.07   13.87   14.63   14.84   15.53   18.22
```

```
summary(y_female)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   9.815  13.559  14.305  14.585  15.712  18.741
```

Classical inference

- No significant difference in group means.

```
t.test(y_male,y_female)
```

```
##
##     Welch Two Sample t-test
##
## data:  y_male and y_female
## t = 1.1204, df = 157.87, p-value = 0.2643
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1947946  0.7054729
## sample estimates:
## mean of x mean of y
##  14.84037  14.58503
```

Simple weighted estimator

- One parameterization that can reflect some relationship between θ_m and θ_f is

$$y_{i,male} \stackrel{iid}{\sim} \mathcal{N}\left(\mu + \delta, \sigma^2\right); \;\; i = 1, \ldots, n_m;$$
$$y_{i,female} \stackrel{iid}{\sim} \mathcal{N}\left(\mu - \delta, \sigma^2\right); \;\; i = 1, \ldots, n_f.$$

where

  - $\theta_m = \mu + \delta$ and $\theta_f = \mu - \delta$,
  - $\mu = \dfrac{\theta_m + \theta_f}{2}$ is the pooled average, and
  - $\delta = \dfrac{\theta_m - \theta_f}{2}$ is half of the population difference in means.

Simple weighted estimator

- Convenient prior:
  - $\pi(\mu, \delta, \sigma^2) = \pi(\mu) \cdot \pi(\delta) \cdot \pi(\sigma^2)$, where
    - $\pi(\mu) = \mathcal{N}(\mu_0, \gamma_0^2)$,
    - $\pi(\delta) = \mathcal{N}(\delta_0, \tau_0^2)$, and
    - $\pi(\sigma^2) = \mathcal{IG}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$.
- We will set the hyper-parameters as:
  - $\mu_0 = 15, \gamma_0 = 5$,
  - $\delta_0 = 0, \tau_0 = 3$,
  - $\nu_0 = 1, \sigma_0 = 5$.
- Do these values seem reasonable to you?

SIMPLE WEIGHTED ESTIMATOR

- Note that we can rewrite

$$y_{i,male} \overset{iid}{\sim} \mathcal{N}\left(\mu + \delta, \sigma^2\right); \;\; i = 1, \ldots, n_m;$$
$$y_{i,female} \overset{iid}{\sim} \mathcal{N}\left(\mu - \delta, \sigma^2\right); \;\; i = 1, \ldots, n_f$$

as

$$(y_{i,male} - \delta) \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right); \;\; i = 1, \ldots, n_m;$$
$$(y_{i,female} + \delta) \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right); \;\; i = 1, \ldots, n_f$$

or

$$(y_{i,male} - \mu) \overset{iid}{\sim} \mathcal{N}\left(\delta, \sigma^2\right); \;\; i = 1, \ldots, n_m;$$
$$(-1)(y_{i,female} - \mu) \overset{iid}{\sim} \mathcal{N}\left(\delta, \sigma^2\right); \;\; i = 1, \ldots, n_f.$$

as needed, so we can leverage past results for the full conditionals.

Full conditionals

- For the full conditionals we will derive today, we will take advantage of previous results from the regular univariate normal model.

- Recall that if we assume

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \;\; i = 1, \ldots, n,$$

and set our priors to be

$$\pi(\mu) = \mathcal{N}\left(\mu_0, \gamma_0^2\right).$$
$$\pi(\sigma^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right),$$

then we have

$$\pi(\mu, \sigma^2 | Y) \propto \left\{ \prod_{i=1}^{n} p(y_i | \mu, \sigma^2) \right\} \cdot \pi(\mu) \cdot \pi(\sigma^2)$$

Full conditionals

- We have

$$\pi(\mu|\sigma^2, Y) = \mathcal{N}\left(\mu_n, \gamma_n^2\right).$$

where

$$\gamma_n^2 = \frac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{\gamma_0^2}}; \qquad \mu_n = \gamma_n^2 \left[\frac{n}{\sigma^2}\bar{y} + \frac{1}{\gamma_0^2}\mu_0\right],$$

- and

$$\pi(\sigma^2|\mu, Y) = \mathcal{IG}\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}\right),$$

where

$$\nu_n = \nu_0 + n; \qquad \sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + \sum_{i=1}^{n}(y_i - \mu)^2\right].$$

Full conditionals

- With $\pi(\mu) = \mathcal{N}(\mu_0, \gamma_0^2)$, we have

$$\mu | Y, \delta, \sigma^2 \sim \mathcal{N}(\mu_n, \gamma_n^2), \quad \text{where}$$

$$\gamma_n^2 = \frac{1}{\dfrac{1}{\gamma_0^2} + \dfrac{n_m + n_f}{\sigma^2}}$$

$$\mu_n = \gamma_n^2 \left[\frac{\mu_0}{\gamma_0^2} + \frac{n_m \overline{(y_{i,male} - \delta)} + n_f \overline{(y_{i,female} + \delta)}}{\sigma^2} \right].$$

where

- $\overline{(y_{i,male} - \delta)} = \dfrac{1}{n_m} \sum_{i=1}^{n_m} (y_{i,male} - \delta)$, and
- $\overline{(y_{i,female} + \delta)} = \dfrac{1}{n_f} \sum_{i=1}^{n_f} (y_{i,female} + \delta)$.

Full conditionals

- With $\pi(\delta) = \mathcal{N}(\delta_0, \tau_0^2)$, we have

$$\delta | Y, \mu, \sigma^2 \sim \mathcal{N}(\delta_n, \tau_n^2), \quad \text{where}$$

$$\tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n_m + n_f}{\sigma^2}}$$

$$\delta_n = \tau_n^2 \left[\frac{\delta_0}{\tau_0^2} + \frac{n_m \overline{(y_{i,male} - \mu)} + (-1) n_f \overline{(y_{i,female} + \mu)}}{\sigma^2} \right].$$

where

- $\overline{(y_{i,male} - \mu)} = \frac{1}{n_m} \sum_{i=1}^{n_m} (y_{i,male} - \mu)$, and
- $\overline{(y_{i,female} - \mu)} = \frac{1}{n_f} \sum_{i=1}^{n_f} (y_{i,female} - \mu)$.

Full conditionals

- With $\pi(\sigma^2) = \mathcal{IG}(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$, we have

$$\sigma^2|Y, \mu, \delta \sim \mathcal{IG}(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2}), \quad \text{where}$$

$$\nu_n = \nu_n + n_m + n_f$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + \sum_{i=1}^{n_m}(y_{i,male} - [\mu + \delta])^2 + \sum_{i=1}^{n_f}(y_{i,female} - [\mu - \delta])^2\right].$$

Application to data

```
#priors
mu0 <- 15; gamma02 <- 5^2
delta0 <- 0; tau02 <- 3^2
nu0 <- 1; sigma02 <- 5^2

#starting values
mu <- (mean(y_male) + mean(y_female))/2
delta <- (mean(y_male) - mean(y_female))/2
#no need for starting values for sigma_squared, we can sample it first

MU <- DELTA <- SIGMA2 <- NULL
```

Application to data

```r
#set seed
set.seed(1234)

#set number of iterations and burn-in
n_iter <- 10000; burn_in <- 0.2*n_iter

##Gibbs sampler
for (s in 1:(n_iter+burn_in)) {
#update sigma2
sigma2 <- 1/rgamma(1,(nu0 + n_m + n_f)/2,
                   (nu0*sigma02 + sum((y_male-mu-delta)^2) + sum((y_female-mu+delta)^2))/2)

#update mu
gamma2n <- 1/(1/gamma02 + (n_m + n_f)/sigma2)
mun <- gamma2n*(mu0/gamma02 + sum(y_male-delta)/sigma2 + sum(y_female+delta)/sigma2)
mu <- rnorm(1,mun,sqrt(gamma2n))

#update delta
tau2n <- 1/(1/tau02 + (n_m+n_f)/sigma2)
deltan <- tau2n*(delta0/tau02 + sum(y_male-mu)/sigma2 - sum(y_female-mu)/sigma2)
delta <- rnorm(1,deltan,sqrt(tau2n))

#save parameter values
MU <- c(MU,mu); DELTA <- c(DELTA,delta); SIGMA2 <- c(SIGMA2,sigma2)
}
```

Posterior summaries

```
#library(coda)
MU.mcmc <- mcmc(MU,start=1)
summary(MU.mcmc)
```

```
##
## Iterations = 1:12000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 12000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##      14.712517       0.118765       0.001084       0.001089
##
## 2. Quantiles for each variable:
##
##  2.5%   25%   50%   75% 97.5%
## 14.48 14.63 14.71 14.79 14.95
```

```
(mean(y_male) + mean(y_female))/2 #compare to data
```

```
## [1] 14.7127
```

Posterior summaries

```
DELTA.mcmc <- mcmc(DELTA,start=1)
summary(DELTA.mcmc)
```

```
##
## Iterations = 1:12000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 12000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##       0.127657       0.119522       0.001091       0.001091
##
## 2. Quantiles for each variable:
##
##     2.5%      25%      50%      75%    97.5%
## -0.10691  0.04791  0.12743  0.20796  0.36407
```

```
summary((2*DELTA)) #rescale as difference in group means
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.63464  0.09582  0.25487  0.25531  0.41592  1.23660
```

```
mean(y_male) - mean(y_female) #compare to data
```

```
## [1] 0.2553392
```

Posterior summaries

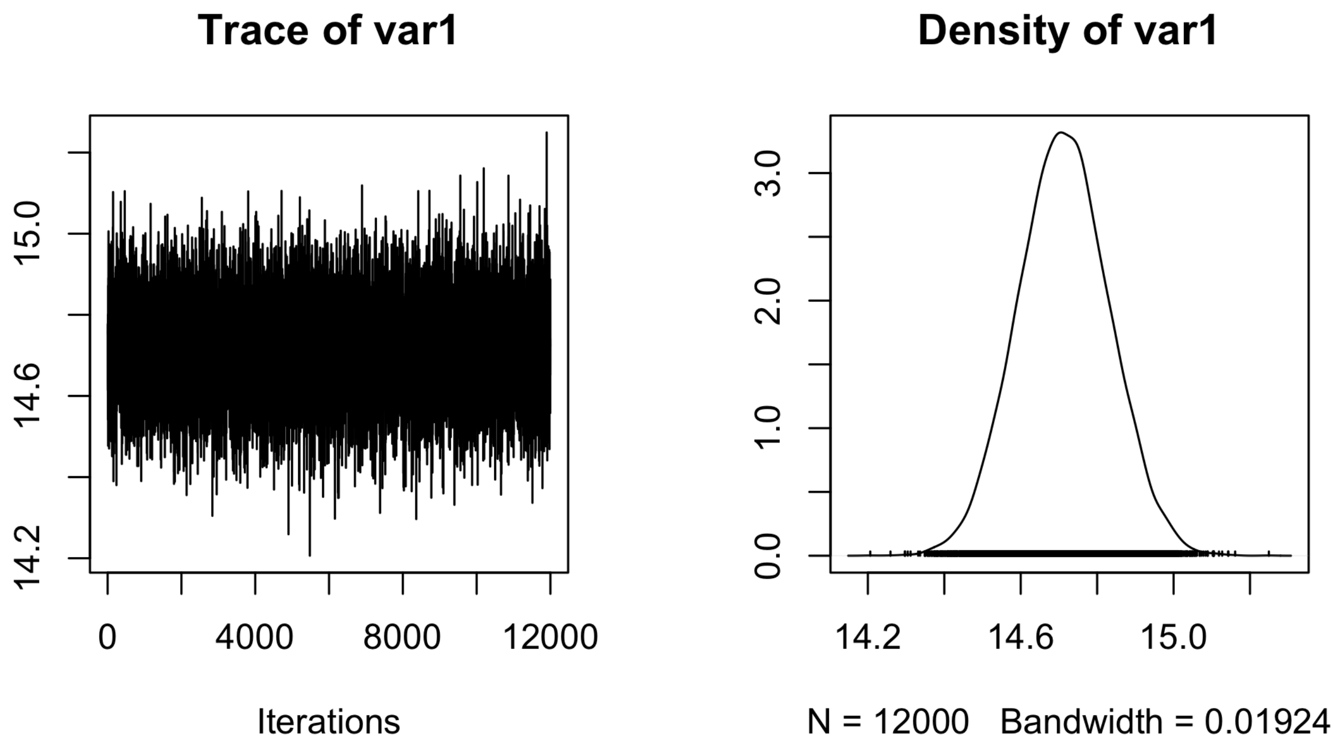
```
SIGMA2.mcmc <- mcmc(SIGMA2,start=1)
summary(SIGMA2.mcmc)
```

```
##
## Iterations = 1:12000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 12000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##       2.287927       0.257689       0.002352       0.002352
##
## 2. Quantiles for each variable:
##
##  2.5%   25%   50%   75% 97.5%
## 1.833 2.107 2.272 2.455 2.841
```

STA 602L

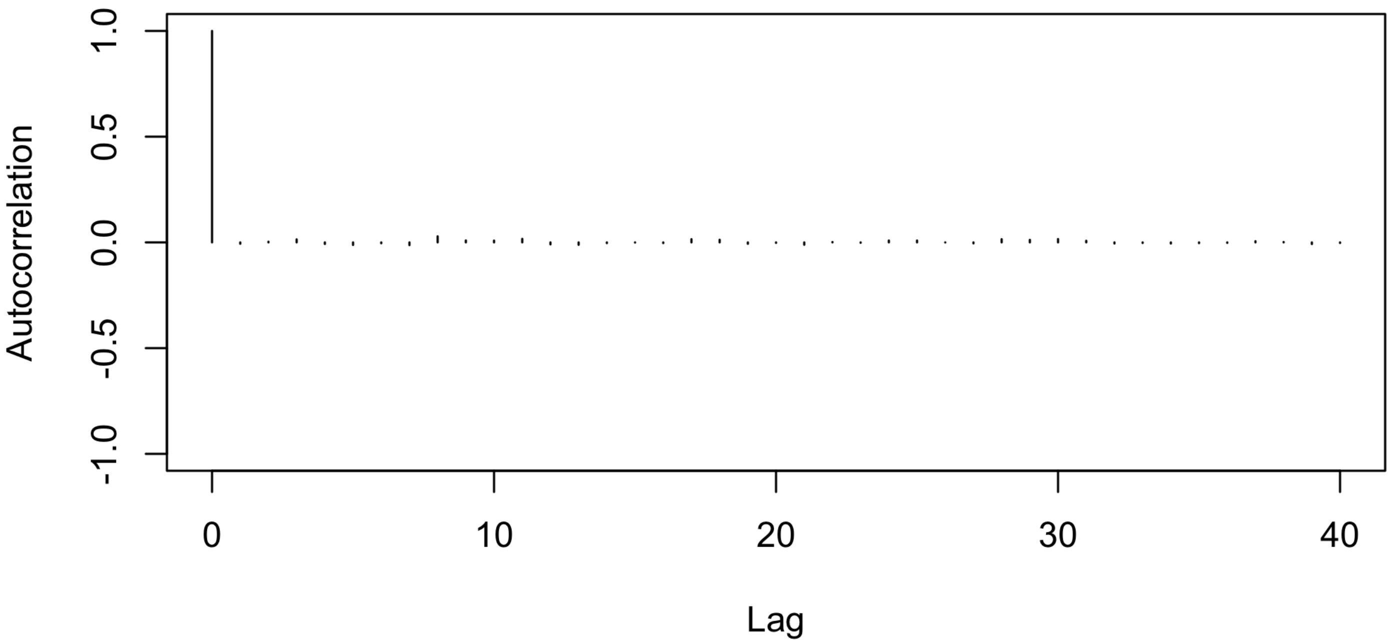
Diagnostics

```
plot(MU.mcmc)
```



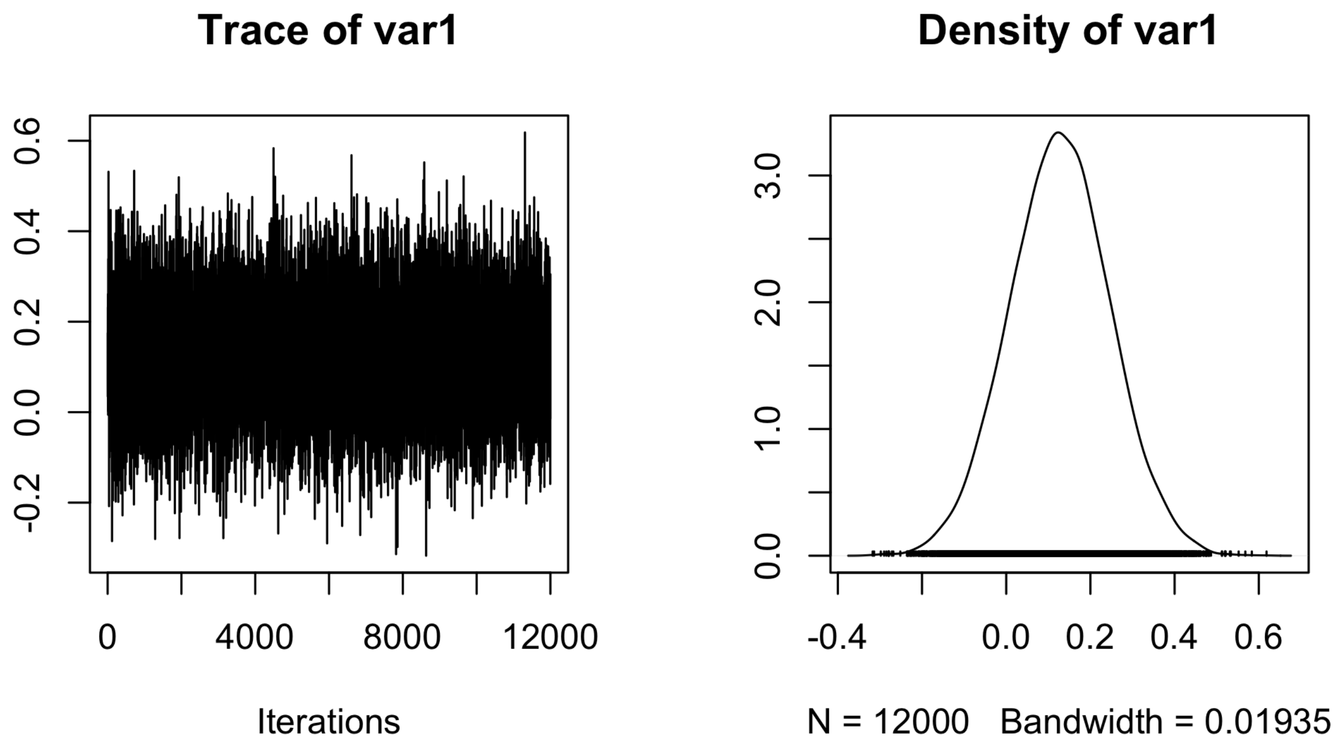
Diagnostics

```
autocorr.plot(MU.mcmc)
```



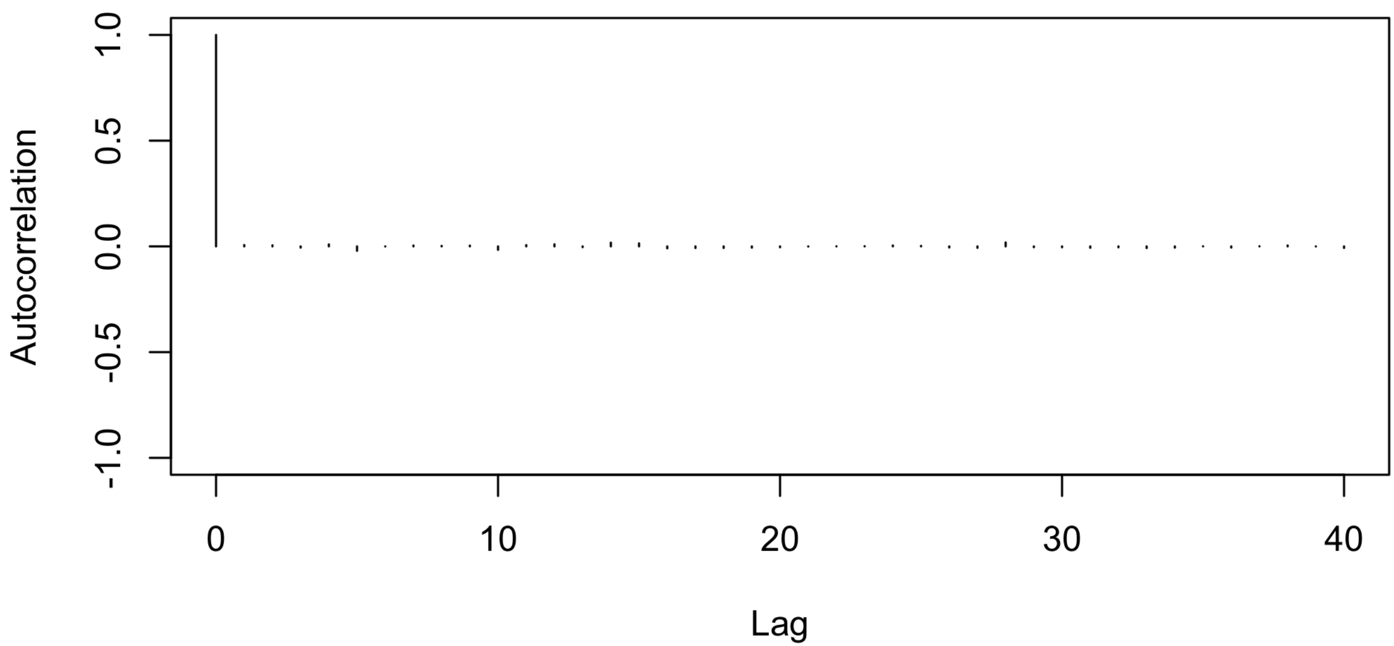
Diagnostics

```
plot(DELTA.mcmc)
```



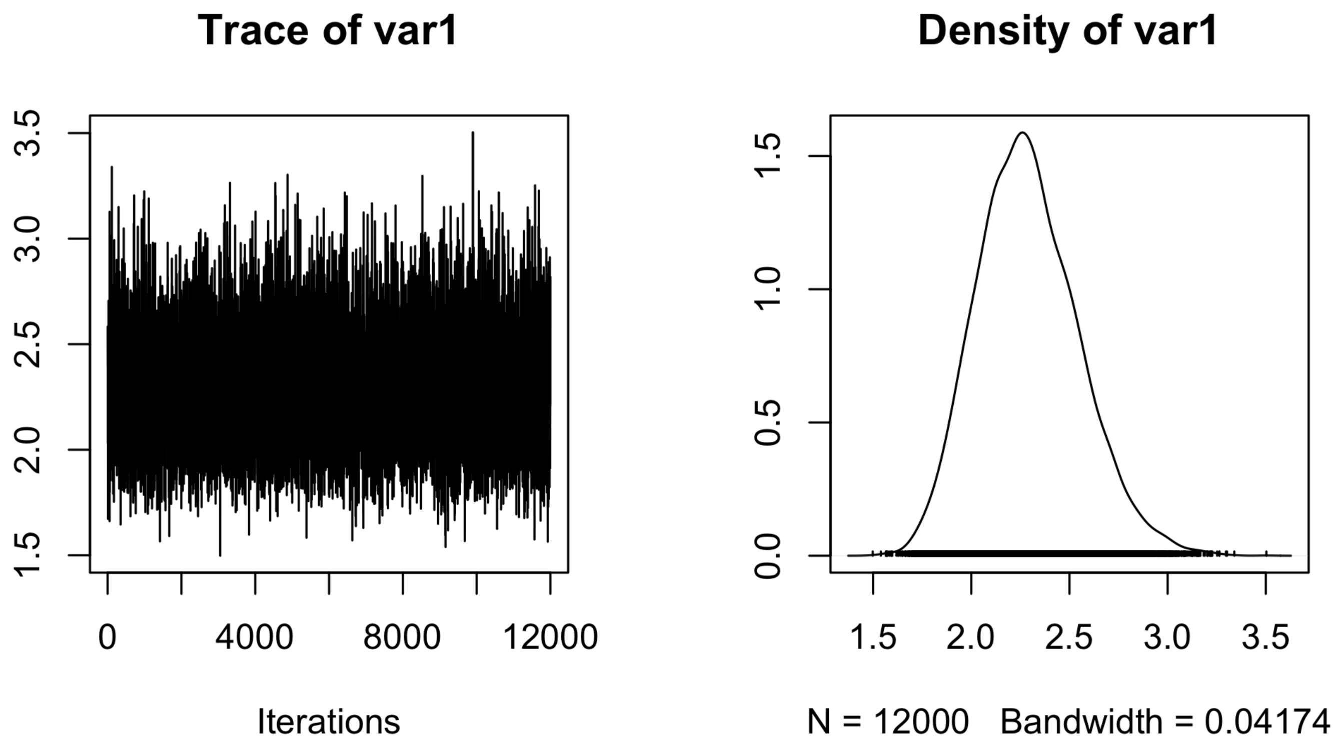
Diagnostics

```
autocorr.plot(DELTA.mcmc)
```



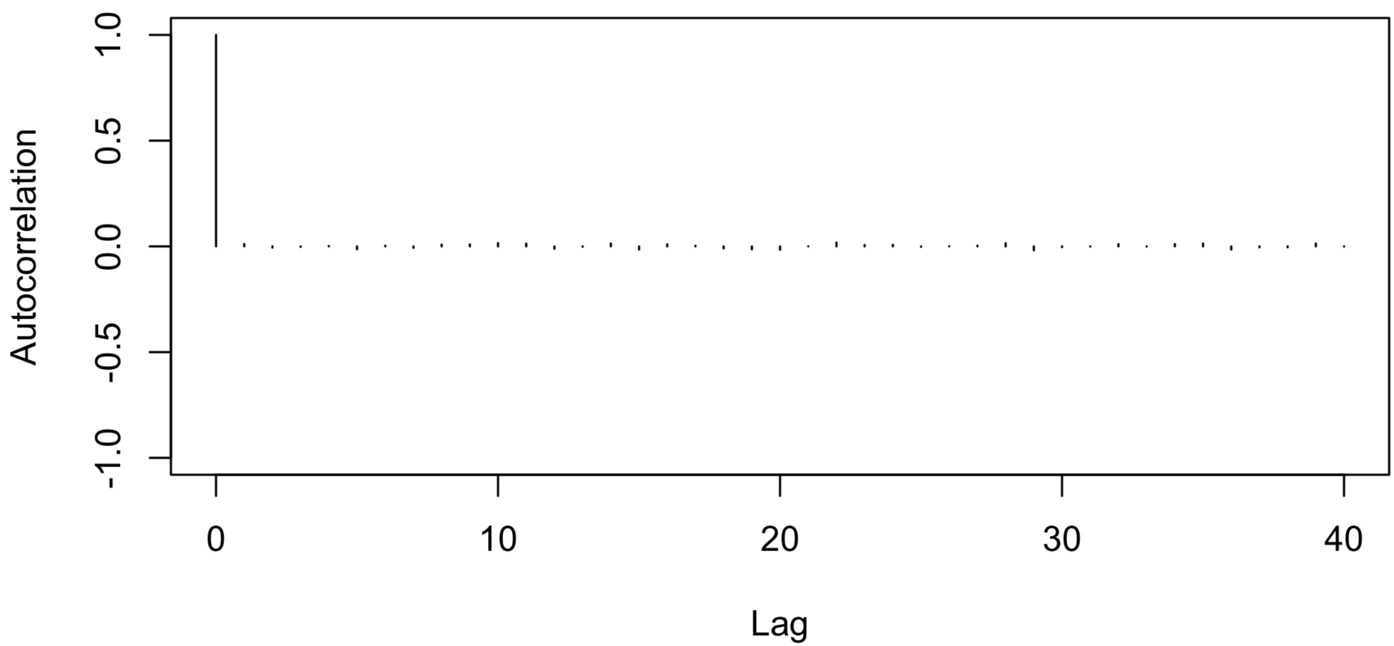
Diagnostics

```
plot(SIGMA2.mcmc)
```



Diagnostics

```
autocorr.plot(SIGMA2.mcmc)
```



Application to data

- Posterior probability that boys have larger average BMI than girls is 0.86!
- Posterior medians and 95% credible intervals for the group means are actually quite similar to the unpooled (gender specific) intervals from classified inference.

```
#mean for boys
quantile((MU+DELTA),probs=c(0.025,0.5,0.975))
```

```
##     2.5%      50%    97.5%
## 14.50255 14.84146 15.17925
```

```
#mean for girls
quantile((MU-DELTA),probs=c(0.025,0.5,0.975))
```

```
##     2.5%      50%    97.5%
## 14.26848 14.58276 14.90761
```

```
#posterior probability girls have larger BMI than boys
mean(DELTA > 0)
```

```
## [1] 0.8571667
```

Application to data

- Let's look at a different sub-population. For older individuals > 75, we only have 8 male and 4 female.

```
y_male <- bmi[Howell1$age > 75 & Howell1$male==1]
y_female <- bmi[Howell1$age > 75 & Howell1$male==0]
n_m <- length(y_male)
n_f <- length(y_female)
n_m
```

```
## [1] 8
```

```
n_f
```

```
## [1] 4
```

Application to data

- A 95% confidence interval for the difference between genders in BMI (estimated as 0.24) is (-4.20,4.68).

```
mean(y_male) - mean(y_female)
```

```
## [1] 0.2408966
```

```
t.test(y_male,y_female)
```

```
##
##     Welch Two Sample t-test
##
## data:  y_male and y_female
## t = 0.13801, df = 5.1869, p-value = 0.8954
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.197948  4.679741
## sample estimates:
## mean of x mean of y
##  18.06751  17.82662
```

Application to data

- Let's apply the shrinkage model with these priors:
  - $\mu_0 = 18, \gamma_0 = 5$,
  - $\delta_0 = 0, \tau_0 = 3$,
  - $\nu_0 = 1, \sigma_0 = 5$.
- Using the shrinkage model, the posterior mean is 0.25 with 95% CI (-3.45, 3.88).

```
mean((DELTA*2))
```

```
## [1] 0.2493733
```

```
quantile((DELTA*2),probs=c(0.025,0.5,0.975))
```

```
##       2.5%        50%      97.5%
## -3.4466931  0.2758598  3.8762543
```

- Our precision has been improved by borrowing of information across the groups. Of course the prior is important here given the sample sizes.

Comparing multiple groups

- Suppose we wish to investigate the mean (and distribution) of test scores for students at J different high schools.
- In each school j, where $j = 1, \ldots, J$, suppose we test a random sample of n_j students.
- Let y_{ij} be the test score for the ith student in school j, with $i = 1, \ldots, n_j$, with

  $$y_{ij} | \theta_j, \sigma_j^2 \sim \mathcal{N}\left(\theta_j, \sigma_j^2\right)$$

  where for each school j, θ_j is the school-wide average test score, and σ_j^2 is the school-wide variance of individual test scores.
- This is what we did for the the Pygmalion study, job training data and the science classroom exercise on homework 3.

School testing example

- Classical inference for each school can be based on large sample 95% CI: $\bar{y}_j \pm 1.96\sqrt{s_j^2/n_j}$, where $\bar{y}_j$ is the sample average in school j, and s_j^2 is the sample variance in school j.
- Clearly, we can overfit the data within schools, for example, what if we only have 4 students from one of the schools? $\bar{y}_j$ can be a good estimate if n_j is large but it may be poor if n_j is small.
- Option II: alternatively, we might believe that $\theta_j = \mu$ for all j; that is, all schools have the same mean. This is the assumption (null hypothesis) in ANOVA models for example. We can also set $\sigma_j^2 = \sigma^2$ for all J.
- Option I ignores that the θ_j's should be reasonably similar, whereas option II ignores any differences between them.
- It would be nice to find a compromise! Borrowing information across, and shrinking our estimate towards a **grand mean** could be very useful here.

School testing example

- For the Pygmalion study and job training data, we focused using priors that are independent between the groups.
- For example, in the conjugate case, we would have

$$\pi(\theta_j|\sigma_j^2) = \mathcal{N}\left(\mu_0, \frac{\sigma_j^2}{\kappa_0}\right)$$
$$\pi(\sigma_j^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$

for some hyperparameters (constants), μ_0, κ_0, ν_0, and σ_0^2.

- In the semi-conjugate case,

$$\pi(\theta_j) = \mathcal{N}\left(\mu_0, \sigma_0^2\right)$$
$$\pi(\sigma_j^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\gamma_0^2}{2}\right)$$

for some hyperparameters (constants), μ_0, σ_0^2, ν_0, and γ_0^2.

Hierarchical normal model

- Instead, we can assume that the θ_j's are drawn from a distribution based on the following: conceive of the schools themselves as being a random sample from all possible schools.

- For now, assume the variance is constant across schools. The hierarchical normal model assumes normal sampling models both within and between groups:

$$
\begin{aligned}
y_{ij}|\theta_j, \sigma^2 &\sim \mathcal{N}\left(\theta_j, \sigma^2\right); \quad i = 1, \ldots, n_j \\
\theta_j|\mu, \tau^2 &\sim \mathcal{N}\left(\mu, \tau^2\right); \quad j = 1, \ldots, J,
\end{aligned}
$$

  which gives us an extra level in the prior on the means, which leads to sharing of information across the groups in estimating the group-specific means.

- We have an extra variance parameter τ^2. Comparing τ^2 to σ^2 tells us how much of the variation in Y is due to within-group versus between-group variation.

Hierarchical normal model

- Standard semi-conjugate priors are given by

$$\pi(\mu) = \mathcal{N}\left(\mu_0, \gamma_0^2\right)$$
$$\pi(\sigma^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$
$$\pi(\tau^2) = \mathcal{IG}\left(\frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right).$$

with

- μ_0: best guess of average of school averages
- γ_0^2: set based on plausible ranges of values of μ
- τ_0^2: best guess of variance of school averages
- η_0: set based on how tight prior for τ^2 is around τ_0^2
- σ_0^2: best guess of variance of individual test scores around respective school means
- ν_0: set based on how tight prior for σ^2 is around σ_0^2.

Exchangeability

- This model relies heavily on exchangeability across units at each level.
- For example, we assume the schools are a random sample from the population of all schools, and the students within schools are a random sample of all the students in each school.
- This is not always completely true.
- Note: we can allow the variance to vary across schools if desired (and we will soon in fact).

Exchangeability

- Turns out that **conditional exchangeability** would be enough if we control for relevant variables in our modeling.
- For example, the schools in Chapel Hill/Carrboro are not entirely exchangeable.
- For example, Phoenix Academy is for students on long-term out-of-school suspension or who need to make up work due to extended absences (e.g., pregnancy), and Memorial Hospital School is for children battling serious illnesses.
- However, if we condition on school type (public, charter, private, special services, home), the schools may then be exchangeable.

Posterior inference

- Recall the model is

$$y_{ij}|\theta_j, \sigma^2 \sim \mathcal{N}\left(\theta_j, \sigma^2\right); \quad i = 1, \ldots, n_j$$
$$\theta_j|\mu, \tau^2 \sim \mathcal{N}\left(\mu, \tau^2\right); \quad j = 1, \ldots, J,$$

- Under our prior specification, we can factor the posterior as follows:

$$\begin{aligned}
\pi(\theta_1, \ldots, \theta_J, \mu, \sigma^2, \tau^2|Y) \propto\ & p(y|\theta_1, \ldots, \theta_J, \mu, \sigma^2, \tau^2) \\
& \times p(\theta_1, \ldots, \theta_J|\mu, \sigma^2, \tau^2) \\
& \times \pi(\mu, \sigma^2, \tau^2) \\
\\
=\ & p(y|\theta_1, \ldots, \theta_J, \sigma^2) \\
& \times p(\theta_1, \ldots, \theta_J|\mu, \tau^2) \\
& \times \pi(\mu) \cdot \pi(\sigma^2) \cdot \pi(\tau^2) \\
\\
=\ & \left\{ \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij}|\theta_j, \sigma^2) \right\} \\
& \times \left\{ \prod_{j=1}^{J} p(\theta_j|\mu, \tau^2) \right\} \\
& \times \pi(\mu) \cdot \pi(\sigma^2) \cdot \pi(\tau^2)
\end{aligned}$$

Full conditional for grand mean

- The full conditional distribution of μ is proportional to the part of the joint posterior $\pi(\theta_1, \ldots, \theta_J, \mu, \sigma^2, \tau^2 | Y)$ that involves μ.

- That is,

$$\pi(\mu|\theta_1, \ldots, \theta_J, \sigma^2, \tau^2, Y) \propto \left\{ \prod_{j=1}^{J} p(\theta_j|\mu, \tau^2) \right\} \cdot \pi(\mu).$$

- This looks like the full conditional distribution from the one-sample normal case, so you can show that

$$\pi(\mu|\theta_1, \ldots, \theta_J, \sigma^2, \tau^2, Y) = \mathcal{N}\left(\mu_n, \gamma_n^2\right) \quad \text{where}$$

$$\gamma_n^2 = \frac{1}{\dfrac{J}{\tau^2} + \dfrac{1}{\gamma_0^2}}; \qquad \mu_n = \gamma_n^2 \left[\frac{J}{\tau^2} \bar{\theta} + \frac{1}{\gamma_0^2} \mu_0 \right]$$

and $\bar{\theta} = \frac{1}{J} \sum\limits_{j=1}^{J} \theta_j$.

Full conditionals for group means

- Similarly, the full conditional distribution of each θ_j is proportional to the part of the joint posterior $\pi(\theta_1, \ldots, \theta_J, \mu, \sigma^2, \tau^2 | Y)$ that involves θ_j.

- That is,

$$\pi(\theta_j|\mu, \sigma^2, \tau^2, Y) \propto \left\{ \prod_{i=1}^{n_j} p(y_{ij}|\theta_j, \sigma^2) \right\} \cdot p(\theta_j|\mu, \tau^2)$$

- Those terms include a normal for θ_j multiplied by a product of normals in which θ_j is the mean, again mirroring the one-sample case, so you can show that

$$\pi(\theta_j|\mu, \sigma^2, \tau^2, Y) = \mathcal{N}\left(\theta_j^\star, \nu_j^\star\right) \quad \text{where}$$

$$\nu_j^\star = \frac{1}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}; \qquad \theta_j^\star = \nu_j^\star \left[\frac{n_j}{\sigma^2} \bar{y}_j + \frac{1}{\tau^2} \mu \right]$$

Full conditionals for group means

- Our estimate for each θ_j is a weighted average of $\bar{y}_j$ and μ, ensuring that we are borrowing information across all levels through μ and τ^2.
- The weights for the weighted average is determined by relative precisions from the data and from the second level model.
- The groups with smaller n_j have estimated $\theta_j^\star$ closer to μ than schools with larger n_j.
- Thus, degree of shrinkage of θ_j depends on ratio of within-group to between-group variances.

Full conditionals for within-group variance

- The full conditional distribution of τ^2 is proportional to the part of the joint posterior $\pi(\theta_1, \ldots, \theta_J, \mu, \sigma^2, \tau^2 | Y)$ that involves τ^2.
- That is,

$$\pi(\tau^2 | \theta_1, \ldots, \theta_J, \mu, \sigma^2, Y) \propto \left\{ \prod_{j=1}^{J} p(\theta_j | \mu, \tau^2) \right\} \cdot \pi(\tau^2)$$

- As in the case for μ, this looks like the one-sample normal problem, and our full conditional posterior is

$$\pi(\tau^2 | \theta_1, \ldots, \theta_J, \mu, \sigma^2, Y) = \mathcal{IG}\left(\frac{\eta_n}{2}, \frac{\eta_n \tau_n^2}{2} \right) \quad \text{where}$$

$$\eta_n = \eta_0 + J; \qquad \tau_n^2 = \frac{1}{\eta_n} \left[\eta_0 \tau_0^2 + \sum_{j=1}^{J} (\theta_j - \mu)^2 \right].$$

Full conditionals for across-group variance

- Finally, the full conditional distribution of σ^2 is proportional to the part of the joint posterior $\pi(\theta_1, \ldots, \theta_J, \mu, \sigma^2, \tau^2 | Y)$ that involves σ^2.
- That is,

$$\pi(\sigma^2 | \theta_1, \ldots, \theta_J, \mu, \tau^2, Y) \propto \left\{ \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij} | \theta_j, \sigma^2) \right\} \cdot \pi(\sigma^2)$$

- We can again take advantage of the one-sample normal problem, so that our full conditional posterior is

$$\pi(\sigma^2 | \theta_1, \ldots, \theta_J, \mu, \tau^2, Y) = \mathcal{IG}\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}\right) \quad \text{where}$$

$$\nu_n = \nu_0 + \sum_{j=1}^{J} n_j; \qquad \sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0 \sigma_0^2 + \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \theta_j)^2\right].$$