GPH-GU2372/3372
Applied Bayesian Analysis in Public Health

Lesson 8: Hierarchical Modeling

Hai Shu, PhD

11/14/2022

Topics

- Introduction to hierarchical models
- Hierarchical normal model
- Shrinkage

Hierarchical models: introduction

- Many statistical applications involve models with multiple parameters that can be assumed to be related or connected by the structure of the problem.
- In these cases, a joint probability model for these parameters is more appropriate than individual probability models as that allows to reflect the dependence among the model parameters.
- Hierarchical models are one way to incorporate dependence structures among model parameters and share information across groups.
  They involve specifying either the prior distribution or the sampling distribution as a series of models.
  Given the fact that hierarchical models often involve models within other models, they are sometimes also called multilevel models.

Hierarchical models: introduction

- Example: Suppose that we take a sample of m hospitals from around the country and within each hospital we sample n_i patients, asking each patient whether or not they were satisfied with their treatment.
- Within each hospital, results should be independent with a probability of satisfaction θ_i that depends on the hospital i.
- Thus, the total number y_i of satisfied patients from hospital i is:

$$y_i|\theta_i \sim \text{Bin}(n_i\theta_i)$$

  independently for $i = 1, \ldots, m$.
- Conditionally on the hospital, individual patient observations should be independent.
- What about unconditionally?

Hierarchical models: introduction

- Unconditionally, observations within a hospital should be more alike than observations from different hospitals.
- The hospitals are a random sample of hospitals, so there should be also a distribution for the hospital satisfaction proportions θ_i.
- One way to model this is to assume that the θ_i's are an i.i.d. sample from some distribution

$$\theta_i | \alpha_1, \alpha_2 \overset{iid}{\sim} \text{Beta}(\alpha_1, \alpha_2)$$

- This is an example of a hierarchical model.

Hierarchical models: introduction

- Example: Suppose that we are interested in assessing the effect of special coaching programs on test scores. Separate randomized experiments were perfomed in each of eight school.
  The outcome variable in each study was the score on a special administration of the SAT-V test, a standardized multiple choice test used to help college make admission decisions.
  The scores can vary between 200 and 800 with a mean of about 500 and a standard deviation of about 100.
- We can model the n_j scores at each of $j = 1, \ldots, 8$ schools as

$$y_{ij} | \mu_j, \sigma_j^2 \sim N(\mu_j, \sigma_j^2)$$

Assuming that there was no reason to believe that any of the eight programs was more effective than any other,

$$\mu_j \overset{iid}{\sim} N(\theta, \tau^2)$$

Additionally, if we believe that the variability in the 8 schools is the same, we can assume $\sigma_j^2 = \sigma^2$ for $j = 1, \ldots, 8$.

Hierarchical models: introduction

- In both of these examples that we considered we had data where there was a hierarchy of nested populations:
  - population of hospitals and population of patients within each hospital
  - population of schools and population of students within each school
- Of course, other situations are also possible, but the key is also the fact that there is a hierarchy of nested populations.
- The simplest multilevel model has only two levels, in which one level consists of groups and the other consists of units within groups.

Hierarchical models

- Let's set up a hierarchical model for out example on the SAT-V scores.
  The data is $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m\}$ where $\mathbf{Y}_j = \{Y_{1,j}, \ldots, Y_{n_j,j}\}$ is the data for school j.
- If we consider the data for a single school j, it makes sense to assume that the $Y_{1,j}, \ldots, Y_{n_j,j}$ are conditionally independent and identically distributed given a parameter ϕ_j, corresponding to school j:

$$Y_{1,j}, Y_{2,j}, \ldots, Y_{n_j,j} | \phi_j = (\theta_j, \sigma_j^2) \overset{iid}{\sim} p(y|\phi_j) = p(y|\theta_j, \sigma_j^2)$$

for $j = 1, 2, \ldots, m$.

Hierarchical models

- Let's consider now the parameters $\phi_1, \ldots, \phi_m$.
- We can think of the schools as a sample from a population of schools. Then, we can model the school-specific parameters $\phi_1, \phi_2, \ldots, \phi_m$ as conditionally independent and identically distributed given a parameter ψ:

$$\phi_1, \ldots, \phi_m | \psi \overset{iid}{\sim} p(\phi|\psi)$$

- Finally, we can place a prior on ψ:

$$\psi \sim p(\psi)$$

Hierarchical models

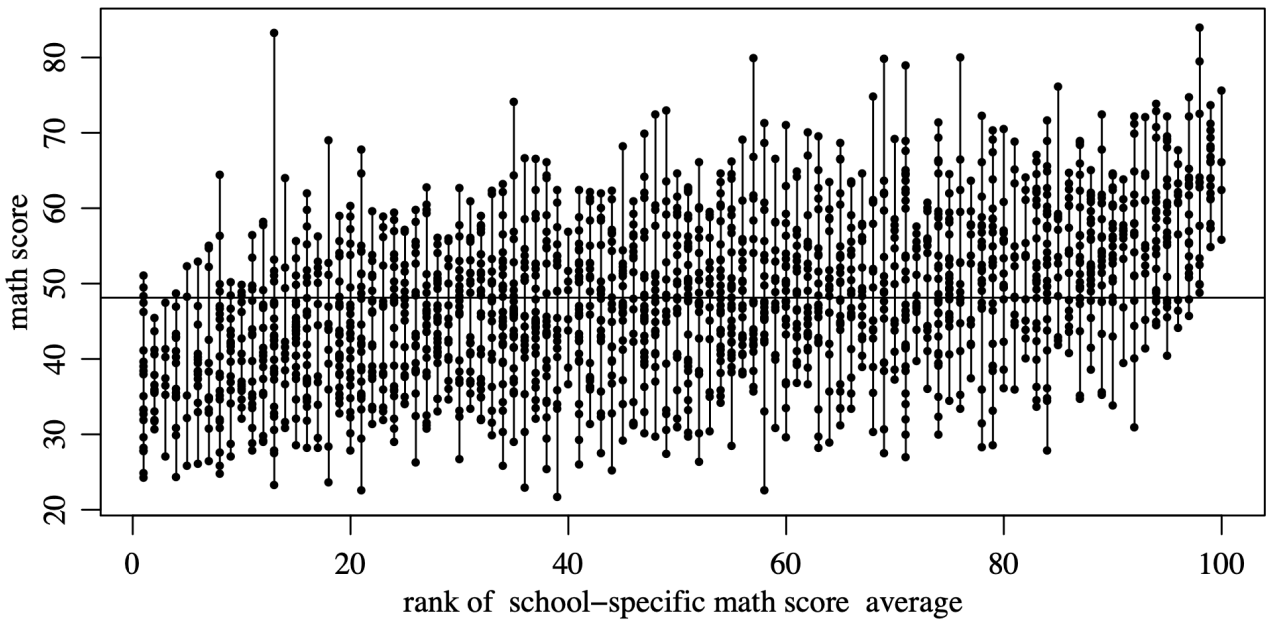
- Therefore our final model looks like:

$$\begin{cases} Y_{1,j}, Y_{2,j}, \ldots, Y_{n_j,j} | \phi_j \overset{iid}{\sim} p(y|\phi_j) \qquad j = 1, \ldots, m & (1) \\ \phi_1, \ldots, \phi_m | \psi \overset{iid}{\sim} p(\phi|\psi) & (2) \\ \psi \sim p(\psi) & (3) \end{cases}$$

- The first stage model, model (1), models the variability of the observations within a group; the second stage model, model (2), models the variability across groups; and the third stage model, model (3), provides the distribution function for the parameter ψ.
- The two distribution in (1) and (2) are sampling distributions while the distribution in (3) is a prior distribution.
- The prior distribution $p(\psi)$ might depend on other parameters, say γ_ψ. Those are called hyperparameters. We can provide a prior distribution for γ_ψ: this would be called a hyperprior.

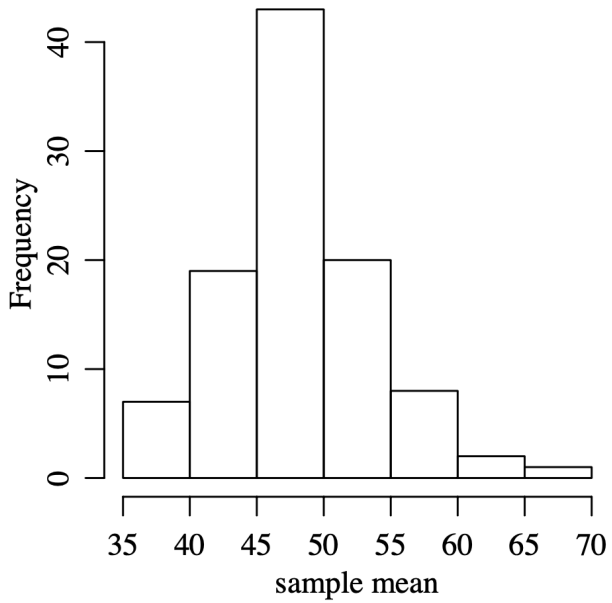
The hierarchical normal model

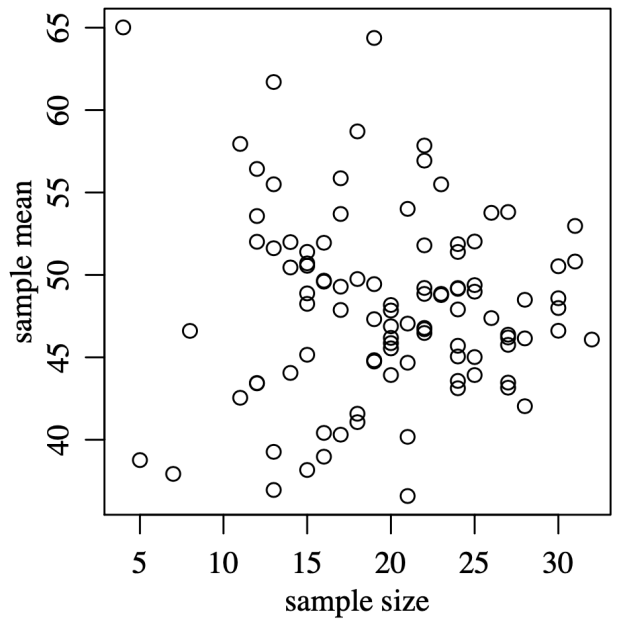
- Consider math scores data for 10-th grade children at 100 different large urban public schools, all having a 10-th grade enrollment of 400 or more students.
- The number of students taking the test in each school is not the same and the sample size varies from a minimum of 5 to a maximum of 32.
- Data from the schools are displayed below with scores for students within the same school plotted along the same common vertical bar.



The hierarchical normal model

- The figure belows shows:
  - a histogram of the sample average score at each school . The histogram appears normal with an average of about 50;
  - a scatterplot of the sample mean in every school versus the number of students for which we have data in the school. Very extreme average math scores are found in schools with small sample sizes. Schools with large(r) sample sizes tends to have average math scores closer to the overall average of 50.





The hierarchical normal model

- Based on this exploratory analysis, and using the following notation:

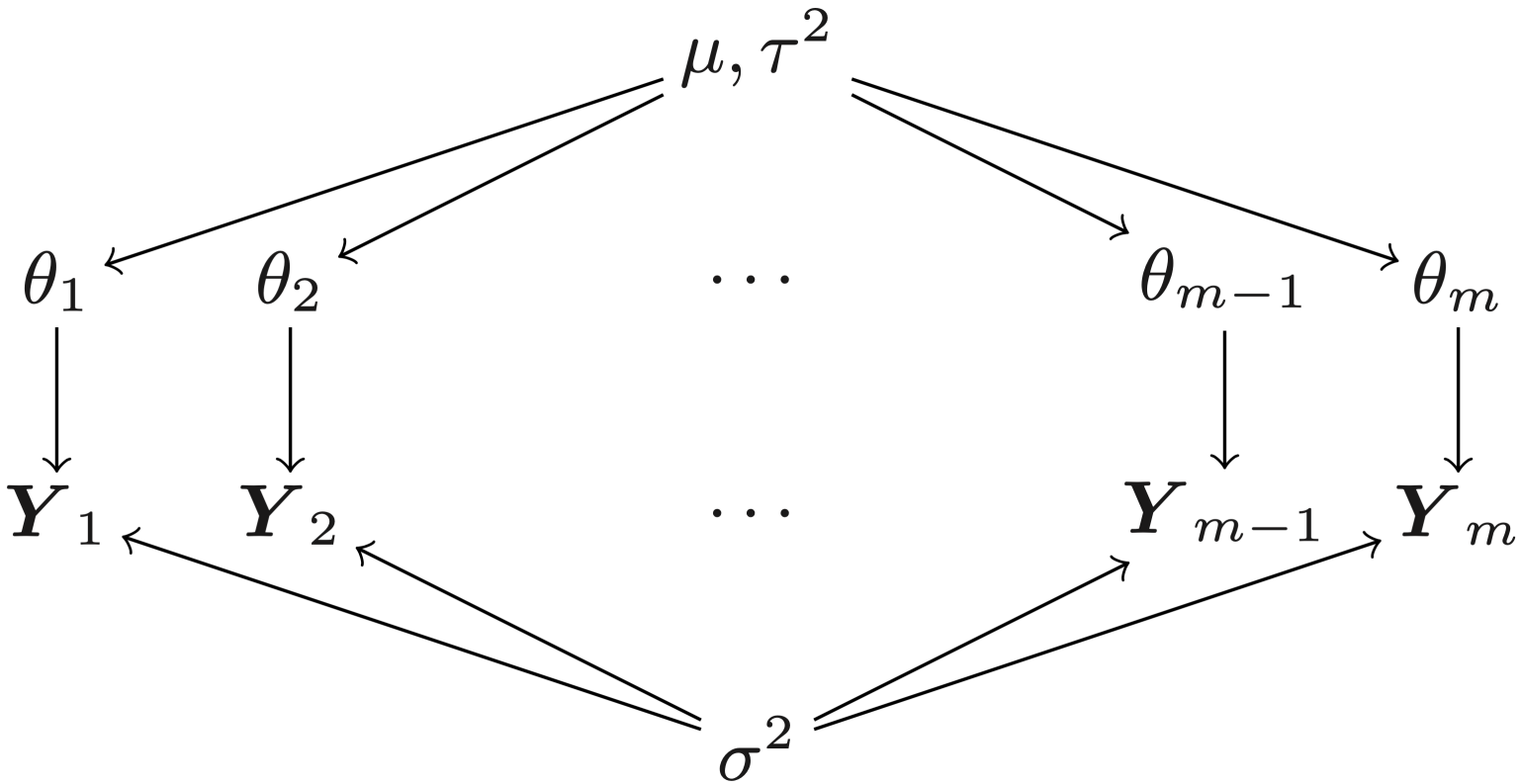
  - $y_{i,j}$: math score for student $i = 1, \ldots, n_j$ in school $j = 1, \ldots, m$
  - θ_j: average math score for school j
  - σ_j^2: variance in math scores in school j.
    [We assume that the variability in math scores in each school is the same. Thus, $\sigma_j^2 = \sigma^2$.]

  we model the data using the following model:

$$
\begin{aligned}
y_{1,j}, \ldots, y_{n_j,j} | \phi_j = (\theta_j, \sigma^2) &\overset{iid}{\sim} N(\theta_j, \sigma^2) \\
\theta_1, \theta_2, \ldots, \theta_m &\perp \sigma^2 \\
\theta_1, \theta_2, \ldots, \theta_m | \psi = (\mu, \tau^2) &\overset{iid}{\sim} N(\mu, \tau^2) \\
\sigma^2 &\sim \text{InverseGamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) \\
\mu &\sim N(\mu_0, \gamma_0^2) \\
\tau^2 &\sim \text{InverseGamma}\left(\frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right)
\end{aligned}
$$

The hierarchical normal model

- Below is a graphical representation of the hierarchical normal model for the math scores data.



The hierarchical normal model

- The unknowns in this model are: $\theta_1, \ldots, \theta_m, \sigma^2, \mu, \tau^2$.
- To infer upon these parameters within a Bayesian framework, we need to derive the joint posterior distribution $p(\theta_1, \ldots, \theta_m, \sigma^2, \mu, \tau^2 | \mathbf{y}_1, \ldots, \mathbf{y}_m)$ where $\mathbf{y}_j$ denotes the vector of n_j observations from school j.
- We approximate the posterior distribution $p(\theta_1, \ldots, \theta_m, \sigma^2, \mu, \tau^2 | \mathbf{y}_1, \ldots, \mathbf{y}_m)$ by devising a Gibbs sampling algorithm that produces a Markov chain that has the joint posterior distribution as its stationary distribution.
- To write the algorithm we need to derive:
  - the full conditional (distribution) of θ_j **given** $\mathbf{y}_1, \ldots, \mathbf{y}_m, \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_m, \sigma^2, \mu, \tau^2$ for $j = 1, \ldots, m$
  - the full conditional (distribution) of σ^2 **given** $\mathbf{y}_1, \ldots, \mathbf{y}_m, \theta_1, \ldots, \theta_m, \mu, \tau^2$
  - the full conditional (distribution) of μ **given** $\mathbf{y}_1, \ldots, \mathbf{y}_m, \theta_1, \ldots, \theta_m, \sigma^2, \tau^2$
  - the full conditional (distribution) of τ^2 **given** $\mathbf{y}_1, \ldots, \mathbf{y}_m, \theta_1, \ldots, \theta_m, \mu, \sigma^2$

The hierarchical normal model

- The joint posterior distribution is given by:

$$
\begin{aligned}
p(\theta_1,\ldots,\theta_m,\sigma^2,\mu,\tau^2|\mathbf{y}_1,\ldots,\mathbf{y}_m) &\propto p(\mathbf{y}_1,\ldots,\mathbf{y}_m|\theta_1,\ldots,\theta_m,\sigma^2,\mu,\tau^2) \\
&\cdot\ p(\theta_1,\ldots,\theta_m|\mu,\sigma^2,\tau^2)\cdot p(\mu,\sigma^2,\tau^2) \\
&= \left\{\prod_{j=1}^{m}\prod_{i=1}^{n_j} p(y_{ij}|\theta_j,\sigma^2)\right\} \\
&\cdot\ \left\{\prod_{j=1}^{m} p(\theta_j|\mu,\tau^2)\right\} \\
&\cdot\ p(\mu)\cdot p(\sigma^2)\cdot p(\tau^2)
\end{aligned}
$$

Full conditional distribution of θ_j

- The full conditional of θ_j is:

$$
\begin{aligned}
p(\theta_j|\mathbf{y}_1,\ldots,\mathbf{y}_m,\theta_1,\ldots,\theta_{j-1},\theta_{j+1},\ldots,\theta_m,\sigma^2,\mu,\tau^2) &\propto \left\{\prod_{i=1}^{n_j} p(y_{ij}|\theta_j,\sigma^2)\right\} \\
&\cdot\ p(\theta_j|\mu,\tau^2) \\
&= \left\{\prod_{i=1}^{n_j} N(y_{ij};\theta_j,\sigma^2)\right\} \\
&\cdot\ N(\theta_j;\mu,\tau^2)
\end{aligned}
$$

- Conditionally on $\mu,\tau^2,\sigma^2,\mathbf{y}_j$, θ_j is independent of the other θ_k's as well as independent of data from other groups. We can see this also in the graph relative to this hierarchical model.
- The full conditional of θ_j is $N(\mu_{n,j},\tau^2_{n,j})$ where

$$
\tau^2_{n,j} = \frac{1}{\frac{n_j}{\sigma^2}+\frac{1}{\tau^2}} \qquad \mu_{n,j} = \frac{\frac{n_j}{\sigma^2}\bar{y}_j + \frac{1}{\tau^2}\mu}{\frac{n_j}{\sigma^2}+\frac{1}{\tau^2}}
$$

Full conditional distribution of σ^2

- The full conditional of σ^2 is:

$$\begin{aligned} p(\sigma^2|\mathbf{y}_1,\ldots,\mathbf{y}_m,\theta_1,\ldots,\theta_m,\mu,\tau^2) &\propto \left\{\prod_{j=1}^m \prod_{i=1}^{n_j} p(y_{ij}|\theta_j,\sigma^2)\right\} \cdot p(\sigma^2) \\ &= \left\{\prod_{j=1}^m \prod_{i=1}^{n_j} N(y_{ij};\theta_j,\sigma^2)\right\} \\ &\cdot \text{InverseGamma}(\sigma^2;\tfrac{\nu_0}{2},\tfrac{\nu_0\sigma_0^2}{2}) \end{aligned}$$

- The full conditional of σ^2 is an $\text{InverseGamma}(\frac{\nu_n}{2},\frac{\nu_n\sigma_n^2}{2})$ where

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

Full conditional distribution of μ

- The full conditional of μ is:

$$\begin{aligned} p(\mu|\mathbf{y}_1,\ldots,\mathbf{y}_m,\theta_1,\ldots,\theta_m,\sigma^2,\tau^2) &\propto \left\{\prod_{j=1}^m p(\theta_j|\mu,\tau^2)\right\}\cdot p(\mu) \\ &= \left\{\prod_{j=1}^m N(\theta_j;\mu,\tau^2)\right\}\cdot N(\mu;\mu_0,\gamma_0^2) \end{aligned}$$

- The full conditional of μ is $N(m_n,\gamma_n^2)$ where

$$\gamma_n^2 = \frac{1}{\frac{m}{\tau^2}+\frac{1}{\gamma_0^2}} \qquad m_n = \frac{\frac{m}{\tau^2}\bar{\theta}+\frac{1}{\gamma_0^2}\mu_0}{\frac{m}{\tau^2}+\frac{1}{\gamma_0^2}}$$

Full conditional distribution of τ^2

- The full conditional of τ^2 is:

$$\begin{aligned} p(\tau^2|\mathbf{y}_1,\ldots,\mathbf{y}_m,\theta_1,\ldots,\theta_m,\sigma^2,\mu,\tau^2) &\propto \left\{\prod_{j=1}^m p(\theta_j|\mu,\tau^2)\right\}\cdot p(\tau^2) \\ &= \left\{\prod_{j=1}^m N(\theta_j;\mu,\tau^2)\right\} \\ &\cdot \text{InverseGamma}(\tau^2;\tfrac{\eta_0}{2},\tfrac{\eta_0\tau_0^2}{2}) \end{aligned}$$

- The full conditional of τ^2 is $\text{InverseGamma}(\frac{\eta_n}{2},\frac{\eta_n\tau_n^2}{2})$ where

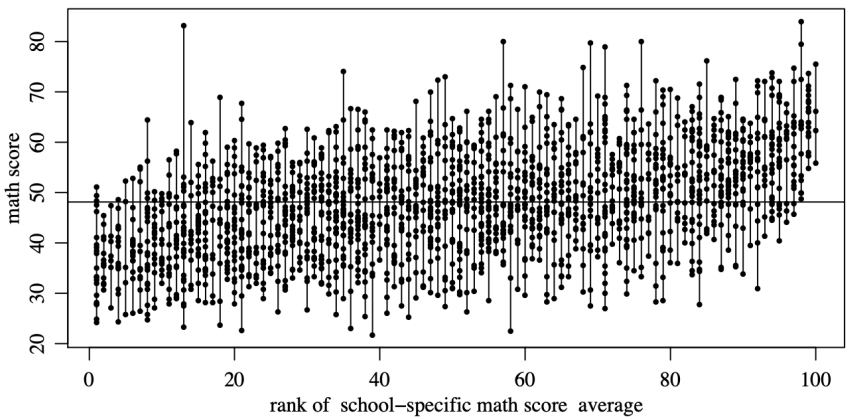
$$\eta_n = m+\eta_0 \qquad \tau_n^2 = \frac{1}{\eta_n}\left[\eta_0\tau_0^2+\sum_{j=1}^m(\theta_j-\mu)^2\right]$$

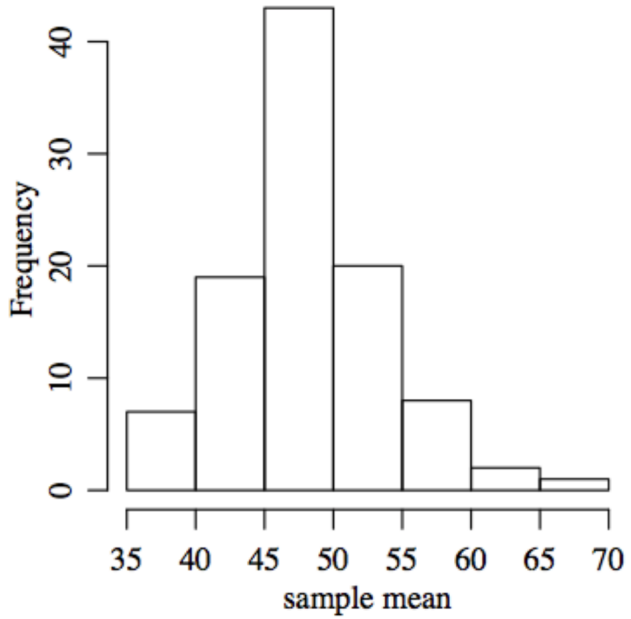
Gibbs sampling algorithm

- The Gibbs sampling algorithm for the basic hierarchical normal model then will proceed as follows:
  - Choose a number S of iterations.
  - Decide starting values $\theta_j^{(0)}$, $j = 1, \ldots, m$, $\sigma^{2(0)}$, $\mu^{(0)}$ and $\tau^{2(0)}$
  - For each iteration $k = 1, \ldots, S$ repeat the following:
    1. sample a new value $\mu^{(k+1)}$ from the full conditional $p(\mu|\theta_1^{(k)}, \ldots, \theta_m^{(k)}, \tau^{2(k)})$
    2. sample a new value $\tau_2^{(k+1)}$ from the full conditional $p(\tau^2|\theta_1^{(k)}, \ldots, \theta_m^{(k)}, \mu^{(k+1)})$
    3. sample a new value $\sigma^{2(k+1)}$ from the full conditional $p(\sigma^2|\mathbf{y}_1, \ldots, \mathbf{y}_m, \theta_1^{(k)}, \ldots, \theta_m^{(k)})$
    4. sample new values $\theta_j^{(k+1)}$ from the full conditionals $p(\theta_j|\mathbf{y}_j, \mu^{(k+1)}, \sigma^{2(k+1)}, \tau^{2(k+1)})$, $j = 1, \ldots, m$
- The order in which the parameters are updated is not important. What is important is that they are generated from the full conditional with the most recent values of the other parameters.

Hierarchical normal model: math scores data

- Let's see an application to the math scores data. Remember the data





- Here we have $m = 100$ groups and different sample sizes n_j for each school.

Math scores data

- We analyze the math scores data using the following priors:

$$\begin{aligned}
\mu &\sim p(\mu) = N(\mu_0 = 50, \gamma_0^2 = 25) \\
\sigma^2 &\sim p(\sigma^2) = \text{InverseGamma}(\tfrac{\nu_0}{2} = \tfrac{1}{2}, \tfrac{\nu_0\sigma_0^2}{2} = \tfrac{1 \cdot 100}{2}) \\
\tau^2 &\sim p(\tau^2) = \text{InverseGamma}(\tfrac{\eta_0}{2} = \tfrac{1}{2}, \tfrac{\eta_0\tau_0^2}{2} = \tfrac{1 \cdot 100}{2})
\end{aligned}$$

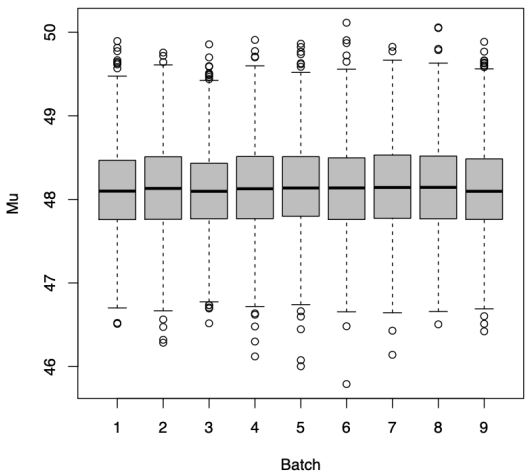
- The math test was designed to have a nationwide variance of 100, so we can set the prior for σ^2 to be centered around 100 and be very vague (this is why we choose $\nu_0 = 1$).
- The nationwide average is 50, so we can set the prior for μ to be centered around 50 with a variance of 25 if we believe that a variance of 100 is an overestimate.
- The variation among schools cannot be larger than the nationwide variance, so we set the prior for τ^2 to be centered around 100 and we take it to be very vague (thus, $\eta_0 = 1$).

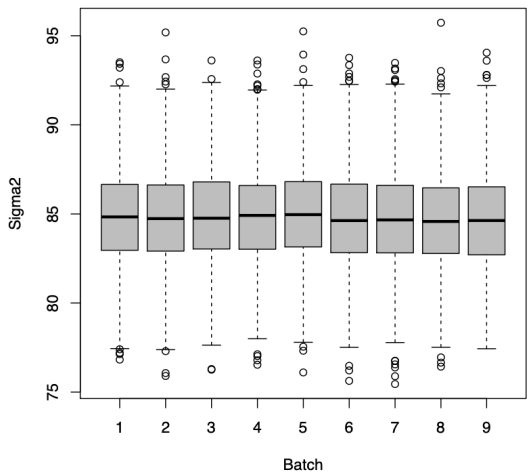
Math scores data

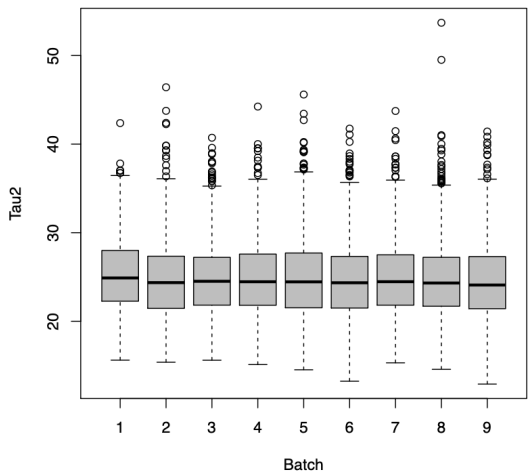
- We ran the Gibbs sampling algorithm for 10,000 iterations using as initial values for
  - θ_j, $j = 1, \ldots, m = 100$ the sample average within each school
  - σ^2, the average of the sample variance for the scores within each school
  - μ, the average of all the school-specific sample averages
  - τ^2, the sample variance among all the school-specific sample averages
- We ran a single chain but we inspected if stationarity was achieved by making boxplots of the MCMC samples for all the parameters every $L = 1000$ iterations after burnin (we chose a burnin of 1000). If stationarity is achieved, the distribution of the MCMC samples when we batch in groups made of L consecutive samples should be approximately the same.

Math scores data: inspecting stationarity

- Box plots of the MCMC samples for μ, σ^2, τ^2 for iterations 1001-10000 grouped in batches of 1000.

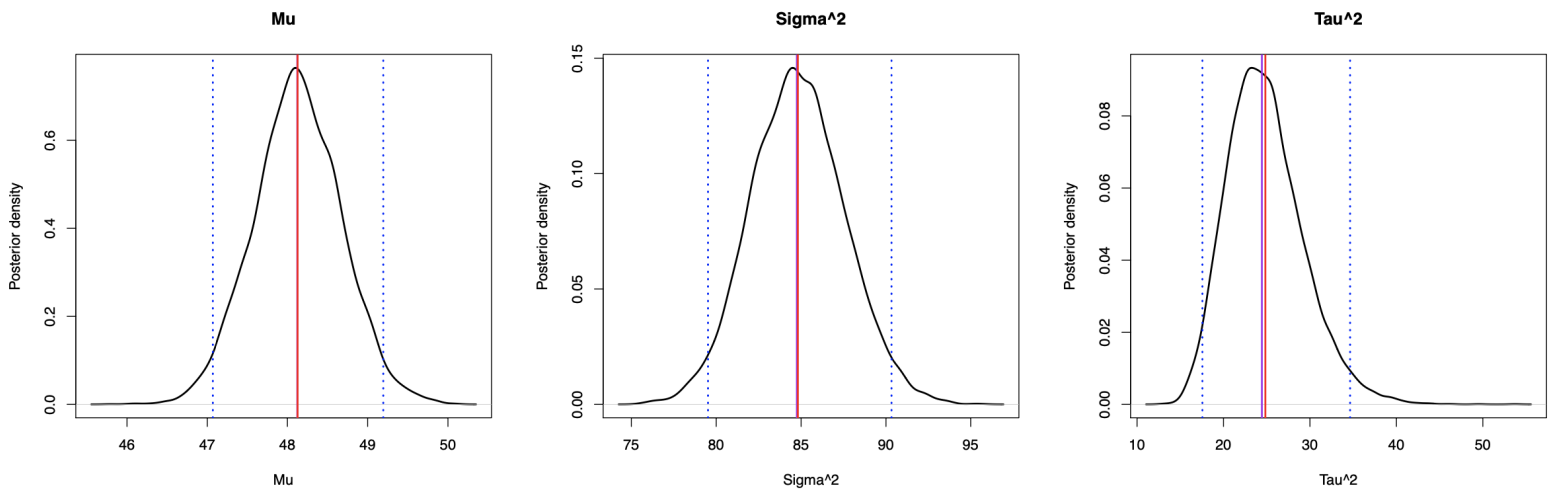






Math scores data: results

- After having performed all the necessary convergence diagnostics (autocorrelation function, effective sample size, trace plots, stationarity check and Gelman-Rubin's convergence criteria R if more than one chain has been run), we can summarize the posterior distribution for each parameter.
- Marginal posterior distribution for μ, σ^2, τ^2 with indicated the posterior mean, posterior median and the 2.5% and 97.5% posterior quantiles.



Math scores data: results

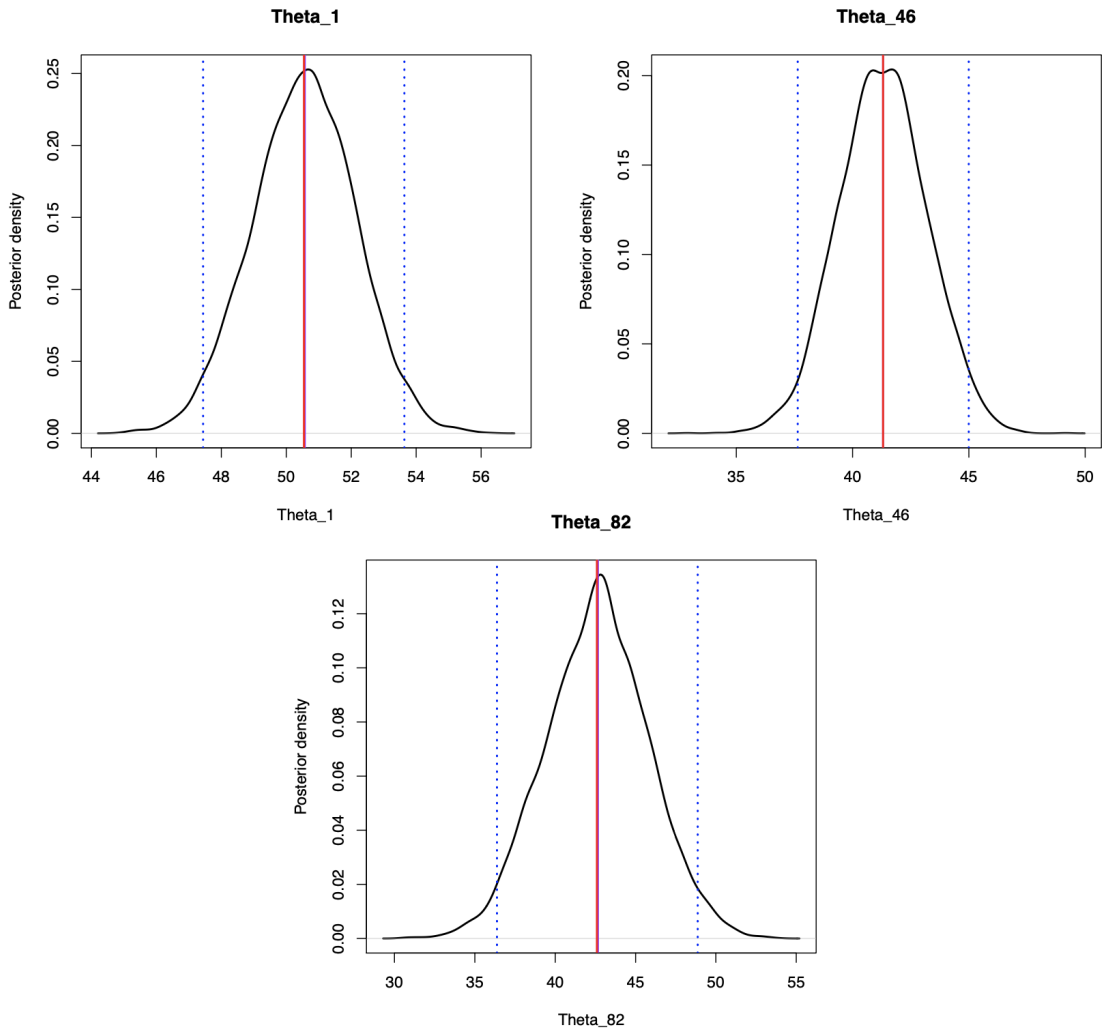
- For μ, σ^2, τ^2 we have the following posterior summaries:

Parameter	Posterior mean	Posterior SD	95% credible interval
μ	48.12	0.54	(47.1,49.2)
σ^2	84.8	2.7	(79.5,90.3)
σ	9.21	0.15	(8.9,9.5)
τ^2	24.8	4.4	(17.6,34.6)
τ	4.97	0.43	(4.2,5.9)

- This means that the overall average math scores in these 100 schools is estimated to be around 48.12. Ninety-five percent of the scores within a school are expected to be within $2 \times 1.96 \times 9.21 \approx 37$ points of each other.
- On the other hand, ninety-five percent of the average school scores are within $2 \times 1.96 \times 4.97 \approx 20$ points of each other.

Math scores data: results

- What about the school-specific averages θ_j?
  We can look at the marginal posterior distribution for some of the θ_j. Below are some.



Math scores data: shrinkage

- From the form of the full conditional for θ_j, we know that **conditional** on $\mathbf{y}_j, \mu, \sigma^2, \tau^2$

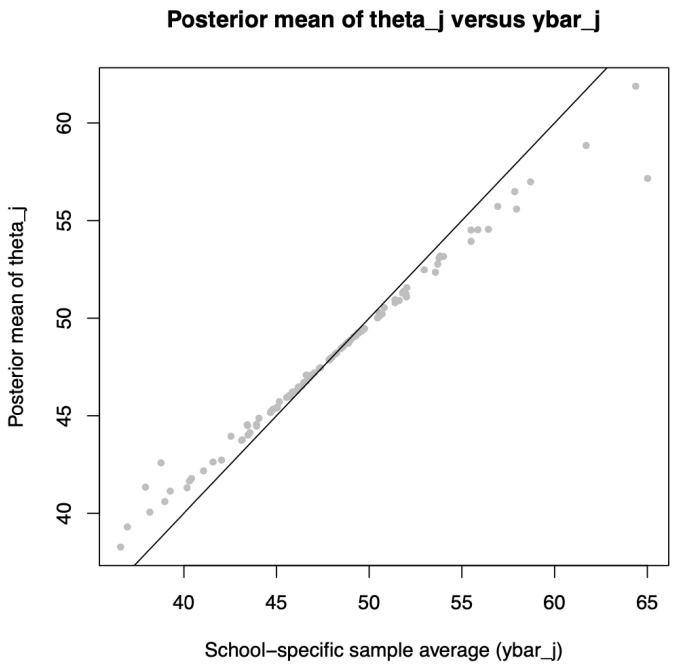
$$E\left[\theta_j | \mathbf{y}_j, \mu, \sigma^2, \tau^2\right] = \frac{\frac{n_j}{\sigma^2}\bar{y}_j + \frac{\mu}{\tau^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$$

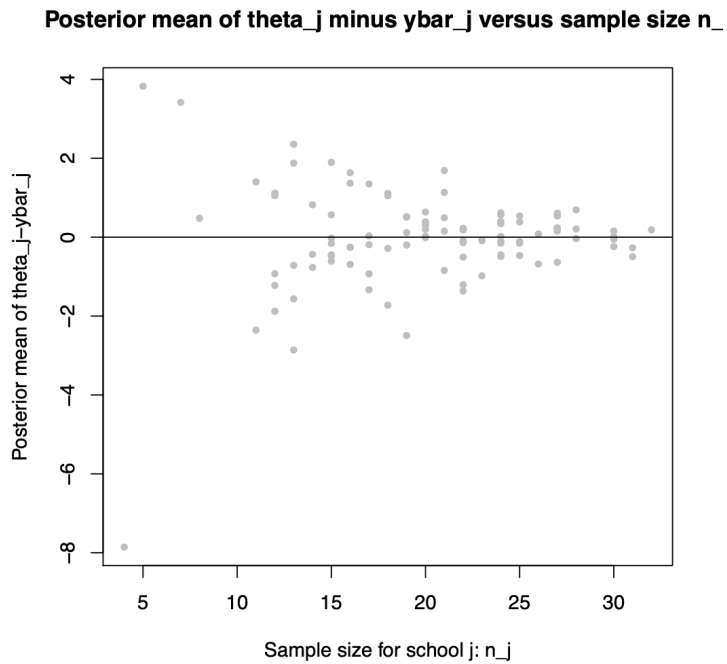
  This implies that the expected value of θ_j is pulled from the school-specific sample average $\bar{y}_j$ towards the overall average μ by an amount that depends on n_j. This is called shrinkage.

- If n_j is smaller the amount of the shrinkage is larger; if n_j is larger, the shrinkage is smaller.
- This says that if there is a lot of data from group j, we do not have to **borrow** information from the rest of the population to make inference about θ_j.
  On the other hand, if we don't have a lot of data in group j, then we can **borrow** information from the rest of the population.

Math scores data: shrinkage

- We can see the effect of shrinkage in the following two plots:
  - a scatterplot of a Bayesian estimate $\hat{\theta}_j$ of θ_j (marginal posterior mean) versus the corresponding school-specific sample average $\bar{y}_j$: the farther the dots are from the 45° degree line, the stronger the shrinkage
  - a scatterplot of the a measure of the shrinkage, $\hat{\theta}_j - \bar{y}_j$ versus the sample size n_j for school j

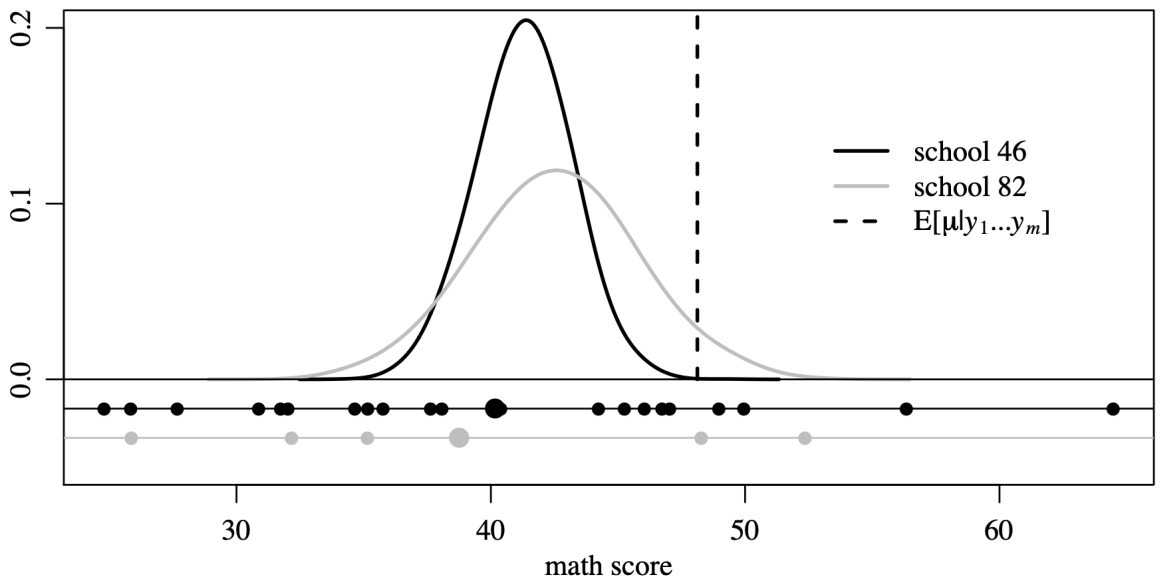




Math scores data: shrinkage

- Suppose that one of our goals is to rank the schools based on what we think the performance of each school would be if all the students had taken the math exam.
- We could rank the schools based on the school-specific sample averages $\bar{y}_1, \ldots, \bar{y}_{100}$ or we could rank based on the posterior means $E(\theta_1|\mathbf{y}_1, \ldots, \mathbf{y}_{100}), \ldots, E(\theta_{100}|\mathbf{y}_1, \ldots, \mathbf{y}_{100})$.
- The two methods produce similar rankings but they are not the same. Why is that?
- As an example, let's look at two schools: school 46 and school 82. Both of these schools were ranked in the bottom 10% of all the schools.

Math scores data: shrinkage



- The plots shows the marginal posterior densities for θ_{46} and θ_{82} along with the raw data on the bottom (dots).
- The larger dots indicate the sample averages $\bar{y}_{46}$ and $\bar{y}_{82}$ for the two schools.
- The marginal posterior density for θ_{46} is more peaked than that for θ_{82}: we have data on 21 students in school 46, while we have only data on 5 students at school 82.
- Note that $\bar{y}_{82} < \bar{y}_{46}$ **BUT** $E[\theta_{46}|\text{data}] < E[\theta_{82}|\text{data}]$. Why is that?

Math scores data: shrinkage

- Does it make sense that the posterior mean for θ_{82} is greater than that for θ_{46}?
- The sample average of school 46 is based on 21 students, while that for school 82 is based on 5 students.
  Therefore, when we estimate θ_{82} we are going to shrink it more towards the population average than in the case of θ_{46}.
- In ranking schools, the hierarchical model takes into account the fact that while there is evidence that the average performance of students in school 46 is exceptionally low, there is not enough evidence for school 82.
  Accounting for that, the hierarchical model would rank school 46 below school 82 differently from the ranking we would have obtained had we used the sample averages.
- See the interesting analogy the book makes with basketball (Section 8.4).

Math scores data

- We have seen in the previous section an analysis of math test scores data collected at 100 schools using a Bayesian hierarchical model.
- The model was as follows

$$
\begin{aligned}
y_{1,j},\dots,y_{n_j,j}|\theta_j,\sigma^2 &\overset{iid}{\sim} N(\theta_j,\sigma^2) \qquad j=1,\dots,m\\
\theta_1,\dots,\theta_m &\perp \sigma^2\\
\theta_1,\dots,\theta_m|\mu,\tau^2 &\overset{iid}{\sim} N(\mu,\tau^2)\\
\sigma^2 &\sim \text{InverseGamma}(\tfrac{\nu_0}{2},\tfrac{\nu_0\sigma_0^2}{2})\\
\mu &\sim N(\mu_0,\gamma_0^2)\\
\tau^2 &\sim \text{InverseGamma}(\tfrac{\eta_0}{2},\tfrac{\eta_0\tau_0^2}{2})
\end{aligned}
$$

  where $y_{1,j},\dots,y_{n_j,j}$ indicate the math scores for students $1,\dots,n_j$ in school j.
- The Deviance Information Criteria (DIC) for this model was 14589.3 (see lecture 7 pp. 62-63).
- One of the main assumption of the model was that the variability σ_j^2 within schools was the same and equal to σ^2.

Math scores data: a second hierarchical model

- One might believe that the variability in the students scores changes from school to school.
- For this reason, we now consider the model

$$
\begin{aligned}
y_{1,j}, \ldots, y_{n_j,j} | \theta_j, \sigma_j^2 &\overset{iid}{\sim} N(\theta_j, \sigma_j^2) && j = 1, \ldots, m \\
\theta_j &\perp \sigma_j^2 && j = 1, \ldots, m \\
\theta_1, \ldots, \theta_m | \mu, \tau^2 &\overset{iid}{\sim} N(\mu, \tau^2) \\
\sigma_1^2, \ldots, \sigma_m^2 &\overset{iid}{\sim} \text{InverseGamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) \\
\mu &\sim N(\mu_0, \gamma_0^2) \\
\tau^2 &\sim \text{InverseGamma}\left(\frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right)
\end{aligned}
$$

- The parts that are different between the two models are highlighted in red.
- In this model, if we take ν_0 and σ_0^2, the role played by $\theta_1, \ldots, \theta_m$ and $\sigma_1^2 \ldots, \sigma_m^2$ is different: the first are additional variables for which we are providing a sampling model, while the second are m independent parameters for which we are providing a prior.

Math scores data: a second hierarchical model

- In other words, if ν_0 and σ_0^2 are fixed, then knowing σ_j^2 provides no information on any σ_k^2 for $k \neq j$.
- That would not be the case if instead we assume the following

$$\begin{aligned} \sigma_1^2, \ldots, \sigma_m^2 | \nu_0, \sigma^2 &\overset{iid}{\sim} \text{InverseGamma}(\tfrac{\nu_0}{2}, \tfrac{\nu_0 \sigma_0^2}{2}) \\ \nu_0 &\sim p(\nu_0) \\ \sigma_0^2 &\sim p(\sigma_0^2) \end{aligned}$$

- We can use a Gamma(a,b) prior on σ_0^2.
- Finding a prior on ν_0 is more difficult, as no distribution will be conjugate.
  We choose to use a discrete prior. The book uses a geometric prior $p(\nu_0) \propto \exp(-\alpha \nu_0)$ on the set of integers $\{1, 2, \ldots, \}$.
  For simplicity, we may use a discrete uniform distribution $p(\nu_0) = \frac{1}{30}$ on the set $\{1, 2, \ldots, 30\}$.

Math scores data: a second hierarchical model

- Thus, the final model that we consider is

$$
\begin{aligned}
y_{1,j}, \ldots, y_{n_j,j} | \theta_j, \sigma_j^2 &\overset{iid}{\sim} N(\theta_j, \sigma_j^2) && j = 1, \ldots, m \\
\theta_j &\perp \sigma_j^2 && j = 1, \ldots, m \\
\theta_1, \ldots, \theta_m | \mu, \tau^2 &\overset{iid}{\sim} N(\mu, \tau^2) \\
\sigma_1^2, \ldots, \sigma_m^2 | \nu_0, \sigma_0^2 &\overset{iid}{\sim} \text{InverseGamma}\left(\tfrac{\nu_0}{2}, \tfrac{\nu_0 \sigma_0^2}{2}\right) \\
\mu &\sim N(\mu_0, \gamma_0^2) \\
\tau^2 &\sim \text{InverseGamma}\left(\tfrac{\eta_0}{2}, \tfrac{\eta_0 \tau_0^2}{2}\right) \\
\sigma_0^2 &\sim \text{Gamma}(a, b) \\
\nu_0 &\sim p(\nu_0)
\end{aligned}
$$

- We will compare the basic hierarchical normal model with common group variance σ^2 seen on pp. 11-33 with this one.
- We will determine which model is more appropriate for the data by looking at their DIC values. The model with the lowest DIC is better.

Posterior inference

- To perform posterior inference for this new model, we need to derive the joint posterior distribution $p(\theta_1, \ldots, \theta_m, \sigma_1^2, \ldots, \sigma_m^2, \mu, \tau^2, \nu_0, \sigma_0^2 | \mathbf{y}_1, \ldots, \mathbf{y}_m)$.
- This is now given by:

$$\begin{aligned} p(\theta_1, \ldots, \theta_m, \sigma_1^2, \ldots, \sigma_m^2, \mu, \tau^2, \nu_0, \sigma_0^2 | \mathbf{y}_1, \ldots, \mathbf{y}_m) \propto & \left\{ \prod_{j=1}^m \prod_{i=1}^{n_j} p(y_{ij} | \theta_j, \sigma_j^2) \right\} \\ & \cdot \left\{ \prod_{j=1}^m p(\theta_j | \mu, \tau^2) \right\} \\ & \cdot \left\{ \prod_{j=1}^m p(\sigma_j^2 | \nu_0, \sigma_0^2) \right\} \\ & \cdot p(\mu) \cdot p(\tau^2) \cdot p(\nu_0) \cdot p(\sigma_0^2) \end{aligned}$$

- As before, we approximate the posterior distribution using an MCMC algorithm, specifically a Gibbs sampling algorithm.

Full conditional of θ_j

- From the form of the joint posterior $p(\theta_1, \ldots, \theta_m, \sigma_1^2, \ldots, \sigma_m^2, \mu, \tau^2, \nu_0, \sigma_0^2 | \mathbf{y}_1, \ldots, \mathbf{y}_m)$, it is clear that the full conditionals for μ and τ^2 remain the same as in the previous model, those of θ_j and σ_j^2 change and we now have to derive the full conditionals for ν_0 and σ_0^2.
- The full conditional of θ_j is:

$$\begin{aligned} p(\theta_j | \mathbf{y}_1, \ldots, \mathbf{y}_m, \theta_{-j}, \sigma_1^2, \ldots, \sigma_m^2, \mu, \tau^2, \nu_0, \sigma_0^2) &\propto \left\{ \prod_{i=1}^{n_j} p(y_{ij} | \theta_j, \sigma_j^2) \right\} \\ &\cdot \; p(\theta_j | \mu, \tau^2) \\ &= \left\{ \prod_{i=1}^{n_j} N(y_{ij}; \theta_j, \sigma_j^2) \right\} \\ &\cdot \; N(\theta_j; \mu, \tau^2) \end{aligned}$$

- Therefore, the full conditional of θ_j is $N(\mu_{n,j}, \tau_{n,j}^2)$ where

$$\tau_{n,j}^2 = \frac{1}{\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2}} \qquad \mu_{n,j} = \frac{\frac{n_j}{\sigma_j^2} \bar{y}_j + \frac{\mu}{\tau^2}}{\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2}}$$

Full conditional of σ_j^2

- The full conditional of σ_j^2 is:

$$\begin{aligned} p(\sigma_j^2|\mathbf{y}_1,\ldots,\mathbf{y}_m,\theta_1,\ldots,\theta_m,\boldsymbol{\sigma}^2{}_{-j},\mu,\tau^2,\nu_0,\sigma_0^2) &\propto \left\{\prod_{i=1}^{n_j} p(y_{ij}|\theta_j,\sigma_j^2)\right\} \\ &\cdot\ p(\sigma_j^2|\nu_0,\sigma_0^2) \\ &= \left\{\prod_{i=1}^{n_j} N(y_{ij};\theta_j,\sigma_j^2)\right\} \\ &\cdot\ \text{InvGamma}(\sigma_j^2;\tfrac{\nu_0}{2},\tfrac{\nu_0\sigma_0^2}{2}) \end{aligned}$$

- Therefore, the full conditional of σ_j^2 is InverseGamma$(\frac{\nu_{n,j}}{2},\frac{\nu_{n,j}\sigma_{n,j}^2}{2})$ where

$$\nu_{n,j}^2 = n_j + \nu_0 \qquad \sigma_{n,j}^2 = \frac{1}{\nu_{n,j}}\left[\nu_0\sigma_0^2 + \sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2\right]$$

- In this model, conditionally on $\theta_j,\mathbf{y}_j,\nu_0,\sigma_0^2$, σ_j^2 is independent of the other σ_k^2 as well as data from other groups.

Full conditional of σ_0^2

- The full conditional of σ_0^2 is:

$$\begin{aligned} p(\sigma_0^2|\mathbf{y}_1,\ldots,\mathbf{y}_m,\theta_1,\ldots,\theta_m,\sigma_1^2,\ldots,\sigma_m^2,\mu,\tau^2,\nu_0) &\propto \left\{\prod_{j=1}^m p(\sigma_j^2|\nu_0,\sigma_0^2)\right\} \\ &\cdot\ p(\sigma_0^2) \\ &= \left\{\prod_{j=1}^m \text{IG}(\sigma_j^2;\tfrac{\nu_0}{2},\tfrac{\nu_0\sigma_0^2}{2})\right\} \\ &\cdot\ \text{Gamma}(a,b) \end{aligned}$$

- Therefore, the full conditional of σ_0^2 is $\text{Gamma}(a_n,b_n)$ where

$$a_n = a + m\frac{\nu_0}{2} \qquad b_n = b + \frac{1}{2}\sum_{j=1}^m \frac{\nu_0}{\sigma_j^2}$$

Full conditional of ν_0

- The full conditional of ν_0 is given by:

$$\begin{aligned} p(\nu_0|\mathbf{y}_1,\ldots,\mathbf{y}_m,\theta_1,\ldots,\theta_m,\sigma_1^2,\ldots,\sigma_m^2,\mu,\tau^2,\sigma_0^2) &\propto \left\{\prod_{j=1}^m p(\sigma_j^2|\nu_0,\sigma_0^2)\right\} \\ &\cdot\ p(\nu_0) \\ &= \left\{\prod_{j=1}^m \mathrm{IG}(\sigma_j^2;\tfrac{\nu_0}{2},\tfrac{\nu_0\sigma_0^2}{2})\right\} \\ &\cdot\ p(\nu_0) \end{aligned}$$

- For both choices of $p(\nu_0)$ [Geometric as in the textbook or discrete] this distribution does not have a closed form.
  However, since we took the sample space for ν_0 to be a discrete set $\{1,2,\ldots,\} = \{\nu_k, k=1,\ldots\}$, we can evaluate the full conditional of ν_0 at those values $\{\nu_k, k=1,\ldots,\}$ and sample from the set $\{1,2,\ldots,\} = \{\nu_k, k=1,\ldots\}$ according to the probabilities

$$\left\{\prod_{j=1}^m \mathrm{IG}(\sigma_j^2;\frac{\nu_0}{2}=\frac{\nu_k}{2},\frac{\nu_0\sigma_0^2}{2}=\frac{\nu_k\sigma_0^2}{2})\right\}\cdot p(\nu_0=\nu_k)$$

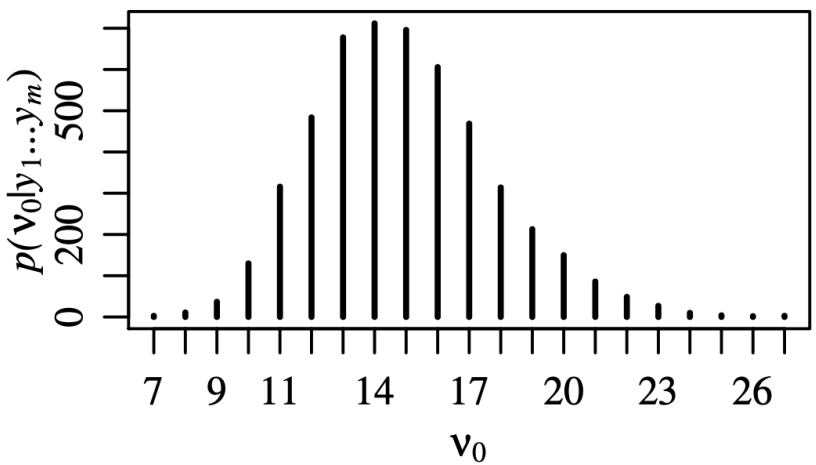
  appropriately normalized.

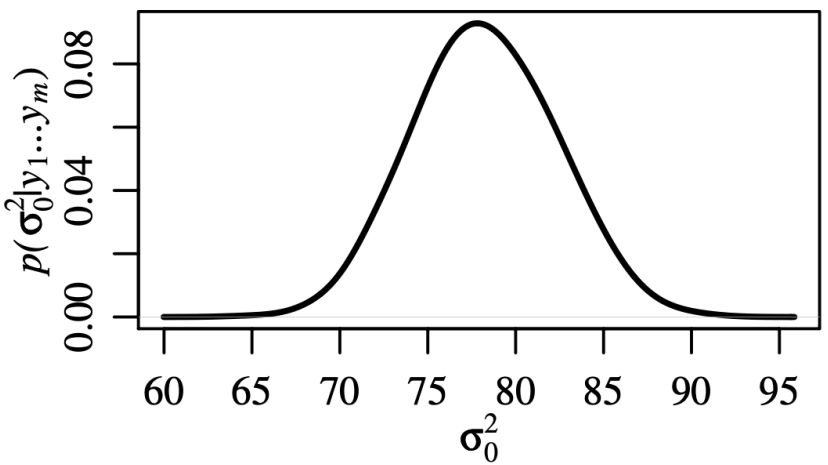
Math scores data: results for new model

- We ran the Gibbs sampling algorithm for this new model for $S = 5{,}000$ iterations using as parameters for the prior of σ_0^2 and ν_0, respectively, $a = 1$ and $b = 100$, and $\alpha = 1$ in the geometric distribution.
- We used the same initial values as in the previous model, except that now the within group variances σ_j^2 are set equal to the sample variances within each school.
- As initial values for ν_0 we used 1, while for σ_0^2 we used the average of all the school-specific sample variances.
- Note that when $\nu_0 \to +\infty$ the current model is equivalent to the previous model where all the group-specific variances were constrained to be equal.
  On the other hand $\nu_0 = 1$ indicates that all group-specific variances are unequal and there is no sharing of information about variances across groups.

Math scores data: results for new model

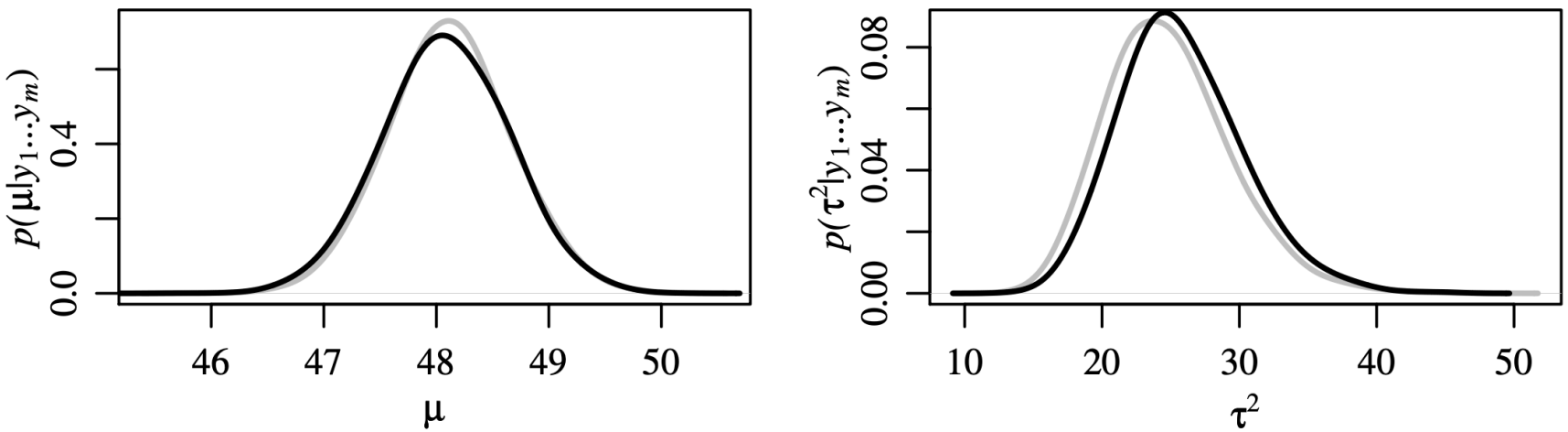
- The situation here is somewhat in between since the marginal posterior density for ν_0 is centered around small values, but ν_0 is different from 1.
- The marginal posterior densities of the two parameters that control the heterogeneity across groups in the variance, that is, ν_0, σ_0^2, are provided below.





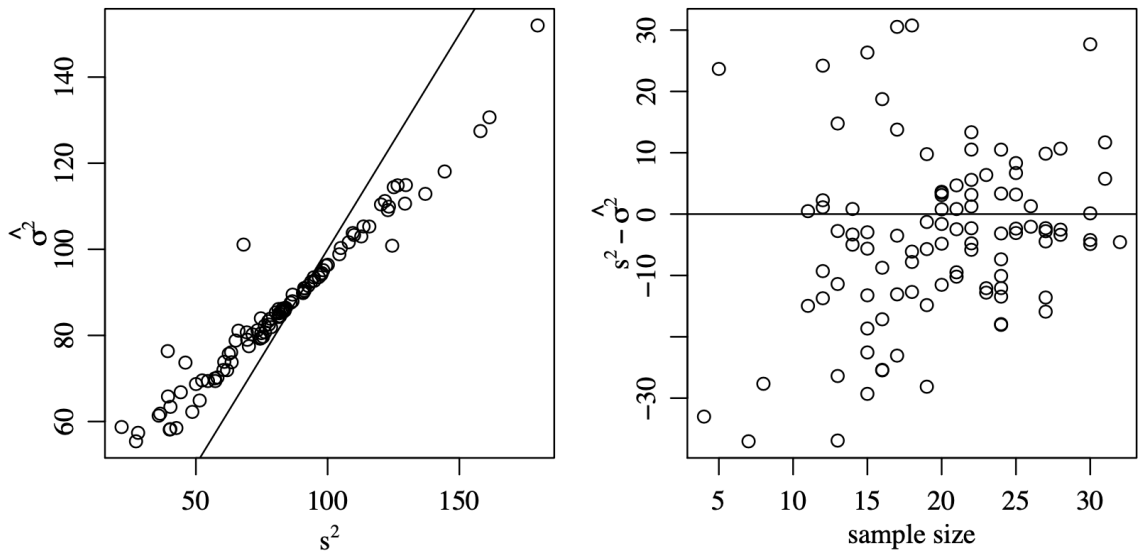
Math scores data: results for new model

- The marginal posterior densities of the two parameters that control the heterogeneity across groups in the average, that is, μ, τ^2, are provided below: the ones obtained under the current model are in black, while the densities obtained under the previous model are in grey.



Math scores data: shrinkage for the variances

- Note that under the new model we have shrinkage towards an overall value not only for the θ_j but also for the σ_j^2.
- As before, the variances σ_j^2 relative to schools with smaller sample size n_j experience a stronger shrinkage towards an overall estimate of the variance, s_0^2, than those relative to schools with larger sample sizes.
- The two plots below illustrate the shrinkage for the variances σ_j^2, $j = 1, \ldots, m$.



Math scores data: DIC comparison

- We compare the two models by looking at their Deviance Information Criteria values: the model with the lowest DIC is better.
- For the first model with common variance across schools, the DIC was equal to: 14589.3 ($p_D = 85.763$, $\bar{D} = 14503.6$ and $\hat{D} = 14417.8$)
- For the second model (where we used a uniform discrete prior on ν_0 on the set $\{1, 2, \ldots, 30\}$), the DIC was equal to: 14647.4 ($p_D = 183.755$, $\bar{D} = 14463.7$ and $\hat{D} = 14279.9$)
- Thus, we conclude that based on the data that we have observed, the hierarchical model with equal variance across school provides a better fit.