COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

The Hierarchical Normal Model

Recall the two-level hierarchical normal data model with $\phi_j = (\theta_j, \sigma^2)$ and $\psi = (\mu, \tau^2)$

$$X_{1j}, \ldots, X_{n_j j} | \phi_j \sim N(\theta_j, \sigma^2)$$
$$\theta_j | \psi \sim N(\mu, \tau^2)$$
$$\psi \sim \pi(\psi)$$

Meaning of these parameters:

θ_j is the average performance of all students withing school j (aka random effects)

σ^2 measures the within-cluster variability (ie, student level)

μ is the average performance of all students in all schools (aka fixed effect)

τ^2 measures the between-cluster variability (ie, school level)

Priors

Two-level hierarchical normal data model with $\phi_j = (\theta_j, \sigma^2)$ and $\psi = (\mu, \tau^2)$ is:

$$X_{1j}, \ldots, X_{n_j j} | \phi_j \sim N(\theta_j, \sigma^2)$$

$$\theta_j | \psi \sim N(\mu, \tau^2)$$

$$\psi \sim \pi(\psi)$$

We will use the following semiconjugate priors:

$$\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

$$\tau^2 \sim IG\left(\frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right)$$

$$\mu \sim N(\mu_0, \gamma_0^2)$$

Conditional posteriors of μ and τ^2

The part of the joint posterior that depends on μ and τ^2 is

$$\pi(\mu)\pi(\tau^2)\left[\prod_{j=1}^{m}\pi\left(\theta_j|\mu,\tau^2\right)\right]$$

This means that:

$$f(\mu|\theta_1,\dots,\theta_m,\tau^2,\sigma^2,\boldsymbol{x}_1,\dots,\boldsymbol{x}_m)$$

$$\propto \pi(\mu)\prod_{j=1}^{m}\pi\left(\theta_j|\mu,\tau^2\right) \sim N\left(\frac{\frac{m\bar{\theta}}{\tau^2}+\frac{\mu_0}{\gamma_0^2}}{\frac{m}{\tau^2}+\frac{1}{\gamma_0^2}}, \frac{1}{\frac{m}{\tau^2}+\frac{1}{\gamma_0^2}}\right)$$

$$f(\tau^2|\theta_1,\dots,\theta_m,\mu,\sigma^2,\boldsymbol{x}_1,\dots,\boldsymbol{x}_m) \propto \pi(\tau^2)\prod_{j=1}^{m}\pi\left(\theta_j|\mu,\tau^2\right)$$

and it can be shown the last is IG distribution

Conditional posteriors of θ_j and σ^2

It can be shown that $\boldsymbol{\theta_j}$s are conditionally independent and

$$f\left(\theta_j \middle| \mu, \tau^2, \sigma^2, \boldsymbol{x}_1, \dots, \boldsymbol{x}_m\right) \propto \pi\left(\theta_j \middle| \mu, \tau^2\right) \prod_{i=1}^{n_j} f\left(x_{ij} \middle| \theta_j, \sigma^2\right)$$

and

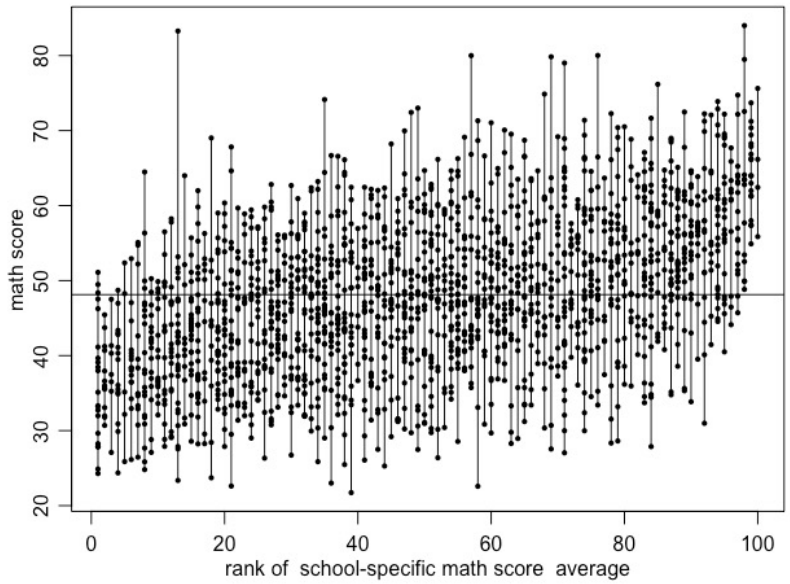
$$\theta_j | \mu, \tau^2, \sigma^2, \boldsymbol{x}_1, \dots, \boldsymbol{x}_m \sim N\left(\frac{\frac{n_j \bar{x}_j}{\sigma^2} + \frac{1}{\tau^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

For the distribution of

$$\sigma^2 | \theta_1, \dots \theta_m, \boldsymbol{x}_1, \dots, \boldsymbol{x}_m \sim IG$$

see the details on p. 135.

Example: Math scores in US public schools



Let's return to the 2002 ELS data described previously. This survey included 10th grade children from 100 different large urban public high schools, all having a 10th grade enrollment of 400 or greater. Data from these schools are shown on the left, with scores from students within the same school plotted along a common vertical bar.

Example: Prior parameters

We need to specify the following parameters for the priors:

- ν_0, σ_0^2 for $\pi(\sigma^2)$
- η_0, τ_0^2 for $\pi(\tau^2)$
- μ_0, γ_0^2 for $\pi(\mu)$

- The exam was designed to have a nationwide variance of 100. Therefore, we set the within-school variance $\sigma_0^2 = 100$.
- We only weakly concentrate the prior distribution around this value by taking $\nu_0 = 1$.
- For the rest of the parameters, see p. 137

Example: Gibbs sampler

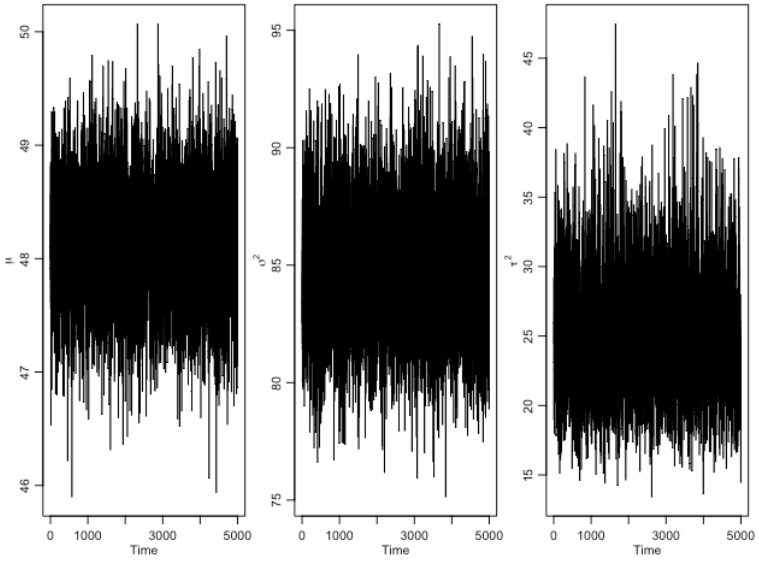
Posterior approximation proceeds by iterative sampling of each unknown quantity from its full conditional distribution. Given a current state of the unknowns $\left\{\theta_1^{(s)}, \ldots, \theta_m^{(s)}, \mu^{(s)}, \tau^{2(s)}, \sigma^{2(s)}\right\}$, a new state is generated as follows:

1. Sample $\mu^{(s+1)} \sim f\left(\mu \middle| \theta_1^{(s)}, \ldots, \theta_m^{(s)}, \tau^{2(s)}\right)$
2. Sample $\tau^{2(s+1)} \sim f(\tau^2 | \theta_1^{(s)}, \ldots, \theta_m^{(s)}, \mu^{(s+1)})$
3. Sample $\sigma^{2(s+1)} \sim f\left(\sigma^2 \middle| \theta_1^{(s)}, \ldots, \theta_m^{(s)}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\right)$
4. For each $j = 1, \ldots, m$, sample $\theta_j^{(s)} \sim f\left(\theta_j \middle| \mu^{(s+)}, \tau^{2(s+1)}, \sigma^{2(s+1)}, \boldsymbol{x}_j\right)$

Example: MCMC diagnostics

Before we make inference using these MCMC samples we should determine if there might be any problems with the Gibbs sampler.

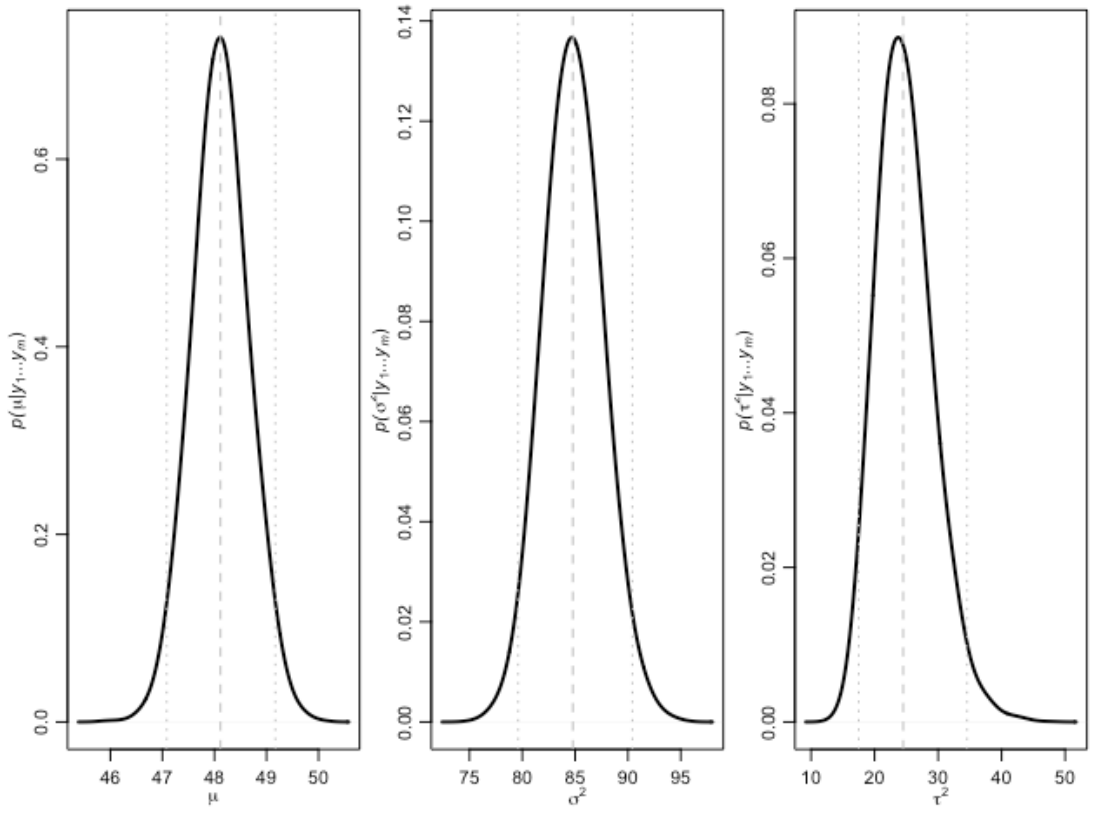
- Check stationarity with a trace plots or boxplots of sequential groups of samples: if the simulated parameter values are moving in a consistent direction, then the chain is not stationary. If stationarity has been achieved, then the distribution of samples in any one boxplot should be the same as that in any other. If not, run the chains longer
- Check autocorrelations
- Check ESS



Example: Posterior summaries

Posterior densities of μ, σ^2, τ^2

Roughly 95% of the scores within a classroom are within $4 \times 9.21 \approx 37$ points of each other, whereas 95% of the average classroom scores are within $4 \times 4.97 \approx 20$ points of each other.



Example: Shrinkage

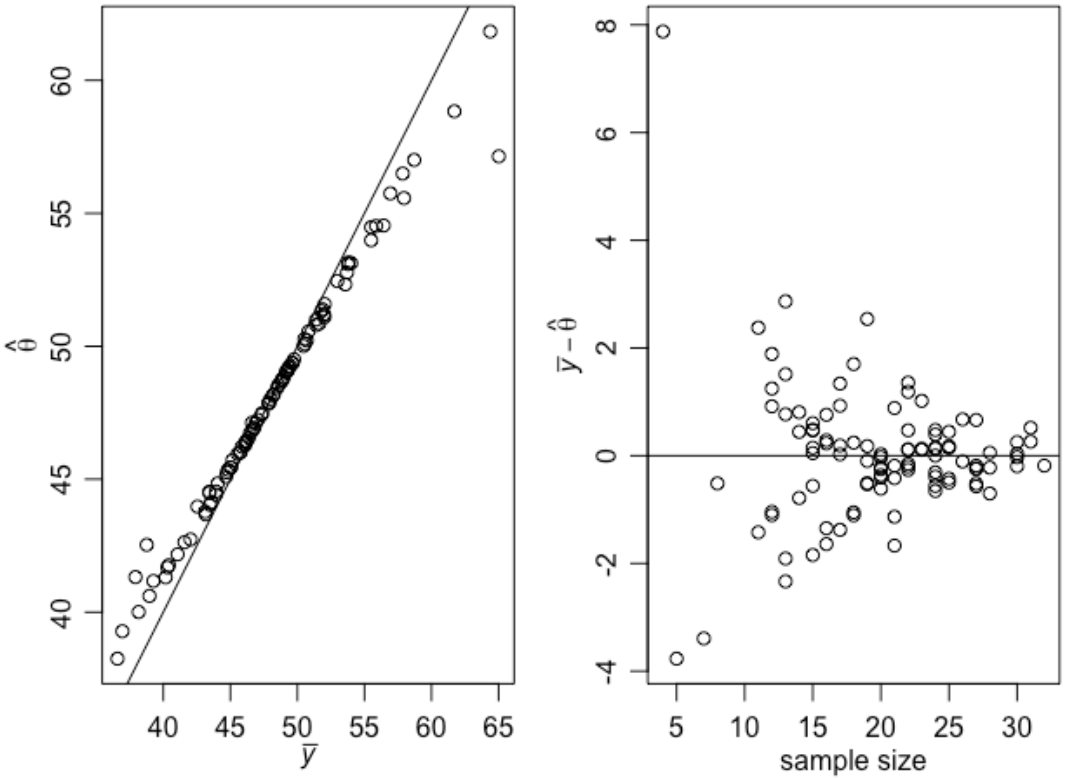
One of the motivations behind hierarchical modeling is that information can be shared across groups. It can be shown that

$$E\left(\theta_j \middle| \boldsymbol{x}_j, \mu, \tau^2, \sigma^2\right) = \frac{\frac{\bar{x}_j n_j}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$$

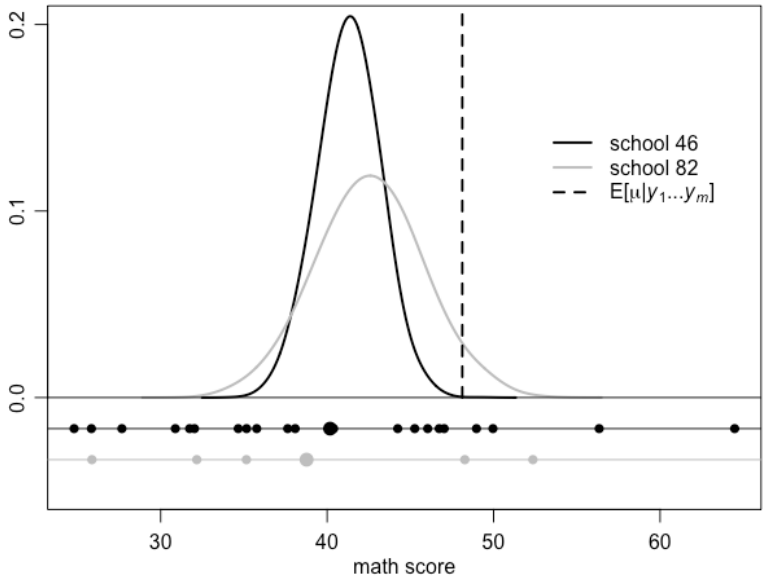
As a result, the expected value of θ_j is pulled a bit from $\bar{x}_j$ towards μ by an amount depending on n_j. This effect is called *shrinkage*.

Example: Shrinkage

- The first panel of the figure plots $\bar{x}_j$ versus $\hat{\theta}_j$. Notice that the relationship roughly follows a line with a slope that is less than one.
- The second panel of the plot shows the amount of shrinkage as a function of the group-specific sample size.



Example: school comparisons



Consider the posterior distributions of θ_{46} and θ_{82}. Both schools have exceptionally low sample means, in the bottom 10% of all schools. The first thing to note is that the posterior density for school 46 is more peaked than that of school 82. This is because the sample size for school 46 is 21 students, whereas that of school 82 is only 5 students.

Hierarchical Normal Model - Extended

If the population means vary across groups, shouldn't we allow for the possibility that the population variances also vary across groups? Then the model with $\phi_j = (\theta_j, \sigma_j^2)$ and $\psi = (\mu, \tau^2)$ is:

$$X_{1j}, \dots, X_{n_j j} | \phi_j \sim N(\theta_j, \sigma_j^2)$$

$$\theta_j | \psi \sim N(\mu, \tau^2)$$

$$\sigma_1^2, \dots, \sigma_m^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

$$\psi = (\mu, \tau^2) \sim \pi(\mu)\pi(\tau^2)$$

Then

$$\theta_j | \mu, \tau^2, \sigma_j^2, \boldsymbol{x}_1, \dots, \boldsymbol{x}_m \sim N\left(\frac{\frac{n_j \bar{x}_j}{\sigma_j^2} + \frac{1}{\tau^2}}{\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2}}\right)$$

Note: ν_0 and σ_0^2 can also be set as parameters to be estimated, which would allow for sharing of info between $\sigma_1^2, \dots, \sigma_m^2$.

Hierarchical Normal Model - Extended

Notice the moderate estimated value of ν_0

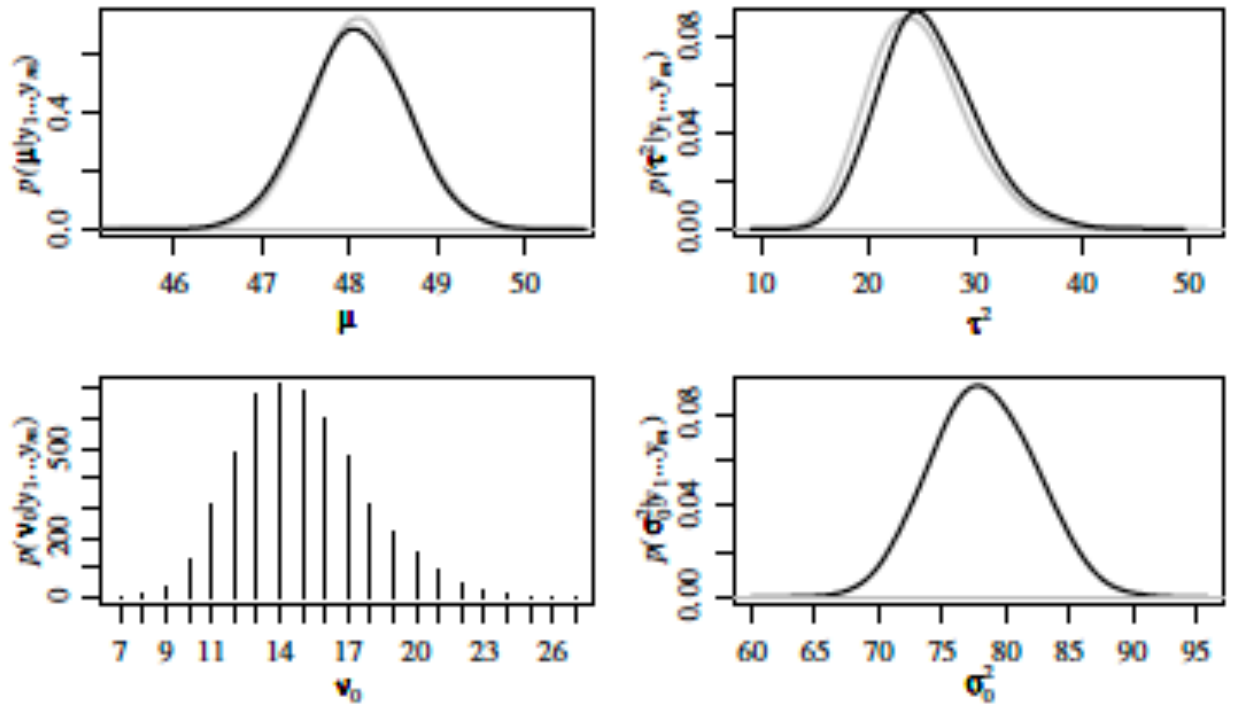


Fig. 8.11. Posterior distributions of between-group heterogeneity parameters.