VE414 Lecture 20

Jing Liu

UM-SJTU Joint Institute

November 18, 2019

- In general, when we have a lot of groups,

  e.g. 100 courses that SJTU offers in a year

  a hierarchical model is typically used to model the between-group variability

  $$\theta_j \mid \{\mu, \tau^2\} \sim \mathrm{Normal}\left(\mu, \tau^2\right)$$

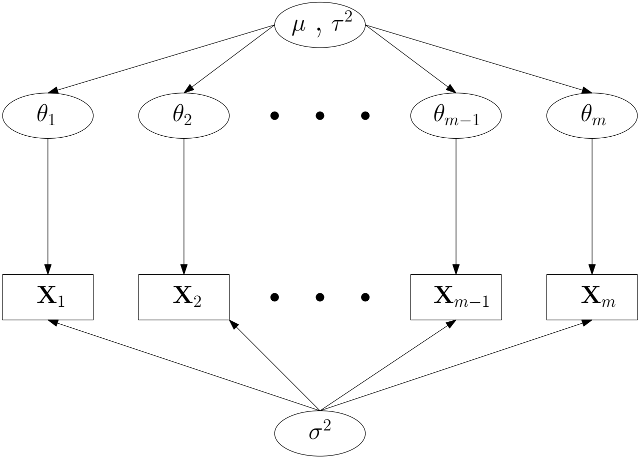
  as well as the within-group variability

  $$X_{ij} \mid \{\theta_j, \sigma^2\} \sim \mathrm{Normal}\left(\theta_j, \sigma^2\right)$$

- For computational convenience, we use the conjugate priors σ^2, μ and τ^2:

  $$\begin{aligned} \sigma^2 &\sim \mathrm{Scaled\,Inverse}\,\chi^2\left(\nu_0, \sigma_0^2\right) \\ \mu &\sim \mathrm{Normal}\left(\mu_0, \gamma_0^2\right) \\ \tau^2 &\sim \mathrm{Scaled\,Inverse}\,\chi^2\left(\eta_0, \tau_0^2\right) \end{aligned}$$

- Here we assume the same within-group sampling variability σ^2 across groups.



- There are $m+3$ number of unknown quantities in this hierarchical model,

$$f_{\{\boldsymbol{\theta},\mu,\tau^2,\sigma^2\}|\{\mathbf{X}_1,\ldots,\mathbf{X}_m\}} \propto \left\{ \prod_{j=1}^{m} \prod_{i=1}^{n_j} f_{X_{ij}|\{\theta_j,\sigma^2\}} \right\} f_\mu f_{\tau^2} f_{\sigma^2} \prod_{j=1}^{m} f_{\theta_j|\{\mu,\tau^2\}}$$

- It can be shown the full conditional posterior of θ_j is given by

$$\begin{aligned}\theta_j \mid \{\mathbf{X}_j, \mu, \sigma^2, \tau^2\} &= \theta_j \mid \{\mathbf{X}_1, \ldots, \mathbf{X}_m, \boldsymbol{\theta}_{-j}, \mu, \sigma^2, \tau^2\} \\ &\sim \mathrm{Normal}\left(\frac{\tau^2 \bar{x}_j + \mu\sigma^2/n_j}{\tau^2 + \sigma^2/n_j}, \frac{\tau^2\sigma^2/n_j}{\tau^2 + \sigma^2/n_j}\right)\end{aligned}$$

and other full conditional posteriors are given by

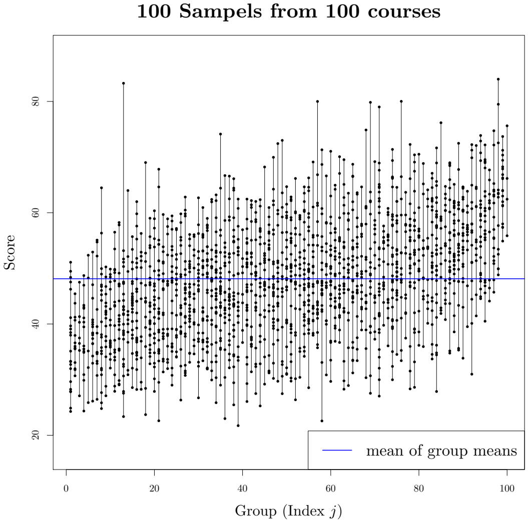
$$\begin{aligned}\mu \mid \{\boldsymbol{\theta}, \tau^2\} &\sim \mathrm{Normal}\left(\mu_n, \gamma_n^2\right) \\ \tau^2 \mid \{\boldsymbol{\theta}, \mu\} &\sim \text{Scaled Inverse}\,\chi^2\left(\eta_n, \tau_n^2\right) \\ \sigma^2 \mid \{\mathbf{X}_1, \ldots, \mathbf{X}_m, \boldsymbol{\theta}\} &\sim \text{Scaled Inverse}\,\chi^2\left(\nu_n, \sigma_n^2\right)\end{aligned}$$

where $\mu_n = \dfrac{\gamma_0^2\bar{\theta} + \mu_0\tau^2/m}{\gamma_0^2 + \tau^2/m}$; $\qquad \gamma_n^2 = \dfrac{\gamma_0^2\tau^2/m}{\gamma_0^2 + \tau^2/m}$

$$\eta_n = \eta_0 + m; \qquad \eta_n\tau_n^2 = \eta_0\tau_0^2 + \sum_{j=1}^m (\theta_j - \mu)^2$$

$$\nu_n = \nu_0 + \sum_{j=1}^m n_j; \qquad \nu_n\sigma_n^2 = \nu_0\sigma_0^2 + \sum_{j=1}^m \sum_{i=1}^{n_j} (x_{ij} - \theta_j)^2$$

- Suppose all data are sampled from the freshman courses, one for each course



- The set of full conditional posteriors are available, so we can again use Gibbs

1. Sample $\mu^{(t+1)} \sim f_{\mu|\{\boldsymbol{\theta},\tau^2\}} = \text{Normal}\left(\mu_n, \gamma_n^2\right)$ where

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)} \qquad \text{and} \qquad \tau^2 = {\tau^2}^{(t)}$$

2. Sample ${\tau^2}^{(t+1)} \sim f_{\tau|\{\boldsymbol{\theta},\mu\}} = \text{Scaled Inverse}\,\chi^2\left(\eta_n, \tau_n^2\right)$ where

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)} \qquad \text{and} \qquad \mu = \mu^{(t+1)}$$

3. Sample ${\sigma^2}^{(t+1)} \sim f_{\tau|\{\mathbf{X}_1,\ldots,\mathbf{X}_m,\boldsymbol{\theta}\}} = \text{Scaled Inverse}\,\chi^2\left(\nu_n, \sigma_n^2\right)$ where

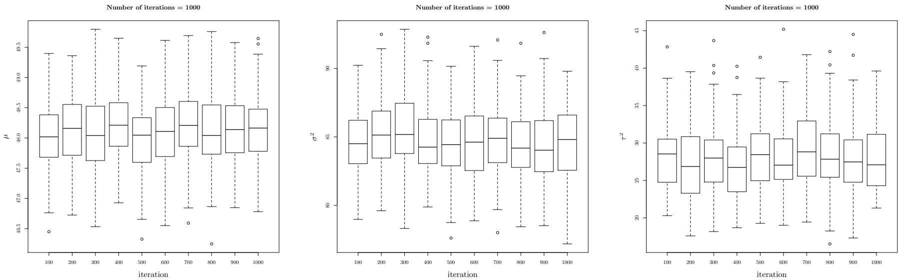
$$\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)} \qquad \text{and} \qquad \mathbf{X}_j = \mathbf{x}_j$$

4. For each $j \in \{1, \ldots, m\}$,

Sample $\theta_j^{(t+1)} \sim f_{\theta_j|\{\mathbf{X}_j,\mu,\sigma^2,\tau^2\}} = \text{Normal}\left(\dfrac{\tau^2\bar{x}_j + \mu\sigma^2/n_j}{\tau^2 + \sigma^2/n_j}, \dfrac{\tau^2\sigma^2/n_j}{\tau^2 + \sigma^2/n_j}\right)$

$$\mu = \mu^{(t+1)}, \qquad \sigma^2 = {\sigma^2}^{(t+1)} \qquad \text{and} \qquad \tau^2 = {\tau^2}^{(t+1)}$$

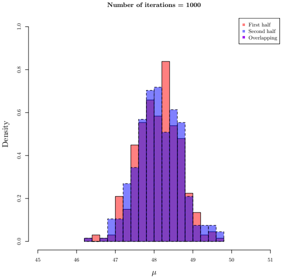
- For a large scale MCMC like this one, diagnostics become more important.

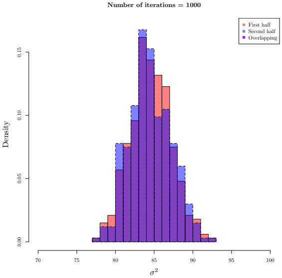


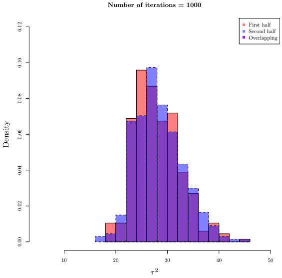
- From the above boxplots, where every 100th sample is plotted, the medians seem to converge really quickly in this case. The following priors were used:

$$\sigma^2 \sim \text{Scaled Inverse}\, \chi^2 \left(\nu_0 = 4, \sigma_0^2 = 10^2\right)$$
$$\mu \sim \text{Normal} \left(\mu_0 = 50, \gamma_0^2 = 5^2\right)$$
$$\tau^2 \sim \text{Scaled Inverse}\, \chi^2 \left(\eta_0 = 4, \tau_0^2 = 10^2\right)$$

- However, the shape of the distributions of μ, σ^2, and τ^2 are not quite stable

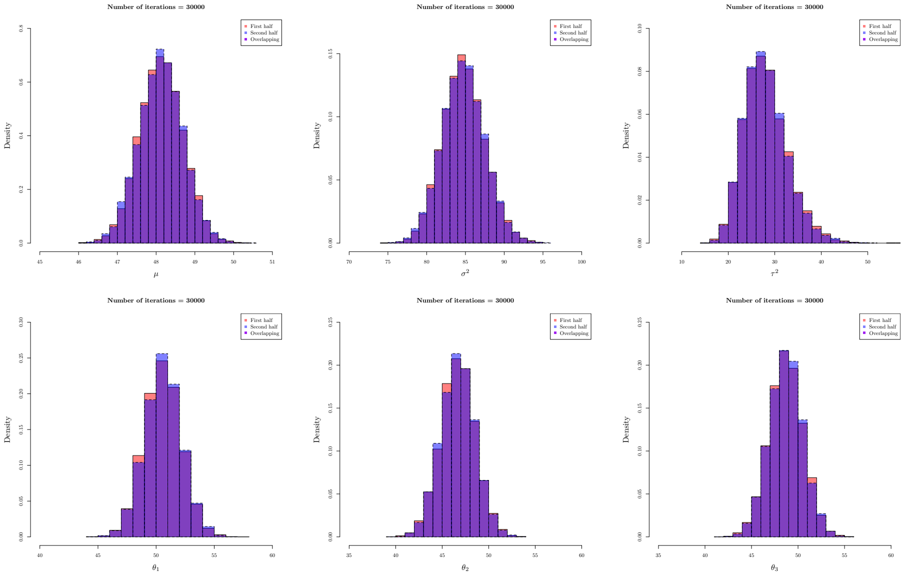






- We treat the first $1/3$ of the Markov chain as the burn-in, and the rest is split into two halves and two histograms are produced, one for each half.
- Depends on the level of accuracy we need, we demand how similar the two histograms need to be. In this example, we have 103 such pairs to consider.
- The histograms for θ_j show similar level of convergence as the ones above.

- The following show some diagnostic plots when I run it for 30000 iterations.



- Recall our primary interest is to estimate θ_j in contrast to

$$\hat{\theta}_1 = w\bar{x}_1 + (1-w)\bar{x}_2 \qquad \text{where} \quad w = \begin{cases} 1 & \text{if p-value} < 0.05, \\ n_1/(n_1+n_2) & \text{otherwise.} \end{cases}$$

and one of the motivations behind this hierarchical model is that information can be shared across groups. Recall full conditional posterior of θ_j is given by

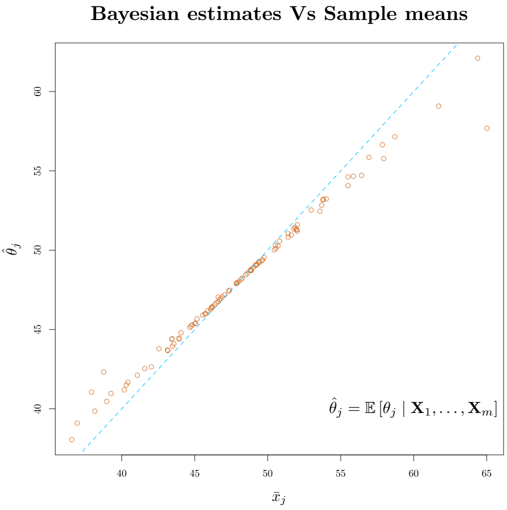
$$\theta_j \mid \{\mathbf{X}_j, \mu, \sigma^2, \tau^2\} \sim \mathrm{Normal}\left(\frac{\tau^2\bar{x}_j + \mu\sigma^2/n_j}{\tau^2 + \sigma^2/n_j}, \frac{\tau^2\sigma^2/n_j}{\tau^2 + \sigma^2/n_j}\right)$$

- Hence the expected value of θ_j conditional μ, σ^2, τ^2 and the data is given by

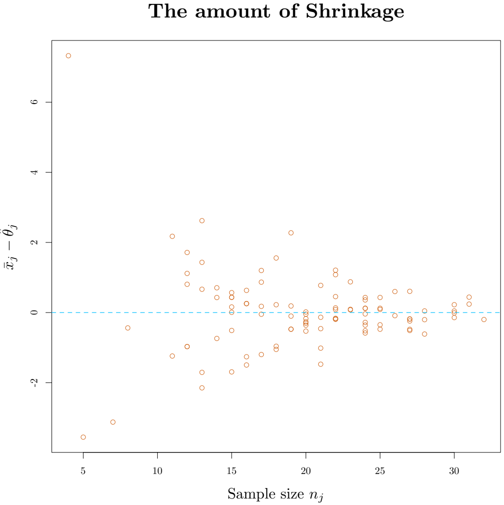
$$\mathbb{E}\left[\theta_j \mid \{\mathbf{X}_j, \mu, \sigma^2, \tau^2\}\right] = \frac{\tau^2\bar{x}_j + \mu\sigma^2/n_j}{\tau^2 + \sigma^2/n_j}$$

- As a result, the mean is pulled a bit from $\bar{x}_j$ towards μ by to some degree depending on n_j as well as other parameters, this is known as shrinkage.

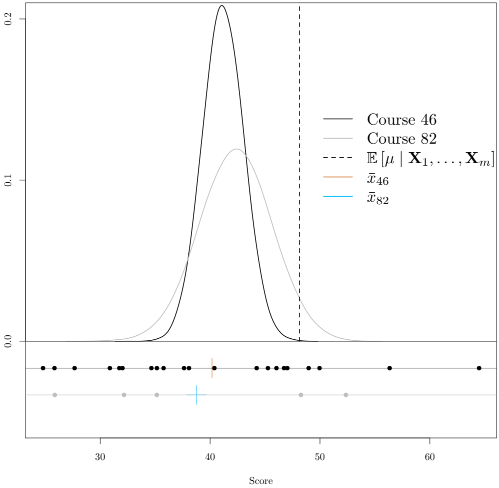
- The Relationship roughly follows a line with a slope slightly less than 1.



- Groups with low sample sizes shrunk the most.



Q: Do you notice anything surprising?



- In general, instead of assuming σ_j^2 to be the same for all groups,

$$X_{ij} \mid \{\theta_j, \sigma^2\} \sim \text{Normal}\left(\theta_j, \sigma^2\right)$$

we could assume they vary from groups to groups, i.e.

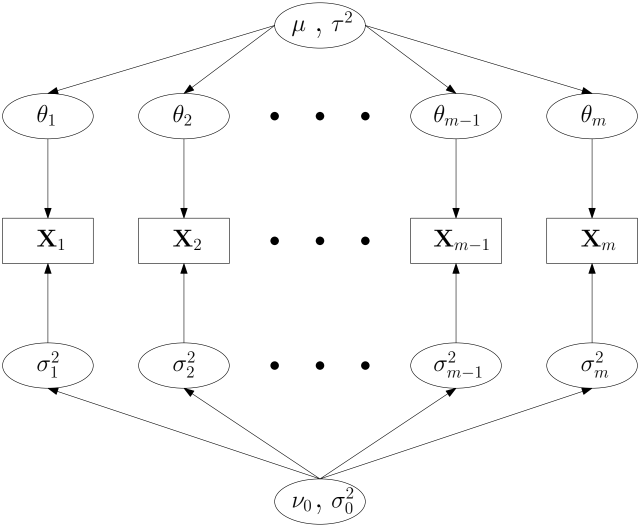
$$X_{ij} \mid \{\theta_j, \sigma_j^2\} \sim \text{Normal}\left(\theta_j, \sigma_j^2\right)$$

then the full conditional posterior of θ_j becomes

$$\theta_j \mid \{\mathbf{X}_j, \mu, \sigma_j^2, \tau^2\} \sim \text{Normal}\left(\frac{\tau^2 \bar{x}_j + \mu\sigma_j^2/n_j}{\tau^2 + \sigma_j^2/n_j}, \frac{\tau^2\sigma_j^2/n_j}{\tau^2 + \sigma_j^2/n_j}\right)$$

- When σ_j^2 are not the same, in order to use information from all the groups to estimate σ_j^2, we will have to add another layer to our hierarchical model:

$$\sigma_j^2 \sim \text{Scaled Inverse}\, \chi^2\left(\nu_0, \sigma_0^2\right)$$



μ , τ^2
θ_1
θ_2
θ_{m-1}
θ_m
$\mathbf{X}_1$
$\mathbf{X}_2$
$\mathbf{X}_{m-1}$
$\mathbf{X}_m$
σ_1^2
σ_2^2
σ_{m-1}^2
σ_m^2
ν_0 , σ_0^2

- Now we have to treat ν_0 and σ_0^2 as random variables, and specify priors.
- A conjugate prior for σ_0^2 is

$$\sigma_0^2 \sim \mathrm{Gamma}\,(a, b)$$

the corresponding full conditional posterior is given

$$\sigma_0^2 \mid \{\boldsymbol{\sigma}^2, \nu_0\} \sim \mathrm{Gamma}\left(a + \frac{1}{2}m\nu_0, b + \frac{1}{2}\sigma_*^2\right), \quad \text{where} \quad \sigma_*^2 = \sum_{j=1}^{m} \frac{1}{\sigma_j^2}$$

- No simple conjugate prior for ν_0 exists, but if we restrict ν_0 to be $\{1, 2, \ldots\}$,

$$\nu_0 \sim \mathrm{Geometric}\left(1 - e^{-\alpha}\right)$$

can be used to have a "simple" full conditional posterior that we can sample

$$\nu_0 \mid \{\sigma_0^2, \boldsymbol{\sigma}^2\} \propto \left(\frac{(\sigma_0^2\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)}\right)^m \left(\prod_{j=1}^{m} \frac{1}{\sigma_j^2}\right)^{\nu_0/2-1} \exp\left(-\frac{\nu_0}{2}\left(2\alpha + \sigma_0^2\sigma_*^2\right)\right)$$