Bayesian Multi-level Regression

Models for Socio-Environmental Data

N. Thompson Hobbs

August 21, 2017

Lecture material

- Background
- Bayesian, multilevel models for grouped data
  - group level intercepts
  - group level intercepts with group level covariate
  - group level slopes and intercepts
  - an essential coding trick
- Priors on group level variances

The simple, Bayesian set-up

Deterministic model:

$$g(\boldsymbol{\theta}, x_i)$$

Stochastic model:

$$\underbrace{[\boldsymbol{\theta}, \sigma^2 | y_i]}_{\text{posterior}} \propto \overbrace{\underbrace{[y_i | g(\boldsymbol{\theta}, x_i), \sigma^2]}_{\text{likelihood}} \underbrace{[\boldsymbol{\theta}][\sigma^2]}_{\text{priors}}}^{\text{joint}}$$

Recall that

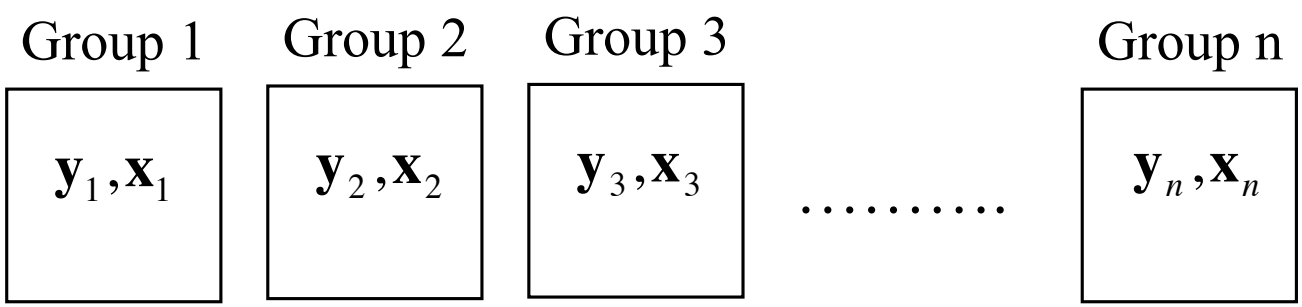
$$\underbrace{[\boldsymbol{\theta}, \sigma^2 | y_i]}_{\text{posterior}} \propto \overbrace{[y_i, \boldsymbol{\theta}, \sigma^2]}^{\text{joint}}$$

Hierarchical models: "modeling parameters"

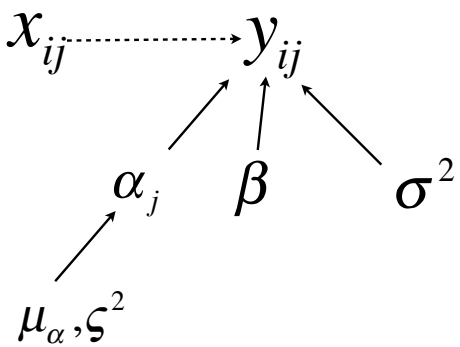
$$
\begin{aligned}
[\theta_1, \boldsymbol{\theta}_2, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2 \mid y_{ij}] &\propto [\theta_1, \boldsymbol{\theta}_2, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2, y_{ij}] \\
[\theta_1, \boldsymbol{\theta}_2, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2 \mid y_{ij}] &\propto [y_{ij} | g(\theta_1, \theta_{2,j}, x_{ij}), \sigma_1^2] \\
&\times [\theta_{2,j} | h(\alpha_1, \alpha_2, u_j), \sigma_2^2] \\
&\times [\theta_1], [\alpha_1], [\alpha_2] [\sigma_1^2] [\sigma_2^2]
\end{aligned}
$$

Draw the DAG.

The problem

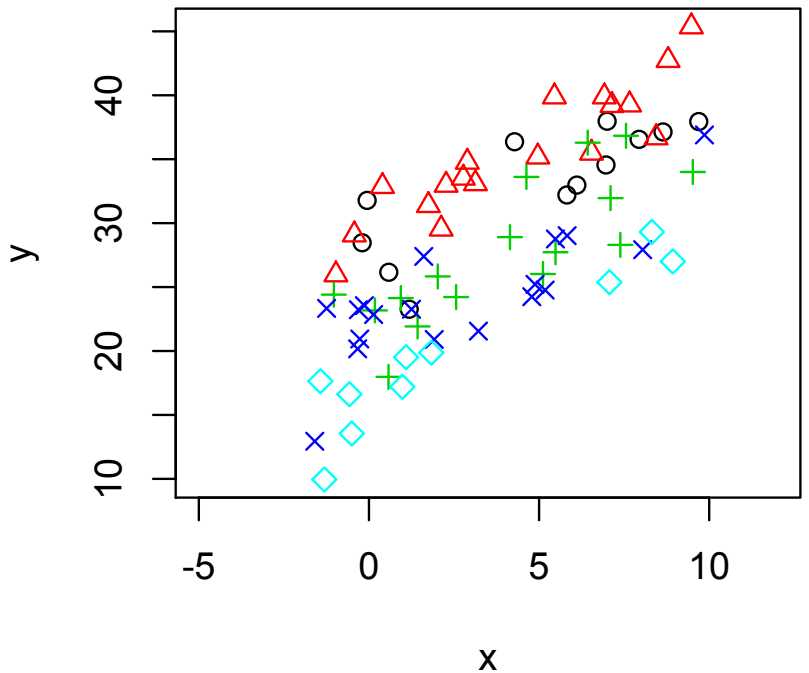

Group 1
Group 2
Group 3
Group n
$\mathbf{y}_1, \mathbf{x}_1$
$\mathbf{y}_2, \mathbf{x}_2$
$\mathbf{y}_3, \mathbf{x}_3$
.....
$\mathbf{y}_n, \mathbf{x}_n$

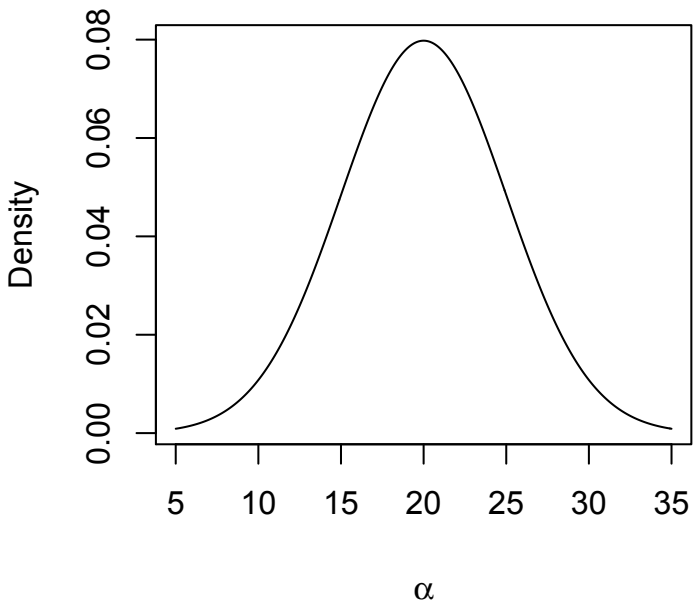
We can model the intercept (or slope):



$$[\beta, \boldsymbol{\alpha}, \sigma^2, \mu_\alpha, \varsigma^2, |\mathbf{y}] \propto \prod_{i=1}^{n_j} \prod_{j=1}^{J} \text{normal}\left(y_{ij} | \alpha_j + \beta x_{ij}, \sigma^2\right)$$
$$\times \text{normal}\left(\alpha_j | \mu_\alpha, \varsigma^2\right)$$
$$\times \text{normal}\left(\beta | 0, 10000\right) \text{normal}\left(\mu_\alpha | 0, 1000\right)$$
$$\times \text{inverse gamma}\left(\sigma^2 | .001, .001\right) \text{inverse gamma}\left(\varsigma^2 | .001, .001\right)$$

We seek to understand the distribution of intercepts.[1]





[1] Remember to talk about borrowing strength.

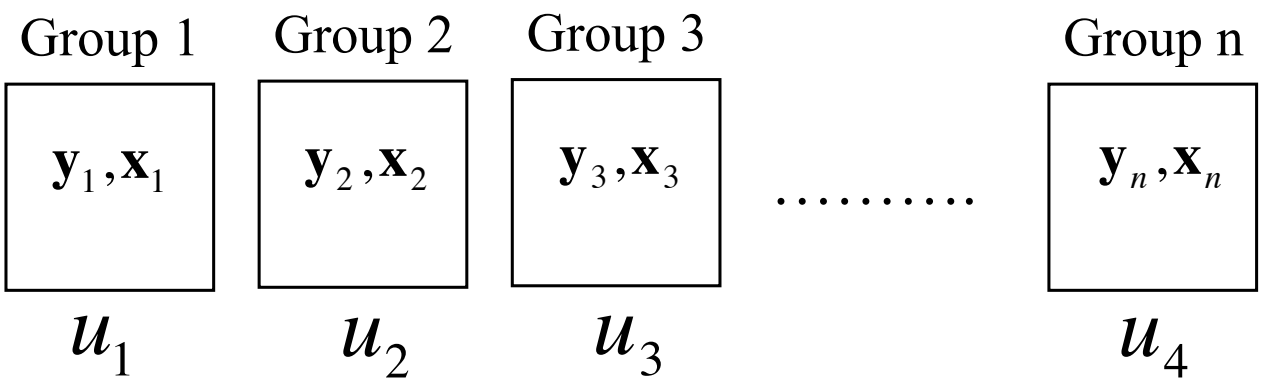
Some notation

$$\mu_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_j$$
$$y_{ij} \sim \text{normal}(\mu_{ij}, \sigma^2)$$
$$\alpha_j \sim \text{normal}(0, \varsigma^2)$$

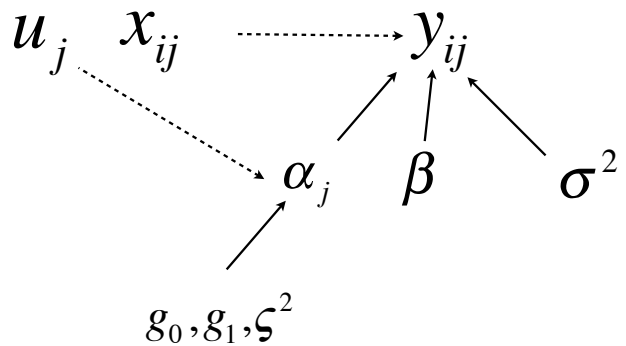
is identical to:

$$\mu_{ij} = \alpha_j + \beta_1 x_{ij}$$
$$y_{ij} \sim \text{normal}(\mu_{ij}, \sigma^2)$$
$$\alpha_j \sim (\mu_\alpha, \varsigma^2)$$

Include data on groups.



We can model the intercept (or slope) as a function of group level data:



$$[\boldsymbol{\alpha}, \beta, \sigma^2, \mathbf{g}, \varsigma^2, |\mathbf{y}] \propto \prod_{i=1}^{n_j} \prod_{j=1}^{J} \text{normal}\left(y_{ij} | \alpha_j + \beta x_{ij}, \sigma^2\right)$$
$$\times \text{normal}\left(\alpha_j | g_0 + g_1 u_j, \varsigma^2\right)$$
$$\times \text{normal}\left(\beta | 0, .001\right) \text{normal}\left(g_0 | 0, 1000\right) \text{normal}\left(g_1 | 0, 1000\right)$$
$$\times \text{inverse gamma}\left(\sigma^2 | .001, .001\right) \text{inverse gamma}\left(\varsigma^2 | .001, .001\right)$$

Modeling intercepts and slopes

A correlation matrix:

Correlations

	Weight in kg	Hours of Sleep	Exposure while Sleeping	Life Span
Weight in kg	1	-.307	.338	.302
Hours of Sleep	-.307	1	-.642	-.410
Exposure while Sleeping	.338	-.642	1	.360
Life Span	.302	-.410	.360	1

[2]

If we multiply this correlation matrix times $\sigma_i \sigma_j$ we obtain a *covariance* matrix.

[2] http://www.theanalysisfactor.com/covariance-matrices/

Recall the correlation coefficient

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Modeling intercepts and slopes

Imagine a vector of 3 random variables, $(z_i, z_2, z_3)'$ The covariance between any two of these random variables is simply an unstandardized version of the correlation between them– it is correlation measured in the units of the random variables. The covariance matrix (aka variance covariance matrix) of the random variable is:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \text{Cov}_{1,2} & \text{Cov}_{1,3} \\ \text{Cov}_{2,1} & \sigma_2^2 & \text{Cov}_{2,3} \\ \text{Cov}_{3,1} & \text{Cov}_{3,2} & \sigma_3^2 \end{pmatrix} \quad (1)$$

Generalizing, a $m \times m$ covariance matrix has the variances of the random variable on the diagonal and the covariance on the off diagonal. The covariance between random variable i and j is $\text{Cov}_{ij} = \rho \sigma_i \sigma_j$ where ρ is the correlation coefficient, which takes on values between -1 and 1. Covariance can take on values between $-\infty$ and $+\infty$.

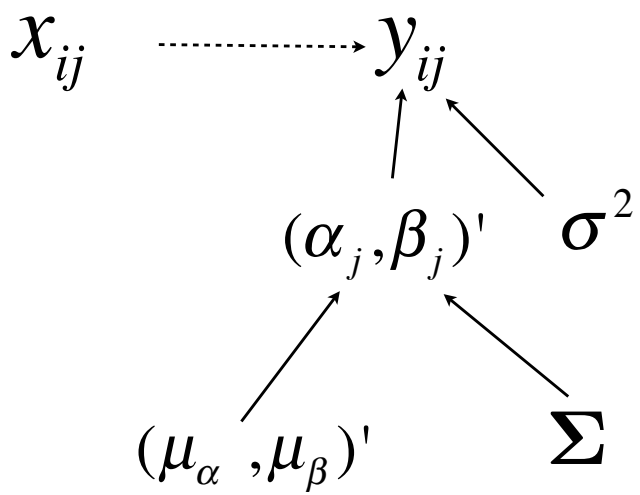
Covariance matrix for two parameter model

Imagine that we have $j = 1, ..., J$ groups with multiple observations within groups and we fit a two parameter linear model to each group, finding J intercepts and slopes. We denote the vector of intercepts as $\boldsymbol{\alpha}$ and the vector of slopes as $\boldsymbol{\beta}$. It is easy to see that we can calculate the variance for each vector $(\sigma_\alpha^2, \sigma_\beta^2)$ as well as the correlation between the vectors ρ. The variance covariance matrix is thus:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\alpha^2 & \text{Cov}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \text{Cov}(\boldsymbol{\beta}, \boldsymbol{\alpha}) & \sigma_\beta^2 \end{pmatrix} \tag{2}$$

where $\text{Cov}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \text{Cov}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \rho \sigma_\alpha \sigma_\beta$

Modeling intercepts *and* slopes



$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim \text{multivariate normal}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \Sigma\right)$$

$$\Sigma = \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}$$

Modeling intercepts *and* slopes

$$\begin{aligned}\left[\boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_\alpha, \mu_\beta, \sigma^2_{\text{reg}}, \sigma^2_\alpha, \sigma^2_\beta, \rho | \mathbf{y}\right] &\propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} \text{normal}(y_{ij} | \alpha_j + \beta_j x_{ij}, \sigma^2_{\text{reg}}) \\ &\times \text{MVN}\left(\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \middle| \begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \boldsymbol{\Sigma} \right) \\ &\times \text{priors on } \mu_\alpha, \mu_\beta, \sigma^2_{\text{reg}}, \sigma^2_\alpha, \sigma^2_\beta, \rho\end{aligned}$$

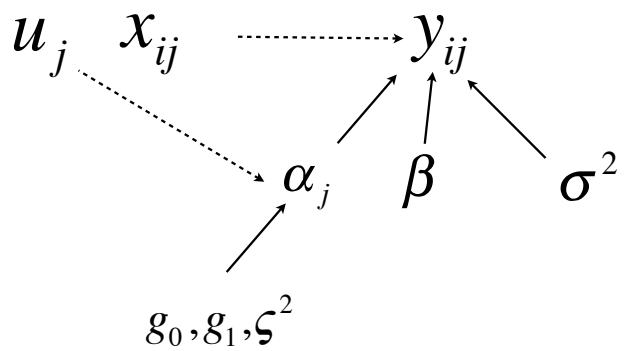
Modeling intercepts and slopes for > 1 slope

See Gelman and Hill, pages 376-380

Some special notation

- We assume that there is a single variance for all random variables such that $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix with ones on the diagonal and zeros elsewhere.
- We assume that each random variable has its own variance σ_i^2 and the random variables are uncorrelated such that $\mathbf{\Sigma} = \mathbf{I}\boldsymbol{\sigma}^2$.

An essential coding trick: Indexing groups



$$[\boldsymbol{\alpha}, \beta, , \sigma^2, \mathbf{g}, \varsigma^2, |\mathbf{y}] \propto \prod_{i=1}^{n_j} \prod_{j=1}^{J} \text{normal}\left(y_{ij} | \alpha_j + \beta x_{ij}, \sigma^2\right)$$
$$\times \text{normal}\left(\alpha_j | g_0 + g_1 u_j, \varsigma^2\right)$$
$$\times \text{normal}\left(\beta | 0, .001\right) \text{normal}\left(g_0 | 0, 1000\right) \times \text{normal}\left(g_1 | 0, 1000\right)$$
$$\times \text{inverse gamma}\left(\sigma^2 | .001, .001\right) \text{inverse gamma}\left(\varsigma^2 | .001, .001\right)$$

Indexing groups

```
> u
[1]  6.215579  8.716296 10.064460 11.292387 14.504154 14.734861
[7] 18.356877 18.910133
```

```
> head(y[,1:4])
     group i        x[i]     y[i]
[1,]     1 1 -0.00266051 13.48934
[2,]     1 2  4.54802848 22.29538
[3,]     1 3  9.86832462 29.03655
[4,]     1 4  0.99869789 18.61136
[5,]     1 5  1.27733200 20.59178
[6,]     1 6  4.32915675 25.37082
> tail(y[,1:4])
       group   i     x[i]     y[i]
[108,]     8 108 4.543959 38.93163
[109,]     8 109 1.287844 34.65796
[110,]     8 110 6.642313 40.62259
[111,]     8 111 7.404183 40.46518
[112,]     8 112 8.252571 41.47995
[113,]     8 113 9.558780 46.14771
```

Indexing groups

```
model{
beta ~ dnorm(0,.0001)
sigma ~ dunif(0,50)
tau.p <- 1/sigma^2
g0 ~ dnorm(0,.0001)
g1 ~ dnorm(0,.0001)
varsigma ~ dunif(0,50)
tau.g <- 1/varsigma^2
 for (i in 1:length(y)){
  mu[i] <- alpha[group[i]]+ beta*x[i]
  y[i] ~ dnorm(mu[i],tau.p)
  }
  for(j in 1:n.group){
  mu.g[j] <- g0 + g1*u[j]
  alpha[j]~dnorm(mu.g[j],tau.g)
    }
```

Reference for priors on group level variance

Gelman. 2006 Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1:1–19

Priors on group-level variances in hierarchical models

The schools data

school	estimate	sd
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Hierarchical model

$$
\begin{aligned}
\theta_j &= \mu + \eta_j \\
y_j &\sim \text{normal}(\theta_j, \text{sd}_j) \\
\eta_j &\sim \text{normal}(0, \sigma_\theta^2) \\
\mu &\sim \text{normal}(0, 100000) \\
\sigma_\theta^2 &\sim ?
\end{aligned}
$$

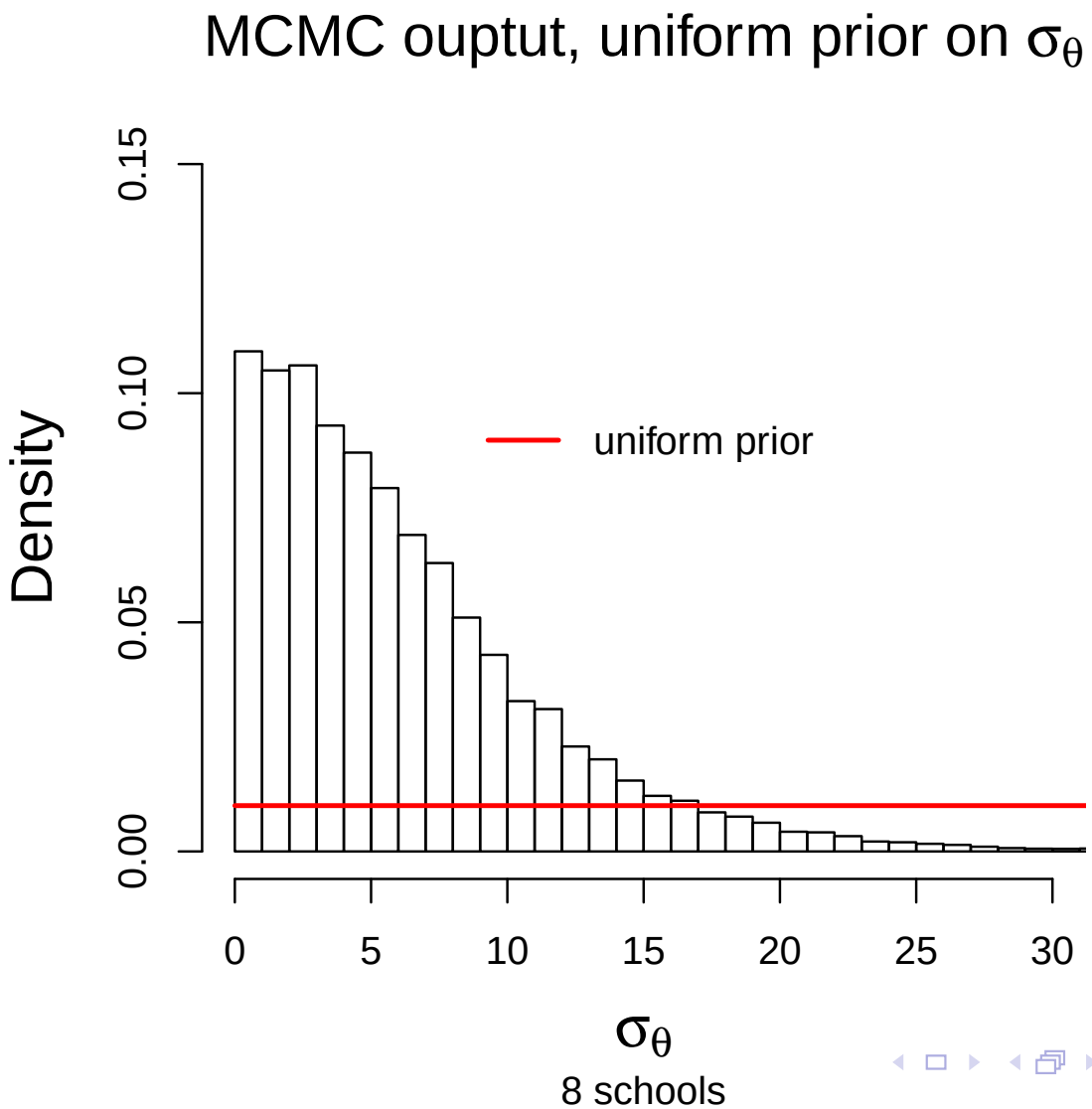
Note that this is identical to:

$$
\begin{aligned}
y_j &\sim \text{normal}(\theta_j, \text{sd}_j) \\
\theta_j &\sim \text{normal}(\mu, \sigma_\theta^2) \\
\mu &\sim \text{normal}(0, 100000) \\
\sigma_\theta^2 &\sim ?
\end{aligned}
$$

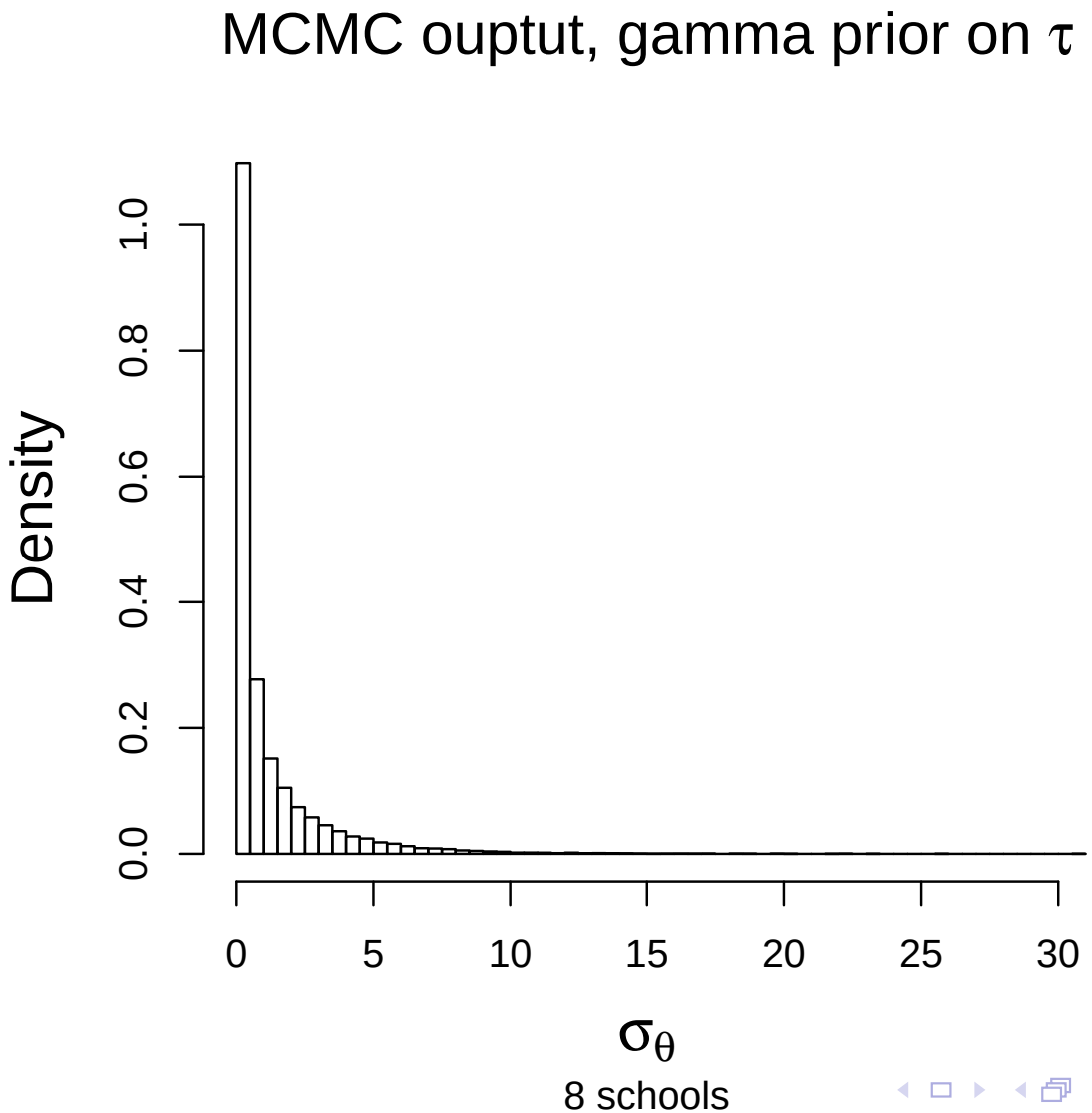
If we had data on individual test scores...

$$
\begin{aligned}
\theta_j &= \mu + \eta_j \\
y_{ij} &\sim \text{normal}(\theta_j, \sigma_j^2) \\
\eta_j &\sim \text{normal}(0, \sigma_\theta^2) \\
\mu &\sim \text{normal}(0,100000) \\
\sigma_\theta^2 &\sim ?
\end{aligned}
$$

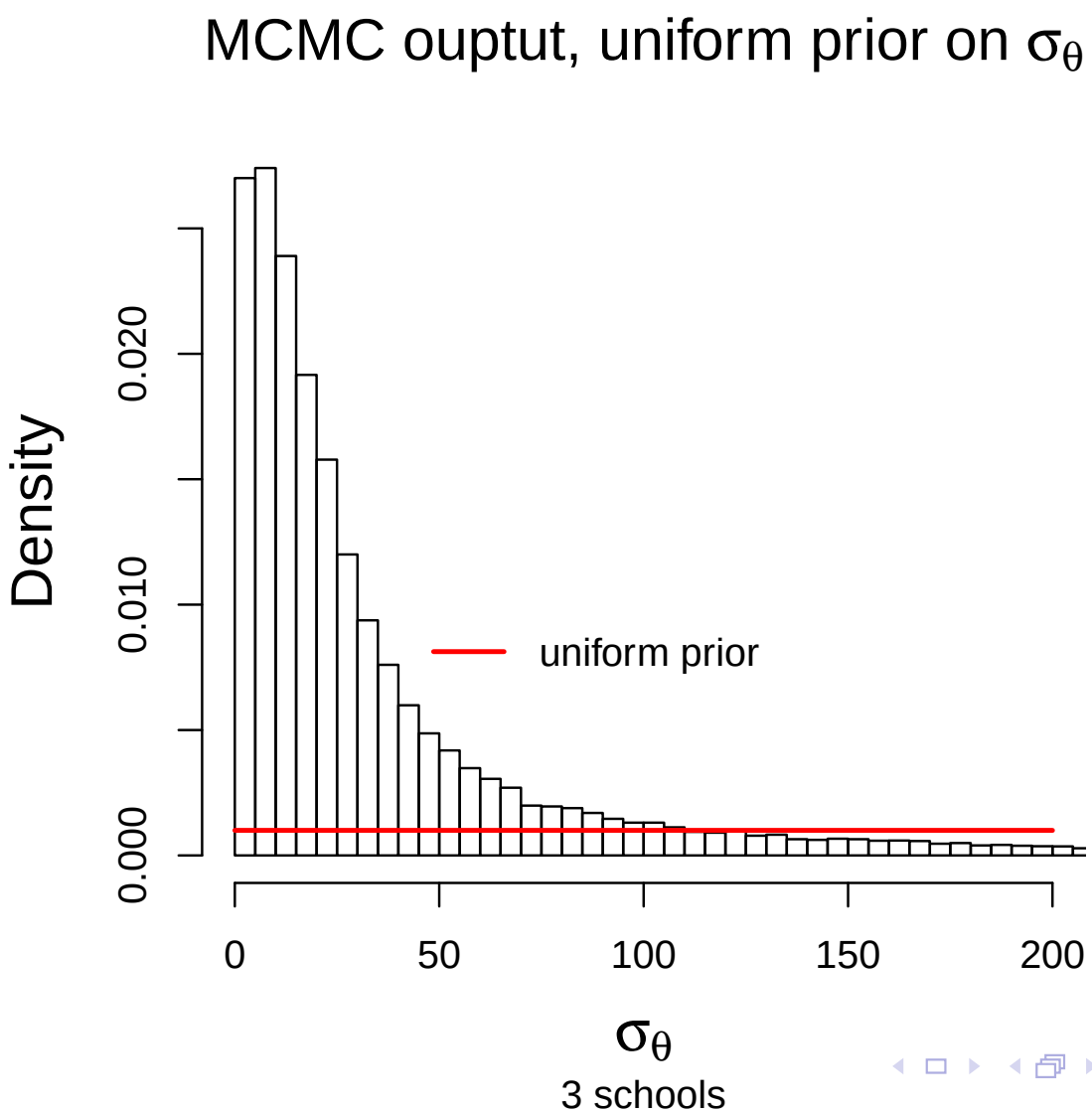
$\sigma_\theta \sim \text{uniform}(0, 100)$, $\tau = \frac{1}{\sigma_\theta^2}$, 8 schools



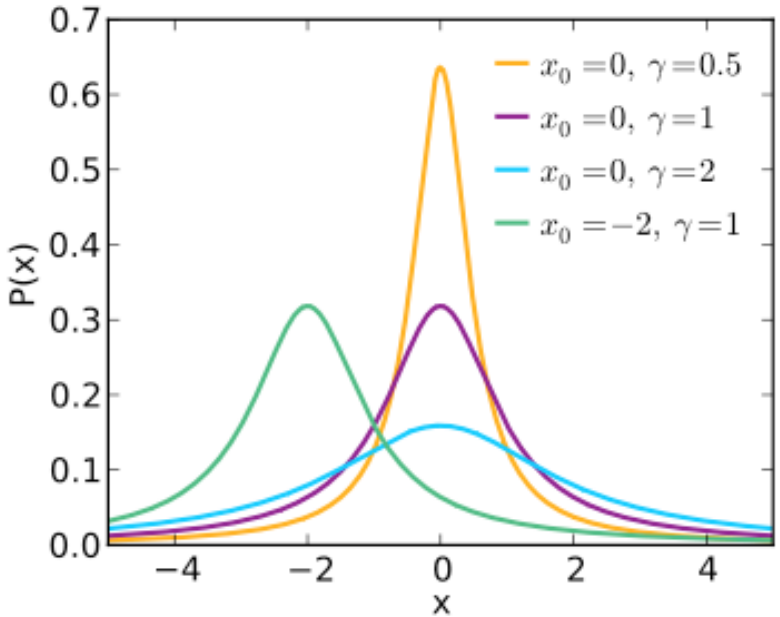
$\tau \sim \text{gamma}(.001, .001)$, 8 schools



$\sigma_\theta \sim \text{uniform}(0, 100)$, $\tau = \frac{1}{\sigma_\theta^2}$, 3 schools



The Cauchy distribution



$$[z|\gamma, z_0] = \frac{1}{\pi\gamma\left[1+\left(\frac{z-z_0}{\gamma}\right)^2\right]}$$

z_0 = location

γ = scale

Represents ratio of two normally distributed random variables

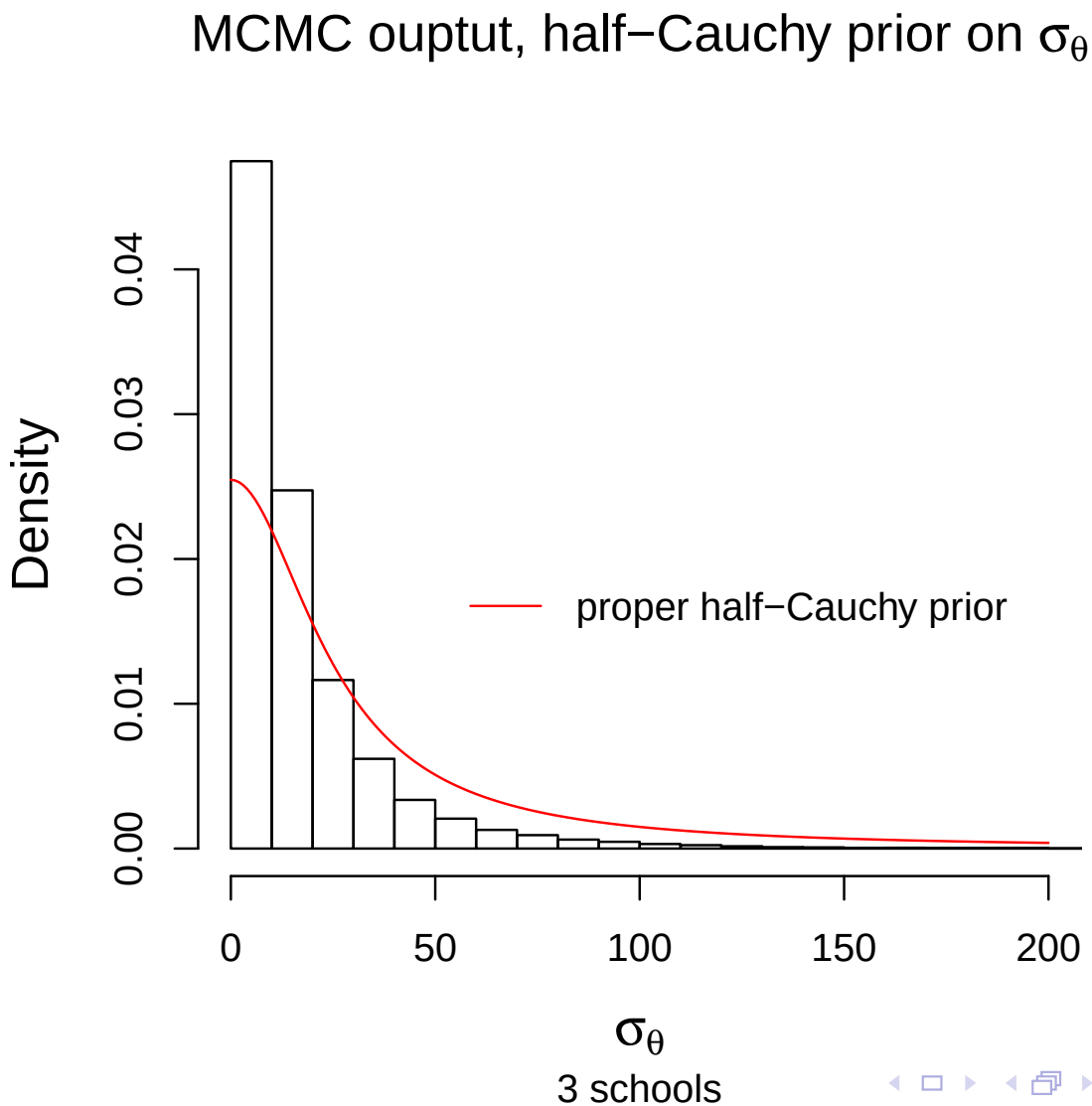
A weakly informative prior on τ

half-Cauchy prior:

$$
\begin{aligned}
z_1 &\sim \text{normal}(0, \text{prior.scale}^{-2}) \\
z_2 &\sim \text{gamma}(.5, .5), \text{ which is } \chi^2 \text{ df=1} \\
\sigma_\theta &= \frac{|z_1|}{\sqrt{z_2}}, \sqrt{z_2} \text{ is half normal} \\
\tau &\sim \frac{1}{\sigma_\theta^2}
\end{aligned}
$$

The parameter prior.scale is chosen based on experience to be a bit higher than we would expect for the standard deviation of the underlying θ_j's. This puts a weak constraint on σ_θ.

A more reasonable posterior



Guidance

- Uniform priors on σ are recommended over gamma priors on group level variances in hierarchical models with at least 4-5 groups.
- When groups are ≤ 4, a half-Cauchy prior can usefully constrain the posterior of group level σ's.
- This illustrates that it can be useful to use weakly informative priors when vague priors produce posteriors with unreasonable values.

A redundant parameterization to speed convergence

$$
\begin{aligned}
\theta_j &= \mu + \xi\,\eta_j \\
y_j &\sim \text{normal}(\theta_j, \text{sd}_j) \\
\eta_j &\sim \text{normal}(0, \sigma_\theta^2)
\end{aligned}
$$

The parameter η_j in the original model corresponds to $\xi\,\eta_j$ in the redundant one. The parameter σ_θ in the original model corresponds to $|\xi|\,\sigma_\theta$ in the redundant one.

A weakly informative alternative to the uniform on σ_θ

$$
\begin{aligned}
\theta_j &= \mu + \xi\,\eta_j \\
y_j &\sim \text{normal}(\theta_j, \text{sd}_j^{-2}) \\
\eta_j &\sim \text{normal}(0, \tau_\theta) \\
\mu &\sim \text{normal}(0,\ .0000001) \\
\xi &\sim \text{normal}(0, \tau_\xi) \\
\tau_\xi &= \frac{1}{\text{prior.scale}^2} \\
\tau_\theta &\sim \text{gamma}(.5, .5) \\
\sigma_\theta &= \frac{|\xi|}{\sqrt{\tau_\theta}}
\end{aligned}
$$

The parameter prior.scale is chosen subjectively to be a bit higher than we would expect for the standard deviation of the underlying θ_j's. This puts a weak constraint on σ_θ.

Some explanation

$$
\begin{aligned}
\xi &\sim \text{normal}(0, \tau_\xi) \\
\tau_\xi &= \frac{1}{\underbrace{\text{prior.scale}^2}_{\text{scale parameter of Cauchy}^2}} \\
\tau_\theta &\sim \underbrace{\text{gamma}(.5, .5)}_{\chi^2 \text{ with 1 df}} \\
\sigma_\theta &= \frac{|\xi|}{\underbrace{\sqrt{\tau_\theta}}_{\text{Cauchy} = \text{normal} \,/\, \sqrt{\chi^2}}}
\end{aligned}
$$