BAYESIAN HIERARCHICAL MODELS

Linda Maokomatanda
Tim Mensinger
Markus Schick

University of Bonn

1. *Bayesian Thinking*

2. *Hierarchical Models*

3. *Monte Carlo Study*

4. *Application*

Bayesian Thinking

Bayes' Theorem

$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta)p(\theta)}{p(\text{data})} \propto p(\text{data} \mid \theta)p(\theta)$$

Bayes' Theorem

$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta)p(\theta)}{p(\text{data})} \propto p(\text{data} \mid \theta)p(\theta)$$

$$\textit{posterior} = \frac{\textit{likelihood} \times \textit{prior}}{\textit{evidence}} \propto \textit{likelihood} \times \textit{prior}$$

Solving for the Posterior Analytically

Setting: $\{y_i : i = 1, \ldots, n\}$ with $y_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and σ^2 known

Solving for the Posterior Analytically

Setting: $\{y_i : i = 1, \ldots, n\}$ with $y_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$
and σ^2 known

Likelihood: $p(y \mid \mu) = \prod_i p(y_i \mid \mu)$

Solving for the Posterior Analytically

Setting: $\{y_i : i = 1, \ldots, n\}$ with $y_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$
and σ^2 known

Likelihood: $p(y \mid \mu) = \prod_i p(y_i \mid \mu)$

Prior: $p(\mu)$

Solving for the Posterior Analytically

Setting: $\{y_i : i = 1, \ldots, n\}$ with $y_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$
and σ^2 known

Likelihood: $p(y \mid \mu) = \prod_i p(y_i \mid \mu)$

Prior: $p(\mu)$

Posterior: $p(\mu \mid y) \propto p(y \mid \mu) p(\mu)$

Solving for the Posterior Analytically

Setting: $\{y_i : i = 1, \dots, n\}$ with $y_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$
and σ^2 known

Likelihood: $p(y \mid \mu) = \prod_i p(y_i \mid \mu)$

Prior: $p(\mu)$

Posterior: $p(\mu \mid y) \propto p(y \mid \mu) p(\mu)$

Goal: Infer distribution of $\mu \mid y$

Solving for the Posterior Analytically

Setting: $\{y_i : i = 1, \ldots, n\}$ with $y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$
and σ^2 known

Likelihood: $p(y \mid \mu) = \prod_i p(y_i \mid \mu)$

Prior: $p(\mu)$

Posterior: $p(\mu \mid y) \propto p(y \mid \mu) p(\mu)$

Goal: Infer distribution of $\mu \mid y$
Why?

Conjugate Prior

Let $\mu \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$

Conjugate Prior

Let $\mu \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$
Then

$$p(\mu \mid y) \propto p(y \mid \mu)p(\mu)$$

Conjugate Prior

Let $\mu \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$
Then

$$\begin{aligned} p(\mu \mid y) &\propto p(y \mid \mu) p(\mu) \\ &\propto \exp\left(\frac{-n}{2\sigma^2}(\mu - \bar{y})^2\right) \exp\left(\frac{-1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \end{aligned}$$

Conjugate Prior

Let $\mu \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$
Then

$$
\begin{aligned}
p(\mu \mid y) &\propto p(y \mid \mu) p(\mu) \\
&\propto \exp\left(\frac{-n}{2\sigma^2}(\mu - \bar{y})^2\right) \exp\left(\frac{-1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma_\mu}(\mu - \bar{\mu})^2\right)
\end{aligned}
$$

Conjugate Prior

Let $\mu \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$
Then

$$
\begin{aligned}
p(\mu \mid y) &\propto p(y \mid \mu) p(\mu) \\
&\propto \exp\left(\frac{-n}{2\sigma^2}(\mu - \bar{y})^2\right) \exp\left(\frac{-1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma_\mu}(\mu - \bar{\mu})^2\right)
\end{aligned}
$$

$$
\implies \mu \mid y \sim \mathcal{N}\left(\bar{\mu}, \sigma_\mu^2\right)
$$

Conjugate Prior

$$\mu \mid y \sim \mathcal{N}\left(\bar{\mu}, \sigma_{\mu}^{2}\right), \text{with}$$

Conjugate Prior

$$\mu \mid y \sim \mathcal{N}\left(\bar{\mu}, \sigma_\mu^2\right), \text{with}$$

$$\sigma_\mu = \left(\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}\right)^{-1}$$

Conjugate Prior

$$\mu \mid y \sim \mathcal{N}\left(\bar{\mu}, \sigma_\mu^2\right), \text{with}$$

$$\sigma_\mu = \left(\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}\right)^{-1}$$

$$\bar{\mu} = \sigma_\mu \left(\frac{1}{\sigma^2/n}\boldsymbol{\bar{y}} + \frac{1}{\sigma_0^2}\boldsymbol{\mu_0}\right)$$

Conjugate Prior

$$
\begin{aligned}
\mu \mid y &\sim \mathcal{N}\left(\bar{\mu}, \sigma_\mu^2\right), \text{with} \\
\sigma_\mu &= \left(\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}\right)^{-1} \\
\bar{\mu} &= \sigma_\mu \left(\frac{1}{\sigma^2/n}\boldsymbol{\bar{y}} + \frac{1}{\sigma_0^2}\boldsymbol{\mu_0}\right) \\
&= \alpha\boldsymbol{\bar{y}} + (1-\alpha)\boldsymbol{\mu_0}
\end{aligned}
$$

A normal model with known variance and no features, really?

🙃

Sampling from the Posterior

Object of interest: $\theta \mid \text{data}$

Sampling from the Posterior

Object of interest: $\theta \mid \text{data}$

Quantity of interest: $\mathbb{E}\left[\, h(\theta) \mid \text{data} \,\right] =: \mathbb{E}_\theta[h]$

Sampling from the Posterior

Object of interest: $\theta \mid \text{data}$

Quantity of interest: $\mathbb{E}\left[\, h(\theta) \mid \text{data} \,\right] =: \mathbb{E}_\theta[h]$

Estimation: Let $\theta^{(1)}, \ldots, \theta^{(n)} \overset{\text{iid}}{\sim} p(\theta \mid \text{data})$, then

$$\frac{1}{\sqrt{n}} \left(\sum_i h(\theta^{(i)}) - \mathbb{E}_\theta[h] \right) \xrightarrow{\text{d}} \mathcal{N}\left(0, \omega\right)$$

Sampling from the Posterior

But how do we sample from $p(\theta \mid \text{data})$?

Sampling from the Posterior

But how do we sample from $p(\theta \mid \text{data})$?
Problems: $p(\theta \mid \text{data})$ might be

Sampling from the Posterior

But how do we sample from $p(\theta \mid \text{data})$?

Problems: $p(\theta \mid \text{data})$ might be

(i.) of unkown form

Sampling from the Posterior

But how do we sample from $p(\theta \mid \text{data})$?

Problems: $p(\theta \mid \text{data})$ might be

(i.) of unkown form

(ii.) very complex

Sampling from the Posterior

But how do we sample from $p(\theta \mid \text{data})$?

Problems: $p(\theta \mid \text{data})$ might be

(i.) of unkown form

(ii.) very complex

(iii.) only known up to an integration const.

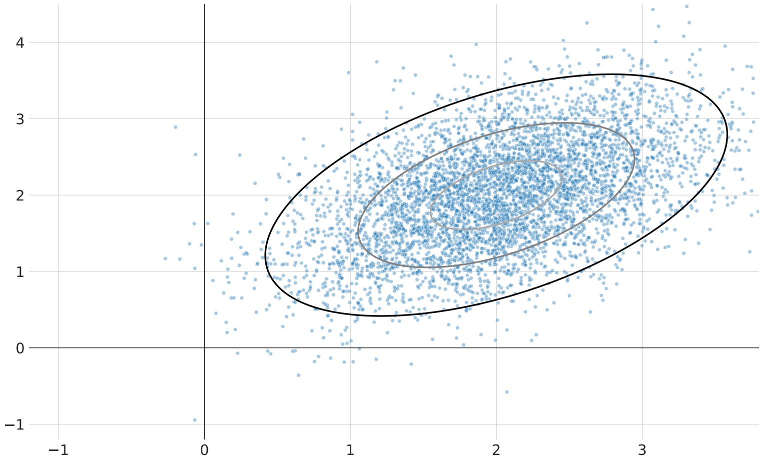
Markov Chain Monte Carlo

Algorithm Metropolis-Hastings (1953, 1970)

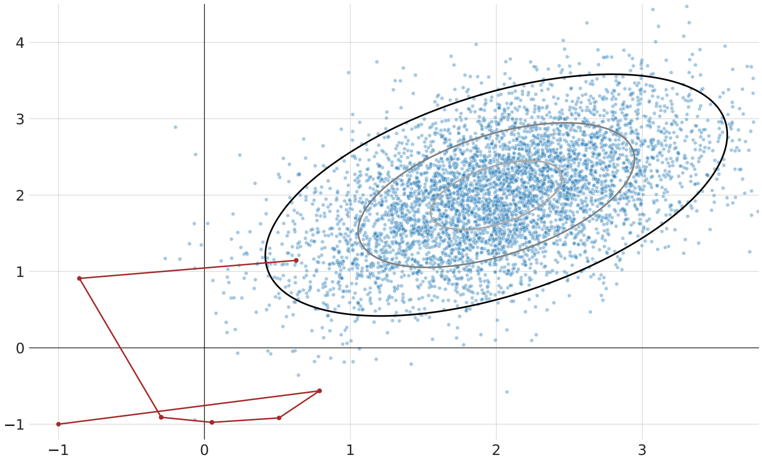
Input: $(\pi, q, T) =$ (target, proposal, no. of samples)

initialize x_0 in $\operatorname{supp} q$
for $t = 0, \ldots, T$ **do**
    candidate: $y \sim q(\cdot \mid x_t)$
    acceptance prob.: $\mathcal{A} \leftarrow \min\left\{\frac{\pi(y)}{\pi(x_t)} \frac{q(y \mid x_t)}{q(x_t \mid y)}, 1\right\}$
    update: $x_{t+1} \leftarrow \begin{cases} y & \text{,with prob. } \mathcal{A} \\ x_t & \text{,with remaining prob.} \end{cases}$
return $\{x_t : t = 1, \ldots, T\}$

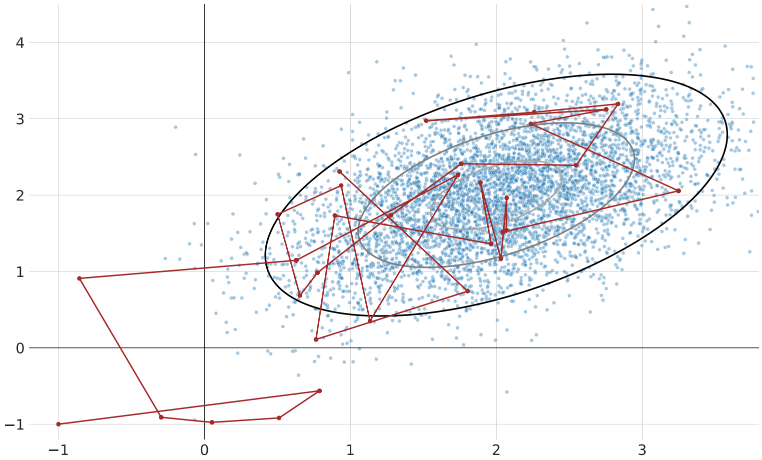
Markov Chain Monte Carlo



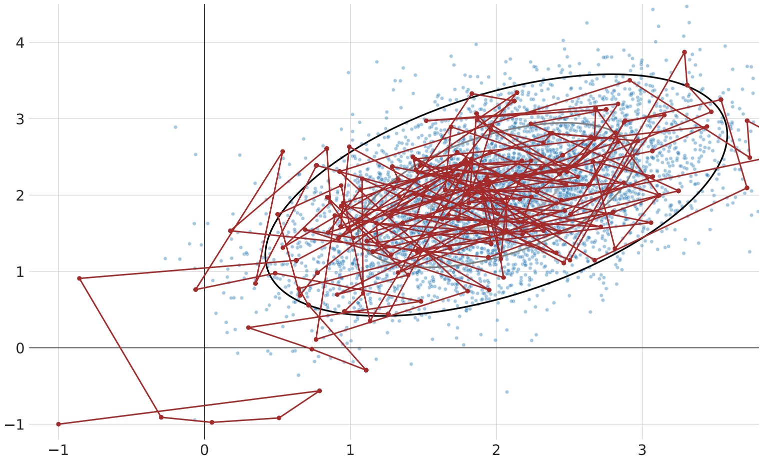
Markov Chain Monte Carlo



Markov Chain Monte Carlo



Markov Chain Monte Carlo



Hierarchical Models

Structure of HM - Setup

Hierarchical Data:

Structure of HM - Setup

Hierarchical Data:

Individual Level: (y_i, x_i) for $i = 1, \ldots, n$

Structure of HM - Setup

Hierarchical Data:

- *Individual Level:* (y_i, x_i) for $i = 1, \ldots, n$
- *Group Level:* u_j for $j = 1, \ldots, J$

Structure of HM - Setup

Hierarchical Data:

Individual Level: (y_i, x_i) for $i = 1, \ldots, n$

Group Level: u_j for $j = 1, \ldots, J$

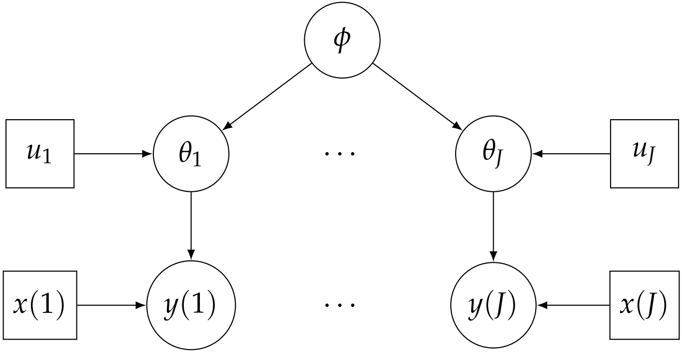
Example:

Test Outcome: y_i

Parental Income: x_i

Num. of Teachers: u_j

Structure of HM



The Prior Revisited

Before:

Model: $p(\text{data} \mid \theta)$

Prior: $p(\theta)$

The Prior Revisited

Before:

Model: $p(\text{data} \mid \theta)$

Prior: $p(\theta)$

Now:

Model: $p(\text{data} \mid \theta, \phi) = p(\text{data} \mid \theta)$

Prior: $p(\theta \mid \phi)$

Hyperprior: $p(\phi)$

The Posterior Revisited

Posterior:

$$p(\theta, \phi \mid \text{data}) \propto p(\text{data} \mid \theta) p(\theta, \phi)$$

The Posterior Revisited

Posterior:

$$\begin{aligned} p(\theta, \phi \mid \text{data}) &\propto p(\text{data} \mid \theta) p(\theta, \phi) \\ &\propto p(\text{data} \mid \theta) p(\theta \mid \phi) p(\phi) \end{aligned}$$

The Posterior Revisited

Posterior:

$$
\begin{aligned}
p(\theta, \phi \mid \text{data}) &\propto p(\text{data} \mid \theta) p(\theta, \phi) \\
&\propto p(\text{data} \mid \theta) p(\theta \mid \phi) p(\phi) \\
p(\phi \mid \text{data}) &\propto \int p(\theta, \phi \mid \text{data}) \mathrm{d}\theta
\end{aligned}
$$

Varying Slopes, Varying Intercepts

Setup: Individual i in group j

Varying Slopes, Varying Intercepts

Setup: Individual i in group j

Individual Level: $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$

Varying Slopes, Varying Intercepts

Setup: Individual i in group j

Individual Level: $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$

Group Level: $\begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} = \gamma_0 + \gamma u_j + \eta_j$

Varying Slopes, Varying Intercepts

Setup: Individual i in group j

Individual Level: $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$

Group Level: $\begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} = \gamma_0 + \gamma u_j + \eta_j$

Priors on: $\gamma_0, \gamma, \epsilon_i, \eta_j$

Varying Slopes, Varying Intercepts

Setup: Individual i in group j

Individual Level: $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$

Group Level: $\begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} = \gamma_0 + \gamma u_j + \eta_j$

Priors on: $\gamma_0, \gamma, \epsilon_i, \eta_j$

Monte Carlo Study

Stan

What is Stan? C++ package (fast run times)

Stan

What is Stan? C++ package (fast run times)

Use cases: Bayesian/Maximum Likelihood estimation of statistical models

Stan

What is Stan? C++ package (fast run times)

Use cases: Bayesian/Maximum Likelihood estimation of statistical models

Interfaces: Pystan, Rstan,Stan.jl,...

Bayesian Models with Stan

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i \,,$$
$$\epsilon \sim \mathcal{N}\left(0, 1\right)$$

Bayesian Models with Stan

$$\begin{aligned} y_i &= \alpha + \beta_{j[i]} x_i + \epsilon_i\,, \\ \epsilon &\sim \mathcal{N}(0, 1) \end{aligned}$$

$$\begin{aligned} \beta_j &= \gamma_0 + \gamma_1 u_j + \eta_j\,, \\ \eta &\sim \mathcal{N}(0, 1) \end{aligned}$$

Bayesian Models with Stan

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i \,,$$
$$\epsilon \sim \mathcal{N}(0, 1)$$

$$\beta_j = \gamma_0 + \gamma_1 u_j + \eta_j \,,$$
$$\eta \sim \mathcal{N}(0, 1)$$

```
data {
vector[N] y;
vector[N] x;
vector[N] u;
int<lower=0> J;
int<lower=0> N;
int<lower=1,upper=J>
          group[N];
}
```

Bayesian Models with Stan

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i \,,$$
$$\epsilon \sim \mathcal{N}(0, 1)$$

$$\beta_j = \gamma_0 + \gamma_1 u_j + \eta_j \,,$$
$$\eta \sim \mathcal{N}(0, 1)$$

```
parameter {
real alpha;
real gamma_0;
real gamma_1;
vector[J] eta_b;
real<lower=0> sigma_b;
real<lower=0> sigma_y;
}
```

Bayesian Models with Stan

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i \,,$$
$$\epsilon \sim \mathcal{N}(0, 1)$$

$$\beta_j = \gamma_0 + \gamma_1 u_j + \eta_j \,,$$
$$\eta \sim \mathcal{N}(0, 1)$$

```
# model
for (i in 1:N) {
  beta[i] = gamma_0 +
       u[i] * gamma_1 +
         eta[group[i]]

  y_hat[i] = alpha +
       x[i] * beta[i];
}
y ~ normal(
       y_hat, sigma_y);
```

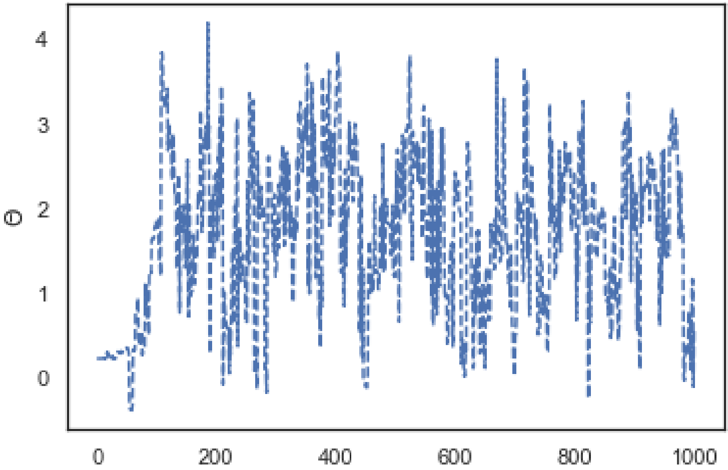
Bayesian Models with Stan

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i\,,$$
$$\epsilon \sim \mathcal{N}(0, 1)$$

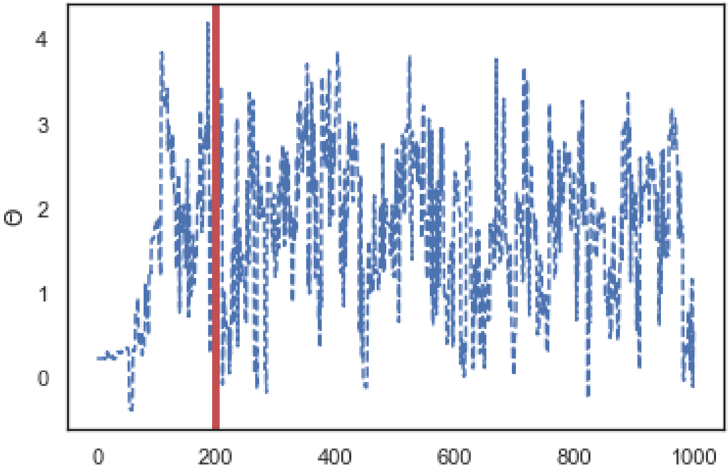
$$\beta_j = \gamma_0 + \gamma_1 u_j + \eta_j\,,$$
$$\eta \sim \mathcal{N}(0, 1)$$

```
# priors
gamma_0 ~ normal(1, 1);
gamma_1 ~ normal(1, 1);
eta ~ normal(0, sigma_b);
sigma_y ~ cauchy(0, 5);
sigma_b ~ cauchy(0, 5);
```

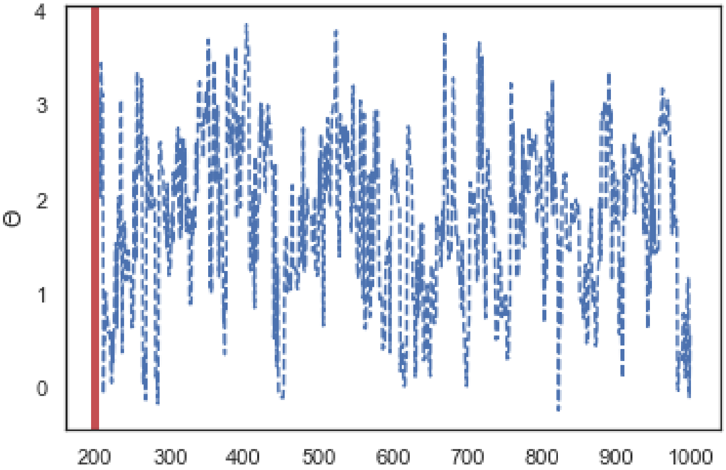
How can we be sure that we sample from the right distribution?



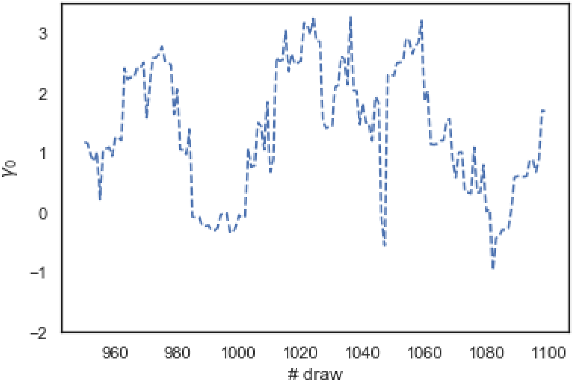
How can we be sure that we sample from the right distribution?



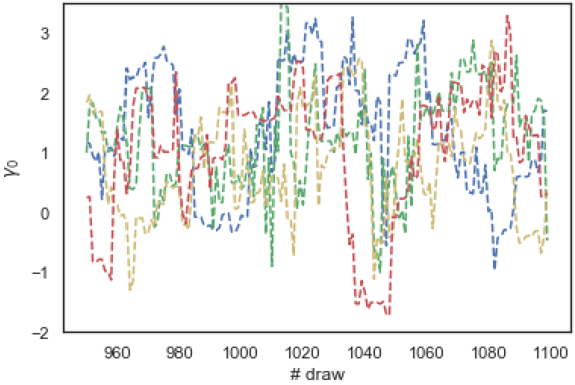
How can we be sure that we sample from the right distribution?



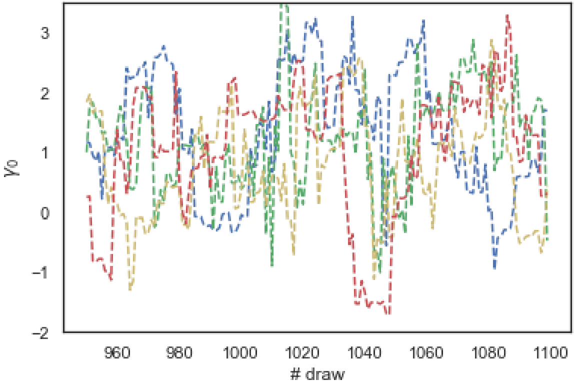
Monitoring Convergence



Monitoring Convergence



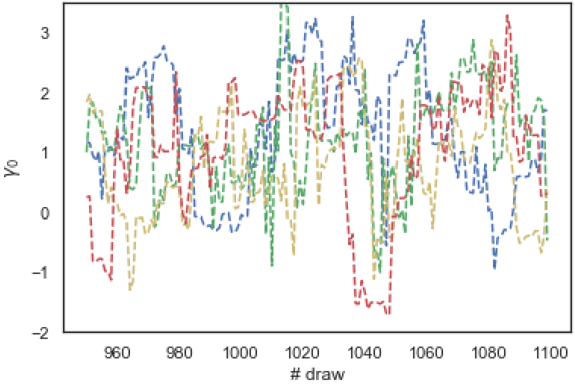
Monitoring Convergence



Variance of a single chain:

$$s_m^2 = \frac{1}{N-1}\sum_{n=1}^{N}(\theta_m^{(n)} - \overline{\theta}_m)^2$$

Monitoring Convergence



Average within chain variance:

$$W = \frac{1}{M} \sum_{m=1}^{M} s_m^2$$

Brooks and Gelman convergence criterium

Average Variance between chains:

$$B/N = \frac{1}{M-1}\sum_{m=1}^{M}(\overline{\theta}_m - \overline{\theta})^2$$

Brooks and Gelman convergence criterium

Average Variance between chains:

$$B/N = \frac{1}{M-1}\sum_{m=1}^{M}(\overline{\theta}_m - \overline{\theta})^2$$

Total Variance:

$$\widehat{\mathrm{Var}}^{+}(\theta \mid y) = \frac{N-1}{N}W + \frac{1}{N}B$$

Monitoring Convergence

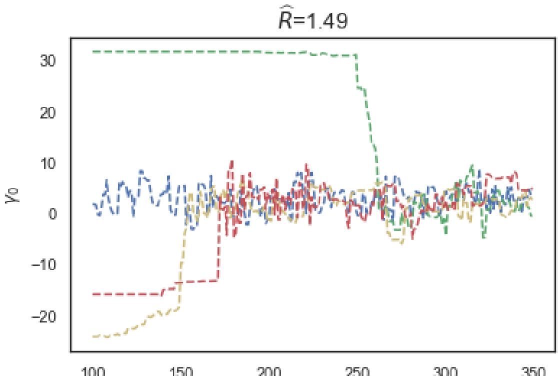
Scale Reducing Factor:

$$\widehat{R} = \sqrt{\frac{\widehat{\text{Var}}^{+}(\theta \mid y)}{W}}$$

Monitoring Convergence

Scale Reducing Factor:

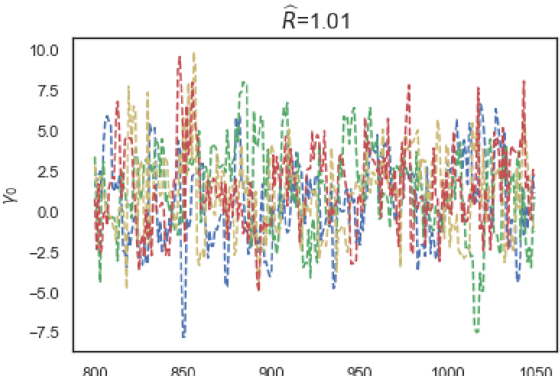
$$\hat{R} = \sqrt{\frac{\widehat{\mathrm{Var}}^{+}(\theta \mid y)}{W}}$$



Monitoring Convergence

Scale Reducing Factor:

$$\widehat{R} = \sqrt{\frac{\widehat{\mathrm{Var}}^{+}(\theta \mid y)}{W}}$$



Prior Design

Good Prior: $\gamma_0 \sim \mathcal{N}(1,1), \gamma_1 \sim \mathcal{N}(1,1)$

Prior Design

Good Prior: $\gamma_0 \sim \mathcal{N}(1,1), \gamma_1 \sim \mathcal{N}(1,1)$

Bad Prior: $\gamma_0 \sim \mathcal{N}(2,1), \gamma_1 \sim \mathcal{N}(2,1)$

Prior Design

Good Prior: $\gamma_0 \sim \mathcal{N}(1,1), \gamma_1 \sim \mathcal{N}(1,1)$

Bad Prior: $\gamma_0 \sim \mathcal{N}(2,1), \gamma_1 \sim \mathcal{N}(2,1)$

Weak, Bad Prior: $\gamma_0 \sim \mathcal{N}(2,3), \gamma_1 \sim \mathcal{N}(2,3)$

Prior Design

Good Prior: $\gamma_0 \sim \mathcal{N}(1, 1), \gamma_1 \sim \mathcal{N}(1, 1)$

Bad Prior: $\gamma_0 \sim \mathcal{N}(2, 1), \gamma_1 \sim \mathcal{N}(2, 1)$

Weak, Bad Prior: $\gamma_0 \sim \mathcal{N}(2, 3), \gamma_1 \sim \mathcal{N}(2, 3)$

Flat Prior: $\gamma_0 \sim \mathcal{U}(-\infty, \infty), \gamma_1 \sim \mathcal{U}(-\infty, \infty)$

Prior Design

Good Prior: $\gamma_0 \sim \mathcal{N}(1,1), \gamma_1 \sim \mathcal{N}(1,1)$

Bad Prior: $\gamma_0 \sim \mathcal{N}(2,1), \gamma_1 \sim \mathcal{N}(2,1)$

Weak, Bad Prior: $\gamma_0 \sim \mathcal{N}(2,3), \gamma_1 \sim \mathcal{N}(2,3)$

Flat Prior: $\gamma_0 \sim \mathcal{U}(-\infty,\infty), \gamma_1 \sim \mathcal{U}(-\infty,\infty)$

In all models: $\sigma_y, \sigma_b \sim \text{half-Cauchy}(0,5)$

Posterior Distribution - good prior

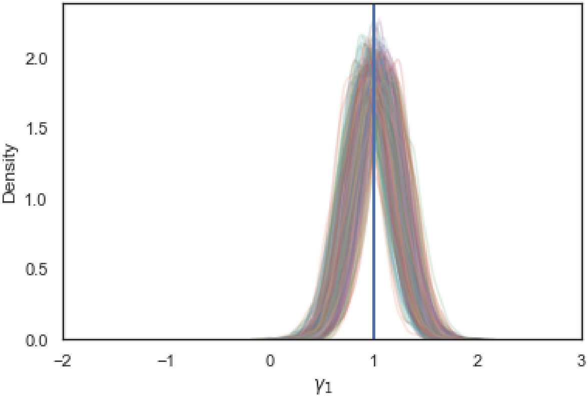


Figure: Posterior Draws of γ_1 with N=200, J=10 and 300 simulations

What happense if we decrease the number of levels J?

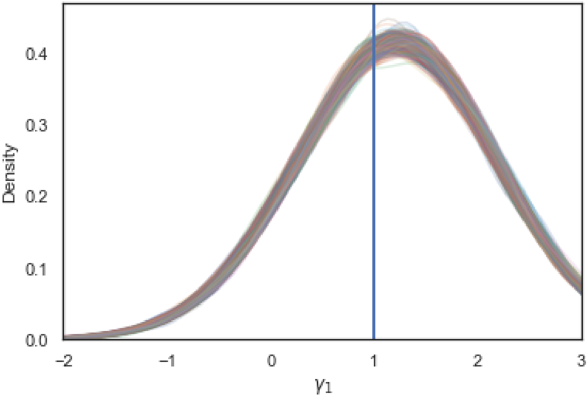


Figure: Posterior Draws of γ_1 with N=50, J=5 and 300 simulations

Posterior Distribution - bad prior

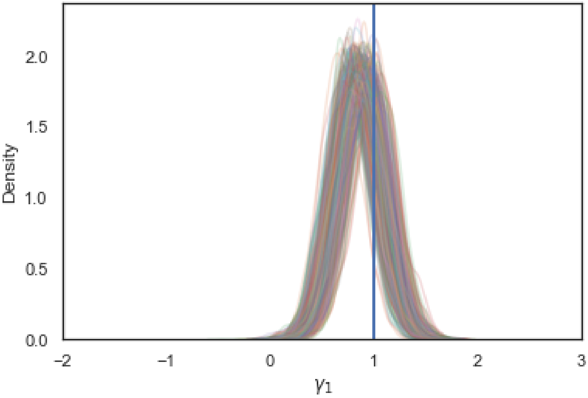


Figure: Posterior Draws of γ_1 with N=200, J=10 and 300 simulations

Posterior Distribution - weak, bad prior

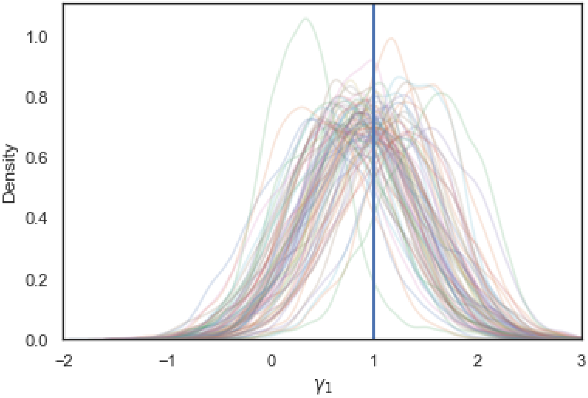


Figure: Posterior Draws of γ_1 with N=200, J=10 and 150 simulations

Not a good idea: Uniform Prior

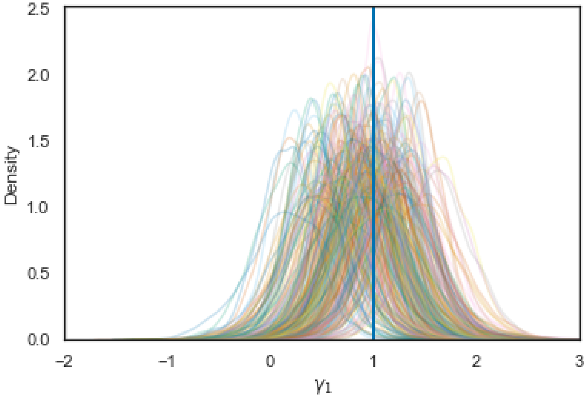


Figure: Posterior Draws of γ_1 with N=200, J=10 and 300 simulations

Increase sample size dramatically

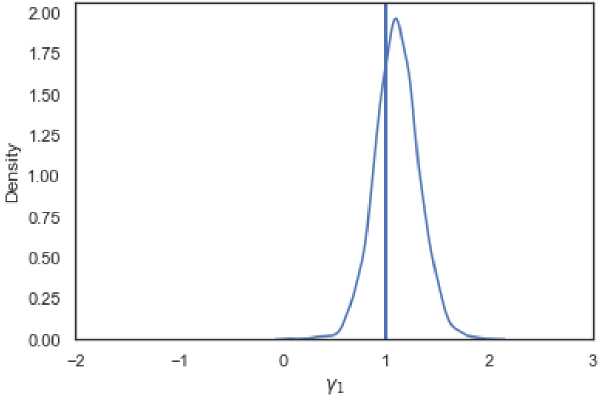


Figure: Single posterior draw for model with wrong prior and N=500, J=50

Application

The Data

Description: General Certificate of Secondary Education (GCSE) exam scores of 1,905 students from 73 schools in England on a science subject

Variables of interest: school identifier, student identifier, gender, total score on written paper and total score of course work.

ML and Bayesian Approach

Application

Comparison

The model: Varying intercept and slope model with a single predictor

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij}, \tag{1}$$

$$\alpha_j = \mu_\alpha + u_j, \tag{2}$$

$$\beta_j = \mu_\beta + v_j, \tag{3}$$

$$y_{ij} = \mu_\alpha + \mu_\beta x_{ij} + u_j + v_j x_{ij} + \epsilon_{ij} \tag{4}$$

Comparison

Maximum Likelihood (ML) Estimation

Package: *lmer*

The lmer Package: combines of ML estimation of model parameters and empirical Bayes (EB) predictions of the varying intercepts and/or slopes resulting in the Best Linear Unbiased Predictions (BLUPs) of the model parameters.

Why lmer?? allows for comparison between parameter estimates

Comparison

Bayesian Estimation:

Package: *stan*

Priors: weekly informative normally distributed priors for hyperparameters

Results

Dependent variable: Course test score		
	ML	Bayes
A. Random effects		
Intercept	–	–
	(10.146)	(10.249)
Female	–	–
	(6.924)	(7.099)
B. Fixed effects		
Intercept	69.425	69.413
	(1.352)	(1.287)
Female	7.128	7.132
	(1.131)	(1.165)
N		
Students	1725	1725
Schools	73	73

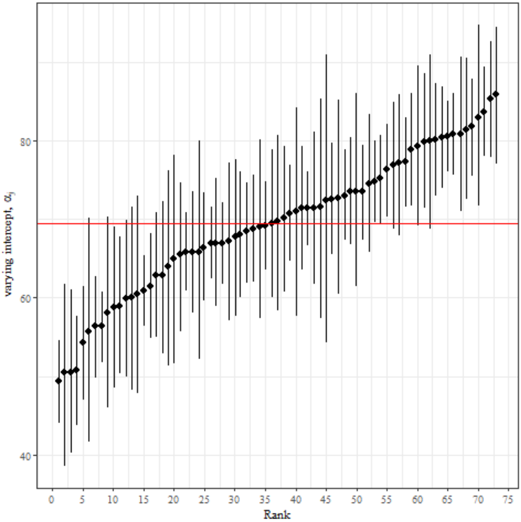
(i.) point estimates almost the same

(ii.) Bayes standard deviations for random effects may be higher because ML does not take into account group level variance

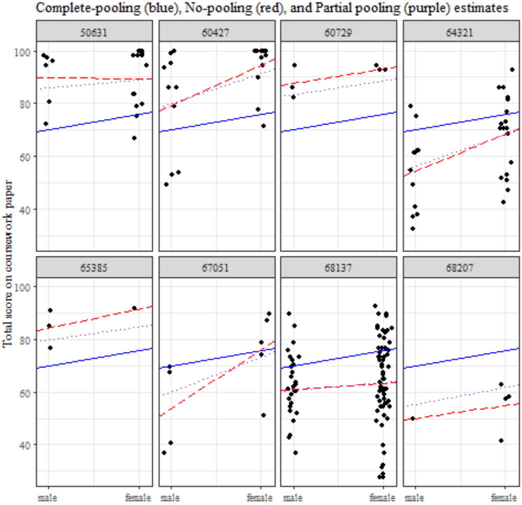
Bayesian Approach: Further Analysis

Application

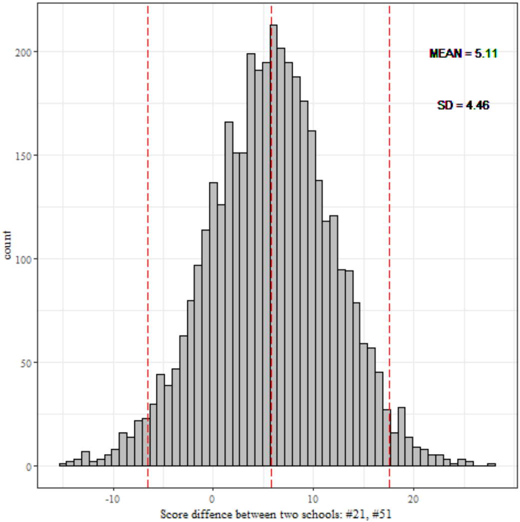
Posterior distribution ranking:



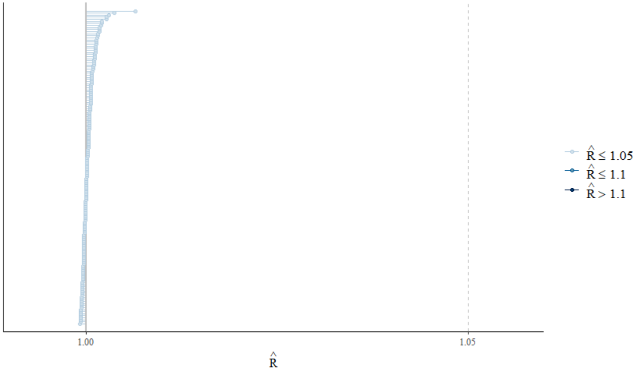
School specific regression lines and pooling:



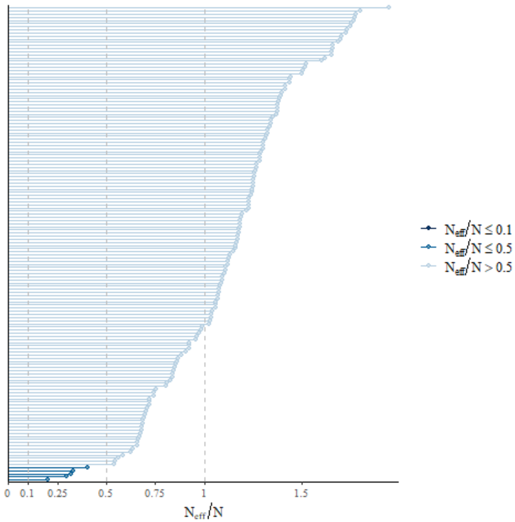
Making comparisons between individual schools:



Convergence



Convergence



https://github.com/timmens/
bayesian-hierarchical-models

http://mfviz.com/hierarchical-models/