A Comparison of Two MCMC Algorithms for Hierarchical Mixture Models

Russell Almond
Florida State University
College of Education
Educational Psychology and Learning Systems
ralmond@fsu.edu
Cognitive Basis

• Multiple cognitive processes involved in writing
• Different processes take different amounts of time
  – Transcription should be fast
  – Planning should be long
• Certain event types should be more common than others
Multi-Process Model of Writing

Mixture Models

- Mixture
- .7*component1 N(0,1)
- .3*component2 N(1,4)
Random Data

Mix = .5, MeanDiff=1, SDratio = 4

Mix = .3, MeanDiff=1, SDratio = 4

Mix = .7, MeanDiff=1, SDratio = 4

N = 500  Bandwidth = 0.4852

N = 500  Bandwidth = 0.4088

N = 500  Bandwidth = 0.4281

N = 500  Bandwidth = 0.7746

N = 500  Bandwidth = 0.6293

N = 500  Bandwidth = 0.7201

N = 500  Bandwidth = 0.3583

N = 500  Bandwidth = 0.351

N = 500  Bandwidth = 0.4217
Within-Word Pauses

Log Within-Word Break Density

- Density

<table>
<thead>
<tr>
<th>N</th>
<th>Bandwidth</th>
<th>Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>859</td>
<td>0.1274</td>
<td>348531460</td>
</tr>
<tr>
<td>190</td>
<td>0.2357</td>
<td>418630983</td>
</tr>
<tr>
<td>404</td>
<td>0.1105</td>
<td>638220089</td>
</tr>
<tr>
<td>218</td>
<td>0.1965</td>
<td>418631122</td>
</tr>
<tr>
<td>424</td>
<td>0.1191</td>
<td>628420501</td>
</tr>
<tr>
<td>618</td>
<td>0.1116</td>
<td>258331066</td>
</tr>
<tr>
<td>272</td>
<td>0.1843</td>
<td>568420801</td>
</tr>
<tr>
<td>354</td>
<td>0.1358</td>
<td>398340580</td>
</tr>
<tr>
<td>740</td>
<td>0.1471</td>
<td>388440137</td>
</tr>
</tbody>
</table>

NCME 2014
Mixture of Lognormals

• Log Pause Time $Y_{ij} = \log(X_{ij})$
  – Student (Level-2 unit) $i=1,\ldots,I$
  – Pause (Level-1 unit) $j=1,\ldots,J_i$

• $Z_{ij} \sim \text{cat}(\pi_{i1},\ldots,\pi_{iK})$ is an indicator for which of K components the $j^{th}$ pause for the $i^{th}$ student is in

• $Y_{ij} \mid Z_{ij} = k \sim N(\mu_{ik}, \tau_{ik})$
Mixture Model

Component $k$

Essay $i$

Event $j$

$\pi_i$

$Z_{ij}$

$Y_{ij}$

$Y^{*}_{ijk}$

$\tau_{ik}$

$\mu_{ik}$
Mixture Model Problems

- If $\pi_{ik}$ is small for some $k$, then category disappears
- If $\tau_{ik}$ is small for some $k$, then category becomes degenerate
- If $\mu_{ik} = \mu_{ik'}$ and $\tau_{ik} = \tau_{ik'}$, then really only have $K-1$ categories
Labeling Components

• If we swap the labels on component $k$ and $k'$, the likelihood is identical
• Likelihood is multimodal
• Often put a restriction on the components:
  \[ \mu_{i1} < \mu_{i2} < \ldots < \mu_{iK} \]
• Frühwirth-Schattner (2001) notes that when doing MCMC, better to let the chains run freely across the modes and sort out post-hoc
• Sorting needs to be done before normal MCMC convergence tests, or parameter estimation
Key Question

• How many components?
• Theory: each component corresponds to a different combination of cognitive processes
• Rare components might not be identifiable from data

Hierarchical models which allow partial pooling across Level-2 (students) might help answer these questions
Hierarchical Mixture Model

Essay $i$ → $\pi_i$ → $Z_{ij}$ → $y_{ij}$ → Event $j$ → $Y^{*ijk}$ → Component $k$ → $\tau_{0k}$ → $\gamma_{0k}$ → $\mu_{0k}$ → $\beta_{0k}$

$\alpha$
Problems with hierarchical models

• If $\gamma_{0k}, \beta_{0k} \rightarrow 0$ we get complete pooling
• If $\gamma_{0k}, \beta_{0k} \rightarrow \infty$ we get no pooling
• Something similar happens with
• Need prior distributions that bound us away from those values.
• $\log(\tau_{0k}), \log(\beta_{0k}), \log(\gamma_{0k}) \sim N(0,1)$
Two MCMC packages

**JAGS**
- Random Walk Metropolis, or Gibbs sampling
- Has a special proposal for normal mixtures
- Can extend a run if insufficient length
- Can select which parameters to monitor

**Stan**
- Hamiltonian Monte Carlo
  - Cycles take longer
  - Less autocorrelation
- Cannot extend runs
- Must monitor all parameters
Add redundant parameters to make MCMC faster

• \( \mu_{ik} = \mu_{0k} + \theta_i \beta_{0k} \)
  - \( \theta_i \sim N(0,1) \)

• \( \log(\tau_{ik}) = \log(\tau_{0k}) + \eta_i \gamma_{0k} \)
  - \( \eta_i \sim N(0,1) \)

• \( \alpha_k = \alpha_{0k} \alpha_N \)
  - \( \alpha_0 \sim \text{Dirichlet}(\alpha_{0m}) \)
  - \( \alpha_N \sim \chi^2(2I) \)
Initial Values

1. Run EM on each student’s data set to get student-level (Level 1) parameters
   – If EM does not converge, set parameters to NA
2. Calculate cross-student (Level 2) as summary statistics of Level 1 parameters
3. Impute means for missing Level 1 parameters

*Repeat with subsets of the data for variety in multiple chains*
Simulated Data Experiment

• Run initial value routine on real data for K=2,3,4
• Generate data from the model using these parameters
• Fit models with K’=2,3,4 to the data from true K=2,3,4 distributions in both JAGS (RWM) and Stan (HMC)
• Results shown for K=2, K’=2
Results (Mostly K=2, K’=2)

• All results (http://pluto.coe.fsu.edu/mcmc-hierMM)
• Deviance/Log Posterior—Good Mixing
• $\mu_{01}$ (average mean of first component)—Slow mixing in JAGS
• $\alpha_{01}$ (average probability of first component)—Poor convergence, switching in Stan
• $\gamma_{01}$ (s.d. of log precisions for first component)—Poor convergence, multiple modes?
Deviance (JAGS)/Log Posterior (Stan)
Rhat (JAGS) = 1   effective sample size (JAGS) = 7992
Rhat (Stan) = 1.08 effective sample size (Stan) = 7388

JAGS

Stan (unconstrained model)
Rhat (JAGS) = 1.14      effective sample size (JAGS) = 143
Rhat (Stan) = 1.15      effective sample size (Stan) = 3680

N = 15000  Bandwidth = 0.0225

N = 15000  Bandwidth = 0.02464
\[ \alpha_{01} \]

- Rhat (JAGS) = 1.3
- effective sample size (JAGS) = 1596
- Rhat (Stan) = 1.57
- effective sample size (Stan) = 5207
JAGS

\[ \gamma_{01} \]

Rhat (JAGS) = 1.73
Rhat (Stan) = 1.73

effective sample size (JAGS) = 1003
effective sample size (Stan) = 3085

Stan (unconstrained model)
Other Results

• Convergence is still an issue
  – MCMC finds multiple modes, where EM (without restarts) typically finds only one
• JAGS (RWM) was about 3 times faster than Stan (HMC)
  – Monte Carlo se in Stan about 5 times smaller (25 time larger effective sample size)
• JAGS is still easier to use than Stan
• Could not use WAIC statistic to recover K
  – Was the same for K’=2,3,4
Implication for Student Essays

• Does not seem to recover “rare components”
• Does not offer big advantage over simpler no pooling non-hierarchical model
• Ignores serial dependence in data
  – Hidden Markov model might be better than straight mixture
Try it yourself

- [http://pluto.coe.fsu.edu/mcmc-hierMM/](http://pluto.coe.fsu.edu/mcmc-hierMM/)
  - Complete source code (R, Stan, JAGS) including data generation
  - Sample data sets
  - Output from all of my test runs, including trace plots for all parameters
  - Slides from this talk.