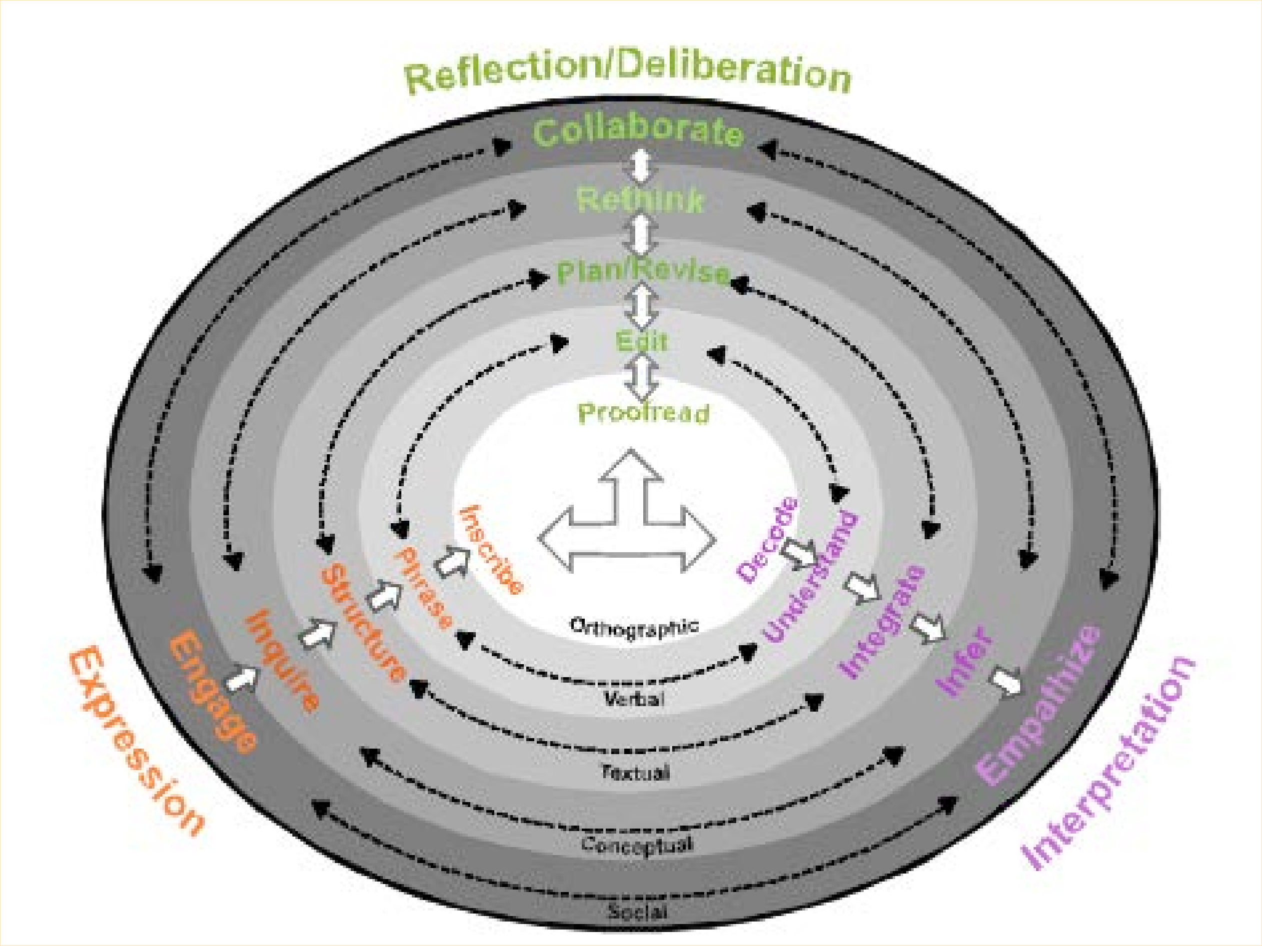
A Comparison of Two MCMC Algorithms for Hierarchical Mixture Models

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

Cognitive Basis

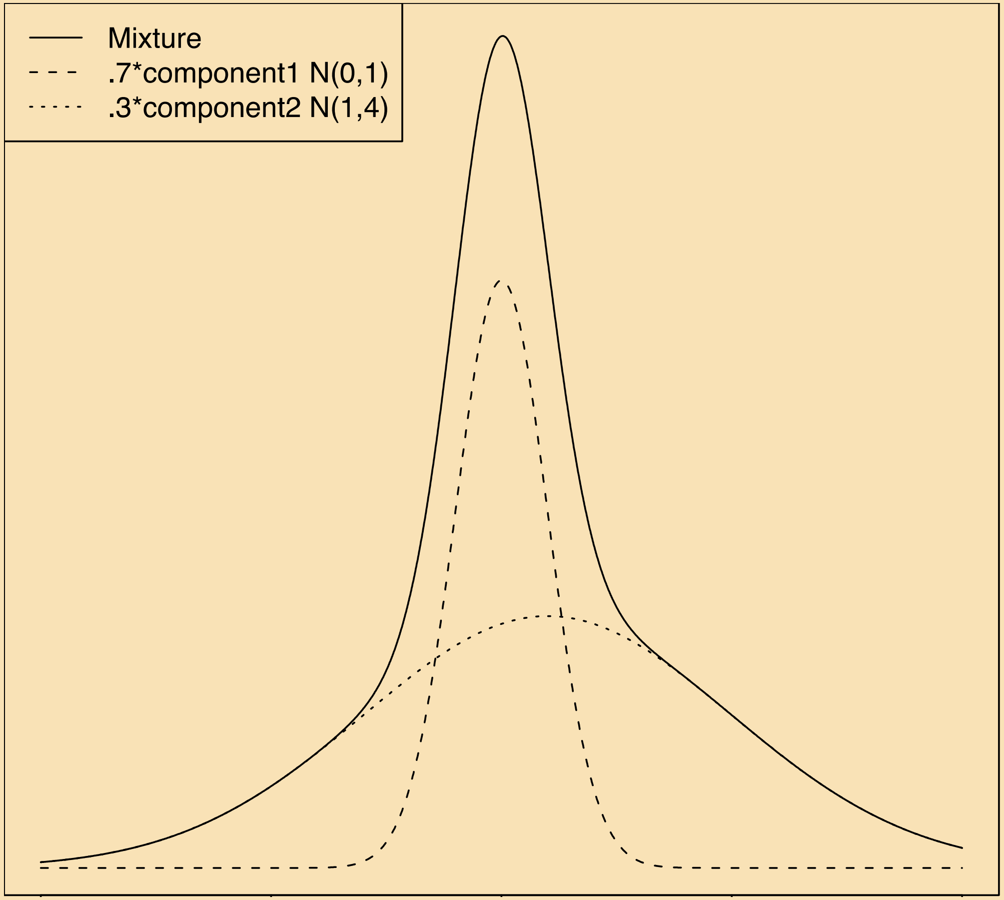
- Multiple cognitive processes involved in writing
- Different processes take different amounts of time
  - Transcription should be fast
  - Planning should be long
- Certain event types should be more common than others

Multi-Process Model of Writing

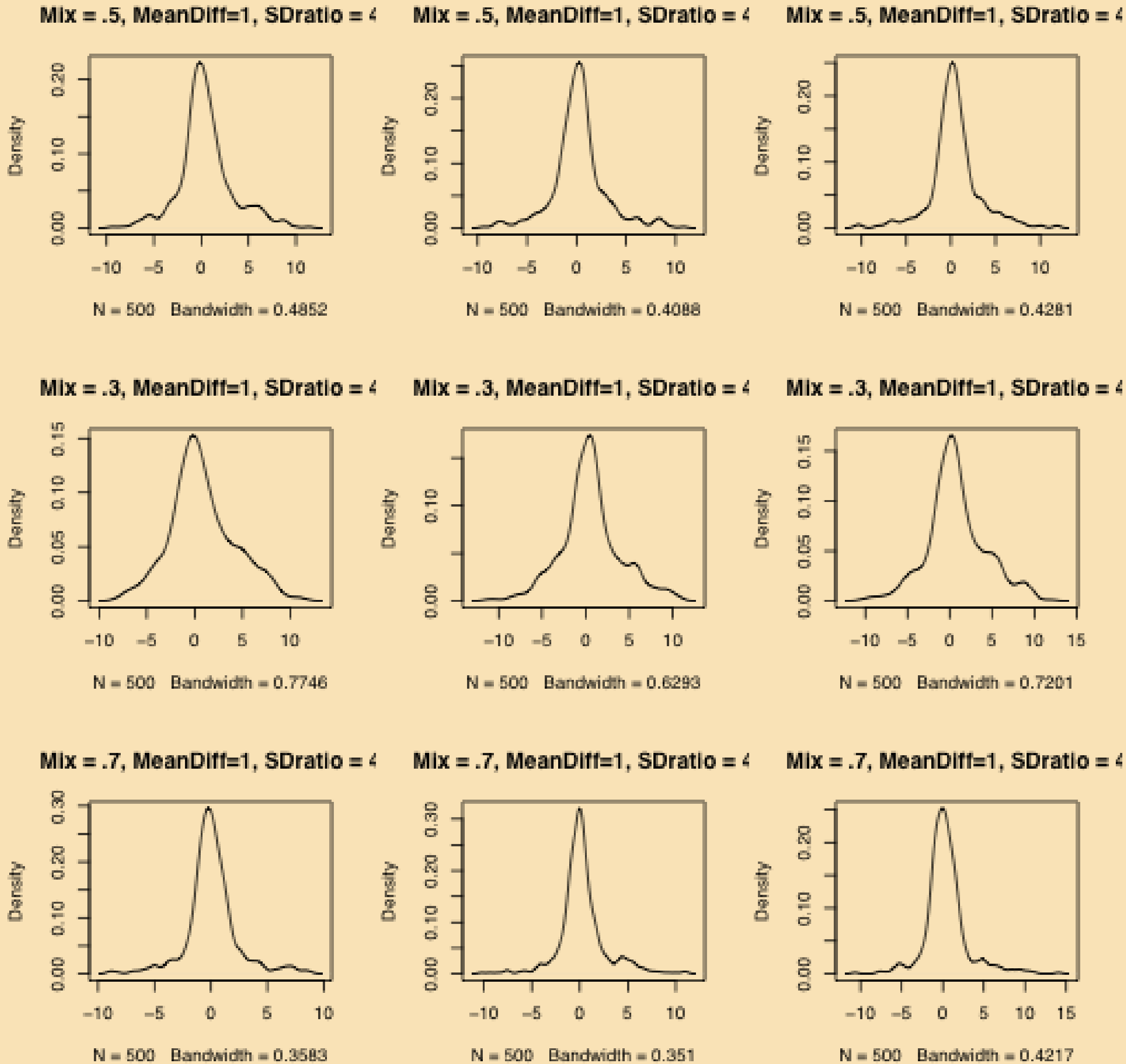


Deane (2009) Model of writing.

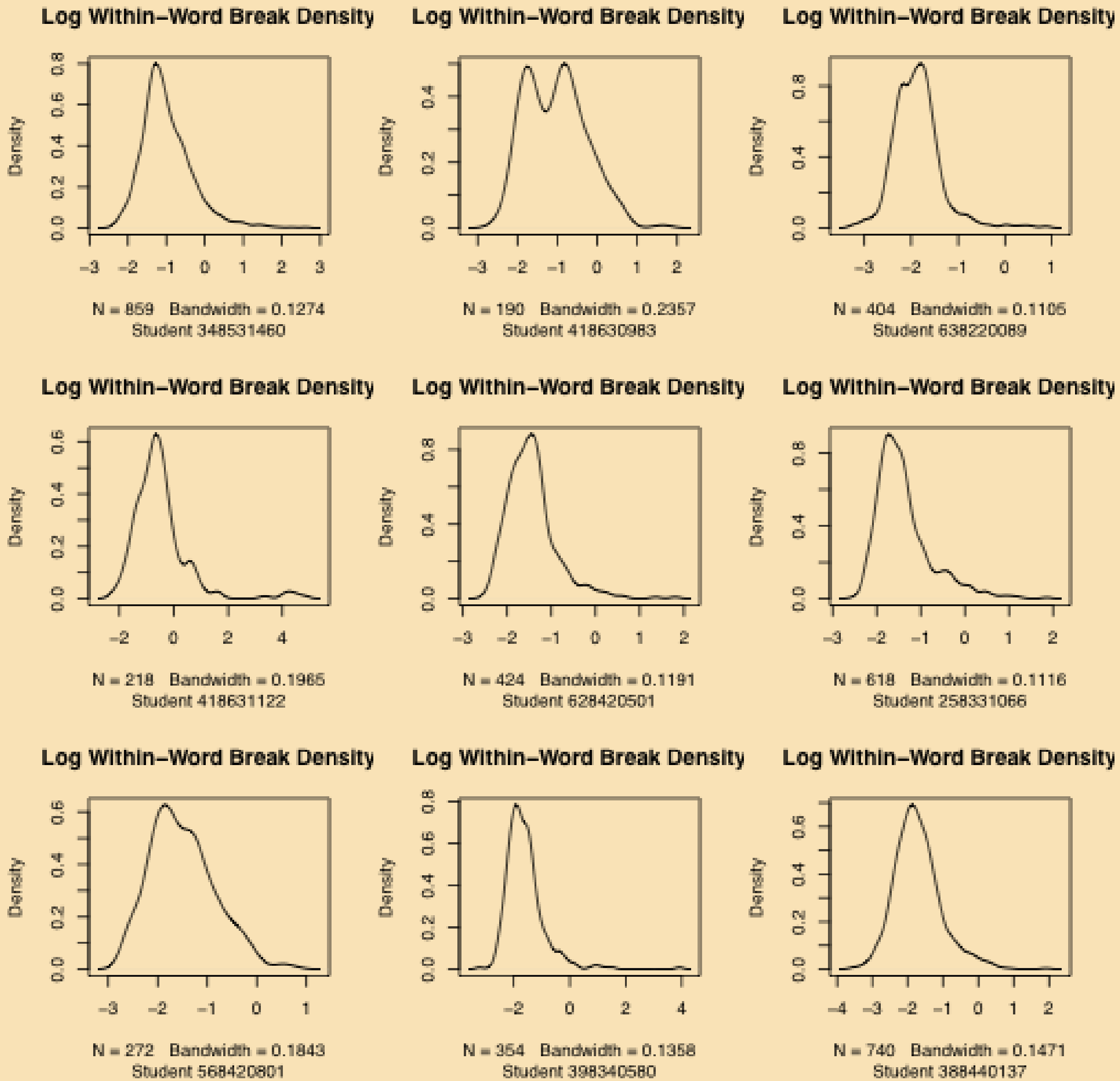
Mixture Models



Random Data


Mix = .5, MeanDiff=1, SDratio = 4
Density
N = 500 Bandwidth = 0.4852
Mix = .5, MeanDiff=1, SDratio = 4
N = 500 Bandwidth = 0.4088
Mix = .5, MeanDiff=1, SDratio = 4
N = 500 Bandwidth = 0.4281
Mix = .3, MeanDiff=1, SDratio = 4
N = 500 Bandwidth = 0.7746
Mix = .3, MeanDiff=1, SDratio = 4
N = 500 Bandwidth = 0.6293
Mix = .3, MeanDiff=1, SDratio = 4
N = 500 Bandwidth = 0.7201
Mix = .7, MeanDiff=1, SDratio = 4
N = 500 Bandwidth = 0.3583
Mix = .7, MeanDiff=1, SDratio = 4
N = 500 Bandwidth = 0.351
Mix = .7, MeanDiff=1, SDratio = 4
N = 500 Bandwidth = 0.4217

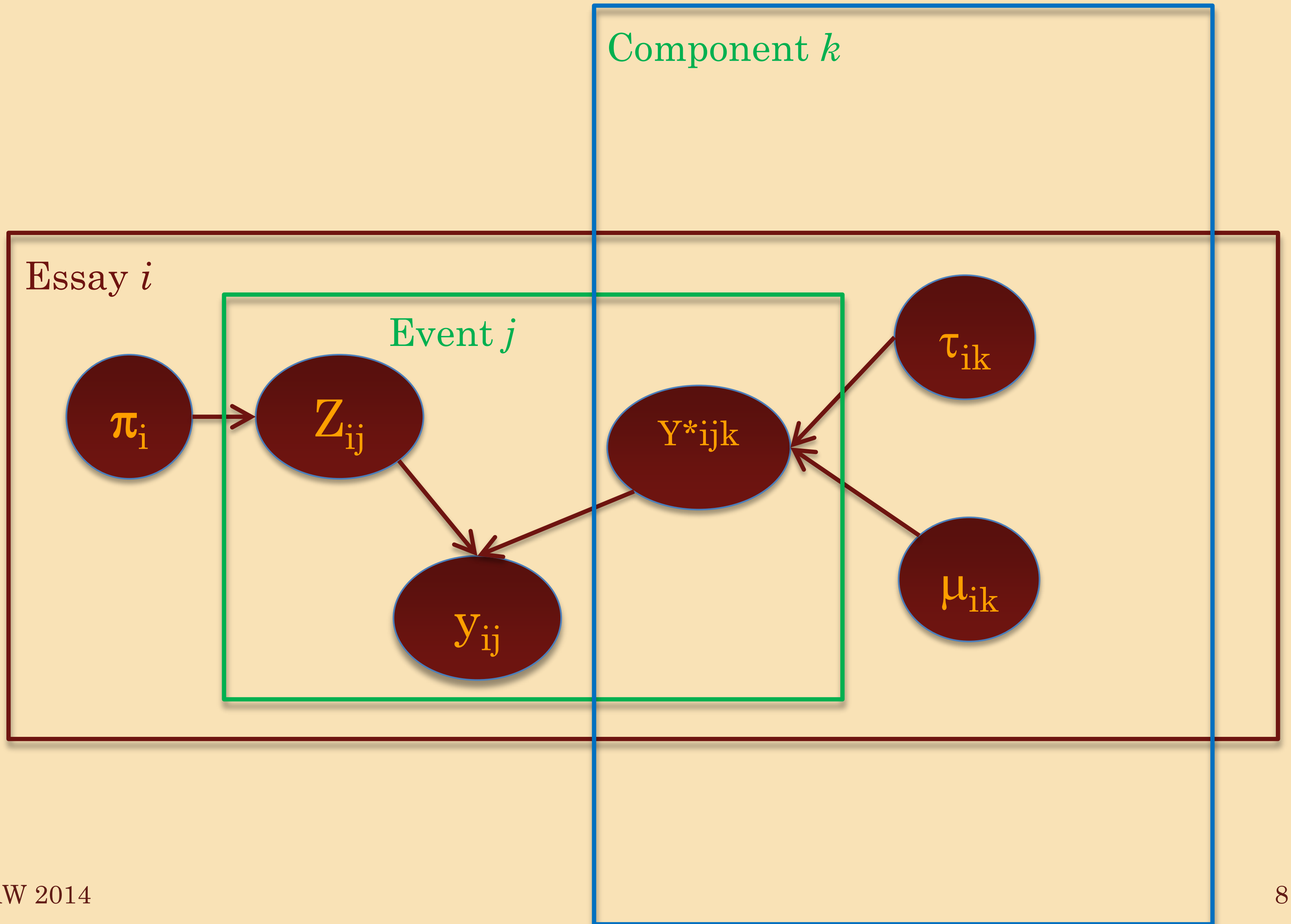
Within-Word Pauses



Mixture of Lognormals

- Log Pause Time $Y_{ij} = \log(X_{ij})$
  - Student (Level-2 unit) $i=1,\ldots,I$
  - Pause (Level-1 unit) $j=1,\ldots,J_i$
- $Z_{ij} \sim \mathrm{cat}(\pi_{i1},\ldots,\pi_{iK})$ 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



Mixture Model Problems

- If π_{ik} is small for some k, then category disappears
- If τ_{ik} is small for some k, then category becomes degenerate
- If μ_{ik}=$\mu_{ik'}$ and τ_{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:

$$\mu_{i1} < \mu_{i2} < \ldots < \mu_{iK}$$

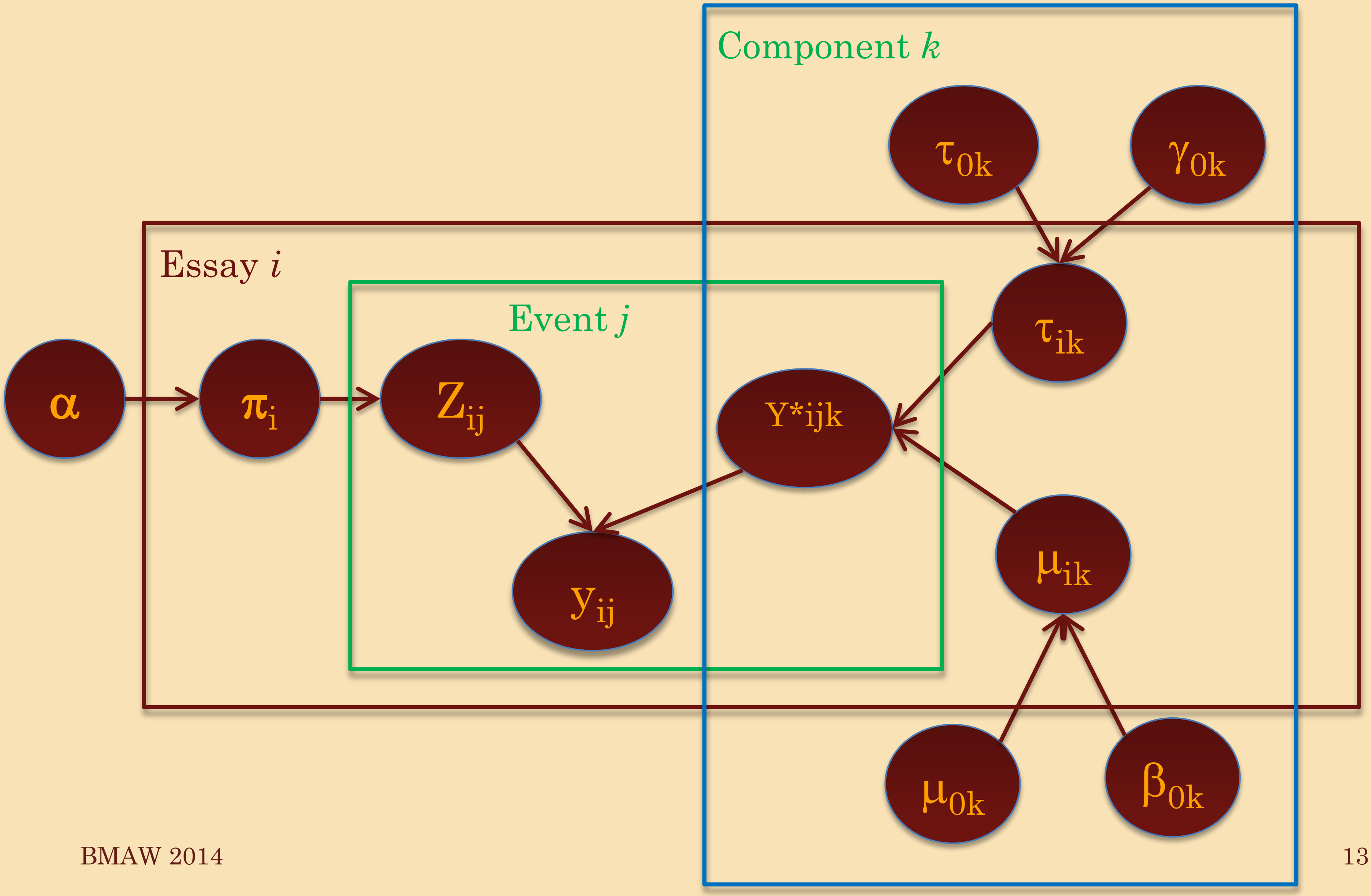
- Früthwirth-Schattner (2001) notes that when doing MCMC, better to let the chains run freely across the modes and sort out post-hoc
- Sorting needs to be done before normal MCMC convergence tests, or parameter estimation

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



Problems with hierarchical models

- If γ_{0k}, β_{0k} ➔ 0 we get complete pooling
- If γ_{0k}, β_{0k} ➔ ∞ 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 paramters

Add redundant parameters to make MCMC faster

- $\mu_{ik} = \mu_{0k} + \theta_i \beta_{0k}$
  - $\theta_i \sim N(0,1)$
- $\log(\tau_{ik}) = \log(\tau_{0k}) + \eta_i \gamma_{0k}$
  - $\eta_i \sim N(0,1)$
- $\alpha_k = \alpha_{0k} \alpha_N$
  - $\boldsymbol{\alpha}_0 \sim \text{Dirichlet}(\boldsymbol{\alpha}_{0m})$
  - $\alpha_N \sim \chi^2(2I)$

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

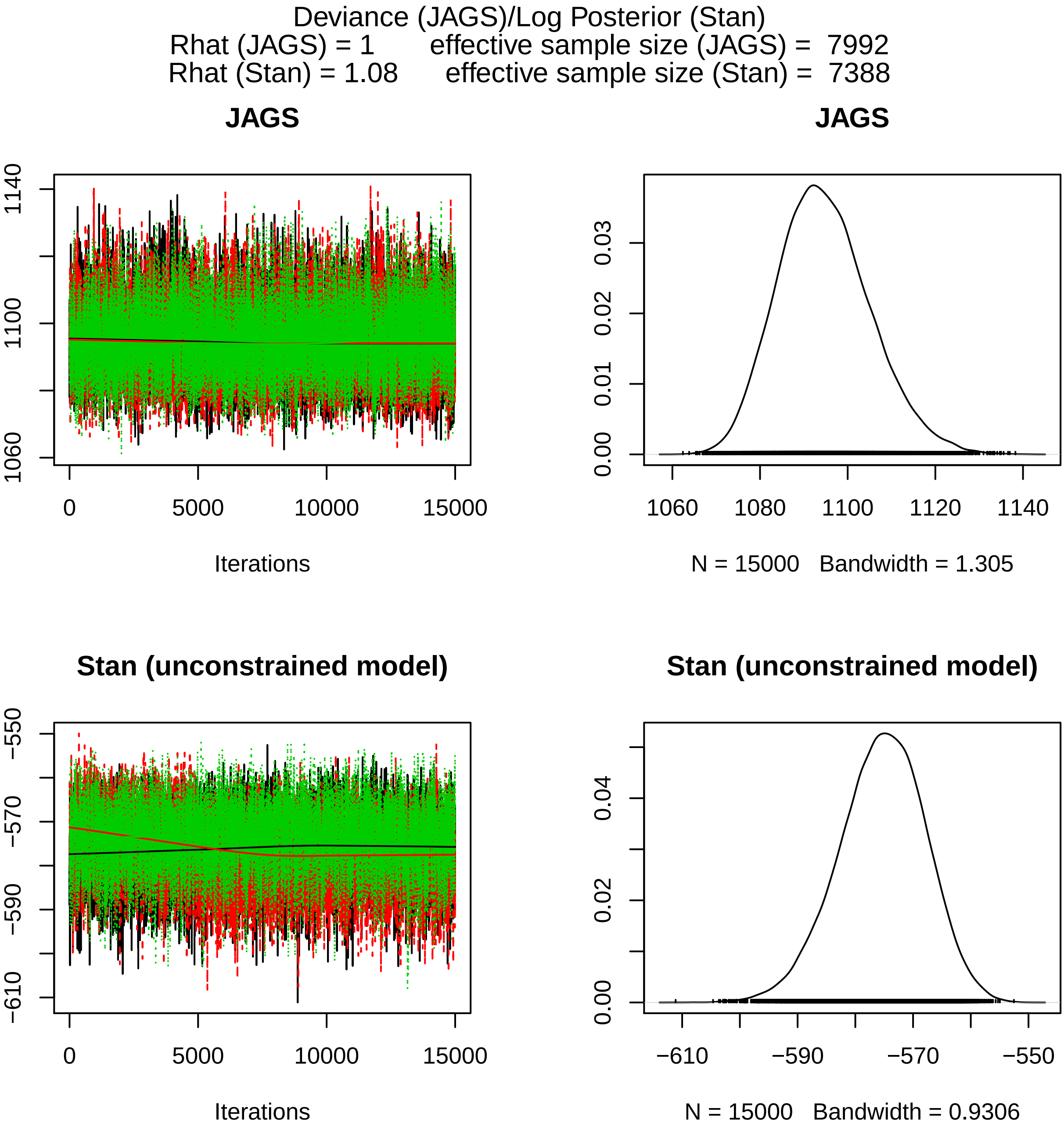
Repeat with subsets of the data for variety in multiple chains

Simulated Data Experiment

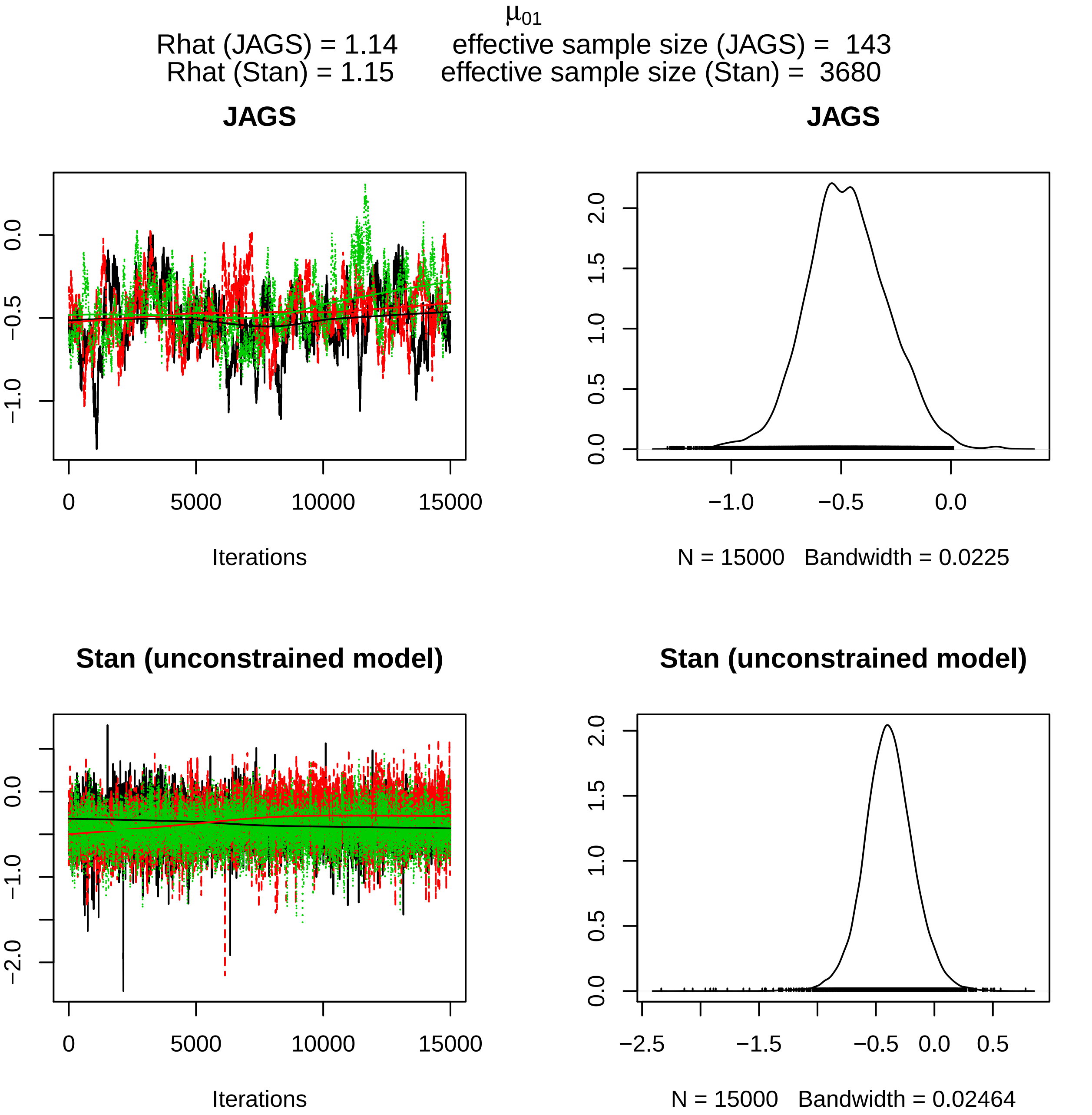
- Run initial value routine on real data for K=2,3,4
- Generate data from the model using these parameters
- Fit models with K'=2,3,4 to the data from true K=2,3,4 distributions in both JAGS (RWM) and Stan (HMC)
- Results shown for K=2, K'=2

Results (Mostly K=2, K'=2)

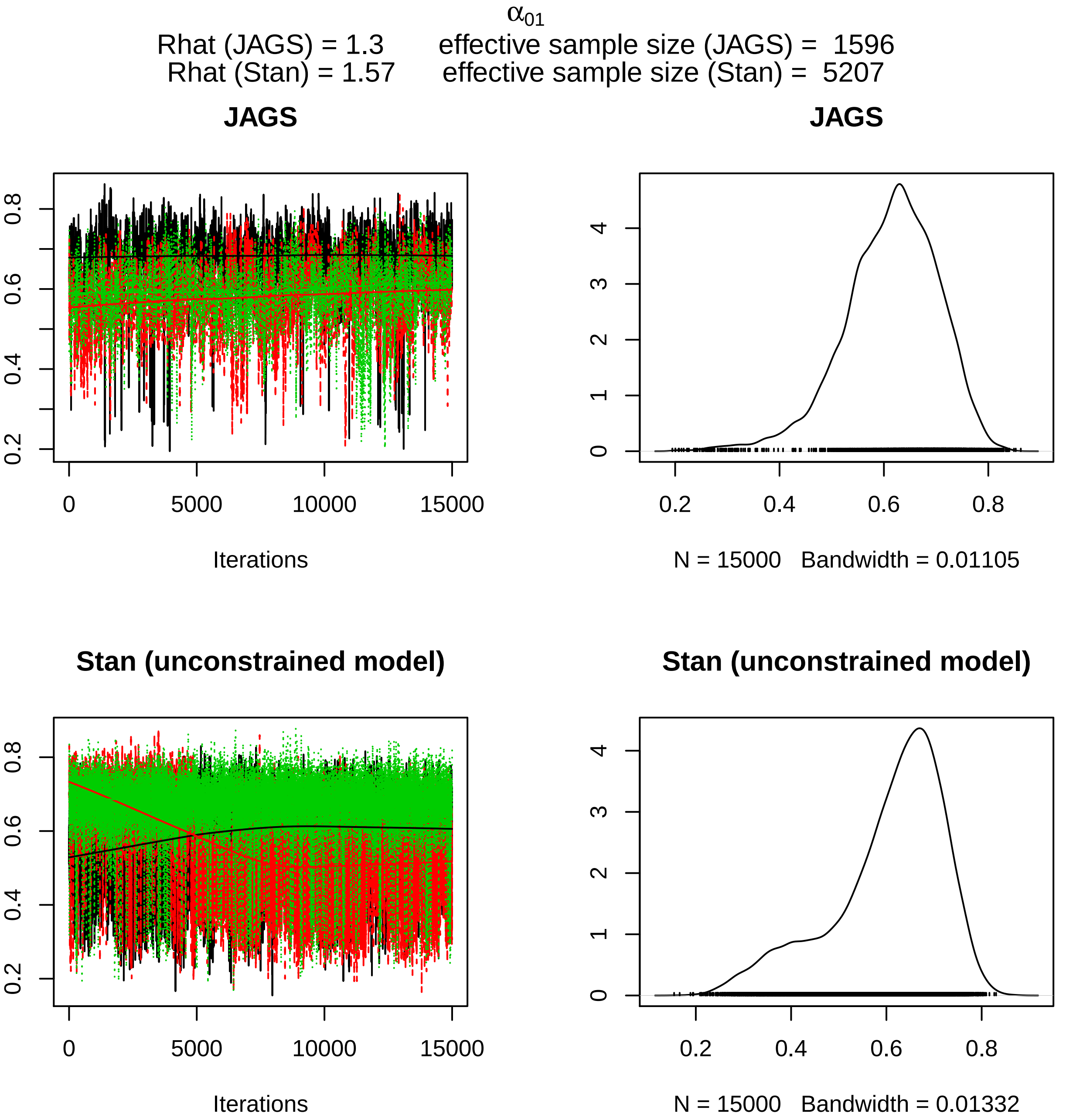
- All results (http://pluto.coe.fsu.edu/mcmc-hierMM)
- Deviance/Log Posterior—Good Mixing
- μ_{01} (average mean of first component)—Slow mixing in JAGS
- α_{01} (average probability of first component)—Poor convergence, switching in Stan
- γ_{01} (s.d. of log precisions for first component)—Poor convergence, multiple modes?

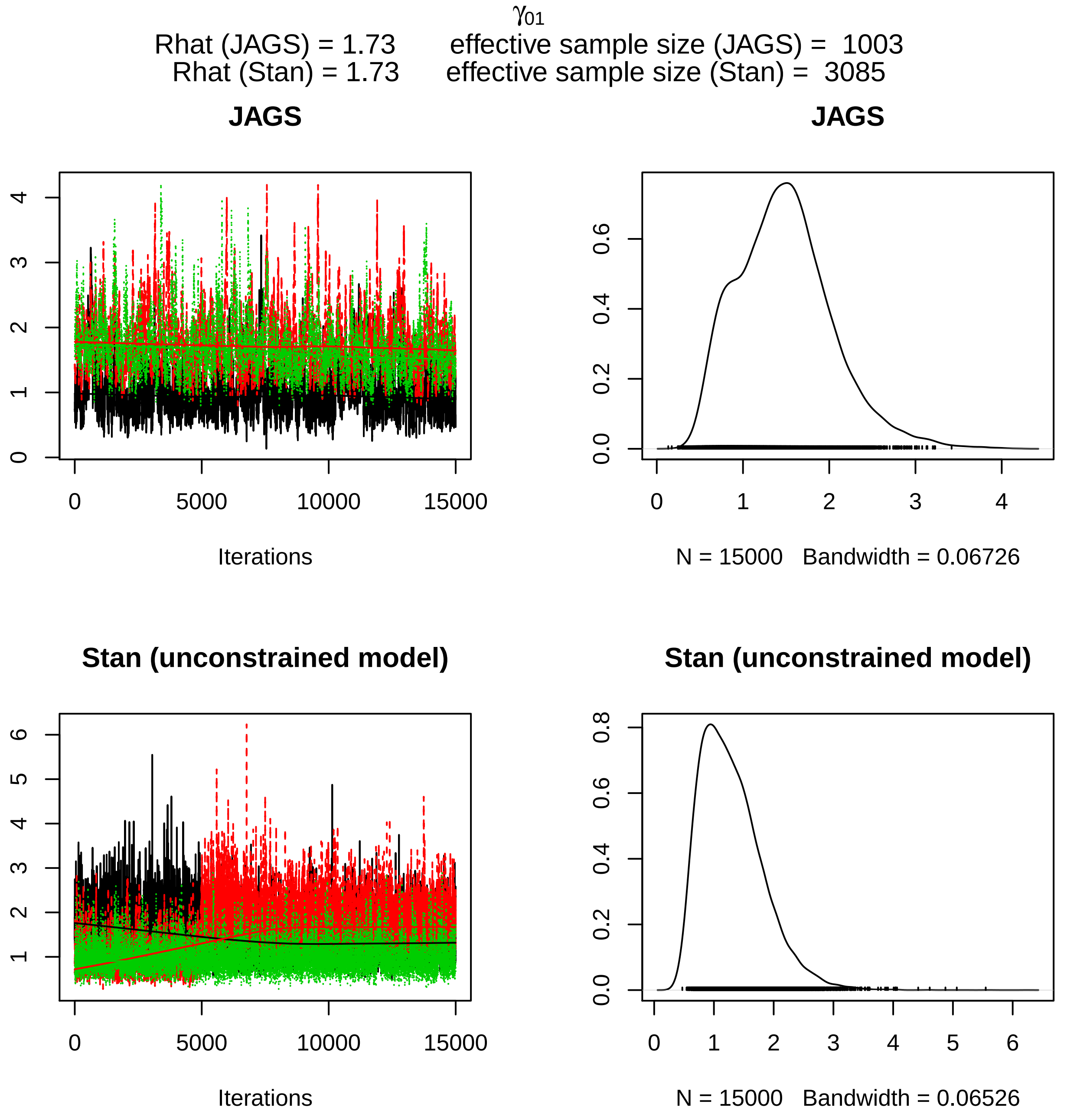


Deviance (JAGS)/Log Posterior (Stan)
Rhat (JAGS) = 1
effective sample size (JAGS) = 7992
Rhat (Stan) = 1.08
effective sample size (Stan) = 7388
JAGS
1060
1100
1140
0
5000
10000
15000
Iterations
JAGS
0.00
0.01
0.02
0.03
1060
1080
1100
1120
1140
N = 15000 Bandwidth = 1.305
Stan (unconstrained model)
−610
−590
−570
−550
0
5000
10000
15000
Iterations
Stan (unconstrained model)
0.00
0.02
0.04
−610
−590
−570
−550
N = 15000 Bandwidth = 0.9306



μ01
Rhat (JAGS) = 1.14 effective sample size (JAGS) = 143
Rhat (Stan) = 1.15 effective sample size (Stan) = 3680
JAGS
JAGS
Iterations
N = 15000 Bandwidth = 0.0225
Stan (unconstrained model)
Stan (unconstrained model)
Iterations
N = 15000 Bandwidth = 0.02464


α_{01}
Rhat (JAGS) = 1.3 effective sample size (JAGS) = 1596
Rhat (Stan) = 1.57 effective sample size (Stan) = 5207
JAGS
JAGS
Iterations
N = 15000 Bandwidth = 0.01105
Stan (unconstrained model)
Stan (unconstrained model)
Iterations
N = 15000 Bandwidth = 0.01332


γ01
Rhat (JAGS) = 1.73
effective sample size (JAGS) = 1003
Rhat (Stan) = 1.73
effective sample size (Stan) = 3085
JAGS
JAGS
Iterations
N = 15000 Bandwidth = 0.06726
Stan (unconstrained model)
Stan (unconstrained model)
Iterations
N = 15000 Bandwidth = 0.06526

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/
  - 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.