Investigating Chain Decisions and Convergence in Monte Carlo Markov-Chain Estimated Analysis

Zachary J. Roman & Holger Brandt

University of Kansas Psychology Department

Introduction

- In MCMC estimated analysis users specify a number of stochastic chains. The first step in determining model success is convergence. Convergence occurs when the random chains reach identical solutions (post burn in). However, the number of chains used has an effect on estimates of convergence [1]. Looking through literature, one is hard pressed to find any substantiated recommendation other than more than one [1]. Although, [2] suggested that multiple chains are more informative due to the possibility of a single chain finding local maxima as a solution. While they suggest choosing a chain number greater than one, no evidence is supplied to suggest a specific number for any given situation.

Objectives

- Explore the relationship between chain number and convergence decisions
  - How does this relationship change in regard to chain length
  - How does this relationship change with varying prior specifications

Simulation Design

1. A balanced clustered data set was simulated N = 900 with 30 groups
   $y_{ij} = 0 + u_{0j} + 0.5x_{1ij} + u_{1j}x_{1ij} + \epsilon_{ij}$
2. An MCMC analysis was conducted using Stan [3], 3 times, under 3 prior conditions with 20 chains specified
3. Entire chains were sampled (disregarding burn-in) in differing amounts (2:10) 1,000 times
4. Each condition (Prior, Chain Number,& Chain Length) was aggregated and summarized

- To describe convergence $\hat{R}$ was computed under each condition
- To account for convergence on improper solutions parameter bias was computed for the fixed effect β_x

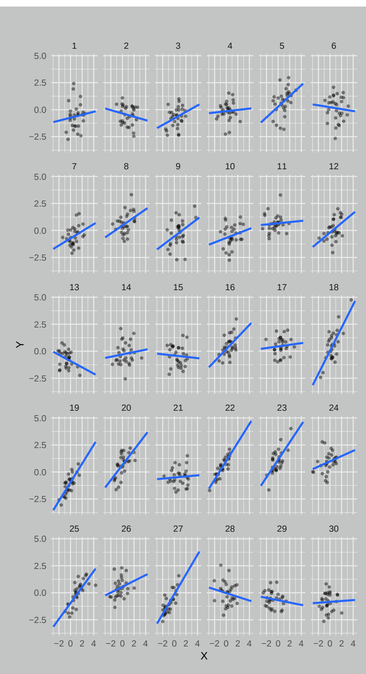

Figure 1: Simulated data by group ID

Simulation Conditions

- 9 chain number conditions
  - 2 through 10 chains
- 2 chain lengths post burn in.
  - 2,000 and 4,000
- 3 prior conditions for β_x
  - Diffuse, Realistic, Strong and Incorrect
- 54 total conditions

- All Prior Conditions:
  - $var(\beta_x) \sim t(3, 0)$
  - $var(\beta_0) \sim t(3, 0)$
  - $\sigma \sim t(3, 0)$
  - $cov(u) \sim LKJ(1)$
- Diffuse
  - $\beta_x \sim unif(-1000, 1000)$
- Realistic
  - $\beta_x \sim N(0, 1)$
- Strong Incorrect
  - $\beta_x \sim N(10, 0.1)$

The specified model includes both random intercepts and slopes for the effect of group $y_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1ij} + u_{1j}x_{1ij} + \epsilon_{ij}$

Measuring Performance

$$\sqrt{\hat{R}} = \sqrt{\frac{\frac{n-1}{n}W + \frac{1}{n}B}{W}} \quad W = \frac{1}{J}\sum_{j=1}^{J} S_J^2 \quad B = \frac{n}{J-1}\sum_{i=1}^{n}(\psi_{\cdot\bar{j}} - \psi_{\cdot\cdot})^2$$

Where: $\psi_{\cdot\bar{j}}$ is the chain mean, $\psi_{\cdot\cdot}$ is the grand mean of the chains S_j^2 summarizes the average squared distance of each retained draw from the chain mean. j is chain length, while n is the number of chains. W is a summary of within chain variation while B is a summary of between chain variation

- Convergence was assessed using the popular $\hat{R}$ statistic [4]. Where $\hat{R}$ values greater than 1.10 are regarded as non convergent, an estimate of 1.0 is said to be perfectly mixed. See the above equation.
- Bias was calculated as the mean squared summed deviation from the population value. Higher values of bias reflect poor estimates, while a bias of 0 reflects a perfect estimate.

Results: Convergence & Bias

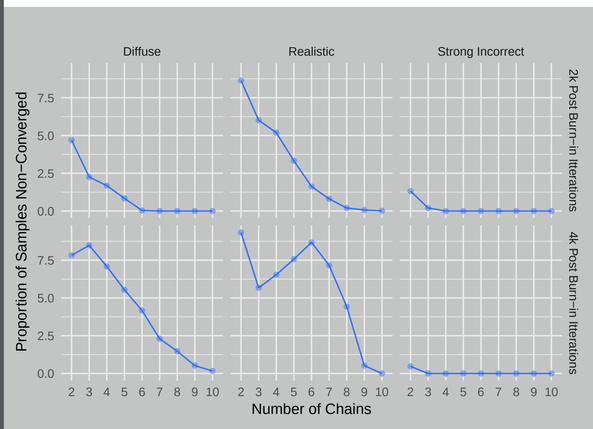

Figure 2: Proportion of samples with $\hat{R} > 1.10$

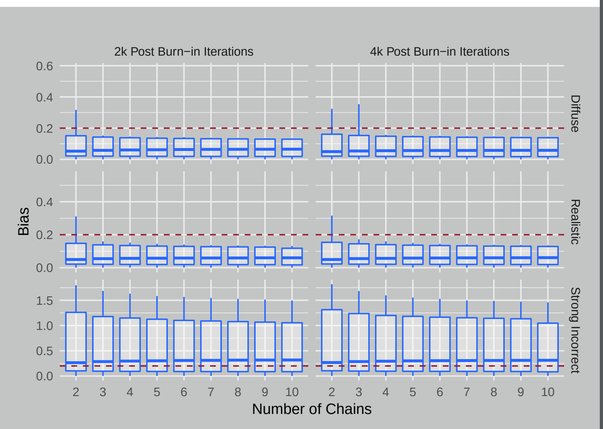

Figure 4: Bias of β_x

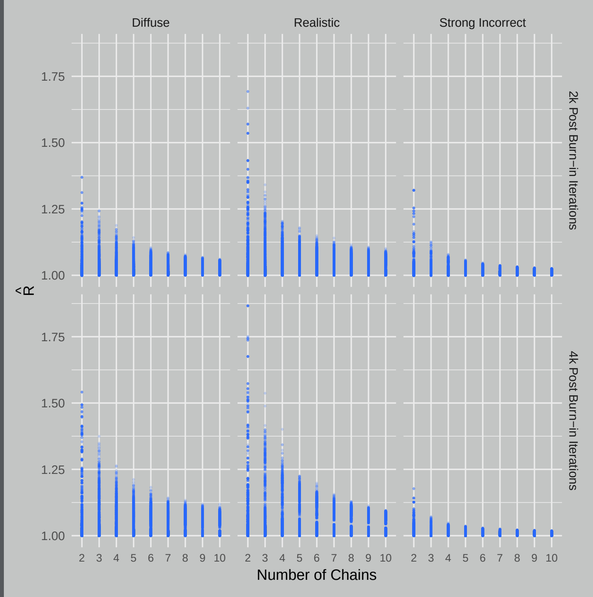

Figure 3: $\hat{R}$ for the overall model

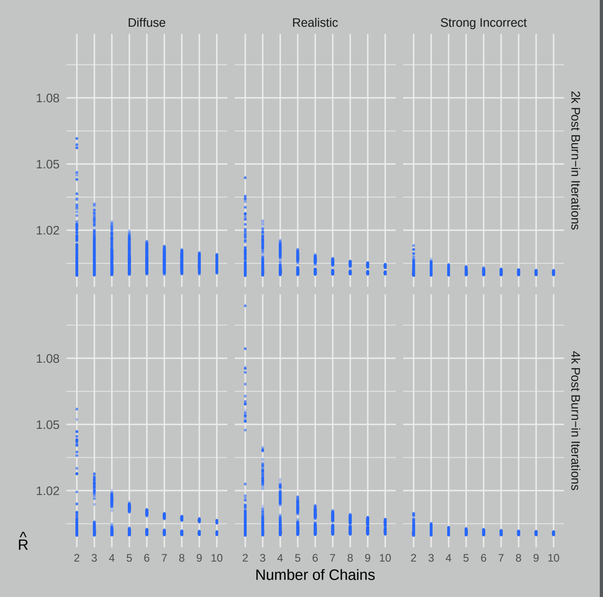

Figure 5: $\hat{R}$ for β_x

Conclusion

- More chains lead to more favorable estimates of model convergence under all prior conditions
  - Under the *Diffuse* & *Realistic* conditions the inclusion of more chains provides more favorable convergence estimates above and beyond the overall iteration count.
  - Under the *Strong Incorrect* condition, the trend is the same, however, the model was more likely to reach favorable convergence on improper solutions.
- Increasing the number of chains (between chain samples) leads to more favorable estimates of convergence as compared to increasing chain length (within chain samples).
  - One Major benefit to between chain information is the ability to run parallel computation. With little effort we can distribute the MCMC chains across multiple cores, whereas increasing iteration counts has no added computational benefit with multi-core processing.

Future Research

- Further simulation studies are needed to better define the benefits of favoring high chain numbers over post burn-in iteration counts
- Reasonably there are limits to the benefits of between chain samples over chain length, where are these boundaries? Under which modeling conditions do they change?

References

[1] Andrew Gelman, Kenneth Shirley, et al.
Inference from simulations and monitoring convergence.
Handbook of markov chain monte carlo, 6:163–174, 2011.

[2] Walter R Gilks, Sylvia Richardson, and David J Spiegelhalter.
Introducing markov chain monte carlo.
Markov Chain Monte Carlo in Practice, 1:19, 1996.

[3] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell.
Stan: A probabilistic programming language.
Journal of statistical software, 76(1), 2017.

[4] Andrew Gelman, Donald B Rubin, et al.
Inference from iterative simulation using multiple sequences.
Statistical science, 7(4):457–472, 1992.

www.zacharyroman.com

zroman@ku.edu