Revisiting the Gelman-Rubin Diagnostic[1]

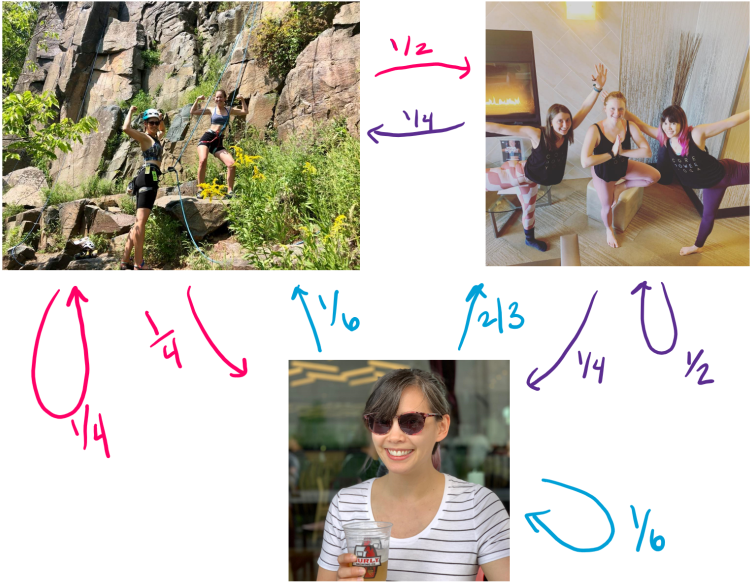
Christina Knudson, Ph.D.

University of St. Thomas
St. Paul, Minnesota

Women in Statistics and Data Science 2019

[1] Joint work with Dootika Vats, Ph.D., IIT Kanpur

Background on Markov Chains (and Me)



Using Markov Chain Monte Carlo

Goal: Use Markov chain Monte Carlo (MCMC) to approximate a target distribution (e.g. an intractible posterior distribution)

Using Markov Chain Monte Carlo

Goal: Use Markov chain Monte Carlo (MCMC) to approximate a target distribution (e.g. an intractible posterior distribution)

Issue: After the chain has started sampling from the target distribution, how long should the sampler run to produce a decent approximation?

Using Markov Chain Monte Carlo

Goal: Use Markov chain Monte Carlo (MCMC) to approximate a target distribution (e.g. an intractible posterior distribution)

Issue: After the chain has started sampling from the target distribution, how long should the sampler run to produce a decent approximation?

Tool: Gelman-Rubin diagnostic (1992)

$$\hat{R} = \sqrt{\frac{\text{chain length} - 1}{\text{chain length}} + \frac{\text{between-chain variance}}{\text{within-chain variance}}}$$

Gelman-Rubin: Is $\hat{R} < 1.1$ Small Enough?

$\hat{R}$ decreases to 1 as the chain length increases, but how small is small enough?

Gelman et al (2004):

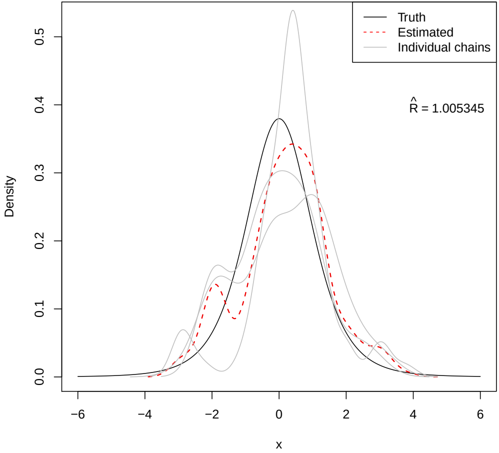
> *For most examples, values below 1.1 are acceptable, but for a final analysis in a critical problem, a higher level of precision may be required.*

$\hat{R}$ thresholds used in 100 papers from 2017:

$\hat{R}$	1.003	1.01	1.02	1.03	1.04	1.05	1.06	1.07	1.1	1.2	1.3
Freq.	1	12	9	9	2	11	2	1	43	9	1

Gelman-Rubin: $\hat{R} < 1.1$?

Reality: stopping at $\hat{R} = 1.1$ can be too early!



Vats and Knudson's Contributions

How can we improve the Gelman-Rubin diagnostic?

1. Stabilize the Gelman-Rubin statistic
2. Construct principled threshold for terminating simulation (move away from $\hat{R} < 1.1$)

Stabilizing the Gelman-Rubin Statistic

Original calculation for between-chain variance is unstable

Lugsail batch means variance estimation is **stable**
and overestimates between-chain variance

$\Rightarrow$ $\hat{R}$ is overestimated

$\Rightarrow$ Chain must run longer for $\hat{R}$ to reach termination threshold

R command: `stable.GR` in R package `stableGR`

Stabilizing the Gelman-Rubin Statistic

An AR(1) process

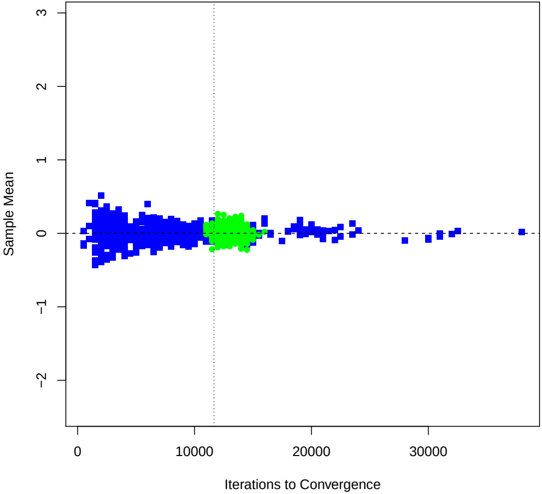
$$Y_t = .95\ Y_{t-1} + \epsilon_t, \quad t = 1, 2, \ldots$$
$$\epsilon_t \sim N(0,\ 1^2)$$

is the same as a Markov chain with distribution $N(0,\ 10.25641)$.

For each of 500 replications, we run five Markov chains until $\hat{R} < 1.001625$ using

- original GR $\hat{R}$ calculation (blue dots)
- VK $\hat{R}$ calculation (green dots)

Stabilizing the Gelman-Rubin Statistic



Blue: original GR $\hat{R}$ calculation. Green: Vats and Knudson's new $\hat{R}$.

A Principled Threshold for Terminating Simulation

Effective sample size: number of uncorrelated samples that produce the same precision as the correlated (MCMC) sample.

V+K identified a one-to-one relationship between ESS and $\hat{R}$:

$$\hat{R} = \sqrt{\frac{\text{chain length} - 1}{\text{chain length}} + \frac{\text{number of chains}}{\text{effective sample size}}}$$

A Principled Threshold for Terminating Simulation

Effective sample size: number of uncorrelated samples that produce the same precision as the correlated (MCMC) sample.

V+K identified a one-to-one relationship between ESS and $\hat{R}$:

$$\hat{R} = \sqrt{\frac{\text{chain length} - 1}{\text{chain length}} + \frac{\text{number of chains}}{\text{effective sample size}}}$$

Upshot:

- Threshold can be calculated *a priori*
  - Similar to introductory statistics sample size calculations for a desired width of a confidence interval
  - Gong and Flegal (2016) and Vats et al. (2019)
- $\hat{R}$ threshold is easily-interpretable

R commands in `stableGR`: `target.psrf`, `n.eff`

A Principled Threshold for Terminating Simulation

Model the log odds of surviving the Titanic's sinking.

Bayesian logistic regression with the following predictors:

- Fare class (3 categories)
- Sex (2 categories)
- Age (quantitative)
- Number of siblings/spouses aboard (quantitative)
- Number of parents/children aboard (quantitative)
- Port of embarkation (3 categories)

A Principled Threshold for Terminating Simulation

Model the log odds of surviving the Titanic's sinking.

Bayesian logistic regression with the following predictors:

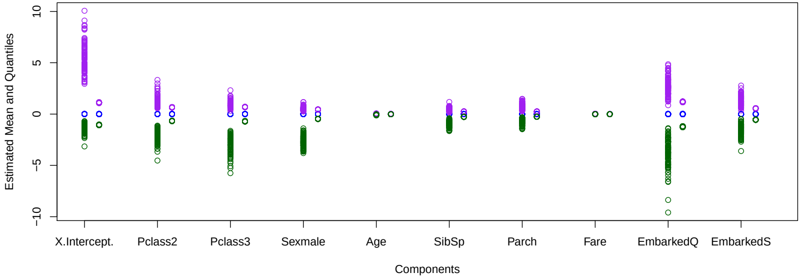
- Fare class (3 categories)
- Sex (2 categories)
- Age (quantitative)
- Number of siblings/spouses aboard (quantitative)
- Number of parents/children aboard (quantitative)
- Port of embarkation (3 categories)

For each of 100 reps, we run 5 chains until convergence is diagnosed according to

- $\hat{R} < 1.1$
- VK's ESS-based $\hat{R}$ termination threshold

using VK's new $\hat{R}$ calculation in both cases.

A Principled Threshold for Terminating Simulation

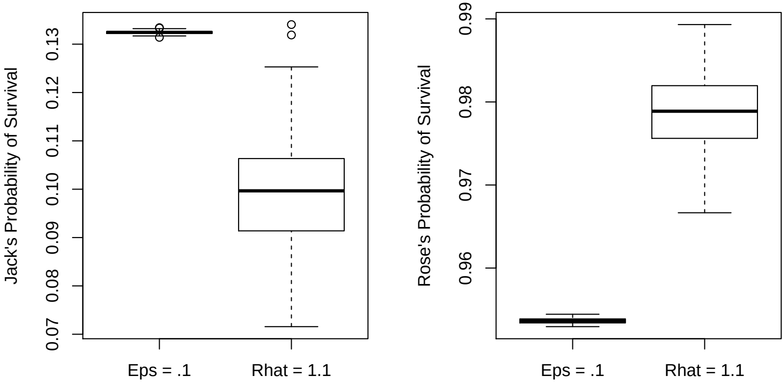


Centered posterior means (blue) and 95% credible interval estimates (green for lower bound, purple for upper bound).

Left points: $\hat{R} < 1.1$. Right points: ESS-based $\hat{R}$ threshold.

What about Jack and Rose?

Bayesian Logistic Regression



Concluding Remarks

To review, we have:

- Stabilized the Gelman-Rubin statistic $\hat{R}$.
- Identified a one-to-one relationship between ESS and $\hat{R}$.
- Created an interpretable stopping rule to replace $\hat{R} < 1.1$.

Additional information:

- Diagnostic is usable for multiple chains or a single chain.
- We have also stabilized the multivariate version of the Gelman-Rubin statistic and produced an interpretable stopping rule for multivariate chains.
- R package `stableGR` not yet available on CRAN.
  You can currently install it from Github.

More Information

`cknudson.com`

for links to

"Revisiting the Gelman-Rubin Diagnostic" on arXiv

and the Github repo for R package `stableGR`

References

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434-455.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7:457-472.

Gong, L. and Flegal, J. M. (2016). A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 25:684-700.

Vats, D. and Flegal, J. M. (2018). Lugsail lag windows and their application to MCMC. *arXiv e-prints.*

Vats, D., Flegal, J. M, and Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2):321-337.

Commands for Installing R Package stableGR from Github

```
  #get some required packages
install.packages("Rcpp")
install.packages("RcppArmadillo")
install.packages("devtools")

  #install mcmcse from github  (rather than CRAN)
library(devtools)
install_github("dvats/mcmcse")

  #install stableGR package
install_github("knudson1/stableGR/stableGR")
library(stableGR)
```

Will be available on CRAN in a couple months.