Revisiting the Gelman-Rubin Diagnostic[1]

Christina Knudson, Ph.D.

University of St. Thomas
St. Paul, Minnesota

BayesComp 2020

[1] Joint work with Dootika Vats, IIT Kanpur

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)

- Run parallel chains with overdispersed starting points.
- Stop when $\hat{R}$ hits a predetermined threshold.

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

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):

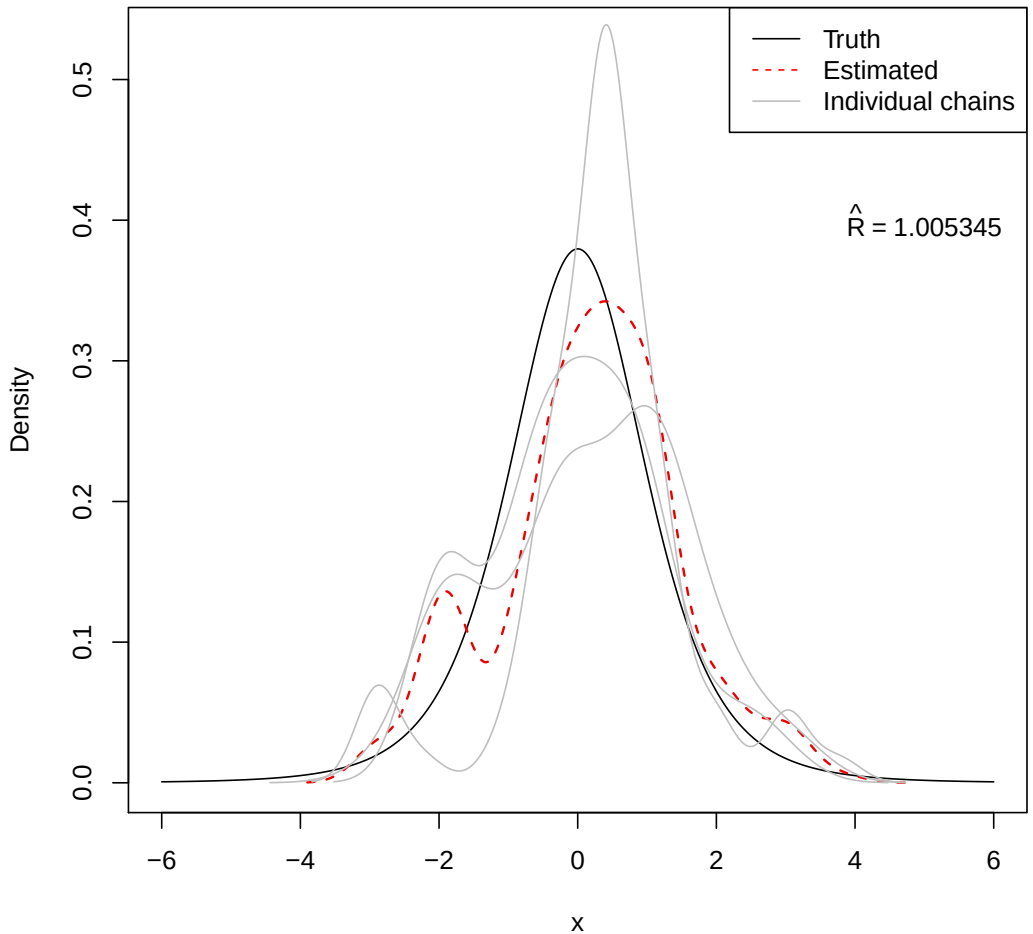
> *For most examples, values below 1.1 are acceptable, but for a final analysis in a critical problem, a higher level of precision may be required.*

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

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
   (by using better variance estimation)
2. Construct a principled threshold for terminating simulation
   (rather than using $\hat{R} < 1.1$)

Stabilizing the Gelman-Rubin Statistic

Instability in between-chain variance $\Rightarrow$ instability in $\hat{R}$:

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

Next slides:

- Why is GR's between-chain variance unstable?
- How can we stabilize $\hat{R}$?

Why is GR's Between-Chain Variance Unstable?

Let X_{it} denote the Markov chain draw from chain i ($i = 1, ..., m$) at time t ($t = 1, ..., n$).

Why is GR's Between-Chain Variance Unstable?

Let X_{it} denote the Markov chain draw from chain i ($i = 1, ..., m$) at time t ($t = 1, ..., n$). Next, define the following means:

$$\bar{X}_{i\cdot} = \frac{1}{n}\sum_{t=1}^{n} X_{it} \qquad \text{and} \qquad \hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} \bar{X}_{i\cdot}$$

Why is GR's Between-Chain Variance Unstable?

Let X_{it} denote the Markov chain draw from chain i ($i = 1, ..., m$) at time t ($t = 1, ..., n$). Next, define the following means:

$$\bar{X}_{i\cdot} = \frac{1}{n}\sum_{t=1}^{n} X_{it} \qquad \text{and} \qquad \hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} \bar{X}_{i\cdot}$$

Then the original GR estimator for between-chain variance is

$$\frac{1}{m-1}\sum_{i=1}^{m} \left(\bar{X}_{i\cdot} - \hat{\mu}\right)^2,$$

which has high variance since usually m is small.

How Can We Stabilize the Gelman-Rubin Statistic?

Look again at the between-chain variance estimator:

$$\frac{1}{m-1}\sum_{i=1}^{m}\left(\bar{X}_{i\cdot}-\hat{\mu}\right)^2.$$

This simply estimates the variance of the MCMC means!

(I heard some of you are experts in this very topic.)

Why not leverage the rich MCMC variance estimation literature?

Many options! Your favorite estimator is probably more stable.

Stabilizing the Gelman-Rubin Statistic

We choose lugsail batch means variance estimation because it stabilizes $\hat{R}$ and makes termination conservative.

Lugsail BM overestimates between-chain variance for finite samples

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

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

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

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

Stabilizing the Gelman-Rubin Statistic

An AR(1) process

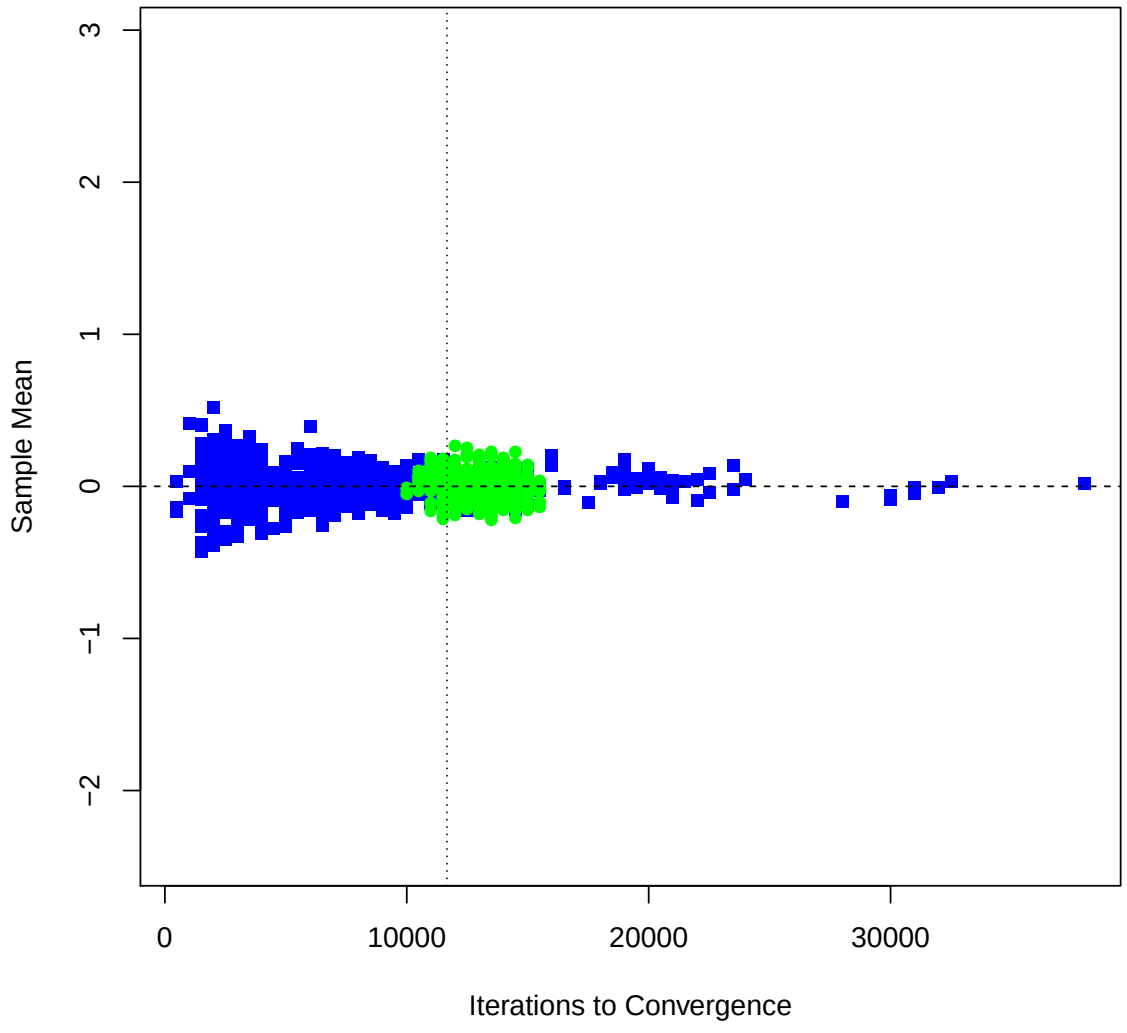
$$Y_t = .95\ Y_{t-1} + \epsilon_t, \quad t = 1, 2, \ldots$$
$$\epsilon_t \sim N(0,\ 1^2)$$

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)
- lugsail-based $\hat{R}$ calculation (green dots)

Stabilizing the Gelman-Rubin Statistic



Blue: original GR $\hat{R}$ calculation.

Green: lugsail-based $\hat{R}$.

A Principled Threshold for Terminating Simulation

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

We identified a one-to-one relationship between estimated ESS and $\hat{R}$:

$$\hat{R} = \sqrt{\frac{\text{chain length} - 1}{\text{chain length}} + \frac{\text{number of chains}}{\text{estimated ESS}}}$$

A Principled Threshold for Terminating Simulation

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

We identified a one-to-one relationship between estimated ESS and $\hat{R}$:

$$\hat{R} = \sqrt{\frac{\text{chain length} - 1}{\text{chain length}} + \frac{\text{number of chains}}{\text{estimated ESS}}}$$

Upshot:

- $\hat{R}$ threshold is interpretable
- 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)

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:

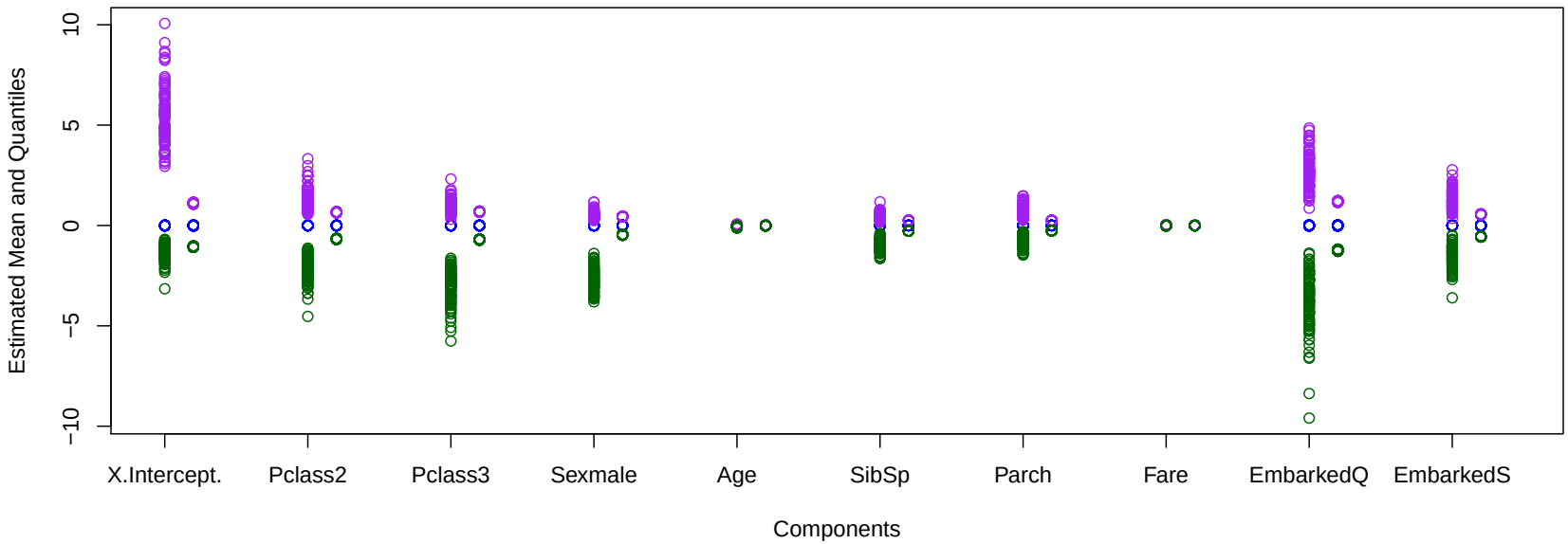
- Fare class (3 categories)
- Sex (2 categories)
- Age (quantitative)
- Number of siblings/spouses aboard (quantitative)
- Number of parents/children aboard (quantitative)
- Port of embarkation (3 categories)

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 our lugsail-based $\hat{R}$ 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.

Concluding Remarks

To review, we have:

- Stabilized $\hat{R}$ with improved variance estimation.
- Identified a one-to-one relationship between ESS and $\hat{R}$.
- Created an ESS-based stopping rule to replace $\hat{R} < 1.1$.

Additional information:

- Our diagnostic works for ≥ 1 chain.
- We have also stabilized the multivariate version of the Gelman-Rubin statistic and produced an interpretable stopping rule for multivariate chains.
- You can install R package `stableGR` from Github.

More Information

Thank you!

`cknudson.com`

links to
these slides,
"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.

FAQ: "Are you saying I should run more/fewer chains?"

No. We are not campaigning for 1 chain or 5 chains.

We are focused on output analysis. In particular, we want to provide tools that help users run their chain(s) long enough to produce a good approximation of a density (regardless of the number of chains).

Whether you run 1 chain or > 1 chain, you can use our $\hat{R}$.

FAQ: "Can I use a different variance estimator?"

Sure!

In fact, we hope that $\hat{R}$ will continue to evolve as researchers develop better methods of variance estimation.

We chose an estimator with low variance, strong consistency, and a slight bias from above for finite samples. Other properties may also be useful.

FAQ: "How did you change the multivariate $\hat{R}$?"

Our change to the multivariate $\hat{R}$ is similar to that for univariate $\hat{R}$.

First, let's introduce multivariate $\hat{R}$ by Brooks and Gelman (1998). Within-chain variance is the multivariate analog:

- $S_i, i = 1, \ldots, n$ is the sample covariance matrix for chain i.
- $S = \frac{1}{n} \sum S_i$.

BG uses the following for between chain variance:

$$\frac{B}{n} = \frac{1}{m-1} \sum_{i=1}^{m} (\bar{X}_{i\cdot} - \hat{\mu})(\bar{X}_{i\cdot} - \hat{\mu})^T$$

and their multivariate analog of $\frac{\text{between-chain variance}}{\text{within-chain variance}}$ is the max eigenvalue of $S^{-1}B/n$. Thus, BG's multivariate $\hat{R}$ is controlled by the slowest-converging component.

FAQ: "How did you change the multivariate $\hat{R}$?"

Our multivariate $\hat{R}$ calculation keeps

- $S_i, i = 1, \ldots, n$ is the sample covariance matrix for chain i.
- $S = \frac{1}{n} \sum S_i$.

and replaces the between-chain variance calculation

$$\frac{B}{n} = \frac{1}{m-1} \sum_{i=1}^{m} (\bar{X}_{i\cdot} - \hat{\mu})(\bar{X}_{i\cdot} - \hat{\mu})^T$$

with the lugsail batch means variance estimator. Finally, rather than using the max eigenvalue of $S^{-1}B/n$, we used the determinant because it better represents the **joint** convergence of the components.