parallelMCMCcombine: An R Package for Bayesian Methods for Big Data and Analytics

Alexey Miroshnikov[1], Erin M. Conlon[1*]

1 Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts, United States of America

* E-mail: econlon@mathstat.umass.edu

Abstract

Recent advances in big data and analytics research have provided a wealth of large data sets that are too big to be analyzed in their entirety, due to restrictions on computer memory or storage size. New Bayesian methods have been developed that partition big data sets into subsets, and perform independent Bayesian Markov chain Monte Carlo analyses on the subsets. The methods then combine the independent subset posterior samples to estimate a posterior density given the full data set. These approaches were shown to be effective for Bayesian models including logistic regression models, Gaussian mixture models and hierarchical models. Here, we introduce the R package **`parallelMCMCcombine`** which carries out four of these techniques for combining independent subset posterior samples. We illustrate each of the methods using a Bayesian logistic regression model for simulation data and a Bayesian Gamma model for real data; we also demonstrate features and capabilities of the R package. The package assumes the user has carried out the Bayesian analysis and has produced the independent subposterior samples outside of the package. We envision this tool will allow researchers to explore the various methods for their specific applications, and will assist future progress in this rapidly developing field.