DOI 10.18699/MM-HPC-BBB-2018-35

Bayesian approach to big data processing: problems and perspectives

M.A. Marchenko
Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia
** e-mail: marchenko@sscc.ru*

Key words: Big Data, Bayesian inference, Markov Chain Monte Carlo, high-performance computations

Motivation and Aim: We desire to apply Bayesian inference (BI) methods for analysis and processing of Big Data arising in system biology, genomics, phylogenetics, etc. The BI methods have already proved their efficiency in many applications, such as nuclear physics, economics, genetics, etc. [1, 2]. In the BI approach, the Markov Chain Monte Carlo (MCMC) methods are used for simulation of the underlying random processes to get the statistical output, which is the evaluation of the model parameters [3, 4].
Unfortunately, standard MCMC methods can scale poorly to big data settings due to the need to evaluate the likelihood at each iteration [5]. There have been a number of approximate MCMC algorithms that use sub-sampling ideas to reduce this computational burden, but with the drawback that these algorithms no longer target the true posterior distribution.
Methods and Algorithms: We developed the PARMONC (PARallel MONte Carlo) solver for distributed stochastic simulation on clusters with massive-parallel and hybrid architectures [6]. A core of the PARMONC is the well-tested, fast and reliable 128-bit parallel random numbers generator, which is in intensive use for more than a decade.
Results: The BI approach and the MCMC methods began to apply on all architectures of computing systems with parallel and distributed computing. We propose the complex methodological approach: the parallel random number generators, libraries, data processing, control programs, etc., i. e. all the stages of creating a "digital product".
Conclusion: New theoretical study and high-performance MCMC methods are needed to make the BI approach a highly effective technique to process Big Data.
Acknowledgements: Supported by the RFBR (18-01-00599, 18-41-540017, 16-01-00755, 16-01-00530).

References

1. Wong K.-C. (2016) Big data analytics in genomics. Springer. ISBN 9783319412788
2. G.M. Allenby et al. (2014) Perspectives on Bayesian Methods and Big Data. Cust. Need. and Solut. 1:169-175.
3. Ahmed S.E. (2016) Big and complex data analysis, contributions to statistics. Springer. ISBN 9783319415727.
4. Insua D. et al. (2012) Bayesian Analysis of Stochastic Process Models. Wiley. ISBN 978-0-470-74453-6.
5. Solonen A. et al. (2012) Efficient MCMC for climate model parameter estimation: parallel adaptive chains and early rejection. Bayesian Anal. 7(3):715-736.
6. Marchenko M. (2014) Efficient Computational Approaches for Parallel Stochastic Simulation on Supercomputers. Parallel Programming: Practical Aspects, Models and Current Limitations. New York: Nova Science Publishers: 117-142.