ABSTRACT

Assessing the quality of estimators is a task of fundamental importance in Data Science. Quality assessments such as confidence region, bias, risk and so on provide much more information than a simple point estimate. This paper aims to study a statistical sampling algorithm Bag of Little Bootstraps (BLB) that solves the same class of problems as general bootstrapping, but which parallelizes better. We do this by implementing this algorithm, as described in “A Scalable Bootstrap for Massive Data,” as well as other bootstrapping methods on simulations of the Uniform distribution and Normal distribution. We found that despite of the fact that the Bag of Little Bootstrap has a more favorable computational profile than other bootstrapping methods, the final approach in assessing/evaluating the estimator’s quality still needs to be discussed further as each assessment is computed based on different data support at different computing node.