Bootstrap Methods in R and C

Ella Kaye Xenia Miscouridou

December 2, 2015

Abstract

We describe a package `BLB` in R which implements the Bag of Little Bootstraps (Kleiner et al. 2014).

The package is available from `https://github.com/EllaKaye/BLB`.

1 Introduction

The bootstrap (Efron 1979), is a statistical tool which gives measures of accuracy of estimators, and can therefore be used to draw inference about the parameters of the sampling distribution. It is a powerful and widely applicable tool. The bootstrap is at the interface between statistical inference and computation, and it was developed at a time when advances in computing technology allowed this computationally intensive method to be used in practice. Shortly after, the Bayesian analogue of the bootstrap was introduced (Rubin 1981). More recently, in this era of 'big data', statistical methodology is again being developed alongside developments in computing technology. Parallel and distributed computing architectures now make it possible to extend the bootstrap so that it can be applied to massive datasets, the so-called bag of little bootstraps (Kleiner et al. 2014).

We have developed a package, `Bootstrap`, with functions to implement the bootstrap, Bayesian bootstrap and bag of little bootstraps. The package can be obtained from GitHub.

2 The Bootstrap

Suppose we observe a sample of n iid realisations $x_1, \ldots, x_n \sim P$, for some probability measure P. Let θ be a parameter of the distribution, and $\hat{\theta}_n$ be an estimator of θ. The goal of the bootstrap is to obtain an assessment, ξ, of the quality of the estimator. For example, θ could be the median, and ξ the standard error. To obtain a bootstrap estimate, we proceed as follows:

1. Repeatedly (B times) sample n points with replacement from the original dataset, giving bootstrap replications (resamples) $(x_1^{*(i)}, \ldots, x_n^{*(i)})$, $i = 1, \ldots, B$

2. Compute $\hat{\theta}_n^{*(i)}$ on each of the B resamples.

3. Compute $\xi^* = \xi(\hat{\theta}_n^{*(1)}, \ldots, \hat{\theta}_n^{*(B)})$ as our bootstrap estimate of ξ.

3 Bag of Little Bootstraps

The original bootstrap arose around the time when increases in computing power allowed the development of statistical tools that had previously been too computationally expensive. In recent years, there has been an influx of 'big data', alongside the development of parallel computing architectures. Kleiner et al. (2014) have developed a scalable bootstrap for massive data, known as the Bag of Little Bootstraps (BLB). With massive datasets, the bootstrap's need for recomputation on resamples of the same size as the original dataset is problematic. Rather than obtain bootstrap samples from the whole dataset, the BLB breaks down the process as follows:

1. Repeatedly (s times) subsample $b(n) < n$ data points *without replacement* from the original dataset of size n.

2. For each of the s subsamples, do the following:

   (a) Repeatedly (r times) resample n data points *with replacement* from the subsample.

   (b) Compute $\hat{\theta}_n^*$ on each resample.

   (c) Compute an estimate of ξ based on these r realisations of $\hat{\theta}_n^*$.

3. We now have one estimate of ξ per subsample, $\xi_1^*, \ldots, \xi_s^*$. Output their average, ξ^*, as the final estimate of ξ for $\hat{\theta}_n$.

Kleiner et al. (2014) recommend taking $b(n) = n^\gamma$, where $\gamma \in [0.5, 1]$. This procedure dramatically reduces the size of each resample. For example, if $n =$ 1 million and $\gamma = 0.6$, the size of the original dataset may be around 1TB, with a bootstrap resample typically occupying approximately 632GB, and a BLB subsample or resample occupying just 4GB.

The Bootstrap package contains three functions which implement BLB in different cases, `BLB.1d, BLB.multi` and `BLB.adapt`. The function `BLB.1d` implements the simplest version of BLB. It takes as input: a 1-dimensional dataset in a vector; γ, which controls the value of b; a function FUN, which computes the parameter estimate, $\hat{\theta}_n$. It also takes as arguments s and r, which default to 20 and 100 respectively (Kleiner et al. (2014) demonstates that these values are likely as large as they'll need to be to obtain convergence). The function returns a BLB estimate of ξ, which is set as the standard error. When $n = 500,000$ the function takes under two minutes to execute (although we compute with a smaller sample here).

References

Efron, Bradley (1979). "Bootstrap methods: another look at the jackknife". In: *Annals of Statistics* 7, pp. 1–26.

Kleiner, Ariel et al. (2014). "A scalable bootstrap for massive data". In: *Journal of the Royal Statistical Society (Series B)* 76.4, pp. 795–816.

Rubin, Donald B (1981). "The Bayesian Bootstrap". In: *The Annals of Statistics* 9.1, pp. 130–134.