A scalable bootstrap for massive data

Kleiner et al.

June 24, 2020

Introduction

Two recent trends are worthy of attention in this regard.
First, the growth in size of data sets is accelerating, with 'massive' data sets becoming increasingly prevalent.
Second, computational resources are shifting towards parallel and distributed architectures, with multicore and cloud computing platforms providing access to hundreds or thousands of processors.
However, from an inferential point of view, it is not yet clear how statistical methodology will transport to a world involving massive data on parallel and distributed computing platforms.

Introduction

The bootstrap would seem ideally suited to exploiting the trend towards parallel and distributed computing: one might imagine using different processors or compute nodes to process different bootstrap resamples independently in parallel.
However, in the massive data setting, computation of even a single point estimate on the full data set can be quite computationally demanding.

Introduction

Another landmark in the development of simulation-based inference is subsampling and the closely related m out of n bootstrap.

- finite sample behaviour can be worse, and their success is sensitive to the choice of resample (or subsample) size.
- Although schemes have been proposed for data-driven selection of an optimal resample size (Bickel and Sakov, 2008), they require significantly greater computation which may eliminate any computational gains.

Introduction

bag of little bootstrap : BIB

Subsample + Boostrap+ Divide and Conquer

Setting and notation

- We assume that we observe a sample $X_1, \ldots, X_n$ drawn i.i.d. from some unknown distribution P
- We denote $\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$, the corresponding empirical distribution.
- On the basis of the observed data, we form an estimate $\hat{\theta}_n = \hat{\theta}_n(\mathbb{P}_n)$ of some unknown population value $P(\theta)$.
- Our end goal is to obtain an assessment $\xi\{Q_n(P)\}$ of the quality of the estimate $\hat{\theta}_n(\mathbb{P}_n)$, which consists of a summary of the distribution $Q_n(P)$ of some quantity $u(\mathbb{P}_n, P)$
- The choice of u depends on one's inferential goals.

Setting and notation

In practice, we cannot compute $\xi\{Q_n(P)\}$ directly because P and $Q_n(P)$ are unknown, and so we must estimate $\xi\{Q_n(P)\}$ on the basis of a single observed data set.

Under this notation, the bootstrap simply computes the plug-in approximation $\xi\{Q_n(\mathbb{P}_n)\} \approx \xi\{Q_n(P)\}$

Note that the vast majority of the bootstrap's computational cost lies in the repeated computation of values of u, which in turn requires costly repeated computation of estimates $\hat{\theta}_n(\mathbb{P}_n^*)$ on resamples.

BLB

- Given a subset size $b < n$, the BLB samples s subsets of size b without replacement from the original n data points, uniformly at random (one can also impose the constraint that the subsets are disjoint).
- Let $\mathcal{I}_1, \ldots, \mathcal{I}_s \subset \{1, \ldots, n\}$ be the corresponding index sets.
- Let $\mathbb{P}_{n,b}^{(j)} = b^{-1} \sum_{i \in \mathcal{I}_j} \delta_{X_i}$ denote the empirical distribution corresponding to subset j.
- The BLB's estimate of $\xi\{Q_n(P)\}$ is then given by

$$s^{-1} \sum_{j=1}^{s} \xi \left\{ Q_n \left(\mathbb{P}_{n,b}^{(j)} \right) \right\}. \tag{1}$$

BLB

Although the terms $\xi\left\{Q_n\left(\mathbb{P}_{n,b}^{(j)}\right)\right\}$ in expression (1) cannot be computed analytically in general, they can be computed numerically via straightforward Monte Carlo approximation in the manner of the bootstrap.

For each term j, repeatedly resample n points IID from $\mathbb{P}_{n,b}^{(j)}$ form the empirical distribution $\mathbb{P}_{n,b}^{*}$ and compute $u\left(\mathbb{P}_{n,b}^{*}, \mathbb{P}_{n,b}^{(j)}\right)$ for each resample, form the empirical distribution $\mathbb{Q}_{n,j}^{*}$ of the computed u-values, and compute $\xi\left(\mathbb{Q}_{n,j}^{*}\right) \approx \xi\left\{Q_n\left(\mathbb{P}_{n,b}^{(j)}\right)\right\}$.

Algorithm 1: the BLB

Input: data $X_1, \ldots, X_n$; b, subset size; s, number of sampled subsets; r, number of Monte Carlo iterations; ξ, u, estimator quality assessment ξ summarizing the distribution of quantity u
Output: an estimate of $\xi\{Q_n(P)\}$
For $j \leftarrow 1$ to s do
  // subsample the data
  randomly sample a set $\mathcal{I} = \{i_1, \ldots, i_b\}$ of b indices from $\{1, \ldots, n\}$ without replacement (or, choose $\mathcal{I}$ to be a disjoint subset of size b from a predefined random partition of $\{1, \ldots, n\}$)
  $\mathbb{P}_{nb}^{(j)} \leftarrow b^{-1} \sum_{i \in \mathcal{I}} \delta_{X_i}$
  // approximate $\xi\{Q_n(\mathbb{P}_{nb}^{(j)})\}$
  for $k \leftarrow 1$ to r do
    sample $(n_1, \ldots, n_b) \sim \text{multinomial}(n, \mathbf{1}_b/b)$
    $\mathbb{P}_{nk}^{*} \leftarrow n^{-1} \sum_{a=1}^{b} n_a \delta_{X_{i_a}}$
    $u_{nk}^{*} \leftarrow u(\mathbb{P}_{nk}^{*}, \mathbb{P}_{nb}^{(j)})$
  end
  $\mathbb{Q}_{n,j}^{*} \leftarrow r^{-1} \sum_{k=1}^{r} \delta_{u_{nk}^{*}}$
  $\xi_{n,j}^{*} \leftarrow \xi(\mathbb{Q}_{n,j}^{*})$
end

The BLB

The BLB straightforwardly permits computation on multiple (or even all) subsamples and resamples simultaneously in parallel: because BLB subsamples and resamples can be significantly smaller than the original data set, they can be transferred to, stored by and processed on individual (or very small sets of) compute nodes.

Theorem 1

Suppose that $\hat{\theta}_n(\mathbb{P}_n) = \phi(\mathbb{P}_n)$ and $\theta(P) = \phi(P)$, where ϕ where ϕ is Hadamard differentiable at P tangentially to some subspace, with $P, \mathbb{P}_n$ and $\mathbb{P}_{n,b}^{(j)}$ viewed as maps from some Donsker class $\mathcal{F}$ to $\mathbb{R}$ such that $\mathcal{F}_\delta$ is measurable for every $\delta > 0$, where $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, \rho_P(f - g) < \delta\}$. Additionally, assume that $\xi\{Q_n(P)\}$ is a function of the distribution of $u(\mathbb{P}_n, P) = n^{1/2}\{\phi(\mathbb{P}_n) - \phi(P)\}$ which is continuous in the space of such distributions with respect to a metric that metrizes weak convergence. Then,

$$s^{-1}\sum_{j=1}^{s}\xi\left\{Q_n\left(\mathbb{P}_{n,b}^{(j)}\right)\right\} - \xi\{Q_n(P)\} \xrightarrow{\mathrm{P}} 0$$

as $n \to \infty$, for any sequence $b \to \infty$ and any fixed s.

Theorem 2

Suppose that $\xi\{Q_n(P)\}$ admits an expansion as an asymptotic series

$$\xi\{Q_n(P)\} = z + \frac{p_1}{\sqrt{n}} + \cdots + \frac{p_k}{n^{k/2}} + o\left(\frac{1}{n^{k/2}}\right),$$

where z is a constant independent of P and the p_k are polynomials in the moments of P. Additionally, assume that the empirical version of $\xi\{Q_n(P)\}$ for any j admits a similar expansion

$$\xi\left\{Q_n\left(\mathbb{P}_{n,b}^{(j)}\right)\right\} = z + \frac{\hat{p}_1^{(j)}}{\sqrt{n}} + \ldots + \frac{\hat{p}_k^{(j)}}{n^{k/2}} + o_P\left(\frac{1}{n^{k/2}}\right).$$

Therem 2 Continue

Then, assuming that $b \leq n$ and $E\left(\hat{p}_k^{(1)}\right)^2 < \infty$ for $k \in \{1, 2\}$,

$$\left| s^{-1} \sum_{j=1}^{s} \xi\left\{ Q_n\left(\mathbb{P}_{n,b}^{(j)}\right)\right\} - \xi\left\{Q_n(P)\right\} \right|$$

$$= \sum_{k=1}^{2} O_P\left[\frac{\sqrt{\left\{ \text{var}\left(\hat{p}_k^{(1)} - p_k | \mathbb{P}_n\right)\right\}}}{n^{k/2}\sqrt{s}} \right] + O_P\left(\frac{1}{n}\right) + O\left(\frac{1}{b\sqrt{n}}\right)$$

Theorem 2 Continue

Therefore, taking
$s = \Omega\left[\max\left\{n \operatorname{var}\left(\hat{p}_1^{(l)} - p_1|\mathbb{P}_n\right), \operatorname{var}\left(\hat{p}_2^{(1)} - p_2|\mathbb{P}_n\right)\right\}\right]$ and
$b = \Omega(\sqrt{n})$ yields

$$\left| s^{-1} \sum_{j=1}^{s} \xi\left\{Q_n\left(\mathbb{P}_{n,b}^{(j)}\right)\right\} - \xi\left\{Q_n(P)\right\} \right| = O_P\left(\frac{1}{n}\right),$$

in which case the BLB enjoys the same level of higher order correctness as the bootstrap.

Theorem 3

Theorem 3. Under the assumptions of theorem 2, and assuming that the BLB uses disjoint random subsets of the observed data (rather than simple random subsamples), we have

$$\left| s^{-1} \sum_{j=1}^{s} \xi \left\{ Q_n \left(\mathbb{P}_{n,b}^{(j)} \right) \right\} - \xi \left\{ Q_n(P) \right\} \right| = O_P \left\{ \frac{1}{\sqrt{(nbs)}} \right\} + O \left(\frac{1}{b\sqrt{n}} \right),$$

Therefore, if $s \sim n/b$ and $b = \Omega(\sqrt{n})$, then

$$\left| s^{-1} \sum_{j=1}^{s} \xi \left\{ Q_n \left(\mathbb{P}_{n,b}^{(j)} \right) \right\} - \xi \left\{ Q_n(P) \right\} \right| = O_P \left(\frac{1}{n} \right).$$

Simulation

- We consider two different settings: regression and classification.
- For both setting, the data have the form $X_i = \left(\tilde{X}_i, Y_i\right) \sim P$, IID for $i = 1, \ldots, n$ where $\tilde{X}_i \in \mathbb{R}^d$; $Y_i \in \mathbb{R}$ for regression whereas $Y_i \in \{0, 1\}$ for classification.
- In each case, $\hat{\theta}_n$ estimates a parameter vector in $\mathbb{R}^d$ for a linear or generalized linear model of the mapping between $\tilde{X}_i$ and Y_i.

Simulation

- We define ξ as a procedure that computes a set of marginal 95% confidence interval, one for each element of the estimated parameter vector.
- In particular, given the distribution $Q_n(P)$ of $u\left(\mathbb{P}_n, P\right) = \hat{\theta}_n\left(\mathbb{P}_n\right)$, ξ forms the boundaries of the relevant confidence intervals as the 2.5th and 97.5th percentiles of the marginal componentwise distributions defined by $Q_n(P)$.
- Averaging across these confidence intervals in the averaging step of the BLB simply consists in averaging these percentile estimates.

Simulation

- To evaluate the various quality assessment procedures on a given estimation task and true underlying data distribution P, we first compute the ground truth $\xi\{Q_n(P)\}$ by generating 2000 realizations of data sets of size n from P, computing $\hat{\theta}_n$ on each, using this collection of $\hat{\theta}_n$s to form a high fidelity approximation to $Q_n(P)$.
- Each estimate is evaluated on the basis of the average relative deviation of its componentwise confidence intervals' widths from the corresponding true width: $|c - c_0|/c_0$.
- For the BLB, the b out of n bootstrap, and subsampling, we consider $b = n^\gamma$ with $\gamma \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$; we use $r = 100$ in all runs of BLB.

Simulation: Regression

- In the regression setting, we generate each data set from a true underlying distribution P consisting of either a linear model $Y_i = \tilde{X}_i 1_d + \varepsilon_i$ or a model $Y_i = \tilde{X}_i^{\mathrm{T}} 1_d + \tilde{X}_i^{\mathrm{T}} \tilde{X}_i + \varepsilon_i$ having a quadratic term, with $d = 1000$ and $n = 20000$.
- The $\tilde{X}_i$ and ε_i are drawn independently from one of the following pairs of distributions:
  - $\tilde{X}_i \sim \mathit{normal}(0, 1_d)$ with $\varepsilon_i \sim \mathit{normal}(0, 10)$,
  - $\tilde{X}_{i,j} \sim \mathit{StudentT}(3)$ IID with $\varepsilon_i \sim \mathit{normal}(0, 10)$,
  - $\tilde{X}_{i,j} \sim \mathit{gamma}\{1 + 5(j-1)/\max(d-1, 1), 2\} - 2\{1 + 5(j-1)/\max(d-1, 1), 2\}$ IID with $\varepsilon_i \sim \mathit{gamma}(1, 2) - 2$.

Simulation : Regression

Fig 1: Relative error versus processing time for the regression setting

Simulation: Classification

- We generate each data set considered from either a linear model $Y_i \sim Bernuoulli[\{1 + \exp(-\tilde{X}_i^T 1_d)\}^{-1}]$ or a model $Y_i \sim Bernuoulli[\{1 + \exp(-\tilde{X}_i^T 1_d - \tilde{X}_i^T \tilde{X}_i)\}^{-1}]$ having a quadratic term, with $d = 10$.
- We use the same distribution on $\tilde{X}_i$ that were used in the regression setting.
- Our estimator,under both the linear and the quadratic data-generating distributions, consists of a linear logistic regression fit via Newton's method.

Simulation: Classification

Fig 2: Relative error versus processing time for the classification setting

Fig 3: Relative error versus processing time for the classification setting

Computation Scalability

- Now, we use $d = 3000$ and $n = 6$ million so the size of full observed data set ia approximately 150 Gbytes.
- $r = 50, s = 5$ and $b = n^{0.7}$.
- 10 worker nodes, each having 6 Gbtyes of memory and eight compute cores.
- Fig 4: Relative error versus time on 150 Gbytes of data

Tuning Parameter

The BLB requires the specification of tuning parameters controlling the number of subsamples and resamples processed.

Fig 5: Results for BIB tuning parameter selection

Real Data

- University of California at Irvine connnect4 data set.
- The model is logistic regression with $d = 42, n = 67557$.
- We now report the average absolute confidence interval width.
- Fig 6: Average (across dimensions) absolute confidence interval width versus processing time

Time Series

- Variants of the bootstrap such as the moving block bootstrap and the stationary bootstrap have been proposed to handle other data analysis settings such as that of time series.
- These bootstrap variants can be used within the BLB.

Time Series

- To extend the BLB in this manner, we must simply alter both the subsample selection mechanism and the resample generation mechanism such that both of these processes respect the underlying data-generating process.
- In particular,for stationary time series data it suffices to select each subsample as a(uniformly)randomly positioned block of length b within the observed time series of length n.
- Given a subsample of size b, we generate each resample by applying the stationary bootstrap to the subsample to obtain a series of length n.

Time Series

Given a $p \in [0, 1]$, we first select uniformly at random a data point in the subsample series and then repeat the following process until we have amassed a new series of length n: with probability 1-p we append to our resample the next point in the subsample series(wrapping around to the beginning if we reach the end of the subsample series),and with probability p we (uniformly at random) select and append a new point in the subsample series.

Time Series

- We generate observed data consisting of a stationary time series $X_1, \ldots, X_n \in \mathbb{R}$ where $X_t = Z_t + Z_{t-1} + Z_{t-2} + Z_{t-3} + Z_{t-4}$ and Z_t are drawn independently from a normal(0,1)distribution.
- We consider the task of estimating the standard deviation of the rescaled mean $\sum_{t=1}^{n} X_t/\sqrt{n}$.
- We set $p = 0.1, n = 5000$.

Time Series

Table 4. Comparison of the standard and stationary bootstrap and the BLB on stationary time series data with $n = 5000$†

Method	*Results for standard method*	*Results for stationary method*
BLB-0.6	2.2 ± 0.1	4.2 ± 0.1
BLB-0.7	2.2 ± 0.04	4.5 ± 0.1
BLB-0.8	2.2 ± 0.1	4.6 ± 0.2
BLB-0.9	2.2 ± 0.1	4.6 ± 0.1
Bootstrap	2.2 ± 0.1	4.6 ± 0.2

†We report the average and standard deviation of estimates (after convergence) of the standard deviation of the rescaled mean aggregated over 10 trials. The true population value of the standard deviation of the rescaled mean is approximately 5.