A resampling technique for massive data in settings of bootstrap inconsistency

Tim White

May 2022

Motivation and roadmap

How do we **efficiently** and **accurately** assess the quality of an estimator in the **massive data** setting under **nonideal** data-generating conditions?

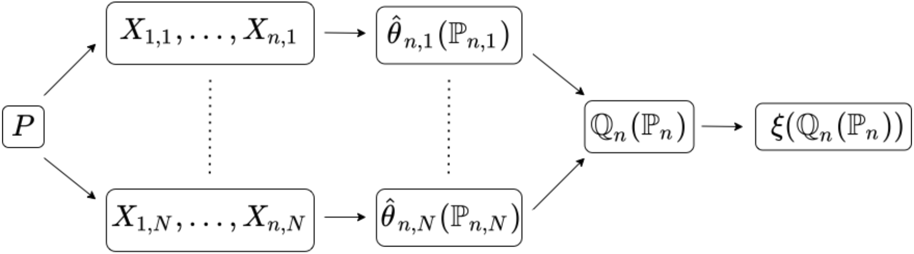
1. Notation
2. Overview of existing methods
    - Frequentist ideal
    - Bootstrap
    - Bag of little bootstraps (BLB)
3. Bag of little *m* out of *n* bootstraps (BLmnB)
4. Simulation framework
5. Simulation results
6. Discussion
    - Summary of results
    - Limitations and future research

Notation

- We adopt our notation from Kleiner et al. (2014) and Garnatz (2015)
- Random sample $X_1, ..., X_n \overset{\text{iid}}{\sim} P$
  - Empirical distribution is $\mathbb{P}_n$
- Parameter of interest: $\theta(P)$
- Estimator: $\widehat{\theta}_n(\mathbb{P}_n)$
- Sampling distribution of $\widehat{\theta}_n(\mathbb{P}_n)$: $Q_n(P)$
- Estimator quality assessment: $\xi(Q_n(P))$
  - e.g., Confidence interval

Frequentist ideal

1. Draw N iid samples from the population
2. Compute a realization of the estimator on each sample
3. Use these N realizations to form an accurate approximation of the sampling distribution
4. Apply ξ to this approximation of the sampling distribution



Frequentist ideal *(cont.)*

Drawback:

We generally do not have the time or resources to collect N independent samples. Often, we have access to just one sample.

Nonparametric bootstrap

1. Draw one sample from the population
2. Generate r resamples of size n from the original sample
3. Compute a realization of the estimator on each resample
4. Form an empirical sampling distribution from these r realizations
5. Apply ξ to this approximation of the sampling distribution

$$
\begin{tikzcd}
 & & X^*_{1,1}, \ldots, X^*_{n,1} \arrow[r] & \hat{\theta}^*_{n,1}(\mathbb{P}^*_{n,1}) \arrow[dr] & & \\
P \arrow[r] & X_1, \ldots, X_n \arrow[ur] \arrow[dr] & \vdots & \vdots & \mathbb{Q}^*_n(\mathbb{P}_n) \arrow[r] & \xi(\mathbb{Q}^*_n(\mathbb{P}_n)) \\
 & & X^*_{1,r}, \ldots, X^*_{n,r} \arrow[r] & \hat{\theta}^*_{n,r}(\mathbb{P}^*_{n,r}) \arrow[ur] & &
\end{tikzcd}
$$

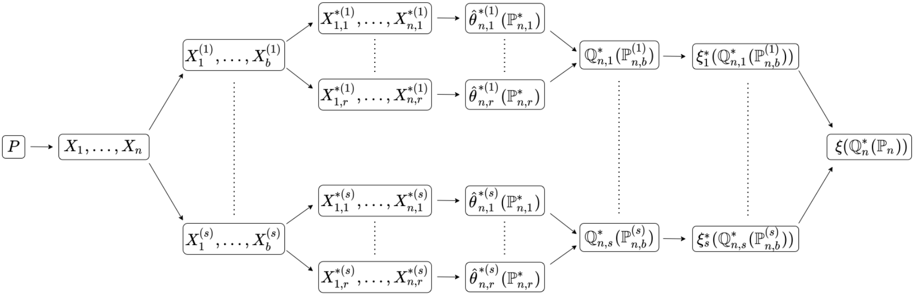
Nonparametric bootstrap *(cont.)*

Drawback:

Each resample of size n contains approximately $0.632n$ distinct data points from the original sample (Efron & Tibshirani, 1993), so the bootstrap becomes computationally inefficient as $n \to \infty$.

Bag of little bootstraps (BLB)

1. Draw one sample from the population
2. Generate s subsamples of size $b << n$ from the original sample
3. For each subsample,
   - Generate r resamples of size n
   - Compute a realization of the estimator on each resample
   - Assemble these r realizations into an empirical sampling distribution
   - Apply ξ to this approximation of the sampling distribution
4. Take the average of the s estimator quality assessments



Bag of little bootstraps (BLB) *(cont.)*

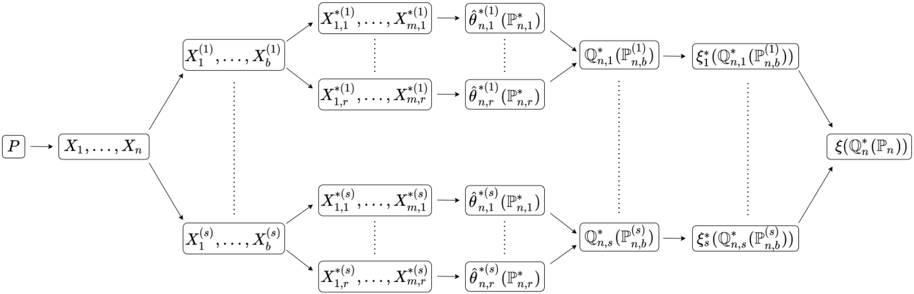
Drawback:

"...although the BLB shares the statistical strengths of the bootstrap, we conversely do *not* expect our procedure to be applicable in cases in which the bootstrap fails.

It would presumably be possible to extend the applicability of the BLB to such edge cases by modifying it to apply the m out of n bootstrap or subsampling, rather than the bootstrap, to each subsample (e.g., thus yielding the 'bag of little m out of n bootstraps')" (Kleiner et al., 2014).

Bag of little m out of n bootstraps (BLmnB)

1. Draw one sample from the population
2. Generate s subsamples of size $b << n$ from the original sample
3. For each subsample,
   - $\boxed{\text{Generate } r \text{ resamples of size } m < n}$
   - Compute a realization of the estimator on each resample
   - Assemble these r realizations into an empirical sampling distribution
   - Apply ξ to this approximation of the sampling distribution
4. Take the average of the s estimator quality assessments



Simulation framework

Setting 1: $N(\mu, 1)$

Case (a): $N(\mu, 1)$ where $-\infty < \mu < \infty$, data-generating $\mu = 0$

- MLE is $\widehat{\mu}_n = \overline{X}_n$
- Bootstrap, BLB, and (presumably) BLmnB are consistent

Case (b): $N(\mu, 1)$ where $\mu \in [0, \infty)$, data-generating $\mu = 0$

- MLE is $\widehat{\mu}_n = \max\{0, \overline{X}_n\}$ (Andrews, 2000)
- Bootstrap and BLB are **inconsistent**

Simulation framework *(cont.)*

Setting 2: Unif$(0, \theta)$**, data-generating** $\theta = 1$

- MLE is $\widehat{\theta}_n = \max\{X_1, ..., X_n\}$ (Hogg et al., 2019)
- Bootstrap and BLB are **inconsistent**

Simulation framework *(cont.)*

We define ξ to be a 95% percentile confidence interval

1. Compute the "ground truth"
   - Generate 10,000 iid samples of size $n = 20,000$ from P
   - Compute a realization of the estimator on each sample
   - Assemble these n realizations into an empirical sampling distribution
   - Apply ξ to this sampling distribution and compute its width w
2. Generate a new sample of size $n = 20,000$ from P
3. Run each method on this sample
   - Obtain $\xi(\mathbb{Q}_n^*(\mathbb{P}_n))$ for each method and compute its width w^*
4. Compute the relative error $|w^* - w|/w$ for each method
5. Repeat steps 2-4 for nine more iid samples, then take the average of the ten relative errors for each method to obtain a final summary of approximation accuracy

Simulation framework *(cont.)*

Bootstrap	BLB	BLmnB
Number of Monte Carlo iterations r	Number of Monte Carlo iterations r	Number of Monte Carlo iterations r
	Number of subsamples s	Number of subsamples s
	Subsample size $b = n^\gamma$	Subsample size $b = n^\gamma$
		Resample size $m = b^\alpha = n^{\alpha\gamma}$

- $r \in \{25, 50, 100, 250, 500\}$
- $s \in \{10, 20, 40, 100, 200\}$
- $\gamma \in \{0.5, 0.7, 0.9\}$
- $\alpha \in [1, 1/\gamma)$
  - Evenly spaced sequence starting $100x\%$ of the way between b and n

Simulation results

Setting 1, case (a): $N(\mu, 1), -\infty < \mu < \infty$, true $\mu = 0$

Bootstrap

r	Average relative error
25	0.17
50	0.18
100	0.07
250	0.07
500	0.04

BLB

$\gamma = 0.7$		s				
		10	20	40	100	200
r	25	0.11	0.12	0.13	0.14	0.13
	50	0.09	0.09	0.08	0.08	0.08
	100	0.05	0.05	0.04	0.04	0.04
	250	0.03	0.03	0.01	0.02	0.02
	500	0.02	0.01	0.01	0.01	0.01

BLmnB

$\gamma = 0.7$		s				
		10	20	40	100	200
m = 11,446	r = 25	0.13	0.13	0.14	0.15	0.14
	50	0.21	0.20	0.22	0.21	0.22
	100	0.25	0.27	0.26	0.26	0.26
	250	0.29	0.30	0.30	0.29	0.30
	500	0.31	0.31	0.31	0.31	0.31
m = 17,267	25	0.06	0.04	0.07	0.07	0.07
	50	0.04	0.02	0.02	0.01	0.01
	100	0.03	0.03	0.04	0.03	0.03
	250	0.06	0.06	0.05	0.06	0.06
	500	0.06	0.06	0.06	0.06	0.07
m = 19,803	25	0.14	0.14	0.14	0.14	0.13
	50	0.06	0.08	0.08	0.08	0.07
	100	0.03	0.04	0.04	0.04	0.04
	250	0.02	0.02	0.01	0.02	0.01
	500	0.01	0.01	0.01	0.01	0.00

Simulation results *(cont.)*

Setting 1, case (b): $N(\mu, 1), \mu \in [0, \infty)$, **true** $\mu = 0$

Bootstrap

r	Average relative error
25	0.46
50	0.39
100	0.42
250	0.38
500	0.40

BLB

$\gamma = 0.7$		s				
		10	20	40	100	200
r	25	0.28	0.16	0.16	0.11	0.16
	50	0.28	0.27	0.16	0.13	0.09
	100	0.45	0.19	0.15	0.22	0.13
	250	0.39	0.18	0.22	0.21	0.20
	500	0.37	0.24	0.29	0.24	0.17

BLmnB

$\gamma = 0.7$		s				
		10	20	40	100	200
m = 11,446	r = 25	0.29	0.37	0.24	0.23	0.21
	50	0.47	0.34	0.41	0.37	0.33
	100	0.42	0.52	0.42	0.40	0.32
	250	0.50	0.45	0.37	0.46	0.49
	500	0.81	0.53	0.50	0.54	0.53
m = 17,267	25	0.19	0.28	0.18	0.10	0.11
	50	0.25	0.20	0.13	0.14	0.10
	100	0.18	0.29	0.22	0.19	0.14
	250	0.28	0.31	0.27	0.23	0.20
	500	0.32	0.41	0.25	0.28	0.28
m = 19,803	25	0.32	0.22	0.16	0.17	0.13
	50	0.40	0.16	0.22	0.10	0.12
	100	0.25	0.27	0.19	0.19	0.12
	250	0.25	0.24	0.25	0.16	0.14
	500	0.34	0.15	0.25	0.25	0.18

Simulation results *(cont.)*

Setting 2: Unif$(0, \theta)$, true $\theta = 1$

Bootstrap

r	Average relative error
25	0.62
50	0.58
100	0.63
250	0.51
500	0.51

BLB

$\gamma = 0.9$		s				
		10	20	40	100	200
r	25	0.69	0.68	0.70	0.63	0.66
	50	0.59	0.57	0.56	0.58	0.59
	100	0.44	0.44	0.53	0.53	0.52
	250	0.44	0.34	0.37	0.44	0.41
	500	0.50	0.25	0.36	0.30	0.32

BLmnB

$\gamma = 0.9$		s				
		10	20	40	100	200
m = 11,486	25	0.33	0.24	0.20	0.20	0.16
	50	0.23	0.19	0.23	0.15	0.14
	100	0.15	0.22	0.24	0.24	0.15
	250	0.25	0.21	0.18	0.15	0.17
	500	0.28	0.23	0.16	0.19	0.15
m = 12,808	25	0.24	0.20	0.25	0.16	0.21
	50	0.19	0.27	0.21	0.17	0.17
	100	0.28	0.22	0.20	0.25	0.21
	250	0.30	0.28	0.27	0.27	0.24
	500	0.36	0.30	0.25	0.26	0.28
m = 14,282	25	0.31	0.38	0.29	0.29	0.27
	50	0.31	0.24	0.25	0.23	0.22
	100	0.25	0.31	0.25	0.24	0.24
	250	0.33	0.37	0.27	0.27	0.30
	500	0.31	0.30	0.27	0.25	0.24

Summary of results

Setting 1, case (a): $N(\mu, 1), -\infty < \mu < \infty$, **true** $\mu = 0$

- Low relative error for all three methods
- BLmnB comparable to BLB and better than bootstrap for sufficiently large *m*

Setting 1, case (b): $N(\mu, 1), \mu \in [0, \infty)$, **true** $\mu = 0$

- Again, BLmnB comparable to BLB and better than bootstrap for sufficiently large *m*
- No advantage to using BLmnB instead of BLB

Summary of results (cont.)

Setting 2: Unif$(0, \theta)$, **true** $\theta = 1$

- Bootstrap and BLB both perform poorly
- BLmnB outperforms bootstrap *and* BLB for $m \in [11,000, 14,000]$
- Performance of BLmnB depends on choice of m

Limitations and future research

1. Did not develop the theory underlying the BLmnB — consistent and higher order correct?
   - Samworth (2003) shows that *m* out of *n* bootstrap is generally consistent in cases of bootstrap inconsistency

Limitations and future research (cont.)

2. Considered only one choice of $\xi(Q_n(P))$
3. Considered only two settings of bootstrap inconsistency
4. Did not compare the runtimes of the three methods

Limitations and future research (cont.)

5. Why does BLmnB perform better in one setting than the other?
6. What is the optimal m and how do we obtain it?

References

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, *68*(2), 399–405.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*(1), 1–26.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Garnatz, C. (2015). *Trusting the black box: Confidence with bag of little bootstraps*. Pomona College.

Hogg, R. V., McKean, J. W., & Craig, A. T. (2019). *Introduction to Mathematical Statistics* (8th ed.). Pearson.

Jordan, M. I. (2017). *A scalable bootstrap for massive data*. https://youtu.be/9v6cMnFNiOI

Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *76*(4), 795–816.

Samworth, R. (2003). A note on methods of restoring consistency to the bootstrap. *Biometrika*, *90*(4), 985–990.