STAT 460/560 Class 7: Bootstrap methods

Ben Bloem-Reddy

Reading: Chapter 8, [Was04]; Chapter 10-11, [EH21].

1. Bootstrap at a high level

The primary motivation for bootstrap methods was—and still is—to estimate uncertainty for estimators that do not admit simple expressions for variance. To illustrate, suppose that we have a statistic $T_n = g(X_1, \dots, X_n)$, where the data have unknown distribution F. If T_n is the sample mean $\bar{X}_n$ then we have a simple mathematical expression for $\mathrm{Var}_F(T_n)$ (the variance under F), and it's straightforward to estimate—usually with $\mathrm{Var}_{\hat{F}_n}(T_n)$. On the other hand, if T_n is the sample median, that is not the case.

At a high level, the bootstrap has two steps:

1. Estimate $\mathrm{Var}_F(T_n)$ with $\mathrm{Var}_{\hat{F}_n}(T_n)$.
2. Use simulation to approximate $\mathrm{Var}_{\hat{F}_n}(T_n)$.

Sometimes the second step is unnecessary, i.e., when we can derive $\mathrm{Var}_{\hat{F}_n}(T_n)$ in closed form.

2. Simulation and Monte Carlo integration

Let Y be a random variable with distribution G and f some function such that $\mathbb{E}(|f(Y)|) < \infty$. The Weak Law of Large Numbers (LLN) tells us that if $Y_1, \dots, Y_B \sim_{\text{IID}} G$, then as $B \to \infty$,

$$\frac{1}{B}\sum_{i=1}^{B} f(Y_i) \xrightarrow{\text{P}} \int f(y)dG(y) = \mathbb{E}(f(Y)) \,. \tag{7.1}$$

The basic idea behind **Monte Carlo integration** is that one can get numerical estimates of integrals by sampling and averaging.

Here's an example. The code below estimates

$$I_1 = \int_{-\infty}^{\infty} x\phi(x)dx$$

and

$$I_2 = \int_{-\infty}^{\infty} xe^{x^2 - x^4}\phi(x)dx \,,$$

where ϕ is the standard normal PDF. (Quick! What should these be?)

```
library(ggplot2)
set.seed(111)
B <- 50000

Y <- rnorm(B, mean = 0, sd = 1)

# define custom function
f <- function(x){
```

```
  return(x*exp(x^2 - x^4))
}

I1 <- sum(Y)/B
I2 <- sum(f(Y))/B

# report estimates
I1 # -0.009006043
I2 # -0.0003849936
```

Exercise 7.1. Let $Y = X^3$, where $X \sim \mathcal{N}(0,1)$. Use $B = 10000$ samples to estimate $\mathrm{Var}(Y)$.

3. Bootstrap variance estimation

We want to estimate $\mathrm{Var}_F(T_n)$. If we knew F, we could just use the ideas from the previous section by simulating datasets $X^*_{1,b}, \ldots, X^*_{n,b}$, $b = 1, \ldots, B$. But we don't know F. So what can we do?

Bootstrapping techniques sample pseudo-datasets from an estimate of F. The most obvious estimate is the eCDF $\hat{F}_n$. This is known as the **nonparametric bootstrap**. The idea here is that if $\hat{F}_n$ is close to F, then $\mathrm{Var}_{\hat{F}_n}(T_n)$ will be close to $\mathrm{Var}_F(T_n)$. What does sampling from $\hat{F}_n$ look like? Recall that $\hat{F}_n$ is the CDF of the distribution that assigns mass $1/n$ to each observed value. So sampling from $\hat{F}_n$ is just sampling with replacement from $X_1, \ldots, X_n$. We denote such a bootstrap sample by $X^*_1, \ldots, X^*_n$, and $T^*_n = T(X^*_1, \ldots, X^*_n)$.

We can use the bootstrap variance as an estimator for $\mathrm{Var}_F(T_n)$,

$$v_{\mathrm{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T^*_{n,b} - \bar{T}^*_n\right)^2 , \quad \text{with} \quad \bar{T}^*_n = \frac{1}{B}\sum_{r=1}^{B} T^*_{n,r} . \tag{7.2}$$

Here's an example: a bootstrap estimate of the variance of the correlation coefficient.

```
#####
library(MASS) # for sampling from MVN
set.seed(101)

# sample size
n <- 100

# generate "data"
Sig.cov <- matrix(c(2,1,1,3), nrow=2, ncol=2)
X <- mvrnorm(n, mu = c(0,0), Sigma = Sig.cov) # this is a nx2 matrix
rho.XY.hat <- cor(X)[1,2]

# bootstrap
B <- 10000
idx <- sample.int(n, size = n*B, replace = TRUE) # get all of the bootstrap indices in one go
bs.idx <- matrix(idx, nrow = B, ncol = n) # convert the indices in a B x n matrix
bs.rho.XY.hat <- rep(0, times = B)
for (b in 1:B){
  Xstar <- X[bs.idx[b,],]
  bs.rho.XY.hat[b] <- cor(Xstar)[1,2]
}

# estimate variance
se.bs.rho <- sqrt(var(bs.rho.XY.hat)*(B-1)/B) # 0.08763904
```

Exercise 7.2. Download the Old Faithful Geyser Data (https://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat) and use $B = 1000$ for a bootstrap estimate of the variance of the sample median waiting time and eruption time (separately). (If you're working in R, you can just use the `faithful` dataset.)

Also get a bootstrap estimate of the variance of the sample mean waiting time, and compare it to the usual estimator,

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \,. \tag{7.3}$$

There are two assumptions/conditions needed here in order for bootstrapping to work well. (We'll talk about theory in detail at the end of the term.) Firstly, we need $\hat{F}_n$ to be a decent estimate of F. When n is large this should be true (recall the uniform convergence bound from last class); for small n we should proceed with caution. Secondly, there's an assumption smuggled in that the map $F \mapsto \mathrm{Var}_F(T_n)$ is continuous so that if $\hat{F}_n$ is a good estimate of F then $\mathrm{Var}_{\hat{F}_n}(T_n)$ is a good estimate of $\mathrm{Var}_F(T_n)$.

4. Bootstrap confidence intervals

There are three simple ways to use bootstrap methods to construct approximate $1-\alpha$ confidence intervals.

1. **The Normal Interval**: The simplest is

   $$T_n \pm z_{\alpha/2}\sqrt{v_{\text{boot}}} \,. \tag{7.4}$$

   Although this is the simplest, it's best to avoid unless one has good reason to think that the distribution of T_n is normal.

2. **Pivotal Interval**: Let $\theta = T(F)$ and $\hat{\theta}_n = T(\hat{F}_n)$. Define the **pivot** $R_n = \hat{\theta}_n - \theta$. If we knew the CDF of the pivot,

   $$H(r) = \mathbb{P}(R_n \le r) \,, \tag{7.5}$$

   then we could compute an exact $1-\alpha$ confidence interval as $C_n^* = (a, b)$, where

   $$a = \hat{\theta}_n - H^{-1}(1-\alpha/2) \quad \text{and} \quad b = \hat{\theta}_n - H^{-1}(\alpha/2) \,. \tag{7.6}$$

   We don't actually know H, but we can estimate it via bootstrapping. Let $R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$, where $\hat{\theta}_{n,b}^*$ is the b-th bootstrap replicate of $\hat{\theta}_n$. Then

   $$\hat{H}(r) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{I}(R_{n,b}^* \le r) \,. \tag{7.7}$$

   Note that because $R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$, the q-th quantile of $(R_{n,1}^*, \ldots, R_{n,B}^*)$, denoted r_q^*, is equal to $\theta_q^* - \hat{\theta}_n$, where θ_q^* is the q-th quantile of $(\hat{\theta}_{n,1}^*, \ldots, \hat{\theta}_{n,B}^*)$.

   Then an approximate $1-\alpha$ confidence interval is $\hat{C}_n = (\hat{a}, \hat{b})$, where

   $$\hat{a} = \hat{\theta}_n - r_{1-\alpha/2}^* = 2\hat{\theta}_n - \theta_{1-\alpha/2}^* \tag{7.8}$$

   $$\hat{b} = \hat{\theta}_n - r_{\alpha/2}^* = 2\hat{\theta}_n - \theta_{\alpha/2}^* \,. \tag{7.9}$$

3. **Percentile interval**: We can just approximate the confidence interval with the bootstrap distribution, so that

   $$C_n = (\theta_{\alpha/2}^*, \theta_{1-\alpha/2}^*) \,. \tag{7.10}$$

Theorem 7.1. *Under weak conditions*[1] *on $T(F)$, the pivotal interval is asymptotically level $1-\alpha$. That is, as $n \to \infty$,*

$$\mathbb{P}_F(T(F) \in \hat{C}_n) \to 1 - \alpha \, . \tag{7.11}$$

Exercise 7.3. Calculate the three different confidence intervals for the sample median of the waiting time from the Old Faithful Geyser dataset. How do they compare?

References

[EH21] B. Efron and T. Hastie. *Computer Age Statistical Inference*. Cambridge University Press, 2021.

[Was04] L. Wasserman. *All of Statistics*. Springer New York, 2004.

[1] Details deferred until much later in the term.