Computation-based Inference: Simulation and Resampling Method

Tengyuan Liang[1]

[1] The University of Chicago Booth School of Business

DLA Lecture 2: Resample

Computation-based, non-formulaic approach to inference. Jackknife and the Bootstrap. Readings: Efron and Hastie [2], Chapter 10, and 11.

[2] Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, 2016

Contents

1 The Jackknife Estimate

A central element of frequentist inference is the standard error. Before the computer age, an applied statistician had to be a master in the Tayler series (or other expansion-based approaches) to produce an analytical formula for the estimate of a parameter of interest. It is a laborious effort, if not impossible, for complicated statistics.

The Jackknife (1957) provides a computation-based, nonformulaic approach to standard errors. Consider a case where the statistician has observed i.i.d. sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from an unknown probability distribution $F \in \mathcal{P}_{\mathcal{X}}$ on some space $\mathcal{X}$,

$$x_i \overset{i.i.d.}{\sim} F,\ i = 1, 2, \ldots, n\ .$$

A real-valued statistics $\hat{\theta}$ (or computation rule) has been computed by executing some algorithm $s(\cdot) : \mathcal{X}^{\otimes\infty} \to \mathbb{R}$

$$\hat{\theta} = s(\mathbf{x})\ . \tag{1.1}$$

TL: Here we assume the algorithm can execute for any sample size n, this is not crucial but convenient for subsequent discussions.

As $\mathbf{x}$ is random, we wish to assign a standard error to θ under the aforementioned i.i.d. sampling model.

Let $\mathbf{x}_{(i)}$ be the sample with x_i removed

$$\mathbf{x}_{(i)} := (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \tag{1.2}$$

and denote the corresponding, leave-one-out execution of the algorithm

$$\hat{\theta}_{(i)} = s(\mathbf{x}_{(i)}).$$

The **Jackknife estimate of the standard error** for $\hat{\theta}$ is

$$\widehat{\text{se}}_{\text{jack}} := \left[\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2\right]^{1/2}, \text{ where } \hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}\ . \tag{1.3}$$

Example. *Consider the case when the test statistic is the sample mean*

TL: The fudge factor $\frac{n-1}{n}$ makes the jackknife estimate coincide with the classic formula when $\hat{\theta} = \bar{x}$.

$$\begin{aligned}
\hat{\theta} &= \bar{x} \\
\hat{\theta}_{(i)} &= \frac{n\bar{x} - x_i}{n-1} \\
\hat{\theta}_{(\cdot)} &= \bar{x} \\
\widehat{\text{se}}_{\text{jack}} &= \left[\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}
\end{aligned}$$

A few remarks on the intuition behind Jackknife follow

- $\widehat{\text{se}}_{\text{jack}}$ can be applied in an automatic way to any statistics $\hat{\theta} = s(\mathbf{x})$, as long as s can be computed based on any sample size. **Computer power is substituted for analytical Taylor series calculations.** For example, when $s(\mathbf{x})$ measure the sample correlation

$$\begin{aligned}
s(\mathbf{x}) &= \frac{\hat{\mu}_{11}}{\sqrt{\hat{\mu}_{20}\hat{\mu}_{02}}} \\
\text{where } \hat{\mu}_{hk} &:= \tfrac{1}{n}\textstyle\sum(x_i - \bar{x})^h(y_i - \bar{y})^k,\ h,k \in \mathbb{N}_{\geq 0}
\end{aligned}$$

  The Taylor series formula looks formidable

$$\widehat{\text{se}}_{\text{taylor}} := \left\{\frac{\frac{\hat{\mu}_{11}^2}{\hat{\mu}_{20}\hat{\mu}_{02}}}{4n}\left[\frac{\hat{\mu}_{40}}{\hat{\mu}_{20}^2} + \frac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2} + \frac{2\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}} + \frac{4\hat{\mu}_{22}}{\hat{\mu}_{11}^2} - \frac{4\hat{\mu}_{31}}{\hat{\mu}_{11}\hat{\mu}_{20}} - \frac{4\hat{\mu}_{13}}{\hat{\mu}_{11}\hat{\mu}_{02}}\right]\right\}^{1/2} \tag{1.4}$$

  it saves us all the tedious analytic calculations.

- It is nonparametric w.r.t F, no assumption needed.

- The algorithm works with data sets of size $n-1$, not n. There is a hidden assumption of smooth behavior across sample sizes. This can be worrisome for statistics such as sample median, which have a different definition for odd and even sample sizes.

- The jackknife standard error is **upwardly biased** as an estimate of the true standard error. We will give a connection to the Efron-Stein-Steele Inequality later.
- Jackknife is approximating the directional derivatives

$$\widehat{\text{se}}_{\text{jack}} = \left[\frac{\sum_{i=1}^n D_i^2}{n^2}\right]^{1/2}, \text{ where } D_i := \frac{\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}}{1/\sqrt{n(n-1)}} \tag{1.5}$$

  Here D_i is measuring as we decrease the weight on data point x_i, how fast the statistics $s(\mathbf{x})$ is changing. We will see a connection in the "Resampling Plans" section soon.

  In particular, this directional derivative viewpoint is also closely related to the influence function, as well as Efron-Stein-Steele inequality.

1.1 A Connection to Efron-Stein-Steele Inequality

Theorem 1 (Efron-Stein-Steele Inequality, or Influence Inequality)**.** *Suppose $x_1, \dots x_n, x_1', \dots, x_n'$ are i.i.d. drawn from F Denote the $i-$th coordinate replacement*

$$\mathbf{x}^{(i)} = (x_1, x_2, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n) \tag{1.6}$$

then

$$\text{var}[s(\mathbf{x})] \leq \sum_{i=1}^n \frac{1}{2}\mathbb{E}\left[\left(s(\mathbf{x}) - s(\mathbf{x}^{(i)})\right)^2\right] \tag{1.7}$$

The proof follows from elementary facts of Martingale differences, see Boucheron, Lugosi, and Massart.

If this is not immediate to see the connection to the jackknife estimate, we simply replace the RHS of the above with the following equality

$$\frac{1}{2}\mathbb{E}\left[\left(s(\mathbf{x}) - s(\mathbf{x}^{(i)})\right)^2\right] = \mathbb{E}\left[\left(s(\mathbf{x}) - \underbrace{\mathbb{E}[s(\mathbf{x})|\mathbf{x}_{(i)}]}_{\text{leave-one-out}}\right)^2\right] \tag{1.8}$$

and therefore

$$\text{var}[s(\mathbf{x})] \leq \sum_{i=1}^n \mathbb{E}\left[\left(s(\mathbf{x}) - \underbrace{\mathbb{E}[s(\mathbf{x})|\mathbf{x}_{(i)}]}_{\text{leave-one-out}}\right)^2\right] \tag{1.9}$$

Contrast with the jackknife

$$\widehat{\text{se}}_{\text{jack}}^2 := \frac{n-1}{n}\sum_{i=1}^n \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2 \tag{1.10}$$

Both measure the influence of i-th coordinate.

2 *The Nonparametric Bootstrap*

The frequentist standard error of an estimate $\hat{\theta} = s(\mathbf{x})$ is, ideally, the standard deviation we would observe by repeatedly sampling new versions of $\mathbf{x}$ from F. This is impossible because we do not know F. **The bootstrap substitutes an estimate $\widehat{F}$ for F and then estimates the frequentist standard error by direct simulation, a feasible tactic only since the advent of electronic computation.**

Recall that given the sample and the statistic

$$\mathbf{x} = (x_1, x_2, \ldots, x_n), \quad \hat{\theta} = s(\mathbf{x}) \; . \tag{2.1}$$

Let us begin with the notion of a **bootstrap sample**, constructed by sampling with replacement where each

$$x_i^\star \overset{i.i.d.}{\sim} \hat{F} := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \; , \tag{2.2}$$

the empirical distribution of F based on the original sample. Each **bootstrap sample** provides a **bootstrap replication** of the statistics

$$\mathbf{x}^\star = (x_1^\star, x_2^\star, \ldots, x_n^\star), \quad \hat{\theta}^\star = s(\mathbf{x}^\star) \; . \tag{2.3}$$

Some large number B of bootstrap samples are independently drawn. The corresponding bootstrap replications are calculated, say

$$\hat{\theta}^{\star b} = s(\mathbf{x}^{\star b}), \; b = 1, 2, \ldots, B. \tag{2.4}$$

The resulting bootstrap estimate of standard error for $\hat{\theta}$ is the empirical sandard deviation of the $\hat{\theta}^{\star b}$ values

$$\widehat{\mathrm{se}}_{\mathrm{boot}} := \left[\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{\star b} - \hat{\theta}^{\star \cdot} \right)^2 \right]^{1/2}, \text{ where } \hat{\theta}^{\star \cdot} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{\star b} \tag{2.5}$$

To see the main motivation behind the bootstrap, we illustrate with the following diagram

$$F \xrightarrow{i.i.d.} \mathbf{x} \xrightarrow{s} \hat{\theta} \tag{2.6}$$

The bootstrap flow diagram reads

$$\hat{F} \xrightarrow{i.i.d.} \mathbf{x}^\star \xrightarrow{s} \hat{\theta}^\star \tag{2.7}$$

To assess the standard error of $\mathrm{sd}(\hat{\theta})$, follow the above diagram, we can define a functional $\mathrm{SD} : \mathcal{P}_{\mathcal{X}} \to \mathbb{R}_{\geq 0}$

$$\mathrm{SD} : F \mapsto \mathrm{sd}(\hat{\theta}) \tag{2.8}$$

where

$$\mathrm{SD}[F] := \left[\mathop{\mathbb{E}}_{\mathbf{x} \sim F^{\otimes n}} \left(s(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathbf{x}' \sim F^{\otimes n}} s(\mathbf{x}') \right)^2 \right]^{1/2} \tag{2.9}$$

In this notation

$$\lim_{B\to\infty} \widehat{\text{se}}_{\text{boot}} = \text{SD}[\widehat{F}] \tag{2.10}$$

Due to the fact that

$$\widehat{F} \to F \tag{2.11}$$

in certain topological sense in the space of probability distributions , if the functional $\text{SD}[F]$ is smooth w.r.t the metric, then

TL: say in Wasserstein metric, weak-* metric, etc.

$$\lim_{B\to\infty} \widehat{\text{se}}_{\text{boot}} = \text{SD}[\widehat{F}] \to \text{SD}[F] \tag{2.12}$$

Some remarks follow

- Automatic based on simulations and resampling
- Nonparametric. The parametric bootstrap will be described later.
- Bootstrap "shakes" the original data more violently than jackknife, producing non-local deviations of $\mathbf{x}^\star$ from $\mathbf{x}$. *The bootstrap is more dependable than the jackknife for unsmooth statistics since it does not depend on local derivatives.*
- There is nothing special about standard errors, say expected absolute error, or any other accuracy measure is fine.
- Might be intensive computationally.

3 *Resampling Plans*

In this section, we consider a resampling framework that unifies bootstrap and jackknife. Again, the whole section is conditioned on the original data vector $\mathbf{x}$, and consider perturbations to the empirical distribution which assigns equal weights to each data point.

A **resampling vector** $\mathbf{P} \in \Delta_n$ is a vector of nonnegative weights that sum up to 1,

$$\mathbf{P} = (P_1, P_2, \ldots, P_n) \tag{3.1}$$

Then the original statistic, the jackknife (leave-one-out), and bootstrap replication can all be represented with a specific resampling vector

$$\hat{\theta} = S(\mathbf{P}_0), \text{ where } \mathbf{P}_0 := (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$$

$$\hat{\theta}_{(i)} = S(\mathbf{P}_{(i)}) \text{ where } \mathbf{P}_{(i)} := (\frac{1}{n-1}, \frac{1}{n-1}, \ldots, \frac{1}{n-1}, \overbrace{0}^{i-th}, \frac{1}{n-1}, \ldots \frac{1}{n-1})$$

$$\hat{\theta}^\star = S(\mathbf{P}^\star) \text{ where } \mathbf{P}^\star := (\frac{N_1}{n}, \frac{N_2}{n}, \ldots, \frac{N_n}{n})$$

Here

$$N_i = \#\{x_j^\star = x_i\}$$

is sampling n balls from a bag of n unique balls with replacement, the number of times when the i-th ball occurs, and follows a multinomial distribution

$$\mathbf{N} = (N_1, N_2, \ldots, N_n) \sim \text{Multi}_n(n, \mathbf{P}_0) \tag{3.2}$$

This gives the bootstrap probability

$$\frac{n!}{N_1!N_2!\cdots N_n!}\frac{1}{n^n} \tag{3.3}$$

on $\mathbf{P}^\star$.

- Bootstrap vs. Jackknife. The Euclidean distance between a jackknife (leave-one-out) to the original data is

$$\|\mathbf{P}_{(i)} - \mathbf{P}_0\| = \frac{1}{\sqrt{n(n-1)}} \tag{3.4}$$

  For the bootstrap, $N_i \sim \text{Binom}(n, 1/n)$ which has mean 1 and variance $(n-1)/n$, then for the bootstrap vector $\mathbf{P}_i^\star = N_i/n$

$$\mathbb{E}[\mathbf{P}_i^\star] = \frac{1}{n} \tag{3.5}$$

$$\text{var}[\mathbf{P}_i^\star] = \frac{n-1}{n^3} \tag{3.6}$$

$$\text{cov}(\mathbf{P}_i^\star, \mathbf{P}_j^\star) = -\frac{1}{n^3} \tag{3.7}$$

  and that

$$\left(\mathbb{E}\,\|\mathbf{P}^\star - \mathbf{P}_0\|^2\right)^{1/2} = \sqrt{\frac{(n-1)}{n^2}} \tag{3.8}$$

  which is $\sqrt{n}$ times larger than the leave-one-out vector.

- The function $S(\mathbf{P})$ has approximate directional derivative

$$D_i = \frac{S(\mathbf{P}_{(i)}) - S(\mathbf{P}_0)}{\|\mathbf{P}_{(i)} - \mathbf{P}_0\|} \tag{3.9}$$

  Jackknife estimate is proportional to the root mean square of the directional derivatives.

3.1 *The Bayesian Bootstrap*

Let $G_1, G_2, \ldots, G_n$ be independent exponential variable with density $\exp(-x)$, the Bayesian bootstrap uses resampling vectors

$$\mathbf{P}^\star = \left(\frac{G_1}{\sum_i G_i}, \frac{G_2}{\sum_i G_i}, \ldots, \frac{G_n}{\sum_i G_i}\right) \tag{3.10}$$

TL: The number of distinct bootstrap weight vector $\mathbf{P}^\star$ is $\binom{2n-1}{n}$, the number of unique ways to put indistinguishable n balls to distinguishable n bars.

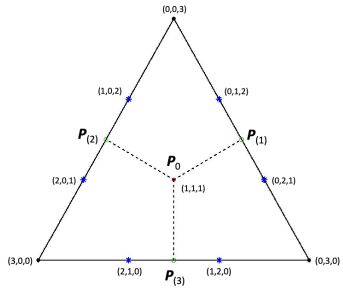

Figure 10.3 Resampling simplex for sample size $n = 3$. The center point is $\boldsymbol{P}_0$ (10.26); the green circles are the jackknife points $\boldsymbol{P}_{(i)}$ (10.28); triples indicate bootstrap resampling numbers (N_1, N_2, N_3) (10.29). The *bootstrap probabilities* are 6/27 for $\boldsymbol{P}_0$, 1/27 for each corner point, and 3/27 for each of the six starred points.

Figure 1: See Efron and Hastie

Note that the above $\mathbf{P}^\star$ follows a Dirichlet distribution, which has the mean and covariance matrix

$$\mathbf{P}^\star \sim [\mathbf{P}_0, \frac{1}{n+1}(\text{diag}(\mathbf{P}_0) - \mathbf{P}_0\mathbf{P}_0^\top)] \tag{3.11}$$

this is almost identical to the mean and covariance of bootstrap resamples $\mathbf{P}^\star \sim \text{Multi}_n(n, \mathbf{p}_0)/n$

$$\mathbf{P}^\star \sim [\mathbf{P}_0, \frac{1}{n}(\text{diag}(\mathbf{P}_0) - \mathbf{P}_0\mathbf{P}_0^\top)] \tag{3.12}$$

4 *Parametric Bootstrap, Infinitesimal Jackknife, and More*

4.1 *Parametric Bootstrap*

$$\hat{F} \xrightarrow{i.i.d.} \mathbf{x}^\star \xrightarrow{s} \hat{\theta}^\star \tag{4.1}$$

Now rather than using the empirical distribution $\hat{F} = \frac{1}{n}\sum_i \delta_{x_i}$, suppose we consider F is from some parametric family

$$\mathcal{F} = \{F_\omega(\mathbf{x}), \omega \in \Omega\} \tag{4.2}$$

Let $\hat{\omega}$ be some estimate of the parameter ω. The parametric bootstrap resamples directly from

$$F_{\hat{\omega}} \xrightarrow{i.i.d.} \mathbf{x}^\star \xrightarrow{s} \hat{\theta}^\star \tag{4.3}$$

The parametric families act as regularizers, namely one may have

$$d(F_{\hat{\omega}}, F) \ll d(\hat{F}, F) \tag{4.4}$$

under some suitable metric $d : \mathcal{P}_\mathcal{X} \times \mathcal{P}_\mathcal{X} \to \mathbb{R}$. Conceptually, the parametric bootstrap smoothes out the raw data and de-emphasizes outliers.

Many simulation-based inference methods can be considered as parametric bootstrap, including the generative models now standard in the machine learning literature.

4.2 *Infinitesimal Jackknife and Influence Functions*

There is an intimate connection between the infinitesimal jackknife and the bootstrap.

Define the linear interpolation between $\mathbf{P}_0$ and $\mathbf{P}_{(i)}$

$$\mathbf{P}_i(\epsilon) = (1-\epsilon)\mathbf{P}_0 + \epsilon\mathbf{P}_{(i)} \tag{4.5}$$

Then

$$\tilde{D}_i = \lim_{\epsilon\to 0} \frac{S(\mathbf{P}_i(\epsilon)) - S(\mathbf{P}_0)}{\epsilon\|\mathbf{P}_{(i)} - \mathbf{P}_0\|} \tag{4.6}$$

The infinitesimal jackknife estimate of the standard error is

$$\widehat{\text{se}}_{\text{IJ}} = \left(\frac{1}{n^2}\sum_{i=1}^{n}\tilde{D}_i^2\right)^{1/2} \tag{4.7}$$

Consider directly the mapping between the probability distribution to the statistic, namely the test statistic functional $T : \mathcal{P}_{\mathcal{X}} \to \mathbb{R}$

$$\theta = T[F] \tag{4.8}$$

$$\hat{\theta} = T[\hat{F}] \tag{4.9}$$

Define the influence function

$$\text{IF}(x) := \lim_{\epsilon \to 0} \frac{T[(1-\epsilon)F + \epsilon\delta_x] - T[F]}{\epsilon} \tag{4.10}$$

A fundamental theorem due to Tukey/Huber claims that

$$\widehat{\theta} \approx \theta + \frac{1}{n}\sum_{i=1}^{n}\text{IF}(x_i), \text{ as } n \to \infty \tag{4.11}$$

which implies

$$\text{var}(\hat{\theta}) \approx \frac{1}{n}\operatorname*{var}_{x\sim F}(\text{IF}(x)) \tag{4.12}$$

Example. *Consider the mean*

$$\theta = T[F] := \int x \, \mathrm{d}F(x) \tag{4.13}$$

$$\hat{\theta} = T[\hat{F}] := \int x \, \mathrm{d}\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.14}$$

then the influence function is

$$\text{IF}(x) = x - \theta \tag{4.15}$$

Example (A failing example)**.** *Consider the median*

$$\theta = T[F] = F^{-1}(1/2) \tag{4.16}$$

where $F(t) := P(x \leq t)$ is the CDF.

Then one can verify that for $x < \theta = F^{-1}(1/2)$

$$T[(1-\epsilon)F + \epsilon\delta_x] = F^{-1}(\tfrac{1/2-\epsilon}{1-\epsilon}) = F^{-1}(1/2) - \frac{1}{2f'(\theta)}\cdot\epsilon + o(\epsilon) \tag{4.17}$$

and thus

$$\text{IF}(x) = -\frac{1}{2f'(\theta)}, \textit{ for } x < \theta \tag{4.18}$$

similarly, one can show

$$\text{IF}(x) = \frac{1}{2f'(\theta)}, \textit{ for } x > \theta \tag{4.19}$$

Thus the influence function has discontinuity at θ. The Infinitesimal Jackknife fail to produce a valid standard error.

5 *Parametric vs. Nonparametric Bootstrap: GFR data exercise*

Table 10.2 in Efron and Hastie [3].

[3] Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, 2016

TL: HW: Reproduce the Table 10.2 in Efron and Hastie's book. Additionally, add the jackknife estimate and compare.

References

Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, 2016.