Subsampling

Mikynová, Pillárová

4. května 2026

Overview

- Bootstrap - reminder
- Subsampling - idea, construction
- Consistency - basic theorem
- Stochastic approximation
- Example
- Comparison with bootstrap
- Consistency - Studentized roots
- Consistency - General parameter space
- Hypothesis testing
- Data dependent choice of block size

Bootstrap-notation

Let $X_1, \ldots, X_n$ be an i.i.d. sample from an unknown distribution F. We consider a statistic

$$R_n = g(X_1, \ldots, X_n, F)$$

with distribution function

$$J_n(x, F) = \mathbb{P}_F(R_n \leq x).$$

Bootstrap

Idea

Approximate the unknown distribution $J_n(\cdot, F)$ by repeatedly recomputing the statistic on samples generated from the empirical distribution $\hat{F}_n$.

Construction

Step 1: Plug-in principle

$$F \approx \hat{F}_n, \qquad (X_1^*, \ldots, X_n^*) \sim \hat{F}_n, \qquad R_n^* = R_n(X_1^*, \ldots, X_n^*, \hat{F}_n).$$

Step 2: Monte Carlo approximation

$$R_{n,j}^* = R_n(X_{1,j}^*, \ldots, X_{n,j}^*, \hat{F}_n), \quad j = 1, \ldots, B.$$

$$\hat{J}_{n,B}^*(x) = \frac{1}{B} \sum_{j=1}^{B} \mathbf{1}\{R_{n,j}^* \leq x\}.$$

Subsampling

Idea

Approximate the sampling distribution $J_n(\cdot, F)$ of R_n by recomputing the same statistic on smaller subsets of the observed data.

Construction

Choose a subsample size $b < n$. Let $Y_1, \ldots, Y_{N_n}$ be all subsets of size b from $\{X_1, \ldots, X_n\}$, where $N_n = \binom{n}{b}$.
For each subset Y_i, compute

$$\hat{\theta}_{n,b,i} = \hat{\theta}_b(Y_i).$$

The subsampling distribution is then defined by

$$L_{n,b}(x) = \frac{1}{N_n} \sum_{i=1}^{N_n} \mathbf{1}\{\tau_b(\hat{\theta}_{n,b,i} - \hat{\theta}_n) \leq x\}.$$

Subsampling – notation

For the estimator $\hat{\theta}_n$, consider the statistic

$$R_n = \tau_n(\hat{\theta}_n - \theta(F)).$$

Hence

$$J_n(x, F) = \mathbb{P}_F \left\{ \tau_n\{\hat{\theta}_n - \theta(F)\} \leq x \right\}.$$

Assumption 1

There exists a limiting law $J(F)$ such that $J_n(F)$ converges weakly to $J(F)$ as $n \to \infty$.
The limiting law $J(F)$ is assumed to be nondegenerate.

Consistency

Theorem 1

Assume Assumption 1 holds. Also assume

$$\frac{\tau_b}{\tau_n} \to 0, \qquad b \to \infty, \qquad \frac{b}{n} \to 0, \quad n \to \infty.$$

1. If x is a continuity point of $J(\cdot, F)$, then

$$L_{n,b}(x) \xrightarrow{P} J(x, F).$$

2. If $J(\cdot, F)$ is continuous, then

$$\sup_x |L_{n,b}(x) - J_n(x, F)| \xrightarrow{P} 0.$$

Stochastic approximation

Computing $L_{n,b}(x)$ exactly means using all $N_n = \binom{n}{b}$ subsamples.

In practice, N_n is often too large.

Instead, choose only B subsamples at random:

$$I_1, \ldots, I_B \in \{1, \ldots, N_n\}.$$

Then use

$$\hat{L}_{n,b}(x) = \frac{1}{B} \sum_{i=1}^{B} \mathbf{1} \left\{ \tau_b(\hat{\theta}_{n,b,I_i} - \hat{\theta}_n) \leq x \right\}.$$

For $B \to \infty$, this Monte Carlo version has the same asymptotic validity as $L_{n,b}(x)$

Bootstrap VS Subsampling

Example: Maximum from $U(0, \theta)$

Let

$$X_1, \ldots, X_n \sim U(0, \theta).$$

We estimate the right endpoint θ by

$$\hat{\theta}_n = \max(X_1, \ldots, X_n).$$

The statistic of interest is the scaled estimation error

$$n(\hat{\theta}_n - \theta).$$

For $\theta = 1$,

$$n(\hat{\theta}_n - \theta) \xrightarrow{d} -Z, \qquad Z \sim \text{Exp}(1).$$

Grafic comparison

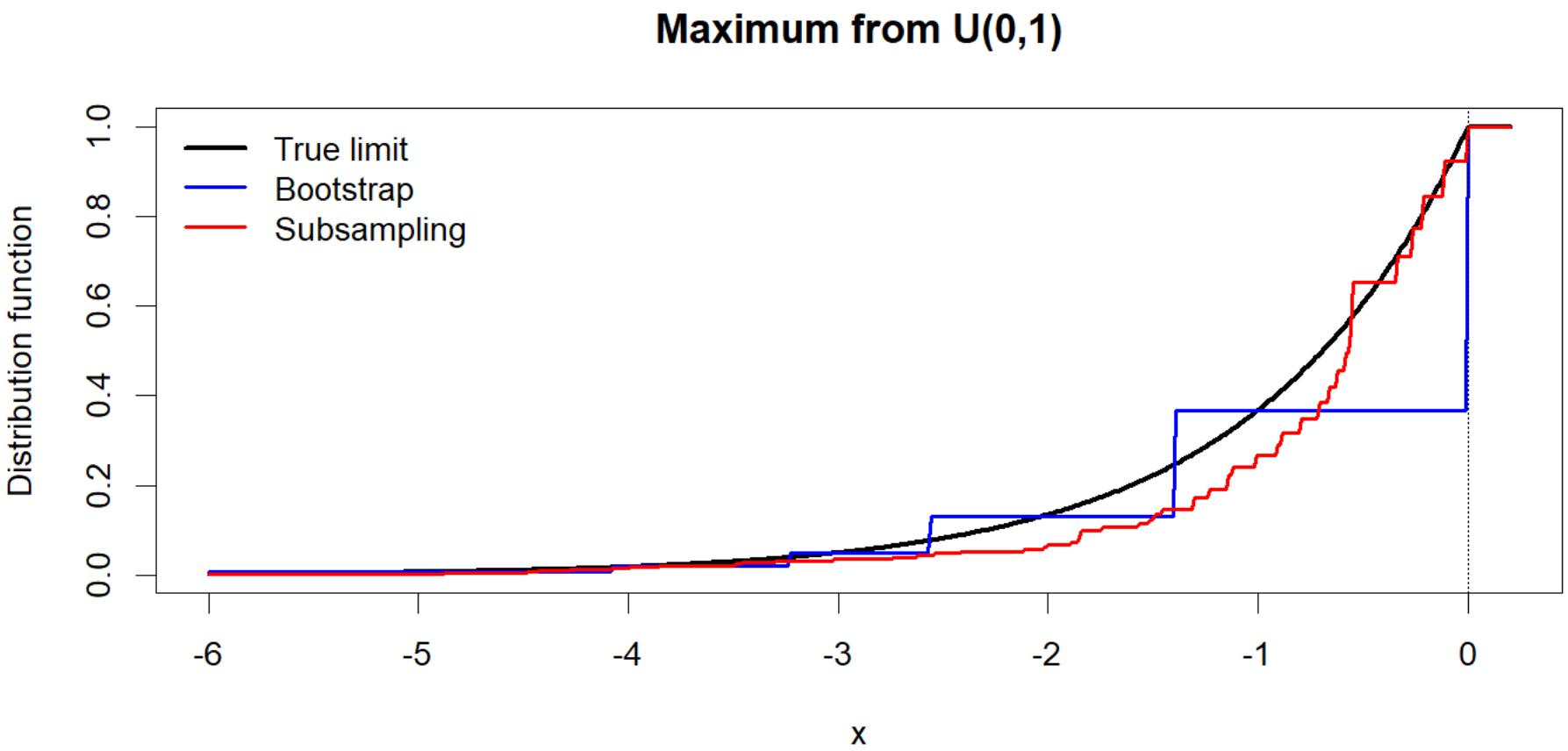

Figure 1: Bootstrap and subsampling approximations for the distribution of the normalized sample maximum.

Example: numerical illustration

In the simulation:

$$n = 500, \qquad b = \lfloor n^{0.6} \rfloor = 41.$$

Probability of reproducing the full-sample maximum

Bootstrap:

$$\mathbb{P}(\hat{\theta}_n^* = \hat{\theta}_n) = 1 - \left(1 - \frac{1}{500}\right)^{500} \approx 0.632.$$

Subsampling:

$$\mathbb{P}(\hat{\theta}_{n,b} = \hat{\theta}_n) = \frac{41}{500} \approx 0.082.$$

The bootstrap distribution therefore assigns substantial probability to the endpoint, whereas this probability is much smaller under subsampling.

Bootstrap vs. Subsampling

Feature	**Bootstrap**	**Subsampling**
Main idea	Resample from the empirical distribution	Use smaller subsamples of the original data
Data selection	With replacement	Without replacement
Resample size	Usually n	$b < n$, typically $b/n \to 0$
Assumptions	Needs stronger regularity conditions	Works under weaker assumptions
Main advantage	Often more accurate when valid	More robust in irregular problems
Main weakness	Can fail for non-smooth statistics	Requires choosing the block size b

Studentized roots

We consider the statistic

$$R_n^* = \tau_n \frac{\hat{\theta}_n - \theta(F)}{\hat{\sigma}_n},$$

where $\hat{\sigma}_n$ is some estimate of scale. Let $\hat{\sigma}_{n,b,i}$ be equal to the estimate of scale based on the i-th subsample of size b from the original data. Now define

$$L_{n,b}^*(x) = \frac{1}{N_n} \sum_{i=1}^{N_n} \mathbf{1}\{\tau_b \frac{\hat{\theta}_{n,b,i} - \hat{\theta}_n}{\hat{\sigma}_{n,b,i}} \leq x\}.$$

Consistency for studentized roots

Theorem 2

Assume Assumption 1 holds. Also assume

$$\frac{\tau_b}{\tau_n} \to 0, \quad b \to \infty, \quad \frac{b}{n} \to 0, \quad n \to \infty, \quad \hat{\sigma}_n \xrightarrow{P} \sigma,$$

where $\sigma = \sigma(F)$ is a positive constant.

1. If $x \cdot \sigma(F)$ is a continuity point of $J(\cdot, F)$, then

$$L^*_{n,b}(x) \xrightarrow{P} J(x \cdot \sigma(F), F).$$

2. If $J(\cdot, F)$ is continuous, then

$$\sup_x |L^*_{n,b}(x) - J_n(x \cdot \sigma(F), F)| \xrightarrow{P} 0.$$

General parameter space

Assume $\theta(F)$ takes values in parameter space Θ with norm denoted by $\|\cdot\|$. Let $\hat{\theta}_n$ be the estimate of $\theta(F)$.
Assume Assumption 1 with the interpretation that $\tau_n(\hat{\theta}_n - \theta(F))$ has a distribution in θ. Let $J_{n,\|\cdot\|}(F)$ denote the distribution of $\tau_n\|\hat{\theta}_n - \theta(F)\|$ under F, with corresponding distribution function $J_{n,\|\cdot\|}(\cdot, F)$.
Assumption 1 implies that $J_{n,\|\cdot\|}(F)$ converges weakly to $J_{\|\cdot\|}(F)$, the distribution of $\|Z\|$, where Z has distribution $J(F)$. $J_{\|\cdot\|}(\cdot, F)$ is the distribution function of $J_{\|\cdot\|}(F)$. Now we define

$$L_{n,b,\|\cdot\|}(x) = \frac{1}{N_n}\sum_{i=1}^{N_n} \mathbf{1}\{\tau_b\|\hat{\theta}_{n,b,i} - \hat{\theta}_n\| \leq x\}.$$

Consistency for general parameter space

Theorem 3

Assume Assumption 1 holds. Also assume

$$\frac{\tau_b}{\tau_n} \to 0, \quad b \to \infty, \quad \frac{b}{n} \to 0, \quad n \to \infty.$$

1. If x is a continuity point of $J_{\|\cdot\|}(\cdot, F)$, then

$$L_{n,b,\|\cdot\|}(x) \xrightarrow{P} J_{\|\cdot\|}(x, F).$$

2. If $J_{\|\cdot\|}(\cdot, F)$ is continuous, then

$$\sup_x |L_{n,b,\|\cdot\|}(x) - J_{\|\cdot\|}(x, F)| \xrightarrow{P} 0.$$

Hypothesis testing

We are testing the hypothesis

$$H_0 : F \in \mathbf{F}_0 \qquad H_1 : F \in \mathbf{F}_1,$$

where $\mathbf{F}$ is a class of distributions, $\mathbf{F}_i \subset \mathbf{F}$ for $i = \{0, 1\}$, $\mathbf{F}_0 \cup \mathbf{F}_1 = \mathbf{F}$.

If the null hypothesis translates into a null hypothesis about a real or vector-valued parameter $\theta(F)$, we can construct a confidence region for $\theta(F)$ and the use the duality between the construction of confidence regions for parameters and the construction of hypothesis tests about those parameters.

Hypothesis testing - confidence interval

Theorem 4

Assume the assumptions of Theorem 1. Let

$$c_{n,b}(1-\alpha) = \inf\{x : L_{n,b}(x) \geq 1-\alpha\}.$$

Correspondingly, define

$$c(1-\alpha, F) = \inf\{x : J(x, F) \geq 1-\alpha\}.$$

If $J(\cdot, F)$ is continuous at $c(1-\alpha, F)$, then

$$P_F(\tau_n(\hat{\theta}_n - \theta(F)) \leq c_{n,b}(1-\alpha)) \to 1-\alpha, \quad \text{as } n \to \infty.$$

Therefore, the asymptotic coverage probability under F of the confidence interval $[\hat{\theta}_n - \tau_n^{-1} c_{n,b}(1-\alpha), \infty)$ is the nominal level $1-\alpha$.

Hypothesis testing

The test statistic is

$$T_n = \tau_n t_n(X_1, ..., X_n).$$

Let

$$G_n(x, F) = P_F(\tau_n t_n(X_1, ..., X_n) \leq x).$$

Choose a subsample size $b < n$. Let $Y_1, \ldots, Y_{N_n}$ be all subsets of size b from $\{X_1, \ldots, X_n\}$, where $N_n = \binom{n}{b}$. Let $t_{n,b,i}$ be equal to the statistic t_b evaluated at the data set Y_i. The sampling distribution of T_n is the approximated by

$$\hat{G}_{n,b}(x) = \frac{1}{N_n} \sum_{i=1}^{N_n} \mathbf{1}\{\tau_b t_{n,b,i} \leq x\}.$$

Hypothesis testing

Define

$$g_{n,b}(1-\alpha) = \inf\{x : \hat{G}_{n,b}(x) \geq 1-\alpha\}.$$

The nominal level α test rejects H_0 if and only if

$$T_n > g_{n,b}(1-\alpha).$$

Hypothesis testing

Theorem 5

- Assume, for $F \in \mathbf{F}_0$, $G_n(F)$ converges weakly to a continuous limit law $G(F)$, whose corresponding distribution function is $G(\cdot, F)$ and whose $1-\alpha$ quantile is $g(1-\alpha, F)$. Assume

$$\frac{b}{n} \to 0, \quad b \to \infty \quad \text{as} \quad n \to \infty.$$

If $G(\cdot, F)$ is continuous at $g(1-\alpha)$ and $F \in \mathbf{F}_0$, then

$$g_{n,b}(1-\alpha) \xrightarrow{P} g(1-\alpha, F)$$

and

$$P_F(T_n > g_{n,b}(1-\alpha)) \to \alpha \text{ as } n \to \infty$$

Hypothesis testing

Theorem 5

- Assume the test statistic is constructed so that $t_n(X_1, ..., X_n) \to t(F)$ in probability, where $t(F)$ is a constant which satisfies $t(F) = 0$ if $F \in \mathbf{F}_0$ and $t(F) > 0$ if $F \in \mathbf{F}_1$. Assume

$$\frac{b}{n} \to 0, \quad b \to \infty, \quad \liminf_n \frac{\tau_n}{\tau_b} > 1.$$

Then, if $F \in \mathbf{F}_1$, the rejection probability satisfies

$$P_F(T_n > g_{n,b}(1-\alpha)) \to 1 \quad \text{as} \quad n \to \infty.$$

Data dependent choice of block size

Theorem 6

Assume Assumption 1. Let $1 \leq j_n \leq k_n \leq n$ be integers satisfying $j_n \to \infty, k_n/n \to 0, \tau_{k_n}/\tau_n \to 0$, and, for every $d > 0, \quad k_n \exp(-d\lfloor \frac{n}{k_n} \rfloor) \to 0 \quad$ as $\quad n \to \infty$. Also, assume $\{\tau_n\}$ is non-decreasing in n.

- If x is a continuity point of $J(\cdot, F)$, then

$$\sup_{j_n \leq b \leq k_n} |L_{n,b}(x) - J(x, F)| \xrightarrow{P} 0.$$

- Hence, if $\{\hat{b}_n\}$ is a data-dependent sequence (that is, a measurable function of $X_1, ..., X_n$), and

$$P_F(j_n \leq \hat{b}_n \leq k_n) \to 1,$$

then

$$L_{n,\hat{b}_n}(x) \xrightarrow{P} J(x, F).$$

Data dependent choice of block size

Theorem 6

- If $J(\cdot, F)$ is continuous, then

$$\sup_x |L_{n,\hat{b}_n}(x) - J(x,F)| \xrightarrow{P} 0.$$

In fact,

$$\sup_{j_n \le b \le k_n} \sup_x |L_{n,b}(x) - J(x,F)| \xrightarrow{P} 0.$$

- Let

$$c_{n,\hat{b}_n}(1-\alpha) = \inf\{x : L_{n,\hat{b}_n}(x) \ge 1-\alpha\}.$$

Then, if $J(\cdot, F)$ is continuous,

$$P_F(\tau_n(\hat{\theta}_n - \theta(F)) \le c_{n,\hat{b}_n}(1-\alpha)) \to 1-\alpha$$

as $n \to \infty$. Therefore, the asymptotic coverage probability under F of the confidence interval $[\hat{\theta}_n - \tau_n^{-1} c_{n,\hat{b}_n}(1-\alpha), \infty)$ is the nominal level

Thank you for your attention!