Bootstrap and Resampling Methods

Lecture 1: Introduction to the Course

Elia Lapenta (CREST & ENSAE)

January 31st

The goals of this course

The goals of this course

- Learn the **principles and ideas** at the basis of bootstrap, jackknife, and resampling
- Learn how to **implement** such methods in practice
- Learn the **mathematical/statistical background** at the basis of bootstrap, jackknife, and resampling methods

To this end, we need to

- know the basics of **probability theory** and **statistics I**
- learn and use the statistical language *R*.
- Ask questions!

Overview of today's class

Lecture 1: Introduction to **Jackknife**, **Bootstrap**, **Subsampling** and **other resampling methods** $\Rightarrow$ These are used to make inference on a parameter that we do not know but we are interested in

The goal of today's class is to broadly present

- What are the motivations for studying these methods? Why are they useful? How we will apply such methods in practice?
- What are the models we will apply these methods to
- What is the approach we will take in the course

Why Bootstrap, Jackknife, and Subsampling methods are useful?

Why Bootstrap, Jackknife, and Subsampling are useful?
A Road map

- The problem of constructing **Confidence Intervals** and **Hypothesis Tests**
- Basic approach: **exact** Confidence Intervals
- Confidence Intervals and Hypothesis Tests by **Asymptotic Approximations**
- The advantages of the **Bootstrap**: Automatic estimation of the distribution of a statistic
- The advantages of the **Jackknife**: Automatic Bias and Variance Estimation
- The motivations for **subsampling**: More robust than the bootstrap

less accurate

A General Problem of Inference

- We observe a sample $\{Z_i\}_{i=1}^n$ of iid data from a probability P.
- We are interested in a population parameter θ_0 that is unknown to us.
- We have an estimator of θ_0, say $\widehat{\theta}$
- Using $\widehat{\theta}$ we want to make inference on θ_0, i.e. we want to "discover features of θ_0"

Example: We want to know $\mathbb{E}Z$ and we estimate it by the sample average $\overline{Z} := \frac{1}{n}\sum_{i=1}^n Z_i$

A general problem of inference: consistency is not enough

- In general, $\widehat{\theta}$ will be a good approximation of θ_0, in the sense that $\widehat{\theta} \xrightarrow{P} \theta_0$
- But sometimes consistency is not enough! Assume we want to know if $\theta_0 \geq 0$
  - We might get $\widehat{\theta} < 0$ just because of sampling errors (we are unlucky!) and this might lead us to the wrong conclusion that $\theta_0 < 0$
- Another example: Assume we want to know if $\theta_0 \neq 0$.
  - If we get $\widehat{\theta} = 0.01$ is this enough to say that $\theta_0 = 0$?

$\Rightarrow$ Because of sampling errors, consistency is not enough to make inference on a parameter θ_0!

In other words, consistency does not tell us if the differences between $\widehat{\theta}$ and θ_0 are **significant** or not !

A general problem of inference: construct CIs

Since consistency of $\widehat{\theta}$ is not enough to make inference about θ_0, we must rely on other tools.

- To make inference on θ_0, it would be better to construct some **confidence intervals (CIs)**
- To obtain the CIs we use **the distribution** of $\widehat{\theta}$ that usually depends on θ_0

Example. Assume we observe $\{Z_i\}_{i=1}^n$ that are iid, with $Z_i \sim \mathcal{N}(\theta_0, V)$ and V known. We want to know if $\theta_0 = 0$.

$$\text{For} \quad \widehat{\theta} := \overline{Z} \quad \text{we have} \quad \widehat{\theta} \sim \mathcal{N}(\theta_0, V/n) \quad \text{so} \quad \frac{\sqrt{n}(\widehat{\theta} - \theta_0)}{\sqrt{V}} \sim \mathcal{N}(0, 1)$$

A general problem of inference: exact CIs

From the exact distribution of $\widehat{\theta}$, we can construct CIs about θ_0: If q_β is the βth quantile of $\mathcal{N}(0,1)$, then

$$P\left(q_{\alpha/2} < \frac{\widehat{\theta} - \theta_0}{\sqrt{V/n}} \leq q_{1-\alpha/2}\right) = 1 - \alpha\,.$$

So the **$1 - \alpha$ Confidence Interval for θ_0** will be

$$\widehat{CI}(1-\alpha) = \left[\widehat{\theta} - q_{1-\alpha/2}\sqrt{\frac{V}{n}}\,,\,\widehat{\theta} - q_{\alpha/2}\sqrt{\frac{V}{n}}\right]$$

In the sense that

$$P\left(\theta_0 \in \widehat{CI}(1-\alpha)\right) = 1 - \alpha$$

This is an **exact** Confidence Interval!

A general problem of inference: CIs by Asymptotic Approximation

However, in the most of the cases, we cannot obtain an exact Confidence Intervals.
In the previous example

- $Z_i \sim \mathcal{N}$ (exact normality) might not hold
- V might be unknown.

When we the distribution of the observations is unknown, we obtain Confidence Intervals for θ_0 by using an **Asymptotic Approximation** based on the CLT:

Given $\quad \widehat{\theta} = \bar{Z}\,,\ \theta_0 = \mathbb{E}Z\,,\ V = Var(Z)\,,$ we have $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$ for $n \to \infty$

so that if $\widehat{V} \xrightarrow{P} V$ then

$$S_n := \frac{\sqrt{n}(\widehat{\theta} - \theta_0)}{\sqrt{\widehat{V}}} \xrightarrow{d} \mathcal{N}(0,1)\,, \text{ as } n \to \infty$$

A general problem of inference: CIs by Asymptotic Approximation, ct'ed

Idea: The **asymptotic distribution** provides an **approximation** to the **true** distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)/\sqrt{\widehat{V}} \implies$ Use it to construct CIs.

Denote with Φ the cdf of $\mathcal{N}(0, 1)$. Then, as $n \to \infty$

cvg in distribution.

$$S_n = \frac{\sqrt{n}(\widehat{\theta} - \theta_0)}{\sqrt{\widehat{V}}} \xrightarrow{d} \mathcal{N}(0, 1) \iff P\left(S_n \leq s\right) \to \Phi(s) \text{ for all } s \in \mathbb{R}$$

Hence, recalling that q_β is the βth quantile of $\mathcal{N}(0, 1)$, we get

$$P\left(q_{\alpha/2} < \frac{\sqrt{n}(\widehat{\theta} - \theta_0)}{\sqrt{\widehat{V}}} \leq q_{1-\alpha/2}\right) \to 1 - \alpha$$

asymptotic

A general problem of inference: CIs by Asymptotic Approximation, ct'ed

So, the $1-\alpha$ **asymptotic** CI for θ_0 is

$$\widehat{CI}^A(1-\alpha) := \left[\widehat{\theta} - q_{1-\alpha/2}\sqrt{\frac{\widehat{V}}{n}}, \widehat{\theta} - q_{\alpha/2}\sqrt{\frac{\widehat{V}}{n}}\right]$$

in the sense that

$$P\left(\theta_0 \in \widehat{CI}^A(1-\alpha)\right) \to 1-\alpha$$

- So far we have focused on CIs
- When we want to test $H_0 : \theta_0 = \overline{\theta}$, the decision rule for a test at the α nominal level is

$$\text{Reject } H_0 \text{ if } \overline{\theta} \notin \widehat{CI}^A(1-\alpha)$$

Issues with the Asymptotic Approximation

$$P\left(\theta_0 \in \widehat{CI}^A(1-\alpha)\right) \xrightarrow{n \to +\infty} 1-\alpha$$

However,

- We need an estimator $\widehat{V}$ of V: in some cases, constructing such estimators is difficult
- For finite samples this is an approximation, so there will be an **approximation error**

$\implies$ In finite samples it might happen that

$$P\left(\theta_0 \in \widehat{CI}^A(1-\alpha)\right) \text{ is far from } (1-\alpha)$$

- So, $\widehat{CI}^A(1-\alpha)$ might not contain θ_0 with a sufficiently high probability: *We do not have a correct coverage probability*!
- If we test $H_0 : \theta_0 = \overline{\theta}$ and the asymptotic approximation is not good, then the probability of rejecting H_0 is higher than α: *We do not have a correct size control*!

The Bootstrap

Why is the bootstrap is useful in inference?

Summarizing the approach based on the *asymptotic approximation*:

- $\widehat{CI}^A$ is called *asymptotic CI* because it has the correct size $(1-\alpha)$ only asymptotically , i.e. when $n \to \infty$.
- The central idea is to use the **asymptotic** distribution to **approximate** the **true** distribution of the statistic: Since

$$P\left(S_n \leq \cdot\right) \approx \Phi(\cdot)$$

  we replace the quantiles of the exact distribution of S_n with the quantiles of $\mathcal{N}(0, 1)$.

However, this is still an approximation! In finite samples there is an **approximation error**

Central Question: Is there a better way to approximate the **true** distribution of the statistic S_n ?

The Bootstrap Principle

Recall that S_n is a function of the data.

- Central Idea The bootstrap approximates the **true** distribution of S_n by ***resampling*** the data
- *Resampling* means to generate artificial data (i.e. **artificial samples**) using the sample data
- Once an **artificial sample** has been created, the statistic S_n is computed using such an artificial sample
- We can repeat this process a very large number of times
- This provides an **artificial distribution** of S_n that is used to approximate the true distribution of S_n

We have **3 main advantages** of the Bootstrap! Before we provide an example

Example: Pairwise Bootstrap

Given the sample $\{Z_i\}_{i=1}^n$ we have the statistic

$$S_n := \sqrt{n}(\overline{Z} - \mathbb{E}Z)$$

The **pairwise bootstrap** works as follows:

- Draw with replacement n observations from $\{Z_i\}_{i=1}^n$, say $\{Z_i^*\}_{i=1}^n$
- Compute the "bootstrap counterpart" of S_n as $S_n^* := \sqrt{n}(\overline{Z^*} - \overline{Z})$, where $\overline{Z^*}$ is the average of $\{Z_i^*\}_{i=1}^n$

  original sample mean
- Repeat the two steps above many times, say B times

We obtain in such a way a distribution of statistics that is used to approximate the true distribution of S_n

get empirical distribution of S_n^* $\longrightarrow$ $\hat{q}_{1-\frac{\alpha}{2}}$, $\hat{q}_{\frac{\alpha}{2}}$

1st advantage of the Bootstrap

e.g. $\sqrt{n}(\bar{z} - \mathbb{E}(z)) \xrightarrow{n \to +\infty} \mathcal{N}(0, Var(z))$

asymptotic distribution

depends on unknown param : $Var(z)$

1st advantage of the bootstrap: asymptotic refinements

When the statistic is **asymptotically pitoval**, the *bootstrap distribution* approximates the *true distribution* of S_n better than the *asymptotic distribution*

- A statistic is *asymptotically pivotal* if its asymptotic distribution does not depend on unknown parameters.
- In our previous example $S_n \xrightarrow{d} \mathcal{N}(0,1)$ and the $\mathcal{N}(0,1)$ does not depend on unknown parameters.
- The bootstrap distribution can be used to construct *bootstrap CIs* for θ_0: In general they are more precise than the *asymptotic CIs*.

When a statistic is not asymptotically pivotal, we can use the **double bootstrap** to obtain refinements

2nd advantage of the Bootstrap

2nd advantage of the bootstrap: no asymptotic approximation

In some cases, the asymptotic distribution of S_n is very intricate $\Rightarrow$ the bootstrap provides an **automatic** way to approximate such a distribution

Example. We will see that in some specification tests $S_n \xrightarrow{d} \mathbb{U}$, where $\mathbb{U}$ is a transformation of a Gaussian process with a continuum of covariances!

- The asymptotic distribution is well defined, but cannot be used to approximate the true distribution of S_n

3rd advantage of the Bootstrap

3rd advantage of the bootstrap: difficult expression of the asymptotic variance

In some cases, the asymptotic distribution depends on a parameter with a complicated expression $\Rightarrow$ the bootstrap provides an **automatic** and **easy** way to approximate the true distribution of the statistic.

Example. We have previously seen that $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$

- If V is difficult to estimate, the asymptotic approximation will be difficult to implement.

Let's be careful though! Bootstrapping doesn't always work

$\Theta = [0,1]$.

There are cases where the bootstrap does not work:

e.g. $\theta_0 = \arg\min_{\theta \in \Theta} \mathbb{E}[m_\theta(z)]$ If $\theta_0 = 0$, Bootstrap does not work.

- When the true parameter θ_0 lies on the boundary of the parameter space
- When making inference on the upperbound of the support of a random variable (lowerbound) e.g. $X \sim \mathcal{U}([0, \theta])$
- With weak Instrumental Variables (if the statistic is not the correct one)
- When it is not able to replicate an asymptotic bias (matching estimator of Average Treatment Effects)

The Jackknife

The Jackknife

The Jackknife is an automatic method used to estimate the (asymptotic) bias and variance of a statistic (or estimator).
However, differently from the bootstrap, it does not estimate the distribution of a statistic

Idea: Each observation is excluded from the sample and the statistic (estimator) is recomputed on the remaining $n-1$ observations. Then, the resulting statistics are "aggregated" to estimate the bias or variance

Applications

- An estimator is biased and we want to reduce its bias
- The asymptotic variance of a statistic has a difficult expression and we want to avoid a difficult computations

Jackknife bias estimation

Let $\widehat{\theta} := G_n(Z_1, \ldots, Z_n)$ be a statistic (estimator) used to estimate θ_0.

- The bias of $\widehat{\theta}$ is defined as $\quad b := \mathbb{E}\widehat{\theta} - \theta_0$
- The jackknife estimator of the bias is

$$b_{jack} := (n-1)\left(\overline{\widehat{\theta}} - \widehat{\theta}\right)$$

where

drop one sample z_i

$$\overline{\widehat{\theta}} := \frac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_{n-1,i} \quad \text{and} \quad \widehat{\theta}_{n-1,i} := G_{n-1}(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$$

(each $\widehat{\theta}_{n-1,i}$ is the statistic based on all the observations except the ith)

- The jackknife bias corrected estimator is $\quad \widehat{\theta} - b_{jack}$

Jackknife variance estimation

As we have previously seen

$$\frac{\widehat{\theta} - \theta_0}{\sigma_n} \xrightarrow{d} \mathcal{N}(0, 1)$$

$\sigma_n^2 := \sigma^2/n$, and σ is the (unknown) asymptotic variance of $\widehat{\theta}$.

- If we have an estimator $\widehat{\sigma} \xrightarrow{P} \sigma$, then

$$\sqrt{n}\frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}} \xrightarrow{d} \mathcal{N}(0, 1)$$

- However, σ might have a very complicated expression $\Rightarrow \widehat{\sigma}$ might be difficult or not convenient to compute when it is based on the expression of σ (i.e. the sample analog of σ)
- The jackknife provides an automatic way to estimate σ

Jackknife variance estimation

- The Jackknife variance estimator is

$$v_{jack}^2 := \frac{(n-1)}{n} \sum_{i=1}^{n} \left[\widehat{\theta}_{n-1,i} - \overline{\widehat{\theta}} \right]^2 \quad \text{where } \overline{\widehat{\theta}} := \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_{n-1,i}$$

and $\widehat{\theta}_{n-1,i}$ is the estimator based on all the observations but the ith.

- We will show that $v_{jack}/\sigma_n \xrightarrow{P} 1$ so that

estimate asymptotique variance

$$\frac{\widehat{\theta} - \theta_0}{v_{jack}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Subsampling

Subsampling

Motivation: As we have seen, the bootstrap does not work in some cases. The **subsampling** is a method that is more robust than the bootstrap and works under weaker conditions.

Idea: Recall from the example on the pairwise bootstrap that the bootstrap is based on drawing n observations from the sample.

Instead of drawing n observations with replacement, the **subsampling** draws $m < n$ ***observations without replacement***. It then repeats this process many times.

Subsampling

Main features of the subsampling

- The subsampling is more robust than the bootstrap and it works under weaker conditions
  - So, it can replace the asymptotic approximation when the bootstrap can't
- However, the robustness comes at a price: when the bootstrap works, subsampling is less accurate than the bootstrap
- Moreover, when performing the subsampling, we need to choose m such that $m/n \to 0$: in general, we do not know how to choose m "optimally"

Applications

OLS

- We observe $\{Y_i, X_i\}_{i=1}^n$, Y_i scalar and X_i vector.

$$Y = X^T \beta_0 + \varepsilon \quad \text{with } \mathbb{E}X\varepsilon = 0$$

β_0 are the coefficients of the linear projection of Y onto X and are estimated by

$$\widehat{\beta} := \left(\overline{XX^T}\right)^{-1} \overline{XY}$$

- From our elementary courses we remember that

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma) \quad \text{where } \Sigma := \mathbb{E}(XX^T)^{-1}\, \mathbb{E}\varepsilon^2 XX^T\, \mathbb{E}(XX^T)^{-1}$$

- We will see that the Jackknife provides a good estimator of Σ, better than the sample analog of Σ (the White-Huber estimator)

2SLS (e.g., Local Average Treatment Effect)

- We observe $\{Y_i, X_i, W_i\}_{i=1}^n$, the model is

$$Y = X^T \beta_0 + \varepsilon \quad \text{with } \mathbb{E}\varepsilon W = 0$$

  $\mathbb{E}X\varepsilon \neq 0$, so X is an endogenous regressor

- The 2 Stages Least Squares estimator of β_0 is

$$\widehat{\beta}_{2SLS} := \left[(\overline{XW^T})(\overline{WW^T})^{-1}(\overline{WX^T})\right]^{-1}(\overline{XW^T})(\overline{WW^T})^{-1}(\overline{WY})$$

- Unlike the OLS estimator, $\widehat{\beta}_{2SLS}$ is biased ($\mathbb{E}\widehat{\beta}_{2SLS} \neq \beta_0$) in finite samples.
- We will see that the Jackknife provides a good estimator of the bias of $\widehat{\beta}_{2SLS}$

M Estimator

We observe a random sample $\{Z_i\}_{i=1}^n$. Given a mapping $(z,\theta) \mapsto m_\theta(z)$, the parameter of interest is

$$\theta_0 = \arg\max_{\theta\in\Theta} \mathbb{E} m_\theta(Z)$$

where Θ is the parameter space.

An **M-estimator** is

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} m_\theta(Z_i)$$

Examples: Maximum Likelihood, Non Linear Least Squares, etc.

- We will study in detail several bootstrap procedures to approximate the distribution of $\widehat{\theta}$.

NP density estimator

We observe a random sample $\{Z_i\}_{i=1}^n$ and f is the density of each Z_i.

Goal: estimate f without imposing parametric assumptions on it.

A popular estimator is the **Kernel estimator**:

kernel smoothing

$$\widehat{f}(z) := \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{Z_i - z}{h}\right)$$

where K is a user-chosen function that is symmetric about zero with $\int K(u)du = 1$.

- We will study several bootstrap schemes for approximating the distribution of $\widehat{f}$. We will then construct CIs for the true f.

Selection of tuning parameters: Smoothing Splines and RKHS

$$Y_i = f_0(X_i) + \varepsilon_i \text{ with } \mathbb{E}\{\varepsilon|X\} = 0$$

Goal: estimate f_0 **nonparametrically**, i.e. without imposing parametric assumptions on f_0 (e.g., it is linear in X). We just assume that f_0 is a *smooth* function.

$$\widehat{f}_\lambda := \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - f(X_i)]^2 + \lambda \int_0^1 \left| f^{(2)}(x) \right|^2 dx$$

where λ is a penalization parameter, $f^{(2)}$ is the second derivative of f, and

$$\mathcal{F} := \text{ Class of functions that are twice continuously differentiable}$$

Choice of λ is crucial: the larger λ the larger the penalty from variations of $f \Longrightarrow$ The **bootstrap** provides a very attractive method for choosing λ

Specification Testing

We observe a random sample $\{Y_i, Z_i\}_{i=1}^{n}$ and the model is assumed to be

$$Y = m_{\theta_0}(Z) + \varepsilon \quad \text{with} \quad \mathbb{E}\{\varepsilon|Z\} = 0$$

for some $\theta_0 \in \Theta$. We want to test is the above model is **correctly specified**, i.e. if $\mathbb{E}\{Y|Z\} = m_{\theta_0}(Z)$. We will see that a statistic we can use is

$$S_n := \int \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[Y_i - m_{\widehat{\theta}}(Z_i) \right] \varphi_t(Z_i) \right|^2 dt$$

where ϕ_t is a "well chosen" weighting function.

If the model is correctly specified, $S_n \xrightarrow{d} \mathbb{U}$, where $\mathbb{U}$ is a transformation of a Gaussian stochastic process.

- The bootstrap is extremely useful in such situation for computing the critical values!

Other applications

There are further applications very similar to those previously listed

- Panel data with fixed or random effects
- GMM estimators
- NP regression estimators
- Semiparametric models (Partly linear, Single-Index, additive separable)
- Models with clusters

Twilight zone: High dimensional models not works well

The approach we will adopt in the course

The approach we will use throughout the course

- Prove the validity of the Jackknife for variance and bias estimation
- Detailed proofs of the consistency of the bootstrap for constructing CIs and tests of hypothesis
- Intuition about the Edgeworth Expansions (that are used to prove the refinements of the bootstrap)
- We will run simulations in R to
  - learn how such methods work
  - realize the advantages of such methods
- We will apply the methods to real data using R

Lecture notes will be provided.