TA Session 10: Nonparametric Bootstrap

Shuowen Chen[1]

EC708: PhD Econometrics I (Spring 2020)

[1] Materials are based on Horowitz (2001, 2017), teaching slides from Hiroaki Kaido and Xiaoxia Shi, and textbook by Bruce Hansen (2020)

Outline

- Why/When does Bootstrap Work?
- Why Use Bootstrap?
  - Bootstrap Standard Errors and Confidence Intervals
  - Estimation: Bias Correction of Nonlinear Estimators
  - Hypothesis Testing: Higher–Order Refinements
- Why/When does Bootstrap NOT Work?
  - General Discussions
  - Two Examples

Caveats

We focus on **nonparametric** bootstrap, but there are many...

- boostrap methods for different types of data/questions
  - Parametric (OLS design matrix invertibility)
  - Wild/residual (Kelly, Pruitt & Su, 2018; Instrumented PCA)
  - Sequential (Non I.I.D. Sample); Block (Dependent Data)
  - Weighted (Chernozhukov, Fernández-Val & Melly, 2013)
  - Numerical (Hong & Li, 2020)
- other resampling methods
  - Subsampling (partial identification)
  - Jackknife (Fernández-Val and Weidner, 2016; panel/trade)
  - Permutation test (regression kink/discontinuity, synthetic control)
- exotic hybrid/ensemble resampling methods
  - Non−regular models (weak IV, post−conservative model selection estimation, nearly integrated processes)[2]
  - Bagging, Boosting (adaboost)
- bells and whistles when implementing bootstrap algorithms
  - Trimming parameter
  - Number of bootstraps: 50? :(500? :| 2000+? Set up a waitbar
  - Dealing with NA values (finite sample issues or coding errors)

[2] D.Andrews and Guggenberger have a series of papers on this topic.

What is Bootstrap?

"A method for estimating the distribution of an estimator or test statistic by resampling one's data or a model estimated from the data."

— Joel L. Horowitz (2017)

Why do we use Bootstrap?

1. An alternative to asymptotic–based inference
2. Bias correct a nonlinear estimator
3. Improve upon first–order approximation
    - Nominal Asymptotic critical value and real values differ a lot (size), but bootstrap reduces the distortion (information matrix test of White)

But Bootstrap is NOT:

1. a substitute of asymptotic theory
    - References: van der Vaart and Wellner (1996), Korosok (2008)
2. a panacea[3]
    - Parameters on the boundary of a parameter set
    - Max or min of random variables
    - Program evaluation methods (matching, IPW)

[3] I want to reiterate that this session focuses on nonparametric bootstrap. Many other bootstrap and resampling methods have been proposed to deal with various problems as selectively listed here.

Data Assumptions and Notations

- A random sample $\{X_1, \ldots, X_n\}$ from a CDF distribution F_0
- Statistic $T_n(X_1, \ldots, X_n)$ (estimator or test statistic)
- $\mathcal{J}_n(\tau, F) \equiv Pr(T_n \leq \tau | F)$: exact, finite sample distributuon of T_n when data sample from CDF F. When $\mathcal{J}_n(\cdot, F)$ doesn't depend on F, we call T_n **pivotal**.
- $\mathcal{J}_n(\tau, F_0)$ unknown as F_0 is unknown
- Approximates $\mathcal{J}_n(\tau, F_0)$ by replacing F_0 with a known estimator $\widehat{F}_n$. Nonparametric bootstrap uses empirical CDF:

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}$$

- Obtain i.i.d. draws from distribution $\mathcal{J}_n(\tau, \widehat{F}_n)$ by simulation

▸ Asymptotically Pivotal Statistics

Nonparametric Bootstrap Algorithm

1. Draw a new sample of size n from $\{X_i\}_{i=1}^n$ with replacement
2. Denote the sample as $\{X_i^*\}_{i=1}^n$, calculate $\widehat{T}_n^1$
3. Repeat steps 1 and 2 B times and obtain a sequence $\{\widehat{T}_n^b\}_{b=1}^B$

Why/When Does Bootstrap Work?

Bootstrap Consistency: Bootstrap distribution of a statistic is a consistent estimator of the statistic's asymptotic distribution

Two sources of randomness:

1. $\{X_i\}_{i=1}^n$ is a **random** sample of **unknown** distribution
2. $\{X_i^*\}_{i=1}^n$ is a **random** sample of $\{X_i\}_{i=1}^n$ of **known** distribution

Heurestics using normalized sample mean as a particular T_n:

- Denote bootstrap conditional expectation $\mathbb{E}^* := \mathbb{E}(\cdot|\widehat{F}_n)$
- Bootstrap sample mean $\overline{X}^* = \frac{1}{n}\sum_{i=1}^n X_i^*$; sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$; population mean $\mu = \mathbb{E}(X)$
- $\mathbb{E}^*(\overline{X}^*) = \overline{X}$, $Var^*(\overline{X}^*) = \frac{1}{n}\overline{\Sigma}$

Connecting the Two Worlds

In the real world

- WLLN: $\overline{X} \xrightarrow{p} \mu$ as $n \to \infty$
- CLT: $\sqrt{n}(\overline{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$
- CMT: $g(\overline{X}) \xrightarrow{p} g(\mu)$ for continuous function $g(\cdot)$
- Delta method: $\sqrt{n}(g(\overline{X} - g(\mu)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial g(u)'}{\partial u}|_{u=\mu} \Sigma \frac{\partial g(u)}{\partial u}|_{u=\mu}\right)$ for continuously differentiable $g(\cdot)$

Connecting the Two Worlds

In the bootstrap world:

- Boostrap WLLN: $\overline{X}^* \xrightarrow{p^*} \overline{X}$
- Bootstrap CLT: $\sqrt{n}(\overline{X}^* - \overline{X}) \xrightarrow{d^*} \mathcal{N}(0, \Sigma)$
- Notations p^* and d^* (from Hansen's textbook) denote convergence in probability/distribution **in probability**. Indeed, the distribution of boostrap sample mean is stochastic **as each sample draw ensues a different distribution**[4], but asymptotically same distribution as the sample mean
- Bootstrap CMT: $g(\overline{X}^*) \xrightarrow{p} g(\overline{X})$ for continuous function $g(\cdot)$
- Boostrap Delta Method[5]:
  $\sqrt{n}(g(\overline{X}^*) - g(\overline{X})) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial g(u)'}{\partial u}|_{u=\mu} \Sigma \frac{\partial g(u)}{\partial u}|_{u=\mu}\right)$ for continuously differentiable $g(\cdot)$
- Can apply Delta method to general smooth functions of sample means, which nests many estimators

[4] This is why we need to use empirical process and weak convergence.

[5] Functional Delta Method

A General Statement (Horowitz, 2001)

Definition: Let P_n denote the joint distribution of $\{X_i\}_{i=1}^n$, then $\mathcal{J}_n(\tau, \widehat{F}_n)$ is consistent if $\forall \varepsilon > 0$

$$\lim_{n\to\infty} P_n\Big[\sup_{\tau} |\mathcal{J}_n(\tau, \widehat{F}_n) - \mathcal{J}_\infty(\tau, F_0)| > \varepsilon\Big] = 0$$

Theorem (Beran and Ducharme, 1991):
$\mathcal{J}_n(\tau, \widehat{F}_n)$ is consistent if $\forall \varepsilon > 0$,

1. $\lim_{n\to\infty} P_n[\rho(\widehat{F}_n, F_0) > \varepsilon] = 0$ (since we use empirical CDF for this session, by SLLN it holds. ρ denotes the Mallows metric)
2. $\mathcal{J}_\infty(\tau, F)$ is a continuous function of τ $\forall F \in \mathcal{F}$
3. $\forall \tau$ and $\forall$ sequence $\{H_n\} \in \mathcal{F}$ such that $\lim_{n\to\infty} \rho(H_n, F_0) = 0$, $\mathcal{J}_n(\tau, H_n) \to \mathcal{J}_\infty(\tau, F_0)$

For sample mean example:

- $T_n = \sqrt{n}(\overline{X} - \mu)$
- $\mathcal{J}_n(\tau, F_0) = P_n(\sqrt{n}(\overline{X} - \mu) \leq \tau)$
- $\mathcal{J}_\infty(\tau, F_0) = \lim_{n\to\infty} P_n(\sqrt{n}(\overline{X} - \mu) \leq \tau)$
- $\mathcal{J}_n(\tau, \widehat{F}_n) = P_n^*(\sqrt{n}(\overline{X}^* - \overline{X}) \leq \tau)$

Use of Bootstrap

Variance and Standard Errors

- Denote $\overline{T}_n = \frac{1}{B}\sum_{i=1}^{B} \widehat{T}_n^b$,

$$\widehat{V}_{boot} = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{T}_n^b - \overline{T}_n)(\widehat{T}_n^b - \overline{T}_n)', \quad \widehat{se}_{boot} = \sqrt{\widehat{V}_{boot}}$$

Confidence Intervals

$$\left[T_n - z_{1-\alpha/2}\widehat{se}_{boot},\ T_n + z_{1-\alpha/2}\widehat{se}_{boot}\right]$$

Critical Values

- Denote $q^*_{1-\alpha}$ as the empirical $(1-\alpha)$th quantile of $\{\widehat{T}_n^b\}_{b=1}^{B}$
- Bootstrap percentile interval:

$$\left[q^*_{\alpha/2},\ q^*_{1-\alpha/2}\right]$$

Remark:

- Functions within T_n need to have bounded order $p \geq 1$ derivatives so that bootstrap variance is consistent (counterexample: $\theta = \mu_1/\mu_2$ where $\mu_i = \mathbb{E}(y_i)$, then $\widehat{\theta}$ has no finte sample variance)
- Need to trim $\widehat{T}_n^b$ by excluding tails.

Asymptotic Refinement

Bootstrap can provide more accurate approximations to the distributions of statistics than does the conventional asymptotic distribution theory. We focus on two aspects

1. First–order bias reduction of nonlinear estimators
2. Higher–order refinements to the rejection probabilities of tests and coverage probabilities of confidence intervals

Refinement: Bias Correction

- Nonlinear estimators are prone to finite–sample bias, and bootstrap can bias correct up to some asymptotic order
- Suppose the estimator using original sample is T_n
- Bias correction version is

$$2T_n - \overline{T}_n,$$

  where $\overline{T}_n$ is the average of bootstrap draws
- Let's look at a concrete example

Bias Correction Example

- For random vector X, consider $\mu = \mathbb{E}(X)$ and $\theta = g(\mu)$, where g is a known continuous nonlinear function
- For random sample $\{X_i\}_{i=1}^n$, define $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and thus a consistent estimator $\widehat{\theta} = g(\overline{X})$
- $\mathbb{E}[\widehat{\theta}] = \mathbb{E}[g(\overline{X})] \neq g(\mu) = \theta$, where the inequality is due to nonlinearity of g
- Hence $\widehat{\theta}$ is a biased estimator of θ
- Denote by G_i the ith order derivative of g, by Taylor expansion:

$$\widehat{\theta} - \theta = G_1(\mu)(\overline{X} - \mu) + \frac{1}{2}(\overline{X} - \mu)' G_2(\mu)(\overline{X} - \mu) + O_p(n^{-2})$$

- Taking expectation:

$$\mathbb{E}(\widehat{\theta} - \theta) = \frac{1}{2}\mathbb{E}\Big[(\overline{X} - \mu)' G_2(\mu)(\overline{X} - \mu)\Big] + O_p(n^{-2})$$

- Blue term: first–order bias

First–Order Bias Reduction

- In the bootstrap world, the true estimate is $\widehat{\theta}$
- Denote the bootstrap estimator as $\widetilde{\theta} := g(\overline{X}^*)$, where $\overline{X}^*$ is the sample mean of bootstrap sample
- Taylor expansion is now

$$\widetilde{\theta} - \widehat{\theta} = G_1(\overline{X})(\overline{X}^* - \overline{X}) + \frac{1}{2}(\overline{X}^* - \overline{X})' G_2(\overline{X})(\overline{X}^* - \overline{X}) + O_p(n^{-2})$$

- Taking conditional expectation (conditional on empirical CDF implied by the observed data)

$$\mathbb{E}^*(\widetilde{\theta} - \widehat{\theta}) = \frac{1}{2}\mathbb{E}^*\Big[((\overline{X}^* - \overline{X})' G_2(\overline{X})((\overline{X}^* - \overline{X})\Big] + O_p(n^{-2})$$

- Red term is the first–order bootstrap bias, which can be estimated, since we know $\overline{X}$
- Can show that $\mathbb{E}$(red term) = blue term + $O_p(n^{-2})$
- Hence the first–order bias correction form:

$$\theta_{bc} := \widehat{\theta} - \mathbb{E}^*(\widetilde{\theta} - \widehat{\theta}) = 2\widehat{\theta} - \mathbb{E}^*(\widetilde{\theta})$$

Higher–Order Refinement

Assume that T_n is a smooth[6] function H of functions of sample moments of X. ▸ Examples of Nonsmooth T_n Recall notations:

1. $\mathcal{J}_n(\tau, F_0)$: exact finite–sample distribution of T_n
2. $\mathcal{J}_\infty(\tau, F_0)$: exact asymptotic distribution of T_n
3. $\mathcal{J}_n(\tau, \widehat{F}_n)$: bootstrap distribution of T_n

Edgeworth expansion:

$$\mathcal{J}_n(\tau, F_0) = \mathcal{J}_\infty(\tau, F_0) + \frac{1}{\sqrt{n}} g_1(\tau, \kappa_1) + \frac{1}{n} g_2(\tau, \kappa_2) + \frac{1}{\sqrt{n^3}} g_3(\tau, \kappa_3) + O_p(n^{-2})$$

$$\mathcal{J}_n(\tau, \widehat{F}_n) = \mathcal{J}_\infty(\tau, \widehat{F}_n) + \frac{1}{\sqrt{n}} g_1(\tau, \kappa_{n1}) + \frac{1}{n} g_2(\tau, \kappa_{n2}) + \frac{1}{\sqrt{n^3}} g_3(\tau, \kappa_{n3}) + O_p(n^{-2})$$

Taking difference:

- $\mathcal{J}_\infty(\tau, F_0) - \mathcal{J}_\infty(\tau, \widehat{F}_n)$ is $O_p(n^{-1/2})$. If T_n is asymptotically pivotal? Zero! In addition, $\frac{1}{\sqrt{n}}(g_1(\tau, \kappa_1) - g_1(\tau, \kappa_{n1}))$ is at least $O_p(n^{-1})$.
- Implication? Error of bootstrap approximation converges to zero more quickly than asymptotic approximation

[6] Has bounded partial derivatives of sufficiently high order with respect to any combination of the components of Z

Two–Sided Hypothesis Testing

- Suppose T_n is an asymptotically pivotal test statistic for testing a null H_0 with size α
- Let $q_{1-\alpha}$ denote the $(1-\alpha)$ quantile of distribution of $|T_n|$, then $Pr(|T_n| \leq q_{1-\alpha}) = 1-\alpha$
- $q_{1-\alpha}$ is unknown, approximate by bootstrap and get the bootstrap critical value as the $(1-\alpha)$ quantile of distribution of $|\widehat{T}_n|$, denote it as $q^*_{1-\alpha}$
- Under the null, real size is $Pr(|T_n| > q^*_{1-\alpha})$ while nominal size is α

$$Pr(|T_n| > q^*_{1-\alpha}) - \alpha = O_p(n^{-2})$$

- What about asymptotic critical value: $1-\alpha/2$ quantile of $\mathcal{N}(0,1)$?

$$Pr(|T_n| > z_{1-\alpha/2}) - \alpha = O_p(n^{-1})$$

- **Takeaway:** Size distortion converges to zero more quickly using bootstrap CV, and in simulations using bootstrap CV usually features smaller distortion

A Partial Guide for Practitioners

- To use bootstrap for standard errors, the estimator must be asymptotically normal[7]
- To use higher–order refinement, the test statstic needs to be asymptotically pivotal[8]
- Trimming parameter affect bootstrap variance, leading to invalid bootstrap for (Ma and Wang, 2019, Inverse Probability Weighting)
- Do not use bootstrap for (weak) IV regressions (I.Andrews, Stock and Sun, 2018)

[7] In some cases the boostrap standard error is inconsistent even though the estimator is asymptotically normal. See Abadie and Imbens (2008) and Adusumilli (2018) on matching estimators.

[8] Under the smoothness assumption, $\sup_\tau |\mathcal{J}_n(\tau, \widehat{F}_n) - \mathcal{J}_n(\tau, F_0)|$ converges to zero as $n \to \infty$ faster than $\sup_\tau |\mathcal{J}_\infty(\tau, F_0) - \mathcal{J}_n(\tau, F_0)|$. Doesn't apply for overidentified GMM estimators. Need to recenter the moments or use empirical likelihood bootstrap. See class slides.

Bootstrap Failure: Parameter on the Boundary

- Consider a random sample $\{X_i\}_{i=1}^n$ from $\mathcal{N}(\mu, 1)$, where $\mu \geq 0$
- MLE estimate of μ is $\widehat{\mu}_{MLE} = \max(0, \overline{X})$, where $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$
- When $\mu > 0$, $\sqrt{n}(\widehat{\mu}_{MLE} - \mu) \xrightarrow{d} \mathcal{N}(0, 1)$
- When $\mu = 0$,

$$Pr(\sqrt{n}(\widehat{\mu}_{MLE} - \mu) \leq \tau) = \begin{cases} \Phi(\tau) & \tau \geq 0 \\ 0 & \tau < 0 \end{cases}$$

  discountinuity in the limiting distribution
- Large literature of moment inequality inference (D.Andrews, 2000)

Bootstrap Failure: 2sls and Weak IV

- Bootstrap 2sls estimator and t-stats have the same asymptotic distribution as the sample estimator (Hahn, 1996)
- Finite–sample distributions of IV estimators non–normal, test statistics non chi–squared
- Under the assumption of normal errors, IV estimator (just–identified) doesn't have finite moments, 2sls estimator (over–identified) number of finite moments equal overidentified restrictions. (Kinal, 1980)[9]
- If 2SLS estimators don't have a finite second moment, bootstrap variance unreliable.
- If IVs are weak, discontinuity in the limiting distribution (D.Andrews and Guggenberger, 2009), and bootstrap errs in estimating the strength of instruments (I.Andrews, Stock and Sun, 2018)

[9] Implication? You need more instruments than endogenous variables for finite moments, but finding IV is hard and subject to weak and many IV problems.

Asymptotically Pivotal Statistics

- When $\mathcal{J}_n(\tau, F)$, the distribution function of T_n, doesn't depend on F, T_n is called pivotal (e.g., t-statistic)
- Many statistics have limiting distributions as standard normal or chi–squared, regardless of the distribution from which the data were sampled[10]
- These statistics are asymptotically pivotal: $\mathcal{J}_\infty(\tau, F) = \mathcal{J}_\infty(\tau)$
- Implication? If n is very large, $\mathcal{J}_n(\tau, F_0)$ can be estimated by $\mathcal{J}_\infty(\tau)$, requiring no knolwedge of F_0
- **Note**: many statstics are asymptotically normal, so depends on mean and variance. Normal distribution with *estimated mean and variance* can be used to approximate $\mathcal{J}_n(\tau, F_0)$ for large n
- This is the rationale behind using asymptotic distribution in inference, but in practice (**finite–sample** simulations and applications) the approximation can be poor Back

[10] Another way of saying this is that their limiting distributions don't depend on some unknown population parameters.

Examples of Non–Smooth T_n

- Quantile regression: objective function is a check function
- Manski's maximum score estimator
- Nonparametric density estimation
- Semiparametric estimators

Refer to Horowitz (2001, 2017) on alternative bootstrap methods Back