STAT 620: Asymptotic Statistics **Spring 2022**

Lecture: Apr 21

Lecturer: Xianyang Zhang

1 Big picture

	Real world	Bootstrap world
distribution	$P \mid \theta$	$\hat{P} \mid \hat{\theta}_n$
samples	$X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} P \mid \theta$	$X_1^*, X_2^*, \ldots, X_n^* \overset{i.i.d}{\sim} \hat{P} \mid \hat{\theta}_n$
parameter estimates	$\hat{\theta}_n$, $\hat{\sigma}_n$ using $X_1, X_2, \ldots, X_n$	$\hat{\theta}_n^*$, $\hat{\sigma}_n^*$ using $X_1^*, X_2^*, \ldots, X_n^*$
pivotal quantity	$\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n}$	$\frac{\hat{\theta}_n^* - \hat{\theta}_n}{\hat{\sigma}_n^*}$

Suppose we know

$$\mathbb{P}\left[\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq x\right] \to F(x).$$

Then to show validity of the Bootstrap method, we show

$$\mathbb{P}\left[\frac{\hat{\theta}_n^* - \hat{\theta}_n}{\hat{\sigma}_n^*} \leq x \,\middle|\, X_1, X_2, \ldots, X_n\right] \overset{a.s.}{\to} F(x).$$

2 Bootstrap consistency: the mean case

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d with mean μ and covariance matrix Σ. From CLT we know

$$\sqrt{n}(\bar{X}_n - \mu) \overset{d}{\to} N(\mathbf{0}, \Sigma).$$

Given this result, we can ask ourselves the following question:

2.1 Question

Can we show

$$\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \overset{d^*}{\to} N(\mathbf{0}, \Sigma)?$$

Here, d^* denotes $\overset{d}{\to}$ almost surely and $\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*$. Almost surely on the sequence $X_1, X_2, \ldots, X_n$, the conditional distribution of the bootstrap estimate of the mean satsifies that

$$\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \overset{d^*}{\to} N(\mathbf{0}, \Sigma).$$

Indeed by the Edgeworth expansion below, it can be shown that, the distribution of $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ is closer to that of $\sqrt{n}(\bar{X}_n - \mu)$. Thus bootstrap estimate provides a better finite sample approximation to $\sqrt{n}(\bar{X}_n - \mu)$ than the normal approximation $N(\mathbf{0}, \Sigma)$.

2.2 Proof

Let us denote $X_1, X_2, \ldots, X_n$ by $X_{1:n}$ Note that,

$$\mathbb{E}\big[\bar{X}_n^* \,\big|\, X_1, X_2, \ldots, X_n\big] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[X_i^* \,\big|\, X_{1:n}\big] = \mathbb{E}\big[X_1^* \,\big|\, X_{1:n}\big] = \bar{X}_n.$$

Fix any $i = 1, \cdots, n$. Similarly,

$$\mathbb{E}\big[X_i^* \, X_i^{*\top} \,\big|\, X_{1:n}\big] = \frac{1}{n}\sum_{i=1}^{n} X_i \, X_i^{\mathrm{T}}.$$

This implies that

$$\mathrm{cov}\big[X_i^* \,\big|\, X_{1:n}\big] = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)\,(X_i - \bar{X}_n)^\top$$

and

$$\mathrm{cov}\big[X_i \,\big|\, X_{1:n}\big] \overset{a.s.}{\to} \Sigma.$$

Following the Lindeberg condition, the goal is to show

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\, ||X_i^*||^2 \,\mathbf{1}\big\{||X_i^*|| > \epsilon\sqrt{n}\big\} \,\big|\, X_{1:n}\big] \overset{a.s.}{\to} 0.$$

To this end, we note that for any ϵ, there exists a large enough M such that $\mathbb{E}||X_1||^2 \,\mathbf{1}\big\{||X_1|| > M\big\}\big] < \epsilon$. Thus when n is large enough, we get

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\, ||X_i^*||^2 \,\mathbf{1}\big\{||X_i^*|| > \epsilon\sqrt{n}\big\} \,\big|\, X_{1:n}\big] &= \mathbb{E}\big[\, ||X_1^*||^2 \,\mathbf{1}\big\{||X_1^*|| > \epsilon\sqrt{n}\big\} \,\big|\, X_{1:n}\big] \\
&= \frac{1}{n}\sum_{i=1}^{n} ||X_i||^2 \,\mathbf{1}\big\{||X_i|| > \epsilon\sqrt{n}\big\} \\
&\le \frac{1}{n}\sum_{i=1}^{n} ||X_i||^2 \,\mathbf{1}\big\{||X_i|| > M\big\} \overset{a.s.}{\to} \mathbb{E}||X_1||^2 \,\mathbf{1}\big\{||X_1|| > M\big\}\big] < \epsilon.
\end{aligned}$$

As ϵ can be arbitrarily small, the Lindeberg condition holds almost surely.

2.3 Bootstrap consistency: the general case

Suppose we can show $\hat{\theta}_n \overset{a.s.}{\to} \theta$ and $\sqrt{n}(\hat{\theta}_n - \theta) \overset{d}{\to} T$. Then it can be shown that

- $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \overset{d^*}{\to} T$ conditional on $X_1, X_2, \cdots, X_n$.
- For any ϕ such that it is continuously differentiable at θ, then $\sqrt{n}\Big(\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\Big) \overset{d^*}{\to} D_\theta T$.

The aforementioned results can also be generalized to empirical process. For example,

- $\big\{\sqrt{n}(\mathbb{P}_n - \mathbb{P})\, f\big\}_{f\in\mathcal{F}} \overset{d}{\to} \big\{G_f\big\}_{f\in\mathcal{F}}$.
- $\big\{\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)\, f\big\}_{f\in\mathcal{F}} \overset{d^*}{\to} \big\{G_f\big\}_{f\in\mathcal{F}}$.

2.4 Edgeworth expansion

The Edgeworth expansion is used to show the high-order accuracy for bootstrap method. Specifically, we have

$$\mathbb{P}\left[\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}_n} \le x\right] = \Phi(x) + \phi(x)\left[\frac{p_1(x;\mu_3)}{\sqrt{n}} + \frac{p_2(x;\mu_3,\mu_4)}{n} + O\left(\frac{1}{n^{3/2}}\right)\right].$$

In the right-hand side of the expression, p_1 and p_2 are polynomials, and μ_3 and μ_4 are the skewness and kurtosis of the population respectively. Similarly, for the bootstrap version, we have,

$$\mathbb{P}\left[\frac{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)}{\hat{\sigma}_n^*} \leq x\right] = \Phi(x) + \phi(x)\left[\frac{p_1(x;\hat{\mu}_3)}{\sqrt{n}} + \frac{p_2(x;\hat{\mu}_3,\hat{\mu}_4)}{n} + O\left(\frac{1}{n^{3/2}}\right)\right].$$

If we compare these two expanded expressions, we can see that bootstrap tries to match higher-order moments which leads to the high-order accuracy and thus faster convergence rate compared to the first order normal approximation.

Reference: For additional details, check the book *The Bootstrap and Edgeworth Expansion* by Peter Hall.

3 Gaussian Sequence Model

Consider the model $Y \sim N(\mu, \sigma^2 I_p)$ or equivalently $Y_i = \mu_i + \sigma\varepsilon_i$, where $\varepsilon_i \overset{iid}{\sim} N(0,1)$ for $i = 1, 2, \ldots, p$. This model is of basic interest in empirical bayes, nonparametric regression, variable selection, multiple hypothesis testing, admisibility (JS estimator) and so on.

3.1 Goal

Estimate μ under the sparsity assumption, i.e. $\|\mu\|_0 \leq k$. For $S \subseteq \{1, \ldots, p\}$, define $\hat{\mu}(S)$ for μ as

$$\hat{\mu}_i(S) = \begin{cases} 0, & \text{if } i \notin S, \\ Y_i, & \text{if } i \in S. \end{cases}$$

Therefore, the L^2-risk of $\hat{\mu}(S)$ is given by

$$R(\mu, \hat{\mu}(S)) = \mathbb{E}||\mu - \hat{\mu}(S)||^2 = \sum_{i\in S}\sigma^2 + \sum_{i\notin S}\mu_i^2.$$

3.2 Ideal risk

$$R^I(\mu) = \min_S R(\mu, \hat{\mu}(S)) = \min_S \left[\sum_{i\in S}\sigma^2 + \sum_{i\notin S}\mu_i^2\right] = \sum_{i=1}^{p}\min(\sigma^2, \mu_i^2).$$

Note that, when $\|\mu\|_0 \leq k$, then $R^I(\mu) \leq k\sigma^2$. When p is large, the risk of MLE (Y) $p\sigma^2 \gg k\sigma^2$.

3.3 Hard Thresholding rule

$$\eta_H(y, \lambda) = \begin{cases} y, & \text{if } |y| \geq \lambda, \\ 0, & \text{if } |y| < \lambda. \end{cases}$$

3.4 Soft Thresholding rule

$$\eta_S(y, \lambda) = \begin{cases} y - \lambda, & \text{if } y \geq \lambda \\ 0, & \text{if } |y| < \lambda \\ y + \lambda, & \text{if } y \leq -\lambda \end{cases} = \text{sgn}(y)\,(|y| - \lambda)_+,$$

where sgn is the sign function and $x_+ = x\,I\{x \geq 0\}$. It can be verified that

$$\eta_S(y, \lambda) = \text{argmin}_\mu \left[\frac{1}{2}(y - \mu)^2 + \lambda|\mu| \right].$$