Statistics 210A Lecture 25 Notes

Daniel Raban

November 23, 2021

1 Introduction to Bootstrap

1.1 Recap: Comparison of bootstrap to other kinds of inference

So far we have done:

- Exact, finite-sample inference
  - Requires special structure
  - No reliance on asymptotic approximation
  - Parametric or non-parametric (e.g. permutation tests)
- Parametric, asymptotic inference
  - Simple ideas, leading to asymptotically optimal results.
  - Only relies on regularity conditions

Today, we will study asymptotic, nonparametric inference.

1.2 Functionals and plug-in estimators

Suppose we have a nonparametric iid sample $X_1, \dots, X_n \overset{\text{iid}}{\sim} P$. We want inference on some "parameter" $\theta(P) \subseteq \mathbb{R}^d$. More precisely, we want a functional $\theta(P)$.

Example 1.1. If the sample space $\mathcal{X} \subseteq \mathbb{R}$, we could look at

$$\theta(P) = \text{median}(P).$$

Example 1.2. If the sample space $\mathcal{X} \subseteq \mathbb{R}^d$, we could look at

$$\theta(P) = \lambda_{\max}(\text{Var}_P(X_i)).$$

Example 1.3. If we are doing linear regression, we could look at

$$\theta(P) = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_P[(Y_i - \theta^\top X_i)^2],$$

where $(X_i, Y_i) \overset{\text{iid}}{\sim} P$.

Example 1.4. More generally, we can set

$$\begin{aligned}\theta(P) &= \arg\min_{\theta \in \Theta} D_{\mathrm{KL}}(P \mid\mid P_\theta) \\ &= \arg\max_{\theta} \mathbb{E}_P[\ell_1(\theta; X_i)].\end{aligned}$$

Note that in these cases, we may have many distribution with the same value $\theta(P)$.

Definition 1.1. The **empirical distribution** of $X_1, \ldots, X_n$ is the random measure

$$\widehat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}, \qquad \widehat{P}_n(A) = \frac{\#\{i : X_i \in A\}}{n}.$$

Definition 1.2. The **plug-in estimator** of $\theta(P)$ is $\widehat{\theta}_n = \theta(\widehat{P}_n)$.

Example 1.5. If $\theta(P)$ is the median, $\widehat{\theta}_n$ is the sample median.

Example 1.6. If $\theta(P) = \lambda_{\max}(\mathrm{Var}_P(X_i))$, then the plug-in estimator is $\lambda_{\max}(\text{samplevariance})$.

Example 1.7. For linear regression, the plug-in estimator is the OLS etimator.

Example 1.8. For the minimizer of the KL-divergence, the plug-in estimator is the MLE for $\{P_\theta : \theta \in \Theta\}$.

1.3 Convergence of plug-in estimators

Does using the plug-in estimator work? It depends. Whether $\widehat{P}_n \xrightarrow{p} P$ depends on what distance we use. We have pointwise convergence, $\widehat{P}_n(A) \xrightarrow{p} P(A)$ for all A by the weak law of large numbers. We can consider convergence in the *total variation distance* (i.e. uniform convergence of these functions):

$$\sup_A |\widehat{P}_n(A) - P(A)| \xrightarrow{p} 0?$$

This is true if the sample space $\mathcal{X}$ is finite. However, if $\mathcal{X} = \mathbb{R}$ and P is continuous, this is not true because if $A^* = \{x_1, \ldots, x_n\}$, then $P(A^*) =)$ but $\widehat{P}_n(A^*) = 1$.

If $X \subseteq \mathbb{R}$, we can look at convergence of the CDFs:

$$\sup_x |\widehat{P}_n((-\infty, x]) - P((-\infty, x])| \xrightarrow{p} 0 \qquad \forall x \in \mathbb{R}.$$

We want the functional $\theta(\cdot)$ to be continuous with respect to some topology in which $\widehat{P}_n \xrightarrow{p} P$, so $\theta(\widehat{P}_n) \xrightarrow{p} \theta(P)$ by the continuous mapping theorem.

Here is a counterexample to keep in mind, so you don't think that bootstrap always works:

Example 1.9. Let

$$\theta(P) = \begin{cases} 1 & \mathbb{P}_{X_1, X_2 \overset{\text{iid}}{\sim} P}(X_1 = X_2) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

If P is continuous, then $\theta(P) = 0$. But $\theta(\widehat{P}_n) = 1$ for all n.

1.4 Bootstrap standard errors

Suppose $\widehat{\theta}_n$ is any estimator of $\theta(P)$. We want to know the standard error $\text{s.e.}(\widehat{\theta}_n) = \sqrt{\text{Var}_P(\widehat{\theta}(X_1, \ldots, X_n))}$.

The only thing here we don't know is P, so we will plug in $\widehat{P}_n$:

$$\widehat{\text{s.e.}}(\widehat{\theta}_n) = \sqrt{\text{Var}_{\widehat{P}_n}(\theta_n^*)}.$$

Here, the star notation is just to make sure we know that $\widehat{\theta}_n^*$ is a random variable drawn from $\widehat{P}_n$. We can write

$$\text{Var}_{\widehat{P}_n}(\widehat{\theta}_n^*) = \text{Var}_{X_i^* \overset{\text{sim}}{\sim} \widehat{P}_n}(\widehat{\theta}_n(X_1^*, \ldots, X_n^*)).$$

Often, bootstrap is defined algorithmically.

How do we calculate this? We will use Monte Carlo with $\widehat{P}_n$ instead of P: For $b = 1, \ldots, B$, Sample $X_1^{*b}, \ldots, X_n^{*b} \overset{\text{iid}}{\sim} \widehat{P}_n$ (resampling n values with replacement from $X_1, \ldots, X_n$), and let $\widehat{\theta}^{*b} = \widehat{\theta}(X_1^{*b}, \ldots, X_n^{*b})$. Then let $\overline{\theta}^* = \frac{1}{B}\sum_{b=1}^B \widehat{\theta}^{*b}$, so the standard error is

$$\widehat{\text{s.e.}}(\widehat{\theta}_n) = \sqrt{\frac{1}{B}\sum_{b=1}^B (\widehat{\theta}^{*b} - \overline{\theta}^*)^2}.$$

1.5 Bootstrap Bias Estimation/Correrction

Let $\widehat{\theta}_n$ be some estimator. What is its bias?

$$\text{Bias}_P(\widehat{\theta}_n) = \mathbb{E}_P[\widehat{\theta}_n - \theta(P)].$$

The idea is to plug in $\widehat{P}_n$ for P:

$$\widehat{\text{Bias}}(\widehat{\theta}_n) = \text{Bias}_{\widehat{P}_n}(\widehat{\theta}_n^*) = \mathbb{E}_{\widehat{P}_n}[\widehat{\theta}_n^* - \theta(\widehat{P}_n)].$$

We can calculate this using Monte Carlo: Sample $X_1^{*b}, \ldots, X_n^{*b} \overset{\text{iid}}{\sim} \widehat{P}_n$, and calculate the estimator $\widehat{\theta}^{*b} = \widehat{\theta}(X^{*b})$. Then we have the average $\overline{\theta}^* = \frac{1}{B}\sum_{b=1}^{B} \widehat{\theta}^{*b}$, so we can calculate

$$\widehat{\text{Bias}}(\widehat{\theta}_n) = \overline{\theta}^* - \theta(\widehat{P}_n).$$

Remark 1.1. The advantage of thinking of this as a plug-in estimator instead of just defining it algorithmically is that it becomes more conceptually clear why we subtract $\theta(\widehat{P}_n)$ instead of $\theta(P)$.

We can then define the **Bias-corrected estimator**

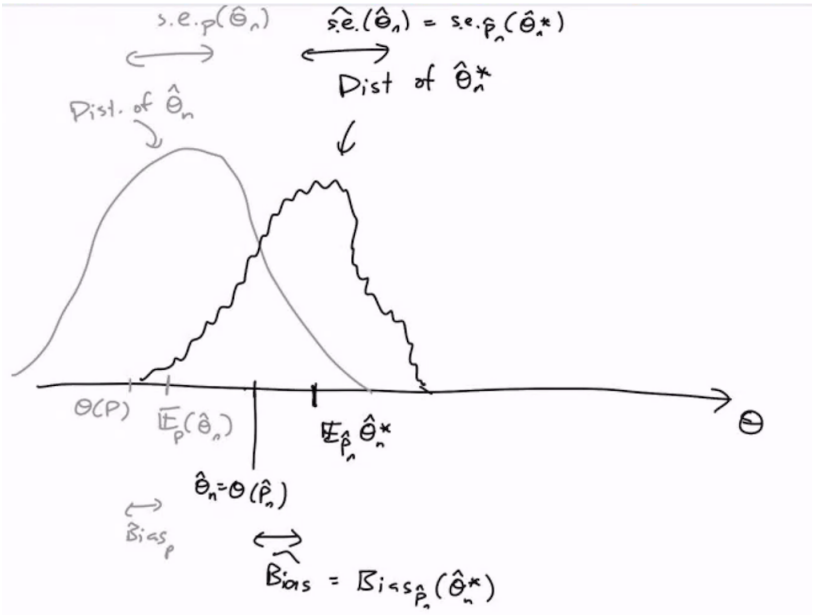
$$\widehat{\theta}_n^{\text{BC}} = \widehat{\theta}_n - \widehat{\text{Bias}}(\widehat{\theta}_n)$$

If $\theta(\widehat{P}_n) = \widehat{\theta}_n$,

$$= 2\widehat{\theta}_n - \overline{\theta}^*.$$

Remark 1.2. If we know the actual bias, it's always better to subtract it because we reduce the bias while keeping the variance the same. However, it is not always better to subtract out the estimated bias because the estimate could be wrong. In particular, the estimate of the bias might be noisy, so it might introduce some variance. Typically, $\widehat{\theta}_n^{\text{BC}}$ has a lower bias but a higher variance than $\widehat{\theta}_n$.

Here is a picture. The things that we can't see are in gray, and what we can see is in black.



Here is a table of analogies between the "real world" and "bootstrap world."

	"Real world"	"Bootstrap world"
Samplingdistribution	P	$\widehat{P}_n$
Parameter	$\theta(P)$	$\theta(\widehat{P}_n)$ (maybe $\widehat{\theta}_n$)
Dataset	$X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$	$X_1^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \widehat{P}_n$
Estimator	$\widehat{\theta}_n(X_1, \ldots, X_n)$	$\widehat{\theta}_n^*(X_1^*, \ldots, X_n^*)$
Standard error of estimator	$\sqrt{\text{Var}_P(\widehat{\theta}_n)}$	$\sqrt{\text{Var}_{\widehat{P}_n}(\widehat{\theta}_n^*)}$
Bias of estimator	$\text{Bias}_P(\widehat{\theta}_n)$	$\text{Bias}_{\widehat{P}_n}(\widehat{\theta}_n^*)$