Lecture 29: Continuation of Bootstrap Discussion

Lecturer: Michael I. Jordan *Scribe: Mike Higgins*

1 Theory of Bootstrap

Oftentimes, we will have a statistic in the form of $\phi_n(F)$ instead of $\phi(F)$, and we will want to estimate performance measures in this setting. Examples of this include:

- CDF: $\lambda_n(F) = P_F(\sqrt{n}(\hat{\theta}_n - \phi(F)) \leq a)$
- Bias: $\lambda_n(F) = E_F(\hat{\theta}_n) - \phi_n(F)$
- Variance: $\lambda_n(F) = \sqrt{n} E_F(\hat{\theta}_n - \phi_n(F))^2$

The basic idea of the bootstrap method is to replace F with $\hat{F}_n$.

Example 1. Suppose $\lambda_n(F) = P_F(\sqrt{n}(\hat{\theta}_n - \phi(F)) \leq a)$. Replace F with $\hat{F}_n$ throughout, thus $\hat{\theta}_n$ becomes a function of "data" $X_1^*, X_2^*, \ldots, X_n^*$ sampled from $\hat{F}_n$. So $\lambda_n(\hat{F}_n) = P_{\hat{F}_n}(\sqrt{n}(\hat{\theta}_n^* - \phi(\hat{F}_n)) \leq a)$.

Example 2 (U-Statistic)**.** Let $\hat{\theta}_n = \frac{2}{n(n-1)} \sum_{i<j} \psi(X_i, X_j)$. We have shown that $\lambda_n(F) = \frac{4(n-2)}{n-1}\gamma_1^2 + \frac{2}{n-1}\gamma_2^2$, where $\gamma_1^2 = E(\psi(X_1, X_2)\psi(X_1, X_3))$ and $\gamma_2^2 = E(\psi(X_1, X_2)^2)$, and so, $\lambda_n(F) \to \lambda(F) = 4\gamma_1^2$. On the other hand, we have that $\lambda_n(\hat{F}_n) = \frac{4(n-2)}{n-1}\gamma_1^{*2} + \frac{2}{n-1}\gamma_2^{*2}$ where $\gamma_1^{*2} = \frac{1}{n^3}\sum_i \sum_j \sum_k \psi(X_i, X_j)\psi(X_i, X_k)$ and $\gamma_2^{*2} = \frac{1}{n^2}\sum_i \sum_j \psi(X_i, X_j)^2$. Let $\gamma_3^2 = E(\psi(X_i, X_i)^2)$. If we have that $\gamma_1^{*2}, \gamma_2^{*2}$, and γ_3^2 are all finite, then we have consistency; $\lambda_n(\hat{F}_n) \to \lambda(F) = 4\gamma_1^2$. However, we will show that if $\gamma_3^2 = \infty$, we may not have consistency.

Let X_i be i.i.d. Uniform$(0, 1)$ variables, and define ψ so that when $i \neq j$, $|\psi(X_i, X_j)| \leq M$ for some real number $M < \infty$, and $\psi(X_i, X_i) = \exp(\frac{1}{X_i})$. For divergence of $\lambda_n(\hat{F}_n)$, we need $P(\frac{1}{n^2}\sum_i e^{\frac{1}{X_i}} > A) \to 1$ for all $A > 0$. Since $\sum_i e^{\frac{1}{X_i}} \geq \max_i e^{\frac{1}{X_i}}$, we can prove divergence by showing $P(\max_i e^{\frac{1}{X_i}} \leq An^2) = \Big(P(e^{\frac{1}{X_1}} \leq An^2)\Big)^n \to 0$. To show this, note $P(e^{\frac{1}{X_i}} \leq An^2) = P(X_i > \frac{1}{\log(An^2)}) = 1 - \frac{1}{\log(An^2)}$. Since $\frac{1}{\log(An^2)} \geq \frac{1}{\sqrt{n}}$ for sufficiently large n, and $(1 - \frac{1}{\sqrt{n}})^n \to 0$, it follows that $P(\max_i e^{\frac{1}{X_i}} \leq An^2) \to 0$, and we have divergence of the bootstrap estimator.

1.1 Comparing weak convergence-based approximations and boostrap.

Suppose $\lambda_n(F) \xrightarrow{d} \lambda$, which is independent of F. We can use λ as an approximation to $\lambda_n(F)$, or we can use $\lambda_n(\hat{F}_n)$. If we suppose $\lambda_n(F) = \lambda + \frac{\alpha(F)}{n} + o(n^{-1})$, where α is a coefficient depending on the distribution, then $\lambda_n(\hat{F}_n) = \lambda + \frac{\alpha(\hat{F})}{n} + o(n^{-1})$. Additionally, if we suppose that $\sqrt{n}(\alpha(\hat{F}_n) - \alpha(F))$ is tight, then we have $\alpha(\hat{F}_n) = \alpha(F) + o_p(1)$, and so, $\lambda_n(\hat{F}_n) = \lambda_n(F) + o_p(n^{-1})$ This is better than our $O_p(n^{-1})$ result obtained from using λ.

If, on the other hand, λ is not independent of F, we get $\lambda_n(\hat{F}_n) = \lambda + \frac{\alpha(\hat{F})}{n} + o(n^{-1})$, which implies $\lambda_n(\hat{F}_n) - \lambda_n(F) = \lambda(\hat{F}_n) - \lambda(F) + \frac{1}{n}(\alpha(\hat{F}_n) - \alpha(F)) + o(n^{-1}) = O(n^{-1})$ since $\lambda(\hat{F}_n) - \lambda(F)$ is $O(n^{-1})$.

Example 3. Suppose $\phi(F) = \sigma^2$. Then $\phi(\hat{F}_n) = \frac{1}{n}\sum_i (X_i - \bar{X}_n)^2 =: M_2$, where M_i is the ith central sample moment.

1. Let $\lambda_n(F) = \text{Var}(\sqrt{n}M_2) = (\mu_4 - \mu_2^2) - \frac{2(\mu_4-\mu_2^2)}{n} + \frac{\mu_4 - 3\mu_2^2}{n^2}$, where μ_i is the ith central moment. The classical estimator is $\lambda(\hat{F}_n) = (M_4 - M_2^2)$, but the bootstrap estimator is $\lambda_n(\hat{F}_n) = (M_4 - M_2^2) - \frac{2(M_4 - M_2^2)}{n} + \frac{M_4 - 3M_2^2}{n^2}$. For both estimators, the error is $(M_4 - M_2^2) - (\mu_4 - \mu_2^2) + O(n^{-1})$, which is $O(n^{-\frac{1}{2}})$ because $M_i = \mu_i + O(n^{-\frac{1}{2}})$.

2. Note that $E(M_2) = \frac{n-1}{n}\sigma^2$, and let $\lambda_n(F)$ be the bias of M_2, that is, $\lambda_n(F) = \frac{n-1}{n}\sigma^2 - \sigma^2 = \frac{\sigma^2}{n}$. We have $\lambda_n(F) \to \lambda = 0$, which is independent of F, and so, it is possible that the bootstrap estimator will converge faster than the classical estimator. We will now show that this is the case. Note that the bootstrap estimator $\lambda_n(\hat{F}_n) = \frac{1}{n}M_2 = \frac{1}{n}(\sigma^2 + O(n^{\frac{1}{2}}))$, which implies $\lambda_n(\hat{F}_n) - \lambda_n(F) = O(n^{-\frac{3}{2}})$, which beats the $O(n^{-1})$ rate of the classical estimator!

1.2 Bootstrap Confidence Intervals

Define a *root* $R_n(X_n, \theta(P))$ as a quantity that can be inverted to obtain a confidence interval. The classical example of a root is $R_n(X_n - \theta(P)) = \frac{\hat{\theta}_n - \theta(P)}{s_n}$, where s_n is some estimate of the standard deviation. To obtain confidence intervals based on R_n, we need the distribution of R_n, which we will call $\lambda_n(P)$. That is, $\lambda_n(P,t) = P(R_n(X_n, \theta(P)) \le t)$. The simplest case occurs when λ_n is independent of P, in which case, we call R_n is called a *pivot*.

Example 4. Suppose $X_i \overset{i.i.d.}{\sim} N(\theta, \sigma^2)$. Then $\lambda_n = \frac{\bar{X} - \theta}{s_n/\sqrt{n}} \sim t_{n-1}$, which is independent of θ and σ^2. In this instance, λ_n is a pivot.

In general, if R_n is a pivot, and there is a t such that $P\left(\left|\frac{\hat{\theta}_n - \theta(P)}{s_n/\sqrt{n}}\right| \le t\right) = 1-\alpha$ for all P, then $\left(\hat{\theta}_n - t\frac{s_n}{\sqrt{n}}, \hat{\theta}_n + t\frac{s_n}{\sqrt{n}}\right)$ is a $(1-\alpha)$ confidence interval for $\theta(P)$ independent of P.

In the case of the bootstrap, we approximate $\lambda_n(P)$ by $\lambda_n(\hat{P}_n)$, and we consider the set $B_n(1-\alpha, X_n) := \{\theta \in \Theta : \lambda_n^{-1}(\frac{\alpha}{2}, \hat{P}_n) \le R_n(X_n, \theta) \le \lambda_n^{-1}(1 - \frac{\alpha}{2}, \hat{P}_n)\}$. We can use a Monte Carlo method to estimate $\lambda_n^{-1}(\cdot, \hat{P}_n)$.

Lemma 5. *(van der Vaart, 1998, Lemma 23.3): Assume* $\frac{\theta_n - \theta}{\hat{\sigma}_n} \xrightarrow{d} T$ *and* $\frac{\theta_n^* - \hat{\theta}_n}{\sigma_n^*} \xrightarrow{d} T$. *Then the bootstrap confidence intervals are asymptotically consistent.*

Theorem 6 (Sample means)**.** *(van der Vaart, 1998, Theorem 23.4): Suppose* X_i *are i.i.d. with* $E(X_i) = \mu$ *and* $Cov(X_i, X_j) = \Sigma$. *Then, conditionally on* $X_1, X_2, \ldots, X_n$, $\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \xrightarrow{d} N(0, \Sigma)$ *for almost every sequence* $X_1, X_2, \ldots$.

Theorem 7 (Delta method for bootstrap)**.** *(van der Vaart, 1998, Theorem 23.5): Let* ϕ *be differentiable in a neighborhood of* θ, *let* $\hat{\theta}_n \xrightarrow{a.s.} \theta$, *and let* $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} T$, $\sqrt{n}(\theta_n^* - \hat{\theta}_n) \xrightarrow{d} T$. *Then* $\sqrt{n}(\phi(\theta_n) - \phi(\theta)) \xrightarrow{d} \phi_\theta'(T)$ *and* $\sqrt{n}(\phi(\theta_n^*) - \phi(\hat{\theta}_n)) \xrightarrow{d} \phi_\theta'(T)$ *conditionally almost surely.*

References

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.