

## Consistency of bootstrap

$G_n(t, \hat{F}) \approx G_\infty(t, \hat{F}) \approx G_\infty(t, F_0) \approx G_n(t, F_0)$  when  $n$  is large  
if  $G_\infty$  is continuous wrt.  $F$

Defn:  $G_n(t, \hat{F})$  is consistent if  $\forall F_0 \in \mathcal{F}$ ,  
 $\sup_t |G_n(t, \hat{F}) - G_\infty(t, F_0)| \xrightarrow{P} 0$ .

Example:  $T_n = \sqrt{n}(\bar{X} - \theta)$ ,  $\theta = E_{F_0}[X]$

We want to know  $G_n(t, F_0) = \Pr_{F_0}(T_n \leq t)$

•  $F_n$  = empirical distribution fcn.

•  $\theta(F_n) = E_{F_n}[X] = \bar{X}$

Bootstrap:  $T_n^* = \sqrt{n}(\bar{X}^* - \bar{X})$ , where  $\bar{X}^*$  is bootstrap sample.

• This gives us  $G_n(t, F_n) = \Pr_{F_n}(T_n \leq t)$

• Should work if  $T_n$  varies smoothly and CLT holds.

Thm (Bickel and Ducharme)  $G_n(t, \hat{F})$  is consistent

if  $\forall F_0 \in \mathcal{F}$ , (i)  $\rho(\hat{F}, F_0) \xrightarrow{P} 0$ , (ii)  $\forall F_n \in \mathcal{F}$

s.t.  $\rho(F_n, F_0) \rightarrow 0$ , we have

$$\sup_t |G_n(t, F_n) - G_\infty(t, F_0)| \rightarrow 0$$

$\rho$  some metric.

Thm (Mammen): Let  $\{\bar{X}_i; i \leq n\}$  be an iid sample.

For a sequence of normalizing constants  $t_n$  and  $\sigma_n$ , define

$$\bar{g}_n = \frac{1}{n} \sum g_n(\bar{X}_i), \quad T_n = \frac{(\bar{g}_n - t_n)}{\sigma_n}$$

$$\bar{g}_n^* = \frac{1}{n} \sum g_n(\bar{X}_i^*), \quad T_n^* = \frac{(\bar{g}_n^* - t_n)}{\sigma_n}$$

Nonparametric bootstrap is consistent iff  $T_n$  is asymptotically normal.

Example: • Let  $\bar{X}_i = \underbrace{(Y_i)}_{\text{dep. var.}}, \underbrace{(W_i)}_{\text{regressor}}$

$$\bullet Y_i = W_i' \beta + \varepsilon_i$$

Interested in  $\theta(F_0) = \beta_j$ ,

$$\text{Statistic: } T_n = \frac{\sqrt{n} (\hat{\beta}_j - \beta_j)}{\text{se}(\hat{\beta}_j)}$$

Want to know  $G_n(t, F_0) = \Pr_{F_0}(T_n \leq t)$

Under bootstrap,  $\theta(F_n) = \hat{\beta}_j$ . ("population parameter")

$$\bullet \text{ compute } T_n^* = \frac{\sqrt{n} (\hat{\beta}_j^* - \hat{\beta}_j)}{\text{se}^*(\hat{\beta}_j)}$$

Bootstrap gives us  $G_n(t, F_n) = \Pr_{F_n}(T_n \leq t)$ , which is consistent.

- Can show that  $T_n \rightarrow N(0, 1)$
- i.e.  $G_n(t, \hat{F}) \approx \Phi(t)$  when  $n$  is large.

Defn:  $T_n$  is asymptotically pivotal relative to a class of DGPs  $\mathcal{F}$  if  $G_\infty(t, F) = G_\infty(t)$   $\forall F \in \mathcal{F}$ . (cf pivotality:  $G_n(t, F) = G_n(t) \forall F \in \mathcal{F}$ )

Claim: The approximation error applied to asymptotically pivotal statistics is smaller than if they are not asymptotically pivotal.

### Failure of nonparametric bootstrap

- What if  $T_n$  is not asymptotically normal?
- example:  $\sum_i \sim \text{Cauchy}$ , then  $\bar{g}_n \sim \text{Cauchy}$  (sum stability of Cauchy). Thus, bootstrap fails by Mammen's theorem.
- There are cases where this could matter.
- What if  $G_\infty(t, F)$  is not continuous wrt  $F$ ?
- Then  $G_\infty(t, \hat{F})$  can deviate from  $G_\infty(t, F_0)$ .

Example:  $\sum_i \sim U[0, \theta_0]$

- $T_n = n(\theta_n - \theta_0)$ ,  $\theta_n = \max\{\sum_1, \dots, \sum_n\}$
- $F_n$  nonparametric bootstrap
- $T_n^* = n(\theta_n^* - \theta_n)$ ,  $\theta_n^* = \max\{\sum_1, \dots, \sum_n\}$

◦  $T_n \xrightarrow{d} \text{exponential} = \ln U[0,1]$

◦  $T_n^* \xrightarrow{d} \text{something else}$

◦  $\Pr[T_n^* = 0] = 1 - \Pr[T_n^* < 0] = 1 - (1 - \frac{1}{n})^n \rightarrow 1 - e^{-1}$

◦ i.e. there is a pointmass

There is failure in most extreme quantile regressions

◦ When most of the information comes from the extremes of the distribution, bootstrapping has problems.

## Subsampling

Draw subsamples of size  $m < n$

◦ less accurate than bootstrap, but works when bootstrap fails. (Sometimes.)

◦ There are also computational advantages

## m out of n bootstrap:

◦ Estimate  $G_n(t, F_0)$  by  $G_m(t, F_n)$ .

◦  $\Pr[T_m^* = 0] = 1 - (1 - \frac{1}{n})^m \approx 1 - e^{-m/n} \approx 0$  as  $n$  gets large  
(or  $\frac{m}{n} \rightarrow 0$ ).

## Bertrand, Duflo, Mullamathan

◦ show that bootstrap works better than robust standard errors.

$$\circ T_n = \frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \quad ; \quad T_n^* = \frac{\hat{\beta}_j^* - \hat{\beta}_j}{\text{se}^*(\hat{\beta}_j)}$$

• use quantiles of simulated distribution of  $T_n^*$  for confidence intervals,

---

---