Subsampling vs Bootstrap

Dimitris N. Politis, Joseph P. Romano, Michael Wolf

$$R_n\left(x_n, \theta\left(P\right)\right) = \tau_n\left(\hat{\theta}_n - \theta\left(P\right)\right)$$

Example:

$$\hat{\theta}_n = \bar{X}_n, \tau_n = \sqrt{n}, \theta = EX = \mu\left(P\right)$$
$$\text{or}\hat{\theta} = \min X_n, \tau_n = n, \theta\left(P\right) = sup\{x : F\left(x\right) \leq 0\}$$

Define: $J_n\left(P\right)$, the distribution of $\tau_n\left(\hat{\theta}_n - \theta\left(P\right)\right)$ under P. For real $\hat{\theta}_n$,

$$J_n\left(x, P\right) \equiv Prob_P\left(\tau_n\left(\hat{\theta}_n - \theta\left(P\right)\right) \leq x\right)$$

Since P is unknown, $\theta\left(P\right)$ is unknown, and $J_n\left(x, P\right)$ is also unknown. The bootrap estimate $J_n\left(x, P\right)$ by $J_n\left(x, \hat{P}_n\right)$, where $\hat{P}_n$ is a consistent estimate of P in some sense. For example, take $hatP_n\left(x\right) = \frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \leq x\right)$ the empirical distribution:

$$sup_x\left|\hat{P}_n\left(x\right) - P\left(x\right)\right| \xrightarrow{a.s.} 0$$

Similarly estimate $(1-\alpha)$th quantile of $J_n\left(x, P\right)$ by $J_n\left(x, \hat{P}_n\right)$: i.e. Estimate $J_n^{-1}\left(x, P\right)$ by $J_n^{-1}\left(x, \hat{P}_n\right)$.

Usually $J_n\left(x, \hat{P}_n\right)$ can't be explicitly calculated(although in some simple case it can be), use Monte Carlo approximation:

$$J_n\left(x, \hat{P}_n\right) \approx \frac{1}{B}\sum_{i=1}^{B} 1\left(\tau_n\left(\hat{\theta}_{n,i} - \hat{\theta}_n\right) \leq x\right)$$

for $\hat{\theta}_{n,i} = \hat{\theta}\left(X_{1,i}^*, \ldots, X_{n,i}^*\right)$.

When bootstrap works(the meaning of "works"), for each x,

$$J_n\left(x, \hat{P}_n\right) - J_n\left(x, P\right) \xrightarrow{p} 0$$
$$\Longrightarrow J_n^{-1}\left(1-\alpha, \hat{P}_n\right) - J_n^{-1}\left(1-\alpha, P\right) \xrightarrow{p} 0$$

When should Bootstrap "work"? Need local uniformity in weak convergence:

1. Usually $J_n(x,P) \longrightarrow J(x,P)$.
2. Also usually $\hat{P}_n \to P$ a.s. in some sense, say $\sup_x \left| \hat{P}_n(x) - P(x) \right| \xrightarrow{a.s.} 0$.
3. Suppose for each sequence P_n s.t. $P_n \to P$, say $\sup_x \left| P_n - P \right| \to 0$, it is also true that $J_n(x,P_n) \longrightarrow J(x,P)$, then it must be true that a.s. $J_n\left(x,\hat{P}_n\right) \longrightarrow J(x,P)$
4. So it ends up having to show for $P_n \to P$, $J_n(x,P_n) \to J(x,P)$, use triangular array formulation.

Case when it works: sample mean with finite variance. It is known that:

1. $\sup_x \left| \hat{F}_n(x) - F(x) \right| \xrightarrow{a.s.} 0$.
2. $\theta\left(\hat{F}_n\right) = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{a.s.} \theta(F) = EX$.
3. $\sigma^2\left(\hat{F}_n\right) = \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2 \xrightarrow{a.s.} \sigma^2(F) = VarX$.
4. Use Linderberg-Feller for the triangular array, applied to the deterministic sequence of P_n such that: 1) $\sup_x \left| P_n(x) - P_n(x) \right| \to 0$; 2) $\theta(P_n) \to \theta(P)$; 3) $\sigma^2(P_n) \to \sigma^2(P)$, it can be shown that $\sqrt{n}\left(\bar{X}_n - \theta(P_n)\right) \xrightarrow{d} N(0,\sigma^2)$ under P_n.
5. Since $\hat{P}_n$ satisfies 1,2,3 a.s., therefore a.s. $J_n\left(x,\hat{P}_n\right) \to J(x,P)$.

Therefore "local uniformity" of weak convergence is satisfied here.

Cases when bootstrap fails:

1. Order Statistics: $F \sim U(0,\theta)$, and $X_{(1)},\ldots,X_{(n)}$ is the order statistics of the sample, so $X_{(n)}$ is the maximum:

$$P\left(n\frac{\theta - X_{(n)}}{\theta} > x\right) = P\left(X_{(n)} < \theta - \frac{\theta x}{n}\right) = P\left(X_i < \theta - \frac{\theta x}{n}\right)^n = \left(\frac{1}{\theta}\left(\theta - \frac{\theta x}{n}\right)\right)^n$$
$$= \left(1 - \frac{x}{n}\right)^n \xrightarrow{n\to\infty} e^{-x}$$

   The bootstrap version:

$$P\left(n\left(X_{(n)} - X^*_{(n)}\right)/X_{(n)} = 0\right) = \left(1 - \left(1 - \frac{1}{n}\right)^n\right) \xrightarrow{n\to\infty} \left(1 - e^{-1}\right) \approx 0.63$$

2. Degenerate U-statistics: Take $w(x,y) = xy$, $\theta(F) = \int\int w(x,y)\, dF(x)\, dF(y) = \mu(F)^2$.

$$\hat{\theta}_n = \theta\left(\hat{F}_n\right) = \frac{1}{n(n-1)} \sum\sum_{i\neq j} X_i X_j$$
$$S(x) = \int xy dF(y) = x\mu(F)$$

If $\mu(F) \neq 0$ it is known that

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N(0, 4Var(S(X))) = N\left(0, 4\left(\mu^2 EX^2 - \mu^4\right)\right)$$

The bootstrap works.

But if $\mu(F) = 0 \Longrightarrow \theta(F) = 0$:

$$\theta\left(\hat{F}_n\right) = \frac{1}{n(n-1)} \sum\sum_{i\neq j} X_i X_j = \bar{X}_n^2 - \frac{1}{n}\frac{1}{n-1}\sum_i \left(X_i - \bar{X}_n\right)^2 = \bar{X}_n^2 - \frac{S_n^2}{n}$$

$$n\left(\theta\left(\hat{F}_n\right) - \theta(F)\right) = n\bar{X}_n^2 - S_n^2 \xrightarrow{d} N\left(0, \sigma^2\right) - \sigma^2$$

However the bootstrap version of $n\left[\theta\left(\hat{F}_n^*\right) - \theta\left(\hat{F}_n\right)\right]$:

$$n\left(\left[\bar{X}_n^{*2} - \frac{1}{n}S_n^{*2}\right] - \left[\bar{X}_n^2 - \frac{1}{n}S_n^2\right]\right) = n\bar{X}_n^{*2} - S_n^{*2} - n\bar{X}_n^2 + S_n^2 \approx n\left(\bar{X}_n^{*2} - \bar{X}_n^2\right)$$
$$= \left[\sqrt{n}\left(\bar{X}_n^* - \bar{X}_n\right)\right]^2 + 2\sqrt{n}\left(\bar{X}_n^* - \bar{X}_n\right)\sqrt{n}\bar{X}_n \xrightarrow{d} N\left(0, \sigma^2\right)^2 + 2N\left(0, \sigma^2\right)\sqrt{n}\bar{X}_n$$

Subsampling: iid case: Y_i block of size b from $(X_1, \ldots, X_n)$, $i = 1, \ldots, q$, for $q = \binom{n}{b}$. Let $\hat{\theta}_{n,b,i} = \hat{\theta}(Y_i)$ calculated with the ith block of data. Use the empirical distribution of $\tau_b\left(\hat{\theta}_{n,b,i} - \hat{\theta}\right)$ over the q pseudo-estimates to approximate the distribution of $\tau_n\left(\hat{\theta} - \theta\right)$: Approximate

$$J_n(x, P) = P\left(\tau_n\left(\hat{\theta}_n - \theta\right) \leq x\right)$$

by

$$L_{n,b}(x) = q^{-1}\sum_{i=1}^{q} 1\left(\tau_b\left(\hat{\theta}_{n,b,i} - \hat{\theta}_n\right) \leq x\right)$$

Claim: If $b \to \infty$, $b/n \to 0$, $\tau_b/\tau_n \to 0$, as long as $\tau_n\left(\hat{\theta}-\theta\right) \xrightarrow{d}$ something,

$$J_n\left(x,P\right)-L_{n,b}\left(x\right)\xrightarrow{p}0$$

Different motivation for Subsampling vs. Bootstrap:

Subsampling: each subset of size b comes from the TRUE model. Since $\tau_n\left(\hat{\theta}_n-\theta\right) \xrightarrow{d} J\left(x,P\right)$, so as long as $b\to\infty$:

$$\tau_b\left(\hat{\theta}_b-\theta\right)\xrightarrow{d}J\left(x,P\right)$$

For n large, the distributions of $\tau_n\left(\hat{\theta}_n-\theta\right)$ and $\tau_b\left(\hat{\theta}_b-\theta\right)$ should be close.
But

$$\tau_b\left(\hat{\theta}_b-\theta\right)=\tau_b\left(\hat{\theta}_b-\hat{\theta}_n\right)+\tau_b\left(\hat{\theta}_n-\theta\right)$$

Since

$$\tau_b\left(\hat{\theta}_n-\theta\right)=O_p\left(\frac{\tau_b}{\tau_n}\right)=o_p\left(1\right)$$

The distributions of $\tau_b\left(\hat{\theta}_b-\theta\right)$ and $\tau_b\left(\hat{\theta}_b-\hat{\theta}_n\right)$ should be close. The distribution of $\tau_b\left(\hat{\theta}_b-\hat{\theta}_n\right)$ is estimated by the empirical distribution over $q=\binom{n}{b}$ pseudo-estimates.

Bootstrap: Recalculate the statistics from the ESTIMATED model $\hat{P}_n$. Given that $\hat{P}_n$ is close to P, hopefully $J_n\left(x,\hat{P}_n\right)$ is close to $J_n\left(x,P\right)$(Or to $J\left(x,P\right)$, the limit distribution). But when bootstrap fails

$$\hat{P}_n\longrightarrow P\nRightarrow J_n\left(x,\hat{P}_n\right)\longrightarrow J\left(x,P\right)$$

Formal Proof of consistency of subsampling:

Assumptions: $\tau_n\left(\hat{\theta}_n-\theta\right)\xrightarrow{d}J\left(x,P\right)$,$b\to\infty$, $\frac{b}{n}\to 0$, $\frac{\tau_b}{\tau_n}\longrightarrow 0$.

Need to show: $L_{n,b}\left(x\right)-J\left(x,P\right)\xrightarrow{p}0$.

Since $\tau\left(\theta_n-\theta\right)\xrightarrow{p}0$, it is enough to show

$$U_{n,b}\left(x\right)=q^{-1}\sum_{i=1}^{q}1\left(\tau_b\left(\hat{\theta}_{n,b,i}-\theta\right)\le x\right)\xrightarrow{p}J\left(x,P\right)$$

$$U_{n,b}\left(x\right)-J\left(x,P\right)=U_{n,b}\left(x\right)-EU_{n,b}\left(x\right)+EU_{n,b}\left(x\right)-J\left(x,P\right)$$

Enough to show

$$U_{n,b}(x) - EU_{n,b}(x) \xrightarrow{p} 0$$

and

$$EU_{n,b}(x) - J(x,P) \to 0$$

But

$$EU_{n,b}(x) - J(x,P) = J_b(x,P) \to 0$$

$U_{n,b}(x)$ is a bth order U-statistics with kernel function bounded by $(-1,1)$. Use Hoeffding exponential-type inequality(Serfling(1980), Thm A. p201):

$$P\left(U_{n,b}(x) - J_b(x,P) \geq \epsilon\right) \leq \exp\left(-2\frac{n}{b}\epsilon^2 / \left[1-(-1)\right]\right) = \exp\left(-\frac{n}{b}t^2\right) \longrightarrow 0$$

as $\frac{n}{b} \longrightarrow \infty$.
So

$$L_{n,b}(x) - J(x,P) = L_{n,b}(x) - U_{n,b}(x) + U_{n,b}(x) - J_b(x,P) + J_b(x,P) - J(x,P) \xrightarrow{p} 0.$$

Q.E.D.
Time Series(!): Respect the ordering of the data to preserve correlation.

$$\hat{\theta}_{n,b,t} = \hat{\theta}_b\left(X_t, \ldots, X_{t+b-1}\right), \quad q = T - b + 1.$$

$$L_{n,b}(x) = \frac{1}{q}\sum_{i=1}^{q} 1\left(\tau_b\left(\hat{\theta}_{n,b,t} - \hat{\theta}_n\right) \leq x\right)$$

Assumption: $\tau_n\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} J(x,P)$, $b \to \infty$, $\frac{b}{n} \to 0$, $\frac{\tau_b}{\tau_n} \to 0$, $\alpha(m) \to 0$.
Result: $L_{n,b}(x) - J(x,P) \xrightarrow{p} 0$.
Most difficult part: To show $\tau_n\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} J(x,P)$.

Can treat iid data as time series, or even using non-overlapping blocks $k = \left[\frac{n}{b}\right]$, but using $\binom{n}{b}$ more efficient. For example, if $\bar{U}_n(x) = k^{-1}\sum_{j=1}^{k} 1\left(\tau_b\left[R_{n,b,j} - \theta(P)\right] \leq x\right)$ then

$$U_{n,b}(x) = E\left[\bar{U}_n(x) | \mathcal{X}_\backslash\right] = E\left[1\left(\tau_b\left[R_{n,b,j} - \theta(P)\right] \leq x\right) | \mathcal{X}_n\right]$$

for $\mathcal{X}_n = \left(X_{(1)}, \ldots, X_{(n)}\right)$. $U_{n,b}(x)$ is better than $\bar{U}_n(x)$ since $\mathcal{X}_n$ is sufficient statistics for iid data.

Hypothesis Testing: $T_n = \tau_n t_n (X_1, \ldots, X_n)$,

$$G_n(x, P) = Prob_p(\tau_n \leq x) \xrightarrow{P \in P_0} J(x, P)$$

$$\hat{G}_{n,b}(x) = q^{-1} \sum_{i=1}^{q} 1(T_{n,b,i} \leq x) = q^{-1} \sum_{i=1}^{q} 1(\tau_b t_{n,b,i} \leq x)$$

As long as $b \to \infty$, $\frac{b}{n} \to 0$, then under $P \in P_0$:

$$\hat{G}_{n,b}(x) \longrightarrow G(x, P)$$

If under $P \in P_1$, $T_n \to \infty$, then $\forall x$, $\hat{G}_{n,b}(x) \to 0$.

Key difference with confidence interval: don't need $\frac{\tau_b}{\tau_n} \to 0$, because don't need to estimate θ_0 but assumed known under the null hypothesis.

Estimating the unknown rate of convergence: Assume that $\tau_n = n^\beta$, for some $\beta > 0$, but β is unknown. Estimate β using different size of subsampling distribution. Key idea: Compare the shape of the empirical distributions of $\hat{\theta}_b - \hat{\theta}_n$ for different values of b to infer the value of β.
Let $q = \binom{n}{b}$ for iid data, or $q = (T - b + 1)$ for time series data:

$$L_{n,b}(x|\tau_b) \equiv q^{-1} \sum_{a=1}^{q} 1\left(\tau_b \left(\hat{\theta}_{n,b,a} - \hat{\theta}_n\right) \leq x\right)$$

$$L_{n,b}(x|1) \equiv q^{-1} \sum_{a=1}^{q} 1\left(\hat{\theta}_{n,b,a} - \hat{\theta}_n \leq x\right)$$

This implies

$$L_{n,b}(x|\tau_b) = L_{n,b}\left(\tau_b^{-1} x|1\right) \equiv t$$

$$x = L_{n,b}^{-1}(t|\tau_b) = \tau_b\left(\tau_b^{-1} x\right) = \tau_b L_{n,b}^{-1}(t|1)$$

Since $L_{n,b}(x|\tau_b) \xrightarrow{p} J(x, P)$, if $J(x, P)$ is continuous and increasing, it can be infered that

$$L_{n,b}^{-1}(t|\tau_b) = J^{-1}(t, P) + o_p(1)$$

Same as

$$\tau_b L_{n,b}^{-1}(t|1) = J^{-1}(t, P) + o_p(1)$$

So

$$b^{\beta} L_{n,b}^{-1}(t|1) = J^{-1}(t, P) + o_p(1)$$

take log(Assuming $J^{-1}(t, P) > 0$, or $t > J(0, P)$), for different b_1 and b_2, then this becomes

$$\beta \log b_1 + \log\left(L_{n,b_1}^{-1}(t|1)\right) = \log J^{-1}(t, P) + o_p(1)$$
$$\beta \log b_2 + \log\left(L_{n,b_2}^{-1}(t|1)\right) = \log J^{-1}(t, P) + o_p(1)$$

Different out the "fixed effect"

$$\beta(\log b_1 - \log b_2) = \log\left(L_{n,b_2}^{-1}(t|1)\right) - \log\left(L_{n,b_1}^{-1}(t|1)\right) + o_p(1)$$

So estimate β by

$$\hat{\beta} = (\log b_1 - \log b_2)^{-1}\left(\log\left(L_{n,b_2}^{-1}(t|1)\right) - \log\left(L_{n,b_1}^{-1}(t|1)\right)\right) = \beta + (\log b_1 - \log b_2)^{-1} \times o_p(1)$$

Take $b_1 = n^{\gamma_1}$, $b_2 = n^{\gamma_2}$, $(1 \geq \gamma_1 > \gamma_2 > 0)$

$$\hat{\beta} - \beta = ((\gamma_1 - \gamma_2)\log n)^{-1} o_p(1) = o_p\left((\log n)^{-1}\right)$$

How to know $t > J(0, P)$

$$L_{n,b}(0|\tau_b) = L_{n,b}(0|1) = J(0, P) + o_p(1)$$

So estimating $J(0, P)$ not a problem.

Alternatively, take $t_2 \in (0.5, 1)$, take $t_1 \in (0, 0.5)$

$$b^{\beta}\left(L_{n,b}^{-1}(t_2|1) - L_{n,b}^{-1}(t_1|1)\right) = J^{-1}(t_2|P) - J^{-1}(t_1|P) + o_p(1)$$

$$\beta \log b + \log\left(L_{n,b}^{-1}(t_2|1) - L_{n,b}^{-1}(t_1|1)\right) = \log\left(J^{-1}(t_2|P) - J^{-1}(t_1|P)\right) + o_p(1)$$

$$\hat{\beta} = (\log b_1 - \log b_2)^{-1}\left[\log\left(L_{n,b_2}^{-1}(t_2|1) - L_{n,b_2}^{-1}(t_1|1)\right) - \log\left(L_{n,b_1}^{-1}(t_2|1) - L_{n,b_1}^{-1}(t_1|1)\right)\right]$$

Take $b_1 = n^{\gamma_1}$, $b_2 = n^{\gamma_2}$, $(1 > \gamma_1 > \gamma_2 > 0)$, As before

$$\hat{\beta} - \beta = o_p\left((\log n)^{-1}\right)$$

Two step subsampling: $\hat{\tau}_n = n^{\hat{\beta}}$

$$L_{n,b}\left(x|\hat{\tau}_b\right) = q^{-1}\sum_{a=1}^{q} 1\left(\hat{\tau}_b\left(\hat{\theta}_{n,b,a} - \hat{\theta}_n\right) \leq x\right)$$

Can show that

$$\sup_x \left| L_{n,b}\left(x|\hat{\tau}_b\right) - J\left(x,P\right) \right| \xrightarrow{p} 0.$$

Problem: imprecise in small samples. E.g. in variation estimation, best choice of b gives error rate of $O\left(n^{-1/3}\right)$ but parameter estimates, if model is true, gives $O\left(n^{-1/2}\right)$ error rate. Bootstrap pivotal statistics, when applicable, gives even better than $O\left(n^{-1/2}\right)$ error rate.