

Bootstrap: Theory and Applications

J.-P. Stockis, Technische Universität Kaiserslautern

References:

- J. Shao, D. Tu (1995): "The Jackknife and Bootstrap", Springer
- B. Efron (1982): "The Jackknife, the Bootstrap and other Resampling Plans", CBMS 38
- P. Hall (1992): "The Bootstrap and Edgeworth Expansion", Springer-Verlag
- B. Efron, R.J. Tibshirani (1993): "An Introduction to the Bootstrap", Chapman and Hall
- A.C. Davison, D.V. Hinkley (1997): "Bootstrap Methods and their Application", Cambridge University Press

Plan of the lectures

1. Introduction to the Jackknife and Introduction to the Bootstrap (I)
2. Introduction to the Bootstrap (II)
3. Techniques for proving consistency (i.i.d. sample)
4. Asymptotic comparison (i.i.d. sample)
5. Bootstrap confidence sets and hypothesis tests
6. Application to linear models
7. Application to nonparametric models
8. Application to financial time series

1. Introduction to the Jackknife and Introduction to the Bootstrap (I)

First, let us have a look at well-known results. We have X_1, \dots, X_n i.i.d. We would like to estimate the mean μ (assumed to exist) of the distribution.

By the strong law of large numbers, the sample mean $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$ and if $\text{var}(X_1) < \infty$, then the central limit theorem provides us with an asymptotic variance of \bar{X}_n and with its asymptotic distribution.

Now, let us suppose we are interested in estimating a median m . The sample median \dot{X}_n defined as $X_{(\frac{n+1}{2})}$ if n is odd ($X_{(\frac{n+1}{2})}$: $\frac{n+1}{2}$ th smallest value of the data) and $\frac{1}{2}(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)})$ if n is even also admits a central limit theorem. If X_1 has a density f , if $f(m) > 0$ and f is continuous around m , then

$$2f(m)(\dot{X}_n - m) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Even for as simple statistics as the sample mean and the sample median only their asymptotic distribution is known. For more complicated statistics, the asymptotic theory becomes more difficult (in terms of assumptions, ...)

Would it be possible in some cases to get better results for the characteristics of an estimator based on X_1, \dots, X_n than the results based on the asymptotic theory and this without too strong assumptions?

Resampling techniques may help to give a positive answer to the question.

First, we have a quick look at the jackknife (leave one out at a time).

We have an estimator $T_n(X_1, \dots, X_n)$ of a "parameter" θ and we are interested in its bias.

The jackknife bias estimator $b_{\text{JACK}} = (n-1)(\frac{1}{n} \sum_{i=1}^n T_{n-1,-i} - T_n)$, where $T_{n-1,-i} = T_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$.

If we denote $\frac{1}{n} \sum_{i=1}^n T_{n-1,-i}$ by \bar{T}_n , we may rewrite $b_{\text{JACK}} = (n-1)(\bar{T}_n - T_n)$ and a better (with smaller bias) estimator is given by $T_{\text{JACK}} = T_n - b_{\text{JACK}} = nT_n - (n-1)\bar{T}_n$.

Why does b_{JACK} make sense?

Here a heuristic justification: if the bias vanishes at $n \rightarrow \infty$ and is linear in $\frac{1}{n}$ then $\frac{E(T_n) - \theta}{E(T_{n-1}) - E(T_n)} = \frac{1/n}{1/n - 1/n}$ and so,

$$\text{Bias} = E(T_n) - \theta = (n-1)(E(T_{n-1}) - E(T_n))$$

and we can replace $E(T_{n-1})$ and $E(T_n)$ by unbiased estimators \bar{T}_n and T_n .

Some examples:

- $T_n = \bar{X}_n$; $b_{\text{JACK}} = 0$ o.k.
- $T_n = \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ (Bias = $-\frac{1}{n-1} \sigma^2$); $b_{\text{JACK}} = -\frac{1}{n(n-1)} \sum_{i=1}^n (X_i - \bar{X}_n)^2$
 $\rightarrow T_{\text{JACK}} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ unbiased estimator of the variance.

Note that the mean squared error (MSE, equal to Bias² + Var) is greater for this unbiased estimator than for $\hat{\sigma}_n^2$ in the case X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma^2)$, so T_{JACK} is not necessarily "better" than T_n .

Now, how can jackknife help in the estimation of the variance of a statistic.

$$\text{var}_{\text{JACK}} = \frac{n-1}{n} \sum_{i=1}^n (T_{n-1,-i} - \bar{T}_n)^2$$

Why does this make sense? If we look at \bar{X}_n , $\text{var}(\bar{X}_n) = \frac{\sigma^2}{n}$ estimated for example by $\frac{1}{n(n-1)} \sum_{i=1}^n (X_i - \bar{X}_n)^2$.

Now, $T_{n-1,-i} = \frac{n\bar{X}_n - X_i}{n-1}$, $\bar{T}_n = \bar{X}_n$ and $T_{n-1,-i} - \bar{T}_n = \frac{\bar{X}_n - X_i}{n-1}$ and thus $\text{var}_{\text{JACK}} = \frac{1}{n(n-1)} \sum_{i=1}^n (X_i - \bar{X}_n)^2$.

The jackknife variance estimator provides us with good results for smooth statistics (e.g. strong consistency $\frac{\text{var}_{\text{JACK}}}{\text{var}(T_n)} \xrightarrow{\text{a.s.}} 1$; see for example chapter 2 of Shao and Tu).

Now, for unsmooth statistics the jackknife variance estimator is not consistent. To illustrate this consider the sample median and the following 9 data: 10, 27, 31, 40, 46, 50, 52, 104, 146. The sample median is 46. The 9 values for $T_{n-1,-i}$ are 48, 48, 48, 48, 45, 43, 43, 43, 43. We see that small changes in the data set do not necessarily cause changes in the statistic (lack of

smoothness). This will lead in this case to an underestimation of $\text{var}(X_n)$. To deal with this kind of unsmooth statistics, we could use so-called delete-d-jackknife (leave d out at a time) with d suitably chosen. Notice also the existence of a jackknife histogram which can help in approximating the distribution of a statistic.

Now, let us suppose we have i.i.d. data X_1, \dots, X_n from F and a statistic $T_n(X_1, \dots, X_n) = T_n$. We are interested in estimating $\text{var}(T_n)$ ($\text{var}_F(T_n) = \int [T_n(x) - \int T_n(y) d \prod_{i=1}^n F] d \prod_{i=1}^n F(x_i)$ typically cannot be obtained exactly!)

What would be great?

To generate, say 100 samples of length n from F

$$\begin{aligned} X_{1,1}, \dots, X_{n,1} &\rightarrow T_{n,1} := T_n(X_{1,1}, \dots, X_{n,1}) \\ X_{1,100}, \dots, X_{n,100} &\rightarrow T_{n,100}. \end{aligned}$$

Then $\text{var}(T_n)$ could very well be estimated:

$$\text{var}(T_n) \cong \frac{1}{100} \sum_{i=1}^{100} (T_{n,i} - \frac{1}{100} \sum_{j=1}^{100} T_{n,j})^2.$$

Problem: F is unknown to us.

Bootstrap ideas:

1. replace F by \hat{F}_n , the empirical distribution function based on X_1, \dots, X_n (or use any good estimator of F) (substitution principle)
2. Typically no exact result is available for $\text{var}_{\hat{F}_n}(T_n | X_1, \dots, X_n)$ either. But \hat{F}_n is known to us (conditioned on X_1, \dots, X_n): it is a discrete distribution with weight $\frac{1}{n}$ to X_i $i = 1, \dots, n$.

So we can draw, say, B samples from \hat{F}_n (conditioned on X_1, \dots, X_n):

$$\begin{aligned} X_{1,1}^* \dots X_{n,1}^* &\rightarrow T_{n,1}^* = T_n(X_{1,1}^*, \dots, X_{n,1}^*) \\ &\vdots \\ X_{1,B}^* \dots X_{n,B}^* &\rightarrow T_{n,B}^* \end{aligned}$$

→ the bootstrap variance estimator $\text{var}_{\text{BOOT}} = \text{var}_*(T_n^*) = \text{var}(T_n(X_1^* \dots X_n^*) | X_1, \dots, X_n) = \text{var}_{\hat{F}_n}(T_n | X_1 \dots X_n)$ can be evaluated by $\frac{1}{B} \sum_{b=1}^B (T_{n,b}^* - \frac{1}{B} \sum_{j=1}^B T_{n,j}^*)^2$

Note that we will also say bootstrap variance estimator and use the same notations (e.g. $\text{var}_*(T_n^*)$) for the numerical approximation.

Remark: When a "*" appears, this means: conditioned on our data (here conditioned on X_1, \dots, X_n).

2. Introduction to the Bootstrap (II)

First we recall some properties of \hat{F}_n , the empirical distribution function of F . Our framework remains the same: X_1, \dots, X_n are assumed to be i.i.d., taken from F . The theorem of Glivenko-Cantelli tells us that

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0.$$

Let $D_n := \sup_x |\hat{F}_n(x) - F(x)|$. Then, for F 1-dimensional and continuous, the distribution of D_n does not depend on F and has a rate of convergence $O(n^{-1/2})$ to 0. More precisely,

$$\lim_{n \rightarrow \infty} P(n^{1/2} D_n \leq d) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2 d^2}, d > 0$$

(this can be obtained by using theory of stochastic processes). So, the substitution of F by \hat{F}_n appears to make sense.

Now, keeping in mind the two ideas behind bootstrap (substitution and (if necessary) simulations), let us have a look at some simple cases.

The bootstrap bias estimator is given by

$$\text{Bias}_{\text{Boot}} = E_* T(X_1^* \dots X_n^*) - T(X_1, \dots, X_n)$$

since $\text{Bias} = E T(X_1, \dots, X_n) - \theta$ and θ in fact depends on $F(\theta(F))$. If we replace F by \hat{F} , $\theta(\hat{F})$ is the plug-in estimator of θ , often $T(X_1, \dots, X_n)$. Even if $T(X_1, \dots, X_n)$ is not the plug-in estimator it typically does not make a significant difference to replace $\theta(\hat{F})$ by $T(X_1, \dots, X_n)$. Then, we replace $E_* T(X_1^*, \dots, X_n^*)$ by $\frac{1}{B} \sum_{b=1}^B T_1^*(X_{1b}^* \dots X_{nb}^*)$.

Notice that $\bar{X}_n, \frac{1}{B} \sum_{b=1}^B T(X_{1b}^* \dots X_{nb}^*) - \bar{X}_n$ is typically different from 0 although $E_* T(X_1^*, \dots, X_n^*) - \bar{X}_n = 0$.

Now, let us come back to the bootstrap variance estimator. There are a few cases when we can get an explicit form for the bootstrap variance estimator (i.e. we do not need simulations). Here are two examples:

$$\text{var}_*(\bar{X}_n^*) = \frac{1}{n} \text{var}_*(X_1^*) = \frac{1}{n} \cdot \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

As a second example we look at the bootstrap variance estimator of the sample median.

Suppose $n = 2m + 1$,

$$\begin{aligned} p_k &:= P_*(X_{(m)}^* = X_{(k)}) \quad , 1 \leq k \leq n \\ &= P_*(X_{(m)}^* \geq X_{(k)}) - P_*(X_{(k)} < X_{(m)}^*) \\ &= P_*(\#\{i : X_{(i)}^* < X_{(k)}\} \leq m - 1) - P_*(\#\{i : X_{(i)}^* \leq X_{(k)}\} \leq m - 1) \\ &= P(\text{Bin}(n, \frac{k-1}{n}) \leq m - 1) - P(\text{Bin}(n, \frac{k}{n}) \leq m - 1) \end{aligned}$$

Now, $\text{var}_*(\dot{X}_n^*) = \sum_{k=1}^n p_k (X_{(k)} - \sum_{j=1}^n p_{(j)} X_{(j)})^2$.

As a next possible application of the bootstrap we say a few words about the so-called parametric bootstrap: X_1, \dots, X_n i.i.d. with distribution P_θ known up to a finite number of parameters θ . Then, the parametric bootstrap is more suitable than the (nonparametric) bootstrap described earlier. We take a good estimator $\hat{\theta}$ for θ and then we draw our bootstrap samples from $P_{\hat{\theta}}$.

In fact, a main idea of the bootstrap is to mimic as well as possible the stochastic model of the data. Let M be this stochastic model (e.g. the joint distribution of X_1, \dots, X_n). Then, the bootstrap technique will use an estimator of the model, say \hat{M} . Then from \hat{M} , it will generate data X^* and by means of these data the statistics we are interested in.

Now, the bootstrap can also help to estimate the distribution of a statistic T_n : it is often possible to show that

$$\sup_x |P_*(T_n(X_1^*, \dots, X_n^*) \leq x) - P(T_n(X_1, \dots, X_n) \leq x)| \xrightarrow{p} 0$$

or even it tends almost surely to 0 (more in Chapter 3).

Note the necessity of a stochastic convergence since the bootstrap distribution is conditioned on X_1, \dots, X_n .

Interestingly, bootstrap approximation provides us sometimes with a better rate of convergence than the normal approximation (more in Chapter 4).

Sometimes (even for X_1, \dots, X_n i.i.d.) the bootstrap technique needs to be modified.

One reason: \hat{F}_n is not a good estimator of F in the extreme tails. Example: X_1, \dots, X_n i.i.d. with abs. continuous distribution function F s.t. $F(\theta) = 1$ for some θ , $F(x) < 1$ for $x < \theta$. Suppose we would like to approximate $X_{(n)} - \theta$ by $X_{(n)}^* - X_{(n)}$ $P(X_{(n)} = \theta) = 0$ but $P_*(X_{(n)}^* = X_{(n)}) = 1 - P_*(X_i^* \neq X_{(n)} \forall i) = 1 - (1 - \frac{1}{n})^n \rightarrow 1 - e^{-1} \neq 0$. A possible solution is to take the length of the bootstrap sample equal to $m < n$ with $\frac{m}{n} \rightarrow 0$. Then

$$P_*(X_{(m)}^* = X_{(n)}) = 1 - (1 - \frac{1}{n})^m \rightarrow 0.$$

Obviously, since the distribution of T_n can be approximated by bootstrapping, bootstrap can also be useful for constructing confidence intervals (more in Chapter 5).

In the case of hypothesis test, we can use the bootstrap to approximate the required distribution under H_0 . But we have to take care since we need our bootstrap data from a distribution under the restrictions specified by H_0 , which excludes the empirical distribution of the original data in most cases (more in Chapter 5).

Now, in the regression case (parametric or nonparametric), the model M does not reduce anymore to the distribution function F . For example, if we have $Y = Xb + \varepsilon$, $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. $(0, \sigma_\varepsilon^2)$, X fixed design matrix, b vector of unknown parameters, M can be summarized as (β, F_ε) . So our bootstrap can be based on $\hat{M} = (\hat{\beta}, \hat{F}_\varepsilon)$ (more in Chapters 6 and 7.)

Also in time series (i.e. X_1, \dots, X_n typically not independent), the distribution function of X_1 is typically not enough to summarize our process. So bootstrapping in the same way as if X_1, \dots, X_n were i.i.d. leads to wrong results (e.g. $\text{var}_*(\bar{X}_n^*) = \frac{\hat{\sigma}_n^2}{n}$ which is obviously typically wrong).

Now, M can be summarized by the joint distribution of X_1, \dots, X_n . A natural bootstrap would then be to bootstrap not single data but blocks of a certain length to keep some structure of the process (more in Chapter 8).

3. Techniques for proving consistency (i.i.d. sample)

In this chapter we are going to illustrate three techniques for proving consistency: by means of Mallows' distance, imitation and linearization. For sake of simplicity we are going to concentrate on the sample mean or on a function of the sample mean.

We would like to show that

$$\sup_x |P(\sqrt{n}(\bar{X}_n - \mu) \leq x) - P_*(\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x)| \xrightarrow{\text{a.s.}} 0.$$

Let F, G be distribution functions. We will denote $\sup_x |F(x) - G(x)|$ by $\rho_\infty(F, G)$, the Kolmogorov distance between F and G . This is a distance between distribution functions: If we write $\rho_\infty(X, Y)$ this in fact means $\rho_\infty(F_X, F_Y)$. In particular $\rho_\infty(aX, aY)$ is equal to $\rho_\infty(X, Y)$ for $a \neq 0$.

Some more properties of ρ_∞ :

let Y_1, \dots, Y_n, Y rvs with distribution functions F_{Y_i}, F_Y $\rho_\infty(F_{Y_n}, F_Y) \downarrow 0$, then $Y_n \xrightarrow{\mathcal{L}} Y$.

The reverse implication does not hold but if $Y_n \xrightarrow{\mathcal{L}} Y$ and F_Y is continuous then $\rho_\infty(F_{Y_n}, F_Y) \downarrow 0$.

Suppose that $Y_n \xrightarrow{\mathcal{L}} Y$. Then, $\lim_{n \rightarrow \infty} E|Y_n|^r = E|Y|^r < \infty$ if and only if $\{|Y_n|^r\}$ is uniformly integrable ($r > 0$).

Now let us introduce another metric which sometimes helps for proving consistency.

Definition: Let \mathbb{R}^s be endorsed with a metric $\|\cdot\|$, let $\mathcal{F}_{r,s}$ be the set of all probability measures defined on the Borel σ -algebra of \mathbb{R}^s with finite r -th moment. Let $G, H \in \mathcal{F}_{r,s}$ (more precisely G and H are distribution functions with corresponding probability measures in $\mathcal{F}_{r,s}$). Then the Mallows' distance $\tilde{\rho}_r(G, H) := \inf_{\mathcal{T}_{X,Y}} (E\|X - Y\|^r)^{1/r}$, where $\mathcal{T}_{X,Y}$ is the collection of all possible joint distributions of the pairs (X, Y) whose marginal distributions are G and H respectively.

For the case $s = 1, r = 2$ and $\|\cdot\|$ the Euclidean norm, it can be shown that

$$\tilde{\rho}_2(G, H) = \left(\int_0^1 |G^{-1}(t) - H^{-1}(t)|^2 dt \right)^{1/2}.$$

In the following we concentrate on $r\tilde{h}o_2$.

Here some properties of $\tilde{\rho}_2$ (more in Bickel, Freedman (1981) Annals of Statistics):

Proposition: let $\alpha_n, \alpha \in \mathcal{F}_{2,s}$. Then $r\tilde{h}o_2(\alpha_n, \alpha) \downarrow 0$ as $n \rightarrow \infty$ if and only if $\alpha_n \xrightarrow{\mathcal{L}} \alpha$ and

$$\int \|x\|^2 \alpha_n(dx) \rightarrow \int \|x\|^2 \alpha(dx).$$

We only prove \rightarrow : let ξ_n have law α_n and ξ have law α and $(E\|\xi_n - \xi\|^2)^{1/2} = r\tilde{h}o_2(\alpha_n, \alpha)$.

Then,

$$\begin{aligned} & \left[\int \|x\|^2 \alpha_n(dx) \right]^{1/2} - \left[\int \|x\|^2 \alpha(dx) \right]^{1/2} \\ &= (E\|\xi_n\|^2)^{1/2} - (E\|\xi\|^2)^{1/2} \leq (E\|\xi_n - \xi\|^2)^{1/2} \downarrow 0. \end{aligned}$$

Likewise, let f be any Lipschitz function, that is $|f(x) - f(y)| \leq K\|x - y\|$, then $|\int f(x)\alpha_n(dx) - \int f(x)\alpha(dx)| = |E(f(\xi_n) - f(\xi))| \leq E|f(\xi_n) - f(\xi)| \leq K E\|\xi_n - \xi\| \leq K(E\|\xi_n - \xi\|^2)^{1/2} \downarrow 0$.

This can be used to prove that for any continuous bounded function f ,

$$\left| \int f(x)\alpha_n(dx) - \int f(x)\alpha(dx) \right| \downarrow 0.$$

Some more properties:

let \hat{F}_n be the empirical distribution function of F , then $\tilde{\rho}_2(\hat{F}_n, F) \xrightarrow{\text{a.s.}} 0$;

let $U, V \in \mathcal{F}_{2,s}$, then $\tilde{\rho}_2(aU, aV) = |a|\tilde{\rho}_2(U, V)$;
 $(\tilde{\rho}_2(U, V))^2 = (\tilde{\rho}_2(U - EU, V - EV))^2 + \|EU - EV\|^2$.

Let $\{U_j\}$ and $\{V_j\}$ be two sequences of independent random vectors whose distributions are in $\mathcal{F}_{2,s}$ and $EU_j = EV_j \forall j$ then

$$(\tilde{\rho}_2(\sum_{j=1}^n U_j, \sum_{j=1}^n V_j))^2 \leq \sum_{j=1}^n (\tilde{\rho}_2(U_j, V_j))^2.$$

Let $\{U_j\}$ and $\{V_j\}$ be two sequences of i.i.d. random variables, whose distributions are in $\mathcal{F}_{2,1}$; let $U = (U_1, \dots, U_n)^T, V = (V_1, \dots, V_n)^T$, let A be a $m \times n$ matrix of scalars, then $(\tilde{\rho}_2(AU, AV))^2 \leq \text{trace}(AA^T)(\tilde{\rho}_2(U_i, V_i))^2$

Now, let X_1, \dots, X_n be i.i.d. random variables from F with mean μ and finite variance, X_1^*, \dots, X_n^* i.i.d. from \hat{F}_n

$$\begin{aligned} & \tilde{\rho}_2(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \sqrt{n}(\bar{X}_n - \mu)) \\ &= \frac{1}{\sqrt{n}} \tilde{\rho}_2(\sum_{i=1}^n (X_i^* - \bar{X}_n), \sum_{i=1}^n (X_i - \mu)) \\ &\leq \left(\frac{1}{n} \sum_{i=1}^n (\tilde{\rho}_2(X_i^* - \bar{X}_n, X_i - \mu))^2\right)^{\frac{1}{2}} \\ &= \tilde{\rho}_2(X_1^* - \bar{X}_n, X_1 - \mu) \\ &= ((\tilde{\rho}_2(X_1^*, X_1))^2 - |EX_1 - E^*X_1^*|^2)^{\frac{1}{2}} \\ &= ((\tilde{\rho}_2(\hat{F}_n, F))^2 - |\mu - \bar{X}_n|^2)^{\frac{1}{2}} = o(1) \text{ a.s.} \end{aligned}$$

Proof based on Mallows' distance; we get then convergence in ρ_∞ since

$$\begin{aligned} & \rho_\infty(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \sqrt{n}(\bar{X}_n - \mu)) \\ &\leq \rho_\infty(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \Phi_{0,\sigma^2}) + \rho_\infty(\Phi_{0,\sigma^2}, \sqrt{n}(\bar{X}_n - \mu)). \end{aligned}$$

Now, $\rho_\infty(\Phi_{0,\sigma^2}, \sqrt{n}(\bar{X}_n - \mu)) \downarrow 0$ and $\rho_\infty(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \Phi_{0,\sigma^2})$ too since $\rho_\infty(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \Phi_{0,\sigma^2}) \leq \tilde{\rho}_2(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \sqrt{n}(\bar{X}_n - \mu)) + \tilde{\rho}_2(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \Phi_{0,\sigma^2}) \downarrow 0$.

The most often method used is the so-called imitation method. In our example, it is well known that $\sqrt{n}(\bar{X}_n - \mu)$ is asympt. $\mathcal{N}(0, \sigma^2)$. It can be proved e.g. that the Lindeberg-Feller condition is satisfied.

Now, we shall prove that $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ is also a.s. asympt. $\mathcal{N}(0, \sigma^2)$ since the Lindeberg-Feller condition is also satisfied in this case. And thus $\rho_\infty(\sqrt{n}(\bar{X}_n - \mu), \sqrt{n}(\bar{X}_n^* - \bar{X}_n)) \leq$

$$\rho_\infty(\sqrt{n}(\bar{X}_n - \mu), \Phi_{0,\sigma^2}) + \rho_\infty(\Phi_{0,\sigma^2}, \sqrt{n}(\bar{X}_n^* - \bar{X}_n)) \xrightarrow{\text{a.s.}} 0.$$

To prove $\rho_\infty(\Phi_{0,\sigma^2}, \sqrt{n}(\bar{X}_n^* - \bar{X}_n)) \xrightarrow{\text{a.s.}} 0$, it is enough to show that $\frac{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)}{\hat{\sigma}_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$ a.s.

since $\hat{\sigma}_n^2 \xrightarrow{\text{a.s.}} \sigma^2$.

We have a Lindeberg-Feller CLT in the bootstrap case if

$$\begin{aligned}
& \hat{\sigma}_n^{-2} E_*(X_1^* - \bar{X}_n) I(|X_1^* - \bar{X}_n| \geq \varepsilon n^{\frac{1}{2}} \hat{\sigma}_n) \xrightarrow{\text{a.s.}} 0 \quad \forall \varepsilon > 0 \\
& = \frac{1}{\hat{\sigma}_n^2} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 I(|X_i - \bar{X}_n| \geq \varepsilon n^{\frac{1}{2}} \hat{\sigma}_n) \\
& \leq \frac{1}{\hat{\sigma}_n^2} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 I(|X_i| \geq \varepsilon n^{\frac{1}{2}} \hat{\sigma}_n - |\bar{X}_n|) \\
& \simeq \frac{1}{\hat{\sigma}_n^2} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 I(|X_i| \geq \varepsilon n^{\frac{1}{2}} \sigma).
\end{aligned}$$

It remains to show that $\sum_{i=1}^n (X_i - \bar{X}_n)^2 I(|X_i| \geq \varepsilon n^{\frac{1}{2}} \sigma) = o(n)$.

Now, $P(|X_i| \geq \varepsilon n^{\frac{1}{2}} \sigma) \leq \frac{E X_1^2}{\varepsilon^2 n \sigma^2}$ (Tschebyscheff) and thus $\sum_{i=1}^n P(|X_i| \geq \varepsilon n^{\frac{1}{2}} \sigma) \leq \frac{E X_1^2}{\varepsilon^2 \sigma^2} \forall n$ which leads to the conclusion by the Borel-Cantelli lemma.

Third possible way of proving consistency: the so-called linearization: As previously, let X_1, \dots, X_n be i.i.d. random variables having finite variance σ^2 .

We know from the previous techniques that $\rho_\infty(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \sqrt{n}(\bar{X}_n - \mu)) \xrightarrow{\text{a.s.}} 0$. (In particular,

$$\rho_\infty(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \Phi_{0, \sigma^2}) \xrightarrow{\text{a.s.}} 0)$$

Now, we would like to show that

$$\rho_\infty(\sqrt{n}(g(\bar{X}_n^*) - g(\bar{X}_n)), \sqrt{n}(g(\bar{X}_n) - g(\mu))) \xrightarrow{\text{p}} 0,$$

with $g : \mathbb{R} \rightarrow \mathbb{R}$, differentiable at μ .

We have

$$\begin{aligned}
g(\bar{X}_n) - g(\mu) &= \frac{\partial g(t)}{\partial t} \Big|_{\mu} (\bar{X}_n - \mu) + o_p(n^{-\frac{1}{2}}) \\
g(\bar{X}_n^*) - g(\mu) &= \frac{\partial g(t)}{\partial t} \Big|_{\mu} (\bar{X}_n^* - \mu) + o_p^*(n^{-\frac{1}{2}})
\end{aligned}$$

and thus $g(\bar{X}_n^*) - g(\bar{X}_n) = \frac{\partial g(t)}{\partial t} \Big|_{\mu} (\bar{X}_n^* - \bar{X}_n) + o_p^*(n^{-\frac{1}{2}}) + o_p(n^{-\frac{1}{2}})$

$\longrightarrow P_*(\sqrt{n}(g(\bar{X}_n^*) - g(\bar{X}_n)) \leq x) \xrightarrow{\text{p}} \Phi_{0, \sigma_g^2}$, with $\sigma_g^2 = (\frac{\partial g(t)}{\partial t} \Big|_{\mu})^2 \sigma^2$ the same limit as $P(\sqrt{n}(g(\bar{X}_n) - g(\mu)) \leq x)$.

If we assume $g : \mathbb{R} \rightarrow \mathbb{R}$ differentiable in a neighbourhood of μ and $\frac{\partial g(t)}{\partial t}$ continuous at μ , then

$$g(\bar{X}_n^*) - g(\bar{X}_n) = \frac{\partial g(t)}{\partial t} \Big|_{\bar{X}_n} (\bar{X}_n^* - \bar{X}_n) + r_n^* \text{ with } n^{\frac{1}{2}} r_n^* \xrightarrow{\text{a.s.}} 0$$

and since $\frac{\partial g(t)}{\partial t} \Big|_{\bar{X}_n} \xrightarrow{\text{a.s.}} \frac{\partial g(t)}{\partial t} \Big|_{\mu}$ we get

$$\rho_\infty(\sqrt{n}(g(\bar{X}_n^*) - g(\bar{X}_n)), \sqrt{n}(g(\bar{X}_n) - g(\mu))) \xrightarrow{\text{a.s.}} 0.$$

Now, if $\frac{\partial g(t)}{\partial t} \Big|_{\mu} = 0$, $\frac{\partial^2 g(t)}{\partial t^2} \Big|_{\mu} \neq 0$, then $\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{\text{p}} 0$ and the interesting statistic is

$n(g(\bar{X}_n) - g(\mu))$ which converges in distribution to

$$\frac{1}{2} \frac{\partial^2 g(t)}{\partial t^2} \Big|_{\mu} Z^2, \text{ with } \mathcal{L}(Z) = \mathcal{N}(0, \sigma^2).$$

Then, $n(g(\bar{X}_n^*) - g(\bar{X}_n))$ is typically not going to be a consistent estimator since

$$g(\bar{X}_n^*) - g(\bar{X}_n) = \frac{\partial g(t)}{\partial t} \Big|_{\bar{X}_n} (\bar{X}_n^* - \bar{X}_n) + \frac{\partial^2 g(t)}{\partial t^2} \Big|_{\bar{X}_n} (\bar{X}_n^* - \bar{X}_n)^2 + o_p^*(n^{-1})$$

(provided some additional assumptions are satisfied) and $\frac{\partial g(t)}{\partial t} \Big|_{\bar{X}_n}$ is typically $\neq 0$. Then, almost surely,

$$\begin{aligned} n \frac{\partial g(t)}{\partial t} \Big|_{\bar{X}_n} (\bar{X}_n^* - \bar{X}_n) &= n \left(\frac{\partial g(t)}{\partial t} \Big|_{\mu} + \frac{\partial^2 g(t)}{\partial t^2} \Big|_{\mu} (\bar{X}_n - \mu) + o_p(n^{-1}) \right) (\bar{X}_n^* - \bar{X}_n) \\ &\xrightarrow{\mathcal{L}} \frac{\partial^2 g(t)}{\partial t^2} \Big|_{\mu} Z^2 \text{ with } \mathcal{L}(Z) = \mathcal{N}(0, \sigma^2) \end{aligned}$$

compared to $n \frac{\partial g(t)}{\partial t} \Big|_{\mu} (\bar{X}_n - \mu) = 0$.

Now, it is easily seen that, if we take $m(g(\bar{X}_m^*) - g(\bar{X}_n))$, $m \rightarrow \infty$, $\frac{m}{n} \rightarrow 0$, we have a consistent estimator of $n(g(\bar{X}_n) - g(\mu))$. In particular, $(m \frac{\partial g(t)}{\partial t} \Big|_{\bar{X}_n} (\bar{X}_m^* - \bar{X}_n) = m^{\frac{1}{2}}(\bar{X}_n - \mu)m^{\frac{1}{2}}(\bar{X}_m^* - \bar{X}_n) \frac{\partial g(t)}{\partial t} \Big|_{\bar{X}_n} + o_p(1)$ and $m^{\frac{1}{2}}(\bar{X}_n - \mu) \xrightarrow{p} 0$).

In this section we have focused on techniques for proving consistency. For more general results, see, e.g., Shao and Tu pp. 72-90.

4. Asymptotic comparisons (i.i.d. sample)

A fourth possible way for proving consistency of the bootstrap is by means of Berry-Esséen inequalities. The advantage of this method is that it provides us with rates of convergence; the drawback is that it is often not possible to derive a Berry-Esséen inequality.

As in chapter 3 we focus on X_1, \dots, X_n i.i.d. with finite mean μ and finite variance σ^2 . Additionally we assume $E|X_1 - \mu|^3 < \infty$. We are still interested in

$$\sup_x |P(\sqrt{n}(\bar{X}_n - \mu) \leq x) - P_*(\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x)|.$$

Now, the Berry-Esséen theorem tells us that

$$\sup_x |P(\frac{1}{\sqrt{n}\sigma} \sum_{i=1}^n (X_i - \mu) \leq x) - \Phi(x)| \leq \frac{C}{n^{\frac{1}{2}}\sigma^{\frac{3}{2}}} E|X_1 - \mu|^3.$$

This means that

$$\rho_{\infty}(\sqrt{n}(\bar{X}_n - \mu), \Phi_{0,\sigma^2}) = O(n^{-\frac{1}{2}}).$$

Now,

$$\sup_x |P_*(\frac{1}{\sqrt{n}\hat{\sigma}_n} \sum_{i=1}^n (X_i^* - \bar{X}_n) \leq x) - \Phi(x)| \leq \frac{C}{n^{\frac{1}{2}}\hat{\sigma}_n^{\frac{3}{2}}} E_*|X_1^* - \bar{X}_n|^3 \text{ a.s.}$$

and $E_*|X_1^* - \bar{X}_n|^3 = \frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}_n|^3 < \infty$ a.s. .

In total $\rho_{\infty}(\sqrt{n}(\bar{X}_n - \mu), \sqrt{n}(\bar{X}_n^* - \bar{X}_n)) \leq \rho_{\infty}(\sqrt{n}(\bar{X}_n - \mu), \Phi_{0,\sigma^2}) + \rho_{\infty}(\Phi_{0,\sigma^2}, \Phi_{0,\hat{\sigma}_n^2}) + \rho_{\infty}(\Phi_{0,\hat{\sigma}_n^2}, \sqrt{n}(\bar{X}_n^* - \bar{X}_n)) \leq \rho_{\infty}(\Phi_{0,\sigma^2}, \rho_{\infty}(\Phi_{0,\hat{\sigma}_n^2})) + O(n^{-\frac{1}{2}})$ a.s.

and $\rho_{\infty}(\Phi_{0,\sigma^2}, (\Phi_{0,\hat{\sigma}_n^2})) = \sup_x |\int_{-\infty}^x \varphi_{0,\sigma^2}(y) - \varphi_{0,\hat{\sigma}_n^2}(y) dy| = O(n^{-\frac{1}{2}})$ a.s. .

(This can easily be seen, e.g., by means of a Taylor expansion of $\varphi_{0, \hat{\sigma}_n^2}(y)$).

So $\rho_\infty(\sqrt{n}(\bar{X}_n - \mu), \sqrt{n}(\bar{X}_n^* - \bar{X}_n)) = O(n^{-\frac{1}{2}})$ a.s., the same rate as for the normal approximation. Now, we will consider $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$. This statistic is asympt. $\mathcal{N}(0, 1)$ distributed and the rate of convergence to $\mathcal{N}(0, 1)$ is $O(n^{-\frac{1}{2}})$ (Berry-Esséen).

Now, we will sketch the proof of the fact that under some conditions $\rho_\infty(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}, \frac{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)}{\hat{\sigma}_n}) = O(n^{-1})$. So, for this statistic, the bootstrap approximation is asymptotically better than the normal approximation.

The idea is the following: (We assume X_1, \dots, X_n i.i.d. with abs. cont. distribution, $EX_1^4 < \infty$.) We will get a so-called Edgeworth expansion for $P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\right) : P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\right) = \Phi(x) + n^{-\frac{1}{2}}p_1(x)\varphi(x) + n^{-1}p_2(x)\varphi(x) + o(n^{-1})$ with $p_1(x)$ polynom of degree 2, $p_2(x)$ polynom of degree 5 in x .

Now, it is possible to show that

$$P\left(\frac{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)}{\hat{\sigma}_n} \leq x\right) = \Phi(x) + n^{-\frac{1}{2}}\hat{p}_1(x)\varphi(x) + n^{-1}\hat{p}_2(x)\varphi(x) + o_p(n^{-1})$$

(see Hall for a proof). In particular

$$\hat{p}_j(x) - p_j(x) = O(n^{-\frac{1}{2}}), j = 1, 2.$$

Thus,

$$\sup_x \left| P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\right) - P_*\left(\frac{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)}{\hat{\sigma}_n} \leq x\right) \right| = O_p(n^{-1}).$$

In the following we are going to sketch the proof of the Edgeworth expansion for $P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\right)$. Let

$$\begin{aligned} \chi_n(t) &:= E(e^{it\sqrt{n}(\frac{\bar{X}_n - \mu}{\sigma})}) \\ &= E(e^{it n^{-\frac{1}{2}} \sum_i (\frac{X_i - \mu}{\sigma})}) \\ &= \prod_{i=1}^n E(e^{it n^{-\frac{1}{2}} (\frac{X_i - \mu}{\sigma})}). \end{aligned}$$

If $\chi(t) := E(e^{it(\frac{X_1 - \mu}{\sigma})})$, then $\chi_n(t) = (\chi(\frac{t}{\sqrt{n}}))^n$.

Now, $\chi(t)$, the characteristic function of $\frac{X_1 - \mu}{\sigma} := Y$ is such that

$$\chi(t) = 1 + E(Y)it + \frac{1}{2}E(Y^2)(it)^2 + \dots$$

On the other hand,

$$\chi(t) = e^{K_1 it + \frac{1}{2}K_2(it)^2 + \dots}$$

with K_j : j th-cumulant of Y .

By Taylor the first expression leads to $\log(\chi(t) = E(Y)it + \frac{1}{2}E(Y^2)it^2 + \dots - \frac{1}{2}(E(Y)it + \dots)^2 + \dots$ and so $K_1 = E(Y)$, $K_2 = \text{var}(Y)$, $K_3 = E(Y - E(Y))^3, \dots$

Now, $E(Y) = 0$, $\text{var}(Y) = 1$ since $Y = \frac{X_1 - \mu}{\sigma}$

$$\longrightarrow \chi(t) = \exp\left\{\frac{1}{2}(it)^2 + \frac{1}{\sigma}(it)^3 K_3 + \dots\right\}$$

$$\begin{aligned} \longrightarrow \chi_n(t) &= \left(\chi\left(\frac{t}{\sqrt{n}}\right)\right)^n = \left(\exp\left\{-\frac{1}{2}\frac{t^2}{n} + \frac{1}{\sigma}K_3 \frac{i^3 t^3}{n^{3/2}} + \dots\right\}\right)^n = \exp\left\{-\frac{1}{2}t^2 + \frac{1}{\sigma}K_3 \frac{i^3 t^3}{n^{1/2}} + \dots\right\} \\ &= \exp\left(-\frac{1}{2}t^2\right) \exp\left\{\frac{1}{\sigma}K_3 \frac{i^3 t^3}{n^{1/2}} + \dots\right\} \end{aligned}$$

$$\begin{aligned}
&= \exp(-\frac{1}{2}t^2)(1 + \frac{1}{\sigma}K_3 \frac{i^3 t^3}{n^{1/2}} + \dots + (\frac{1}{\sigma}K_3 \frac{i^3 t^3}{n^{1/2}} + \dots)^2 + \dots) \\
&= e^{-\frac{t^2}{2}}(1 + n^{-1/2}r_1(it) + n^{-1}r_2(it) + o(n^{-1})) \\
&\text{with } r_1(u) = \frac{1}{6}K_3 u^3, r_2(u) = \frac{1}{24}K_4 u^4 + \frac{1}{72}u^6 K_3^2.
\end{aligned}$$

Now,

$$\chi_n(t) = e^{-t^2/2} + n^{-1/2}r_1(it)e^{-t^2/2} + n^{-1}r_2(it)e^{-t^2/2} + o(n^{-1})$$

$$\text{and } \chi_n(t) = \int_{-\infty}^{\infty} e^{itx} d P(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x), e^{-t^2/2} = \int_{-\infty}^{\infty} e^{itx} d \Phi(x).$$

So, in case of an abs. continuous distribution, the possibility of an "inverse expansion":

$P(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x) = \Phi(x) + n^{-1/2}R_1(x) + n^{-1}R_2(x) + o(n^{-1})$ with $\int_{-\infty}^{\infty} e^{itx} d R_j(x) = r_j(it)e^{-t^2/2}$ $j = 1, 2$. $R_1(x)$ and $R_2(x)$ can easily be obtained. We have, e.g., for $R_1(x) = \frac{-1}{6}K_3(x^2 - 1)\varphi(x)$. And in total we get a rate of convergence for the bootstrap approximation of $O(n^{-1})$ as briefly explained above.

A last comment to conclude the chapter: to use the same way, it is possible to show that

$$\begin{aligned}
&P(\sqrt{n}(\bar{X}_n - \mu) \leq x) = \Phi(\frac{x}{\sigma}) + n^{-1/2}q_1(\frac{x}{\sigma})\varphi(\frac{x}{\sigma}) + n^{-1}q_2(\frac{x}{\sigma})\varphi(\frac{x}{\sigma}) + o(n^{-1}) \\
&\text{and } P_*(\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x) = \Phi(\frac{x}{\hat{\sigma}_n}) + n^{-1/2}\hat{q}_1(\frac{x}{\hat{\sigma}_n})\varphi(\frac{x}{\hat{\sigma}_n}) + n^{-1}\hat{q}_2(\frac{x}{\hat{\sigma}_n})\varphi(\frac{x}{\hat{\sigma}_n}) + o(n^{-1}).
\end{aligned}$$

Now, if we look at $|P(\sqrt{n}(\bar{X}_n - \mu) \leq x) - P_*(\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x)|$, we get a rate $O(n^{-1/2})$ (since $\Phi(\frac{x}{\sigma}) - \Phi(\frac{x}{\hat{\sigma}_n}) = O(n^{-1/2})$) : the same rate as when using Berry-Esséen and the same rate as for the normal approximation). The difference between $\sqrt{n}(\bar{X}_n - \mu)$ and $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ is that the second statistic is "pivotal" (i.e. its limiting distribution does not depend on unknown quantities) and for pivotal statistics Edgeworth expansions enable us to get better rates than the normal approximation.

5. Bootstrap confidence sets and hypothesis tests

We have our data X_1, \dots, X_n i.i.d. and we would like to construct a $(1 - \alpha)$ -confidence interval (CI) for an unknown "parameter" θ :

$$P(I_1(X_1, \dots, X_n) \leq \theta \leq I_2((X_1, \dots, X_n))) = 1 - \alpha.$$

We denote an estimator of θ by T_n . Often, there exists a central limit theorem of the form

$$\frac{T_n - \theta}{\sqrt{\hat{\text{var}}(T_n)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1),$$

and by means of this we get an approximate $(1 - \alpha)$ -CI with bounds $T_n \pm u_{1-\frac{\alpha}{2}} \sqrt{\hat{\text{var}}(T_n)}$, where $u_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$ -th quantile of a $\mathcal{N}(0, 1)$ dist. random variable.

How can bootstrap be used in this context?

There are different possible approaches.

First, often it can be shown (see, e.g., previous chapters) that

$$\rho_{\infty} \left(\frac{T_n - \theta}{\sqrt{\hat{\text{var}}(T_n)}}, \frac{T_n^* - T_n}{\sqrt{\text{var}_*(T_n^*)}} \right) \xrightarrow{\text{P}} 0.$$

Let us denote $\mathcal{L} \left(\frac{T_n^* - T_n}{\sqrt{\text{var}_*(T_n^*)}} \right)$ by H_{Boot} .

Now, we can by means of simulations get very good estimates of $H_{\text{Boot}}^{-1}(1 - \frac{\alpha}{2})$ and $H_{\text{Boot}}^{-1}(\frac{\alpha}{2})$. We plug-in these estimates into our bootstrap confidence interval:

$$P(T_n - H_{\text{Boot}}^{-1}(1 - \frac{\alpha}{2})\sqrt{\hat{\text{var}}(T_n)} \leq \theta \leq T_n - H_{\text{Boot}}^{-1}(\frac{\alpha}{2})\sqrt{\hat{\text{var}}(T_n)}) \simeq 1 - \alpha.$$

This is called a bootstrap $-t$ CI. The CI is not necessarily symmetric around T_n and the rate of convergence to $1 - \alpha$ is typically faster than in the traditional case. It is a very useful method if $\hat{\text{var}}(T_n)$ is a good estimator of $\text{var}(T_n)$. If it is not the case another method should be chosen. We can also notice that bootstrap $-t$ CI are not transformation preserving: if we construct a bootstrap $-t$ CI for $g(\theta)$, g strictly increasing, say $P(I_1^* \leq g(\theta) \leq I_2^*) \simeq 1 - \alpha$, then $P(g^{-1}(I_1^*) \leq \theta \leq g^{-1}(I_2^*)) \simeq 1 - \alpha$ but $g^{-1}(I_1^*)$ and $g^{-1}(I_2^*)$ may be very different from the bounds obtained above.

If we cannot get a good enough estimator for $\text{var}(T_n)$, then we can use the so-called hybrid bootstrap. Let us suppose we have $\rho_\infty(c_n(T_n - \theta), c_n(T_n^* - T_n)) \xrightarrow{p} 0$, where $c_n \rightarrow \infty$ as

$n \rightarrow \infty$ (e.g. $c_n = \sqrt{n}$ for $T_n = \bar{X}_n$).

Then, if $\mathcal{L}(c_n(T_n^* - T_n)) = H_{\text{Boot}}$,

$$P(H_{\text{Boot}}^{-1}(\frac{\alpha}{2}) \leq c_n(T_n - \theta) \leq H_{\text{Boot}}^{-1}(1 - \frac{\alpha}{2})) \simeq 1 - \alpha$$

and it remains to isolate θ in the middle.

A third possibility is to use the so-called percentile bootstrap. Denote $\mathcal{L}(T_n^*)$ by K_{Boot} .

Then a one-sided percentile bootstrap is given by $[K_{\text{Boot}}^{-1}(\alpha), +\infty)$.

(Of course, a two-sided percentile bootstrap is constructed in the same way: $[K_{\text{Boot}}^{-1}(\frac{\alpha}{2}), K_{\text{Boot}}^{-1}(1 - \frac{\alpha}{2})]$; we are using one-sided CI because the proof of the following claim is then easier in terms of notations.)

Claim: the percentile bootstrap $[K_{\text{Boot}}^{-1}(\alpha), +\infty)$ is an exact $(\frac{1-\alpha}{2})$ -CI for θ if there exists a strictly increasing function m_n such that $\mathcal{L}(m_n(T_n) - m_n(\theta)) = \mathcal{N}(0, 1)$ and $\mathcal{L}_*(m_n(T_n^*) - m_n(T_n)) = \mathcal{N}(0, 1)$.

Proof $P(m_n(T_n) - m_n(\theta) \leq u_{1-\alpha}) = 1 - \alpha$

$\rightarrow P(m_n^{-1}(m_n(T_n) - u_{1-\alpha}) \leq \theta) = 1 - \alpha$ (exact $(1 - \alpha)$ CI).

We have to show that

$$P_*(T_n^* \geq m_n^{-1}(m_n(T_n) - u_{1-\alpha})) = 1 - \alpha \text{ (i.e. } K_{\text{Boot}}^{-1}(\alpha) = m_n^{-1}(m_n(T_n) - u_{1-\alpha})).$$

The left-hand side is equal to

$$P_*(m_n(T_n^*) - m_n(T_n) \geq u_\alpha) = 1 - \alpha.$$

So, if there exists such a function m_n we do not need to find her: we still can get an exact CI by means of percentile bootstrap. Obviously it is very rarely the case that such m_n exists for finite n ; asymptotically however the assumptions in the claim are in many cases satisfied.

The bootstrap percentile is transformation preserving.

Note that to get a better rate of convergence to $1 - \alpha$ the so-called bootstrap accelerated bias-corrected percentile is quite often used. It is based on the assumption that there exists a strictly increasing function m_n such that

$$\mathcal{L}\left(\frac{m_n(T_n) - m_n(\theta)}{1 + a_n m_n(\theta)} + z_n\right) = \mathcal{N}(0, 1).$$

Drawback of this method: z_n and a_n may be difficult to estimate.

Next possible method: the iterative bootstrap (also called double bootstrap).

Let $R_n^{(0)} = c_n(T_n - \theta)$ (c_n has the same meaning as for the hybrid bootstrap); let $H_n^{(0)}$ be the distribution function of $R_n^{(0)}$; let $H_{\text{Boot}}^{(0)}$ be the bootstrap estimation of $H_n^{(0)}$.

Now, $H_n^{(0)}$ is not pivotal (i.e. depends on unknown quantities) and so $H_{\text{Boot}}^{(0)}$ is not as efficient as if $H_n^{(0)}$ was pivotal. How to make the statistic more pivotal? $H_n^{(0)}(R_n^{(0)})$ is $U(0, 1)$ distributed (in particular, it is pivotal) but $H_n^{(0)}$ is unknown to us. Instead we consider $H_{\text{Boot}}^{(0)}(R_n^{(0)}) := R_n^{(1)}$ (close to $U(0, 1)$; more pivotal than $R_n^{(0)}$).

Let $H_n^{(1)}$ be the distribution function of $R_n^{(1)}$; let $H_{\text{Boot}}^{(1)}$ be the bootstrap estimation of $H_n^{(1)}$ and so on $R_n^{(2)} = H_{\text{Boot}}^{(1)}(R_n^{(1)})$...

Now, the (one-sided) $(1 - \alpha)$ confidence interval obtained after j iterations is given by $C_{IB}^{(j)} = \{\theta : R_n^{(j)} \leq H_{\text{Boot}}^{(j-1)}(1 - \alpha)\}$. (Of course, here θ is not the true value of the parameter.)

Drawback: a lot of simulations are needed.

Example: $H_{\text{Boot}}^{(0)}(x)$ is estimated by

$$\frac{1}{B_1} \sum_{i=1}^{B_1} I(R_n^{(0)}(X_{1i}^*, \dots, X_{ni}^*) \leq x)$$

so $H_{\text{Boot}}^{(0)}(R_n^{(0)}(X_1, \dots, X_n))$ is estimated by

$$\frac{1}{B_1} \sum_{i=1}^{B_1} I(R_n^{(0)}(X_{1i}^*, \dots, X_{ni}^*) \leq R_n^{(0)}(X_1, \dots, X_n)).$$

Now, for $1 \leq b \leq B_1$, let

$$z_b^* = \frac{1}{B_2} \sum_{k=1}^{B_2} I(R_n^{(0)}(X_{1bk}^{**}, \dots, X_{nbk}^{**}) \leq R_n^{(0)}(X_{1b}^*, \dots, X_{nb}^*))$$

($X_{1bk}^{**}, \dots, X_{nbk}^{**}$ taken i.i.d. from $X_{1b}^*, \dots, X_{nb}^*$).

By means of these z_b^* we can estimate $H_{\text{Boot}}^{(1)}$. Roughly,

$$C_{IB}^{(1)} = \left\{ \theta : R_n^{(0)} \leq (H_{\text{Boot}}^{(0, B_1)})^{-1} (H_{\text{Boot}}^{(1, B_1, B_2)})^{-1} (1 - \alpha) \right\}.$$

Before looking at tests a few general words about importance sampling, which can also be useful for bootstrap (estimation of tail probabilities).

Let us suppose we would like to estimate the value of the integral $\int f(z)g(z)d(z)$, f density function (the integral is assumed to exist). A possibility is to use $\hat{e}_1 = \frac{1}{B} \sum_{i=1}^B g(z_i)$, z_1, \dots, z_B taken from f .

Now, let us suppose there exists a function $h(z)$, also a density function, which is such that

$$\frac{f(z)g(z)}{h(z)} \simeq C, \text{ constant. } (h(z) \neq 0 \forall z).$$

Then, $\int f(z)g(z)dz = \int \frac{f(z)g(z)}{h(z)} h(z)dz$ and a second estimator is given by

$$\hat{e}_2 = \frac{1}{B} \sum_{i=1}^B \frac{f(z_i)g(z_i)}{f(z_i)}, z_1, \dots, z_B \text{ from } h.$$

Now, $E\hat{e}_1 = E\hat{e}_2 = \int f(z)g(z)dz$ but $\text{var } \hat{e}_1 \gg \text{var}\hat{e}_2$ (since $\frac{f(z_i)g(z_i)}{h(z)} \simeq C$).

Example: $L(Z) = N(0,1)$,

$$P(Z > 1.96) = \int 1_{(z>1.96)}\varphi(z)dz.$$

Let $h(z) = \varphi_{1.96,1}(z)$.

Then $\text{var}\hat{e}_1 \simeq 17 \text{var}\hat{e}_2$.

For bootstrap implications, see, e.g., Shao-Tu pp. 223-227.

Tests

Before speaking about bootstrap tests a few words about related tests, the so-called permutation tests (first introduced by Fisher in the 30's). An example is given now.

We have data X_1, \dots, X_m i.i.d. from F , Y_1, \dots, Y_n i.i.d. from G . We would like to test $H_0 : F = G$ against $H_1 : F \neq G$. (the significance level α is fixed). We first select a test statistic, say T (e.g. the mean difference). Let us suppose we expect this statistic to be large.

For our data we have a value for T , say T_r . Now, we would like to compare $P_{H_0}(T \geq T_r)$ with α ($P_{H_0}(T \geq T_r)$ is the probability, H_0 being true, that $T \geq T_r$ sometimes called achieved significance level (ASL) or p value). If $P_{H_0}(T \geq T_r) < \alpha$ we reject H_0 .

Now, how to estimate $P_{H_0}(T \geq T_r)$?

A few notations are needed.

Let $Z = (X_1, \dots, X_m, Y_1, \dots, Y_n)$; v be the combined and ordered vector of values; $g = (g_1, \dots, g_{m+n})$ be the vector that indicates which group each ordered observation belongs to. The vector g consists of m X 's and n Y 's. There are $\binom{m+n}{n}$ possibilities of having m X 's and n Y 's in a vector of length $m+n$.

Now, the permutation lemma tells us that, under H_0 , the vector g has probability $1/\binom{m+n}{n}$ of equaling any one of the possible ways of having m X 's and n Y 's.

It means that for a permutation test, $P_{H_0}(T \geq T_r) = \frac{\#\{s:T_s \geq T_r\}}{\binom{m+n}{n}}$.

Note that we have used kind of symmetry in the null hypothesis to be able to get a permutation distribution ($\frac{1}{\binom{m+n}{n}}$ to any possibility) and then to get an estimate for $P_{H_0}(T \geq T_r)$.

A nice feature of these permutation tests is that P_{H_0} (we reject H_0) $\simeq \alpha$ since P_{H_0} (ASL of permutation test = $\frac{k}{n+m}$) = $\frac{1}{n+m} k = 1, \dots, n+m$ (if there are no ties among the T_s .)

The permutation tests provide us with good results. Unfortunately, they are not widely applicable.

Now, for the permutation tests, in the $\binom{m+n}{n}$ different configurations the drawings are made without replacement. For bootstrap tests the drawings are made with replacement which makes these tests much more widely applicable.

The strategy of a bootstrap test is to estimate the probability mechanism under the null hypothesis, then sample from it and finally estimate the achieved significance level.

Here a few examples of computation of bootstrap tests:

Computation of bootstrap test statistic for testing $F = G$

Data X_1, \dots, X_m from F Y_1, \dots, Y_n from G

Let $Z = (X_1, \dots, X_m, Y_1, \dots, Y_n)$

– Draw B samples of size $n + m$ with replacement from Z . Call the first m observations X^* , the remaining ones Y^* .

– Select a test statistic T (e.g. difference of means). Let T_r be the value for the original data. Evaluate T on each bootstrap sample: T_1^*, \dots, T_B^*

– estimation of the achieved significance level for the bootstrap test by

$$\frac{\#\{T_b^* \geq T_r\}}{B}.$$

Computation of bootstrap test statistic for testing equality of means

Data as above.

Our test statistic can be chosen as

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_1^2}{m} + \frac{\hat{\sigma}_2^2}{n}}}$$

– Let \tilde{F} put equal probability on the points $\tilde{X}_i = X_i - \bar{X} + \bar{Z}$ $i = 1, \dots, m$; let \tilde{G} put equal probability on the points $\tilde{Y}_i = Y_i - \bar{Y} + \bar{Z}$.

– Form B bootstrap data sets (X^*, Y^*) where X^* is a vector of length m obtained by a drawing with replacement from the \tilde{X}_i , Y^* is a vector of length n obtained by a drawing with replacement from the \tilde{Y}_i

– we get B bootstrap statistics of the form: $\frac{\bar{X}^* - \bar{Y}^*}{\sqrt{\frac{\hat{\sigma}_1^2}{m} + \frac{\hat{\sigma}_2^2}{n}}}$

– we estimate the achieved significance level.

Notice the importance to draw from distributions with the same mean since we want to estimate the probability mechanism under the null hypothesis. Moreover notice that implicitly it is assumed that both F and G belong to a translation family.

If we do not believe that F and/or G belong to a translation family other methods exist for estimating the probability mechanism under H_0 (see, e.g., Efron-Tibshirani, p. 235).

A last remark: it is clearly possible to obtain a hypothesis test by constructing an appropriate confidence set. Still bootstrap hypothesis testing is an important topic for some reasons, one of the main ones being that a bootstrap test takes into account the particular structure of H_0 and is therefore often more efficient (e.g. in terms of power) as a test based on confidence set.

6. Application to linear models

Our model is $Y = Xb + \varepsilon$, where X is a fixed design matrix, $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. $(0, \sigma_\varepsilon^2)$, Y is the vector of observations and b is the vector of unknown parameters.

A usual way to estimate b is to use the least-squares (LS) estimator $\hat{b} = (X^T X)^{-1} X^T Y$ (if $X^T X$ is invertible).

Clearly $E\hat{b} = b$ and under very general conditions \hat{b} is asymptotically normally distributed with mean b and variance $\sigma_\varepsilon^2(X^T X)^{-1}$.

Our statistical model can be summarized by $M = (F_\varepsilon, b)$, F_ε has mean 0.

How to use bootstrap in this context? We can mimick M by $\hat{M} = (\hat{F}_\varepsilon, \hat{b})$, with \hat{b} the LS estimator and \hat{F}_ε putting mass $\frac{1}{n}$ to $\hat{e}_i - \frac{1}{n} \sum_{i=1}^n \hat{e}_i$ $i = 1, \dots, n$.

$\hat{e} = Y - X\hat{b}$ is the vector of the empirical residuals.

\hat{F}_ε puts mass $\frac{1}{n}$ to $\hat{e}_i - \frac{1}{n} \sum_{i=1}^n \hat{e}_i$ and not to \hat{e}_i to ensure that $E_* X^* = 0$ with X^* having dist. \hat{F}_ε .

Notice that $\frac{1}{n} \sum_{i=1}^n \hat{e}_i = 0$ if the constant vector $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ belongs to X .

Now, we draw e_1^*, \dots, e_n^* i.i.d. from \hat{F}_ε and we get $Y^* = X\hat{b} + e^*$ which leads to

$$\hat{b}^* = (X^T X)^{-1} X^T Y^*.$$

Clearly, $E_* \hat{b}^* = \hat{b}$. We can go further:

Theorem: Suppose trace $(X X^T) \uparrow \infty$, then

$$\tilde{\rho}_2(H_{\text{Boot}}, H_n) \xrightarrow{\text{a.s.}} 0,$$

where H_n is the distribution of $(X^T X)^{1/2}(\hat{b} - b)$, H_{Boot} is the bootstrap distribution of $(X^T X)^{1/2}(\hat{b}^* - \hat{b})$.

Proof: (only in the case $\frac{1}{n} \sum_{i=1}^n \hat{e}_i = 0$; the more general case takes a few more lines, see, e.g., Shao-Tu pp. 320-321).

$$\begin{aligned} \tilde{\rho}_2(H_{\text{Boot}}, H_n) &= \tilde{\rho}_2((X^T X)^{-1/2} X^T e^*, (X^T X)^{-1/2} X^T \varepsilon) \\ &\leq \tilde{\rho}_2(e^*, \varepsilon) = \tilde{\rho}_2(\hat{F}_\varepsilon, F_\varepsilon) \xrightarrow{\text{a.s.}} 0. \end{aligned}$$

Now, let us suppose X is not fixed anymore but random. It is still possible to use residual bootstrap, i.e. $Y^* = X\hat{b} + e^*$. An alternative, if $(Y_i, X_i, \dots, X_{ip}) := (Y_i, X_i^T) i = 1, \dots, n$ are i.i.d. is to use the so-called paired bootstrap.

In this case and if we are interested in the least-squares estimator of b the statistical model M can be identified by the joint distribution of (Y_i, X_i^T) and estimated by the empirical distribution function putting mass $\frac{1}{n}$ to $(Y_i, X_i^T) i = 1, \dots, n$.

Then, we get Y^*, X^* and $\hat{b}^* = (X^{*T} X^*)^{-1} X^{*T} Y^*$.

Under some conditions we get consistency of $(X^{*T} X^*)^{1/2}(\hat{b}^* - \hat{b})$ (see, e.g., Shao-Tu p. 322). Clearly, the paired bootstrap is robust against heteroskedasticity of the residuals contrary to the residual bootstrap.

On the other hand, it is not as efficient as the residual bootstrap if the model is correct and it is typically not consistent in the nonparametric case (i.e. $Y_i = m(X_i) + \varepsilon_i$).

Now, a bootstrap which applies to both random and fixed design, which is robust against heteroskedasticity and which can also be consistent in the nonparametric case is the so-called wild bootstrap.

Let $Y = Xb + \varepsilon$, $\varepsilon_1, \dots, \varepsilon_n$ independent with mean 0. Now, let $e_i^* i = 1, \dots, n$ be i.i.d. from a distribution with mean 0 and variance 1. Then

$$Y_i^* = x_i^T \hat{b} + \frac{|\hat{e}_i|}{\sqrt{1 - h_i}} e_i^*$$

with $h_i = x_i^T (X^T X)^{-1} x_i$.

The coefficient $\frac{|\hat{\epsilon}_i|}{\sqrt{1-h_i}}$ ensures that $\text{var}_* Y_i^* = \text{var} Y_i$. The variance of the observation i is estimated just by means of observation i . Of course this allows for (strong) heteroskedasticity. On the other hand it is not as efficient as residual bootstrap if $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d..

7. Application to nonparametric models

We focus in this chapter on kernel estimation and more particularly we shall concentrate on the kernel estimation in the case of nonparametric regression with fixed design:

$$Y_j = m(x_j) + \varepsilon_j, \quad \varepsilon_1, \dots, \varepsilon_n \text{ i.i.d. } (0, \sigma_\varepsilon^2), \quad m : [0, 1] \rightarrow \mathbb{R}, \quad x_j = \frac{j}{n} \quad j = 1, \dots, n.$$

Notice that result similar to those we are going to review also exist for nonparametric regression with stochastic design or for density estimation.

To estimate m at x we may use a kernel estimator: $\hat{m}(x, h) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) Y_i$, where h is the bandwidth and K is a kernel function (i.e. a continuous bounded function s.t. $\int K(u) du = 1$).

Under some general conditions (in particular $m \in C^2$), it can easily be shown that the mean squared error (MSE) of $\hat{m}(x, h)$ is asymptotically equal to:

$$\text{MSE}(\hat{m}(x, h)) = \sigma_\varepsilon^2 \frac{1}{nh} \int K^2(u) du + (m''(x))^2 \frac{h^4}{4} \left(\int u^2 K(u) du \right)^2 + o\left(\frac{1}{nh}\right) + o(h^4)$$

which leads to an optimal asymptotic h of order $O(n^{-1/5})$. We get a same order for the optimal asymptotic h if we take a global measure of accuracy, the integrated mean squared error (IMSE).

Now, for finite n , which h should we choose?

A naive idea would be to take the h which minimizes $\sum_{i=1}^n (Y_i - \hat{m}(x_i, h))^2$. Clearly this leads to a kind of interpolation. A much better idea is to use cross-validation (leave one out at a time - similar to jackknife idea). We take h which minimizes $\sum_{i=1}^n (Y_i - \hat{m}_{-i}(x_i, h))^2$, where $\hat{m}_{-i}(x_i, h)$ is the estimator of m at x_i obtained by means of all data except (x_i, Y_i) .

This technique provides us with some nice asymptotic properties (see, e.g., W. Härdle: "Applied Nonparametric Regression" (1990), Chapter 5).

Now, what can bootstrap bring us in this context? First, we should answer the question: how to bootstrap?

We have empirical residuals $\hat{\epsilon}_i = Y_i - \hat{m}(x_i, h)$ $i = 1, \dots, n$. We remove those empirical residuals which correspond to x_i close to the boundaries (because \hat{m} is not a very good estimator close to the boundaries), say we remove ηn data corresponding to x_i close 0 and also ηn data corresponding to x_i close to 1.

We center the remaining residuals. Let us denote \hat{F}_ε the empirical distribution function corresponding to these centered remaining residuals. We draw $\varepsilon_1^*, \dots, \varepsilon_n^*$ i.i.d. from \hat{F}_ε and we get

$$Y_i^* = \hat{m}(x_i, h) + \varepsilon_i^*.$$

Now, we could hope that $E_*(\hat{m}^*(x, h) - \hat{m}(x, h))^2$ is close to $E(\hat{m}(x, h) - m(x))^2$, with $\hat{m}^*(x, h) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) Y_i^*$.

But this is not the case for $h \sim n^{-1/5}$. We see now why.

$E_*(\hat{m}^*(x, h)) = \hat{m}(x, h) + \frac{h^2}{2} \hat{m}''(x, h) \int u^2 K(u) du + o(h^2)$ a.s. (exactly in the same way as in the traditional case.)

Now, is $\hat{m}''(x, h)$ a good estimator of $m''(x)$?

$$\hat{m}''(x, h) = \frac{1}{nh^3} \sum_{i=1}^n K''\left(\frac{x-x_i}{h}\right) Y_i$$

$$\text{and } \text{var}(\hat{m}''(x, h)) = \frac{1}{n^2 h^6} \sum_{i=1}^n K''^2\left(\frac{x-x_i}{h}\right) \sigma_\varepsilon^2 \simeq \frac{\sigma_\varepsilon^2}{nh^6} \int K''^2\left(\frac{u}{h}\right) du = \frac{\sigma_\varepsilon^2}{nh^5} \int K''^2(v) dv$$

and for $h \sim n^{-1/5}$,

$$\text{var}(\hat{m}''(x, h)) \rightarrow 0.$$

So, for $h \sim n^{-1/5}$, the bootstrap bias is not "close" to the true bias. As a consequence we have also $\mathcal{L}_*(\sqrt{nh}(\hat{m}^*(x, h) - \hat{m}(x, h))) \rightarrow \mathcal{L}(\sqrt{nh}(\hat{m}(x, h) - m(x)))$.

Now, in the following of the chapter, we shall present two possible ways of dealing with this problem of bias.

We are interested at $E \hat{m}(x, h) - m(x)$ $h \sim n^{-1/5}$ and more generally at $\mathcal{L}(\sqrt{nh}(\hat{m}(x, h) - m(x)))$ $h \sim n^{-1/5}$. Let \hat{F}_ε be defined as earlier in this chapter.

Now, we consider $Y_i^* = \hat{m}(x_i, g) + \varepsilon_i^*$, with g a bandwidth (may be different from h ; the importance of this will become clear for the second approach) and $\hat{m}_g^*(x, h) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) Y_i^*$, with $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$.

Let us look at the first approach (Härdle, Bowman JASA, 88).

$$\begin{aligned} E_* \hat{m}_g^*(x, h) &= E_* \left(\frac{1}{n} \sum_{i=1}^n K_h(x - x_i) \left(\frac{1}{n} \sum_{j=1}^n K_g(x_i - x_j) Y_j + \varepsilon_i^* \right) \right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n K_h(x - x_i) K_g(x_i - x_j) Y_j. \end{aligned}$$

Now, for j fixed,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) K_g(x_i - x_j) &\simeq \int_0^1 K_h(x - u) K_g(u - x_j) du \\ &= \int_0^0 K_h(v) K_g(x - x_j - v) dv \\ &=: \tilde{K}(x - x_j) \end{aligned}$$

and it is possible to get $\tilde{K}(x - x_j)$ analytically.

Now, $E_* \hat{m}_g^*(x, h) \simeq \hat{m}_{g,c}(x, h) := \frac{1}{n} \sum_{j=1}^n \tilde{K}(x - x_j) Y_j$. So we can get analytically a very good approximation to $E_* \hat{m}_g^*(x, h)$.

But, of course, $\mathcal{L}_*(\sqrt{nh}(\hat{m}_g^*(x, h) - \hat{m}_{g,c}(x, h)))$ is not a good approximation of $\mathcal{L}(\sqrt{nh}(\hat{m}(x, h) - m(x)))$ since $E(\sqrt{nh}(\hat{m}(x, h) - m(x))) \rightarrow c \frac{m''(x)}{2} \int u^2 K(u) du$ for $h \sim n^{-1/5}$.

Now, we can find a consistent estimator of $m''(x)$, say $\hat{m}_{\text{cons}}''(x)$.

It is then possible to prove that

$$\forall x, \tilde{\rho}_2(\sqrt{nh}(\hat{m}(x, h) - m(x)), \sqrt{nh}(\hat{m}_g^*(x, h) - \hat{m}_{g,c}(x, h) + \frac{h^2}{2} \hat{m}_{\text{cons}}''(x) \int u^2 K(u) du)) \xrightarrow{p} 0$$

under some conditions ($m \in C^2, h \sim n^{-1/5}, g \sim n^{-1/5} \dots$)

The second approach (Franke, Härdle, Ann. Stat. 1992) takes the following point of view: remember that $Y_i^* = \hat{m}(x_i, g) + \varepsilon_i^*$. Now g is taken to be s.t. $\frac{h}{g} \rightarrow 0$ (i.e. g goes slower to 0 than $n^{-1/5}$).

Then, $\text{var}(\hat{m}''(x, g)) \rightarrow 0$ and we do not have a bias problem any more: the bootstrap bias is close to the true bias.

If $\frac{h}{g} \rightarrow 0$ and under some more conditions, then it is possible to show that

$$\forall x, \tilde{\rho}_2(\sqrt{nh}(\hat{m}_g^*(x, h) - \hat{m}(x, g)), \sqrt{nh}(\hat{m}(x, h) - m(x))) \xrightarrow{p} 0.$$

If our model allows for heteroskedasticity:

$$Y_j = m(x_j) + \varepsilon_j \quad \varepsilon_1 \dots \varepsilon_n \text{ independent with mean } 0,$$

$$m : [0, 1] \rightarrow \mathbb{R}, \quad x_j = \frac{j}{n} \quad j = 1, \dots, n,$$

then the wild bootstrap can be very useful (see the idea of wild bootstrap in chapter 6 - more in Mammen (1992): "When does bootstrap work?")

8. Application to financial time series

As already briefly mentioned at the end of chapter 2 the usual bootstrap typically does not work for time series sequences since X_1, \dots, X_n are then not i.i.d..

Let us first assume our process $\{X_t\}$ is strictly stationary with $\text{var}X_t < \infty$.

We would like to approximate $\mathcal{L}(\sqrt{n}(\bar{X}_n - \mu))$. There are some possible bootstrap techniques to deal with this: we mention two of the simplest ones.

First the moving block bootstrap (Künsch): by means of our data X_1, \dots, X_n we can construct $n - b + 1$ overlapping blocks of b successive observations; i.e. $B_1 = \{X_1, \dots, X_b\}, B_2 = \{X_2, \dots, X_{b+1}\}, \dots, B_{n-b+1} = \{X_{n-b+1}, \dots, X_n\}$.

Now, let us assume $\frac{n}{b}$ is an integer, then we draw $\frac{n}{b}$ blocks with replacement and therefore obtain a new time series of length n : X_1^*, \dots, X_n^* . Notice that, e.g., X_1^*, \dots, X_b^* are the ordered elements of a same block.

We get a consistency result. If $\{X_t\}$ is α -mixing with $\alpha(n) \downarrow 0$ sufficiently fast and if other conditions are satisfied then

$$\sup_x |P_*(\sqrt{n}(\bar{X}_n^* - E_*\bar{X}_n^*) \leq x) - P(\sqrt{n}(\bar{X}_n - \mu) \leq x)| \xrightarrow{p} 0,$$

as $n \rightarrow \infty, b \rightarrow \infty, \frac{b}{n} \rightarrow 0$.

Notice that we need $\bar{X}_n^* - E_*\bar{X}_n^*$ and not $\bar{X}_n^* - \bar{X}_n$ since $E_*\bar{X}_n^* \neq \bar{X}_n$ and $E\bar{X}_n = \mu$.

Recall: A strictly stationary process X_t is α -mixing if

$$\sup_{A, B} |P(A \cap B) - P(A)P(B)| \rightarrow 0 \text{ as } k \rightarrow \infty$$

with $A \in \mathcal{F}_{-\infty}^0 = \sigma(X_0, X_{-1}, \dots), B \in \mathcal{F}_k^\infty = \sigma(X_k, X_{k+1}, \dots)$.

The next technique is a slight modification of moving block bootstrap and allows us to look at $\bar{X}_n^* - \bar{X}_n$: in the circular block bootstrap to the $n - b + 1$ blocks $B_1 \dots B_{n-b+1}$ we add $B_{n-b+2} = \{X_{n-b+2}, \dots, X_n, X_1\} \dots B_n = \{X_n, X_1, X_2, \dots\}$ and so $E_*\bar{X}_n^* = \bar{X}_n$.

The rest is similar to the previous technique.

Up to now we did not assume any structure in our time series but often (e.g. in financial time series) we have a model of the form $X_t = m(X_{t-1}) + \sigma(X_{t-1}) \varepsilon_t, \varepsilon_t$ i.i.d. $(0, 1)$ with density f_ε , with m the so-called trend function and σ the volatility function ($\text{var}(X_t|X_{t-1})$ is typical not a const. when $|X_{t-1}|$ is high, the market is more volatile at time t) and our goal is to estimate

the distribution of estimators of m and σ . $\{X_t\}$ is assumed to satisfy some mixing properties. In particular, it is strictly stationary.

m and σ can be estimated in many ways. We focus here on kernel estimators; first we need to estimate the density function of X_t :

$$\begin{aligned}\hat{f}(x, h) &= \frac{1}{n-1} \sum_{i=1}^{n-1} K_h(x - X_i). \text{ Then,} \\ \hat{m}(x, h) &= \frac{1}{n-1} \sum_{i=1}^{n-1} K_h(x - X_i) X_{i+1} / \hat{f}(x, h) \\ \hat{\sigma}^2(x, h') &= \frac{1}{n-1} \sum_{i=1}^{n-1} K'_h(x - X_i) (X_{i+1} - \hat{m}(x_i, h))^2 / \hat{f}(x, h')\end{aligned}$$

h and h' are typically not the same but in the following, to simplify the notations, we will simply write h for both h and h' .

Under general conditions, these estimators are asymptotically normally distributed.

We are interested in $\mathcal{L}(\sqrt{nh}(\hat{m}(x, h) - m(x)))$ and $\mathcal{L}(\sqrt{nh}(\hat{\sigma}^2(x, h) - \sigma^2(x)))$ with $h = O(n^{-1/5})$. How to bootstrap in this case?

Let us first suppose we want to mimick the whole structure of the process (Autoregression bootstrap): we construct empirical residuals

$$e_j = \frac{X_j - \hat{m}(X_{j-1}, k)}{\hat{\sigma}(X_{j-1}, k)} \text{ with } k \sim n^{-\alpha} \quad \alpha \leq \frac{2}{15}$$

(if we would take $k = h = O(n^{-1/5})$ we could not prove consistency of the bootstrap).

We remove those \hat{e}_j corresponding to $|X_{j-1}|$ very large (\hat{m} and $\hat{\sigma}$ not reliable for $|X_{j-1}|$ very large) and we center the remaining residuals. For these centered remaining residuals we have an empirical distribution function, we smooth it slightly; let us denote \hat{F}_ε the smoothed empirical distribution function.

Now, we draw $\varepsilon_1^* \dots \varepsilon_n^*$ i.i.d. from \hat{F}_ε .

$X_t^* = \hat{m}(X_{t-1}^*, h) + \hat{\sigma}(X_{t-1}^*, h) \varepsilon_t^*$, $X_1^* = X_1$ would not work for the same reason as in the previous chapter (bias problem)

$X_t^* = \hat{m}(X_{t-1}^*, g) + \hat{\sigma}(X_{t-1}^*, g) \varepsilon_t^*$, $X_1^* = X_1$, $\frac{g}{h} \rightarrow 0$ (more particularly $g \sim n^{-\alpha}$, $\alpha \leq \frac{2}{15}$) does not work either because there is a non-negligible probability that the process goes out of control (very large $|X_{t-1}^*|$, unreliable \hat{m} and $\hat{\sigma}$ and so X_t^* not reliable and so on ...).

We need to truncate \hat{m} and $\hat{\sigma}$: let \tilde{m} and $\tilde{\sigma}$ be equal to \hat{m} and $\hat{\sigma}$ on a increasing compact (as $n \rightarrow \infty$) and equal to 0 (for \tilde{m}) and 1 (for $\tilde{\sigma}$) outside this compact

$$X_t^* = \tilde{m}(X_{t-1}^*, g) + \tilde{\sigma}(X_{t-1}^*, g) \varepsilon_t^*, X_1^* = X_1$$

(if $|X_{t-1}^*|$ very large X_t^* may be very different from what $m(X_{t-1}^*) + \sigma(X_{t-1}^*) \varepsilon_t$ is but then at time $t + 1$ everything is again in order.)

Now, Franke, Kreiss, Mammen (Bernoulli(2002)) have proven that, under some conditions, $\forall x$

$$\rho_\infty(\sqrt{nh}(\hat{m}_g^*(x, h) - \hat{m}(x, g)), \sqrt{nh}(\hat{m}(x, h) - m(x))) \xrightarrow{p} 0$$

with $\hat{m}_g(x, h) = \frac{1}{n-1} \sum_{i=1}^{n-1} K_h(x - X_i^*) X_{i+1}^* / \hat{f}(x, h)$. There is a similar result for the volatility function.

Now, we do not need to mimick the whole structure of the process to have a good approximation of $\mathcal{L}(\sqrt{nh}(\hat{m}(x, h) - m(x)))$ and $\mathcal{L}(\sqrt{nh}(\hat{\sigma}^2(x, h) - \sigma^2(x)))$ as shows the regression bootstrap: we take $\varepsilon_1^* \dots \varepsilon_n^*$ i.i.d. from \hat{F}_ε as for the previous method.

Then, $X_t^* = \hat{m}(X_{t-1}, g) + \hat{\sigma}(X_{t-1}, g) \varepsilon_t^*$. Notice that we do not need here to truncate \hat{m} and $\hat{\sigma}$ (we do not base X_{t+1}^* on X_t^* but on X_t).

Under some conditions we get as for the previous method consistency of the bootstrap approximations. (Obviously $\hat{m}_g^*(x, g) = \frac{1}{n-1} \sum_{i=1}^{n-1} K_h(x - X_i) X_{i+1}^* / \hat{f}(x, h) \dots$) (Franke, Kreiss, Mammen (2002)).

The third method, the wild bootstrap, focuses on $\mathcal{L}(\sqrt{nh}(\hat{m}(x, h) - m(x)) : \eta_1^* \dots \eta_n^*)$ are taken i.i.d. from a distribution with mean 0 and variance 1. Then $\varepsilon_t^* = (X_t - \hat{m}(X_{t-1}, h))\eta_t^*$ and $X_t^* = \hat{m}(X_{t-1}, g) + \varepsilon_t^*$.

Then, under some conditions, we get consistency of the bootstrap approximation (Franke, Kreiss, Mammen (2002)).

Now, the previous results allow us to construct confidence intervals for $m(x)$ and $\sigma^2(x)$. The next step is to construct simultaneous confidence bands for m and σ^2 . Franke, Neumann, Stockis (Journal of Econometrics (2004)) deal with this problem for the autoregressive bootstrap. The main result allowing to get these bands is that, under some conditions, the stationary distribution of X_t is uniformly approximated by the stationary distribution of X_t^* , with a rate equal to some power of T^{-1} . See Franke, Neumann, Stockis (2004) for more details.