Bootstrap for Dependent Data

Barbora Šáchová, Alena Ulmanová

18. května 2026

Motivation

- Classical bootstrap assumes independent and identically distributed observations.
- Time series and dependent data violate the IID assumption.
- Resampling individual observations destroys dependence structure.
- Block bootstrap methods preserve local dependence by resampling blocks instead of single observations.
- One of the most important approaches is the **Moving Block Bootstrap (MBB)**.

Why IID Bootstrap Fails for Dependent Data

Suppose we observe a stationary time series:

$$X_1, X_2, \ldots, X_n$$

The IID bootstrap resamples observations independently:

$$X_1^*, X_2^*, \ldots, X_n^*$$

This destroys serial correlation:

$$\mathrm{Cov}(X_t, X_{t+h}) \neq \mathrm{Cov}(X_t^*, X_{t+h}^*)$$

Corollary

Suppose $\{X_n\}_{n\geq 1}$ is a sequence of stationary m-dependent random variables with $\mathbb{E}X_1 = \mu$ and $\sigma^2 = \text{var}(X_1) \in (0, \infty)$. If $\sum_{i=1}^{m} \text{cov}(X_1, X_{1+i}) \neq 0$ and $\sigma_\infty^2 = \sigma^2 + 2\sum_{i=1}^{\infty} \text{cov}(X_1, X_{1+i}) \neq 0$, then

$$\lim_{n\to\infty} [P_*(T_n^* \leq x) - P(T_n \leq x)] = [\Phi(x/\sigma) - \Phi(x/\sigma_\infty)] \neq 0 \text{ a.s.}$$

Basic Idea of the Moving Block Bootstrap

- Divide the observed time series into overlapping blocks.
- Each block preserves local dependence.
- Resample blocks with replacement.
- Join selected blocks to get bootstrap samples.

Construction of the Bootstrap Sample

Choose: $b = \lfloor \frac{n}{\ell} \rfloor$, where:

- ℓ = block length
- b = number of blocks sampled

Procedure:

1. Randomly sample b blocks with replacement.
2. Concatenate sampled blocks.
3. Form bootstrap series:

$$X_1^*, X_2^*, \ldots, X_{b\ell}^*$$

The resulting sample approximately preserves dependence structure.

Key intuition:

- Dependence is mainly local.
- Overlapping blocks capture short-range dependence.
- Resampled blocks mimic the original process.

In MBB, we can generalize $\hat{\theta}_n = T(F_n)$ by defining

$$F_{p,n} = (n-p+1)^{-1} \sum_{j=0}^{n-p+1} \delta_{Y_j},$$

where $Y_j = (X_j, ..., X_{j+p-1})$. We then define $\hat{\theta}_n = T(F_{p,n})$.

MBB Estimator

- Fix block size ℓ such that $1 < \ell < n - p + 1$
- $\tilde{B}_j = (Y_j, \ldots, Y_{j+\ell-1}),\ 1 \le j \le n - p - \ell + 2$

- Randomly sample k blocks with replacement from $\{\widetilde{B}_1, \ldots, \widetilde{B}_{n-p-\ell+2}\}$ to generate the bootstrap observations $Y_1^*, Y_2^*, \ldots, Y_m^*,\ m = k\ell$.

- Define the bootstrap empirical distribution

$$\widetilde{F}_{m,n}^* = \frac{1}{m} \sum_{j=1}^{m} \delta_{Y_j^*},$$

- The Moving Block Bootstrap estimator is

$$\theta_{m,n}^* = T(\widetilde{F}_{m,n}^*).$$

- Small ℓ:
  - Weak dependence preservation
  - High bias
- Large ℓ:
  - Better dependence capture
  - Higher variance

In general, we need $\ell \to \infty$ and $\frac{\ell}{n} \to 0$ as $n \to \infty$.
Popular choice: $\ell \propto n^{1/3}$.
Alternative methods:

- Plug-in estimators
- Cross-validation
- Minimum MSE selection

Non-Overlapping Block Bootstrap (NBB)

- Define the non-overlapping blocks

$$B_1 = (X_1, \ldots, X_\ell), \quad B_2 = (X_{\ell+1}, \ldots, X_{2\ell}), \quad \ldots$$

for

$$j = 1, \ldots, b, \qquad b = \left\lfloor \frac{n}{\ell} \right\rfloor .$$

- Sample b blocks independently with replacement from

$$\{B_1, \ldots, B_b\}.$$

- Joint the selected blocks to obtain the bootstrap series

$$X_1^*, \ldots, X_{b\ell}^*.$$

MBB vs Nonoverlapping Block Bootstrap

Advantages of MBB:

- Uses more blocks
- Better finite-sample performance
- Improved variance estimation

Disadvantage:

- More computationally intensive

Generalized Block Bootstrap (GBB)

- Define new series $\{Y_{n,i}\}_{l\geq 1}$ by periodic extention as $Y_{n,i} = X_{j_i}$ where $i = j_i$ (modulo n).
- Define the blocks $B(i,j) = (Y_{n,i}, \ldots, Y_{i+j-1})$ for $i \geq 1,\ j \geq 1$.
- Let Γ_n be a transition probability function on the set $\mathbb{R}^n \times \bigotimes_{t=1}^{\infty}(\{1,\ldots,n\} \times \mathbb{N})$, i.e., for each $x \in \mathbb{R}^n$, $\Gamma_n(x;\cdot)$ is a probability measure on

$$\bigotimes_{t=1}^{\infty}(\{1,\ldots,n\} \times \mathbb{N}) \equiv \{\{i_t, l_t\}_{t=1}^{\infty} : 1 \leq i_t \leq n, 1 \leq l_t < \infty \text{ for all } t \geq 1\}$$

  and for any set $A \subset \bigotimes_{t=1}^{\infty}(\{1,\ldots,n\} \times \mathbb{N})$, $\Gamma_n(\cdot; A)$ is a Borel measurable function from $\mathbb{R}^n$ into $[0,1]$.
- GBB resamples blocks from $\{B(i,j) : i \geq 1, j \geq 1\}$ as $B(I_1,J_1), B(I_2,J_2), \ldots$, where $(I_1,J_1), (I_2,J_2), \ldots$ is a sequence of random vectors with conditional joint distribution $\Gamma_n(\mathbb{X}_n;\cdot)$, given $\mathbb{X}_n$
- Let $X^*_{G,1}, X^*_{G,2}, \ldots$ denote the elements of these resampled blocks. Then, $\theta^{*(G)}_{m,n} = T(F^{*(G)}_{m,n})$ for a suitable choice of $m \geq 1$.

Circular Block Bootstrap (CBB)

- For $i = 1, \ldots, n$, define
$$B_i = (X_i, X_{i+1}, \ldots, X_{i+\ell-1}),$$
where indices are taken modulo n, i.e.
$$X_{n+j} \equiv X_j.$$

- **Bootstrap procedure:**
  - Sample blocks $B_{I_1}, \ldots, B_{I_b}$ with replacement from $\{B_1, \ldots, B_n\}$,
  - Join the sampled blocks to form $X_1^*, \ldots, X_{b\ell}^*$.
- Eliminates boundary effects present in MBB and NBB.
- Ensures every index has the same role in block formation.

- Let $X_1, \ldots, X_n$ be a stationary time series. Fix a parameter $p \in (0, 1)$ and define a geometric distribution:

$$P(L = k) = (1 - p)^{k-1} p, \qquad k \in \mathbb{N}.$$

- **Block construction:** Choose a starting index $I_1 \sim \text{Uniform}\{1, \ldots, n\}$. Generate a block of random length L_1:

$$B_1 = (X_{I_1}, X_{I_1+1}, \ldots, X_{I_1+L_1-1}),$$

with indices taken modulo n (circular extension).

Stationary Block Bootstrap (SBB)

- **Bootstrap sample generation:**
  - Generate i.i.d. pairs (I_j, L_j), where

$$I_j \sim \text{Uniform}\{1, \ldots, n\}, \quad L_j \sim \text{Geom}(p),$$

  - Form blocks

$$B_j = (X_{I_j}, \ldots, X_{I_j+L_j-1}),$$

  - Concatenate blocks until at least n observations are obtained,
  - Truncate to obtain $X_1^*, \ldots, X_n^*$.
- $(X_t^*)_{t\geq 1}$ is strictly stationary conditional on the data.
- Random block lengths remove fixed segmentation effects and improve preservation of dependence structure compared to fixed-length block bootstrap methods.

Smooth Function Model

- $\{X_{0i}\}_{i\in\mathbb{Z}}$ is a $\mathbb{R}^{d_0}$-valued stationary process
- Borel measurable function $f : \mathbb{R}^{d_0} \to \mathbb{R}^d$, smooth function $H : \mathbb{R}^d \to \mathbb{R}$
- Parameter of interest is $\theta = H(Ef(X_{0i}))$, $\widehat{\theta}_n = H\left(\frac{1}{n}\sum_{i=1}^n f(X_{0i})\right)$
- $T_{1n} = \sqrt{n}\left(\widehat{\theta}_n - \theta\right)$
- $X_i = f(X_{0i}), \quad i \in \mathbb{Z}, \qquad \boldsymbol{X}_n = (X_1, \ldots, X_n)$
- $\boldsymbol{X}_n^*$ is the set of n_1 bootstrap samples based on b blocks of length ℓ from $\boldsymbol{X_n}$
- $\theta_n^* = H\left(\overline{X}_n^*\right), \quad \widetilde{\theta}_n = H\left(E_*\overline{X}_n^*\right)$
- Bootstrap version: $T_{1n}^* = \sqrt{n_1}\left(\theta_n^* - \widetilde{\theta}_n\right)$

Theorem

Suppose that the function H is differentiable in some neighborhood N_H of EX_1, $\sum_{|\alpha|=1} |D^\alpha H(EX_1)| \neq 0$, and that the first-order partial derivatives of H satisfy a Lipschitz condition of order $\kappa > 0$ on N_H. Then

$$\sup_{x \in \mathbb{R}} |P_*(T_{1n}^* \leq x) - P(T_{1n} \leq x)| \xrightarrow{p} 0, \quad n \to \infty$$

Example

- $\{X_{0i}\}_{i\in\mathbb{Z}}$ a stationary time series with autocovariance function

$$\theta = \gamma(k) = Cov(X_{0i}, X_{0(i+k)}), \quad i, k \in \mathbb{Z}$$

- $$\widehat{\theta}_n = \widehat{\gamma}_n(k) = \frac{1}{n-k}\sum_{i=1}^{n-k} X_{0i}X_{0(i+k)} - \overline{X}^2_{0(n-k)}, \qquad \overline{X}_{0(n-k)} = \frac{1}{n-k}\sum_{i=1}^{n-k} X_{0i}$$

- $T_{1n} = \sqrt{n-k}\left(\widehat{\theta}_n - \theta\right) = \sqrt{n-k}\,(\widehat{\gamma}_n(k) - \gamma(k))$
- MSE of $\widehat{\gamma}_n(2)$:

$$\varphi_n = ET^2_{1n} = (n-2)MSE\,(\widehat{\gamma}_n(2))$$

Example

$$X_{0i} - 0.4X_{0(i-1)} - 0.2X_{0(i-2)} - 0.1X_{0(i-3)} = \varepsilon_i + 0.2\varepsilon_{i-1} + 0.3\varepsilon_{i-2} + 0.2\varepsilon_{i-3} + 0.1\varepsilon_{i-4}$$

- $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ iid N(0,1), $n = 102$

Block Size	4	6	8	10	15	20
MBB	1.159	1.085	0.881	0.820	1.078	0.884
NBB	1.299	0.904	1.093	0.763	0.879	1.030
CBB	1.020	1.106	0.951	0.812	0.968	0.808
SB	0.935	0.941	0.898	0.810	0.746	0.642

Table: Block bootstrap estimates of φ_n based on different block sizes. True value of φ_n is given by 1.058. $B = 800$.

- $\{X_i\}_{i\in\mathbb{Z}}$ is a stacionary process taking values in $\mathbb{R}^d$
- Parameter of interest θ is a solution to the equation

$$E\Psi(X_1,\ldots,X_m,\theta)=0$$

- M-estimator $\widehat{\theta}_n$ of θ is a solution to the estimating equation

$$\frac{1}{n-m+1}\sum_{i=1}^{n-m+1}\Psi(X_i,\ldots,X_{i+m-1},\widehat{\theta}_n)=0$$

- $T_{2n}=\sqrt{n}\left(\widehat{\theta}_n-\theta\right)$

M-Estimators - bootstrap version

- $Y_i = (X_i^\top, \ldots, X_{i+m-1}^\top)^\top, \qquad i = 1, \ldots, n-m+1$
- $Y_1^*, \ldots, Y_{n+m-1}^*$ denote the block bootstrap sample of size $n+m-1$ drawn from Y
- Bootstrap version θ_n^* of $\widehat{\theta}_n$ is a solution to the equation

$$\frac{1}{n-m+1} \sum_{i=1}^{n-m+1} \Psi(Y_i^*, \theta_n^*) = 0$$

- Alternative bootstrap version θ_n^{**} of $\widehat{\theta}_n$ is a solution to the equation

$$\frac{1}{n-m+1} \sum_{i=1}^{n-m+1} \left(\Psi(Y_i^*, \theta_n^{**}) - \frac{1}{n-m+1} E_* \left[\sum_{i=1}^{n-m+1} \Psi(Y_i^*, \widehat{\theta}_n) \right] \right) = 0$$

- $T_{2n}^* = \sqrt{n-m+1}(\theta_n^* - \widetilde{\theta}_n), \qquad T_{2n}^{**} = \sqrt{n-m+1}(\theta_n^{**} - \widehat{\theta}_n)$

Theorem

Assume that $\Psi(y,t)$ is differentiable with respect to t for almost all y and the first-order partial derivatives of Ψ satisfy a Lipschitz condition of order $\kappa \in (0,1]$. Also, assume that $E\Psi(Y_i,\theta) = 0$ and Σ_Ψ and D_Ψ are regular. There exists a $\delta > 0$ such that $E||D^\alpha \Psi(Y_1,\theta)||^{2r+\delta} < \infty$ for all $\alpha \in \mathbb{Z}_+^s$ with $|\alpha| = 0,1$, and $\Delta(r,\delta) < \infty$. Assume that $\ell^{-1} + n^{-1/2}\ell = o(1)$ and $p + (n^2 p)^{-1} = o(1)$. Then,

- $\{\widehat{\theta}_n\}_{n\geq 1}$ is a consistent for θ and

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, D_\Psi \Sigma_\Psi D_\Psi^\top)$$

-

$$\sup_{x\in\mathbb{R}^s} |P_*(T_{2n}^* \leq x) - P(T_{2n} \leq x)| \xrightarrow{p} 0, \quad n \to \infty$$

Stationary AR processes

- $\{X_i\}_{i\in\mathbb{Z}}$ is a stationary autoregressive process of order p

$$AR(p) : X_i = \beta_1 X_{i-1} + \cdots + \beta_p X_{i-p} + \varepsilon_i, \quad i \in \mathbb{Z}$$

- Infinite-order moving average representation :

$$X_i = \sum_{j=0}^{\infty} b_j \varepsilon_{i-j}$$

- Suppose $X_1, \ldots, X_n$ is observed
- Let $\widehat{\beta}_{1n}, \ldots, \widehat{\beta}_{pn}$ denote the least squares estimators of $\beta_1, \ldots, \beta_p$ based on $X_1, \ldots, X_n$

- $\widehat{\varepsilon}_i = X_i - \widehat{\beta}_{1n} X_{i-1} - \ldots - \widehat{\beta}_{pn} X_{i-p}, \qquad i = p+1, \ldots, n$
- $\widetilde{\varepsilon}_i = \widehat{\varepsilon}_i - \overline{\varepsilon}_n, \qquad \overline{\varepsilon}_n = \frac{1}{n-p} \sum_{i=p+1}^{n} \widehat{\varepsilon}_i$
- Generate the bootstrap error variables ε_i^* by sampling randomly with replacement from $\{\widetilde{\varepsilon}_{p+1}, \ldots, \widetilde{\varepsilon}_n\}$
- $X_i^* = \widehat{\beta}_{1n} X_{i-1}^* + \cdots + \widehat{\beta}_{pn} X_{i-p}^* + \varepsilon_i^*, \quad i \in \mathbb{Z}$
- Let $\{X_i^*\}_{i \in \mathbb{Z}}$ be a stationary solution
- The autoregressive bootstrap (ARB) of $T_n = t_n(X_1, \ldots, X_n, \beta_1, \ldots, \beta_p, F)$ is given by

$$T_n^* = t_n(X_1^*, \ldots, X_n^*, \widehat{\beta}_{1n}, \ldots, \widehat{\beta}_{pn}, \widehat{F}_n)$$

- Let β_{jn}^* denote the bootstrap version of $\widehat{\beta}_{jn}$

Theorem

Assume that $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ is a sequence of iid random variables such that $E\varepsilon_1 = 0$, $E\varepsilon_1^2 = 1, E\varepsilon_1^8 < \infty$ and

$$\limsup_{||t||\to\infty} |E\exp(i(\varepsilon_1, \varepsilon_1^2)t)| < 1.$$

Also, suppose that all roots of the characteristic polynomial $z^p - \beta_1 z^{p-1} - \ldots - \beta_p$ lie inside the unit circle. Then

$$\sup_{x\in\mathbb{R}^p} \left| P_* \left(\sqrt{n}\widehat{\Sigma}_n^{1/2}(\beta_n^* - \widehat{\beta}_n) \le x \right) - P\left(\sqrt{n}\Sigma^{1/2}(\widehat{\beta}_n - \beta) \le x \right) \right| = o\left(\frac{1}{\sqrt{n}}\right) \quad a.s.$$

ARB Example

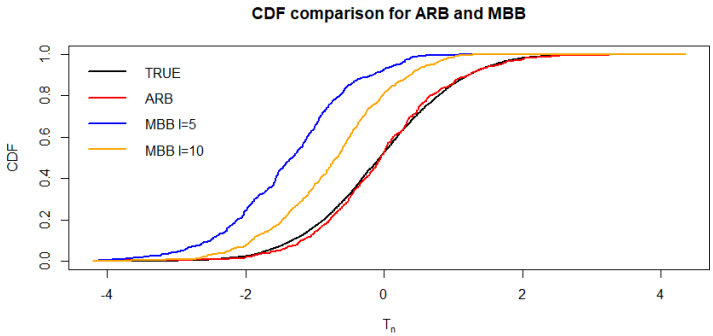
$AR(1) : X_i = 0.5X_{i-1} + \varepsilon_i, \qquad \varepsilon_i \sim N(0,1)$ iid

- n=100
- We want to approximate the sampling distribution of

$$T_n = \left(\sum_{t=1}^{n-1} X_t^2 \right)^{1/2} \left(\widehat{\beta}_{1n} - \beta_1 \right)$$

- ARB version T_n^* based on B=500 bootstrap replicates
- True distribution of T_n using 10 000 simulation runs

$AR(1) : X_i = 0.5X_{i-1} + \varepsilon_i$



ARB Example

$AR(1) : X_i = 0.5X_{i-1} + \varepsilon_i$

Method	CI lower	CI upper
True	0.4639	0.7456
ARB	0.4717	0.7423
MBB $l = 5$	0.3511	0.6137
MBB $l = 10$	0.4092	0.6594

Table: 90% confidence intervals for β_1 under different methods

- $\{X_i\}_{i\in\mathbb{Z}}$ is a stationary $ARMA(p,q)$ process

$$X_i = \beta_1 X_{i-1} + \cdots + \beta_p X_{i-p} + \varepsilon_i + \alpha_1 \varepsilon_{i-1} + \cdots + \alpha_q \varepsilon_{i-q}, \quad i \in \mathbb{Z}$$

- $$\varepsilon_i = \sum_{j=1}^{i} a_{j-1} \left(- \sum_{k=0}^{p} \beta_k X_{i+1-j-k} \right) + \sum_{s=0}^{q-1} \varepsilon_{-s} \left(\sum_{k=0}^{s} a_{i+1+s-k} \alpha_k \right), \qquad i \geq 1-q$$

- Suppose $X_{1-p}, \ldots, X_n$ is observed
- Let $\widehat{\beta}_{1n}, \ldots, \widehat{\beta}_{pn}, \widehat{\alpha}_{1n}, \ldots, \widehat{\alpha}_{qn}$ denote the estimators of the parameters based on $X_{1-p}, \ldots, X_n$ such that

$$\sum_{j=1}^{p} \left| \widehat{\beta}_{jn} - \beta_j \right| + \sum_{j=1}^{q} \left| \widehat{\alpha}_{jn} - \alpha_j \right| \xrightarrow{p} 0, \quad n \to \infty$$

- $$\widehat{\varepsilon}_i = \sum_{j=1}^{i} \widehat{a}_{j-1,n} \left(- \sum_{k=0}^{p} \widehat{\beta}_{kn} X_{i+1-j-k} \right), \qquad i = 1, \ldots, n$$
- $\widetilde{\varepsilon}_i = \widehat{\varepsilon}_i - \overline{\varepsilon}_n, \qquad \overline{\varepsilon}_n = \frac{1}{n} \sum_{i=1}^{n} \widehat{\varepsilon}_i$
- Generate iid bootstrap error variables ε_i^* by sampling randomly with replacement from $\{\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_n\}$
- $X_i^* = \widehat{\beta}_{1n} X_{i-1}^* + \cdots + \widehat{\beta}_{pn} X_{i-p}^* + \varepsilon_i^* + \widehat{\alpha}_{1n} \varepsilon_{i-1}^* + \cdots + \widehat{\alpha}_{qn} \varepsilon_{i-q}^*, \quad i \geq 1 - \max(p, q)$
- $X_i^* = 0, \varepsilon_i^* = 0, \quad i \leq - \max(p, q)$
- $T_n^* = t_n(X_{1-p}^*, \ldots, X_n^*, \widehat{\beta}_{1n}, \ldots, \widehat{\beta}_{pn}, \widehat{\alpha}_{1n}, \ldots, \widehat{\alpha}_{qn}, \widehat{F}_n)$

Thank you for your attention!