# Block Bootstrap Methods

Agnieszka Tracz

October 2023

## 1 Moving Block Bootstrap

### 1.1 Introduction

The Moving Block Bootstrap (MBB) is suitable for dependent data without making parametric model assumptions. MBB resamples consecutive blocks of observations not only one observation at a time as before and in this way it preserves the original observations dependence structure within each block.
Moreover, as the sample size increases, the block size also increases. Consequently, when dealing with data generated by a weakly dependent process, MBB asymptotically replicates the underlying dependence structure of the process.

### 1.2 Description of the MBB

We denote:
$X_1, X_2, ...$ - sequence of stationary random variables
$\mathrm{X}_n = \{X_1, ..., X_n\}$ - observations
$\hat{\theta_n} = T(F_n)$ - MBB version of estimators
$F_n$ - the empirical distribution function of $X_1, ..., X_n$
$T(\cdot)$ - (real-valued) functional of $F_n$.
$l \equiv l_n \in [1, n]$ - an integer
For dependent data, we typically require that:
$l \to \infty$ and $n^{-1}l \to 0$ as $n \to \infty$
However, a description of the MBB can be given without this restriction.

**The process of obtaining MBB samples:**

1. Take the observations $\{X_1, ..., X_n\}$ and choose the integer $l$ - length of blocks.

2. Place the observations in blocks $B_i = (X_i, ..., X_{i+l-1})$ , $1 \le i \le N$ where $N = n - l + 1$ (each block has length $l$)

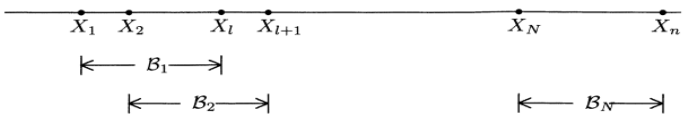



FIGURE 2.1. The collection $\{\mathcal{B}_1, \ldots, \mathcal{B}_N\}$ of overlapping blocks under the MBB.

3. Fix number $k$ and take $B_1^*, ..., B_k^*$ - simple random sample drawn with replacement from $\{B_1, ..., B_N\}$

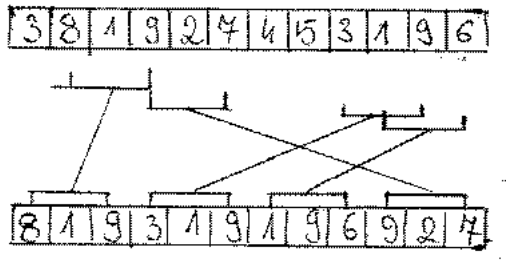
4. We get $\left(X^*_{(i-1)l+1}, ..., X^*_{il}\right)$ - elements in $B_i^*, i = 1, ..., k$

5. Then $X_1^*, ..., X_m^*$ constitute the MBB sample of size $m = kl$

The MBB version $\theta^*_{m,n}$ of $\hat{\theta_n}$ is defined as $\theta^*_{m,n} = T\left(F^*_{m,n}\right)$, where $F^*_{m,n}$ denotes the empirical distribution of $X_1^*, ..., X_m^*$

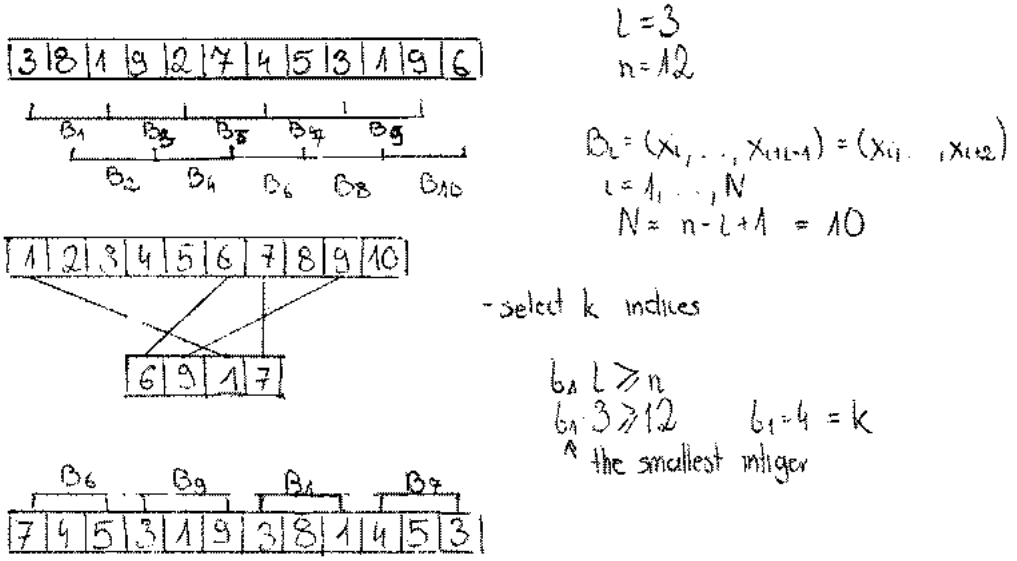
**MBB sample size is typically chosen to be of the same order as the original sample size.** If $b_1$ denotes the smallest integer such that $b_1 l \ge n$, then one may select $k = b_1$ blocks to generate the MBB samples,

and use only the first $n$ values to define the bootstrap version of $T_n$



## 1.3 An alternative formulation of the MBB

Selecting the blocks $B_i^*$'s randomly from $\{B_1, ..., B_N\}$ is equivalent to selecting $k$ indices at random from the set $\{1, ..., N\}$.
Let $I_1, ..., I_k$ be iid random variables with the *discrete uniform distribution* on $\{1, ..., N\}$. If we set $B_i^* = B_{I_i}$ for $i = 1, ..., k$, then $B_1^*, ..., B_k^*$ represent a simple random sample drawn with replacement from $\{B_1, ..., B_N\}$. The bootstrap sample $X_1^*, ..., X_m^*$ can be defined using the resampled blocks $B_1^*, ..., B_k^*$ as before.




`Note:` Conditional on the data $\mathrm{X}_n$, the resampled blocks of observations $(X_1^*, ..., X_l^*)', (X_{l+1}^*, ..., X_{2l}^*)', ..., (X_{(k-1)l+1}^*, ..., X_{kl}^*)'$ are iid $l$-dimensional random vectors with:

$$P_*((X_1^*, ..., X_l^*)' = (X_j, ..., X_{j+l-1})') = P_*(I_1 = j) = N^{-1}, \text{ for } 1 \le i \le N$$

$P_*$ denotes the conditional probability given $\mathrm{X}_n$.

## 1.4 General version of the MBB

### 1.4.1 Intorduction

Estimators of the form $\hat{\theta_n} = T(F_n)$ considered above include many commonly used estimators, e.g., the sample mean, but they are not sufficiently rich for applications in the time series context.
( $\hat{\theta_n}$ depends only on the one-dimensional marginal empirical distribution $F_n$, and hence does not cover standard statistics like the sample lag correlations)

Given the observations $\mathrm{X}_n$, let $F_{p,n}$ denote the $p$-dimensional empirical measure:

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

where $Y_j = (X_j, ..., X_{j+p-1})$ and where for any $y \in \mathbf{R}^p$, $\gamma_y$ denotes the probability measure on $\mathbf{R}^p$ putting unit mass on $y$.

The general version of the MBB concerns estimators of the form :

$$\hat{\theta_n} = T(F_{p,n}),$$

where $T(\cdot)$ is now a functional defined on a (rich) subset of the set of all probability measures on $\mathbf{R}^p$. Here, $p \geq 1$ may be a fixed integer, or it may tend to infinity with $n$ suitably.

### 1.4.2 "ordinary" approach of the MBB

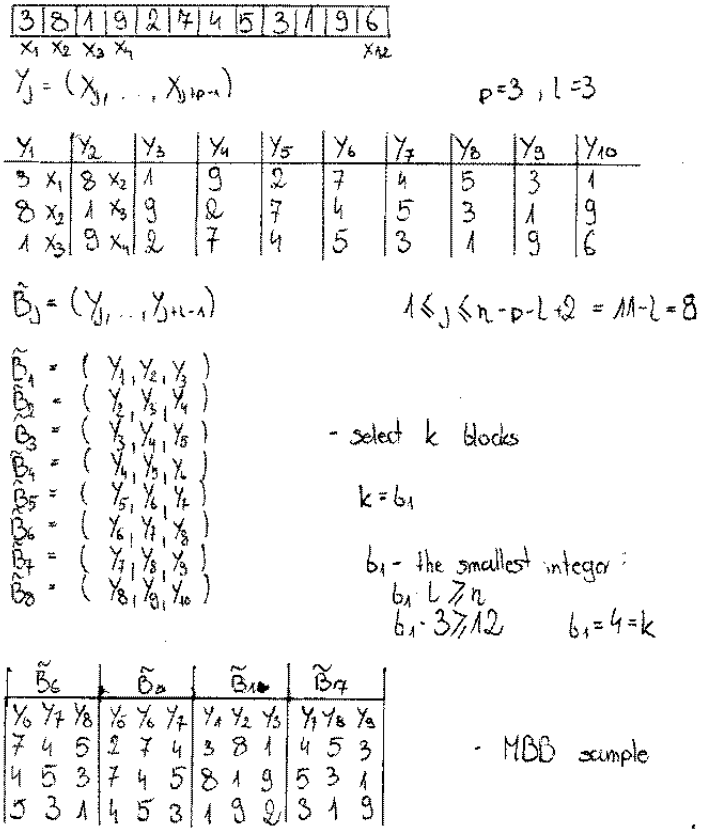
To define the "ordinary" MBB version we fix a block size $l$, $1 < l < n - p + 1$, and define the blocks in terms of $Y_i$'s as

$$\tilde{B}_j = (Y_j, ..., Y_{j+l-1}), 1 \leq j \leq n - p - l + 2.$$

For $k \geq 1$, select $k$ blocks randomly from the collection: $\tilde{B}_i : 1 \leq i \leq n - p - l + 2$, to generate the MBB observations: $Y_1^*, ..., Y_l^*, Y_{l+1}^*, ..., Y_m^*$, where $m = kl$.
The MBB version of $\hat{\theta_n} = T(F_{p,n})$ is now defined as

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

where $\tilde{F}_{m,n}^* \equiv m^{-1} \sum_{j=1}^m \gamma_{Y_j^*}$ denotes the empirical distribution of $Y_1^*, ..., Y_m^*$. Thus, for estimators of the form $\hat{\theta_n} = T(F_{p,n})$, the MBB version is defined by resampling from blocks of $Y$-values instead of blocks of $X$-values themselves.

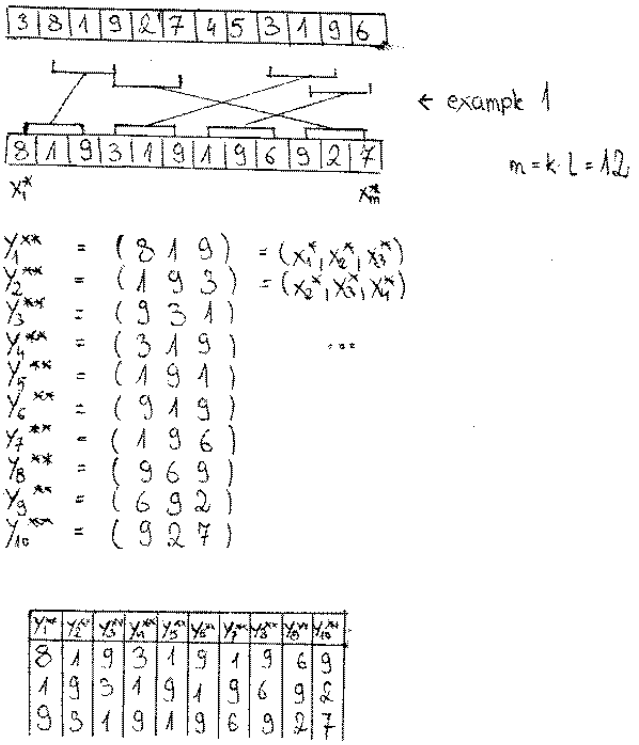



### 1.4.3 "naive" approach of the MBB

For the more general class of statistics $\hat{\theta_n}$ given by $\hat{\theta_n} = T(F_{p,n})$ for some $p \geq 2$, there is an alternative way of defining the bootstrap version of $\hat{\theta_n}$ .
Since the estimator $\hat{\theta_n}$ can always be expressed as a function of the given observations $X_1, ..., X_n$ , we define the bootstrap version of $\hat{\theta_n}$ by resampling from$X_1, ..., X_n$ directly.
Suppose that the block bootstrap observations $X_1^*, ..., X_m^*$ are generated by resampling from the blocks $B_i = (X_i, ..., X_{i+l-1})$, $i = 1, ..., N$ of X-values. Define bootstrap "analogs" of the $p$-dimensional variable $Y_i \equiv (X_i, ..., X_{i+p-1})'$ in terms of $X_1^*, ..., X_m^*$ as $Y_i^{**} \equiv (X_i^*, ..., X_{i+p-1}^*)'$, $i = 1, ..., m - p + 1$. Then, the bootstrap version of $\hat{\theta_n}$ under this alternative approach is defined as:

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

where $\tilde{F}^{**}_{m,n} = \sum_{i=1}^{m-p+1} \gamma_{Y_i^{**}}$.




| $Y_1^{**}$ | $Y_2^{**}$ | $Y_3^{**}$ | $Y_4^{**}$ | $Y_5^{**}$ | $Y_6^{**}$ | $Y_7^{**}$ | $Y_8^{**}$ | $Y_9^{**}$ | $Y_{10}^{**}$ |
|---|---|---|---|---|---|---|---|---|---|
| 8 | 1 | 9 | 3 | 1 | 9 | 1 | 9 | 6 | 9 |
| 1 | 9 | 3 | 1 | 9 | 1 | 9 | 6 | 9 | 2 |
| 9 | 3 | 1 | 9 | 1 | 9 | 6 | 9 | 2 | 7 |

## 2 Nonoverlapping Block Bootstrap

In this section, we consider the blocking rule due to Carlstein. For simplicity, here we consider estimators given by $\hat{\theta_n} = T(F_{p,n})$, with $p = 1$ only.
The key feature of Carlstein's blocking rule is to **use nonoverlapping segments of the data** to define the blocks. The corresponding block bootstrap method will be called the **nonoverlapping block bootstrap (NBB)**

Notation:
$l \equiv l_n \in [1, n]$ - an integer (length of blosks)
$b \geq 1$ - the largest integer satisfying: $lb \leq n$

**The process of obtaining NBB samples:**

1. Take the observations $\{X_1, ..., X_n\}$ and choose the integer $l$ - length of blocks.

2. Place the observations in blocks $B_i^{(2)} = \left(X_{(i-1)l+1}, ..., X_{il}\right)'$ , $1 \leq i \leq b$ (each block has length $l$)

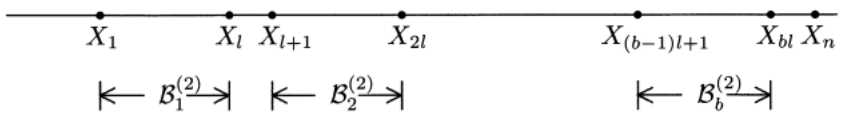



FIGURE 2.2. The collection $\{\mathcal{B}_1^{(2)}, \ldots, \mathcal{B}_b^{(2)}\}$ of nonoverlapping blocks under Carlstein's (1986) rule.

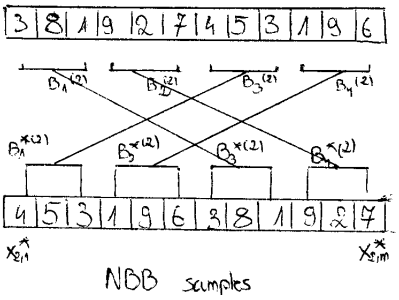
**Note:** Blocks in the MBB overlap, but the blocks $B_i^{(2)}$'s under the NBB do not. The collection of blocks from which the bootstrap blocks are selected is smaller than the collection for the MBB.

3. Calculate the number $k$ and take $B_1^{*(2)}, ..., B_k^{*(2)}$ - simple random sample drawn with replacement from $\{B_1^{(2)}, ..., B_b^{(2)}\}$

4. We get $\left(X^*_{(i-1)l+1}, ..., X^*_{il}\right)$ - elements in $B^*_i, i = 1, ...k$

5. Then $X^*_{2,1}, ..., X^*_{2,l}, ...X^*_{2,(b-1)l+1}, ..., X^*_{2,m}$ constitute the NBB sample of size $m = kl$

$F^{*(2)}_{m,n}$ denote the empirical distribution of the bootstrap sample $X^*_{2,1}, ..., X^*_{2,m}$, obtained by writing the elements of $B^{*(2)}_1, ..., B^{*(2)}_k$ in a sequence.
Then, the bootstrap version of estimator $\hat{\theta_n}$ is given by: $\theta^{*(2)}_{m,n} = T(F^{*(2)}_{m,n})$




Even though the definition of the bootstrapped estimators are very similar for the MBB and for the NBB, the resulting bootstrap versions $\theta^*_{m,n}$ and $\theta^{*(2)}_{m,n}$ have different distributional properties.

We illustrate the point with the simplest case, where $\hat{\theta_n} = n^{-1}\sum_{j=1}^n X_j$ is the sample mean. The bootstrap version of $\hat{\theta_n}$ under the two methods are respectively given by:

$$\theta^*_{m,n} = m^{-1}\sum_{j=1}^m X^*_j, \text{ and } \theta^{*(2)}_{m,n} = m^{-1}\sum_{j=1}^m X^*_{2,j}$$

$$\begin{aligned} E_*(\theta^*_{m,n}) &= E_*(\ell^{-1}\sum_{i=1}^{\ell} X^*_i) \\ &= N^{-1}\sum_{j=1}^{N}\left(\ell^{-1}\sum_{i=1}^{\ell} X_{j+i-1}\right) \\ &= N^{-1}\left\{n\bar{X}_n - \ell^{-1}\sum_{j=1}^{\ell-1}(\ell-j)(X_j + X_{n-j+1})\right\}. \end{aligned}$$

To obtain a similar expression for $E_*(\theta^{*(2)}_{m,n})$ , note that under the NBB, the bootstrap variables: $(X^*_{2,1}, ..., X^*_{2,l}), ..., (X_{2,(b-1)l+1}, ..., X^*_{2,m})$ are iid, with common distribution:

$$P_*((X^*_{2,1}, ...X^*_{2,l}) = (X_{(j-1)l+1}, ..., X_{jl})) = b^{-1},$$

for $j = 1, ...b$

$$\begin{aligned} E_*(\theta^{*(2)}_{m,n}) &= E_*(\ell^{-1}\sum_{i=1}^{\ell} X^*_{2,i}) \\ &= b^{-1}\sum_{j=1}^{b}\left(\ell^{-1}\sum_{i=1}^{\ell} X_{(j-1)\ell+i}\right) \\ &= (b\ell)^{-1}\left\{n\bar{X}_n - \sum_{i=b\ell+1}^{n} X_i\right\}, \end{aligned}$$

which equals $\bar{X_n}$ if $n$ is a multiple of $l$.

The bootstrapped estimators have different (conditional) means under the two methods.
If the process $\{X_n\}_{n\geq 1}$ satisfies some standard moment and mixing conditions, then $E\{E_*(\theta^*_{m,n}) - E_*(\theta^{*(2)}_{m,n})\}^2 = O(\frac{l}{n^2})$.
The difference between the two is negligible for large sample sizes.

$$\Theta^*_{min} = m^{-1} \sum_{j=1}^{m} X_j^*$$

$$E_*(\Theta^*_{min}) = E_*(m^{-1} \sum_{j=1}^{m} X_j^*) = E_*(l^{-1} \sum_{i=1}^{l} X_i^*) = N^{-1} \sum_{j=1}^{N} (l^{-1} \sum_{i=1}^{l} X_{j+i-1})$$

$$P_*((X_1^*, \ldots, X_l^*) = (X_j, \ldots, X_{j+l-1})) = P(I_1 = j) = N^{-1} \quad 1 \le j \le N$$

$$= \frac{1}{Nl} \sum_{j=1}^{N} \sum_{i=1}^{l} X_{j+i-1} = \frac{1}{Nl} \left( \sum_{i=1}^{l} X_i + \sum_{i=2}^{l+1} X_i + \ldots + \sum_{i=1}^{l} X_{i+N-1} \right) = \qquad N = n-l+1$$

$$= \frac{1}{Nl} \left( \sum_{i=1}^{l} X_i + \sum_{i=1}^{l} X_{i+1} + \ldots + \sum_{i=1}^{l} X_{i+n-l} \right) = \ldots$$

$$= \frac{1}{Nl} \left( X_1 + \ldots + X_l + X_2 + \ldots + X_{l+1} + \ldots + X_{n-l+1} + \ldots + X_n \right)$$

$$= \frac{1}{Nl} \begin{bmatrix} X_1 + X_2 + \ldots + X_l \\ X_2 + X_3 + \ldots + X_{l+1} \\ \ldots \\ \ldots + X_{n-1} \\ X_{n-l+1} + X_{n-l+2} + \ldots + X_n \end{bmatrix}$$

$$= \frac{1}{Nl} \left( \sum_{i=1}^{2} X_i - X_{n-l+2} - \ldots - X_n + \sum_{i=1}^{3} X_i - X_1 - X_{n-l+3} - \ldots - X_n + \sum_{i=1}^{n} X_i - X_1 - \ldots - X_{n-1} \right)$$

$$= \frac{1}{Nl} \left( l \sum_{i=1}^{n} X_i - (l-1)(X_1 + X_n) - (l-2)(X_2 + X_{n-1}) - \ldots - 1 \cdot (X_{l-1} + X_{n-l+2}) \right)$$

$$= \frac{1}{Nl} \left( l \cdot n \cdot \frac{1}{n} \sum_{i=1}^{n} X_i - \sum_{j=1}^{l-1} (l-j)(X_j + X_{n-j+1}) \right)$$

$$= \frac{1}{N} \left( n \cdot \bar{X}_n - \frac{1}{l} \sum_{j=1}^{l-1} (l-j)(X_j + X_{n-j+1}) \right)$$

$$\Theta^{*(2)}_{min} = m^{-1} \sum_{j=1}^{m} X_{2j}^*$$

Under the NBB, the bootstrap variables $(X_{21}^*, \ldots, X_{2l}^*), \ldots, (X_{2(m-1)l+1}^*, \ldots, X_{2ml}^*)$ are iid with common distribution:

$$P_*((X_{2l+1}^*, \ldots, X_{2l}^*) = (X_{(j-1)l+1}, \ldots, X_{jl})) = 1/b \qquad j = 1, \ldots, b$$

$$E_*(\Theta^{*(2)}_{min}) = E_*(m^{-1} \sum_{j=1}^{m} X_{2j}^*) = E_*(l^{-1} \sum_{i=1}^{l} X_{2i}^*) = b^{-1} \sum_{j=1}^{b} (l^{-1} \sum_{i=1}^{l} X_{(j-1)l+i}) =$$

$$= \frac{1}{bl} \sum_{j=1}^{b} \sum_{i=1}^{l} X_{(j-1)l+i} =$$

$$= \frac{1}{bl} \left( \sum_{i=1}^{l} X_i + \sum_{i=1}^{l} X_{i+l} + \ldots + \sum_{i=1}^{l} X_{i+(b-1)l} \right)$$

$$= \frac{1}{bl} \left( X_1 + X_2 + \ldots + X_l + X_{l+1} + \ldots + X_{2l} + \ldots + X_{1+(b-1)l} + \ldots + X_{bl} \right)$$

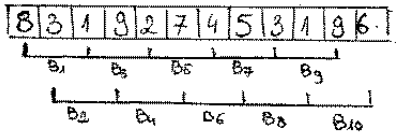
$$= \frac{1}{bl} \left( \sum_{i=1}^{bl} X_i \right) = \frac{1}{bl} \left( \sum_{i=1}^{n} X_i - \sum_{i=bl+1}^{n} X_i \right) =$$

$$= \frac{1}{bl} \left( n \bar{X}_n - \sum_{i=bl+1}^{n} X_i \right)$$

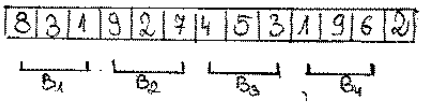
↑ equals $\bar{X}_n$ if $n$ is a multiple of $l$

# 3 Problems with MBB and NBB

The MBB resampling scheme suffers from an undesirable boundary effect as it assigns lesser weights to the observations toward the beginning and the end of the data set than to the middle part.
Indeed, for $l \le j \le n-l$, the $j$th observation $X_j$ appears in exactly $l$ of the blocks $\{B_1, ..., B_N\}$, whereas for $1 \le j \le l-1$, $X_j$ and $X_{n-j+1}$ appear only in $j$ blocks.
Since there is no observation beyond $X_n$ (or prior to $X_1$), we cannot define new blocks to get rid of this boundary effect.




A similar problem also exists under the NBB with the observations near the end of the data sequence when $n$ is not a multiple of $l$.



Politis and Romano suggested a simple way out of this boundary problem by **wrapping the data around a circle and forming additional blocks using the 'circularly defined' observations**. They put forward two resampling schemes based on circular blocks, called the 'circular block bootstrap' (CBB) and the 'stationary bootstrap' (SB).

## 3.1 Circular Block Bootstrap

The Circular Block Bootstrap (CBB) method resamples overlapping and periodically extended blocks of a given length $l$ satisfying $1 \leq l \leq n$.

**The process of obtaining CBB samples:**

1. Given the variables $\mathrm{X}_n = \{X_1, ..., X_n\}$, define a new time series $Y_{n,i},\ i \geq 1$ by **periodic extension**. (for any $i \geq 1$, there are integers $k_i \geq 0$ and $j_i \in [1, n]$ such that $i = k_i n + j_i$, then $i = j_i$ (modulo $n$).

Define the variables $Y_{n,i},\ i \geq 1$ by the relation:

$$Y_{n,i} = X_{j_i}$$

(This is equivalent to writing the variables $X_1, ..., X_n$ repeatedly on a line and labeling them serially as $Y_{n,i},\ i \geq 1$)

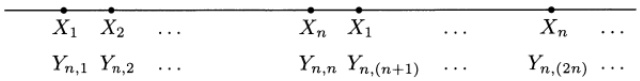



FIGURE 2.3. The periodically extended time series $Y_{n,i},\ i \geq 1$.

3. Define the blocks:

$$B(i, j) = (Y_{n,i}, ..., Y_{n,(i+j-1)}) \text{ for } i \geq 1,\ j \geq 1,$$

In this method we will use the parametr $j : j = l$

4. Then, the CBB resamples blocks from the collection $\{B(i, l) : i \geq 1, \}$

5. Let $(I_1, J_1), (I_2, J_2), ....$ be a sequence of random vectors with conditional joint distribution.
The blocks selected by the CBB are given by $B(I_1, J_1), B(I_2, J_2), ...$ .

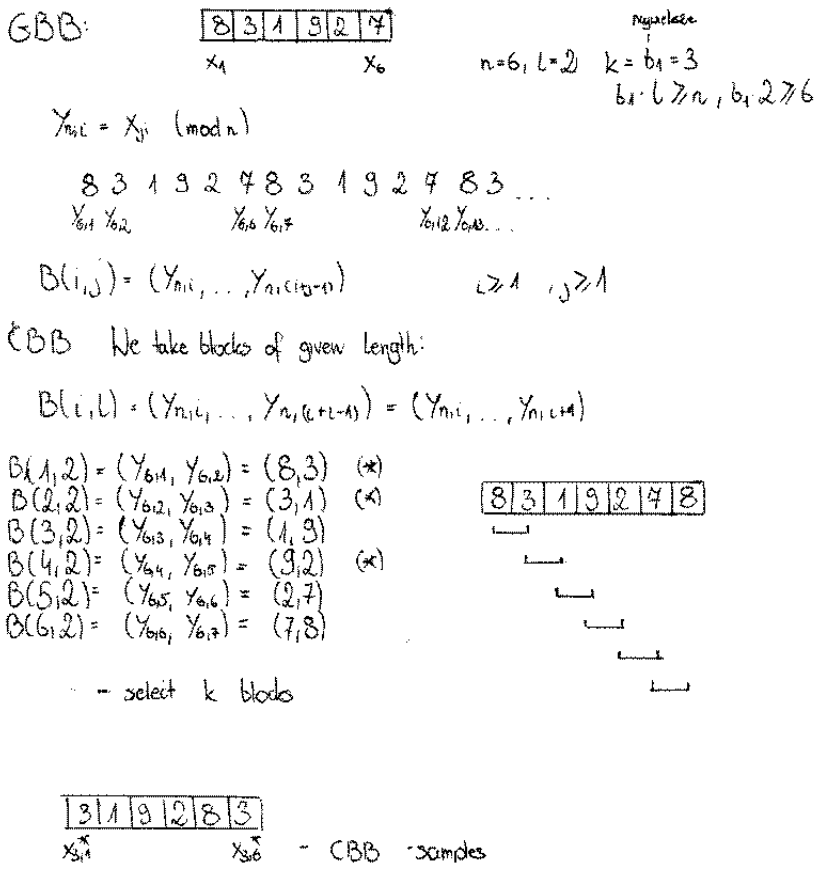
6. $X^*_{C,1}, X^*_{C,2}, ...$ denote the elements of these resampled blocks.

The bootstrap version of an estimator $\hat{\theta_n} = T(F_n)$ under the CBB is defined as $\theta^{*(C)}_{m,n} = T(F^{*(C)}_{m,n})$ for a suitable choice of $m \geq 1$ , where $F^{*(C)}_{m,n}$ denotes the empirical distribution of $X^*_{C,1}, ..., X^*_{C,m}$.

We denote the resampling block indices for the CBB by $I_{3,1}, I_{3,2}, ....$ (the variables $I_i$'s in the collection $(I_1, J_1), (I_2, J_2), ....$
Variables $I_{3,1}, I_{3,2}, ....$ are **conditionally iid** with:

$$P_*(I_{3,1} = i) = n^{-1} \text{ and } P_*(J_i = l) = 1 \text{ for all } i = 1, ..., n.$$

Since each $X_i$ appears exactly $l$ times in the collection of blocks $\{B(1, l), ..., B(n, l)\}$, and since the CBB resamples the blocks from this collection with equal probability, each of the original observations $X_1, ..., X_n$ receives equal weight under the CBB.




Let $X^*_{3,1}, X^*_{3,2}, \ldots$ denote the CBB observations obtained by arranging the elemets of the resampled blocks $\{B(I_{3,i}, l) : i \geq 1\}$ and let $\bar{X}^{*(3)}_m$ denote the CBB sample mean based on $m$ bootstrap observations, where $m = kl$ for some integer $k \geq 1$. Then:

$$
\begin{aligned}
E_* \bar{X}^{*(3)}_m &= E_*\left[m^{-1}\sum_{i=1}^{m} X^*_{3,i}\right] \\
&= \ell^{-1} E_*\left[\sum_{i=1}^{\ell} X^*_{3,i}\right] \\
&= \ell^{-1}\left[n^{-1}\sum_{j=1}^{n}\left\{\sum_{i=1}^{\ell} Y_{n,(j+i-1)}\right\}\right] \\
&= \ell^{-1}\left[\ell \bar{X}_n\right] \\
&= \bar{X}_n.
\end{aligned}
$$

$E_* \bar{X}^{*(3)}_m = E_*[m^{-1}\sum_{i=1}^{m} X^*_{3,i}] = E_*[\ell^{-1}\sum_{i=1}^{l} X^*_{3,i}] =$

$= \ell^{-1} E_*[\sum_{i=1}^{l} X^*_{3,i}] = \ell^{-1}[\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{l} Y_{n,(j+i-1)}] = (*)$

$P(I_{3,1}=i) = \frac{1}{n}$, $P_*(Y_i = l) = \frac{1}{n}$, $l = 1, \ldots, n$

from $\{B(I_{3,i}, l) : i \geq 1\}$ from $\{B(i,j) : i \geq 1, j \geq 1\}$

$B(i,j) = (Y_{n,i}, \ldots, Y_{n,(i+j-1)})$

$(*) = \ell^{-1} n^{-1}[\sum_{i=1}^{l} Y_{n,i} + \sum_{i=1}^{l} Y_{n,i+1} + \ldots + \sum_{i=1}^{l} Y_{n,i+n-1}] =$

$= Y_{n,1} + Y_{2n} \ldots + Y_{n,l}$
$Y_{2,n} + Y_{3,n} + \ldots \quad Y_{n,l+1}$

$Y_{n,n} = \sum_{i=1}^{n} X_i$ $\quad Y_{n,1+n-1} = \sum_{i=1}^{n} X_i$

$= \ell^{-1} n^{-1}(\ell \cdot \sum_{i=1}^{n} X_i) = n^{-1}\sum_{i=1}^{n} X_i = \bar{X}_n$

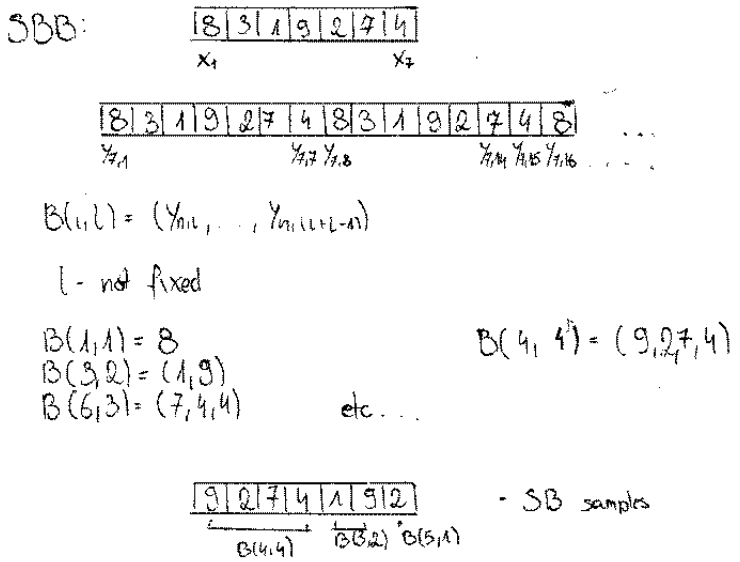
## 3.2 Stationary Block Bootstrap

The stationary bootstrap (SB) differ from the earlier block bootstrap methods in that **it uses blocks of random lengths rather than blocks of a fixed length** $l$.

Let $p \equiv p_n \in (0,1)$ be such that $p \to 0$ and $np \to \infty$ as $n \to \infty$. Then the SB resamples the blocks $B(I_{4,1}, J_{4,1}), B(I_{4,2}, J_{4,2}), \ldots$ where the index vectors $(I_{4,1}, J_{4,1}), (I_{4,2}, J_{4,2}), \ldots$ are conditionally iid with $I_{4,1}$

having **the discrete uniform distribution** on $\{1, ..., n\}$ , and $J_{4,1}$ having **the geometric distribution** $\nu_n$ with parameter $p$:

$$P_*(J_{4,1} = j) \equiv \nu_n(j) = p(1-p)^{j-1} \ , \ j = 1, 2, ...$$




An important property of the SB method is that conditional on $\mathbf{X}_n$, the bootstrap observations $\{X^*_{4,i}\}_{i\in\mathbf{N}}$ are stationary (which is why it is called the "stationary" bootstrap).
For a given resample size $m$, the conditional expectation of the SB sample mean $\bar{X}^{*(4)}_m \equiv m^{-1}\sum_{i=1}^m X^*_{4,i}$ is given by:

$$E_*(\bar{X}^{*(4)}_m) = E_* X^*_{4,i} = \bar{X}_n$$

# 4 Subsampling

Subsampling is a widely used technique in statistical analysis, particularly in scenarios with independently and identically distributed (iid) observations. It involves the selection of different subsets of data to approximate the bias and variance of a statistical estimator.

However, subsampling is not limited to iid observations. It can also be applied to subseries of dependent observations to yield valid estimations of not only bias and variance but also the entire sampling distribution of a statistic. Surprisingly, these valid estimations can be achieved under very mild assumptions.

Suppose that $\hat{\theta_n} = t_n(\mathbf{X_n})$ is an estimator of a parameter $\theta$, such that for some normalizing constant $a_n > 0$, the probability distribution

$$Q_n(x) = P(a_n(\hat{\theta}_n - \theta) \leq x)$$

of the centered and scaled estimator $\hat{\theta_n}$ converges weakly to a limit distribution $Q(x)$:

$$Q_n(x) \to Q(x)$$

for all continuity points $x$ of $Q$. Furthermore, assume that $a_n \to \infty$ as $n \to \infty$ and that $Q$ is not degenerate at zero ( $Q(\{0\}) < 1$ ).
Let $1 \leq l \leq n$ be a given integer and let:

$$B_i = (X_i, ..., X_{i+l-1})', \ 1 \leq i \leq N$$

denote the overlapping blocks of length $l$ where $N = n - l + 1$. (the same as in MBB)
Then, the subsampling estimator of $Q_n$, based on the overlapping version of the subsampling method, is given by :

$$\hat{Q}_n(x) = N^{-1}\sum_{i=1}^{N} 1(a_l(\hat{\theta_{i,l}} - \hat{\theta_n}) \leq x), \ x \in \mathbf{R}$$

where $\hat{\theta_{i,l}}$ is a "copy" of the estimator $\hat{\theta_n}$ on the block $B_i$, defined by $\hat{\theta_{i,l}} = t_l(B_i), \ i = 1, ..., N$.

**Note:** We used constant $a_l$ instead of $a_n$ and $t_l(\cdot)$ (in place of $t_n(\cdot)$) to define the subsample copy $\hat{\theta_{i,l}}$ as the $i$th block $B_i$ contains only $l$ observations.

**Note:** The overlapping version of the subsampling method is a special case of the MBB where a single block is resampled.

The estimator $\hat{Q}_n$ of the distribution function $Q_n(x)$ can be used to obtain subsampling estimators of the bias and the variance of $\hat{\theta_n}$.

The bias of $\hat{\theta_n}$ is given by:

$$\text{Bias}(\hat{\theta_n}) = E\hat{\theta_n} - \theta = a_n^{-1} \int x dQ_n(x)$$

The subsampling estimator of $\text{Bias}(\hat{\theta_n})$ is then obtained by replacing $Q_n(\cdot)$ by $\hat{Q}_n(\cdot)$

$$\widehat{\text{Bias}}(\hat{\theta_n}) = a_n^{-1} \int x d\hat{Q}_n(x) = a_l a_n^{-1} (N^{-1} \sum_{i=1}^{N} \hat{\theta_{i,l}} - \hat{\theta_n})$$

Similarly, the subsampling estimator of the variance of $\hat{\theta_n}$ is given by :

$$\widehat{\text{Var}}(\hat{\theta_n}) = a_l^2 a_n^{-2} [N^{-1} \sum_{i=1}^{N} \hat{\theta}_{i,l}^2 - (N^{-1} \sum_{i=1}^{N} \hat{\theta}_{i,l})^2],$$

which is the sample variance of $\hat{\theta}_{i,l}$ multiplied by the scaling factor $a_l^2 a_n^{-2}$. We need to use the correction factors $\frac{a_l}{a_n}$ and $(\frac{a_l}{a_n})^2$ to scale up from the level of $\hat{\theta}_{i,l}$'s, which are defined using $l$ observations, to the level of $\hat{\theta_n}$, which is defined using $n$-observations.
In applying a bootstrap method, we typically use a resample size that is comparable to the original sample size, and therefore, such explicit corrections of the bootstrap bias and variance estimators are usually unnecessary.
In analogy to the bootstrap methods, one may attempt to apply the subsampling method to a centered variable of the form $T_{1n} \equiv (\hat{\theta_n} - \theta)$ However, this may not be the right thing to do.
(When we don't take the scaling factor we will overestimate the variance)

As noted previously, the subsampling method is a special case of the MBB where the number of resampled blocks is identically equal to 1. Exploiting this fact, we may similarly define other versions of the subsampling method based on nonoverlapping blocks or circular blocks.