KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIWERSYTET GDAŃSKI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Bootstrap methods in data processing.

PWP Interdysciplinary Doctoral Studies in Mathematical Modeling
UDA-POKL.04.01.01-00-026/13-00
Projekt jest współfinansowany przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

1 Introduction

We are going to present one of non-classical methods of mathematical statistics: the *bootstrap method*. It is usually used in the situation when the data sample is small or we are not able to say much about the distribution of the attribute we are trying to measure, in the entire population. We are going to present the method and briefly describe its pros and cons. But before we proceed we should have a quick look at the statistics itself.

What are we thinking about when we mention *statistics*? First of all we should mention that we are going to describe the population – or at least one of the features of objects in the population – by some number (or numbers). But we cannot check or measure *all* objects in the population – we can do it only for a relatively small subset of the entire population – for a *sample*. So we start with:

- certain measurable attribute of objects belonging to a population;
- the sample drawn from the population in order to represent the measured attribute.

So the statistics tells us something about the distribution of the measured attribute in the entire population, based on the data collected from the sample extracted from the population. Mathematical statistics gives a lot of arguments to support the following attitude: taking the sample, that is big enough, allows us to say a lot about the relation between the distribution of the measured property in the sample and in the entire population. Let us briefly remind how it works.

Assume X_1, X_2, …, X_N is the sequence of independent random variables, with the same distribution. This distribution is described by the cumulative distribution function F. This is just the mathematical description of the situation presented above – here F tells us how the value we are interested at is distributed and X_i corresponds to the result of i-th measurement. Of course, when we look at the samples we get (i.e. the values of X_i), the only distribution we may observe is the empirical distribution $F_N(x)$ given by

$$F_N(x) = \frac{\#\{1 \le i \le N : \; X_i \le x\}}{N}, \quad x \in \mathbb{R}.$$

It is a very good moment to recall the basic theorem of mathematical statistics – i.e. Glivenko-Cantelli theorem that says that if the sample is large enough, then the empirical distribution F_N approximates unknown distribution F uniformly for $x \in \mathbb{R}$. More precisely:

$$P_F\left(\lim_{N\to\infty} \sup_{x\in\mathbb{R}} |F_N(x) - F(x)| = 0\right) = 1.$$

We must realize though, that this theorem does not say that F_N is close to F – it says that if the sample is large enough, then there is a big chance that for observations $x_1, x_2, \ldots, x_N$ the empirical distribution is not far from the distribution F.

We will now briefly describe how the typical statistical reasoning looks like. We are interested in certain parameter θ of the investigated distribution (this is a parameter of the entire population). Usually this parameter is the mean, variance, different correlation coefficients, linear regression coefficients etc. Unfortunately, due to various practical limitations we are not able to find the distribution F in the entire population but we have to draw conclusions from the empirical distribution F_N that we have from the collected sample. We would like to know parameter θ for the entire population but we have to live with some estimation of this value based on the sample we have. How this may be done? We have to know certain function $t = T(x_1, x_2, ..., x_N)$, that assigns a value to a sample, a value that may be considered to be the estimation of the parameter θ. We call the function T *the estimator of* θ. So the estimator is the method to calculate the estimation based on the sample we have.

How to find the estimator? Usually we expect that an estimator satisfies several natural properties

- It is *unbiased*, i.e. $\mathbb{E}t = \theta$. The average value of the estimator over all considered samples equals to the value of the estimated parameter.
- It is *consistent*, or more strictly speaking, the sequence of estimators t_n is consistent (where n usually denotes the sample size). It means that it converges in probability to the estimated value, i.e.
$$\lim_{n\to+\infty} P\{|t_n - \theta| < \varepsilon\} = 1.$$
- It is *efficient*, i.e. for any $\varepsilon > 0$ it has the smallest possible variance. In case of unbiased estimator we want to minimize the value of $\operatorname{Var} t$ over all possible estimators (or unbiased estimators).

Finding *any* estimator is usually not a problem, but finding *good* estimator sometimes is. We will not focus on this issue, though – it is enough to mention that the natural estimators of the mean and variance given by

$$T(X_1, .., X_N) = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

and

$$T(X_1, .., X_N) = S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2.$$

respectively, are unbiased, consistent and efficient (please note that for variance estimator we have to use $N - 1$ – otherwise it is not unbiased).

But the parameter estimation is usually the beginning of the trip: the estimated value is just the number – nothing else – and the question appears: how *reliable* it is. Is it far or close to the real value of the parameter? How can we know it? Can we say anything about the difference $t - \theta$?

The first answer that comes to our mind is probably – "how could we say anything"? If we knew θ there would be no need to estimate it! And this is right... but we can look at the problem from the different perspective: we can ask what is the probability that the error we make is small. The best case is when we know the distribution of the estimator $t = T(X_1, ..., X_N)$ over all possible samples of the size N.

Sometimes we are able to find the distribution of t – as a basic example of this situation we can look at the theorem:

Theorem 1. *If the distribution of the certain characteristics in the population is $N(m, \sigma)$ (i.e. normal with mean m and variance σ^2), then the distribution of the mean value estimator given by*

$$T(X_1, ..., X_N) = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

is $N(m, \frac{\sigma}{\sqrt{N}})$.

On the other hand if we are going to estimate the variance we usually use one of the two formulas: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ or $\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ (as we mentioned above only the second one is unbiased). For these estimators we have the following theorem:

Theorem 2. *If the distribution of certain characteristics in the entire population equals to $N(m, \sigma)$, then the statistics*

$$U^2 = \frac{nS^2}{\sigma^2}, \quad \hat{U}^2 = \frac{(n-1)\hat{S}^2}{\sigma^2}$$

have χ^2 distribution with $n-1$ degrees of freedom, while the statistics $t = \frac{\bar{X}-\mu}{S}\sqrt{n-1}$ has t-Student's distribution with $n-1$ degrees of freedom.

There are plenty of other theorems similar to the mentioned above but all of them have to satisfy certain assumptions. And very often we face situations that do not match any theoretically investigated case – e.g. we are not sure if the characteristics is normally distributed (or how it is distributed). Or we have relatively small sample that does not let us draw any valuable conclusions from the general theorems we have.

Below we are going to present the other attitude to the estimation of the distribution of statistics t – it will not be based on the theoretical investigations but rather look at things in the experiment-oriented way.

2 Bootstrap method

Let us briefly summarize what we know from the previous section: we have a population with certain characteristics with the unknown distribution F. We want to find out the value of a certain parameter θ of the distribution (like mean or variance). To do it we may draw a sample from the population, and – based on this sample – find the estimation $t = T(X_1, ..., X_N)$ of the parameter θ. But now we want to say how reliable our estimation is. To do so we need to know the distribution of the estimator t over all possible samples of the size N. The other question we may face is how many elements should the sample consists of, so the estimation is good enough. The rule of thumb says that we should take at least 30 elements in the sample (see. [1]) – but it is just the rule of thumb. If we don't know the underlying distribution we are not able to strictly justify any value, and have to rely on some heuristics.

But what can we do if we cannot assume anything on the underlying statistics? Or the sample does not contain the reasonable (whatever it means) number of attributes? The *bootstrap method* appeared as a way to face these problems. At the end of 1970s and beginning of 1980s several important papers appeared on the subject starting numerous works and related ideas. The key paper is 1979 paper of Bradley Efron [4] that describes the method of investigates its basic properties. It is not hard to get skeptical about the presented ideas – but a lot of (not only heuristic) arguments justifies such attitude. Well, sometimes we just have what we have, and in order to say something about the data, we need to use non-classical tools. More on the historical and theoretical background to the method may be found in monographs [3] and [5].

We will now describe the bootstrap method. First of all we assume to have the sequence $x_1, x_2, ..., x_N$ of values of independent random variables $X_1, X_2, ..., X_N$ having the same distribution F. We want to find the value of the certain parameter $\theta \in \mathbb{R}$ of this distribution and we are using the estimator T to estimate it. The value of T for specific set of data

$x_1, x_2, ..., x_N$ is the estimation $t = T(x_1, x_2, ..., x_N)$ of the parameter θ. Of course for other values in the sample we would receive different value of t – but this sample is everything we have. We cannot repeat the sampling procedure forever to find the distribution of statistics t, one sample is everything we get.

Is it possible to create more data from the sample we have? It looks like a crazy idea but this is the way bootstrap method starts. We look at the initial sample like it is the entire population and draw samples from the sample. We build the sequence of independent random variables with empirical distribution $\hat{F}_N$, i.e.

$$P(X_i^* = x_k) = \frac{1}{N} \quad \text{for} \quad k = 1, 2, \ldots, N.$$

for $i = 1, 2, \ldots, N$.

So the bootstrap method builds samples from the sample. Hence the distribution $\hat{F}_N$ is called *the sample bootstrap distribution*, the vector random variable $X^* = (X_1^*, X_2^*, ..., X_N^*)$ – is called the *bootstrap sample*. The values of the bootstrap sample will be denoted by $(x_1^*, x_2^*, ..., x_N^*)$. We should observe that the numbers x_i^* are drawn independently, so they may be repeating. The empirical distribution implies that each of the values may be drawn with the same probability (this may be changed as described below).

Having the bootstrap sample we may apply estimator T to it and find the value of $t^* = T(X^*) = T(X_1^*, X_2^*, ..., X_N^*)$ estimating the value of t. One may think that we are repeating the population-sample schema, but this case is different. Now we know the entire population, we know the distribution, we know the value t. We are in the situation of the small population meaning that we may find out the distribution of t^*. Hence we know how much the value of t^* may differ from the value of t. Why couldn't we assume now that t^* differs from t as much as t differs from the real value of θ?

To briefly summarize this:

Assumption 1 (Basic assumption of the bootstrap method)**.** *The distribution of $t^* - t$ resembles the distribution of $t - \theta$.*

With this assumption in hand we may use the distribution of $t^* - t$ to approximate the distribution of $t - \theta$. This is how we may check how reliable is the estimation of θ by t. Let us think about the confidence intervals for a moment. Assume that there is 95% chance of $t^* \in (t - \alpha^-, t + \alpha^+)$. By the basic assumption given above we can conclude that there is 95% chance of $t \in (\theta - \alpha^-, \theta + \alpha^+)$. This looks like very heuristic attitude (well, it is) but there are some strong and strict arguments supporting this.

Let us have a look at how the distribution of $t^* - t$ (or t^*) may be found. We may point out several ideas:

- Analytically. We analyze all the cases that may appear and find the analytic description of the t^* distribution. Actually this may be done in some very simple cases (like Bernoulli distribution) but gets practically impossible for bigger samples.

- The distribution t^* is supported on the finite set – so we may just generate the entire set. Practically – even for the relatively small sample size – it is very difficult to handle. Let us observe that even for $N = 10$ there is 10^{10} all possible samples from the initial sample. The number is quite big – but what can we say if $N = 20$?

- Monte Carlo attitude – we may draw relatively many (let's say $R = 1000$) bootstrap samples and for each of them we find the value of t^*. The values of $t_1^*, t_2^*, \ldots t_R^*$ may be used to estimate $t^* = T(X^*)$. We don't get the exact and full distribution of t^* but we find some approximation of it. The bigger R we have, the better approximation we get.

Practically the last attitude is the one that is the most popular one.

Let us also review some modifications of the bootstrap method described above.

1. We may assign each value a different weight. We just may treat some values as more important – or reliable – then others. Then, when the bootstrap sample is created, we may draw these values with higher probability. In the basic bootstrap method each value is drawn with the same probability. We may see that it may be a good way to handle *outliers*, i.e. the values that – as we see – are far too distanced to what we expect to be a typical value. The single outlier may seriously influence our estimation – and even if we cannot remove it from the sample, we may try to limit its influence to the final results.

2. Normalization. In a typical case we are comparing distributions of $t - \theta$ and $t^* - t$. But sometimes it may be more natural to compare normalized values, i.e. $\frac{|t-\theta|}{\sigma}$ to $\frac{|t^*-t|}{s}$.

3. Smoothing the empirical distribution $\hat{F_N}$. There is no need to assume that the bootstrap distribution is discrete. The values in the sample are all we have but we may assume that the bootstrap samples will be drawn from the continuous distribution $\tilde{F_N}$, that is concentrated in the neighbourhood of the values from the sample. This makes the boostrap distribution more smooth.

4. Symmetrization. Sometimes we know that the distribution we are sampling is symmetric (e.g. with respect to the median). So having collected the data, we introduce some symmetry to it – for each value in the sample we add the value symmetric to it with respect to the median.

5. Balanced bootstrap, introduced in [2], suggests the selection of samples, so each value from the initial sample, $x_1, ..., x_N$ appears in the bootstrap sample X_i^* $(i = 1, ..., R)$ with the same frequency. This looks like very difficult to achieve but with some clever ideas this may be implemented.

6. Antithetic bootstrap, introduced in [6] is another method to introduce some symmetry to bootstrap samples – but now with respect to how they are ordered as real numbers. Loosely speaking: for each randomly chosen bootstrap sample we create the 'anti-sample' in such way that if the bootstrap sample contains the second lowest element from $\{x_1, ..., x_N\}$, then the anti-sample must contain the second largest element from this set. Assuming we have $x_1 \leq x_2 \leq ... \leq x_{N-1} \leq x_N$ this means that if bootstrap sample contains x_k, then anti-sample contains x_{N-k}.

3 Estimation of the mean

We will now look more closely to the estimation of the mean – which is probably the most typical case to handle. The good thing is that there exists the natural estimator (given by the arithmetic mean of the sample).

Let us now try to estimate $\mathbb{E}(F)$ by means of the most natural estimator, i.e. the average from the sample. Starting from the observation $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ and the bootstrap distribution

$$\hat{F}_N : \quad P(x_i) = \frac{1}{N} \quad \text{for} \quad i = 1, 2, \ldots, N.$$

We calculate the means $t_1^*, t_2^*, \ldots, t_R^*$ for all bootstrap samples drawn. And we take the bootstrap estimator of $\mathbb{E}(F)$ given by

$$\bar{t}^* = \frac{1}{R}\sum_{i=1}^{R} t_i^*,$$

where R is big enough. How big? We need to find empirically and have to consider the limitations in our resources (e.g. time needed to generate bootstrap samples) – in practice it is enough to take $R = 1000$. The estimation $\bar{t}^*$ we arrived to may be considered to be the new estimation of $\mathbb{E}(F)$. We should note that the value t^* may not be treated as the *better* estimation than $\bar{x}$ (i.e. the average of the sample). Such attitude is definitely incorrect: the R bootstrap samples are not *all* bootstrap samples! If we took all bootstrap samples then we would have $t^* = \bar{x}$. And with $R \to +\infty$ we have $t^* \to \bar{x}$. The bootstrap method will not fix the estimation – it just helps to say how reliable the estimation is.

In this situation the basic assumption of the bootstrap method is that the distribution of $\bar{X} - \mathbb{E}(F)$ is similar to the distribution of $\bar{X}^* - \bar{x}$. The last distribution (similarly as the distribution of $\bar{X}^*$), may be found theoretically and using the Monte Carlo attitude is not necessary in this case. Actually it may be shown that the mean value of $\bar{X}^*$ equals to $\bar{x}$. What do we gain though? First of all – we receive the distribution of $\bar{X}^* - \bar{x}$ so we can estimate how the value $\bar{x}$ may differ from the actual value of $\mathbb{E}(F)$.

4 Bootstrap in `R`

In this section we will briefly present how to use the bootstrap methods in `R` package (`http://www.r-project.org`). The simulations will be performed with the `boot` library, so the first thing we should do is to load the library.

```
library(boot)
```

In this section we are going to present how the standard boot library may be used and interpreted. We are not showing the details of the procedures, we are not describing all parameters of the functions from the `boot` package. In search of details we suggest to go to the package documentation (e.g. `http://cran.r-project.org/web/packages/boot/boot.pdf`). What follows is just a simple example to perform some preliminary tests and simulations.

As an example let us consider the sample with 10 values

```
x <- c(1.55,5.28,2.83,1.47,7.89,2.03,0.18,1.82,0.14,6.62)
```

In order to do the bootstraping easily, we should define the statistics function that takes two arguments – the list of all bootstrap samples (so the list of all generated 10-element sequences) and the index (ordinal number of the sample in the list that should be considered). The example of the definition is

```
samplemean <- function(y, d) mean(y[d])
```

or

```
samplesd <- function(x, d) sd(x[d])
```

or anything more complex, maybe quite nonstandard.

Then the actual bootstraping is performed by the procedure

```
b=boot(x,samplemean,R=1000)
```

In the presented example it takes the three parameters:

1. `x` is the data vector (sample) we are bootstraping from;
2. `samplemean` is the function used to calculate the bootstrap statistics from the bootstrap sample (in this case it is just the mean value of the sample);
3. `R` is the number of bootstrap samples generated.

The function `boot` returns the special object keeping all necessary information. There are some easy methods to extract data from the returned object:

- `print(b)` presents basic information about the output of the bootstrap procedure

  ```
  ORDINARY NONPARAMETRIC BOOTSTRAP

  Call:
  boot(data = x, statistic = samplesd, R = 1000)

  Bootstrap Statistics :
           original      bias       std.  error
  t1*      2.690495   -0.1961996   0.5421074
  ```

  The meaning of the values is as follows:

  - `original` is the value of the statistics calculated on the sample passed to the procedure;
  - `bias` is the difference between the mean value of the statistics calculated for all bootstrap samples and the original;
  - `std.  error` is just the standard error for all bootstrap samples.

- `plot(b)` shows the graphical representation of the generated bootstrap data

- `boot.ci(b,conf=0.95)` shows bootstrap confidence intervals – identified with several different methods (we will not describe them here – for details refer to [5])

  ```
  BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
  Based on 1000 bootstrap replicates

  CALL :
  boot.ci(boot.out = b, conf = 0.95)

  Intervals :
  Level Normal          Basic
  95% ( 1.824, 3.949 ) ( 2.028, 4.299 )

  Level Percentile          BCa
  95% ( 1.082, 3.353 ) ( 1.784, 3.577 )
  Calculations and Intervals on Original Scale
  Some BCa intervals may be unstable
  Warning message:
  In boot.ci(b, conf = 0.95) :
  ```

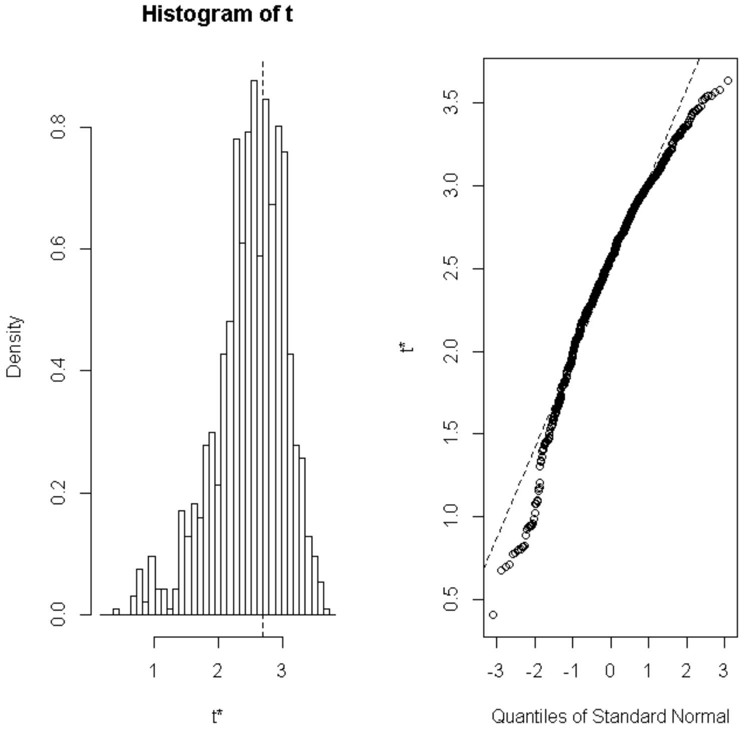

Figure 1: Sample results of the function `plot(b)` for a mean value bootstrap estimation.

```
bootstrap variances needed for studentized intervals
```

We should not worry about the warning message – one of the methods for estimation of confidence intervals (so called *studentized bootstrap method* requires some extra input to the procedure – the input which was not supplied. But nevertheless, we can see the results of the four methods estimating the confidence intervals (normal, basic, percentile and BCa). As mentioned before we are not going to describe the methods at all, but it is good to mention that the *percentile* method corresponds to the natural attitude mentioned earlier. In order to estimate to-sided 95% confidence interval we just estimate such values $t_{2.5\%}$ and $t_{97.5\%}$ that

$$P(t^* \leq t_{2.5\%}) = 2.5\% \quad P(t^* \geq t_{97.5\%}) = 2.5\%.$$

Let us also look at the example in `R` that generates the bootstrap samples for the standard deviation statistics. We call the sequence of operations

```
x <- c(1.55,5.28,2.83,1.47,7.89,2.03,0.18,1.82,0.14,6.62)
samplesd <- function(x, d) sd(x[d])
b=boot(x,samplesd,R=1000)
print(b)
plot(b)
boot.ci(b,conf=0.95)
```

And end up with the following information:

```
> print(b)
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = x, statistic = samplesd, R = 1000)

Bootstrap Statistics :
         original     bias       std.  error
t1*      2.690495   -0.1717596   0.5024649
```

Then `plot(b)` creates the image presented in Figure 2

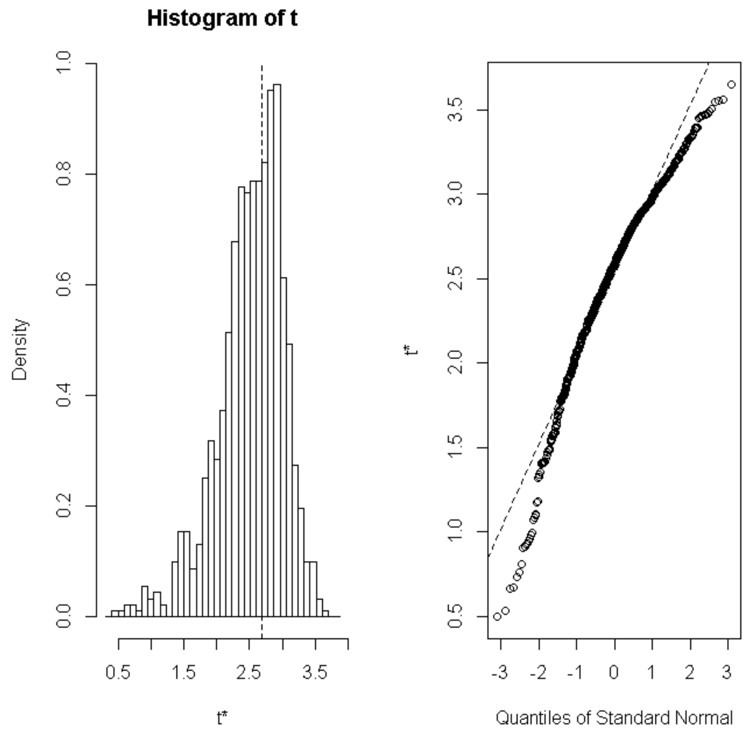

Figure 2: Sample results of the function `plot(b)` for a standard deviation bootstrap estimation.

And for the confidence intervals we have

```
   BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = b, conf = 0.95)

Intervals :
Level Normal          Basic
95% ( 1.877, 3.847 ) ( 2.059, 4.023 )

Level Percentile          BCa
95% ( 1.358, 3.322 ) ( 1.810, 3.555 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
Warning message:
```

```
In boot.ci(b, conf = 0.95) :
bootstrap variances needed for studentized intervals
```

References

[1] Belle G. v., *Statistical Rules of Thumb*, Wiley-Interscience, 2008

[2] Davison A.C., Hinkley D.V., Schechtman E., *Efficient bootstrap simulation*, Biometrika 73, 555-566, (1986)

[3] Davison A.C., Hinkley D.V., *Bootstrap Methods and their Application*, Cambridge University Presss, 1997

[4] Efron B., *Bootstrap Methods: Another Look at the Jackknife*, The Annals of Statistics 7, 1979

[5] Efron B., Tibshirani R.J., *An Introduction to the Bootstrap*, Chapman & Hall, 1993

[6] Hall P., *Antitethic resampling for the bootstrap*, Biometrika, 76, 713-124 (1989)