BOOTSTRAP METHODS AND THEIR APPLICATIONS

Summary of PhD Dissertation

LÁSZLÓ VARGA

Supervisor: András Zempléni
Associate professor, CSc

Doctoral School of Mathematics
Director: István Faragó

Doctoral Programme of Applied Mathematics
Director: János Karátson

Department of Probability Theory and Statistics

Eötvös Loránd University

Faculty of Science

2017

1 Introduction

The dissertation is about a computationally demanding topic of theoretical/applied statistics: bootstrap methods. My purpose with my PhD thesis is multifaceted: I would like to outline the basic concept of bootstrap methods, enlighten the mathematical difficulties behind the techniques, broaden the theory of bootstrap with new methods and show their practical applicability.

This summary follows mostly the structure of the dissertation. The PhD thesis is based on the papers [1], [2] and [3] of the author, which apply the theoretical results to different meteorological phenomena.

2 Special topics from probability theory and statistics

This chapter contains the main concepts and results from probability theory, time series analysis (stationary processes, vector autoregression), copulas (goodness-of-fit tests based on Kendall's transform, homogeneity test) and extreme value theory (one- and bivariate maxima and threshold models) and effective sample size needed for the applications and the further investigation and expansion of the theory of bootstrap methods.

3 Bootstrap methods

The bootstrap (or bootstrapping) is a statistical technique based on random sampling with replacement for solving a wide range of statistical problems: estimating the distribution of a statistic of interest, reducing bias, testing statistical hypothesis, constructing confidence intervals or confidence sets, making forecasts for time series and so on.

Bootstrap has been developed in the last two decades of the previous century. The main concept was introduced in the classical article [6] by Bradley Efron and since then – thanks to the numerous extensions of the concept – it has become one of the most widely used Monte Carlo methods in a number of areas of applied sciences. The applicability of bootstrap methods increased exponentially along with the development of computer hardware and programming languages. In the years after its introduction, several limitations of the bootstrap have been found, which served as motivation to modify the original concept, leading to a range of extensions: parametric/semiparametric bootstrap, residual bootstrap, block bootstrap, multiplier or weighted bootstrap, double/triple bootstrap and m-out-of-n bootstrap. The greatest impact on my thinking and on my dissertation had the textbooks [9], [8] and [11].

3.1 The bootstrap principle

The basic idea of bootstrap is to produce new samples from the original one via resampling with replacement. Here absolutely nothing is assumed about the distribution of the

sample. Formally, let $\mathcal{X}_n = (X_1, \ldots, X_n)^T$ be a sequence of i.i.d. random variables with unknown common distribution function F and let $T_n = t_n(\mathcal{X}_n; F)$ be a statistic of interest (like the sample mean $\overline{X}$). The X_i random variables could also be random vectors – in this case, $\mathcal{X}_n$ would be a matrix – , however, the essence of the technique remains the same. As F is unknown, the distribution of the statistic T_n is unknown, too. Our main purpose is to approximate the distribution of T_n or its function of interest – for example the standard deviation of T_n or some of its quantiles for estimating p-values.

We will denote by P_*, E_*, D_*^2 and Cov_* the conditional probability, the conditional expectation, the conditional variance and the conditional covariance given the sample $\mathcal{X}_n$, for instance formally $P_*(\,\cdot\,) = P(\,\cdot\,|\mathcal{X}_n)$. The i.i.d. bootstrap technique is the following. For a given $\mathcal{X}_n$, we draw a random sample $\mathcal{X}_m^* = \{X_1^*, \ldots, X_m^*\}$ of size m with replacement from $\mathcal{X}_n$:

$$P_*(X_j^* = X_i) = P(X_j^* = X_i \,|\, \mathcal{X}_n) = \frac{1}{n} \qquad i = 1, \ldots, n \quad j = 1, \ldots, m,$$

thus the elements of the bootstrap sample are conditionally (on the original sample) independent and identically distributed. Therefore, the common distribution of the X_i^* is given by the empirical distribution function $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$. The resampling size m usually equals with the original sample size n. In the next step, we define the bootstrap version of the statistic: $T_{m,n}^* = t_m(\mathcal{X}_m^*; F_n)$. By repeating this procedure enough times, we can approximate the unknown distribution function G_n of T_n by its bootstrap counterpart G_m^*. In most of the cases G_m^* cannot be determined explicitly, but it can be approximated by simulation.

Our goal is usually to show that the bootstrap distribution of the test statistic is close enough to the original distribution of the test statistic. The bootstrap is called weakly/strongly consistent if with a properly chosen metric, the distance of the bootstrap and the original distribution converges to zero in probability/almost surely – see Chapter 3.1 in [14] for a thorough description and ideas.

3.2 Block bootstrap methods

If we have stationary dependent data, the favourite resampling method is the block bootstrap, see [11] for a thorough presentation. The general idea behind block bootstrap is that if the data are dependent, then instead of individual sample elements, let us take blocks from the sample and put them together. We describe now the block bootstrap resampling more precisely:

1. Wrap the data $X_1, \ldots, X_n$ around a circle, i.e. define the pseudo-series $\widetilde{X}_t = X_{t_{\mathrm{mod(n)}}}$ $(t \in \mathbb{Z}_+)$, where $\mathrm{mod(n)}$ denotes division "modulo n".
   This means that $\widetilde{X}_k = \widetilde{X}_{k+n} = \widetilde{X}_{k+2n} = \ldots = X_k$ for all $k \in \{1, 2, \ldots, n\}$.

2. Specify the starting indices of the blocks: $I_1, I_2, \ldots$ series of random variables concentrated on subset $A \subseteq \{1, \ldots, n\}$.

3. Specify the lengths of the blocks: $L_1, L_2, \ldots$ series of non-negative integer valued random variables.

4. Define the blocks: $B(I_i, L_i) = \{\widetilde{X}_{I_i}, \widetilde{X}_{I_i+1}, \ldots, \widetilde{X}_{I_i+L_i-1}\} \quad i = 1, 2, \ldots$

5. Put the blocks together: $\mathcal{X}^* = \{B(I_1, L_1), B(I_2, L_2), \ldots\}$.

The classical block bootstrap is the moving block bootstrap (MBB), where block size is a fixed integer number $1 \leq b \leq n$, the blocks are taken only from the original sample and each block can be chosen with equal probability. Hence the starting indices of the blocks are uniform on the set $A = \{1, 2, \ldots, n - b + 1\}$. The circular block bootstrap (CBB) differs from MBB just in $A = \{1, 2, \ldots, n\}$. Stationary block bootstrap (SBB) is a generalisation of CBB in the way that the block lengths are independent and geometrically distributed with parameter $p \in (0, 1]$.

Generalised block bootstrap

We introduced in the PhD thesis a block bootstrap method, which helps overcoming the problem that originally the block size was supposed to be a natural number, see 3.4 for the concrete motivation. In our extension, the block size is a random variable, and it contains circular block bootstrap as a special case.

In case of $1 \leq b \in \mathbb{R}$, let the generalised block bootstrap sample be defined as follows. Let k be a random integer between 1 and the sample size n and, again, let us wrap the sample around the circle. The bootstrap blocks are either of length $\lfloor b \rfloor$ or $\lceil b \rceil$:

$$\begin{array}{lll} \{X_k, X_{k+1}, \ldots, X_{k+\lfloor b \rfloor - 1}\} & \text{with probability} & 1 - b + \lfloor b \rfloor \\ \{X_k, X_{k+1}, \ldots, X_{k+\lceil b \rceil - 1}\} & \text{with probability} & b - \lfloor b \rfloor \end{array}$$

where $\lceil b \rceil$ denotes the upper and $\lfloor b \rfloor$ the lower integer part of b. At last, we put the blocks together. We can formalise the "parameters" of our **generalised block bootstrap** (GBB) with the notations used at block bootstrap:

- $1 \leq b \in \mathbb{R}$ is the expected block size, fixed in advance
- $A = \{1, 2, \ldots, n\}$
- $I_i | \mathcal{X}_n \sim \text{Unif}(A) \quad i = 1, 2, \ldots$ are conditionally independent from each other
- $P_*(L_i = \lceil b \rceil) = 1 - P_*(L_i = \lfloor b \rfloor) = b - \lfloor b \rfloor \quad i = 1, 2, \ldots$ are conditionally independent from each other
- I_i and L_j are conditionally independent for all i and j

Proposition 1. *Using the GBB technique, we get for all $i = 1, 2, \ldots$ that $E_* L_i = b$ and $D_*^2 L_i = (b - \lfloor b \rfloor)(1 - b - \lfloor b \rfloor)$.*

In the same way as the circular block bootstrap sample, our generalised bootstrap sample is usually not a stationary process, conditionally on the original sample. From now on in this subsection, we assume that the bootstrap sample size equals the original one, i.e. $m = n$. Let us define the following random variables:

- N_s: the number of blocks with block size $\lfloor b \rfloor$ (s in the subscript refers to small);
- N_l: the number of blocks with block size $\lceil b \rceil$ (l in the subscript refers to large);
- R: length of the remainder block size, i.e. $R = n - N_s \cdot \lfloor b \rfloor - N_l \cdot \lceil b \rceil$.

The following proposition gives the distribution of N_s, which is enough to determine the distributions of N_l and R.

Proposition 2. *Let $p = b - \lfloor b \rfloor$, then the distribution of N_s is as follows for $j = 0, 1, \ldots, \left\lfloor \frac{n}{\lfloor b \rfloor} \right\rfloor$:*

$$P_*(N_s{=}j) = \begin{cases} 0 & \text{if } \frac{n-(j+1)\lfloor b \rfloor}{\lceil b \rceil} \text{ is integer} \\ \left[p^{\frac{n-j\lfloor b \rfloor}{\lceil b \rceil}} (1-p)^{j-1} \right] \left[\binom{j + \frac{n-j\lfloor b \rfloor}{\lceil b \rceil} - 1}{j-1} + \binom{j + \frac{n-j\lfloor b \rfloor}{\lceil b \rceil} - 1}{j} (1-p) \right] & \text{if } \frac{n-j\lfloor b \rfloor}{\lceil b \rceil} \text{ is integer} \\ \binom{j + \lfloor \frac{n-j\lfloor b \rfloor}{\lceil b \rceil} \rfloor}{j} p^{\lfloor \frac{n-j\lfloor b \rfloor}{\lceil b \rceil} \rfloor} (1-p)^j & \text{otherwise} \end{cases}$$

In the applications, we will need the trace of the covariance matrix of the bootstrap mean.

Theorem 1. *The covariance matrix of the bootstrap mean can be calculated as*

$$\begin{aligned} \mathrm{Cov}_*(\overline{\mathbf{X}}^*_b) = & \frac{\lfloor b \rfloor^2}{n^2} \left[\mathrm{Cov}_*(\overline{\mathbf{X}}^*_{\lfloor b \rfloor, i}) \cdot E_* N_s + D^2_* N_s \cdot \overline{\mathbf{X}}_n (\overline{\mathbf{X}}_n)^T \right] + \\ & + \frac{\lceil b \rceil^2}{n^2} \left[\mathrm{Cov}_*(\overline{\mathbf{X}}^*_{\lceil b \rceil, i}) \cdot E_* N_l + D^2_* N_l \cdot \overline{\mathbf{X}}_n (\overline{\mathbf{X}}_n)^T \right] + \\ & + \frac{1}{n^2} \left[\sum_{i=0}^{\lfloor b \rfloor - 1} i^2 P_*(R = i) \cdot \mathrm{Cov}_*(\overline{\mathbf{X}}^*_{i,1}) + D^2_* R \cdot \overline{\mathbf{X}}_n (\overline{\mathbf{X}}_n)^T \right], \end{aligned}$$

where $\overline{\mathbf{X}}^*_{b,i}$ is the mean of the ith block ($i = 1, 2, \ldots$) with size b.

3.3 Weighted bootstrap

The weighted (or multiplier/wild) bootstrap is an extension of the i.i.d. bootstrap scheme. The idea of the classical weighted bootstrap appeared first in Chapter 10 of [7] and have been applied – with appropriate modifications – to a number of areas in the theory of bootstrap.

Our research in the past years was about a special topic, the **weighted likelihood bootstrap** and its applications. The bootstrap weights will be random variables, belonging to the sample $\mathcal{X}_n$, we denote the weights as $\boldsymbol{\tau}_n = (\tau_{n,1}, \tau_{n,1}, \ldots, \tau_{n,n})$. [13] applied weighted bootstrap to maximum likelihood estimation, simply by multiplying the elements of the log-likelihood function by the weights. In this context, $P_*(\cdot)$ will denote the conditional probability, when the weights are considered as random variables and the sample is fixed. In the dissertation, we proved a further generalisation with random weights of Wilks' classical result ([15]) about the asymptotics of the generalised likelihood ratio test statistic.

Let us assume that we have a distribution family with density $f_{\boldsymbol{\vartheta}}(x)$, where $\boldsymbol{\vartheta} \in \Theta \subseteq \mathbb{R}^p$ is the unknown parameter. We will denote the log-likelihood function of an i.i.d. $\mathcal{X}_n = (X_1, \ldots, X_n)^T$ sample by $l(\boldsymbol{\vartheta}|\mathcal{X}_n) = l(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{\vartheta}}(X_i)$ and the maximum likelihood estimate of the parameter by $\widehat{\boldsymbol{\vartheta}}_n = \arg\max_{\boldsymbol{\vartheta}} l(\boldsymbol{\vartheta})$. We can define the (bootstrap) weighted version of the log-likelihood function:

$$l^*(\boldsymbol{\vartheta}|\mathcal{X}_n) = l^*(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \tau_{n,i} \log f_{\boldsymbol{\vartheta}}(X_i),$$

and be $\widehat{\boldsymbol{\vartheta}}^*_n$ the weighted ML estimate. The assumptions on the weights may vary from context to context, that's why we will only write the assumptions needed for our problem. Let us suppose the following *assumptions for the bootstrap weights*:

A1. they are independent from the data-generating process;

A2. they have finite second moments for all $n = 1, 2, \ldots$;

A3. $P(\tau_{n,i} \geq 0) = 1; \quad i = 1, \ldots, n; \;\; n = 1, 2, \ldots$;

A4. $E\tau_{n,i} = 1 \quad i = 1, \ldots, n; \;\; n = 1, 2, \ldots$;

A5. There exist $\gamma \in \mathbb{R}$ for which $\frac{1}{n} \sum_{i=1}^{n} \tau_{n,i}^2 \xrightarrow[n\to\infty]{p} \gamma$;

A6. There exists a $|q| < 1$ real number for which $\mathrm{Cov}(\tau_{n,i}, \tau_{n,j}) \leq q^{|i-j|} \quad 1 \leq i \neq j \leq n; \;\; n = 1, 2, \ldots$

Several distributions fulfil the assumptions above, we shall use the multinomial and i.i.d. exponential weights in the applications:

$$(\tau_{n,1}, \ldots, \tau_{n,n}) \sim \text{ Multinomial}\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right) \quad \text{and} \quad (\tau_{n,1}, \ldots, \tau_{n,n}) \sim \text{ i.i.d. Exp(1).}$$

They were chosen for their simplicity, and because we were curious, whether weak dependence between the coordinates (multinomial) has a considerable effect on our results or not, compared to i.i.d. weights.

Let us assume that standard strong regularity conditions hold for the distribution family, for example (RR) of [4], page 191. Let us partition the parameter vector into two parts: $\boldsymbol{\vartheta} = \begin{pmatrix} \boldsymbol{\sigma} \\ \boldsymbol{\rho} \end{pmatrix} \begin{matrix} \}q \\ \}p-q \end{matrix}$ and be $\boldsymbol{\sigma}_\bullet \in \text{int}(Pr_H(\Theta))$, where H is the subspace according to the first q coordinates of Θ. Let us define the constrained ML estimate

$$\widetilde{\boldsymbol{\vartheta}}_n = \begin{pmatrix} \boldsymbol{\sigma}_\bullet \\ \widetilde{\boldsymbol{\rho}}_n \end{pmatrix} = \arg\max_{\boldsymbol{\rho}} l\left(\begin{pmatrix} \boldsymbol{\sigma}_\bullet \\ \boldsymbol{\rho} \end{pmatrix}\right). \tag{1}$$

Let us denote by $\widetilde{\boldsymbol{\vartheta}}^*_n$ the weighted version of (1). By the result of Wilks, we know that in case $\boldsymbol{\sigma} = \boldsymbol{\sigma}_\bullet$, $T_n := 2\left[l(\widehat{\boldsymbol{\vartheta}}_n) - l(\widetilde{\boldsymbol{\vartheta}}_n)\right] \xrightarrow[n\to\infty]{d} \chi^2_q$.

Theorem 2. ***Asymptotic distribution of the generalised likelihood ratio test statistic with random weights.***
Under strong regularity conditions and using the notations of this section; if $\boldsymbol{\sigma} = \boldsymbol{\sigma}_\bullet$, then

$$T^*_n := \frac{2}{\gamma}\left[l^*(\widehat{\boldsymbol{\vartheta}}^*_n) - l^*(\widetilde{\boldsymbol{\vartheta}}^*_n)\right] \xrightarrow[n\to\infty]{d} \chi^2_q.$$

We needed two propositions to justify Theorem 2. The first one can be thought as a generalisation of the multidimensional central limit theorem, while the second one is a generalisation of the weak law of large numbers.

Proposition 3. *Be the $\boldsymbol{\tau}_n$ random variables (weights) as defined before and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ i.i.d. p dimensional random vectors with expectation vector $\mathbf{0}_p$ and covariance matrix $\boldsymbol{\Sigma}$. If the weights are independent from the random vectors, then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tau_{n,i} \mathbf{Y}_i \xrightarrow[n\to\infty]{d} N_p(\mathbf{0}_p, \gamma\boldsymbol{\Sigma}).$$

Proposition 4. *Be the $\boldsymbol{\tau}_n$ random variables (weights) as defined before and $Y_1, Y_2, \ldots$ i.i.d. random variables with finite first two moments. If the weights are independent from the Y_i random variables, then* $\frac{1}{n}\sum_{i=1}^{n} \tau_{n,i} Y_i \xrightarrow[n\to\infty]{p} EY_1$.

3.4 Block size determination in practice

There are two general strategies for block size selection that can be applied in practice. These methods are based on either subsampling ([10]) or nonparametric plugin ([12]). In Sections 4.1 and 4.3, we follow a *different*, model-based approach: we try to find the best block size by fitting a proper vector autoregression – VAR(p) – model to the multidimensional data and then searching the block size, for which the covariance matrix of the block bootstrap mean is "closest" to the covariance matrix of the VAR-sample mean. The methodology is general enough to be compatible with other, more complex classes of time series models as well.

In Section 4.1, the optimal block size $\widehat{b}$ is the integer number, for which the estimated trace of the covariance matrix is closest to the one derived from the fitted VAR model:

$$\widehat{b} = \underset{1 \leq b \in \mathbb{Z}}{\operatorname{argmin}} \left| \operatorname{tr}\left(\operatorname{Cov}\left(\overline{\mathbf{X}}_{VAR}\right)\right) - \operatorname{tr}\left(\operatorname{Cov}_*(\overline{\mathbf{X}}_b^*)\right)\right|, \tag{2}$$

where $\operatorname{Cov}_*(\overline{\mathbf{X}}_b^*)$ is the covariance matrix of the bootstrap mean with block size b and $\operatorname{Cov}\left(\overline{\mathbf{X}}_{VAR}\right)$ is calculated the following way. It is enough to care with VAR(1) processes, because a d-dimensional VAR(p) can be rewritten as a pd-dimensional VAR(1). If we have a d-dimensional VAR(1) process in the form of $\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t$ with $\operatorname{Cov}(\boldsymbol{\varepsilon}_t) = \mathbf{C}$, then

$$\operatorname{Cov}(\overline{\mathbf{X}}_{VAR}) = \frac{1}{n}\left\{\boldsymbol{\Gamma}_X(0) + \sum_{h=1}^{n-1}\left(1 - \frac{h}{n}\right)\left[\mathbf{A}^h\boldsymbol{\Gamma}_X(0) + (\boldsymbol{\Gamma}_X(0))^T(\mathbf{A}^h)^T\right]\right\}, \tag{3}$$

where $\boldsymbol{\Gamma}_X$ is the autocovariance matrix and $\operatorname{vec}(\boldsymbol{\Gamma}_X(0)) = (\mathbf{I}_{d^2} - \mathbf{A} \otimes \mathbf{A})^{-1}\operatorname{vec}(\mathbf{C})$.

In the literature simulations are naturally based on integer block sizes. But using the block length of (2), the estimated trace of covariance may be not be close enough to the theoretical trace of covariance. The same is true for other methods for block size determination. This may cause substantial bias, especially for smaller b. This can be overcome by the generalisation of the block bootstrap methodology described in Section 3.2. In this case, instead of approximating as in (2), we simply solve the following equation in the unknown variable $1 \leq b \in \mathbb{R}$

$$\operatorname{tr}\left(\operatorname{Cov}(\overline{\mathbf{X}}_{VAR})\right) = \operatorname{tr}\left(\operatorname{Cov}_*(\overline{\mathbf{X}}_b^*)\right). \tag{4}$$

In Section 4.3 we will follow this approach to determine the block size.

3.5 Profile likelihood and bootstrapping the extremes

In Section 4.2, we use Theorem 2 to construct confidence intervals for the return values in the univariate extreme value models based on threshold exceedances. We know from the Pickands–Balkema–de Haan theorem that exceedances over a given threshold may be considered as a sample from the generalized Pareto distribution (GPD), which has the cumulative distribution function

$$H(x) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - e^{-\frac{x}{\sigma}} & \text{if } \xi = 0 \end{cases},$$

where parameter ξ is called the shape parameter, while σ is the scale parameter. In the applications, we used a different parametrisation with ξ and the q-quantile (return value) $H^{-1}(q)$ as parameters, so the log-likelihood function becomes

$$l(\xi, H^{-1}(q)|\mathbf{X}_n) = \sum_{i=1}^{n} \log h_{\xi, H^{-1}(q)}(X_i),$$

where $h_{\xi,H^{-1}(q)}(z) = \frac{(1-q)^{-\xi}-1}{\xi H^{-1}(q)} \left(1 + z\frac{(1-q)^{-\xi}-1}{H^{-1}(q)}\right)^{-\frac{1}{\xi}-1}$ is the density function with the new parameter setting. The ML estimates are denoted by $\hat{\xi}$ and $\widehat{H^{-1}(q)}$. Now we apply the weights – described in Section 3.3 – to the log-likelihood function:

$$l^*(\xi, H^{-1}(q)|\mathcal{X}_n) = \sum_{i=1}^{n} \tau_{ni} \log h_{\xi,H^{-1}(q)}(X_i).$$

The profile likelihood is widely used to construct confidence intervals for return values (quantiles) or other relevant parameters. The used main concept is the so-called *profile log-likelihood function* ([5], p. 33-36) defined the following way for our problem:

$$l_p(H^{-1}(q)|\mathbf{X}_n) = \max_{\xi} \; l(\xi, H^{-1}(q)|\mathbf{X}_n). \tag{5}$$

The profile log-likelihood function is the maximized log-likelihood with respect to ξ, so it gives the local maxima of the log-likelihood function for different $H^{-1}(q)$ values.

We combined the weighted bootstrap with the profile likelihood method to construct confidence regions for the return values. The bootstrap version of the profile log-likelihood function is simply the appropriate modification of (5):

$$l_p^*(H^{-1}(q)|\mathbf{X}_n) = \max_{\xi} \; l^*(\xi, H^{-1}(q)|\mathbf{X}_n).$$

Let γ be the stochastic limit of the average of the weights' second moments, defined in **A5**. Theorem 2 states that under strong regularity conditions and with assumptions **A1**-**A6** about the weights,

$$\frac{2}{\gamma}\left[l^*(\hat{\xi}, \widehat{H^{-1}(q)}|\mathbf{X}_n) - l_p^*(H^{-1}(q)|\mathbf{X}_n)\right] \xrightarrow[n\to\infty]{d} \chi_1^2. \tag{6}$$

This asymptotic result can be used to construct confidence regions for the return values. In the sequel the level of confidence will be denoted by $1-\alpha$, typically chosen as 0.95 or 0.99; the $(1-\alpha)$-quantile of the χ_1^2-distribution by $c_{1-\alpha}$; and the empirical sample by $\mathbf{x} = (x_1, \ldots, x_n)$. So using (6), the we can construct the weighted profile confidence interval I_α^* as

$$I_\alpha^* = \left\{H^{-1}(q) : l_p^*(H^{-1}(q)|\mathbf{x}) \geq l^*(\hat{\xi}, \widehat{H^{-1}(q)}|\mathbf{x}) - \frac{\gamma \cdot c_{1-\alpha}}{2}\right\}, \tag{7}$$

which is usually wider than the conventional profile likelihood confidence interval and it may outperform the traditional methods. For example simulations showed (they were needed in Section 4.2) much higher coverage percentages for this method, when we constructed confidence intervals for return values of mixed GPD samples.

4 Applications

4.1 Copula fitting and bootstrap to wind speed modelling

This subsection rests on article [1]. We modelled daily maxima measured for about 50 years at sites Hamburg and Fehmarn, two locations in North Germany. Our main goal

was to model the data with copulas and to construct confidence sets.

We fitted different copula models and checked their goodness-of-fit based on Kendall's function, but as the data are dependent, we had to modify the traditional testing procedure. The critical values were computed via circular block bootstrap simulations with a smaller sample size (the notion of effective sample size was applied). This smaller sample and the block size was determined the following way. We fitted a VAR(1) time series to the data, which tended to be appropriate. Then we looked for the optimal block length, solving (2), hence we got $\widehat{b} = 8$. The effective sample size practically means the size of an *independent* sample from the distribution of the data, for which the variance of the sample mean coincides with its observed variance. In higher dimensions, the trace of the covariance matrix may be used. In our case, the effective sample size is $n_e = n \cdot \frac{\frac{\mathrm{tr}(\Sigma)}{n}}{\mathrm{tr}(\mathrm{Cov}^{*8}(\overline{\mathbf{X}}))} = 2580 \cdot \frac{0.3715}{0.6101} = 1571$. Overall, the graphical methods and also the test showed that the Gumbel copula produced the best fit, but without sample size correction, it would have been rejected very strongly.

4.2 Threshold models and the weighted bootstrap in meteorology

This subsection lies on article [2]. The observations we have used, are 63 years of daily precipitation data from the grid E-OBS. We chose 5 grid points in Hungary, close to Budapest, Tapolca, Várpalota, Székesfehérvár and Hatvan. Our aim was to model the extreme precipitations and to investigate whether there is a clear change in our climate or not, furthermore how return values should be expected in the future.

First, we worked with univariate threshold extreme value models. As threshold, 10 mm was chosen, and it turned out that our weighted profile confidence interval (7) for return values performed better in some cases, than other traditional methods, however, the type of weights did not play an important role. Looking at the time dependence of the parameters, there seemed to be significant change during the 63 years long time period, and the increase in the more extreme events appeared to be rather prominent. We got similar conclusions modelling bivariate data with the BGPD II extreme value model. For example, the dependence parameter of the model increased significantly for the Tapolca–Budapest pair. We estimated the ratio of the probabilities of the joint exceedances of the marginal 10-year return values in the time interval 1993–2012, in comparison with 1965–1984, and half of our data pairs showed a substantial increase.

4.3 Generalised block bootstrap in temperature data modelling

This subsection is based on article [3]. We modelled the dependence structure of 5 site pairs in the Carpathian Basin of the gridded temperature database of E-OBS. Our main goal was to investigate the change of dependence structure with testing the homogeneity of two copulas (first and second half of the samples).

First of all, we conducted simulations regarding the properties of the proposed homogeneity test. It turned out that it is consistent and it has reasonable power for relatively

small sample sizes. We have also investigated the effect of the block size for the properties of the test.

After that, we have demonstrated the use of our generalisation of the block bootstrap for determining the p-values of a homogeneity test. We chose the block size by simulations, solving equation (4). The VAR model performed again pretty well, so the the trace of the covariance matrix of the sample mean could be calculated by formula (3). Our main meteorological result is that we have found some significant changes in the dependence structure between the standardised temperature values, which is more obvious at grid point pairs lying farther from each other.

The PhD thesis is based on the following papers of the author:

[1] P. Rakonczai, L. Varga, and A. Zempléni. Copula fitting to autocorrelated data with applications to wind speed modelling. *Annales Universitatis Scientarium de Rolando Eotvos Nominatae, Sectio Computatorica*, 43:3–20, 2014.

[2] L. Varga, P. Rakonczai, and A. Zempléni. Applications of threshold models and the weighted bootstrap for hungarian precipitation data. *Theoretical and applied climatology*, 124(3-4):641–652, 2016.

[3] L. Varga and A. Zempléni. Generalised block bootstrap and its use in meteorology. *Advances in Statistical Climatology, Meteorology and Oceanography*, 3(1):55–66, 2017.

Further references:

[4] A. A. Borovkov and A. M. Mathematical statistics. *Gordon Breach, Amsterdam*, 1998.

[5] S. Coles. *An introduction to statistical modeling of extreme values*. Springer Verlag, 2001.

[6] B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1), 1979.

[7] B. Efron. *The jackknife, the bootstrap and other resampling plans*. CBMS-NFS, 1982.

[8] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[9] P. Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

[10] P. Hall, J. L. Horowitz, and B.-Y. Jing. On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574, 1995.

[11] S. N. Lahiri. *Resampling methods for dependent data*. Springer Science & Business Media, 2003.

[12] S. N. Lahiri, K. Furukawa, and Y.-D. Lee. A nonparametric plug-in rule for selecting optimal block lengths for block bootstrap methods. *Statistical Methodology*, 4(3):292–321, 2007.

[13] Michael A Newton and Adrian E Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48, 1994.

[14] J. Shao and D. Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012.

[15] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.