BOOTSTRAP MÓDSZEREK ÉS ALKALMAZÁSAIK

Doktori értekezés tézisei

VARGA LÁSZLÓ

Témavezető: Zempléni András
Egyetemi docens, CSc

Matematika Doktori Iskola
Vezető: Faragó István

Alkalmazott Matematika Doktori Program
Vezető: Karátson János

Eötvös Loránd Tudományegyetem

Valószínűségelméleti és Statisztika Tanszék

Természettudományi Kar

2017

1. Bevezetés

A disszertáció az elméleti/alkalmazott statisztikai eljárások egy számításigényes családjáról szól: bootstrap módszerekről. A PhD dolgozat több nézőpontot mutat be: vázolja a legfontosabb bootstrap módszereket, megvilágítja az elméleti eredmények mögött rejlő matematika nehézségeit, új módszerekkel bővíti a bootstrap elméletét és bemutatja azok gyakorlati alkalmazhatóságát.

Ez a tézisfüzet nagyrészt a disszertáció felépítését követi. A PhD dolgozat a szerző [1], [2] és [3] publikációin nyugszik, melyek az elméleti eredményeket különböző meteorológiai jelenségek modellezésére alkalmazzák.

2. Fejezetek valószínűségelméletből és statisztikából

Ez a szakasz valószínűségelméletből, idősorok elméletéből (stacionárius folyamatok, vektor autoregresszió), a kopulák elméletéből (illeszkedésvizsgálat a Kendall-függvény segítségével, kopulák homogenitásvizsgálata) és extrém érték elméletből (egy- és kétváltozós maximumon alapuló és küszöbmeghaladási modellek) tartalmaz a későbbi fejezetek számára szükséges megközelítéseket és eredményeket.

3. Bootstrap módszerek

A bootstrap egy rendszerint visszatevéses mintavételen alapuló statisztikai eljárás, amit számos statisztikai feladat megoldására lehet használni: a bennünket érdeklő statisztika eloszlásának becslésére, torzítás csökkentésére, hipotézisvizsgálatra, konfidenciaintervallumok és -halmazok készítésére, idősorok előrejelzésére stb.

A bootstrap módszereket az elmúlt évszázad utolsó két évtizedében fejlesztették ki kiváló tudósok. A fő koncepciót Bradley Efron vezette be klasszikus cikkében ([6]), és azóta – köszönhetően a számos kiterjesztésnek és általánosításnak – az egyik legszélesebb körben elterjedt Monte–Carlo–módszerré vált. A bootstrap módszerek gyakorlati alkalmazhatósága exponenciálisan megnőtt a számítógépes hardware és a programozási nyelvek gyors fejlődésének köszönhetően. A bevezetését követő években a bootstrap számos korlátjára derült fény, melyek az eredeti koncepció módosításához és rengeteg kiterjesztéshez vezettek, így megszületett a paraméteres/félparaméteres bootstrap, reziduális bootstrap, blokk bootstrap, súlyozott bootstrap, dupla/tripla bootstrap és az n-ből-m (angolban m-out-of-n) bootstrap. Gondolkozásomra és ezáltal a disszertációra a [9], [8] és [11] tankönyvek gyakorolták a legnagyobb hatást.

3.1. A bootstrap alapelve

Az **i.i.d. bootstrap** alapötlete az, hogy az eredeti mintából visszatevéses mintavételezéssel további mintákat veszünk. Formálisan felírva, legyen $\mathcal{X}_n = (X_1, \ldots, X_n)^T$ egy

i.i.d., valószínűségi változókból álló sorozat ismeretlen F eloszlásfüggvénnyel és legyen $T_n = t_n(\mathcal{X}_n; F)$ egy bennünket érdeklő statisztika (például az $\overline{X}$ mintaátlag). Az X_i valószínűségi változók akár vektorváltozók is lehetnek – ilyenkor $\mathcal{X}_n$ mátrix lesz. Rendszerint az a fő célkitűzés, hogy T_n egy bizonyos függvényének az eloszlását megbecsüljük, például gyakran van szükségünk T_n szórására vagy egy magas kvantilisére.

Adott $\mathcal{X}_n$ mintára, P_*, E_*, D_*^2 and Cov_* fogja jelölni a feltételes valószínűséget, a feltételes várható értéket, a feltételes varianciát és a feltételes kovarianciát, például $P_*(\,\cdot\,) = P(\,\cdot\,|\mathcal{X}_n)$. Az i.i.d. bootstrap módszert formalizálhatjuk is: egy adott $\mathcal{X}_n$ mintából mint alaphalmazból m elemű $\mathcal{X}_m^* = \{X_1^*, \ldots, X_m^*\}$ véletlen mintákat veszünk:

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

így a bootstrap minta elemei feltételesen függetlenek és azonos eloszlásúak lesznek. Ezáltal az X_i^* boostrap mintaelemek közös eloszlását az $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x)$ empirikus eloszlásfüggvény határozza meg. A bootstrap minta nagysága rendszerint megegyezik az eredeti minta méretével. A következő lépés a statisztika bootstrap verziójának definiálása: $T_{m,n}^* = t_m(\mathcal{X}_m^*; F_n)$. Ha az eljárást sokszor megismételjük, akkor T_n ismeretlen G_n eloszlását a bootstrap verziók G_m^* eloszlásával becsülhetjük. A problémák többségében a G_m^*-tól függő mennyiségek kiszámításához számítógépes szimulációkra van szükség.

A matematikai elmélet fejlesztése során az egyik legfontosabb szempont annak a vizsgálata (legalább szimulációkkal), hogy a statisztika bootstrap eloszlása elég közel van-e az eredeti eloszláshoz. Azt mondjuk, hogy a bootstrap gyengén/erősen konzisztens, amennyiben egy alkalmasan választott metrikában a két eloszlás távolsága sztochasztikusan/egy valószínűséggel 0-hoz tart – lásd [14] 3.1 fejezetét a téma bővebb kifejtéséért.

3.2. Blokk bootstrap módszerek

Amennyiben adataink összefüggők és stacionáriusak, akkor a blokk bootstrap a legelfogadottabb újramintavételezési módszer; [11] alaposan tárgyalja ezt az eljárást. A blokk bootstrap alapötlete az a szándék, hogy megpróbáljuk az összefüggőséget átörökíteni a mintákra. Ennek érdekében egy-egy mintaelem helyett egész blokkokból veszünk újabb mintákat, majd ezeket a blokkokat összerakjuk, precízebben leírva:

1. Tekerjük fel az $X_1, \ldots, X_n$ adatokat egy körvonalra, azaz definiáljuk az $\widetilde{X}_t = X_{t_{\mathrm{mod(n)}}}$ $(t \in \mathbb{Z}_+)$ sorozatot, ahol $\mathrm{mod(n)}$ a "modulo n" osztást jelöli. Ez azt jelenti, hogy $\widetilde{X}_k = \widetilde{X}_{k+n} = \widetilde{X}_{k+2n} = \ldots = X_k$ minden $k \in \{1, 2, \ldots, n\}$-re.

2. Határozzuk meg a blokkok kezdőindexeit: az $A \subseteq \{1, \ldots, n\}$ halmazra koncentrált $I_1, I_2, \ldots$ valószínűségi változó sorozatot.

3. Határozzuk meg a blokkok hosszát: az $L_1, L_2, \ldots$ nemnegatív egész értékű valószínűségi változó sorozatot.

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

5. Rakjuk össze a blokkokat: $\mathcal{X}^* = \{B(I_1, L_1), B(I_2, L_2), \ldots\}$.

A klasszikus blokk bootstrap a mozgó blokk bootstrap (MBB), ahol a blokkméret egy rögzített $1 \leq b \leq n$ egész szám és a blokkokat az eredeti mintából veszik azonos valószínűséggel, azaz a blokkok kezdőindexei egyenletes eloszlásúak az $A = \{1, 2, \ldots, n-b+1\}$ halmazon. A cirkuláris blokk bootstrap (CBB) mindössze annyiban különbözik az MBB-től, hogy $A = \{1, 2, \ldots, n\}$. A stacionárius blokk bootstrap (SBB) a CBB általánosítása, a blokkméretek független geometriai eloszlásúak $p \in (0, 1]$ paraméterrel.

Általánosított blokk bootstrap

A PhD dolgozatban egy olyan blokk bootstrap módszert vezettünk be, amely kiküszöbölte azt a problémát, amit az egész blokknagyságok okoztak – a konkrét motivációt lásd a 3.4 fejezetben. Kiterjesztésünkben a blokkméret valószínűségi változó, a módszer pedig a CBB-t is magában foglalja speciális esetként.

Amennyiben $1 \leq b \in \mathbb{R}$, akkor legyen az általánosított blokk bootstrap minta a következő. Tekerjük fel most is a mintát egy körvonalra. Tetszőleges $k \in \{1, 2, \ldots, n\}$ esetén a blokkok legyenek az alábbiak (hosszuk vagy $\lfloor b \rfloor$, vagy $\lceil b \rceil$) :

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

ahol $\lceil b \rceil$ jelöli b felső, míg $\lfloor b \rfloor$ az alsó egészrészét. Végül illesszük össze a blokkokat. Az előzőekben leírt **általánosított blokk bootstrap** (GBB) "paramétereit" a blokk bootstrap-nél bevezetett jelölésekkel is felírhatjuk:

- $1 \leq b \in \mathbb{R}$ az elvárt blokkméret, amit előre rögzítünk
- $A = \{1, 2, \ldots, n\}$
- $I_i | \mathcal{X}_n \sim \text{Unif}(A) \quad i = 1, 2, \ldots$ feltételesen függetlenek egymástól
- $P_*(L_i = \lceil b \rceil) = 1 - P_*(L_i = \lfloor b \rfloor) = b - \lfloor b \rfloor \quad i = 1, 2, \ldots$ feltételesen függetlenek egymástól
- I_i és L_j feltételesen függetlenek minden i és j esetén

1. Állítás. *A GBB módszer esetén minden $i = 1, 2, \ldots$-re azt kapjuk a blokkméretekre, hogy $E_* L_i = b$ és $D_*^2 L_i = (b - \lfloor b \rfloor)(1 - b - \lfloor b \rfloor)$.*

A cirkuláris blokk bootstrap mintához hasonlóan a mi általánosított bootstrap mintánk is rendszerint nem stacionárius folyamat (az eredeti mintára feltételesen). Mostantól fel fogjuk tenni, hogy a bootstrap minta hossza megegyezik az eredeti mintamérettel, azaz $m = n$. Definiáljuk a következő valószínűségi változókat:

- N_s: a $\lfloor b \rfloor$ méretű blokkok száma;

- N_l: a $\lceil b \rceil$ méretű blokkok száma;
- R: a maradék blokkméret hossza, azaz $R = n - N_s \cdot \lfloor b \rfloor - N_l \cdot \lceil b \rceil$.

Az alábbi állítás megadja N_s eloszlását, amiből N_l és R eloszlása már könnyedén kiszámolható.

2. Állítás. *Legyen $p = b - \lfloor b \rfloor$, ekkor N_s eloszlása a következő: $j = 0, 1, \ldots, \left\lfloor \frac{n}{\lfloor b \rfloor} \right\rfloor$-re*

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

Az alkalmazásokban a bootstrap mintaátlag kovarianciamátrixának nyomára lesz szükség, ebben segít a következő tétel.

1. Tétel. *A bootstrap átlag kovarianciamátrixát az alábbi módon lehet kiszámítani:*

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

ahol $\overline{\mathbf{X}}_{b,i}^*$ az i-edik b méretű blokk átlaga ($i = 1, 2, \ldots$).

3.3. Súlyozott bootstrap

A súlyozott (weighted vagy multiplier) bootstrap az i.i.d. bootstrap kiterjesztésének tekinthető. A klasszikus súlyozott bootstrap ötlete először a [7] könyv 10. fejezetében jelent meg és a későbbiekben számos alkalmazásra lelt.

Az elmúlt években kutatásaim egyik fókuszpontjában ennek az elméletnek egy részterülete, a **súlyozott likelihood bootstrap** állt. Úgynevezett bootstrap súlyokat vezetünk be, melyeket $\boldsymbol{\tau}_n = (\tau_{n,1}, \tau_{n,1}, \ldots, \tau_{n,n})$-nel jelölünk és feltesszük róluk, hogy az $\mathcal{X}_n$ mintához tartozó valószínűségi változók. [13] a súlyozott bootstrap-et a maximum likelihood becsléssel kombinálta úgy, hogy a log-likelihood függvény elemeit megszorozta a megfelelő súlyokkal. Ebben a kontextusban $P_*(\cdot)$ olyan feltételes valószínűséget jelöl, amikor a súlyok véletlenek, a minta viszont rögzített. A disszertációban Wilks klasszikus, az általánosított likelihood-hányados tesztstatisztikára vonatkozó eredményének ([15]) egy további általánosítását és annak bizonyítását mutatjuk be.

Tegyük fel, hogy adott egy eloszláscsalád $f_{\boldsymbol{\vartheta}}(x)$ sűrűségfüggvénnyel, ahol $\boldsymbol{\vartheta} \in \Theta \subseteq \mathbb{R}^p$ ismeretlen paraméter. Egy $\mathcal{X}_n = (X_1, \ldots, X_n)^T$ i.i.d. minta log-likelihood függvényét $l(\boldsymbol{\vartheta}|\mathcal{X}_n) = l(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{\vartheta}}(X_i)$ fogja jelölni, a paraméter maximum likelihood becslését pedig $\widehat{\boldsymbol{\vartheta}}_n = \arg\max_{\boldsymbol{\vartheta}} l(\boldsymbol{\vartheta})$. Definiáljuk a log-likelihood függvény (bootstrap) súlyozott verzióját az alábbi módon:

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

és legyen $\widehat{\boldsymbol{\vartheta}}_n^*$ a súlyozott ML-becslés.

A súlyokra tett feltételek rendszerint kontextusról kontextusra változnak, ezért csak a mi feladatunkra vonatkozó feltételrendszert fogjuk bemutatni. Tegyük fel, hogy az alábbi *feltételek* teljesülnek *a bootstrap súlyokra*:

A1. függetlenek az adatgeneráló folyamattól;

A2. véges második momentummal rendelkeznek minden $n = 1, 2, \ldots$ esetén;

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. Létezik egy olyan $\gamma \in \mathbb{R}$, amire $\quad \frac{1}{n} \sum_{i=1}^{n} \tau_{n,i}^2 \xrightarrow[n\to\infty]{p} \gamma;$

A6. Létezik egy olyan $|q| < 1$ valós szám, amire $\mathrm{Cov}(\tau_{n,i}, \tau_{n,j}) \leq q^{|i-j|} \quad 1 \leq i \neq j \leq n; \;\; n = 1, 2, \ldots$

A fenti feltételeknek számos eloszlás eleget tesz, mi az alkalmazásoknál az i.i.d. exponenciális és a polinomiális (multinomiális) eloszlást használtuk:

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

Egyrészt az egyszerűségük miatt választottuk ezeket, másrészt azért, hogy megnézzük, a koordináták közötti gyenge összefüggőség (polinomiális eloszlás) jelentős hatást gyakorol-e az adott probléma esetén a végeredményre (például a 4.2 fejezetben a konfidenciaintervallumok lefedési valószínűségére).

Tegyük fel, hogy az eloszláscsaládra standard erős regularitási feltételek teljesülnek, például a [4] 191. oldalán lévő (RR). Készítsünk a paramétertérből egy két részből álló partíciót: $\boldsymbol{\vartheta} = \begin{pmatrix} \boldsymbol{\sigma} \\ \boldsymbol{\rho} \end{pmatrix} \begin{matrix} \}q \\ \}p-q \end{matrix}$ és legyen $\boldsymbol{\sigma}_\bullet \in \mathrm{int}(Pr_H(\Theta))$, ahol H a Θ paramétertér első q koordinátájának megfelelő altér. Definiáljuk a korlátozott ML-becslést az alábbi módon:

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

Jelölje $\widetilde{\boldsymbol{\vartheta}}_n^*$ a fenti (1) súlyozott verzióját. Wilks eredménye nyomán tudjuk, hogy $\boldsymbol{\sigma} = \boldsymbol{\sigma}_\bullet$ esetén $T_n := 2\left[l(\widehat{\boldsymbol{\vartheta}}_n) - l(\widetilde{\boldsymbol{\vartheta}}_n)\right] \xrightarrow[n\to\infty]{d} \chi_q^2$.

2. Tétel. ***A véletlen súlyozású általánosított likelihood-hányados statisztika aszimptotikus eloszlása.***
Erős regularitási feltételek esetén és használva az eddigi jelöléseket; ha $\boldsymbol{\sigma} = \boldsymbol{\sigma}_\bullet$, akkor

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

A 2. tétel bizonyításához további állításokra volt szükségünk. Az első a többdimenziós határeloszlás-tétel, míg a második a nagy számok gyenge törvényének egy általánosítása véletlen súlyokkal.

3. Állítás. *Legyenek $\boldsymbol{\tau}_n$-ek az ebben a fejezetben bevezetett, **A1-A6** feltételeknek eleget tevő valószínűségi változók (súlyok); $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ i.i.d. p dimenziós valószínűségi vektorváltozók $\mathbf{0}_p$ várható érték vektorral és $\boldsymbol{\Sigma}$ kovarianciamátrixszal. Amennyiben a súlyok függetlenek az $\mathbf{Y}_i$ valószínűségi vektorváltozóktól, akkor*

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

4. Állítás. *Legyenek $\boldsymbol{\tau}_n$-ek az ebben a fejezetben bevezetett, **A1-A6** feltételeknek eleget tevő valószínűségi változók (súlyok); $Y_1, Y_2, \ldots$ i.i.d. valószínűségi változók véges első két momentummal. Amennyiben a súlyok függetlenek az Y_i valószínűségi változóktól, akkor*

$$\frac{1}{n}\sum_{i=1}^{n} \tau_{n,i}Y_i \xrightarrow[n\to\infty]{p} EY_1.$$

3.4. Blokkméret megállapítása a gyakorlatban

A szakirodalomban két általános stratégiát szoktak javasolni az ideális blokkméret kiválasztására, az egyik almintavételezésen ([10], subsampling), a másik nemparaméteres behelyettesítésen ([12], nonparametric plugin) alapul. A 4.1 és 4.3 fejezetekben egy ezektől eltérő, modell alapú megközelítést mutatunk be: úgy próbáljuk megtalálni a legjobb blokkméretet, hogy először egy reményeink szerint megfelelően illeszkedő vektor autoregressziós – VAR(p) – modellt illesztünk a többdimenziós adatokra, majd megnézzük, melyik blokkméretre lesz az eredeti mintából blokk bootstrap-pel vett minta mintaátlagának kovarianciamátrixa legközelebb a VAR-modellből származó minta mintaátlagának kovarianciamátrixához. Ez az eljárás kellően általános ahhoz, hogy más, a VAR-nál akár jóval bonyolultabb sztochasztikus folyamatokra is alkalmazni lehessen.

A 4.1 fejezetben az optimális $\widehat{b}$ blokkméretet a következő képlettel számítjuk:

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

ahol $\mathrm{Cov}_*(\overline{\mathbf{X}}_b^*)$ a blokk bootstrap átlag kovarianciamátrixa b blokknagyság esetén és a $\mathrm{Cov}\left(\overline{\mathbf{X}}_{VAR}\right)$ mennyiséget az alábbi módon kapjuk meg. Elegendő csak VAR(1) folyamatokkal foglalkozni, mert egy d dimenziós VAR(p) felírható pd dimenziós VAR(1)-ként. Ha van egy d dimenziós VAR(1) folyamatunk $\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t$, $\mathrm{Cov}(\boldsymbol{\varepsilon}_t) = \mathbf{C}$ alakban, akkor

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

ahol $\boldsymbol{\Gamma}_X$ az autokovariancia mátrix és $\mathrm{vec}(\boldsymbol{\Gamma}_X(0)) = (\mathbf{I}_{d^2} - \mathbf{A}\otimes\mathbf{A})^{-1}\mathrm{vec}(\mathbf{C})$.

A szakirodalomban rendszerint egész blokkmérettel végzik a szimulációkat. Azonban azt tapasztaltuk, hogy a (2) képlettel kapott egész blokkméretek esetén néha igencsak jelentős a két nyom közti eltérés, ami akár jelentős torzítást is okozhat, főleg kisebb blokknagyságok esetén. Sajnos hasonló a helyzet az alfejezet elején megemlített, a szakirodalomban széles népszerűségnek örvendő két általános technikával is, ez volt a fő motivációja a 3.2 fejezetben bevezetett általánosított blokk bootstrap-nek. Ezért aztán a (2) képlet helyett az alábbi egyenlet megoldását javasoljuk az ismeretlen $1 \leq b \in \mathbb{R}$ változó szerint:

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

A 4.3 fejezetben ezt a megközelítést követjük a blokkméret meghatározása során.

3.5. Profil likelihood és az extrémumok bootstrap-ezése

A 4.2 fejezetben a korábban kimondott 2. tételt fogjuk arra használni, hogy konfidenciaintervallumot készítsünk a küszöbmeghaladáson alapuló egyváltozós extrémérték-eloszlás visszatérési értékeire. A Pickands–Balkema–de Haan tétel alapján tudjuk, hogy egy adott küszöbérték felett a megfigyelések általánosított Pareto-eloszlással (GPD) közelíthetők, amely eloszlás az alábbi eloszlásfüggvénnyel rendelkezik:

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

ahol ξ-t alakparaméternek, σ-t pedig skálaparaméternek hívják. Mi az alkalmazásokban egy másfajta paraméterezéssel dolgoztunk: ξ és a q-kvantilis (visszatérési érték) $H^{-1}(q)$ voltak a paramétereink, ekkor a log-likelihood függvény a következő alakot ölti:

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

ahol $h_{\xi,H^{-1}(q)}(z) = \frac{(1-q)^{-\xi}-1}{\xi H^{-1}(q)}\left(1 + z\frac{(1-q)^{-\xi}-1}{H^{-1}(q)}\right)^{-\frac{1}{\xi}-1}$ az új paraméterezésű sűrűségfüggvény. Jelölje az ML-becsléseket $\hat{\xi}$ és $\widehat{H^{-1}(q)}$. Most bevetjük a 3.3 fejezetben beve-

zetett súlyokat – a log-likelihood függvény elemeit szorozzuk meg velük:

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

A profil likelihood egy széles körben használt módszer arra, hogy visszatérési értékek (kvantilisek) vagy más fontos paraméterek értékeire konfidenciaintervallumot konstruáljunk. Az eljárás alapja az ún. *profil log-likelihood függvény* ([5], p. 33-36), amit jelen esetben a következőképp definiálhatunk:

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

Tehát az l_p függvény rögzített kvantilis értékekre a log-likelihood függvény ξ szerinti lokális maximumát adja meg.

A súlyozott bootstrap-et a profil likelihood módszerrel kombináltuk, hogy a visszatérési értékekre konfidenciaintervallumot határozzunk meg. A profil log-likelihood függvény bootstrap verziója (5) értelemszerű módozata:

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

Legyen γ a 3.3 fejezetbeli **A5** feltételben szereplő konstans, ami a súlyok második momentumának átlagából számolt sztochasztikus határérték. A 2. tétel szerint erős regularitási feltételek mellett, amennyiben a súlyokra az **A1–A6** feltételek teljesülnek, akkor

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

Ezt az aszimptotikus eredményt felhasználhatjuk arra, hogy a visszatérési értékekre konfidenciaintervallumot konstruáljunk. A továbbiakban jelölje $1-\alpha$ a megbízhatósági szintet, $c_{1-\alpha}$ a χ_1^2-eloszlás $(1-\alpha)$-kvantilisét és $\mathbf{x} = (x_1, \ldots, x_n)$ a tapasztalati mintát. Ezáltal (6)-t felhasználva, az alábbi I_α^* súlyozott profil konfidenciaintervallumot készíthetjük:

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

amely rendszerint szélesebb a hagyományos profil likelihood konfidenciaintervallumnál és gyakran jobban is teljesít nála. Szimulációink azt mutatták, jóval pontosabb lefedési valószínűséggel rendelkezik a hagyományos profil intervallumhoz képest, amennyiben a minta kevert GPD eloszlásból származik (a 4.2. fejezetben volt rá szükség).

4. Alkalmazások

4.1. Kopulaillesztés és bootstrap szélsebességi adatok modellezésében

Ez az alfejezet [1] cikk alapján készült. Két észak-német állomás, Hamburg és Fehmarn 50 éves napi szélsebességi maximumait modelleztük. Fő célunk az volt, hogy az össze-

függőségi struktúrát kopulákkal elemezzük és előrejelzéseket készítsünk.

Különböző kopula modelleket illesztettünk és az illeszkedést a Kendall-függvény segítségével ellenőriztük, de mivel adataink összefüggők voltak, a hagyományos tesztelési eljárás módosításra szorult. A kritikus értékeket CBB elven vett, kisebb elemszámú mintákból generáltuk, felhasználva az effektív mintaméret fogalmát. Ezt a kisebb mintaelemszámot és a blokkméretet a következő módon határoztuk meg. Először egy VAR(1) folyamatot illesztettünk az adatokra, ami jónak bizonyult, majd a (2) képletet megoldva, optimális blokknagyságnak $\widehat{b} = 8$ adódott. Az effektív mintaméret azt a mintanagyságot jelenti, amivel egy *független* mintából vett minta mintaátlagának a varianciája megegyezik a megfigyelt, összefüggőséget is magában tartalmazó mintaátlag varianciájával. Több dimenziós megfigyelések esetén a variancia helyett a kovarianciamátrix nyomát lehet használni. A mi esetünkben az effektív mintaméretre $n_e = n \cdot \frac{\frac{\mathrm{tr}(\Sigma)}{n}}{\mathrm{tr}(\mathrm{Cov}^{*8}(\overline{\mathbf{X}}))} = 2580 \cdot \frac{0.3715}{0.6101} = 1571$ adódott. Összességében azt kaptuk, hogy a grafikus módszerek és a tesztek szerint egyöntetűen a Gumbel kopula illeszkedett a legjobban, de mintaméret-korrekció nélkül a Gumbel kopula illeszkedését is erősen elutasítottuk volna.

4.2. Küszöbmeghaladási modellek és a súlyozott bootstrap a meteorológiában

Ez az alfejezet a [2] cikket mutatja be. A felhasznált megfigyelések az E-OBS adatbázis 63 éves napi csapadékadataiból származnak, öt magyarországi állomást választottunk Budapest, Tapolca, Várpalota, Székesfehérvár és Hatvan településekhez közel. Az elemzés célja az volt, hogy modellezzük az csapadékok kiugró értékeit; megvizsgáljuk, ebből a szempontból megfigyelhető-e változás a klímánkban; illetve magas visszatérési szintekhez tartozó visszatérési értékekre intervallumbecslést adjunk.

Először egyváltozós küszöbérték-modellekkel foglalkoztunk. Küszöbértéknek 10 mm-t választottuk, azt kaptuk, hogy a (7) súlyozott profil likelihood intervallum számos esetben jobban teljesített, mint a hagyományos profil intervallumbecslés, viszont a súlyok eloszlása nem bizonyult fontos tényezőnek. A vizsgált 63 év alatt a GPD eloszlás paraméterei az idő függvényében szignifikáns módon megváltoztak, ez a változás pedig a magas visszatérési értékeknél különösen jelentős volt, megerősítve azt, hogy jóval gyakrabbá váltak a szélsőséges időjárási események. Hasonló eredményekre jutottunk a kétváltozós BGPD II extrém érték modell illesztése során is, például a Tapolca–Budapest párok esetén az összefüggőségi paraméter értéke szignifikáns módon megnőtt, illetve a vizsgált állomáspárok felénél a 10 éves visszatérési értékeknek megfelelő extrém események bekövetkezésének együttes valószínűsége erőteljes emelkedést mutatott.

4.3. Általánosított blokk bootstrap alkalmazása hőmérsékleti adatok modellezésében

Ez az alfejezet [3] cikk alapján készült. Az E-OBS adatbázisban található 5 kárpát-medencei állomáspár összefüggőségi struktúráját modelleztük. Azt a célt tűztük ki ma-

gunk elé, hogy a minták első és második felének összefüggőségi struktúráját kopulák homogenitásvizsgálatával összehasonlítsuk egymással.

Mindenekelőtt szimulációkat hajtottunk végre annak érdekében, hogy a kopula homogenitásvizsgálat teszt erejét bootstrap esetén is megvizsgáljuk. Azt kaptuk, hogy a próba konzisztens és még kis mintára is elfogadható ereje van. Megnéztük továbbá a blokkméret tesztre gyakorolt hatását, ami az esetek többségében meglehetősen gyengének bizonyult.

Ezután a disszertációban bevezetett általánosított blokk bootstrap segítségével p-értékeket szimuláltunk. A blokkméretet a (4) képlet megoldásával határoztuk meg. A VAR modell most is jól illeszkedett, így a mintaátlag kovarianciamátrixát (3) képlettel lehetett számítani. Meteorológiai szempontból [3] cikkünknek az volt a fő következtetése, hogy a hőmérséklet-adatok összefüggőségi struktúrájában változást lehet megfigyelni, ami annál erősebb, minél távolabb van egymástól a két állomáspár.

A PhD értekezés alapjául szolgáló publikációk:

[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.

További hivatkozások:

[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.