

Tőzsdeindexek elemzése extrémérték-elmélettel

Kovács Dávid

témavezető: Dr. Zempléni András

Bevezetés

Jelen dolgozat célja az extrémérték-elmélet módszereivel modellezni tőzsdeindexek extrém veszteségeit. Az első fejezet a Dow Jones Ipari Átlag és Dow Jones Közszolgáltatói Átlag extrém veszteségeit vizsgálja, különös tekintettel arra a kérdésre, hogy homogénnak tekinthetőek-e az egyes adatsorok. Ez a vártnál bonyolultabb kérdésnek bizonyult. A második fejezet a két tőzsdeindex együttes elemzését tárgyalja. Az első fejezet váratlan terjedelme miatt, csak gazdasági válságok detektálásra jutott idő a második fejezetben.

1. A marginálisok elemzése

1.1. A cél

Rendelkezésünkre állnak 1970-től a Dow Jones Ipari Átlag (DJIA) és 1997-től a Dow Jones Közszolgáltatói Átlag (DJUA) napi árfolyamai. Ebből képezzük a napi negatív loghozamokat, majd blokkmaximumot veszünk (ily módon a nagy veszteségeket vizsgáljuk). Általánosan egy tetszőleges adatsorra a blokkmaximumok legyenek X_1, X_2, \dots, X_N . Megfelelő blokkméret esetén feltehetjük, hogy $X_t \sim GEV(\mu_t, \sigma_t, \xi_t)$. Azonban μ_t, σ_t, ξ_t nem konstans paraméterek. Azt gondolhatjuk, hogy $\exists 1 = t_1, t_2, \dots, t_m = N$ időpontok, melyek mellett μ_t, σ_t, ξ_t konstans, ha $t \in [t_k, t_{k+1} - 1]$. Továbbá feltételezhetjük, hogy bármely időpontot elhagyva, már nem teljesül, ez a szakaszonkénti homogenitás. Célunk becslést adni ezekre az időpontokra, valamint a paraméterekre az egyes intervallumokon belül. Tehát azonos eloszlású részekre szeretnénk osztani a mintánkat.

1.2. A hipotézisvizsgálat

Legyen $[s, t]$ egy adott szakasz. A nullhipotézis (H_0) legyen az, hogy azonos GEV eloszlású a szakasz, azaz

$$X_s, X_{s+1}, \dots, X_t \sim GEV(\mu, \sigma, \xi)$$

valamilyen $\mu, \xi \in \mathbb{R}$ és $\sigma \in \mathbb{R}_+$ paraméterekre. Az ellenhipotézis pedig legyen az, hogy két különböző GEV eloszlású részre bontható a szakasz, azaz $\exists u^* \in$

$[s + M, t - M] \cap \mathbb{N}$, melyre

$$X_s, X_{s+1}, \dots, X_{u^*} \sim GEV(\mu', \sigma', \xi')$$

$$X_{u^*+1}, X_{u^*+2}, \dots, X_t \sim GEV(\mu'', \sigma'', \xi'')$$

valamilyen $\mu', \mu'', \xi', \xi'' \in \mathbb{R}$ és $\sigma', \sigma'' \in \mathbb{R}_+$, a $(\mu' \neq \mu'', \sigma' \neq \sigma'', \xi' \neq \xi'')$ relációk közül legalább egyet teljesítő paraméterekre. Azért a $[s + M, t - M]$ intervallumon keressük u^* -t, mert a felosztás eredményeként nem akarunk túl kicsi intervallumokat kapni, valamint túl kevés értékre nem érdemes eloszlást illeszteni. M alkalmas megválasztása is a feladatunk lesz.

Írjuk fel a megfelelő loglikelihoodok különbségét. Legyen

$$l_0 = \max \left\{ \sum_{i=s}^t \ln(f_{\mu, \sigma, \xi}(X_i)) : \mu, \xi \in \mathbb{R}, \sigma \in \mathbb{R}_+ \right\}$$

Rögzített $u \in [s + M, t - M] \cap \mathbb{N}$ mellett legyen

$$l_1(u) = \max \left\{ \sum_{i=s}^u \ln(f_{\mu', \sigma', \xi'}(X_i)) + \sum_{i=u+1}^t \ln(f_{\mu'', \sigma'', \xi''}(X_i)) : \mu', \mu'', \xi', \xi'' \in \mathbb{R}, \sigma', \sigma'' \in \mathbb{R}_+ \right\}$$

Továbbá legyen $l_1 = \max\{l_1(u) : u \in [s + M, t - M] \cap \mathbb{N}\}$, $u^* = \operatorname{argmax} l_1(u)$, $l^* = l_1 - l_0$. Az optimális u^* osztóindex is egy paraméter, amit becsülnünk kell, így nem alkalmazható a likelihood-hányados statisztikára vonatkozó aszimptotikus eredmény. Tehát nem igaz, hogy $2l^* \sim \chi_3^2$ aszimptotikusan. Így a p -érték meghatározásához szimulációs módszerekhez folyamodunk.

1.3. A szimuláció

Legyenek $X_1, X_2, \dots, X_N \sim GEV(\mu, \sigma, \xi)$ valószínűségi változók. Vizsgáljuk, hogy az ezekből származó l^* statisztika eloszlása függ-e μ, σ, ξ és N értékétől. Nézzünk egy konkrét, az adatsor alapján reális példát. Legyen $\mu_1 = 0, \sigma_1 = 0,001, \xi_1 = 0,3$ és $N_1 = 150$. Valamint legyen $\mu_2 = 0,1, \sigma_2 = 0,02, \xi_2 = 0,6$ és $N_2 = 75$. Szimuláljunk mindkét paraméter-négyes mellett 1000 adatsort, majd számoljuk ki mindegyik szimulált sorozatra a megfelelő l^* -statisztikát. A kapott két adatsorra így alkalmazhatunk egy χ^2 homogenitás vizsgálatot. Itt minden észszerű csoportszám mellett $p < 10^{-5}$ adódik, így határozottan elutasítható a homogenitás hipotézise.

Legyenek adottak X_s, X_{s+1}, \dots, X_t és tekintsük az előző alfejezetben leírt nullhipotézist és ellenhipotézist. Számoljuk ki a mintára az l^* -statisztikát. Célunk szimulációval meghatározni az l^* -próbat statisztika p -értéket. Legyen $\hat{\theta} = (\hat{\mu}, \hat{\sigma}, \hat{\xi})$ a H_0 -beli közös μ, σ, ξ paraméterekre kapott ML becslés, I pedig a paraméterekre a Fisher-féle információs mátrix. Generáljunk $N_3(\hat{\theta}, I^{-1}(\hat{\theta}))$ eloszlással n -szer véletlen μ, σ, ξ paramétereket és ezen paraméterekkel generáljunk

$t-s+1$ hosszú adatsorokat. Mind az n adatsorra számoljuk ki az l^* -statisztikát. Ha az i . generált adatsorból kapott l^* -statisztikát l_i^* jelöli, akkor a becült p -érték $\hat{p} = \frac{1}{n} \sum_{i=1}^n I(l^* < l_i^*)$.

1.4. Az algoritmus

Definíció. Egy $[s, t]$ szakaszt nevezünk homogénnek, ha ott elfogadható H_0 az l^* -próba alapján. Valamint nevezük inhomogénnek, ha H_0 elutasítható

Definíció. Legyen $\mathcal{U} = \{u_1, u_2, \dots, u_n\}$ osztópontok egy halmaza. Nevezük \mathcal{U} -t elfogadhatónak, ha $\forall i = 1, 2, \dots, n-1$ esetén $[u_i, u_{i+1}]$ homogén. Valamint nevezük \mathcal{U} -t minimálisnak, ha $\forall i = 1, 2, \dots, n-2$ esetén $[u_i, u_{i+2}]$ inhomogén.

Célunk minimális elfogadható \mathcal{U} -t találni. Legyen H egy olyan függvény, mely adott \mathcal{U} indexhalmazt elfogadhatóvá tesz, oly módon, hogy addig bővíti a halmazt az optimális u^* indexekkel, amíg az elfogadható nem lesz. M pedig legyen egy olyan függvény, mely adott \mathcal{U} -nak eltávolítja a fölösleges indexeit, azaz minimálissá teszi azt. Az algoritmus során eltávolítjuk minden $[u_i, u_{i+1}]$ és $[u_i, u_{i+2}]$ intervallum p -értékét, így M és H gyorsabban lefut. H algoritmus:

1. lépés: Megnézzük \mathcal{U} -nak eddig mely $[u_i, u_{i+1}]$ intervallumjának p -értéke ismert, amelyiknek nem ismert, annak kiszámoljuk.

2. lépés: HA $\forall i [u_i, u_{i+1}] p$ -értéke $\geq \alpha$, AKKOR VÉGE

EGYÉBKÉNT: $k^* := \operatorname{argmin}(k \mapsto [u_k, u_{k+1}]\text{-nak a } p\text{-értéke})$

$\mathcal{U} := \mathcal{U} \cup \{[u_{k^*}, u_{k^*+1}] u^* \text{ indexe}\}$

VISSZA az 1. lépéshez

Minden új index után \mathcal{U} -nak legfeljebb 2 új $[u_i, u_{i+1}]$ szakasza lesz csak, így H -nak esetleg az első iteráció kivételével minden 1. lépésben legfeljebb 2 szimulációt kell végeznie. M algoritmus:

1. lépés: Megnézzük \mathcal{U} -nak eddig mely $[u_i, u_{i+2}]$ intervallumjának p -értéke ismert, amelyiknek nem ismert, annak kiszámoljuk.

2. lépés: HA $\forall i [u_i, u_{i+2}] p$ -értéke $\leq \alpha$, AKKOR VÉGE

EGYÉBKÉNT: $k^* := \operatorname{argmax}(k \mapsto [u_{k-1}, u_{k+1}]\text{-nek a } p\text{-értéke})$

$\mathcal{U} := \mathcal{U} \setminus \{u_{k^*}\}$

VISSZA az 1. lépéshez

Minden elhagyott index után \mathcal{U} -nak legfeljebb 2 új $[u_i, u_{i+2}]$ szakasza lesz csak, így esetleg az első iteráció kivételével M -nek minden 1. lépésben legfeljebb 2 szimulációt kell végeznie. A fő algoritmus:

1. lépés: $\mathcal{U}_1 := \mathcal{U}_2 := \{1, N\}$

2. lépés: $\mathcal{U}_1 := \mathcal{U}_2, \mathcal{U}_2 := M(H(\mathcal{U}_2))$

3. lépés: HA $\mathcal{U}_1 = \mathcal{U}_2$, AKKOR VÉGE

EGYÉBKÉNT: VISSZA a 2. lépésre

Ha leáll az algoritmus, akkor biztosan minimális elfogadható felosztást ad. Az eljárás a bináris CPD (Change Point Detection) algoritmusok családjába sorolható, melyek lehetséges javításairól [1] is ír. Egy lehetséges változtatás az lehet, hogy a szétválasztási próbáknál figyelembe vesszük \mathcal{U} elemszámát és

nagyobb $|\mathcal{U}|$ értékek esetén csökkentjük a próba szintjét. Ezzel kevésbé sűrű felosztásokat fogunk kapni. Egy másik lehetséges finomítás l^* helyett valami más homogenitást jellemző $\lambda(I)$ mérőszám bevezetése, oly módon I_1 és I_2 intervallumok esetén $\lambda(I_1) + \lambda(I_2) = \lambda(I_1 \cup I_2)$ teljesüljön. Ez az additivitás lényegesen gyorsabbá tenné az algoritmust, ugyanakkor az egyes próbák gyengébbek lennének.

1.5. Az eljárás eredménye

Az eredményt befolyásolja a blokkok (konstans) B elemszáma, a minimális megengedett M intervallumhossz és a próba α szintje. Továbbá azt is meg kell adnunk, hogy hány szimulációt végzünk, ez a szám a továbbiakban $n = 1000$. A kapott eredményt az alábbi táblázat foglalja össze, ahol $B = 20$ és $M = 24$. Vagyis havi maximum mellett a megengedett minimális intervallumhossz 2 év volt. A választás oka az, hogy $B = 10$ esetén nem adódtak reális eredmények.

Adatsor	α	kapott intervallumok száma
DJIA	0,01	2
DJIA	0,05	3
DJUA	0,01	1
DJUA	0,05	1

Az $\alpha = 0,01$ szint mellett a DJIA adatsor egyedüli osztópontjának 1986 adódik, az 1997-től vizsgált DJUA adatsort pedig egyáltalán nem osztja fel az algoritmus. Tehát ilyen szignifikancia mellett 1997-től napjainkig azonos eloszlásúnak tekinthető a DJIA és DJUA is. Ezen az időintervallumon a ML becslés eredményét az alábbi táblázat foglalja össze:

Adatsor	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	5% VaR	1 % VaR
DJIA	0,0133	0,0076	0,2430	0,1057	0,1692
DJUA	0,0146	0,0074	0,2522	0,1020	0,1607

A VaR paramétereit évből értendők. Szóval például az 1% VaR az veszteség, amit átlagban minden 100. évben haladunk csak meg.

Ellenőrzésképpen vizsgáljuk azt a nullhipotézist, hogy 1997-től napjainkig a két adatsor *GEV* eloszlású a kapott paraméterekkel. Anderson–Darling próbával rendre $p = 0,685$ és $p = 0,735$ adódik. Tehát elfogadható mindkét nullhipotézis.

2. Többdimenziós elemzés

Ezek után rátérhetünk a DJUA és DJIA adatsorok együttes elemzésére. Láttuk az előző fejezetben, hogy 1997-től napjainkig homogénnek tekinthető mindkét adatsor. Transzformáljuk az adatsorokat *standard Frechet* eloszlásúvá. Ezután együttes eloszlásukat próbáljuk meg a többek között [2] által is tárgyalt *Gumbel* (logisztikus) eloszlással modellezni. Azaz $G(x, y) = \exp\left(-\left(x^{-1/\alpha} + y^{-1/\alpha}\right)^\alpha\right)$ alakban keressük az együttes eloszlásfüggvényt. Itt α az összefüggés mértékét kvantifikálja: $\alpha = 0$ a teljes összefüggés, $\alpha = 1$ a függetlenség.

2.1. Válság detektálása

Próbáljunk meg gazdasági válságokat detektálni a két tőzsdeindex együttes viselkedésének vizsgálatával. Erre a feladatra az előző algoritmus még nem bizonyult alkalmasnak, további finomítások szükségesek. Az algoritmus sikeretelenségének oka az lehet, hogy válságok detektálásánál természetesebb lehet a három részre osztás ellenhipotézise. Legyen adott a standard *Frechet* marginálisú Z_1, Z_2, \dots, Z_N kétdimenziós sorozat. Hasonlóan az eddigiekhez legyen $[s, t]$ egy adott szakasz. A nullhipotézis legyen az, hogy azonos *Gumbel* eloszlású a szakasz, azaz

$$Z_s, Z_{s+1}, \dots, Z_t \sim \text{Gumbel}(\alpha)$$

valamilyen $\alpha \in [0, 1]$ paraméterrel. Az ellenhipotézis pedig legyen az, hogy három részre bontható a szakasz a következőképpen: $\exists u^* \in [s+M, t-M] \cap \mathbb{N}$, melyre

$$Z_s, Z_{s+1}, \dots, Z_{u^*}, Z_{u^*+K}, Z_{u^*+K+1}, \dots, Z_t \sim \text{Gumbel}(\alpha')$$

$$Z_{u^*+1}, Z_{u^*+2}, \dots, Z_{u^*+K-1} \sim \text{Gumbel}(\alpha'')$$

valamilyen $\alpha', \alpha'' \in [0, 1]$, $\alpha' \neq \alpha''$ paraméterekre. Itt K egy előre meghatározott szám, mely a detektálandó válság hosszának felel meg. Az előző pontban látott gondolatmenettel analóg módon képezhetjük az l^* maximális loglikelihood különbséget, mely az u^* és $u^* + K$ indexek mentén történő harmadolással adódik. Ha kellően nagy l^* , akkor elvetjük H_0 -t. A próba p -értékét szimuláció segítségével határozzuk meg. Legyen $\hat{\alpha}$ a H_0 -beli közös paraméterre kapott ML becslés. Generáljunk $N(\hat{\alpha}, I^{-1}(\hat{\alpha}))$ eloszlással n -szer véletlen α paramétereket és ezen paraméterekkel generáljunk $t - s + 1$ hosszú adatsorokat. Mind az n adatsorra számoljuk ki az l^* -statisztikát. Ekkor $\hat{p} = \frac{1}{n} \sum_{i=1}^n I(l^* < l_i^*)$.

2.2. Az eljárás eredménye

Legyen most $K = 24$. Továbbra is blokkmaximumokkal dolgozunk, tehát ez azt jelenti, hogy a detektálandó válság hossza 2 év. Ekkor legjobb harmadolásnak $u^* = 125$ adódik, ami közelítőleg 2007 szeptemberét jelöli ki a válság kezdetének. Az l^* -próbánál pedig $p = 0,001$ adódik. Ez alapján határozottan elutasíthatjuk a nullhipotézist. Ezen harmadolás mellett becsült paraméterek: $\hat{\alpha}_1 = 0,6915$ és $\hat{\alpha}_2 = 0,3599$. Tehát valóban erősebb összefüggés adódik válság idején.

Hivatkozások

- [1] Gerrit J. J. van den Burg, Christopher K. I. Williams. *An Evaluation of Change Point Detection Algorithms*. arXiv:2003.06222, 2020.
- [2] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag London, 2001.