Repni indeks i zavisnost

Stipetić, Ognjen

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Ognjen Stipetić

REPNI INDEKS I ZAVISNOST

Diplomski rad

Voditelj rada:
prof. dr. sc. Bojan Basrak

Zagreb, rujan, 2016.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

U ovom radu proučavamo distribucije teškog repa, odnosno distribucije za koje je vjerojatnost postizanja jako velikih vrijednosti relativno velika. Takve se distribucije koriste primjerice za modeliranje odštete u osiguranju, log-povrata financijskih instrumenata, maksimalnih dnevnih količina padalina, količine nafte u naftnim poljima, površine izgorjele u šumskim požarima i veličine datoteka poslanih internetom. Također, u teoriji vjerojatnosti, takve se distribucije pojavljuju kao granične distribucije graničnih teorema za parcijalne maksimume.

Proučavamo dva statistička problema vezana za distribucije teškog repa. Jedan je procijeniti repni indeks koji mjeri koliko je rep težak, a drugi je izmjeriti repnu zavisnost slučajnog vektora čije su komponente teškog repa.

U poglavlju 1 imamo pregled definicija i pripremnih rezultata koji će nam biti potrebni u ostatku rada. Definiramo regularnu varijaciju i iskazujemo neke osnovne rezultate o njoj uključujući poznati Karamatin teorem. Zatim definiramo slabu i slabašnu konvergenciju slučajnih mjera te iskazujemo nekoliko teorema o njima koji će nam kasnije biti potrebni. Na kraju definiramo Poissonovu slučajnu mjeru.

U poglavlju 2 definiramo Hillov procjenitelj koji za slučajnu varijablu čija je funkcija doživljenja "slična" $x^{-\alpha}$ procjenjuje $1/\alpha$ i dokazujemo njegovu konzistentnost. Zatim proučavamo njegovo ponašanje u praksi na primjeru podataka iz Paretove, normalne i Cauchyjeve distribucije te stvarnih podataka o dnevnom trgovanom volumenu dionica tehnoloških kompanija.

U poglavlju 3 proučavamo zavisnost višedimenzionalnih podataka iz distribucija teškog repa. Definiramo kutnu mjeru i koeficijente kutne zavisnosti χ i $\bar{\chi}$ koji svi mjere zavisnost višedimenzionalnih podataka. Isprobali smo te mjere zavisnosti na generiranim podacima iz bivarijatne normalne distribucije i bivarijatne logističke distribucije ekstremnih vrijednosti te podacima pravim podacima o dnevnom volumenu dionica.

Za razumijevanje ovog rada potrebno je predznanje iz teorije mjere i vjerojatnosti.

Poglavlje 1

Osnovni pojmovi i rezultati

1.1 Regularna varijacija

U ovom potpoglavlju navodimo neke rezultate iz analize regularno varirajućih funkcija, odnosno onih koje se ponašaju "slično" kao x^ρ. Takve su nam funkcije zanimljive jer se tako ponašaju funkcije doživljenja nekih slučajnih varijabli. Započnimo s definicijom

Definicija 1.1. *Kažemo da je izmjeriva funkcija* $U : \mathbb{R}_+ \mapsto \mathbb{R}_+$ regularno varirajuća *u* ∞ *s indeksom* ρ *(i pišemo* $U \in RV_\rho$*) ako za svaki* $x > 0$

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\rho.$$

Broj ρ *se nekad zove i* eksponent varijacije.

Ako je $\rho = 0$, kažemo da je U *sporo varirajuća*. Sporo varirajuće funkcije obično označavamo s $L(x)$. Za regularno varirajuću funkciju $U \in RV_\rho$ lako provjerimo da je $U(x)/x^\rho$ sporo varirajuća:

$$\lim_{t\to\infty} \frac{U(tx)/(tx)^\rho}{U(t)/t^\rho} = \lim_{t\to\infty} \frac{U(tx)}{x^\rho U(t)} = \frac{x^\rho}{x^\rho} = 1.$$

Dakle svaku regularno varirajuću funkciju možemo zapisati kao $L(x)x^\rho$ gdje je L sporo varirajuća.

Najjednostavniji primjer regularno varirajuće funkcije je $U(x) = x^\rho$. Sve konvergentne funkcije čiji je limes konačan i strogo pozitivan su sporo varirajuće, ali postoje i funkcije koje ne konvergiraju, ali jesu sporo varirajuće - primjerice $\log(1 + x)$. Navodimo karakterizaciju regularne varijacije koja izgleda puno slabije od definicije, ali se ispostavlja da je ekvivalentna.

Propozicija 1.2. *Izmjeriva funkcija $U : \mathbb{R}_+ \to \mathbb{R}_+$ je regularno varirajuća ako i samo ako postoji funkcija h takva da za svaki $x > 0$.*

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = h(x)$$

U tom slučaju je $h(x) = x^\rho$ za neki $\rho \in \mathbb{R}$.

Navodimo još jedan veoma važni teorem o regularnoj varijaciji - Karamatin teorem, koji nam govori kako integrirati regularno varirajuće funkcije.

Teorem 1.3. *(Karamatin teorem)*

1. *Neka je $\rho > -1$ i $U \in RV_\rho$. Tada je $\int_0^x U(t)dt \in RV_{\rho+1}$ i*

$$\lim_{x\to\infty} \frac{xU(x)}{\int_0^x U(t)dt} = \rho + 1.$$

Ako je $\rho < -1$ (ili ako je $\rho = -1$ i $\int_x^\infty U(s)ds < \infty$), onda $U \in RV_\rho$ povlači da je $\int_x^\infty U(t)dt$ konačan, $\int_x^\infty U(t)dt \in RV_{\rho+1}$, i

$$\lim_{x\to\infty} \frac{xU(x)}{\int_x^\infty U(t)dt} = -\rho - 1.$$

2. *Ako za U vrijedi*

$$\lim_{x\to\infty} \frac{xU(x)}{\int_0^x U(t)dt} = \lambda \in (0, \infty)$$

Onda je $U \in RV_{\lambda-1}$. Ako je $\int_x^\infty U(t)dt < \infty$ i

$$\lim_{x\to\infty} \frac{xU(x)}{\int_x^\infty U(t)dt} = \lambda \in (0, \infty),$$

tada je $U \in RV_{-\lambda-1}$.

Ovo potpoglavlje završavamo iskazom jedne posljedice Karamatinog teorema koju ćemo kasnije koristiti.

Propozicija 1.4. *Neka je $U \in RV_\rho$, $\rho \in \mathbb{R}$ i $\varepsilon > 0$. Tada postoji t_0 takav da za svaki $x \geq 1$ i $t \geq t_0$*

$$(1 - \varepsilon)x^{\rho-\varepsilon} < \frac{U(tx)}{U(t)} < (1 + \varepsilon)x^{\rho+\varepsilon}.$$

1.2 Slaba i slabašna konvergencija

Neka je $\mathbb{S}$ potpun separabilan metrički prostor s metrikom d, $\mathcal{S}$ Borelova σ-algebra na $\mathbb{S}$ generirana otvorenim skupovima, te neka je $(\Omega, \mathcal{A}, \mathbb{P})$ vjerojatnosni prostor. Slučajni element X u $\mathbb{S}$ je izmjerivo preslikavanje iz $(\Omega, \mathcal{A})$ u $(\mathbb{S}, \mathcal{S})$. Neka je $(X_n)_{n\in\mathbb{N}}$ niz slučajnih elemenata u $\mathbb{S}$. Tada odgovarajući niz *vjerojatnosnih distribucija* definiramo kao

$$P_n = \mathbb{P} \circ X_n^{-1} = \mathbb{P}(X_n \in \cdot).$$

Za naše će potrebe $\mathbb{S}$ najčešće biti $\mathbb{R}$ (pa je X slučajna varijabla), $\mathbb{R}^d$ (pa je X slučajni vektor), $\mathbb{R}^\infty$ (pa je X slučajni niz) ili $M_+(\mathbb{E})$, skup Radonovih mjera koji ćemo kasnije definirati (pa je X slučajna mjera). U $\mathbb{R}$ i $\mathbb{R}^d$ koristimo euklidsku metriku, a u $\mathbb{R}^\infty$ ćemo koristiti metriku definiranu s

$$d(x, y) = \sum_{k=1}^{\infty} \min\{|x_k - y_k|, 1\}2^{-k}.$$

Definicija 1.5. *Uz oznake iz gornjeg odlomka, kažemo da X_n konvergira po distribuciji u X_0, odnosno da P_n slabo konvergira u P_0 (i pišemo $X_n \Rightarrow X_0$ i $P_n \Rightarrow P_0$) ako za svaku ograničenu i neprekidnu funkciju $f : \mathbb{S} \to \mathbb{R}$ vrijedi*

$$\mathbb{E}f(X_n) = \int_{\mathbb{S}} f(x)P_n(dx) \to \mathbb{E}f(X_0) = \int_{\mathbb{S}} f(x)P_n(dx). \tag{1.1}$$

U slučaju $\mathbb{S} = \mathbb{R}$, slaba konvergencija je ekvivalentna konvergenciji po distribuciji.

U nastavku će nam biti potrebna još jedna vrsta konvergencije. Definirat ćemo *slabašnu konvergenciju* (ne nužno vjerojatnosnih) mjera. Prvo definiramo *Radonovu mjeru*.

Definicija 1.6. *Neka je $\mathcal{E}$ Borelova σ-algebra na $\mathbb{E}$. Kažemo da je mjera μ na $\mathbb{E}$* Radonova *ako je $\mu(K) < \infty$ za svaki kompaktni skup $K \subset \mathbb{E}$. Definiramo $M_+(\mathbb{E})$ kao skup svih Radonovih mjera na $\mathbb{E}$.*

Pokazuje se, $M_+(\mathbb{E})$ možemo promatrati kao metrički prostor u takozvanoj *slabašnoj metrici*. Ovdje nećemo definirati samu metriku, nego samo konvergenciju u njoj.

Kada smo govorili o slaboj konvergenciji, konačnost vjerojatnosne mjere i ograničenost testne funkcije f nam je garantirala da su očekivanja (odnosno integrali) konačni. Budući da sada govorimo o mjerama koje nisu nužno konačne, moramo na neki način osigurati konačnost integrala. Budući da su naše mjere Radonove, dovoljno je tražiti da funkcije (osim što su neprekidne) iščezavaju na komplementima kompaktnih skupova. Definiramo

$$C_K^+(\mathbb{E}) = \{f : \mathbb{E} \mapsto \mathbb{R}_+ : f \text{ je neprekidna s kompaktnim nosačem}\}.$$

Uočimo da uvjet da je f neprekidna s kompaktnim nosačem povlači ograničenost

Definicija 1.7. *Kažemo da niz mjera μ_n u $M_+(\mathbb{E})$* slabašno konvergira *(vague convergence) u μ_0 i pišemo $\mu_n \xrightarrow{v} \mu_0$ ako za svaku $f \in C_K^+(\mathbb{E})$ vrijedi*

$$\int_{\mathbb{E}} f(x)\mu_n(dx) \to \int_{\mathbb{E}} f(x)\mu_0(dx),$$

kad $n \to \infty$. Integrale na lijevoj i desnoj strani skraćeno označavamo s $\mu_n(f)$ odnosno $\mu_0(f)$.

Često će se pojavljivati mjere koncentrirane u jednoj ili nekoliko točaka. Mjeru koncentriranu u točki x ćemo označavati s ϵ_x, odnosno za svaki $x \in \mathbb{E}$ i $A \in \mathcal{E}$ definiramo

$$\epsilon_x(A) = \begin{cases} 1 & x \in A \\ 0 & x \notin A. \end{cases}$$

Definicija 1.8. *Kažemo da je mjera m na $M_+(\mathbb{E})$* točkovna mjera *ako je oblika*

$$m = \sum_i \epsilon_{x_i},$$

gdje i ide po konačnom ili prebrojivom skupu.

U nastavku navodimo bez dokaza nekoliko dovoljnih uvjeta za slabu konvergenciju mjera.

Lema 1.9. *Niz slučajnih mjera $(\mu_n)_{n\in n}$ u $M_+(\mathbb{E})$ slabo konvergira k mjeri μ ako i samo ako za svaku familiju $(h_j)_{j\in\mathbb{N}} \subset C_K^+(\mathbb{E})$*

$$(\mu_n(h_j), j \geq 1) \Rightarrow (\mu(h_j), j \geq 1)$$

u $\mathbb{R}^\infty$.

Time smo konvergenciju u vrlo apstraktnom prostoru $M_+(\mathbb{E})$ sveli na konvergenciju u nešto razumljivijem prostoru $\mathbb{R}^\infty$. Sljedeća lema pokazuje da je za konvergenciju u $\mathbb{R}^\infty$ dovoljno provjeriti konvergenciju u još jednostavnijim prostorima $\mathbb{R}^d$.

Lema 1.10. *Niz slučajnih nizova $\mathbf{X_n} = (X_n^{(1)}, X_n^{(2)}, ...)$ slabo konvergira u niz $\mathbf{X}$ u $\mathbb{R}^\infty$ ako za svaki $d \geq 1$*

$$(X_n^{(1)}, X_n^{(2)}, ..., X_n^{(d)}) \Rightarrow (X^{(1)}, X^{(2)}, ..., X^{(d)})$$

u $\mathbb{R}^d$.

Konvergenciju u $\mathbb{R}^d$ možemo još pojednostavniti. Sljedeća lema kaže da niz nenegativnih slučajnih vektora u $\mathbb{R}_+^d$ konvergira čim njegove Laplaceove transformacije konvergiraju.

Lema 1.11. *Neka je* $\mathbf{X}_n \geq \mathbf{0}$ *niz slučajnih vektora u* $\mathbb{R}^d$ *takav da za svaki vektor* $\lambda > \mathbf{0}$ *iz* $\mathbb{R}^d$ *vrijedi*

$$\mathbb{E}e^{-\lambda\cdot\mathbf{X}_n} \to \mathbb{E}e^{-\lambda\cdot\mathbf{X}}$$

Tada $\mathbf{X}_n$ *slabo konvergira u* $\mathbf{X}$*, odnosno* $\mathbf{X}_n \Rightarrow \mathbf{X}$*.*

Navodimo još jedan bitan teorem o slabašnoj konvergenciji vjerojatnosnih distribucija slučajnih varijabli čije komplementarne funkcije distribucije regularno variraju, za dokaz vidi [7], Theorem 3.6.

Teorem 1.12. *Neka je* X_1 *nenegativna slučajna varijabla s funkcijom distribucije* $F(x)$ *te neka je* $\overline{F} = 1 - F$*. Tada su sljedeće tri tvrdnje ekvivalentne:*

(i) $\overline{F} \in RV_{-\alpha}, \quad \alpha > 0.$

(ii) Postoji niz b_n *takav da* $b_n \to \infty$ *i*

$$\lim_{n\to\infty} n\overline{F}(b_n x) = x^{-\alpha}, \quad x > 0.$$

(iii) Postoji niz b_n *takav da* $b_n \to \infty$ *takav da*

$$\mu_n(\cdot) := n\mathbb{P}\Big[\frac{X_1}{b_n} \in \cdot\Big] \xrightarrow{v} \nu_\alpha(\cdot)$$

u $M_+(0,\infty]$*, gdje je* $\nu_\alpha(x,\infty] = x^{-\alpha}$*.*

U nastavku navodimo neke teoreme o zajedničkoj konvergenciji, također bez dokaza (vidi [1] Theorem 2.7., Theorem 3.9.)

Propozicija 1.13. *Neka su* $\mathbb{E}$ *i* $\mathbb{E}'$ *dva potpuna separabilna metrička prostora, i neka su* $(\xi_n)_{n\geq 0}$ *i* $(\eta_n)_{n\geq 0}$ *nizovi slučajnih elemenata u* $\mathbb{E}$ *odnosno* $\mathbb{E}'$ *na istom vjerojatnosnom prostoru. Pretpostavimo* $\xi_n \Rightarrow \xi_0$ *u* $\mathbb{E}$ *i* $\eta_n \xrightarrow{P} e_0'$*, gdje je* e_0' *neslučajni element* $\mathbb{E}'$*. Tada*

$$(\xi_n, \eta_n) \Rightarrow (\xi_0, e_0')$$

u $\mathbb{E}\times\mathbb{E}'$ *kada* $n \to \infty$*.*

Teorem 1.14. *Neka su* $(\mathbb{S}_1, d_1)$ *i* $(\mathbb{S}_2, d_2)$ *dva metrička prostora, i neka je* $(X_n)_{n\geq 0}$ *niz slučajnih elemenata u* $\mathbb{S}_1$ *za koji vrijedi* $X_n \Rightarrow X_0$*. Ako za funkciju* $h : \mathbb{S}_1 \mapsto \mathbb{S}_2$ *vrijedi*

$$\mathbb{P}(X_0 \in D(h)) = \mathbb{P}(X_0 \in \{s_1 \in \mathbb{S}_1 : h \text{ ima prekid u } s_1\}) = 0,$$

onda

$$h(X_n) \Rightarrow h(X_0)$$

u $\mathbb{S}_2$*.*

Navodimo još jedan bitan teorem o slaboj konvergenciji, korišten u nastavku (za dokaz vidi [7] Theorem 3.5.).

Teorem 1.15. *Neka su* $(X_{Mn})_{M\in\mathbb{N},n\in\mathbb{N}}, (X_M)_{M\in\mathbb{N}}, (Y_n)_{n\in\mathbb{N}}$ *i* X *slučajni elementi metričkog prostora* $(\mathbb{S}, \mathcal{S})$ *na istom vjerojatnosnom prostoru. Pretpostavimo da za svaki* M

$$X_{Mn} \Rightarrow X_M$$

kad $n \to \infty$ *i*

$$X_M \Rightarrow X$$

kad $M \to \infty$. *Također, pretpostavimo da za svaki* $\varepsilon > 0$,

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}(d(X_{Mn}, Y_n) > \varepsilon) = 0.$$

Tada, vrijedi

$$Y_n \Rightarrow X$$

kad $n \to \infty$.

1.3 Poissonov proces kao slučajna mjera

U ovom ćemo potpoglavlju definirati Poissonovu slučajnu mjeru. Za to nam je prvo potrebna definicija slučajnog procesa.

Neka je $\mathbb{E}$ metrički prostor s Borelovom σ-algebrom $\mathcal{E}$ i neka je $(X_n, n \in \mathbb{N})$ slučajni niz elemenata u $\mathbb{E}$, a ϵ_{X_n} niz njima pridruženih (slučajnih) točkovnih mjera. Definiramo slučajnu brojeću mjeru N sa

$$N(\cdot) = \sum_{n\in\mathbb{N}} \epsilon_{X_n}(\cdot).$$

Kažemo da je tako definirana mjera *N točkovni proces*. Sada možemo definirati Poissonovu slučajnu mjeru. [6]

Definicija 1.16. *Neka je N točkovni proces na skupu stanja* $\mathbb{E}$. *Kažemo da je N Poissonov proces sa srednjom mjerom* μ, *ili Poissonova slučajna mjera (Poisson random measure, PRM(*μ*)) ako vrijedi:*

(i) Za svaki $A \in \mathcal{E}$

$$P(N(A) = k) = \begin{cases} \frac{e^{-\mu(A)}(\mu(A))^k}{k!} & \mu(A) < \infty \\ 0 & \mu(A) = \infty. \end{cases}$$

(ii) Ako su $A_1, ..., A_k \in \mathcal{E}$ *disjunktni, onda su* $N(A_1), ..., N(A_k)$ *nezavisne slučajne varijable.*

Svojstvo (ii) u slučaju $\mathbb{E} = \mathbb{R}$ zovemo svojstvo nezavisnih prirasta jer povlači da su za $t_1 < t_2 < ... < t_k$ slučajne varijable $(N((t_i, t_{i+1}]), i = 1, ..., k-1)$ nezavisne.

Poglavlje 2

Hillov procjenitelj

U ovom poglavlju definiramo Hillov procjenitelj.

2.1 Motivacija

Promatramo niz jednako distribuiranih slučajnih varijabli $X, X_1, X_2, ...$ s funkcijom distribucije $F(x)$, odnosno funkcijom doživljenja $\overline{F}(x) = 1 - F(x)$, za koji mislimo da je teškog repa. U literaturi se težak rep definira na različite načine, među ostalim pojavljuju se sljedeće dvije pretpostavke:

- Možemo pretpostaviti da distribucija ima egzaktan Paretov (desni) rep iza nekog x_l, odnosno da vrijedi

$$\mathbb{P}(X > x) = cx^{-\alpha}, \quad x > x_l \tag{2.1}$$

- Alternativno, možemo uvesti samo poluparametarsku pretpostavku da F ima regularno variraući desni rep, to jest

$$\mathbb{P}(X > x) = 1 - F(x) = \overline{F}(x) = L(x)x^{-\alpha} \tag{2.2}$$

  gdje je L sporo varirajuća u $+\infty$.

U ovom ćemo radu uglavnom raditi uz slabiju, poluparametarsku, pretpostavku (2.2). Označimo sa $X_{(i)}, i = 1, ..., n$ i-tu najveću observaciju u nizu $X_1, ..., X_n$. Pretpostavimo još, radi jednostavnosti, da su observacije X_i (ili barem njih nekoliko najvećih) pozitivne. Tada možemo definirati Hillov procjenitelj.

Definicija 2.1. Hillov procjenitelj *za* α^{-1} *baziran na k gornjih uređajnih statistika definiramo kao*

$$H_{n,k} = \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{X_{(i)}}{X_{(k+1)}}\right). \tag{2.3}$$

Uočimo da se ovdje, osim broja observacija n pojavljuje i parametar k kojeg treba nekako izabrati. U narednim ćemo poglavljima komentirati metode izbora k. Prokomentirajmo zašto tako definiran procjenitelj ima smisla. Zamislimo privremeno da su podaci nezavisni i dolaze iz Paretove distribucije, to jest $\overline{F}(x) = x^{-\alpha}$, s nosačem $[1, +\infty\rangle$. Tada vrijedi

$$\mathbb{P}(\log X_i > x) = \mathbb{P}(X_i > e^x) = e^{-\alpha x}.$$

Dakle, $\log X_i$ dolazi iz ekponencijalne distribucije s parametrom α, pa je $\widehat{\alpha^{-1}} = \sum_{i=1}^{n} \log(X_i)$ procjenitelj maksimalne vjerodostojnosti za α^{-1}. Zamislimo sada da vrijedi samo pretpostavka (2.1), da je rep distribucije Paretov samo iza nekog x_l. Tada, zbog svojstva zaboravljivosti eksponencijalne slučajne varijable, za svaki $u > x_l$ i $x > 0$ vrijedi

$$\mathbb{P}(\log \frac{X_i}{u} > x | X_i > u) = \frac{\mathbb{P}(\log X_i > x + \log u)}{\mathbb{P}(\log X_i > \log u)} = \frac{e^{-\alpha(\log u + x)}}{e^{-\alpha \log u}} = e^{-\alpha x}.$$

Dakle, slučajna varijabla $\log \frac{X_i}{u} | X_i > u$ je eksponencijalno distribuirana (s parametrom α) za dovoljno veliku konstantu u, što sugerira da ima smisla promatrati samo nekoliko najvećih observacija i normirati konstantom dovoljno velikom da ostale observacije budu manje od nje. Također, vidimo da parametar eksponencijalne distribucije α ne ovisi o izboru u, nego samo o distribuciji slučajnih varijabli. Bilo bi poželjno da je u što manji tako da imamo više observacija $X_i > u$, pa da α procjenjujemo na temelju što više podataka, ali istovremeno moramo paziti da u nije premali jer smo pretpostavili da je veći od x_l.

Jasno, ovdje nismo argumentirali zašto u definiciji Hillovog procjenitelja smijemo konstantu u zamijeniti slučajnom varijablom $X_{(k+1)}$, iako vidimo da je $X_{(k+1)}$ odličan izbor za u jer će, uz dobre asimptotske uvjete ($k \to \infty$, $k/n \to 0$), $X_{(k+1)}$ gotovo sigurno prerasti x_l. Teoreme koji nam garantiraju dobro ponašanje Hillovog procjenitelja i u poluparametarskom slučaju (2.2) dokazujemo u sljedećim točkama. Započnimo s nekoliko definicija koje će nam biti potrebne.

Definicija 2.2. *Neka su dane observacije* $x_1, ..., x_n$ *i prag u. Kažemo da je observacija* x_j prekoračenje praga *ako je* $x_j > u$*, te u tom slučaju broj* $x_j - u$ *zovemo* višak.

Definicija 2.3. *Neka su* X_n *nezavisne jednakodistribuirane slučajne varijable i neka je u prag. Definiramo vremena prekoračenja* τ_j *sa*

$$\tau_1 = \inf\{j \geq 1 : X_j > u\}$$
$$\tau_r = \inf\{j > \tau_{r-1} : X_j > u\}, \quad r \geq 2.$$

Jasno je da su slučajne varijable X_{τ_j} prekoračenja.

2.2 Repna empirijska mjera

Teorem 1.12 sugerira da, umjesto procjenjivanja repnog indeksa α možemo procijeniti mjeru ν_α na $(0, \infty]$ koja bi dala istu informaciju. Započnimo s nekoliko definicija.

Definicija 2.4. *Neka je $F : \mathbb{R} \mapsto [0, 1]$ nepadajuća zdesna neprekidna funkcija. Definiramo njen* generalizirani inverz $F^{-1} : [0, 1] \mapsto \mathbb{R} \cup \{\infty, -\infty\}$ *sa*

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}.$$

Uočimo da ovako definirana funkcija zaista je generalizacija inverza - ako je F bijekcija, onda je ovako definirana F^{-1} jednaka inverzu od F, ali ova defincija ima smisla i za funkcije koje nisu injektivne ili surjektivne. Ako je F funkcija distribucije neke slučajne varijable X, onda funkcija $F^{-1}(p)$ vraća najmanji broj x takav da je $\mathbb{P}(X \leq x)$ barem p.

Definicija 2.5. *Neka je X_j niz jednakodistribuiranih slučajnih varijabli sa zajedničkom funkcijom distribucije F i regularno varirajućom funkcijom doživljenja*

$$\overline{F}(x) = \mathbb{P}(X > x) = 1 - F(x) = L(x)x^{-\alpha} \quad \alpha > 0.$$

Zbog jednostavnosti, pretpostavimo da su slučajne varijable nenegativne. Definiramo funkciju kvantila $b(t)$ *kao*

$$b(t) = \left(\frac{1}{1 - F}\right)^{-1}(t) = F^{-1}(1 - \frac{1}{t}).$$

Grubo govoreći, funkcija kvantila b za broj $t > 1$, daje broj $b(t)$ takav da je vjerojatnost da je X_j veća od $b(t)$ jednaka $1/t$. Očito, poznavanje funkcije kvantila b je ekvivalentno poznavanju funkcije distribucije F.

Definicija 2.6. Repnu empirijsku mjeru ν_n *definiramo kao slučajni element $M_+(0, \infty]$, prostora nenegativnih Radonovih mjera na* $(0, \infty]$ *sa*

$$\nu_n := \frac{1}{k} \sum_{i=1}^{n} \epsilon_{X_i / b(\frac{n}{k})}.$$

Uočimo da ova definicija ovisi i o parametru k, odnosno o broju najvećih opažanja koje smatramo relevantnima za procjenu α. Također, u primjenama nećemo znati niti pravu funkciju distribucije F, odnosno funkciju kvantila b, dakle trebamo i nju procijeniti. Međutim, sljedeći teoremi pokazuju da ti problemi i nisu toliko veliki, odnosno da repna empirijska mjera ν_n asimptotski dobro aproksimira pravu repnu mjeru ν_α. Prvo ćemo dokazati slabiji teorem, u kojem pretpostavljamo da znamo teorijsku funkciju kvantila b.

Teorem 2.7. *Neka je $(X_j)_{j\in\mathbb{N}}$ niz nezavisnih jednakodistribuiranih nenegativnih slučajnih varijabli čija je funkcija doživljenja regularno varirajuća, što povlači (po teoremu 1.12) da*

$$\frac{n}{k}\mathbb{P}\Big(\frac{X_1}{b(n/k)} \in \cdot\Big) \xrightarrow{v} \nu_\alpha(\cdot)$$

u $M_+(0,\infty]$ kad $n \to \infty$, $k = k(n) \to \infty$ i $n/k \to \infty$. Tada u $M_+(0,\infty]$

$$\nu_n \Rightarrow \nu_\alpha. \tag{2.4}$$

Dokaz. Po lemi 1.9, da bismo dokazali konvergenciju u $M_+(0,\infty]$, dovoljno je dokazati da za svaki niz funkcija h_j u $C_K^+(0,\infty]$

$$(\nu_n(h_j), j \geq 1) \Rightarrow (\nu_\alpha(h_j), j \geq 1)$$

u $\mathbb{R}^\infty$, što po lemi 1.10 vrijedi ako i samo ako za svaki $d \in \mathbb{N}$

$$(\nu_n(h_j), 1 \leq j \leq d) \Rightarrow (\nu_\alpha(h_j), 1 \leq j \leq d)$$

u $\mathbb{R}^d$. Po lemi 1.11, da bismo to dokazali, dovoljno je dokazati da za svaki vektor $\lambda = (\lambda_1 ... \lambda_n) > \mathbf{0}$ u $\mathbb{R}^d$ vrijedi

$$\mathbb{E}e^{-\sum_{j=1}^d \lambda_j \nu_n(h_j)} \to \mathbb{E}e^{-\sum_{j=1}^d \lambda_j \nu_\alpha(h_j)}. \tag{2.5}$$

Iz linearnosti Lebesguevog integrala slijedi

$$\sum_{j=1}^d \lambda_j \nu_n(h_j) = \nu_n(\sum_{j=1}^d \lambda_j h_j),$$

te ista tvrdnja za ν_α. Definiramo $h = \sum_{j=1}^d \lambda_j h_j$. Sada (2.5) postaje

$$\mathbb{E}e^{-\nu_n(h)} \to \mathbb{E}e^{-\nu_\alpha(h)} \tag{2.6}$$

Budući da je h linearna kombinacija neprekidnih funkcija s kompaktnim nosačem, i ona je neprekidna funkcija s kompaktnim nosačem, dakle ostaje dokazati (2.6) za $h \in C_K^+(0,\infty]$. Lijeva strana u izrazu (2.6) jednaka je

$$\begin{aligned}
\mathbb{E}e^{-\frac{1}{k}\sum_{j=1}^n h(X_j/b(n/k))} &= \text{(nezavisnost i jednaka distribuiranost)} \\
&= \big(\mathbb{E}e^{-\frac{1}{k}h(X_1/b(n/k))}\big)^n \\
&= \left(1 - \int_{(0,\infty]} (1 - e^{-\frac{1}{k}h(x)})\mathbb{P}\Big(\frac{X_1}{b(n/k)} \in dx\Big)\right)^n \\
&= \left(1 - \frac{\int_{(0,\infty]} (1 - e^{-\frac{1}{k}h(x)}) n\mathbb{P}\Big(\frac{X_1}{b(n/k)} \in dx\Big)}{n}\right)^n.
\end{aligned}$$

Dokazujemo da konvergira u $e^{-\nu_\alpha(h)}$. Za integral u zagradi vrijedi aproksimacija

$$\int_{(0,\infty]} (1 - e^{-\frac{1}{k}h(x)}) n\mathbb{P}\Big(\frac{X_1}{b(n/k)} \in dx\Big) \approx \int_{(0,\infty]} h(x)\frac{n}{k}\mathbb{P}\Big(\frac{X_1}{b(n/k)} \in dx\Big). \tag{2.7}$$

Detalje o toj približnoj jednakosti možemo naći u [7]. Po pretpostavci teorema,

$$\frac{n}{k}\mathbb{P}\Big(\frac{X_1}{b(n/k)} \in \cdot\Big) \xrightarrow{v} \nu_\alpha(\cdot),$$

odnosno, za neprekidnu funkciju h s kompaktnim nosačem,

$$\int_{(0,\infty]} h(x)\frac{n}{k}\mathbb{P}\Big(\frac{X_1}{b(n/k)} \in dx\Big) \to \int_{(0,\infty]} h(x)d\nu_\alpha(x) = \nu_\alpha(h).$$

Označimo li lijevu stranu sa a_n, a desnu sa a, ostaje nam dokazati da iz $a_n \to a$ slijedi

$$\lim\Big(1 - \frac{a_n}{n}\Big)^n = e^{-a}.$$

Za proizvoljni $\varepsilon > 0$ i svaki n, osim eventualno njih konačno mnogo, vrijedi $a_n \geq a - \varepsilon$, pa vrijedi i

$$\Big(1 - \frac{a_n}{n}\Big)^n \leq \Big(1 - \frac{a-\varepsilon}{n}\Big)^n, \quad n \geq n_0.$$

Uzimanjem limes superiora obje strane dobivamo

$$\limsup\Big(1 - \frac{a_n}{n}\Big)^n \leq \limsup\Big(1 - \frac{a-\varepsilon}{n}\Big)^n = e^{-(a-\varepsilon)}.$$

Te analogno dobijemo donju ogradu na limes inferior.

$$e^{-(a+\varepsilon)} \leq \liminf\Big(1 - \frac{a_n}{n}\Big)^n.$$

Zbog proizvoljnosti ε i neprekidnosti eksponencijalne funkcije vrijedi

$$e^{-a} \leq \liminf\Big(1 - \frac{a_n}{n}\Big)^n \leq \limsup\Big(1 - \frac{a_n}{n}\Big)^n \leq e^{-a}.$$

Dakle,

$$\lim\Big(1 - \frac{a_n}{n}\Big)^n = e^{-a}.$$

Što je i trebalo dokazati. □

2.3 Konzistentnost Hillovog procjenitelja

U prethodnom smo potpoglavlju dokazali da repna empirijska mjera slabo konvergira u mjeru ν_α iz koje vidimo parametar α koji nas zanima. Sada ćemo taj rezultat primijeniti kako bismo dokazali konzistentnost Hillovog procjenitelja.

U prethodnom smo poglavlju pretpostavljali da znamo $b(n/k)$, što u praksi nećemo znati jer ne znamo funkciju b, nego ćemo raditi s konzistentnim procjeniteljem $\hat{b}(n/k)$ za $b(n/k)$ definiranim kao

$$\hat{b}(n/k) = X_{(k)},$$

za koji se pokazuje da je konzistentan. Sada procjena za $\hat{\nu}_n$ iznosi

$$\hat{\nu}_n = \frac{1}{k}\sum_{i=1}^{n} \epsilon_{X_i/\hat{b}(n/k)}.$$

Dokazujemo glavni rezultat ovog poglavlja, da Hillov procjenitelj definiran relacijom (2.3) po vjerojatnosti konvergira u $1/\alpha$.

Teorem 2.8. *Ako vrijedi (2.4), tada*

$$H_{n,k} \xrightarrow{P} \frac{1}{\alpha}$$

kada $n \to \infty$, $k \to \infty$, $k/n \to 0$.

Dokaz. Prvo dokazujemo da (2.4) povlači

$$\frac{X_{(k)}}{b(n/k)} = \frac{\hat{b}(n/k)}{b(n/k)} \xrightarrow{P} 1$$

kada $n \to \infty$, $k \to \infty$, $k/n \to 0$.

$$\begin{aligned}\mathbb{P}\left(\left|\frac{X_{(k)}}{b(n/k)} - 1\right| > \epsilon\right) &= \mathbb{P}(X_{(k)} > (1+\varepsilon)b(n/k)) + \mathbb{P}(X_{(k)} < (1-\varepsilon)b(n/k)) \\ &= \mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{n} \epsilon_{X_i/b(n/k)}(1+\varepsilon,\infty] \geq 1\right) \\ &+ \mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{n} \epsilon_{X_i/b(n/k)}[1-\varepsilon,\infty] < 1\right)\end{aligned} \tag{2.8}$$

Vidimo da se u zagradama nalazi baš ν_n, koji po pretpostavci slabo konvergira u ν_α. Zbog toga $\nu_n(1+\varepsilon,\infty] \Rightarrow \nu_\alpha(1+\varepsilon,\infty] = (1+\varepsilon)^{-\alpha}$. Ovdje govorimo o slaboj konvergenciji

slučajnih varijabli za koju znamo da je ekvivalentna konvergenciji po distribuciji, a konvergencija po distribuciji u konstantu povlači konvergenciju po vjerojatnosti. Dakle,

$$\nu_n(1+\varepsilon,\infty] \xrightarrow{P} (1+\varepsilon)^{-\alpha} < 1$$

i

$$\nu_n[1-\varepsilon,\infty] \xrightarrow{P} (1-\varepsilon)^{-\alpha} > 1,$$

iz čega slijedi da oba člana u (2.8) konvergiraju u 0, pa $X_{(k)}/b(n/k) \xrightarrow{P} 1$.

Sada dokazujemo da u $M_+(0,\infty]$

$$\hat{\nu}_n \xrightarrow{P} \nu_\alpha$$

kada $n \to \infty$, $k \to \infty$, $k/n \to 0$. Da bismo to dokazali, definiramo operator

$$T : M_+((0,\infty]) \times (0,\infty) \mapsto M_+((0,\infty])$$

sa

$$T(\mu, x)(A) = \mu(xA).$$

Iz pretpostavke teorema, ranije dokazane slabe konvergencije i propozicije (1.13) slijedi

$$\left(\nu_n, \frac{X_{(k)}}{b(n/k)}\right) \Rightarrow (\nu_\alpha, 1)$$

u $M_+((0,\infty]) \times (0,\infty)$. Iz defincije $\hat{\nu}_n$,

$$\hat{\nu}_n(\cdot) = \nu_n\left(\frac{X_{(k)}}{b(n/k)} \cdot\right) = T\left(\nu_n, \frac{X_{(k)}}{b(n/k)}\right),$$

pa, po teoremu (1.14) (ako ga smijemo primijeniti), $\hat{\nu}_n \Rightarrow T(\nu_\alpha, 1) = \nu_\alpha$. Još treba dokazati da vrijede uvjeti teorema (1.14), odnosno da je operator T neprekidan u $(\nu_\alpha, 1)$. Dokazujemo neprekidnost u (ν_α, x) za proizvoljni $x > 0$. Neka su $\mu_n \in M_+(0,\infty]$ i x_n, $x \in (0,\infty)$ takvi da $\mu_n \xrightarrow{v} \nu_\alpha$ i $x_n \to x$. Dovoljno je dokazati da za svaku funkciju $f \in C_K^+(0,\infty]$ vrijedi

$$\int_{(0,\infty]} f(t)\mu_n(x_n dt) = \int_{(0,\infty]} f(y/x_n)\mu_n(dy) \to \int_{(0,\infty]} f(y/x)\nu_\alpha(dy).$$

Ograničimo,

$$\begin{aligned}\left|\int_{(0,\infty]} f(y/x_n)\mu_n(dy) - \int_{(0,\infty]} f(y/x)\nu_\alpha(dy)\right| \leq \\ \leq \left|\int_{(0,\infty]} f(y/x_n)\mu_n(dy) - \int_{(0,\infty]} f(y/x)\mu_n(dy)\right| + \\ + \left|\int_{(0,\infty]} f(y/x)\mu_n(dy) - \int_{(0,\infty]} f(y/x)\nu_\alpha(dy)\right|\end{aligned}$$

Ovdje druga apsolutna vrijednost konvergira u 0 po definiciji slabašne konvergencije (ako je $f(y)$ neprekidna s kompaktnim nosačem, tada je i $f(y/x)$ za fiksni x. Dokažimo još da prvi integral možemo učiniti proizvoljno malim. Uzmimo δ_0 takav da su nosači funkcija $f(\frac{\cdot}{x})$ i $f(\frac{\cdot}{x_n})$ sadržani u $[\delta_0, \infty]$. Budući da je f neprekidna s kompaktnim nosačem, ona je uniformno neprekidna na $(0, \infty])$ (u metrici $d(s,t) = |s^{-1} - t^{-1}|$). Tada za svaki $\delta > 0$ zbog $x_n \to x$ možemo odabrati dovoljno veliki n da

$$d(y/x_n, y/x) = y^{-1}|x_n - x| < \delta$$

za svaki $y > \delta_0$. Neka je $\varepsilon > 0$. Uzmemo δ iz definicije uniformne neprekidnosti funkcije f i dovoljno veliki n. Tada za $y > \delta_0$ vrijedi

$$|f(y/x_n) - f(y/x)| < \varepsilon.$$

Budući da je $\mu_n(\delta_0, \infty]$ ograničen, onda i

$$\int_{(0,\infty])} |f(y/x_n) - f(y/x)|\mu_n(dy) \to 0.$$

Iz čega slijedi da je T zaista neprekidno, a time i $\hat{\nu}_n \xrightarrow{P} \nu_\alpha$.

Sada Hillov procjenitelj želimo zapisati preko mjere $\hat{\nu}_n$. Računamo:

$$\begin{aligned}
H_{n,k} &= \frac{1}{k}\sum_{i=1}^{k} \log \frac{X_{(i)}}{X_{(k)}} \\
&= \frac{1}{k}\sum_{i=1}^{k} \int_1^{X_{(i)}/X_{(k)}} x^{-1}dx \\
&= \frac{1}{k}\sum_{i=1}^{n} \int_1^{\max\{X_{(i)}/\hat{b}(n/k),1\}} x^{-1}dx \\
&= \int_1^{\infty} \frac{1}{k}\sum_{i=1}^{n} \epsilon_{X_i/\hat{b}(n/k)}(x,\infty]x^{-1}dx \\
&= \int_1^{\infty} \hat{\nu}_n(x,\infty]x^{-1}dx.
\end{aligned}$$

Da bismo dokazali (slabu) konzistentnost Hillovog procjenitelja, još ostaje dokazati da

$$\int_1^{\infty} \hat{\nu}_n(x,\infty]x^{-1}dx \xrightarrow{P} \int_1^{\infty} \hat{\nu}_\alpha(x,\infty]x^{-1}dx = \alpha^{-1}.$$

To ćemo dokazati pomoću teorema 1.15. Stavimo

$$X_{Mn} = \int_1^{M} \hat{\nu}_n(x,\infty]x^{-1}dx,$$

$$X_M = \int_1^M \nu_\alpha(x,\infty]x^{-1}dx,$$
$$Y_n = \int_1^\infty \hat{\nu}_n(x,\infty]x^{-1}dx,$$
$$X = \int_1^\infty \nu_\alpha(x,\infty]x^{-1}dx,$$

Vidimo da nam zaključak teorema daje $Y_n \Rightarrow X$, upravo ono što nam je ostalo dokazati. Provjerimo sada vrijede li pretpostavke teorema. Vidimo da za svaki M, $X_{Mn} \Rightarrow X_M$ jer je integral po kompaktnom skupu neprekidan operator. Također, $X_M \Rightarrow X$ po Lebesguevom teoremu o monotonoj konvergenciji. Ostaje provjeriti treću pretpostavku. Neka je $\delta > 0$. Dokazujemo da je

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}\Big(\int_M^\infty \hat{\nu}_n(x,\infty]x^{-1}dx > \delta\Big) = 0.$$

Zapišimo to ovako:

$$\begin{aligned}
&\mathbb{P}\Big(\int_M^\infty \hat{\nu}_n(x,\infty]x^{-1}dx > \delta\Big) \\
&\quad = \mathbb{P}\Big(\int_M^\infty \hat{\nu}_n(x,\infty]x^{-1}dx > \delta, \frac{\hat{b}(n/k)}{b(n/k)} \in (1-\eta, 1+\eta)\Big) \\
&\quad + \mathbb{P}\Big(\int_M^\infty \hat{\nu}_n(x,\infty]x^{-1}dx > \delta, \frac{\hat{b}(n/k)}{b(n/k)} \notin (1-\eta, 1+\eta)\Big) \\
&\quad = I + II.
\end{aligned}$$

Odmah vidimo da je

$$II \le P\Big(\Big|\frac{\hat{b}(n/k)}{b(n/k)} - 1\Big| \ge \eta\Big) \to 0$$

Ograničimo još I, na sljedeći način

$$\begin{aligned}
I &\le \mathbb{P}\Big(\int_M^\infty \nu_n((1-\eta)x,\infty]x^{-1}dx > \delta\Big) \\
&= \mathbb{P}\Big(\int_{M(1-\eta)}^\infty \nu_n(x,\infty]x^{-1}dx > \delta\Big) \\
&\le \delta^{-1}\mathbb{E}\Big(\int_{M(1-\eta)}^\infty \nu_n(x,\infty]x^{-1}dx\Big) \\
&= \delta^{-1}\int_{M(1-\eta)}^\infty \frac{n}{k}\mathbb{P}(X_1 > b(n/k)x)x^{-1}dx.
\end{aligned} \tag{2.9}$$

Ovdje prva nejednakost slijedi iz monotonosti vjerojatnosti, drugi red je drugačije zapisan prvi, treći red slijedi iz Markovljeve nejednakosti, a u četvrtom redu je raspisana definicija ν_n. Iz propozicije 1.4 slijedi da je za proizvoljni ε i dovoljno velik n,

$$\begin{aligned}(1-\varepsilon)x^{-\alpha-\varepsilon} &= \frac{n}{k}\overline{F}(b(n/k))(1-\varepsilon)x^{-\alpha-\varepsilon} \\ &< \frac{n}{k}\overline{F}(b(n/k)x) \\ &< \frac{n}{k}\overline{F}(b(n/k))(1+\varepsilon)x^{-\alpha+\varepsilon} = (1+\varepsilon)x^{-\alpha+\varepsilon}.\end{aligned}$$

Uzimanjem $\varepsilon \to 0$, iz Lebesguevog teorema o monotonoj konvergenciji slijedi da kada $n \to \infty$, izraz (2.9) konvergira u

$$\delta^{-1}\int_{M(1-\eta)}^{\infty} x^{-\alpha-1}dx = \frac{1}{\alpha}(1-\eta)^{-\alpha}M^{-\alpha},$$

što konvergira u 0 kada $M \to \infty$.

Iz dokazanog slijedi da možemo primijeniti teorem 1.15, čime je dokaz konzistentnosti Hillovog procjenitelja završen. □

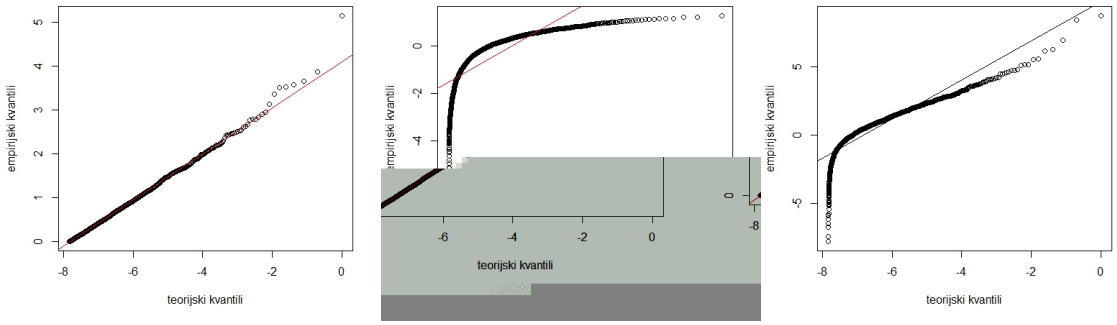
2.4 Hillov procjenitelj u praksi

Do sad smo promatrali asimptotska svojstva Hillovog procjenitelja, a sada ćemo promotriti kako se on koristi u praksi na konačnom uzorku. Do sad smo za $k = k(n)$ uzimali bilo koji niz takav da $k \to \infty$ i $k/n \to 0$, a u primjenama za dani uzorak moramo odabrati neki k koji ćemo koristiti za procjenu. Za to koristimo takozvani Hillov graf (Hill plot),

$$((k, H_{k,n}^{-1}), 1 \leq k \leq n),$$

i nadamo se da graf izgleda stabilno pa da po tome možemo odabrati α [4]. Očekivano, Hillov procjenitelj najbolje radi kada podaci dolaze baš iz Paretove distribucije, bez sporo varirajućeg dijela. Simulirali smo 2500 podataka iz Paretove distribucije s indeksom $\alpha = 2$, te jednako toliko podataka iz standardne normalne i standardne Cauchyjeve distribucije od kojih smo uzeli apsolutnu vrijednost. Prvo provjeravamo dolaze li podaci iz distribucije slične Paretovoj pomoću Q-Q grafa (vrijednosti na obje osi su logaritmirane).

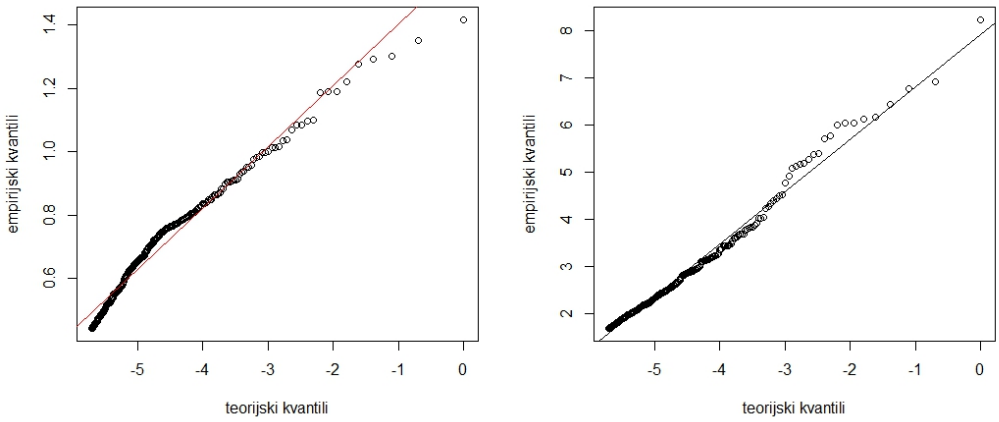
Iz slike 2.1 bismo mogli zaključuiti da podaci iz Paretove distribucije jesu teškog repa a oni iz normalne i Cauchyjeve nisu. Međutim, Cauchyjeva funkcija doživljenja je, za razliku od normalne, regularno varirajuća, pa bismo htjeli moći donijeti zaključak da je ona teškog repa. Budući da smo više zainteresirani za ekstremne događaje nego za observacije blizu nule, pogledajmo na slici 2.2 takav Q-Q graf za najvećih 300 observacija iz normalne i



(a) Paretova (b) Normalna (c) Cauchyjeva

Slika 2.1: Q-Q graf generiranih podataka u usporedbi s teorijskom Paretovom distribucijom

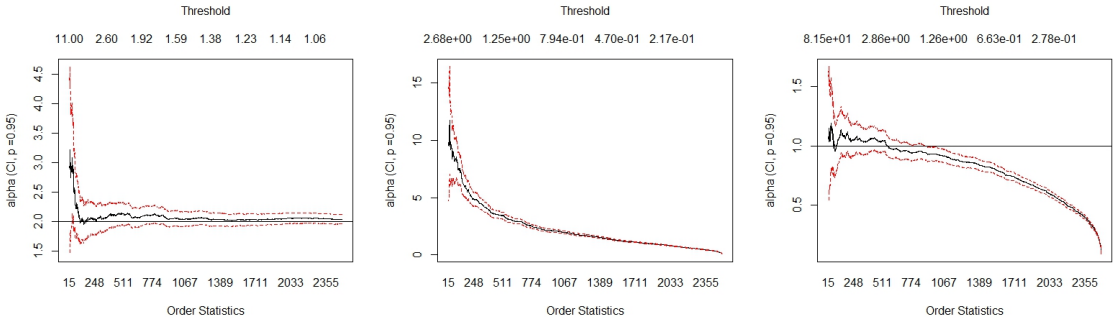
Cauchyjeve. Točke na grafu za normalnu čine opisuju neku krivulju koja nije sasvim na pravcu, pa možemo zaključiti da je pretpostavka regularne varijacije opravdanija u slučaju Cauchyjeve, ali nije jasno koliko bismo takvih odstupanja kod stvarnih podataka trebali tolerirati. Osim toga, nemamo neki dobar način izbora koliko ćemo najvećih observacija promatrati. Sve to dovodi do zaključka da je u praksi jako teško procijeniti kada uopće koristiti metode opisane u ovom radu.



(a) Normalna (b) Cauchyjeva

Slika 2.2: Q-Q graf za 300 najvećih podataka

Na slici 2.3 vidimo da se za podatke iz Paretove distribucije graf vrlo brzo stabilizira blizu točne vrijednosti 2, dok se za normalno distribuirane podatke (čija funkcija



(a) Paretova (b) Normalna (c) Cauchyjeva

Slika 2.3: Hillovi grafovi za dvije distribucije teškog i jednu laganog repa

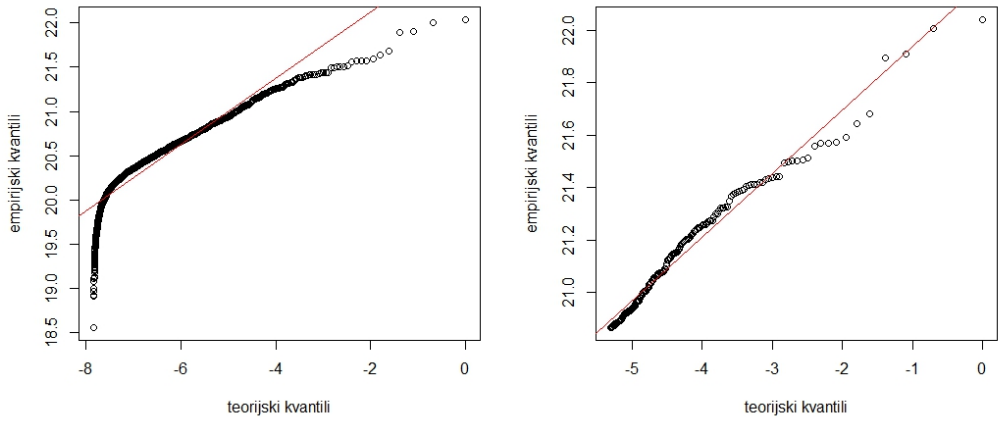
doživljenja nije regularno varirajuća) ne stabilizira niti oko jedne vrijednosti, što smo i očekivali jer repni indeks ne postoji za distribucije tako laganog repa. Hillov graf za Cauchyjevu distribuciju se za male k (ali dovoljno velike da varijanca padne) stabilizira oko točne vrijednosti 1, ali kada se odmaknemo od repa i Cauchyjeva se distribucija počinje sve više razlikovati od Paretove indeksa 1 konvergira u 0. Međutim, budući da nas zanima indeks u repu, možemo zaključiti da je Hillov graf prilično dobar u za Cauchyjevu distribuciju.

U praksi se distribucije regularno varirajućeg repa koriste za mnoge stvari uključujući modeliranje godišnjih maksimuma padalina, veličine preuzimanih datoteka na internetu, vrijednost kupljenih i prodanih financijskih objekata na burzi u jednom danu i druge podatke gdje se očekuje da će nekoliko ekstremnih točaka biti puno veće od većine ostalih. U ovom radu promatramo podatke o ukupnom dnevnom volumenu dionica kompanija Apple, Google, Microsoft i IBM u periodu od 1.1.2006. do 1.1.2016. Ukupno imamo 2517 podataka za svaku dionicu.

Kao i za generirane podatke, prvo pomoću Q-Q grafa provjeravamo jesu li podaci teškog repa. Iz grafa je očito da distribucija podataka za IBM nije Paretova, međutim, kada uzmemo samo nekoliko najvećih observacija (budući da nas zanima samo rep), točke na Q-Q grafu leže prilično blizu pravcu (iako opisuju konkavnu krivulju koja sugerira da je rep nešto lakši od repa Paretove), tako da u nastavku koristimo metode prilagođene teškim repovima iako pretpostavka regularne varijacije nije sasvim opravdana.

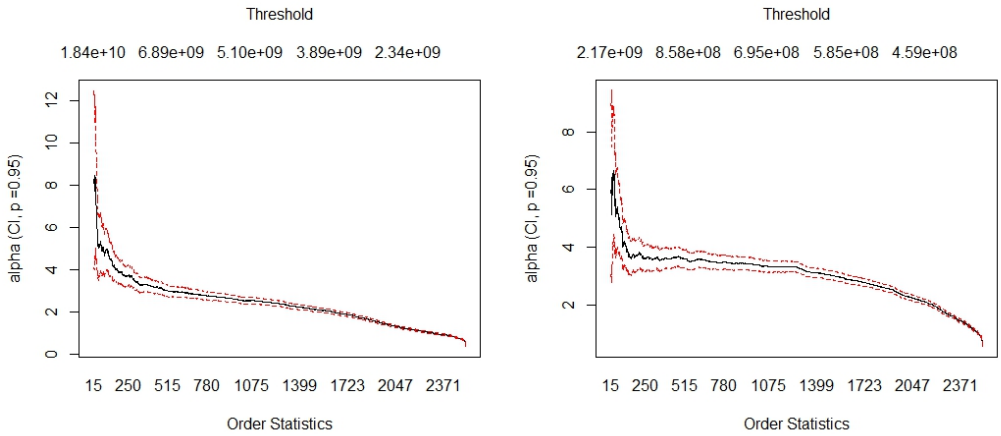
Promotrimo sada Hillov graf za jedno i drugo.

Vidimo da je Hillov graf za IBM prilično stabilan oko 3.5 u lijevoj polovici nakon kratkog dijela u visoke varijance za mali k (varijanca je visoka jer zaključujemo na temelju malog broja podataka). U desnoj polovici (za $k > n/2$) vrijednosti procjenitelja padaju u 0, ali to i nije jako bitno jer smo se odmakli daleko od repa. Dakle, zaključujemo da je repni



(a) Svih 2500 podataka (b) Samo najvećih 500 podataka

Slika 2.4: Q-Q graf za dnevni volumen dionice IBM-a



(a) AAPL (b) IBM

Slika 2.5: Hillovi grafovi za dnevni volumen dionica

indeks za IBM otprilike jednak 3.5. S druge strane, graf za Apple nije toliko informativan, nakon što se varijanca stabilizira, procjena pada od skoro 4 do ispod 2. Procjena repnog indeksa bi mogla ponovo biti oko 3, ali ne znamo koliko je ta procjena točna niti ima li uopće smisla to modelirati distribucijom sličnoj Paretovoj. Hillovi grafovi za druge dionice su rijetko toliko stabilni kao za IBM, ali su često bolji nego za Apple.

U literaturi [7] se predlažu još dvije varijante grafa Hillovog procjenitelja. Budući da nas više zanima lijevi dio grafa (i u dokazima za asimptotsko ponašanje procjenitelja pretpostavljamo $k/n \to 0$, odnosno k je puno manji od n), ima smisla x-os transformirati tako da na većem dijelu grafa vidimo manje vrijednosti k. Zato ponekad koristimo graf

$$((\Theta, H^{-1}_{\lceil n^{\Theta} \rceil, n}), 0 \leq \Theta \leq 1)$$

koji nazivamo altHill.

Drugi problem koji se pojavljuje kod Hillovog procjenitelja je visoka varijanca. To možemo smanjiti tako da koristimo takozvani smooHill, koji uzima prosjek vrijednosti Hillovih procjenitelja za nekoliko okolnih vrijednosti k. Odaberimo prirodni broj r, najčešće 2 ili 3, i definiramo

$$smooH_{k,n} = \frac{1}{(r-1)k} \sum_{j=k+1}^{rk} H_{j,k}.$$

Ovo je također konzistentan procjenitelj za $1/\alpha$, a graf izgleda glatko, bez toliko naglih skokova kao graf običnog Hillovog procjenitelja. I altHill i smooHill su detaljnije opisani u [7] u potpoglavlju 4.4.3.

Poglavlje 3

Mjere repne zavisnosti

U ovom poglavlju promatramo razne mjere repne zavisnosti komponenata slučajnog vektora teškog repa. Jednostavan način mjerenja zavisnosti slučajnih varijabli je Pearsonov koeficijent korelacije, ali u slučaju vektora teškog repa on ne daje neke informacije koje bi nam mogle biti od interesa - primjerice je li zavisnost koncentrirana u malim vrijednostima, a asimptotski su nezavsine ili je zavisnost koncentrirana u repu. Ovdje opisujemo neke mjere zavisnosti koje su bolje prilagođene mjerenju asimptotske zavisnosti komponenti slučajnih vektora teškog repa. Kako bismo mogli govoriti o tome, potreban nam je koncept višedimenzionalne regularne varijacije.

3.1 Višedimenzionalna regularna varijacija

U ovom potpoglavlju definiramo višedimenzionalnu regularnu varijaciju i iznosimo neke osnovne rezultate o njoj, od čega je najbitniji višedimenzionalni analogon teoremu 1.12. Definirajmo prvo pojmove koji će nam biti potrebni.

Kažemo da je skup $C \subset \mathbb{R}^d$ *konus* ako za svaki $\mathbf{x} \in C$ i $t > 0$ vrijedi $t\mathbf{x} \in C$. Kažemo da je funkcija $h : C \mapsto (0, \infty)$ monotona ako je ili nerastuća po svim komponentama ili nepadajuća po svim komponentama.

Neka je $h \geq 0$ izmjeriva funkcija na konusu C. Pretpostavimo da je $\mathbf{1} = (1, ..., 1) \in C$. Kažemo da je funkcija h *višedimenzionalna regularno varirajuća s funkcijom limesa* $\lambda : C \mapsto (0, \infty)$ ako za svaki $\mathbf{x} \in C$ vrijedi

$$\lim_{t \to \infty} \frac{h(t\mathbf{x})}{h(t\mathbf{1})} = \lambda(\mathbf{x}).$$

Vidimo da je ova definicija zaista poopćenje definicije jednodimenzionalne regularne varijacije. Također, očito je $\lambda(\mathbf{1}) = 1$. Fiksirajmo $\mathbf{x} \in C$ i definirajmo funkciju $U_{\mathbf{x}} : (0, \infty) \mapsto [0, \infty)$ sa $U_{\mathbf{x}}(t) = h(t\mathbf{x})$. Za $s > 0$

$$\lim_{t\to\infty}\frac{U_{\mathbf{x}}(ts)}{U_{\mathbf{x}}(t)}=\lim_{t\to\infty}\frac{h(ts\mathbf{x})}{h(t\mathbf{x})}=\lim_{t\to\infty}\frac{h(ts\mathbf{x})}{h(t\mathbf{1})}\Big/\frac{h(t\mathbf{x})}{h(t\mathbf{1})}=\frac{\lambda(s\mathbf{x})}{\lambda(\mathbf{x})}.$$

Dakle, funkcija $U_{\mathbf{x}}$ je regularno varirajuća s nekim indeksom $\rho(\mathbf{x})$, pa iz propozicije 1.2 slijedi da je $\lambda(s\mathbf{x})/\lambda(\mathbf{x}) = s^{\rho(\mathbf{x})}$. Dokažimo sada da ρ ne ovisi o $\mathbf{x}$.

$$s^{\rho(\mathbf{y})}=\lim_{t\to\infty}\frac{h(ts\mathbf{y})}{h(t\mathbf{y})}=\lim_{t\to\infty}\frac{\frac{h(ts\mathbf{y})}{h(ts\mathbf{x})}}{\frac{h(t\mathbf{y})}{h(t\mathbf{x})}}\cdot\frac{h(ts\mathbf{x})}{h(t\mathbf{x})}=\frac{\frac{\lambda(\mathbf{y})}{\lambda(\mathbf{x})}}{\frac{\lambda(\mathbf{y})}{\lambda(\mathbf{x})}}\cdot s^{\rho(\mathbf{x})}=s^{\rho(\mathbf{x})}.$$

Budući da ovaj račun vrijedi za svaki $\mathbf{y}$, ρ je konstanta, pa je $\lambda(s\mathbf{x}) = s^{\rho}\lambda(\mathbf{x})$. Sada možemo dokazati jednu karakterizaciju višedimenzionalne regularne varijacije.

Propozicija 3.1. *Funkcija $h : C \mapsto (0,\infty)$ je višedimenzionalno regularno varirajuća s graničnom funkcijom λ ako i samo ako postoji regularno varirajuća funkcija $V : (0,\infty) \mapsto (0,\infty)$ takva da za svaki $\mathbf{x}$ vrijedi*

$$\lim_{t\to\infty}\frac{h(t\mathbf{x})}{V(t)}=\lambda(\mathbf{x}). \tag{3.1}$$

Dokaz. Pretpostavimo da je h regularno varirajuća. Stavimo $V(t) = h(t\mathbf{1})$ i tvrdnja slijedi iz definicije ako dokažemo da je funkcija $t \mapsto h(t\mathbf{1})$ regularno varirajuća. Zaista,

$$\lim_{t\to\infty}\frac{h(t(s\mathbf{1}))}{h(t\mathbf{1})}=\lambda(s\mathbf{1})=s^{\rho}.$$

Obratno, pretpostavimo da vrijedi (3.1). Tada je

$$\lim_{t\to\infty}\frac{h(t\mathbf{x})}{h(t\mathbf{1})}=\lim_{t\to\infty}\frac{h(t\mathbf{x})}{V(t)}\Big/\frac{h(t\mathbf{1})}{V(t)}=\lambda(\mathbf{x})/\lambda(\mathbf{1})$$

Tada je h regularno varirajuća s graničnom funkcijom $\lambda(\mathbf{x})$ (do na multiplitativnu konstantu), što je trebalo dokazati. □

Često je jednostavnije raditi s transformacijom u polarne koordinate. Za neku fiksnu normu na R^d, definiramo jediničnu sferu

$$S := \{\mathbf{x} : \|\mathbf{x}\| = 1\}.$$

Označimo sa S_+ nenegativni dio jedinične sfere, $S_+ = S \cap [\mathbf{0},\infty)$. Definiramo preslikavanje vektora u polarne koordinate $T : R^d\backslash\{\mathbf{0}\} \mapsto (0,\infty) \times S$ sa

$$T(\mathbf{x})=\Big(\|\mathbf{x}\|,\frac{\mathbf{x}}{\|\mathbf{x}\|}\Big).$$

Morali smo isključiti nulu iz domene kako bi T bila bijekcija. Ovako postoji inverz T^{-1} : $(0,\infty)\times S \mapsto R^d\backslash\{\mathbf{0}\}$ dan sa

$$T^{-1}(r,\mathbf{a}) = r\mathbf{a}.$$

I T i T^{-1} su neprekidne bijekcije. Ako je $\mathbf{X}$ slučajni vektor, pišemo $T(\mathbf{X}) = (R,\Theta)$.

Iskažimo sada analogon teoremu (1.12). Označimo $\mathbb{E} := [\mathbf{0},\infty]\backslash\{\mathbf{0}\}$ i promatrajmo to kao topološki prostor s relativnom topologijom s obzirom na euklidsku. U iskazu teorema pojam Radonova mjera podrazumjeva da mjera nije nul-mjera niti je degenerirana u jednu točku.

Teorem 3.2. *Neka je $\mathbf{Z} \geq \mathbf{0}$ d-dimenzionalni nenegativni slučajni vektor s funkcijom distribucije F. Tada su sljedeće tvrdnje ekvivalentne:*

(i) Postoji Radonova mjera ν na $\mathbb{E}$ takva da je

$$\lim_{t\to\infty}\frac{1-F(t\mathbf{x})}{1-F(t\mathbf{1})} = \lim_{t\to\infty}\frac{\mathbb{P}(\mathbf{Z}/t\in[\mathbf{0},\mathbf{x}]^c)}{\mathbb{P}(\mathbf{Z}/t\in[\mathbf{0},\mathbf{1}]^c)} = \nu([\mathbf{0},\mathbf{x}]^c)$$

za sve točke $\mathbf{x}\in[\mathbf{0},\infty)\backslash\{0\}$ u kojima je funkcija $\nu([\mathbf{0},\cdot]^c)$ neprekidna.

(ii) Postoji funkcija $b(t)\to\infty$ i Radonova mjera ν na $\mathbb{E}$ koju zovemo granična mjera *takva da u $M_+(\mathbb{E})$ vrijedi*

$$t\mathbb{P}(\mathbf{Z}/b(t)\in\cdot)\xrightarrow{v}\nu,\quad t\to\infty.$$

(iii) Postoji niz $b_n\to\infty$ i Radonova mjera ν na $\mathbb{E}$ takva da u $M_+(\mathbb{E})$ vrijedi

$$n\mathbb{P}(\mathbf{Z}/b_n\in\cdot)\xrightarrow{v}\nu,\quad n\to\infty.$$

(iv) Postoji vjerojatnosna mjera μ na S_+ koju zovemo kutna mjera, funkcija $b(t)\to\infty$ i konstanta $c>0$ takva da za $(R,\Theta) = (\|\mathbf{Z}\|,\mathbf{Z}/\|\mathbf{Z}\|)$ vrijedi

$$t\mathbb{P}\Big(\Big(\frac{R}{b(t)},\Theta\Big)\in\cdot\Big)\xrightarrow{v}c\nu_\alpha\times\mu$$

u $M_+((0,\infty]\times S_+)$.

(v) Postoji vjerojatnosna mjera μ na S_+ koju zovemo kutna mjera, niz $b_n\to\infty$ i konstanta $c>0$ takva da za $(R,\Theta) = (\|\mathbf{Z}\|,\mathbf{Z}/\|\mathbf{Z}\|)$ vrijedi

$$n\mathbb{P}\Big(\Big(\frac{R}{b_n},\Theta\Big)\in\cdot\Big)\xrightarrow{v}c\nu_\alpha\times\mu$$

u $M_+((0,\infty]\times S_+)$.

Dokaz teorema može se naći u [7], Theorem 6.1. Često nas zanima jesu li repovi podjednako teški, odnosno želimo izračunati

$$\lim_{x\to\infty} \frac{\mathbb{P}(Z^{(i)} > x)}{\mathbb{P}(Z^{(j)} > x)} =: r_{ij} \in [0, \infty].$$

Ako taj limes postoji i vrijedi $0 < r_{ij} < \infty$, kažemo da su marginalne distribucije *ekvivalentne u repu*

U teorijskim je razmatranjima jednostavnije raditi kada su sve marginalne distribucije ekvivalente u repu. Ako to nije slučaj, onda će kutna mjera biti koncentrirana u presjeku S_+ i potprostora generiranog onim osima kojima su pridruženi najmanji repni indeksi. Kada je $b(t) = t$ odnosno $b_n = n$, kažemo da smo u *standardnom slučaju*. Tada su sve marginalne distribucije u repu ekvivalentne Paretovoj distribuciji s $\alpha = 1$. U praksi, obično ne možemo normirati sve komponente istom funkcijom, pa ćemo morati na neki način dobiti jednake težine repova, o čemu će biti govora kasnije.

Povežimo sada ekvivalencije iz teorema 3.2 sa slabom konvergencijom odgovarajućih točkovnih mjera u Poissonovu slučajnu mjeru.

Teorem 3.3. *Neka je* $\{\mathbf{Z}, \mathbf{Z}_1, ...\}$ *niz nezavisnih jednakodistribuiranih slučajnih varijabli i neka je* $\{(R, \Theta), (R_1, \Theta_1), ...\}$ *niz njihovih polarnih transformacija. Tada su ekvivalencije iz teorema 3.2 također ekvivalentne sljedećim tvrdnjama:*

(vi) Postoji niz $b_n \to \infty$ *takav da*

$$\sum_{i=1}^{n} \epsilon_{\mathbf{Z}_i/b_n} \Rightarrow PRM(\nu)$$

u $M_p(\mathbb{E})$.

(vii) Postoji niz $b_n \to \infty$ *takav da*

$$\sum_{i=1}^{n} \epsilon_{R_i/b_n, \Theta_i} \Rightarrow PRM(c\nu_\alpha \times \mu)$$

u $M_p((0, \infty] \times S_+)$.

Bilo koji od uvjeta (i) - (vii) također povlači

(viii) Vrijedi

$$\frac{1}{k} \sum_{i=1}^{n} \epsilon_{\mathbf{Z}_i/b(n/k)} \Rightarrow \nu$$

u $M_+(\mathbb{E})$

(ix) Vrijedi

$$\frac{1}{k}\sum_{i=1}^{n}\epsilon_{(R_i/b(n/k),\Theta_i)} \Rightarrow c\nu_\alpha \times \mu$$

u $M_+((0,\infty]\times S_+)$.

Također, (viii) i (ix) su ekvivalentni (i) - (vii) ako za $k(\cdot)$ *vrijedi* $k(n)\sim k(n+1)$.

Promotrimo sada problem različite težine repova. Uvjet

$$n\mathbb{P}(\mathbf{Z}_1/b_n\in\cdot)\xrightarrow{v}\nu$$

u $M_+(\mathbb{E})$ povlači da za svaki $i=1,...,d$ postoji c_i takav da

$$n\mathbb{P}(\mathbf{Z}_1^{(i)}/b_n\in\cdot)\xrightarrow{v}c_i\nu_\alpha$$

u $M_+((0,\infty])$. Tada će sve komponente kojima je c_i različit od 0 imati isti repni indeks α, ali se može dogoditi da ima i komponenti s takvih da je $c_i=0$, odnosno da imaju veći repni indeks i za takve komponente naš model neće biti informativan. Primjerice, ako je P slučajna varijabla s jediničnom Paretovom distribucijom i $\mathbf{Z}=(P,P^2)$, onda skaliranjem istom funkcijom dobijemo

$$n\mathbb{P}(\left(\left(\frac{P}{n^2},\frac{P^2}{n^2}\right)\in\cdot\right)\xrightarrow{v}\epsilon_0\times\nu_{1/2}.$$

Kada bismo prvu komponentu normirali sa n, a drugu sa n^2, dobili bismo

$$n\mathbb{P}(\left(\left(\frac{P}{n},\frac{P^2}{n^2}\right)\in\cdot\right)\xrightarrow{v}\nu_1\circ T_{1,1^2}^{-1},$$

gdje $T_{1,1^2}:(0,\infty]\mapsto(0,\infty]\times(0,\infty]$ definiramo sa $T_{1,1^2}(x)=(x,x^2)$.

Jasno, ovo nam daje više informacija nego prethodni izraz, pa ćemo često htjeti normirati različite komponente različitim nizovima. Kako bismo to mogli kompaktnije zapisati, definiramo dijeljenje vektora iste dimenzije kao dijeljenje po komponentama $\mathbf{a}/\mathbf{b}=(a_1/b_1,...,a_d/b_d)$. Sljedeći nam teorem daje jednostavan način kako svesti opću regularno varirajuću funkciju na standardni slučaj.

Teorem 3.4. *Neka je* $\mathbf{Z}=(Z^{(1)},...,Z^{(d)})$ *slučajni vektor s nenegativnim komponentama i marginalnim funkcijama doživljenja* $\overline{F}_{(i)}$ *i* $\mathbb{E}=[\mathbf{0},\infty]\backslash\{\mathbf{0}\}$. *Pretpostavimo da za svaki* $i=1,...d$ *postoji niz* $b_n^{(i)}\to\infty$ *takav da vrijedi:*

(i) Marginalna regularna varijacija: za svaki $i=1,...d$ *postoji* $\alpha_i>0$ *takav da*

$$nP(Z^{(i)}/b_n^{(i)}\in\cdot)\xrightarrow{v}\nu_{\alpha_i}$$

u $M_+(0,\infty]$.

(ii) Nestandardna globalna regularna varijacija: Postoji mjera ν na Borelovoj σ-algebri na $\mathbb{E}$ takva da

$$nP(\mathbf{Z}/b_n \in \cdot) \xrightarrow{v} \nu$$

u $M_+(\mathbb{E})$.

Neka je $b^{(i)}$ funkcija kvantila pridružena $F_{(i)}$ i neka je $b_n^{(i)} := b^{(i)}(n)$. Tada vrijedi:

(i) Standardna globalna regularna varijacija:

$$nF_*(n\cdot) := n\mathbb{P}\left(\left(\frac{(b^{(i)})^{-1}(Z^{(i)})}{n}, i = 1, ..., d\right) \in \cdot\right) \xrightarrow{v} \nu_*(\cdot)$$

u $M_+(\mathbb{E})$ gdje je

$$\nu_*(t\cdot) = t^{-1}\nu_*(\cdot)$$

na Borelovim podskupovima od $\mathbb{E}$.

(ii) Standardna marginalna konvergencija: Za svaki $i = 1, ...d$ vrijedi

$$n\mathbb{P}\left(\frac{(b^{(i)})^{-1}(Z^{(i)})}{n} > x\right) \to x^{-1}, \quad x > 0.$$

3.2 Kutna mjera

U iskazu teorema 3.2 definirali smo kutnu mjeru kao mjeru na nenegativnom dijelu jedinične sfere koja opisuje zajedničko ponašanje repova slučajnih varijabli. Također, komentirali smo da višedimenzionalnu regularnu varijaciju funkcije doživljenja slučajnog vektora nužno da su sve funkcije doživljenja njegovih komponenata regularno varirajuće s istim α. Zato ćemo prvo pretpostaviti da smo u standardnom slučaju, (odnosno da su svi α jednaki 1), a zatim ćemo diskutirati kako svesti nestandardni slučaj (koji se u praksi gotovo uvijek pojavljuje) na standardni.

Neka je $\mathbf{Z}_j$ niz nezavisnih jednakodistribuiranih slučajnih vektora čija je zajednička funkcija distribucije F višedimenzionalna regularno varirajuća i da se sve komponente mogu skalirati istom funkcijom $b(t)$. Iz ekvivalencija danih teoremima 3.2 i 3.3 slijedi da postoji funkcija $b(t) \to \infty$ i Radonova mjera ν na $\mathbb{E} = [\mathbf{0}, \infty]\backslash\{\mathbf{0}\}$ takva da $t\mathbb{P}(\mathbf{Z}_1/b(t) \in \cdot) \xrightarrow{v} \nu$, to jest da za $n \to \infty, k \to \infty, k/n \to 0$,

$$\frac{1}{k}\sum_{i=1}^{n} \epsilon_{\mathbf{Z}_i/b(n/k)} \Rightarrow \nu, \tag{3.2}$$

odnosno u polarnim koordinatama

$$\frac{1}{k}\sum_{i=1}^{n} \epsilon_{(R_i/b(n/k),\Theta_i)} \Rightarrow c\nu_\alpha \times \mu. \tag{3.3}$$

Slično kao u prethodnom poglavlju i ovdje se pojavljuje problem što u praksi ne znamo funkciju kvanrila b, no kada bismo je znali, iz (3.2) bismo mogli dobiti konzistentan procjenitelj granične mjere ν. Ovo nam također daje konzistentan procjenitelj kutne mjere μ. Uvrštavanjem cijelog intervala $[1, \infty]$ u prvi argument mjere u izrazu (3.3) dobijemo

$$\frac{\sum_{i=1}^{n} \epsilon_{(R_i/b(n/k),\Theta_i)}([1,\infty] \times \cdot)}{\sum_{i=1}^{n} \epsilon_{R_i/b(n/k)}[1,\infty]} \Rightarrow \mu(\cdot)$$

zbog činjenice da se nazivnik asimptotski približava k. Slično, ako u drugi argument mjere u izrazu (3.2) uvrstimo cijeli S_+ dobijemo

$$\frac{1}{k}\sum_{i=1}^{n} \epsilon_{R_i/b(n/k))} \Rightarrow c\nu_\alpha.$$

Kao u dokazu teorema 2.8 dobivamo

$$R_{(k)}/b(n/k) \xrightarrow{P} 1.$$

gdje je $R_{(k)}$ k-ti najveći od brojeva $R_1, ..., R_n$, pa možemo (slično kao u prethodnom poglavlju) nepoznati teorijski kvantil $b(n/k)$ zamijeniti empirijskim kvantilom $\hat{b}(n/k) = R_{(k)}$. U teoriji, kada bismo znali da smo u standardnom slučaju, mogli bismo uvrstiti jednostavno $b(t) = t$, ali u praksi se ispostavlja da je bolje skalirati sa $R_{(k)}$.

Pretpostavka da možemo skalirati svaku komponentu istom funkcijom kvantila b je vrlo jaka i u praksi je rijetko ispunjena, pa moramo naći metoda za procjenu kutne mjere i u nestandardnom slučaju.

Najjednostavnija takva metoda je ponovo uvesti pretpostavku da za svaki i, i-ta marginalna distribucija ima egzaktan Paretov rep s indeksom α_i od nekog mjesta nadalje i zatim transformirati svaku komponentu podizanjem na α_i. Dakle, ako je X slučajna varijabla čija je distribucija asimptotski Paretova, odnosno $\mathbb{P}(X > x) \sim x^{-\alpha}$ kad $x \to \infty$, onda

$$\mathbb{P}(X^\alpha > x) = \mathbb{P}(X > x^{1/\alpha}) \sim x^{-1}, \quad x \to \infty.$$

Ovime smo slučajnu varijablu proizvoljnog repnog indeksa transformirali u varijablu repnog indeksa 1. Kada istu transformaciju izvršimo na svim komponentama, novom vektoru su sve komponente jednako teške, pa ga možemo analizitati kao u standardnom slučaju, odnosno nadati se da je

$$\frac{1}{k}\sum_{i=1}^{n} \epsilon_{((Z_i^{(j)}/\hat{b}^{(j)}(n/k))^{\alpha_j};j=1...d)} \tag{3.4}$$

dobar procjenitelj granične mjere ν_* što onda možemo iskoristiti za procjenu kutne mjere μ. U praksi naravno ne znamo vektor $\alpha = (\alpha_1, ..., \alpha_d)$, nego ga moramo zamijeniti procijenjenim vektorom $\hat{\alpha}$ dobivenim analizom jednodimenzionalnih podataka kao u prethodnom poglavlju. Ovdje se pojavljuje problem što uopće počinjemo analizirati višedimenzionalne podatke nakon što smo ih transformirali procijenjenim vektorom α čime povećavamo grešku.

3.3 Koeficijent zavisnosti χ

Ponekad želimo repnu zavisnost izmjeriti jednim brojem koji će nam dati mjeru zavisnosti zajedničkog pojavljivanja ekstremnih vrijednosti, iako očito ne možemo sve informacije o zavisnosti spremiti u jedan broj (osim u vrlo posebnim situacijama kao na primjer kod bivarijatne normalne distribucije kada cijelu zavisnost možemo opisati koeficijentom korelacije). Koristan alat za opisivanje strukture zavisnosti je funkcija kopule.

Definicija 3.5. *Neka je (X, Y) slučajni vektor s neprekidnom funkcijom distribucije F i marginalnim funkcijama distribucije F_X i F_Y. Tada funkciju $C : [0,1] \times [0,1] \mapsto [0,1]$ definiranu sa $C(u,v) = F(F_X^{-1}(u), F_Y^{-1}(v))$ zovemo* kopula slučajnog vektora (X, Y).

Intuitivno, kopula za brojeve u i v daje vjerojatnost da je X manja ili jednaka od svog u-tog kvantila i Y manja ili jednaka od svog v-tog kvantila. Kopula nam daje sve informacije o zavisnosti dvije slučajne varijable, ali nam ne daje nikakve informacije o marginalnim distribucijama. Za nju vrijedi, $F(x,y) = C(F_X(x), F_Y(y))$, odnosno C je funkcija distribucije vektora (X, Y) nakon što mu obje komponente monotono transformiramo u uniformne slučajne varijable na $[0,1]$. Pogledajmo sada nekoliko primjera kopula. Ako su slučajne varijable X i Y nezavisne, kopula je jednaka

$$C(u,v) = F(F_X^{-1}(u), F_Y^{-1}(v)) = F_X(F_X^{-1}(u))F_Y(F_Y^{-1}(v)) = uv.$$

Ako je jedna od njih jednaka monotonoj transformaciji druge, $Y = F_Y^{-1}(F_X(X))$, tada $F(x,y) = \mathbb{P}(X < x, F_Y^{-1}(F_X(X))) = \mathbb{P}(F_X(X) < F_X(x), F_X(X) < F_Y(y)) = \min\{F_X(x), F_Y(y)\}$, odnosno

$$C(u,v) = \min\{F_X(F_X^{-1}(u)), F_Y(F_Y^{-1}(v))\} = \min\{u,v\}.$$

Ako su slučajne varijable iz bivarijatne logističke distribucije ekstremnih vrijednosti, tada je kopula jednaka

$$C(u,v) = e^{-((-\log u)^{1/\alpha} + (-\log v)^{1/\alpha})^{\alpha}}.$$

Za neki $\alpha \in (0,1]$. U tom slučaju $\alpha = 1$ odgovara nezavisnim slučajnim varijablama, a u limesu $\alpha \to 0$ dobivamo savršenu monotonu zavisnost. Ako je (X, Y) vektor iz bivarijatne normalne distribucije s koeficijentom korelacije ρ, tada je

$$C(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\{-\frac{1}{2(1-\rho^2)}(s^2 - 2\rho st + t^2)\} ds\, dt$$

gdje je Φ funkcija distribucije standardne normalne slučajne varijable.

Ako su slučajne varijable X i Y jednako distribuirane, ima smisla promatrati vjerojatnost da jedna od njih postiže ekstremnu vrijednost uz uvjet da druga postiže ekstremnu vrijednost, $\lim_{z\to z^*} \mathbb{P}(Y > z|X > z)$. Gdje je $z^* \in \mathbb{R} \cup \{\infty\}$ gornja granica nosača slučajne varijable X, odnosno Y. U slučaju da X i Y nisu jednako distribuirane, nema ih smisla obje uspoređivati s istim brojem z. Međutim, ako (X, Y) monotono transformiramo u uniformne (U, V) (što je upravo ono što kopula radi), ekstremni događaji (primjerice postizanje vrijednosti veće od 99. kvantila) ostaju sačuvani, pa možemo definirati

$$\chi = \lim_{u\to 1} \mathbb{P}(V > u|U > u)$$

ako taj limes postoji. Kažemo da su slučajne varijable X i Y *asimptotski nezavisne* ako je $\chi = 0$. U praksi je χ jednostavnije računati kao limes jedne druge, asimptotski ekvivalentne, funkcije.

$$\begin{aligned}
\lim_{u\to 1} \mathbb{P}(V > u|U > u) &= \lim_{u\to 1} \frac{\mathbb{P}(U > u, V > u)}{P(U > u)} \\
&= \lim_{u\to 1} \frac{1 - 2u + C(u, u)}{1 - u} \\
&= \lim_{u\to 1} 2 - \frac{1 - C(u, u)}{1 - u} \\
&= \lim_{u\to 1} 2 - \frac{1 - C(u, u)}{\log C(u, u)} \frac{\log u}{1 - u} \frac{\log C(u, u)}{\log u} \\
&= \lim_{u\to 1} 2 - \frac{\log C(u, u)}{\log u}.
\end{aligned}$$

Dakle ako definiramo

$$\chi(u) = 2 - \frac{\log \mathbb{P}(U < u, V < u)}{\log \mathbb{P}(U < u)},$$

slijedi da je $\chi = \lim_{u\to 1} \chi(u)$. Za ovako definiranu funkciju $\chi(u)$ vrijedi

$$2 - \frac{\log(2u - 1)}{\log(u)} \leq \chi(u) \leq 1,$$

gdje izraz na lijevoj strani shvaćamo kao $-\infty$ za $u < 1/2$, a 0 za $u = 1$. Izračunajmo sada χ za neke primjere. Ako su slučajne varijable nezavisne, tada je

$$\begin{aligned}
\lim_{u\to 1} \chi(u) &= \lim_{u\to 1} 2 - \frac{\log \mathbb{P}(U < u, V < u)}{\log \mathbb{P}(U < u)} \\
&= \lim_{u\to 1} 2 - \frac{\log \mathbb{P}(U < u)\mathbb{P}(V < u))}{\log \mathbb{P}(U < u)} \\
&= \lim_{u\to 1} 2 - \frac{2 \log \mathbb{P}(U < u)}{\log \mathbb{P}(U < u)} = 0.
\end{aligned}$$

Ako su X i Y monotona transformacija jedna druge, odnosno $U = V$, tada je

$$\begin{aligned}\lim_{u\to 1}\chi(u) &= \lim_{u\to 1} 2 - \frac{\log \mathbb{P}(U < u, V < u)}{\log \mathbb{P}(U < u)} \\ &= \lim_{u\to 1} 2 - \frac{\log \mathbb{P}(U < u)}{\log \mathbb{P}(U < u)} = 1\end{aligned}$$

Ako je (X, Y) iz bivarijatne logističke distribucije ekstremnih vrijednosti,

$$\begin{aligned}\lim_{u\to 1}\chi(u) &= \lim_{u\to 1} 2 - \frac{\log \mathbb{P}(U < u, V < u)}{\log \mathbb{P}(U < u)} \\ &= \lim_{u\to 1} 2 - \frac{\log \exp(-((-\log u)^{1/\alpha} + (-\log v)^{1/\alpha})^{\alpha})}{\log u} \\ &= \lim_{u\to 1} 2 - \frac{-(2(-\log u)^{1/\alpha})^{\alpha}}{\log u} = 2 - 2^{\alpha}.\end{aligned}$$

U ova tri primjera je $\chi(u)$ konstanta, ali to općenito nije istina. Primjerice, za bivarijatnu normalnu distribuciju $\chi(u)$ (s korelacijom različitom od -1, 0 i 1) je $\chi(u)$ integral koji se mora izračunati numerički. Za vrijednosti u daleko od 1, $\chi(u)$ je različit on 0 i veći je što je veći koeficijent korelacije ρ, ali za svaki $-1 < \rho < 1$ je $\lim_{u\to 1}\chi(u) = 0$, što sugerira da χ nije jako informativan kada podaci ne dolaze iz distribucije teškog repa.

U praksi $\chi(u)$ procjenjujemo koristeći [5]

$$\hat{\chi}_n(u) = \frac{1}{n(1-u)} \sum_{i=1}^{n} \mathbb{1}_{(x_i > x_{(\lfloor nu \rfloor)}, y_i > y_{(\lfloor nu \rfloor)})}, \tag{3.5}$$

gdje sa $x_{(k)}$ označavamo k-ti najveći x. Koeficijent χ procjenjujemo kao $\chi(u)$ za neki u blizu 1. Ako odaberemo prevelik u, riskiramo da donosimo zaključke na premalo podataka, a ako odaberemo premali, $\chi(u)$ može biti loša aproksimacija limesa.

3.4 Koeficijent zavisnosti $\bar{\chi}$

U praksi često želimo na neki način izmjeriti repnu zavisnost nekih slučajnih varijabli koje su asimptotski nezavisne. Ovdje nam mjera χ ne može pomoći jer je po definiciji za asimptotski nezavisne slučajne varijable $\chi = 0$. Definirajmo zato drugi koeficijent zavisnosti $\bar{\chi}$ tako da bude informativan u slučaju $\chi = 0$. Definiramo funkciju $\bar{C}$ kao

$$\bar{C}(u, v) = 1 - u - v + C(u, v).$$

Za tako definiranu funkciju vrijedi

$$\bar{F}(x, y) = 1 - F_X(x) - F_Y(y) + F(x, y) = \bar{C}(F_X(x), F_Y(y)).$$

Sada, analogno definiciji koeficijenta χ definiramo

$$\bar{\chi}(u) = \frac{2\log \mathbb{P}(U > u)}{\log \mathbb{P}(U > u, V > v)} - 1 = \frac{2\log(1-u)}{\log \bar{C}(u,u)} - 1$$

za $u \in [0,1]$. Tada je $\bar{\chi}(u) \in (-1,1]$. Koeficijent zavisnosti $\bar{\chi}$ onda definiramo kao

$$\bar{\chi} = \lim_{u\to 1} \bar{\chi}(u).$$

Očito vrijedi $-1 \leq \bar{\chi} \leq 1$. Izračunajmo $\bar{\chi}$ za naše ranije primjere kopula. Ako su X i Y nezavisne,

$$\begin{aligned}\bar{\chi} &= \lim_{u\to 1} \frac{2\log(1-u)}{\log(\mathbb{P}(U>u)\mathbb{P}(V>u))} - 1 \\ &= \lim_{u\to 1} \frac{2\log(1-u)}{2\log(1-u)} - 1 = 0.\end{aligned}$$

Ako je Y monotona funkcija do X, onda je

$$\bar{\chi} = \lim_{u\to 1} \frac{2\log(1-u)}{\log(1-u)} - 1 = 1.$$

Ako je (X,Y) iz bivarijatne logističke distribucije ekstremnih vrijednosti,

$$\begin{aligned}\bar{\chi} &= \lim_{u\to 1} \frac{2\log(1-u)}{\log(1-2u+\exp(2^\alpha \log u))} - 1 \\ &= \lim_{u\to 1} \frac{2\log(1-u)}{\log(1-2u+u^{2^\alpha})} - 1 \\ &= \text{(dvaput primijenimo L'Hopitalovo pravilo)} = 1.\end{aligned}$$

za $\alpha > 0$. Može se dokazati da sve asimptotski zavisne slučajne vektore vrijedi $\bar{\chi} = 1$, dok za asimptotski nezavisne vrijedi $\bar{\chi} < 1$. Za bivarijatni normalni slučajni vektor može se dokazati da je $\bar{\chi} = \rho$.

Analogno (3.5), koeficijent $\bar{\chi}$ procjenjujemo kao [5]

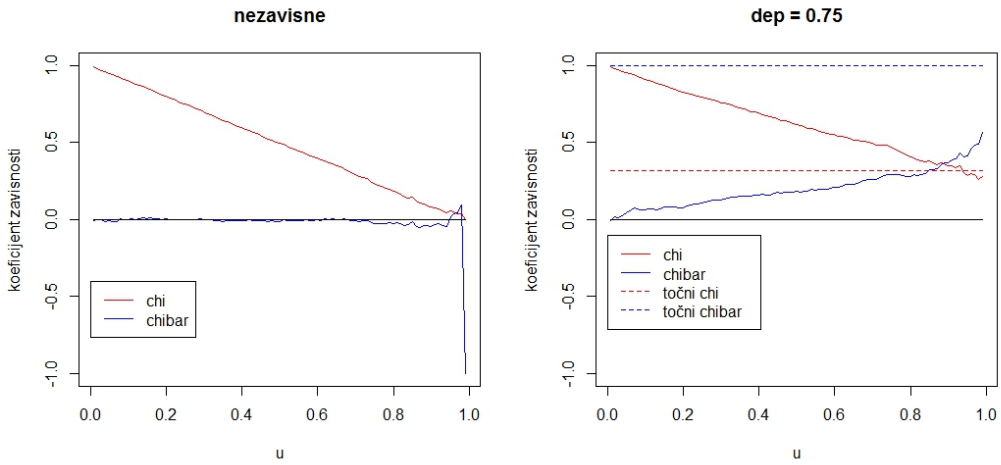
$$\hat{\bar{\chi}}_n(u) = \frac{2\log(1-u)}{\log\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(x_i > x_{(\lfloor nu \rfloor)}, y_i > y_{(\lfloor nu \rfloor)})}\right)} - 1, \tag{3.6}$$

a $\bar{\chi}$ procjenjujemo kao $\bar{\chi}(u)$ za dovoljno velik u.

Dakle, $\chi \in [0,1]$ je jednak 0 ako su slučajne varijable asimptotski nezavisne, a u intervalu $(0,1]$ ako su asimptotski zavisne, tako da on daje dodatne informacije o asimptotski zavisnim slučajnim vektorima, dok je $\bar{\chi} \in [-1,1]$ jednak 1 ako su asimptotski zavisne, a u intervalu $[-1,1)$ ako su asimptotski nezavisne, pa on daje više informacija o zavisnosti u asimptotski nezavisnom slučaju.

3.5 Koeficijenti repne zavisnosti u praksi

U ovom potpoglavlju procjenjujemo koeficijente repne zavisnosti na nekim stvarnim podacima. Pogledajmo prvo kako procjenitelji dani sa (3.5) i (3.6) procjenjuju zavisnost na podacima iz distribucija za koje znamo što bismo trebali dobiti. Pomoću paketa evd za R, generirali smo 2500 parova podataka iz bivarijatne logističke distribucije ekstremnih vrijednosti, sa koeficijentima zavisnosti 1 (nezavisne), 0.75 i 0.5 i 0.25.

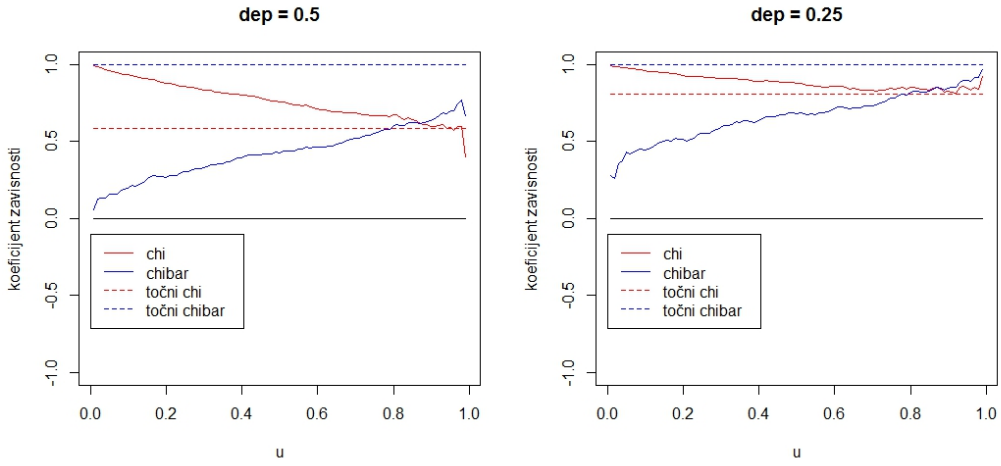


Slika 3.1: Grafovi $\chi(u)$ i $\bar{\chi}(u)$ za podatke iz bivarijatne logističke distribucije ekstremne vrijednosti

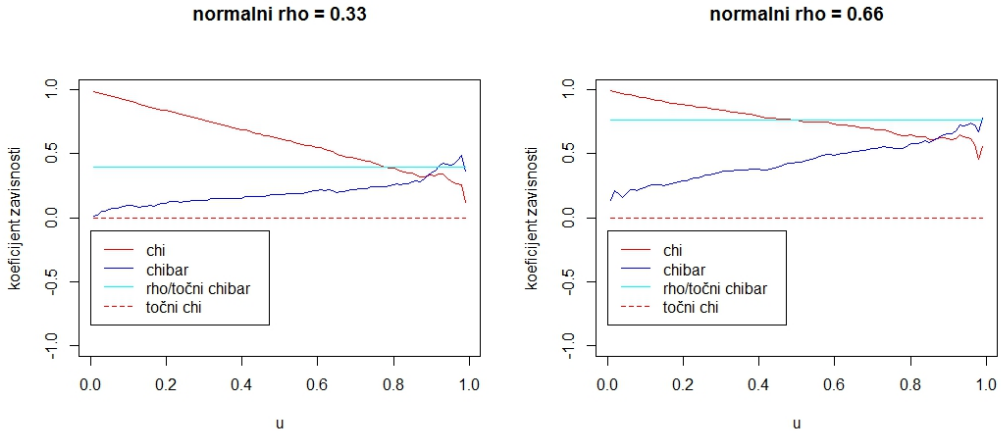
Vidimo da je u slučaju asimptotski zavisnih varijabili $\chi(0.9)$ već prilično blizu točnoj teorijskoj vrijednosti, što sugerira da je na uzorku ove veličine $u = 0.9$ dovoljno blizu 1 da bismo koristili $\chi(0.9)$ kao procjenu za χ, a istovremeno je i dovoljno daleko od 1 da varijanca procjenitelja nije prevelika. Za razliku od toga, $\bar{\chi}$ ne prilazi blizu vrijednosti 1 čak niti na ovoliko velikom uzorku niti za velike u kad graf postane volatilan. Pogledajmo sada slučaj bivarijatne normalne distribucije sa koeficijentom korelacije 0.33 i 0.66.

Vidimo da u ovom slučaju, kada je teorijski χ jednak 0, $\chi(u)$ zaista pada, ali ne izgleda kao da konvergira u 0. S druge strane, $\bar{\chi}(u)$ je vrlo blizu pravoj vrijednosti za u blizu 0.9. Pogledajmo sada mjere zavisnosti za ukupnu dnevnu vrijednost trgovanih dionica firmi Apple, Google, Microsoft i IBM, za koje smo ranije zaključili da su teškog repa. Promatramo korelacijsku matricu i za standardni Pearsonov koeficijent korelacije i za Spearmanov koeficijent korelacije rangova, te graf $\chi(u)$ i $\bar{\chi}(u)$ za svaki par dionica.

Relativni odnos veličina je sličan za Pearsonov i Spearmanov koeficijent korelacije, te oni sugeriraju da je najjača povezanost između Applea i IBM-a, a najslabija između Applea



Slika 3.2: Grafovi $\chi(u)$ i $\bar{\chi}(u)$ za podatke iz bivarijatne logističke distribucije ekstremne vrijednosti



Slika 3.3: Grafovi $\chi(u)$ i $\bar{\chi}(u)$ za generirane normalne podatke

i Googlea. Također, prilično je jaka povezanost Microsofta i s Googleom i sa IBM-om. Pogledajmo koeficijente repne zavisnosti.

Vidimo da za svaki par dionica koeficijent $\chi(u)$ konvergira u 0, pa zaključujemo da se radi o asimptotski nezavisnim slučajnim varijablama i zato promatramo $\bar{\chi}$ kako bismo izmjerili njihovu repnu zavisnost.

Rezultati za $\bar{\chi}$ su prilično slični kao i za koeficijente korelacije, Apple i Google su skoro

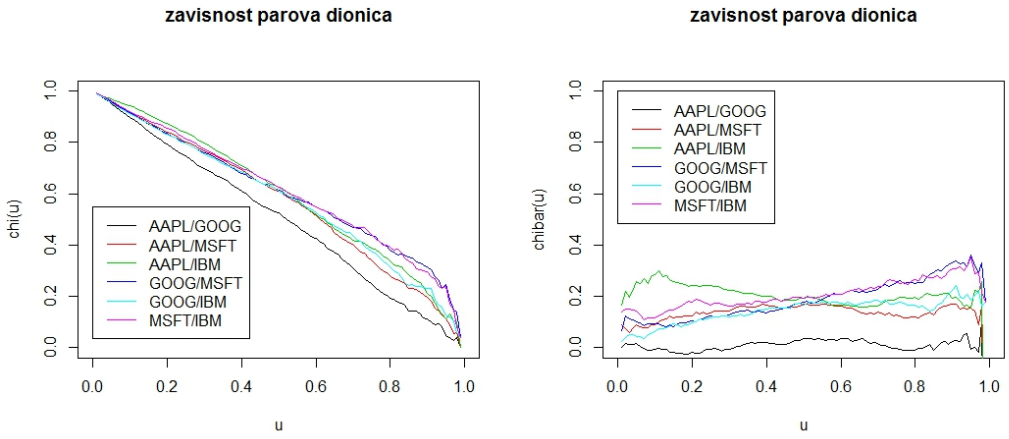
	AAPL	GOOG	MSFT	IBM
AAPL	1.00	0.00	0.17	0.30
GOOG	0.00	1.00	0.27	0.18
MSFT	0.17	0.27	1.00	0.26
IBM	0.30	0.18	0.26	1.00

(a) Pearsonov koeficijent

	AAPL	GOOG	MSFT	IBM
AAPL	1.00	0.02	0.28	0.39
GOOG	0.02	1.00	0.34	0.24
MSFT	0.28	0.34	1.00	0.36
IBM	0.39	0.24	0.36	1.00

(b) Spearmanov koeficijent

Slika 3.4



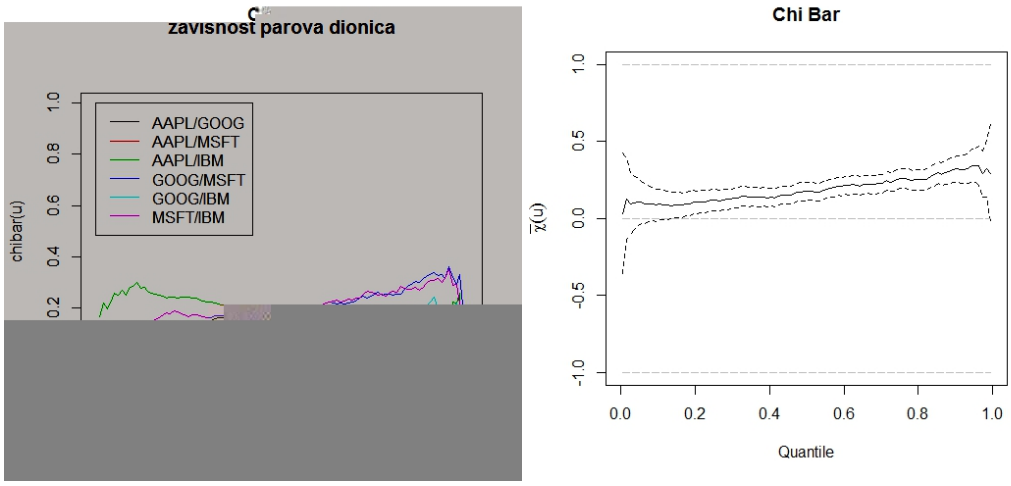
(a) Graf $\chi(u)$ za svaki par dionica

(b) Graf $\bar{\chi}(u)$ za svaki par dionica

Slika 3.5

potpuno nezavisni, a Microsoft je prilično zavisan i sa Googleom i sa IBM-om. Jedino što se jako razlikuje od običnog koeficijenta korelacije je zavisnost Applea i IBM-a koja je jako visoka za male u, a za velike postaje dosta slabija od zavisnosti Microsofta i Googlea odnosno IBM-a. To sugerira da su Apple i IBM jače zavisni blizu nuli, a mnogo slabije u repu, dok su druge zavisnosti koncentrirane u repu. Grafovi $\bar{\chi}(u)$ za parove Google i Microsoft te Apple i IBM sa 95% pouzdanim intervalima pokazuju da je njihova repna zavisnost značajno različita od 0. Za konstrukciju intervala pouzdanosti koristili smo softverski paket "texmex" za R koji koristi pretpostavke iz [2] da su observacije nezavisne, da empirijske marginalne distribucije savršeno aproksimiraju stvarne marginalne distribucije i da asimptotska distribucija procjenitelja dobro aproksimira stvarnu distribuciju procjenitelja na konačnom uzorku te ih računa koristeći delta-metodu [3]. Te su pretpostavke

prilično jake i vjerojatno u ovom slučaju nisu ispunjene, tako da te intervale treba shvatiti samo informativno.



(a) Apple i IBM (b) Google i Microsoft

Slika 3.6: 95%-tni intervali pouzdanosti za $\bar{\chi}$

Graf indicira da su svi parovi dionica asimptotski nezavisni, a koeficijent repne zavisnosti $\bar{\chi}$ se kreće od 0 (za Google i Apple) do 0.3.

Bibliografija

[1] P. Billingsley, *Convergence of Probabilty Measures*, Second Edition, John Wiley and Sons, 1999.

[2] S. Coles, J. Heffernan, J. Tawn, *Dependence Measures for Extreme Value Analyses*, Kluwer Academic Publishers, 1999.

[3] A. C. Davison, *Statistical Models*, Cambridge university press, 2003.

[4] H. Drees, L. de Haan, S. Resnick, *How to Make a Hill Plot*, The Annals of Statistics volume 28, number 1 (2000), 254-274.

[5] R.-D. Reiss, M. Thomas, *Statistical analysis of Extreme Values*, Third Edition, Birkhäuser, 1997.

[6] S. Resnick, *Adventures in Stochastic Processes*, Fourth Edition, Birkhäuser, 2005.

[7] S. Resnick, *Heavy-Tail Phenomena, Probabilistic and Statistical Modeling*, Springer, 2007.

Sažetak

Cilj ovog rada bio je istražiti metode procjene repnog indeksa te mjera zavisnosti kod distribucija teškog repa te ih isprobati u praksi, ali i upozoriti na probleme u njihovoj primjeni. Uveli smo pojam regularne varijacije iz matematičke analize i slabe i slabašne konvergencije iz teorije mjere te smo iskazali neka njihova svojstva.

Zatim smo definirali Hillov procjenitelj za repni indeks α slučajne varijable čija je funkcija doživljenja regularno varirajuća i dokazali njegovu konzistentnost. Da bismo ga isprobali u praksi, koristili smo slučajno generirane podatke iz Paretove, normalne i Cauchyjeve distribucije te podatke o dnevnom dolarskom volumenu dionica nekolicine tehnoloških kompanija. Tu su se pojavili slični problemi na koje literatura upozorava - nije lagano uopće ustvrditi kada je model teškog repa primjeren, a čak niti jednom kad smo ga odlučili koristiti nije jasno kako točno interpretirati Hillov procjenitelj, odnosno koji prag koristiti. Za neke je dionice Hillov graf jasno sugerirao vrijednost α između 3 i 3.5, dok za neke to nije bilo jednostavno isčitati.

Kako bismo mogli govoriti o repnoj zavisnosti slučajnih vektora, morali smo definirati višedimenzionalnu generalizaciju regularne varijacije i iskazati neka njezina svojstva. Zatim smo definirali tri mjere repne zavisnosti - kutnu mjeru (s kojom je teško raditi na konačnom uzorku jer se procjenjuje mjera na jediničnoj sferi, beskonačnodimenzionalan objekt), te koeficijente χ i $\bar{\chi}$ koji imaju slične probleme izbora praga kao Hillov procjenitelj. Njihove su procjene prilično dobro funkcionirale na generiranim podacima, te su dali procjenu vrijednosti χ blizu nule za svaki par dionica, a $\bar{\chi}$ za različite parove između 0 i 0.3.

Summary

The aim of this thesis was to study estimators of the tail index and measures of dependence for heavy-tailed random variables and to test those methods on some real life data, but also to discuss some problems in this context. We introduced the concept of regular variation from mathematical analysis as well as weak and vague convergence from measure theory and stated some of their properties.

We defined Hill's estimator for the tail index of a random variable whose tail distribution is regularly varying and proved its consistency. We used randomly generated data from Pareto, normal and Cauchy distributions to illustrate performance of Hill's estimator and then applied it to tech company stocks daily dollar volume data to see how it works in practice. We encountered the same problems literature warns us about - it may not be simple to determine whether such a heavy tail model is appropriate, and even once we decided to use it, it's not easy to select appropriate threshold in the corresponding Hill plot. For some stocks Hill plot clearly suggests value of α in the interval between 3 and 3.5, for others it is very hard conclude anything.

We introduced multivariate regular variation in order to measure tail dependence between components of random vectors. We defined three measures of tail dependence - angular measure (which is hard to use on a finite sample because, as a measure on unit sphere, it is very difficult to estimate), and coefficients χ and $\bar{\chi}$ which have problems with choosing threshold similar to the Hill estimator. Standard estimators of the quantities worked quite well on simulated data from bivariate logistic extreme value distribution, and when applied to the stocks, gave us estimates for χ very close to 0 for each pair of stocks, and $\bar{\chi}$ varying between 0 and 0.3.

Životopis

Rođen sam 14. kolovoza 1992. u Zagrebu. Završio sam Osnovnu školu Ksavera Šandora Gjalskoga i Petu gimnaziju u Zagrebu. Tijekom srednjoškolskog obrazovanja sudjelovao sam na državnim natjecanjima iz matematike, fizike, logike, programiranja i osnova informatike, te osvojio pohvalu i brončanu medalju na Međunarodnoj matematičkoj olimpijadi. Zatim sam 2011. upisao preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu, a 2014. godine diplomski studij Matematička statistika. Tijekom studija, nagrađen sam trećom i prvom nagradom na među-narodnom studentskom natjecanju IMC 2012. i 2013. godine te Priznanjem za izniman uspjeh u studiju 2015. i 2016. godine.