TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Ants Kaasik

Ekstremaalväärtuste lävendimudelid

Bakalaureusetöö

Juhendaja:

prof. Kalev Pärna

TARTU

2005

Sisukord

Sissejuhatus

Käesolev töö on jätkuks autori klassikalise ekstremaalväärtuste teooria mõningaid aspekte tutvustavale semestritööle [4]. Sedakorda on vaatluse all ekstremaalväärtuste teooria teine oluline haru klassikalise suuna kõrval – lävendimudelid (*threshold models*), mille kasutamise pioneerideks olid hüdroloogid. Lävendimudelite teooria valdkonnas kesksel kohal oleva üldistatud Pareto jaotuse formuleeris 1975. aastal James Pickands [5].

Bakalureusetöös selgitab autor täpsemalt motivatsiooni lävendimudelite kasutuselevõtuks. Vaadeldakse mitmeid valdkonna olulisemaid tulemusi, mida autor omalt poolt näidetega täiendab. Töö teises pooles käsitletakse lävendimudeli kasutamist reaalsete vaatlusandmete korral. Autori täiendavaks panuseks lisaks erinevatest allikatest pärit faktide kokkuviimisele ja mõningate tulemuste tõestamisele on erinevate programmide kirjutamine, mis võimaldavad mudelite kasutamist.

Töö kirjutamiseks on kasutatud tekstitöötlusprogrammi MiKTeX. Programmid on koostatud statistikapaketi R abil.

Kasutatud allikatele on töös viidatud nurksulgude abil. Esimene pool näitab allika numbrit töö lõpus asuvas kirjanduse loetelus ja teine pool lehekülge või lehekülgi, kus viidatud faktist juttu on.

Autor tänab väärtpaberituru ettevõtteid Tallinna Börs ja Eesti Väärtpaberikeskus töös kasutatud Hansapanga aktsia hindu puudutavate andmete kasutamise loa ja professor Kalev Pärnat arvukate paranduste ja täienduste ning mitmete töö struktuuri puudutavate ideede eest.

1. Lävendimudelite kasutamise otstarve

Enne lävendimudelite teooriasse süvenemist on paslik selgitada valdkonna seost klassikalise ekstremaalväärtuste teooriaga. Klassikalist ekstremaalväärtuste teooria lähenemist (*classical extreme value approach*) tähistataksegi sageli lühendi EVT (*extreme value theory*) abil, samas kui lävendimudelitel põhinev suund on tuntud kui POT lähenemine (*peaks-over-threshold approach*).

1.1. Sarnasused klassikalise ekstremaalväärtuste teooriaga

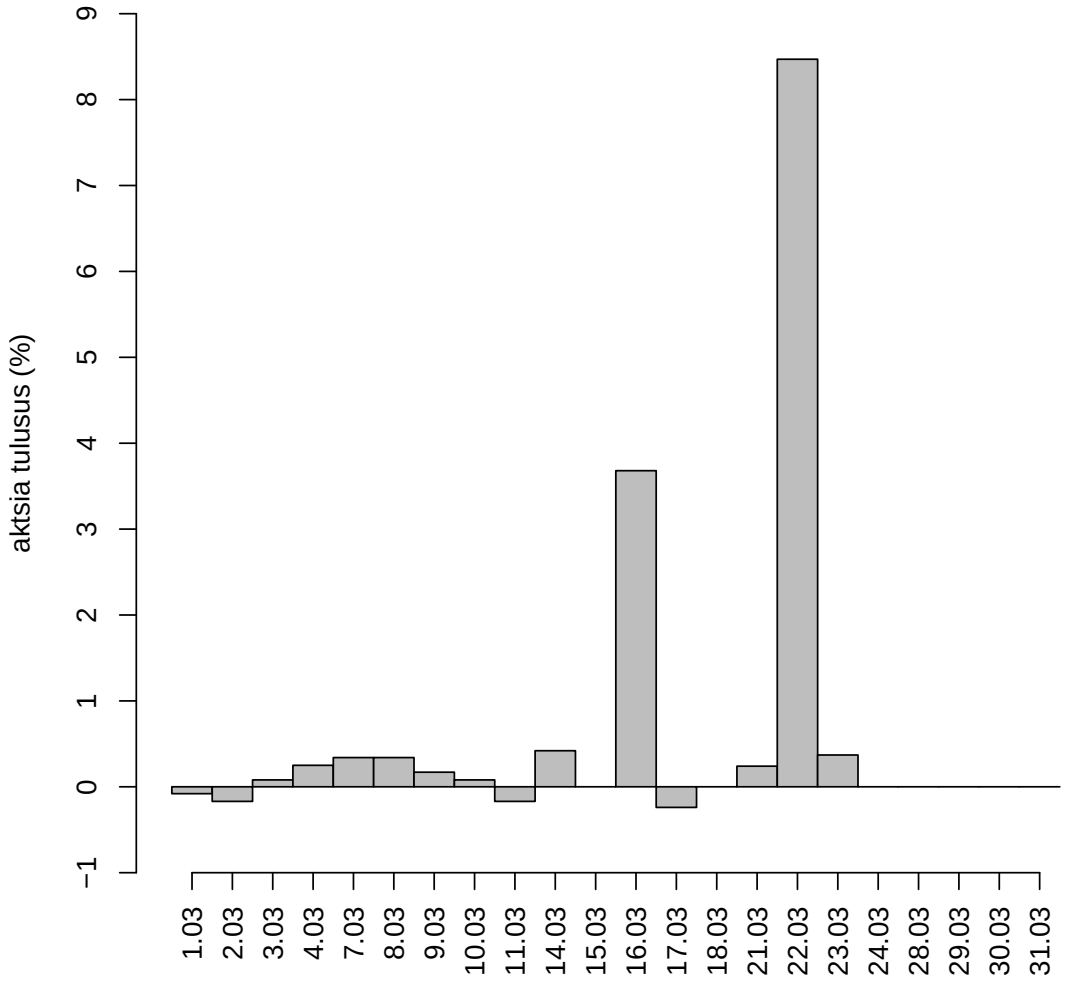
Ekstremaalväärtuste teooria rakendamisel on sõltumata valdkonnast enamasti üks eesmärk: hinnata riski. Lisaks samale põhieesmärgile on mõlema lähenemise puhul kesksel kohal üks jaotuste klass. Kui klassikalise ekstremaalväärtuste teooria puhul on selleks üldistatud ekstremaalväärtuste jaotus (*generalized extreme value distribution* ehk GEVD) siis lävendimudelite korral omab keskset rolli üldistatud Pareto jaotus (*generalized Pareto distribution* ehk GPD). Kehtima jääb ka põhimõte, et kui huvi pakuvad minimaalsed väärtused, siis tuleb lähtuda vastandmärgiga vaatlustest ning seejärel rakendada neile maksimaalsetele väärtustele mõeldud mudelit. Üleminek leidis täpsemat käsitlemist semestritöös [4: 7].

1.2. Erinevused klassikalisest ekstremaalväärtuste teooriast

Klassikalisest ekstremaalväärtuste teooria alusest – Fisher-Tippeti teoreemist – vahetult tulenev mudel võimaldab modelleerida suure valimi blokimaksimumi või -miinimumi. Semestritöös läbi viidud simuleerimiskatsed näitasid, et asümptootika hakkab hästi tööle juba siis, kui blokis on orienteeruvalt 20 vaatlust [4: 15-21]. Ometi tähendab see, et suur osa andmetest võib jääda kasutamata. Lävendimudelite kontseptsioon seisneb nivoo fikseerimises ning ekstremaalseteks loetakse kõik vaatlused, mis antud taset ületavad (kui tegu on minimaalseid väärtusi modelleeriva juhuga, siis vaatlused, mis on allpool fikseeritud nivood). Selline lähenemine võimaldab sageli andmeid efektiivsemalt kasutada – blokiekstreemumit modelleeriva mudeli korral heidetakse kõrvale kõik teised blokki kuuluvad vaatlused peale ekstremaalse, olgugi et ka variatsioonreas blokimaksimumi või -miinimumi kõrval paiknev vaatlus võib olla suurem (või väiksem kui uuritakse minimaalseid väärtusi) kõikide teiste blokkide ekstremaalsetest väärtustest. Selle väite illustreerimiseks vaatleme järgmist näidet.

Näide 1.1 Joonisel 1.1 on näha, et 2005. aasta märtsikuus oli börsil kaks kauplemispäeva, mil Hansapanga aktsia hind tugevalt tõusis. Suuri hinnalanguseid ei esinenud. Kui huvi pakuksid ekstremaalsed aktsiahinna kasvud, siis POT-mudeli korral loetaks ekstremaalseks nii 16. kui ka 22. märtsi aktsia sulgemishinna muutus, samas kui EVT-mudelis kasutataks vaid 22. märtsi vaatlust kui bloki suurimat elementi.

□



Joonis 1.1 *Hansapanga aktsia sulgemishinna muutused märtsis 2005. Väikeaktsionäridele tehtud ülevõtupakkumine tingis aktsia hinna muutumatuse kuu lõpus.*

EVT mudeli abil saame vastuse küsimusele: kui tõenäone on, et antud ajavahemiku jooksul ei esine mingist teatud nivoost suuremat (väiksemat) vaatlust. POT mudel aga võimaldab leida lahenduse probleemile: kui kriitiline tase ületatakse (ja seda võib bloki suurusele

vastava ajavahemiku jooksul kindlasti toimuda mitu korda), siis milline on sellise lävendit ületava väärtuse keskmine, aga ka kõikidele teistele nivood ületavate väärtuste jaotusega seonduvatele küsimustele.

Kui veepiirile rajatud kaitsetammi ületab aasta kõrgeim veetase kaduvväikese tõenäosusega, siis ei ole eriti oluline, kui kõrge see üleujutust põhjustav veetase keskmiselt on. Samas ei ole näiteks investoril eriti huvitav teada, et on vaid üks võimalus sajast, et tema portfelli kuuluv aktsia kaotab kuu suurimal languspäeval enam kui 3% oma väärtusest. Hoopis huvipakkuvam oleks teada, kas päeval, mil aktsia kaotab enam kui 2% oma väärtusest, on keskmiseks languseks 3% või hoopis 5%. Kokkuvõtvalt võibki märkida, et vajadus lävendimudeli järele on tingitud asjaolust, et sageli on mingi protsessi vaatlemisel oluline hinnata iga ekstremaalse väärtuse jaotust, mitte piirduda vaid teatud ajaperioodi suurima või vähima väärtuse jaotusega.

2. Lävendit ületavate valimi väärtuste jaotus

Kogu peatükis vaatleme tõenäosusruumi $(\Omega,\mathbb{F},P)$, kus $X_1, X_2, \ldots$ on sõltumatud sama jaotusega mittekonstantsed juhuslikud suurused jaotusfunktsiooniga $F(x) = P\{X_i \leqslant x\}$. Jaotusfunktsiooni täiendfunktsiooni tähistab $\overline{F}(x) = 1 - F(x)$.

2.1. Lävendit ületavate teadaoleva jaotusega vaatluste jaotus

Tundub mõistlik lugeda ekstremaalseteks need väärtused X_i, mis ületavad mingit kõrget nivood u. Meile pakub huvi nende ekstremaalsete väärtuste jaotus, mida aga kirjeldab $\forall x > 0$ korral järgmine tinglik tõenäosus

$$P\{X_i > u+x|\ X_i > u\} = \frac{P\{X_i > u+x, X_i > u\}}{P\{X_i > u\}} = \frac{P\{X_i > u+x\}}{P\{X_i > u\}} = \frac{\overline{F}(u+x)}{\overline{F}(u)}, \quad (2.1)$$

mis oma olemuselt on lävendit ületavate valimi elementide, mis on kohandatud nivoo u lahutamise abil, jaotusfunktsiooni täiendfunktsioon. Teades kõigi valimi elementide jaotusfunktsiooni F, oleks vahetult leitav ka lävendit ületavate väärtuste jaotus.

Näide 2.1 Semestritöös veendusime, et samade parameetritega Pareto jaotusest pärinevate sõltumatute valimielementide sobivalt normeeritud valimimaksimumi piirjaotuseks on Frechet' ekstremaalväärtuste jaotus [4: 16]. Vaatleme nüüd lävendit u ületavate sõltumatute valimi elementide X_i jaotust. Pareto jaotusega juhusliku suuruse jaotusfunktsiooni täiendfunktsioon on kujul $\overline{F}(x) = (x/k)^{-a}$, kus $k > 0$ ja $a > 0$ on jaotuse parameetrid ning $x \geqslant k$. Vastavalt võrdusele (2.1) saame, et lävendit ületavate vaatluste X_i, mida on vähendatud nivoo väärtuse võrra, jaotusfunktsiooni täiendfunktsioon avaldub

$$P\{X_i - u > x|\ X_i > u\} = \frac{[(u+x)/k]^{-a}}{[u/k]^{-a}} = \left(1 + \frac{x}{u}\right)^{-a}, \quad (2.2)$$

kus $u \geqslant k$ ja $x > 0$.

□

Näide 2.2 Semestritöös näitasime, et samade parameetritega eksponentjaotusest pärinevate sõltumatute valimielementide sobivalt normeeritud valimimaksimumi piirjaotuseks on Gumbeli ekstremaalväärtuste jaotus [4: 18-19]. Arvestades, et $\overline{F}(x) = e^{-\lambda \cdot x}$, kus $\lambda > 0$

ja $x \geqslant 0$, siis saame (2.1) abil, et $u \geqslant 0$ korral kehtib

$$P\{X_i - u > x |\ X_i > u\} = \frac{e^{-\lambda \cdot (u+x)}}{e^{-\lambda \cdot u}} = e^{-\lambda \cdot x}, \tag{2.3}$$

kus $x > 0$. Näeme, et lävendit u ületavate väärtuste jaotus on sama, mis esialgne jaotus. Seetõttu nimetatakse eksponentjaotust mäluta jaotuseks.

□

Näide 2.3 Semestritöös nägime, et samade parameetritega ühtlasest jaotusest pärinevate sõltumatute valimielementide sobivalt normeeritud valimimaksimumi piirjaotuseks on Weibulli ekstremaalväärtuste jaotus [4: 20]. Lõigul $[k, l]$ antud ühtlase jaotuse korral on $\overline{F}(x) = (l - x)/(l - k)$, kus $k \leqslant x \leqslant l$, seetõttu saame seost (2.1) kasutades

$$P\{X_i - u > x |\ X_i > u\} = \frac{l - (u + x)}{l - k} \cdot \frac{l - k}{l - u} = 1 + \frac{x}{u - l}\ , \tag{2.4}$$

kus $k \leqslant u < l$ ja $0 < x < l - u$.

□

Näide 2.4 Semestritöös on ära toodud tarvilik ja piisav tingimus normeeritud valimimaksimumi piirjaotuse leidumiseks. Osutus, et see tingimus ei ole täidetud Poissoni jaotusest valimi korral [4: 13-14]. Samas ei ole mingit takistust leidmaks lävendit u ületavate väärtuste jaotust. Kuivõrd $P\{X_i = k\} = e^{-\lambda} \cdot \lambda^k / k!$, kus $\lambda > 0$ ja $k \in \{0, 1, 2, \dots\}$ siis saame seost (2.1) kasutades, et iga mittenegatiivse täisarvulise u ja naturaalarvulise x korral kehtib

$$\begin{aligned} P\{X_i - u > x |\ X_i > u\} &= \frac{P\{X_i \geqslant u + x + 1\}}{P\{X_i \geqslant u + 1\}} = \frac{\sum_{r=u+x+1}^{\infty} e^{-\lambda} \cdot \lambda^r / r!}{\sum_{r=u+1}^{\infty} e^{-\lambda} \cdot \lambda^r / r!} = \\ &= \frac{\sum_{r=u+x+1}^{\infty} \lambda^{r-u} / r!}{\sum_{r=u+1}^{\infty} \lambda^{r-u} / r!} = 1 - \frac{\sum_{r=u+1}^{u+x} \lambda^{r-u} / r!}{\sum_{r=u+1}^{\infty} \lambda^{r-u} / r!}\ . \end{aligned} \tag{2.5}$$

□

Avaldise (2.5) kasutamine on küll arvutuslikult mõneti keerukam seostest (2.2), (2.3) ja (2.4), ent ometi on Poissoni jaotusest pärit valimi korral võimalik leida lävendit ületavate väärtuste jaotusfunktsioon. Seega ei ole tarvilik jaotuse F kuulumine Frechet', Gumbeli või Weibulli tõmbepiirkonda selleks, et oleks leitav antud jaotusest F pärit ja lävendit u ületavate väärtuste jaotus. Järgnevalt näeme aga, et jaotuse kuulumine mingi ekstremaalväärtuste jaotuse tõmbepiirkonda pakub mõningaid lisavõimalusi.

2.2. Üldistatud Pareto jaotus

Toome nüüd formaalselt sisse lävendimudelite teoorias kesksel kohal oleva jaotuse.

Definitsioon 2.1 Juhusliku suuruse X jaotust nimetatakse üldistatud Pareto jaotuseks kui X jaotusfunktsioonil on kuju

$$G(x) = \begin{cases} 1 - \left(1 + \dfrac{\varepsilon \cdot x}{\tilde{\sigma}}\right)^{-1/\varepsilon}, & \varepsilon \neq 0, \\ 1 - \exp\left(-\dfrac{x}{\tilde{\sigma}}\right), & \varepsilon = 0, \end{cases}$$

kus $x > 0$, $\tilde{\sigma} > 0$ ja $1 + \varepsilon \cdot x/\tilde{\sigma} > 0$.

□

Leides jaotusfunktsiooni tuletise, saame üldistatud Pareto jaotuse tihedusfunktsiooni:

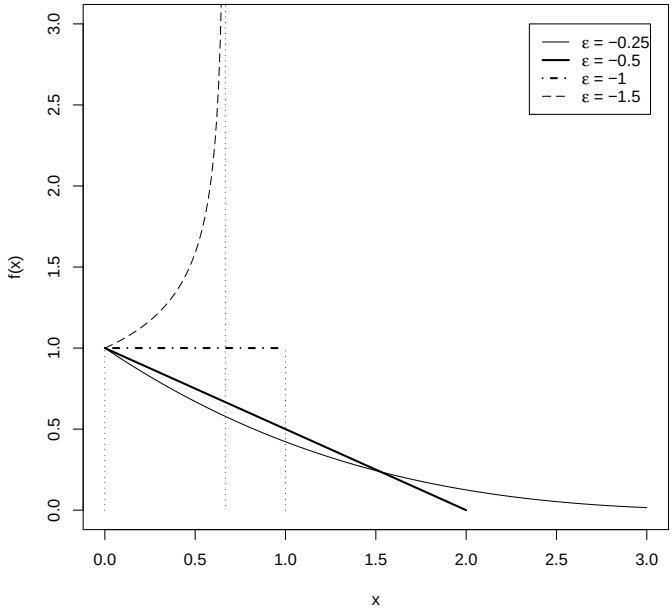
$$g(x) = \begin{cases} \dfrac{1}{\tilde{\sigma}}\left(1 + \dfrac{\varepsilon \cdot x}{\tilde{\sigma}}\right)^{-(1+\varepsilon)/\varepsilon}, & \varepsilon < 0,\ 0 < x < -\tilde{\sigma}/\varepsilon, \\ \dfrac{1}{\tilde{\sigma}}\left(1 + \dfrac{\varepsilon \cdot x}{\tilde{\sigma}}\right)^{-(1+\varepsilon)/\varepsilon}, & \varepsilon > 0,\ x > 0, \\ \dfrac{1}{\tilde{\sigma}} \cdot \exp\left(-\dfrac{x}{\tilde{\sigma}}\right), & \varepsilon = 0,\ x > 0, \\ 0, & \text{mujal.} \end{cases}$$

Joonistel 2.1 ja 2.2 on graafiliselt kujutatud üldistatud Pareto jaotuse tihedusfunktsioon parameetrite erinevate väärtuste korral. Analoogiliselt üldistatud ekstremaalväärtuste jaotusega on üldistatud Pareto jaotus tõkestatud negatiivse ε korral. On näha, et parameetri ε kasvades muutub jaotuse parempoolne saba järjest raskemaks. Järgnevalt leiame üldistatud Pareto jaotuse keskväärtuse ja veendume, et kui $\varepsilon \geqslant 1$, siis on jaotuse parempoolne saba sedavõrd raske, et keskväärtus ei ole enam lõplik.

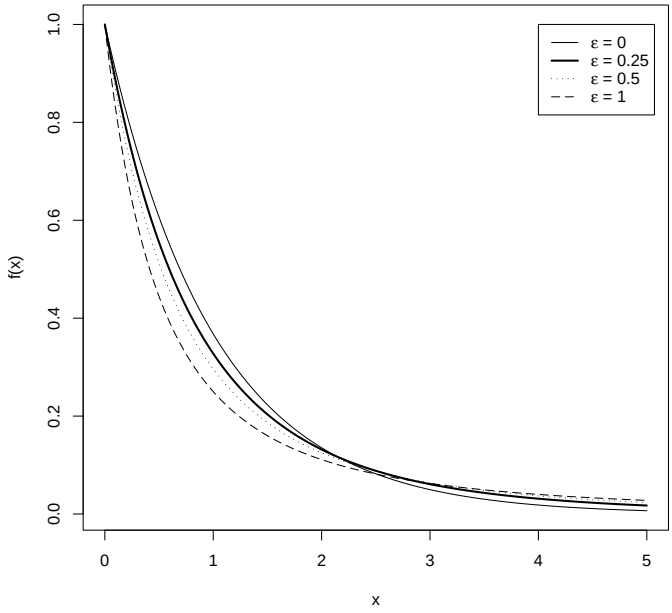
Vaatleme esmalt olukorda, kus $\varepsilon < 0$. Sellisel juhul saame tihedusfunktsioonist lähtudes ja ositi integreerimise valemi ning seejärel muutuja vahetuse abil

$$\mathbb{E}X = \int_0^{-\tilde{\sigma}/\varepsilon} \underbrace{\frac{x}{\tilde{\sigma}}}_{u} \underbrace{\left(1 + \frac{\varepsilon \cdot x}{\tilde{\sigma}}\right)^{-(1+\varepsilon)/\varepsilon} dx}_{dv} = \left[-x\left(1 + \frac{\varepsilon \cdot x}{\tilde{\sigma}}\right)^{-1/\varepsilon}\right]\Bigg|_0^{-\tilde{\sigma}/\varepsilon} + \int_0^{-\tilde{\sigma}/\varepsilon} \left(\underbrace{1 + \frac{\varepsilon \cdot x}{\tilde{\sigma}}}_{y}\right)^{-1/\varepsilon} dx =$$

$$= 0 + 0 + \frac{\tilde{\sigma}}{\varepsilon} \int_1^0 y^{-1/\varepsilon} dy = \frac{\tilde{\sigma}}{\varepsilon - 1}\left(y^{(\varepsilon-1)/\varepsilon}\right)\Big|_1^0 = 0 - \frac{\tilde{\sigma}}{\varepsilon - 1} = \frac{\tilde{\sigma}}{1 - \varepsilon}\ .$$



Joonis 2.1 *Üldistatud Pareto jaotuse tihedusfunktsioon erinevate* $\varepsilon < 0$ *väärtuste korral kui* $\tilde{\sigma} = 1$.



Joonis 2.2 *Üldistatud Pareto jaotuse tihedusfunktsioon erinevate* $\varepsilon \geqslant 0$ *väärtuste korral kui* $\tilde{\sigma} = 1$.

Kui $\varepsilon = 0$, siis tunneme ära eksponentjaotuse tiheduse parameetriga $1/\tilde{\sigma}$. Sellise jaotuse keskväärtus on aga $\tilde{\sigma}$.

Vaatleme nüüd situatsiooni, kus $\varepsilon > 0$. Muutuja vahetuse abil saame

$$\mathbb{E}X = \int_0^\infty \frac{x}{\tilde{\sigma}} \Bigg(\underbrace{1 + \frac{\varepsilon \cdot x}{\tilde{\sigma}}}_{y} \Bigg)^{-(1+\varepsilon)/\varepsilon} dx = \frac{\tilde{\sigma}}{\varepsilon^2} \int_1^\infty (y-1) \cdot y^{-(1+\varepsilon)/\varepsilon} dy = \frac{\tilde{\sigma}}{\varepsilon^2} \int_1^\infty y^{-1/\varepsilon} +$$

$$+ \frac{\tilde{\sigma}}{\varepsilon} \cdot y^{-1/\varepsilon} \Big|_1^\infty = \frac{\tilde{\sigma}}{\varepsilon^2} \int_1^\infty y^{-1/\varepsilon} - \frac{\tilde{\sigma}}{\varepsilon} \, .$$

Kui $0 < \varepsilon < 1$, siis on integraali väärtuseks $\varepsilon/1-\varepsilon$ ja keskväärtuseks seega $\tilde{\sigma}/(1-\varepsilon)$. Kui $\varepsilon \geqslant 1$, siis on integraali väärtus lõpmata suur, seega on lõpmatu ka keskväärtus. Kokkuvõttes oleme tõestanud lemma:

Lemma 2.1 *Üldistatud Pareto jaotuse, mille parameetrid on $\tilde{\sigma}$ ja ε, keskväärtus avaldub*

$$\mathbb{E}X = \begin{cases} \tilde{\sigma}/(1-\varepsilon), & \varepsilon < 1, \\ \infty, & \varepsilon \geqslant 1. \end{cases}$$

□

2.3. Lävendit ületavate valimi väärtuste piirjaotus

Osutub, et kui jaotus F kuulub ühe ekstremaalväärtuste jaotuse tõmbepiirkonda (st normeeritud valimimaksimumi piirjaotuseks on üldistatud ekstremaalväärtuste jaotus) ja piirjaotuse parameeter on teada, siis on kergesti leitav ka lävendit u ületavate valimiväärtuste ligikaudne jaotus. Nimelt kehtib

Teoreem 2.1 *Olgu $\{X_n\}$ sõltumatute sama jaotusega F juhuslike suuruste jada ja $M_n = \max\{X_1, \ldots, X_n\}$. Leidugu normeerimiskonstandid $c_n > 0$, $d_n \in \Re$ ja mittekonstantne juhuslik suurus Y nii, et $(M_n - d_n)/c_n \xrightarrow{D} Y$ ehk*

$$\lim_{n\to\infty} P\{\frac{M_n - d_n}{c_n} \leqslant y\} = H(y).$$

Kui juhusliku suuruse Y jaotusfunktsioon on kujul

$$H(y) = \exp\left[-(1+\varepsilon \cdot y)^{-1/\varepsilon}\right],$$

kus $\varepsilon \neq 0$ on piirjaotuse parameeter, siis kehtib

$$\lim_{n\to\infty} P\{X - (c_n \cdot y + d_n) > c_n \cdot x | X > c_n \cdot y + d_n\} = \left(1 + \frac{\varepsilon \cdot x}{\tilde{\sigma}}\right)^{-1/\varepsilon}, \qquad (2.6)$$

kus y on selline, et $H(y) > 0$, $x > 0$ ja $1 + \varepsilon \cdot x/\tilde{\sigma} > 0$ ning $\tilde{\sigma} = 1 + \varepsilon \cdot y$. Kui juhusliku suuruse Y jaotusfunktsioon on kujul

$$H(y) = \exp[-\exp(-y)],$$

siis kehtib

$$\lim_{n\to\infty} P\{X - (c_n \cdot y + d_n) > c_n \cdot x | X > c_n \cdot y + d_n\} = \exp(-x), \tag{2.7}$$

kus $x > 0$.

Tõestus. Vaatleme alguses olukorda, kus $H(y) = \exp\left[-(1+\varepsilon \cdot y)^{-1/\varepsilon}\right]$. On selge, et tänu sõltumatusele kehtib

$$P\{M_n \leqslant y\} = P\{X_1 \leqslant y, \ldots, X_n \leqslant y\} = F^n(y),$$

seega teoreemi eelduste kehtides kehtib ka

$$\lim_{n\to\infty} F^n(c_n \cdot y + d_n) = \exp\left[-(1+\varepsilon \cdot y)^{-1/\varepsilon}\right]. \tag{2.8}$$

Saadud seosest järeldub, et kui $H(y) > 0$, siis $F(c_n \cdot y + d_n) \xrightarrow{n\to\infty} 1$, sest kui viimane piirväärtus oleks ühest väiksem, siis oleks piirväärtus seoses (2.8) võrdne nulliga. Samuti ei ole võimalik, et jada $F(c_n \cdot y + d_n)$ ei koondu, sest siis piirväärtus (2.8) ei eksisteeriks või oleks võrdne nulliga. Seega peab kehtima $F(c_n \cdot y + d_n) \xrightarrow{n\to\infty} 1$. Logaritmime seose (2.8) mõlemaid pooli ja saame

$$\lim_{n\to\infty} n \cdot \ln F(c_n \cdot y + d_n) = -(1+\varepsilon \cdot y)^{-1/\varepsilon}. \tag{2.9}$$

Kirjutame välja Taylori valemi 2 esimest liiget funktsiooni $\ln y$ jaoks punktis $y = 1$ ja saame

$$\ln x = \ln 1 + (x-1) + r_n(x),$$

kus $r_n(x) = -(x-1)^2 \cdot \xi^{-2}/2$ ja $x < \xi < 1$. Seega on $x - 1$ ja $\ln x$ protsessis $x \to 1$ asümptootiliselt ekvivalentsed. Eeldame nüüd, et $H(y) > 0$. Seosest (2.8) tehtud järelduse tõttu saame ülatoodu põhjal anda seosele (2.9) kuju

$$\lim_{n\to\infty} n[F(c_n \cdot y + d_n) - 1] = -(1+\varepsilon \cdot y)^{-1/\varepsilon},$$

ehk siis

$$\lim_{n\to\infty} n \cdot \overline{F}(c_n \cdot y + d_n) = (1 + \varepsilon \cdot y)^{-1/\varepsilon}.$$

Analoogiliselt saame, et $x > 0$ korral

$$\lim_{n\to\infty} n \cdot \overline{F}[c_n(y + x) + d_n] = [1 + \varepsilon(y + x)]^{-1/\varepsilon}.$$

Seega kehtib

$$\lim_{n\to\infty} \frac{\overline{F}[c_n(y + x) + d_n]}{\overline{F}(c_n \cdot y + d_n)} = \frac{[1 + \varepsilon(y + x)]^{-1/\varepsilon}}{[1 + \varepsilon \cdot y]^{-1/\varepsilon}} = \left(1 + \frac{\varepsilon \cdot x}{1 + \varepsilon \cdot y}\right)^{-1/\varepsilon},$$

kus paremal pool tunneme ära üldistatud Pareto jaotuse jaotusfunktsiooni täiendfunktsiooni. Antud juhul $\tilde{\sigma} = 1 + \varepsilon \cdot y$. Seost (2.1) arvestades oleme saanud

$$\lim_{n\to\infty} P\{X - (c_n \cdot y + d_n) > c_n \cdot x | X > c_n \cdot y + d_n\} = \left(1 + \frac{\varepsilon \cdot x}{\tilde{\sigma}}\right)^{-1/\varepsilon}.$$

Kui $H(y) = \exp[-\exp(-y)]$ ja $x > 0$, siis saame analoogiliselt esimese osaga, et

$$\lim_{n\to\infty} \frac{\overline{F}[c_n(y + x) + d_n]}{\overline{F}(c_n \cdot y + d_n)} = \frac{\exp[-(x + y)]}{\exp(-y)} = \exp(-x).$$

Tunneme ära üldistatud Pareto jaotuse jaotusfunktsiooni täiendfunktsiooni, kus $\varepsilon = 0$ ja $\tilde{\sigma} = 1$. Seost (2.1) arvestades oleme tõestanud, et

$$\lim_{n\to\infty} P\{X - (c_n \cdot y + d_n) > c_n \cdot x | X > c_n \cdot y + d_n\} = \exp(-x).$$

□

Teoreemi alternatiivne piirkuju on ära toodud [3: 158-160].

Märkus. Eelduse kohaselt on teoreemis 2.1 normeerimiskonstandid valitud nii, et normeeritud valimimaksimum koondub üldistatud ekstremaalväärtuste jaotuseks, mille parameetrid $\mu = 0$ ja $\sigma = 1$. Taoline eeldus pole siiski kitsendav. Täpsemalt, olgu meil teada üks teine normeerimiskonstantide komplekt c_n ja d_n, mille kasutamise korral koondub valimimaksimum üldistatud ekstremaalväärtuste jaotuseks teistsuguste parameetritega μ ja σ. Siis on selge, et normeerimiskonstandid $c_n^* = \sigma \cdot c_n$ ja $d_n^* = d_n + \mu \cdot c_n$ rahuldavad juba teoreemi 2.1 eeldusi.

Märgime, et teoreemi 2.1 praktilisel rakendamisel on lävend u, mille rolli tõestatud teoree-

mis mängib $c_n \cdot y + d_n$, vaja valida piisavalt suur, et üldistatud Pareto jaotuse kasutamine lähendina oleks otstarbekas. Liiga madala lävendi korral on $F(u)$ ühest selgelt väiksem ja Taylori valemi jääkliige pole enam nullilähedane.

Näide 2.5 Semestritöös leidsime piirjaotuseks koondumise tagavad normeerimiskonstandid Pareto jaotusest pärit sõltumatute vaatluste valimimaksimumile: $c_n = k \cdot n^{1/a}$ ja $d_n = 0$ [4: 16]. Koondumine toimub üldistatud ekstremaalväärtuste jaotuseks parameetritega $\mu = 1$, $\sigma = 1/a$ ja $\varepsilon = 1/a$. Eelneva märkuse põhjal võime kasutada normeerimiskonstante $c_n^* = k \cdot n^{1/a}/a$ ja $d_n^* = k \cdot n^{1/a}$, mis tagavad koondumise üldistatud ekstremaalväärtuste jaotuseks parameetritega $\mu = 0$, $\sigma = 1$ ja $\varepsilon = 1/a$. Teoreemist 2.1 (seos (2.6)) saame n fikseerimisel

$$P\{X > (k \cdot n^{1/a}/a) \cdot y + k \cdot n^{1/a} + (k \cdot n^{1/a}/a) \cdot x | X > (k \cdot n^{1/a}/a) \cdot y + k \cdot n^{1/a}\} \approx$$
$$\approx \left(1 + \frac{x/a}{1 + y/a}\right)^{-a}$$

ehk

$$P\{X > u + z | X > u\} \approx \left(1 + \frac{z}{u}\right)^{-a}, \tag{2.10}$$

kus $u = (k \cdot n^{1/a}/a) \cdot y + k \cdot n^{1/a}$ ja $z = (k \cdot n^{1/a}/a) \cdot x$. Samal ajal teame näite 2.1 põhjal, et (2.10) kehtib täpse võrdusena. Seega kehtib Pareto jaotuse korral (2.6), sõltumata n väärtusest.

□

Näide 2.6 Semestritöös leidsime piirjaotuseks koondumise tagavad normeerimiskonstandid eksponentjaotusest pärit sõltumatute vaatluste valimimaksimumile: $c_n = 1/\lambda$ ja $d_n = \ln n/\lambda$ [4: 19]. Fikseerides n saame teoreemist 2.1 (seos (2.6))

$$P\{X > (1/\lambda) \cdot y + \ln n/\lambda + (1/\lambda) \cdot x | X > (1/\lambda) \cdot y + \ln n/\lambda\} \approx \exp(-x)$$

ehk

$$P\{X > u + z | X > u\} \approx \exp(-\lambda \cdot z), \tag{2.11}$$

kus $u = (1/\lambda) \cdot y + \ln n/\lambda$ ja $z = (1/\lambda) \cdot x$. Samas näite 2.2 põhjal on (2.11) täpne tulemus. Järelikult kehtib eksponentjaotuse korral (2.6), sõltumata n väärtusest.

□

Näide 2.7 Semestritöös leidsime piirjaotuseks koondumise tagavad normeerimiskonstandid eksponentjaotusest pärit sõltumatute vaatluste valimimaksimumile: $c_n = (l - k)/n$ ja

$d_n = l$ [4:20]. Koondumine toimub üldistatud ekstremaalväärtuste jaotuseks parameetritega $\mu = -1, \sigma = 1$ ja $\varepsilon = -1$. Eelneva märkus põhjal võime kasutada normeerimiskonstante $c_n^* = (l-k)/n$ ja $d_n^* = l - (l-k)/n$, mis tagavad koondumise üldistatud ekstremaalväärtuste jaotuseks parameetritega $\mu = 0$, $\sigma = 1$ ja $\varepsilon = -1$. Teoreemist 2.1 (seos (2.6)) saame n fikseerimisel

$$P\{X > [(l-k)/n] \cdot y + l - (l-k)/n + [(l-k)/n] \cdot x | X > [(l-k)/n] \cdot y + l - (l-k)/n\} \approx$$
$$\approx 1 + \frac{x}{y-1}$$

ehk

$$P\{X > u + z | X > u\} \approx 1 + \frac{z}{u-l}, \tag{2.12}$$

kus $u = [(l-k)/n] \cdot y + l - (l-k)/n$ ja $z = [(l-k)/n] \cdot x$. Teame näite 2.3 põhjal, et (2.12) on täpne tulemus. Seega kehtib ühtlase jaotuse korral (2.6), sõltumata n väärtusest.

□

Praktikas on teoreemi 2.1 tähtsus ka asjaolus, et teades normeeritud valimimaksimumi piirjaotuse parameetrit ε, on vahetult leitavad ka üldistatud Pareto jaotuse parameetrid lävendit ületavate väärtuste jaotuse jaoks. Seega ei ole blokimaksimumi modelleeriva jaotuse ja lävendi ületamist modelleeriva jaotuse jaoks vaja mitut korda parameetreid hinnata.

Võime liikuda ka vastupidises suunas. Olgu teada, et jaotus F kuulub mingi ekstremaalväärtuste jaotuse tõmbepiirkonda ja olgu meil teada antud jaotusest F pärit lävendit ületavate väärtuste jaotus. Siis on normeeritud valimimaksimumi piirjaotus samuti teada, sest selle parameetri ε väärtus on sama, mis vastaval lävendit ületavate väärtuste jaotusel.

Lemmas 2.1 saadud tulemuse põhjal võime veel mainida, et kui sõltumatute vaatlustega valim on võetud jaotusest, mis kuulub Frechet' ekstremaalväärtuste jaotuse tõmbepiirkonda, kusjuures piirjaotuse parameeter $\varepsilon \geqslant 1$, siis kõrget lävendit u ületavad väärtused, millest on lävendi väärtus maha lahutatud, on keskmiselt lõpmatult suured.

2.4. Lävendit ületavate väärtuste esinemissagedus

Teades lävendit ületavate väärtuste jaotust, tekib loomulik küsimus, kui tihti selliseid vaatluseid esineb. Iga X_i korral võime vaadelda Bernoulli jaotusega juhuslikku suurust $I_{\{X_i > u\}}$, mille keskväärtus väljendabki lävendit ületava väärtuse esinemistõenäosust. Se-

da suurust võib vaadelda kui üht täiendavat parameetrit p, mida on tarvis valimi põhjal hinnata. Lävendit ületavate vaatluste arv seerias pikkusega n on seega binoomjaotusega parameetritega n ja p. Poissoni piirteoreemi põhjal saame seda jaotust pika seeria korral lähendada Poissoni jaotusega parameetriga λ eeldusel, et $n \cdot p_n \to \lambda$.

Sõltumatute vaatluste seerias kirjeldab esimese ekstremaalse väärtuse tuleku tõenäosust geomeetriline jaotus parameetriga p.

Nii peame hindama kokku nelja suurust: sobiva lävendi u väärtust, valitud lävendi ületamise tõenäosust p ja lävendit ületavate väärtuste jaotuse parameetreid $\tilde{\sigma}$ ja ε.

3. Lävendimudeli kasutamine

Vaatleme jällegi tõenäosusruumi $(\Omega,\mathbb{F},P)$, kus $X_1,\ldots,X_n$ on sõltumatud sama jaotusega mittekonstantsed juhuslikud suurused jaotusfunktsiooniga $F(x)=P\{X_i\leqslant x\}$. Olgu meil teada, et antud andmete korral on võimalik mingit kõrget lävendit u ületavate väärtuste jaotust üldistatud Pareto jaotuse abil modelleerida. See teadmine üksi ei anna aga vastust küsimustele, mis puudutavad sobiva lävendi u leidmise põhimõtteid, mudeli parameetrite hindamist ja ka mudeli tõlgendamist.

3.1. Lävendi valik

Teoreem 2.1 ei ütle, kui suur peaks lävend u olema. On vaid teada, et üldistatud Pareto jaotus on modelleerimiseks kasutatav u suure väärtuse korral. Näidetes 2.5, 2.6 ja 2.7 aga nägime, et teoreem 2.1 annab meile täpse tulemuse, sõltumata lävendi valikust.

Näide 3.1 Vaatleme piirjaotuseks koondumise kiirust standardsest normaaljaotusest pärit ekstremaalväärtuste puhul. On teada, et normaaljaotusega juhuslikud suurused kuuluvad Gumbeli ekstremaalväärtuste jaotuse tõmbepiirkonda, seega peab teoreemi 2.1 kohaselt lävendit ületavate väärtuste jaotus olema GPD, mille parameeter $\varepsilon=0$. Rakendame l'Hospitali reeglit suhtele $x\cdot\overline{\Phi}(x)/\varphi(x)$ protsessis $x\to\infty$, kus $\overline{\Phi}(x)$ on standardse normaaljaotuse jaotusfunktsiooni täiendfunktsioon ja $\varphi(x)$ tihedusfunktsioon. Kasutades, et $\varphi'(x)=-x\cdot\varphi(x)$, saame

$$\lim_{x\to\infty}\frac{\overline{\Phi}(x)}{\varphi(x)/x}=\lim_{x\to\infty}\frac{-\varphi(x)}{\varphi'(x)/x-\varphi(x)/x^2}=\lim_{x\to\infty}\frac{x^2}{x^2+1}=1.$$

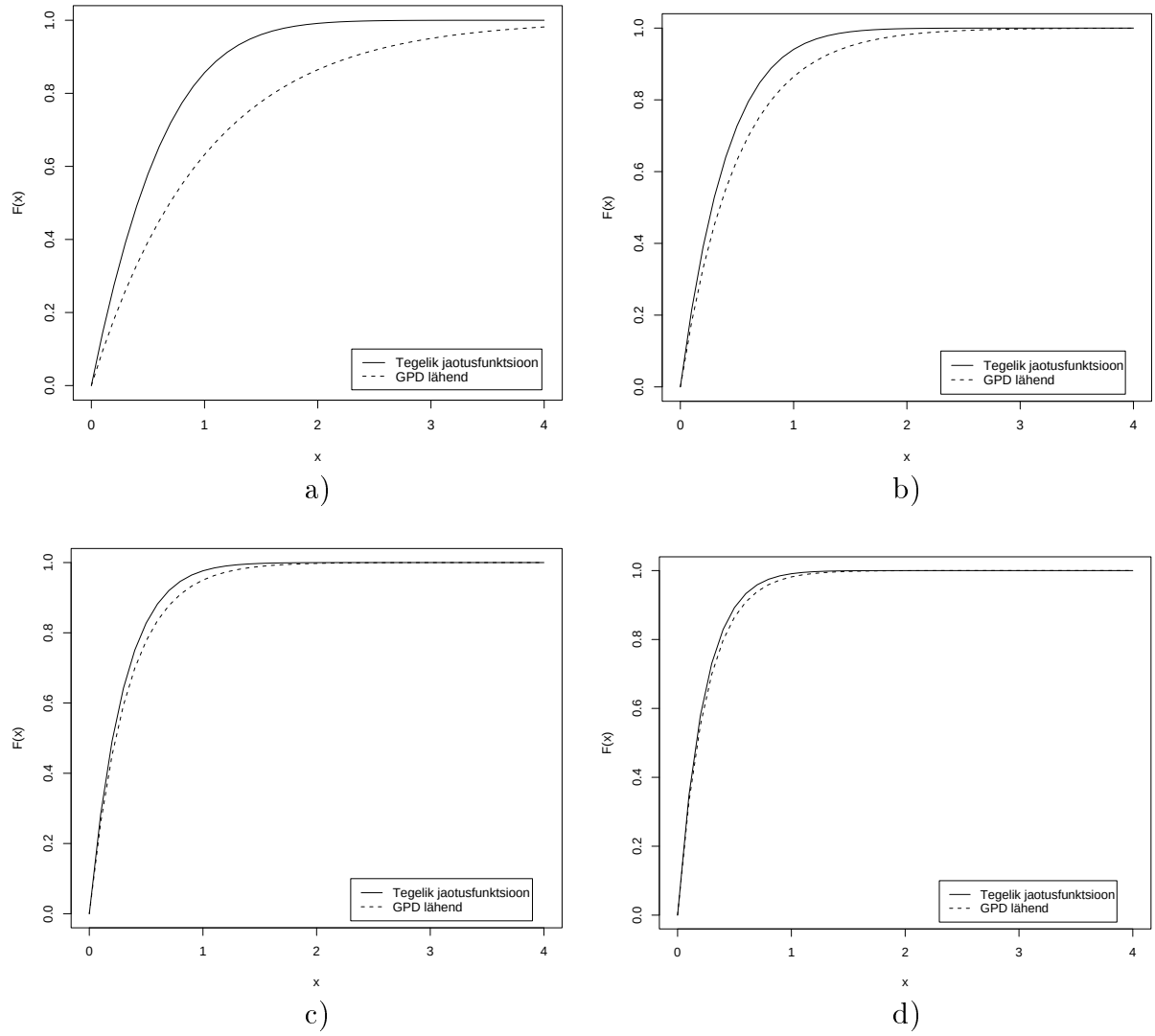
Seega võime kirjutada [6: 12]

$$\lim_{u\to\infty}\frac{\overline{\Phi}(u+z/u)}{\overline{\Phi}(u)}=\lim_{u\to\infty}\frac{\varphi(u+z/u)\cdot u}{\varphi(u)\cdot(u+z/u)}=\lim_{u\to\infty}\left\{\left(1+\frac{z}{u^2}\right)^{-1}\cdot\exp\left[\frac{-(u+z/u)^2+u^2}{2}\right]\right\}=$$
$$=\lim_{u\to\infty}\left[\left(1+\frac{z}{u^2}\right)^{-1}\cdot\exp\left(-z-\frac{z^2}{2u^2}\right)\right]=e^{-z}.$$

Kui võtame $x=z/u$, siis oleme saanud

$$\lim_{u\to\infty}\frac{\overline{\Phi}(u+x)}{\overline{\Phi}(u)}=e^{-u\cdot x},\tag{3.1}$$

milles tunneme ära GPD jaotusfunktsiooni täiendfunktsiooni, kus $\tilde{\sigma}=1/u$.



Joonis 3.1 *Üldistatud Pareto jaotuse lähendi koondumine normaaljaotusest pärit lävendit ületavate väärtuste jaotusfunktsiooniks erinevate lävendi väärtuste korral: a)* $u = 1$*, b)* $u = 2$*, c)* $u = 3$ *ja d)* $u = 4$*.*

Jooniselt 3.1 on näha, et koondumine on üpris aeglane – alles lävendi $u = 3$ korral saavutatakse rahuldav lähend, kus suhteline viga jääb alla 10% piiri. Kuivõrd aga $\overline{\Phi}(3) = 0.00134$ ehk keskmiselt 13 vaatlust 10 000 sõltumatust standardsest normaaljaotusest pärit valimielemendist ületab lävendit $u = 3$, siis on vaja hiiglaslikku valimimahtu, et GPD oleks lävendit ületavatele väärtustele sobitatav.

□

Üldjuhul ei ole meil andmed teadaolevast jaotusest ja siis ei ole analüütilised tulemused leitavad. On selge, et kui seame lävendi u väga kõrgele, siis ei ole eriti palju vaatlusi, mis lävendit ületaksid ja siis on tagajärjeks mudeli parameetrite väga laiad usalduspiirid. Teisalt, kui lävend u on liiga madal, siis ei ole teoreemi 2.1 eeldused täidetud ja me ei tohiks lävendit ületavate väärtuste jaotust üldistatud Pareto jaotusega modelleerida. Seega oleks vaja leida sedavõrd väike u väärtus, mis tagaks juba asümptootika normaalse töö.

On näidatud, et kui samast jaotusest pärit sõltumatute vaatluste normeerimata valimimaksimumi ligikaudne jaotus on üldistatud ekstremaalväärtuste jaotus parameetritega μ, σ ja ε, siis on lävendit u ületavate väärtuste ligikaudne jaotus üldistatud Pareto jaotus parameetritega $\tilde{\sigma} = \sigma + \varepsilon(u - \mu)$ ja ε [1: 76-77]. Seega muutub lävendi muutmisel ka $\tilde{\sigma}$ väärtus. Olgu $\tilde{\sigma}_u$ üldistatud Pareto jaotuse parameetri väärtus lävendi u korral. Eeldades, et $\varepsilon < 1$, teame lemma 2.1 põhjal, et

$$\mathbb{E}(X_i - u|\ X_i > u) \approx \frac{\tilde{\sigma}_u}{1-\varepsilon} = \frac{\sigma + \varepsilon(u-\mu)}{1-\varepsilon} = \underbrace{\frac{\sigma - \varepsilon \cdot \mu}{1-\varepsilon}}_{const} + u \cdot \frac{\varepsilon}{1-\varepsilon}. \tag{3.2}$$

Saadud võrduse parem pool on u lineaarne funktsioon, seose vasak pool aga lävendit u ületavate väärtuste, millest on u maha lahutatud, keskväärtus. Seega on üldistatud Pareto jaotus lähendina kasutatav alates lävendi väärtusest u_0, kui lävendit ületavate väärtuste keskmise graafik on ligikaudu sirge $u > u_0$ korral. Praktikas kasutatakse mõistagi lävendit ületavate väärtuste valimikeskmist. Graafikul esitatakse punktid

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < \max\{x_1, \ldots, x_n\} \right\},$$

kus n_u on lävendit u ületavate valimi väärtuste arv ja $x_{(i)}$ on i-s lävendit u ületav valimi vaatlus. Tekkivat graafikut nimetatakse keskmise jääkeluea graafikuks (*mean residual life plot*). Selline nimetus tuleneb lävendimudelite kasutamisest süsteemide vastupidavuse modelleerimisel, kus suurus $\mathbb{E}(X_i - u|\ X_i > u)$ kirjeldab süsteemi i keskmist oodatavat tööiga pärast u ajaühikut kestnud süsteemi kasutamist.

Väärtus u, alates millest graafik on ligikaudu sirge, võetakse lävendina kasutusele, kusjuures paremaks tõlgendamiseks kantakse graafikule ka vastava valimikeskmise usalduspiirid (vt. joonis 4.2).

Usalduspiiridega keskmise jääkeluea graafikut võimaldab joonistada statistikapaketis R

realiseeritud programm (vt lisa A), millele tuleb sisenditena ette anda andmevektor ja olulisusnivoo.

Sageli on graafiku lineaarseks muutumise punkti hindamine siiski piisavalt raske ja subjektiivne ülesanne. Küll aga võime ε väärtust (või ε hinnangut) teades saada jääkeluea graafiku sirge osa tõusu hinnangu. Selleks leiame võrduses (3.2) esineva u kordaja väärtuse, ning see peab ligikaudu ühtima keskmise jääkeluea graafikul tekkiva sirge tõusunurga tangensiga. Kui on teada ε usalduspiirid, siis saame leida kaks tõusu, mille vahele peaks jääma keskmise jääkeluea graafiku sirge osa tõus. Paraku võib ka selline "tõusukoridor" osutuda väga laiaks ja seega jääb sobiva lävendi valik ikkagi mõneti subjektiivseks.

3.2. Parameetrite hindamine

Pärast lävendi fikseerimist saame hinnata üldistatud Pareto jaotuse parameetrid suurima tõepära meetodil. Olgu meil lävendit ületavad vaatlused $y_1, \ldots, y_{n_u}$, millest on lävendi väärtus maha lahutatud ($y_i = x_{(i)} - u$). Vaadeldes olukorda, kus $\varepsilon \neq 0$, saame, et logaritmilisel tõepärafunktsioonil on kuju

$$\begin{aligned} \ell(\tilde{\sigma}, \varepsilon) = &\ln \prod_{i=1}^{n_u} \left[\frac{1}{\tilde{\sigma}} \left(1 + \frac{\varepsilon \cdot y_i}{\tilde{\sigma}} \right)^{-(1+\varepsilon)/\varepsilon} \right] = \ln \frac{1}{\tilde{\sigma}^{n_u}} + \sum_{i=1}^{n_u} \ln \left(1 + \frac{\varepsilon \cdot y_i}{\tilde{\sigma}} \right)^{-(1+\varepsilon)/\varepsilon} = \\ = &- n_u \cdot \ln \tilde{\sigma} - \left(\frac{1+\varepsilon}{\varepsilon} \right) \sum_{i=1}^{n_u} \ln \left(1 + \frac{\varepsilon \cdot y_i}{\tilde{\sigma}} \right), \end{aligned} \tag{3.3}$$

kusjuures iga $i = 1, \ldots, n_u$ korral peab olema täidetud $1 + \varepsilon \cdot y_i / \tilde{\sigma} > 0$. Juhul kui $\varepsilon = 0$ on logaritmiline tõepärafunktsioon kujul

$$\ell(\tilde{\sigma}) = \ln \prod_{i=1}^{n_u} \left[\frac{1}{\tilde{\sigma}} \cdot \exp\left(- \frac{y_i}{\tilde{\sigma}} \right) \right] = \ln \frac{1}{\tilde{\sigma}^{n_u}} + \sum_{i=1}^{n_u} \ln \left[\exp\left(- \frac{y_i}{\tilde{\sigma}} \right) \right] = -n_u \cdot \ln \tilde{\sigma} - \frac{1}{\tilde{\sigma}} \sum_{i=1}^{n_u} y_i.$$

Pärast logaritmilise tõepärafunktsiooni maksimeerimist ja seeläbi suurima tõepära hinnangute leidmist oleks järgmine loomulik samm leida parameetrite vahemikhinnagud. Suurima tõepära hinnangud on asümptootiliselt normaalsed ainult teatavate regulaarsuse tingimuste kehtides, mille täidetust siinkohal eeldame.

Vaatluste jaotuse regulaarsustingimuste täidetuse korral kehtib järgmine tulemus [1: 31-32]

Teoreem 3.1 *Olgu $X_1, ..., X_n$ sõltumatud vaatlused samast d-mõõtmelise vektorparameet-*

riga jaotusest. Olgu parameetri tegelik väärtus θ_0 ja $\hat{\theta}_0$ vastav suurima tõepära hinnang. Siis kehtib

$$\hat{\theta}_0 \dot{\sim} N_d(\theta_0, [I_n(\theta_0)]^{-1}), \tag{3.4}$$

kus $I_n(\theta_0)$ on Fisheri informatsioon parameetri väärtuse θ_0 korral ehk maatriksi $I_n(\theta_0)$ i-ndas reas ja j-ndas veerus asub element $I_{i,j}(\theta_0) = -\mathbb{E}[\ell_{\theta_i\theta_j}(\theta_0)]$.

□

On näidatud, et mitmete suurima tõepära meetodiga analoogiliste meetodite korral on saadavad hinnangud üldistatud Pareto jaotuse korral asümptootiliselt normaalsed [2].

Parameetrile θ vahemikhinnagut leides kasutatakse praktikas lähenemist, kus leitakse logaritmilise tõepärafunktsiooni vajalikud tuletised ning arvutatakse need välja konkreetse valimi korral, kusjuures θ_0 rolli võetakse tema eelnevalt leitud suurima tõepära hinnang.

Järgnevalt leiame ülalkirjeldatud Fisheri informatsioonimaatriksi elemendid üldistatud Pareto jaotuse korral lähtudes vaatlustest $y_1, \ldots, y_{n_u}$ ja eeldusest, et $\varepsilon \neq 0$. Esmalt leiame logaritmilise tõepärafunktsiooni esimese ja teise tuletise $\tilde{\sigma}$ järgi. Saame, et

$$\ell_{\tilde{\sigma}} = -\frac{n_u}{\tilde{\sigma}} - \frac{1+\varepsilon}{\varepsilon}\sum_{i=1}^{n_u}\frac{\tilde{\sigma}}{\tilde{\sigma}+\varepsilon\cdot y_i}\cdot\left(-\frac{\varepsilon\cdot y_i}{\tilde{\sigma}^2}\right) = -\frac{n_u}{\tilde{\sigma}} + \frac{1+\varepsilon}{\varepsilon}\sum_{i=1}^{n_u}\left(\frac{1}{\tilde{\sigma}} - \frac{1}{\tilde{\sigma}+\varepsilon\cdot y_i}\right)$$

ja

$$\ell_{\tilde{\sigma}\tilde{\sigma}} = \frac{n_u}{\tilde{\sigma}^2} + \left(1+\frac{1}{\varepsilon}\right)\left[-\frac{n_u}{\tilde{\sigma}^2} + \sum_{i=1}^{n_u}\frac{1}{(\tilde{\sigma}+\varepsilon\cdot y_i)^2}\right] = \frac{1+\varepsilon}{\varepsilon}\sum_{i=1}^{n_u}\frac{1}{(\tilde{\sigma}+\varepsilon\cdot y_i)^2} - \frac{n_u}{\tilde{\sigma}^2\cdot\varepsilon}\ . \tag{3.5}$$

Järgnevalt leiame logaritmilise tõepärafunktsiooni esimese ja teise tuletise ε järgi. Saame, et

$$\ell_{\varepsilon} = \frac{1}{\varepsilon^2}\sum_{i=1}^{n_u}\ln\left(1+\frac{\varepsilon\cdot y_i}{\tilde{\sigma}}\right) - \frac{1+\varepsilon}{\varepsilon}\sum_{i=1}^{n_u}\frac{y_i}{\tilde{\sigma}+\varepsilon\cdot y_i}$$

ja

$$\ell_{\varepsilon\varepsilon} = -\frac{2}{\varepsilon^3}\sum_{i=1}^{n_u}\ln\left(1+\frac{\varepsilon\cdot y_i}{\tilde{\sigma}}\right) + \frac{2}{\varepsilon^2}\sum_{i=1}^{n_u}\frac{y_i}{\tilde{\sigma}+\varepsilon\cdot y_i} + \frac{1+\varepsilon}{\varepsilon}\sum_{i=1}^{n_u}\frac{y_i^2}{(\tilde{\sigma}+\varepsilon\cdot y_i)^2}\ . \tag{3.6}$$

Viimaseks leiame logaritmilise tõepärafunktsiooni segatuletise. Tulemuseks on

$$\ell_{\varepsilon\tilde{\sigma}} = -\frac{1}{\varepsilon}\sum_{i=1}^{n_u}\frac{y_i}{\tilde{\sigma}(\tilde{\sigma}+\varepsilon\cdot y_i)} + \frac{1+\varepsilon}{\varepsilon}\sum_{i=1}^{n_u}\frac{y_i}{(\tilde{\sigma}+\varepsilon\cdot y_i)^2}\ . \tag{3.7}$$

Tulemusi (3.5), (3.6) ja (3.7) kasutab statistikapaketis R realiseeritud programm (vt lisa B), mis võimaldab leida teoreemis 3.1 esineva pöördmaatriksi hinnangu, saades sisenditena ette lävendit ületavad väärtused, millest on lävendi väärtus maha lahutatud, ja üldistatud Pareto jaotuse parameetrite suurima tõepära hinnangud.

Lävendit ületavate väärtuste esinemissageduse p kui binoomjaotuse parameetri suurima tõepära hinnanguks on

$$\hat{p} = \frac{n_u}{n} \tag{3.8}$$

ning hinnangu standardviga avaldub kui

$$\sqrt{\hat{D}\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ . \tag{3.9}$$

3.3. Mudeli tõlgendamine

Lisaks traditsioonilistele mudeli tõlgendamise abivahenditele nagu seda on tõenäosuspaber ja kvantiil-kvantiil graafik leiab lävendimudelite korral kasutamist ka realiseerumisgraafik (*return level plot*), mis sisuliselt näitab mingist tasemest suuremate väärtuste esinemissagedusi.

Olgu lävendimudel kasutatav nivoo u korral ja $\varepsilon \neq 0$. See tähendab, et kehtib (2.6). Nagu nägime seose (2.1) tuletamise juures, võime siis iga $x > 0$ korral kirjutada

$$P\{X_i > u + x\} \approx P\{X_i > u\}\left(1 + \frac{\varepsilon\cdot x}{\tilde{\sigma}}\right)^{-1/\varepsilon}.$$

Olgu meile huvipakkuv selline väärtus x_m, et lävendi ja selle väärtuse x_m summa ületaks vaatlusalune juhuslik suurus tõenäosusega $1/m$, ehk siis (kuna vaatlused on sõltumatud) keskmiselt ületatakse tase $u + x_m$ üks kord m vaatluse jooksul. Saame ligikaudse võrduse

$$P\{X_i > u\}\left(1 + \frac{\varepsilon\cdot x_m}{\tilde{\sigma}}\right)^{-1/\varepsilon} \approx \frac{1}{m}\ ,$$

kust

$$x_m \approx \frac{\tilde{\sigma}}{\varepsilon}\left[(m \cdot P\{X_i > u\})^{\varepsilon} - 1\right] \tag{3.10}$$

ja seega oleme saanud, et väärtus $u + \tilde{\sigma}\left[(m \cdot P\{X_i > u\})^{\varepsilon} - 1\right]/\varepsilon$ ületatakse keskmiselt kord m vaatluse jooksul, kusjuures väärtus m peab olema sedavõrd suur, et x_m oleks positiivne. Suurus m kujutatakse graafiku horisontaalteljel ning $u + x_m$ vertikaalteljel. Võrdluseks kantakse graafikule empiirilised tulemused (vt joonis 4.3).

Realiseerumisgraafikut võimaldab joonistada ka kirjutatud programm (vt lisa C), mis vajab sisendina andmeid, üldistatud Pareto jaotuse parameetrite suurima tõepära hinnanguid ja vastavat fikseeritud lävendi väärtust.

4. Ekstremaalväärtuste lävendimudelid praktikas

Viimases peatükis vaatame, kuidas toimub töös kirjeldatu vahetu rakendamine reaalse andmestiku korral, kus eeldame, et vaatlused on sõltumatud ja samast jaotusest, ent see jaotus on teadmata.

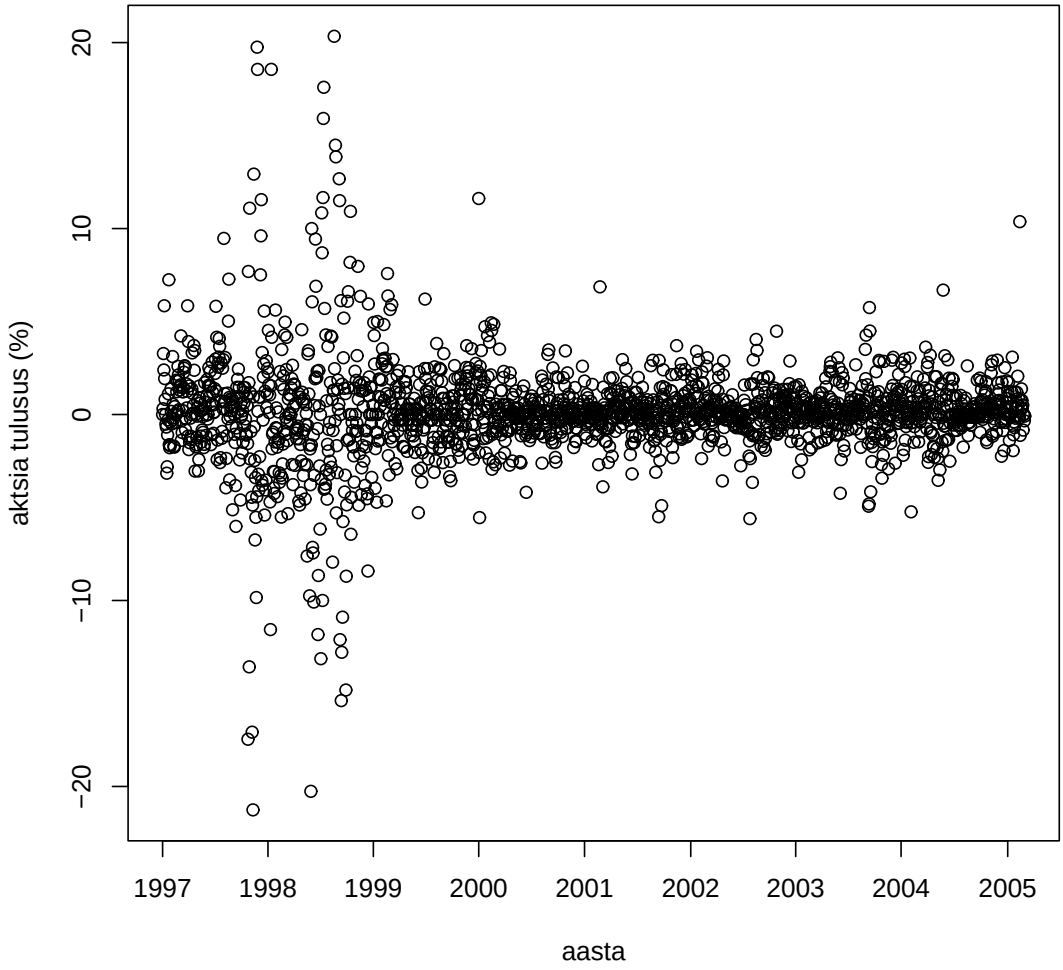
4.1. Lävendimudeli sobitamine Hansapanga aktsia tulususele

Semestritöös kasutasime perioodi 1. jaanuar 1997 kuni 1. märts 2004 Hansapanga aktsia sulgemishindade põhjal leitud aktsia päevaseid tulususi, jagasime need blokkideks ning modelleerisime blokkide miinimumi jaotust üldistatud ekstremaalväärtuste jaotuse abil [4: 23-26].

Käesolevas töös on kasutusel samad andmed, ent vaatlusperiood on ühe täiendava aasta võrra pikenenud, väldates nüüd 1. märtsini 2005. Esialgse valimi moodustavad 2071 vaatlust. Näide andmetabelist on toodud lisas D. Tuletame meelde, et aktsia kauplemispäevale i vastav tulusus on defineeritud kui päevade i ja $i-1$ aktsia sulgemishindade vahe suhe aktsia sulgemishinda kauplemispäeval $i-1$ ehk $r_i = (p_i - p_{i-1})/p_{i-1}$. Selline teisendus peaks ligikaudselt tagama eelduse, et andmed on sõltumatud ja sama jaotusega.

Joonis 4.1 aga kinnitab, et eeldus, nagu oleksid vaatlused samast jaotusest, on ilmselgelt rikutud. Aastatel 1997 ja 1998, mil esmalt toimus aktsiahindade kiire tõus ning hiljem veelgi järsem langus, on väga suuri ja väga väikeseid tulususi selgelt enam kui hilisematel aastatel. Seetõttu tundub mõistlik esimese kahe aasta vaatluste eemaldamine valimist. Uue valimi maht on 1565 vaatlust.

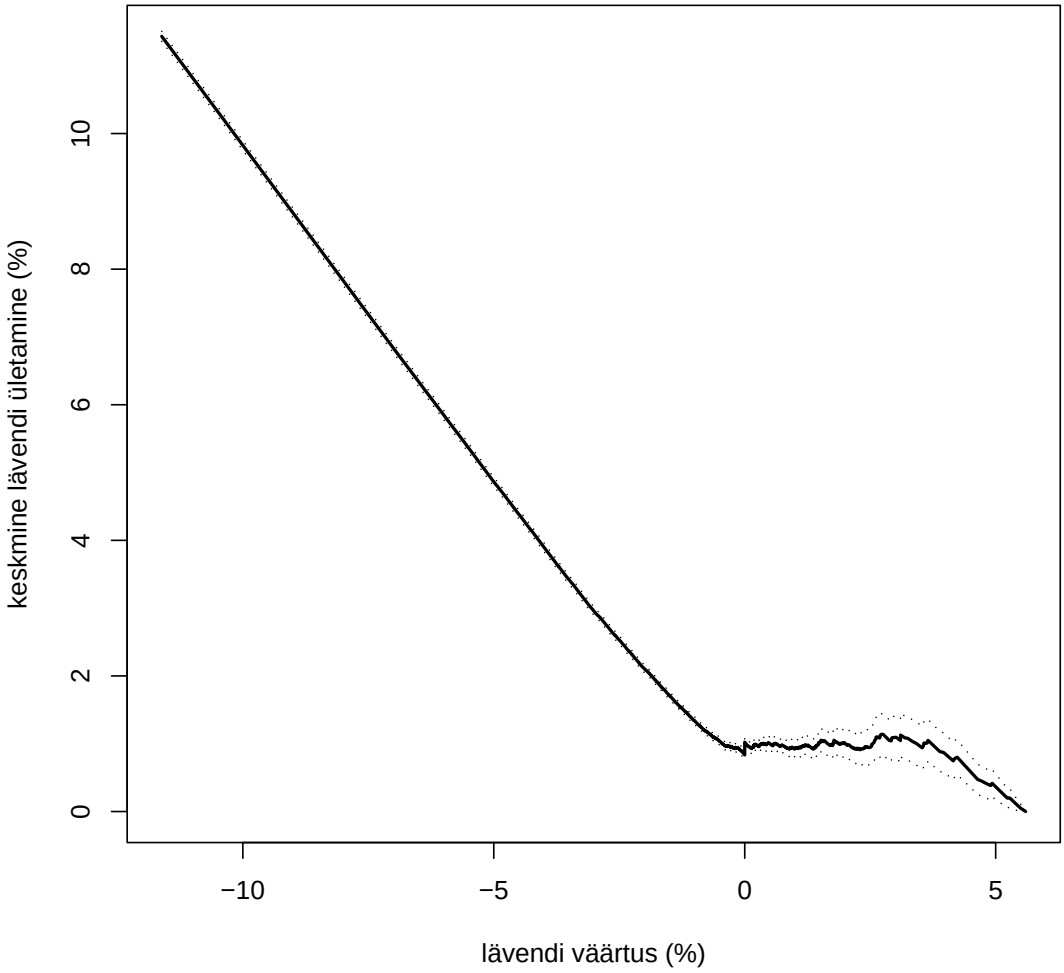
Töötame välja mudeli, mis iseloomustab ainult Hansapanga aktsiat sisaldava portfelliga seotud riske. Tulemusena saadav mudel võimaldab meil ühelt poolt hinnata suurust VaR (oma olemuselt tulususte α-kvantiil etteantud α väärtuse korral) kui üht klassikalist finantsriski hindamisel kasutatavat suurust, mis väljendab minimaalset kaotust protsentides, mida kanname halval, tõenäosusega α esineval börsipäeval. Samuti saame hinnata tinglikku keskväärtust $\mathbb{E}(X_i|\ X_i < q_\alpha)$, kus q_α on juhusliku suuruse X_i jaotuse α-kvantiil. See keskväärtus väljendab keskmist kaotust protsentides, mida kanname halval, tõenäosusega α esineval börsipäeval.



Joonis 4.1 *Hansapanga aktsia sulgemishinnapõhine tulusus aastatel 1997 – 2005.*

Riski hindamisel on huvipakkuvamad negatiivsed tulusused (kauplemispäevad, mil aktsia hind langeb). Praktikas tähendab see, et esmalt tuleb GPD sobitamisel andmetel märki muuta. Seega modelleerime nüüd üldistatud Pareto jaotuse abil juhuslikku suurust $Y_i = -X_i$, ent Y_i kohta saadavad tulemused on hõlpsasti ülekantavad suurusele X_i.

Eelmises peatükis nägime, et järgmiseks sammuks mudeli sobitamisel on sobiva lävendi valik. Selleks rakendame jääkeluea graafiku leidmise programmi muudetud märgiga tulusustele. Tulemuseks on joonis 4.2, kust minimaalse lävendi u, millest alates graafik muutub sirgeks, määramine on oodatult raske.



Joonis 4.2 *Muudetud märgiga Hansapanga aktsia tulususe keskmise jääkeluea graafik.*

On selge, et allapoole kaarduv (ka usalduspiire arvestades) graafiku lõpuosa on igal juhul paratamatu, ehkki võrduse (3.2) põhjal peaks $0 < \varepsilon < 1$ korral lävendi kasvades kasvama ka keskmine lävendi ületamine. Lõpliku valimi korral see tegelikkuses nii muidugi olla ei saa.

Tundub, et väärtus $u = 2$ näib olevat piir, millest alates on jääkeluea graafik tõlgendatav sirgena (usalduspiire arvestades). Eraldades vaatlused, mis ületavad fikseeritud lävendit (kokku 82 vaatlust), ning maksimeerides nende korral funktsiooni (3.3) numbriliselt, on tulemuseks punkthinnangud $\hat{\tilde{\sigma}} = 1.12475$ ja $\hat{\varepsilon} = -0.10713$. Saadud hinnangutele usalduspiiride leidmiseks kasutame Fisheri informatsiooni pöördmaatriksit leida võimaldavat

programmi, mille väljundiks on antud juhul maatriks

$$[I_n(\hat{\theta}_0)]^{-1} = \begin{pmatrix} 0.01221 & -0.00131 \\ -0.00131 & 0.00582 \end{pmatrix}.$$

Teoreemi 3.1 põhjal avaldub $(1-\alpha)$-protsendiline usaldusvahemik parameetrile $\tilde{\sigma}$ kui $1.12475 \pm z_{\frac{\alpha}{2}}\sqrt{0.01221}$ ja parameetrile ε kui $-0.10713 \pm z_{\frac{\alpha}{2}}\sqrt{0.00582}$, kus $z_{\frac{\alpha}{2}}$ on standardse normaaljaotuse $\alpha/2$-täiendkvantiil. Võttes $\alpha = 0.05$ saame 95% usaldusintervallid parameetritele $\tilde{\sigma}$ ja ε vastavalt (0.90818, 1.34132) ja (-0.25665, 0.04239). Parameetri ε usaldusvahemikust, milles sisaldub väärtus 0, järeldub, et võimalikud kandidaadid lävendit ületavate väärtuste jaotuse kohale on nii tõkestatud kui ka tõkestamata üldistatud Pareto jaotus. Reaalselt ei saa aktsia muidugi kaotada rohkem kui 100% oma väärtusest. STP-hinnangutega määratud üldistatud Pareto jaotuse parempoolseks otspunktiks on aga ligikaudu 10.5.

Kokkuvõttes oleme Hansapanga aktsia tulususte X_i tinglikku jaotust hindama sobitanud tõkestatud otspunktiga üldistatud Pareto jaotuse. Selle tulemuse võtab kokku seos

$$P\{X_i < -(2+x) |\ X_i < -2\} \approx (1 - 0.09525 \cdot x)^{9.33445}, \tag{4.1}$$

kus $0 < x < 10.49893$.

Lävendi $u = 2$ ületamise tõenäosuse hinnang on võrduse (3.8) põhjal $\hat{p} = 0.05240$ ja hinnangu standardviga võrduse (3.9) põhjal $\sqrt{\hat{D}\hat{p}} = 0.00563$. Seega võib hinnanguliselt öelda, et kahekümnest börsipäevast ühel langeb Hansapanga aktsia enam kui 2%. Ligikaudne $(1-\alpha)$-protsendiline usaldusvahemik aga avaldub kui $0.05240 \pm z_{\frac{\alpha}{2}} 0.00563$. Seega on ligikaudne 95% usaldusintervall parameetrile p vahemik $(0.04137, 0.06343)$.

Teises peatükis sooritatud arutelu põhjal on suure n väärtuse korral lävendit ületavate väärtuste arvu N_u jaotus lähendatav Poissoni jaotusega. Kehtib ligikaudne võrdus

$$P\{N_u = k\} \approx \frac{\lambda^k}{k!}\, e^{-\lambda}, \tag{4.2}$$

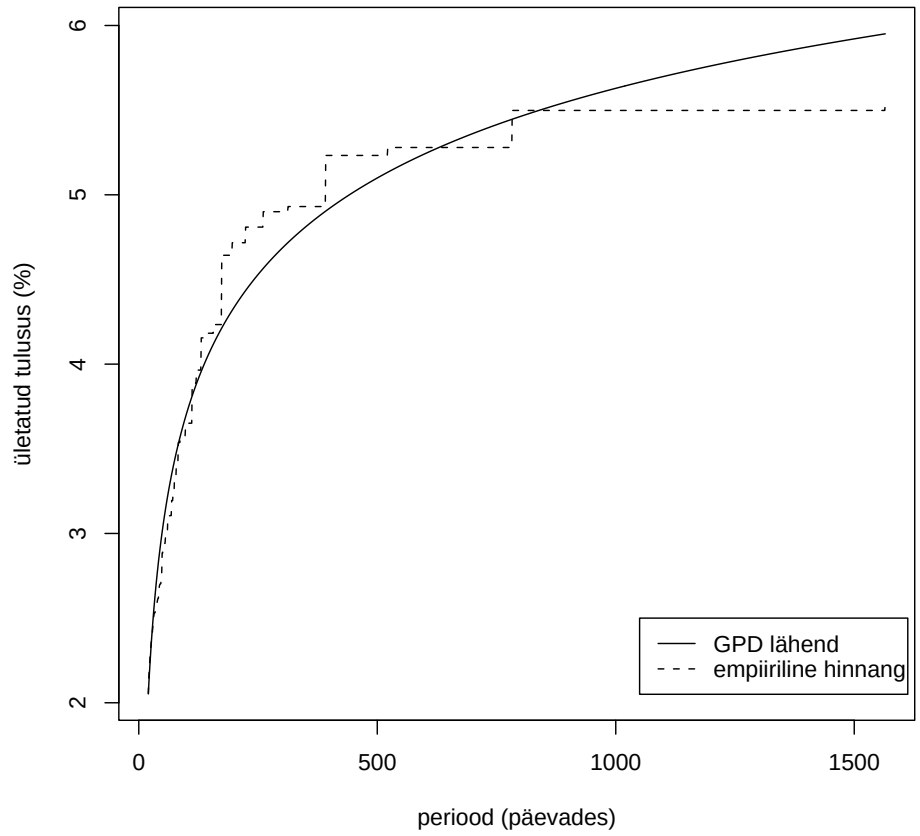
kus $k \in \{0, 1, 2, \dots\}$ ja $\lambda = 0.05240 \cdot n$. Võttes eelduseks, et aastas on 250 börsipäeva, saame vahetu arvutamise teel, et sellise sündmuse, kus aastase perioodi vältel langeb Hansapanga aktsia päevaga üle 2% vähem kui kümnel korral, esinemistõenäosus on alla ühe kuuendiku.

4.2. Sobitatud mudeli diagnostika ja tõlgendamine

Järgmisena pakub huvi eelmises peatükis kirjeldatud realiseerumisgraafik, mis on kujutatud joonisel 4.3. Et antud juhul on tulususte märk muudetud, siis võib öelda, et GPD lähend alahindab riski piirkonnas, kus ta asub allpool empiirilist hinnangut, ja ülehindab mujal. Nii võib joonise põhjal hinnata, et keskmiselt korra tuhande kauplemispäeva jooksul langeb Hansapanga aktsia väärtus enam kui 5.5%, poolteist tuhande börsipäeva jooksul leidub aga keskmiselt üks päev, mil aktsia kaotab koguni 6% oma väärtusest.

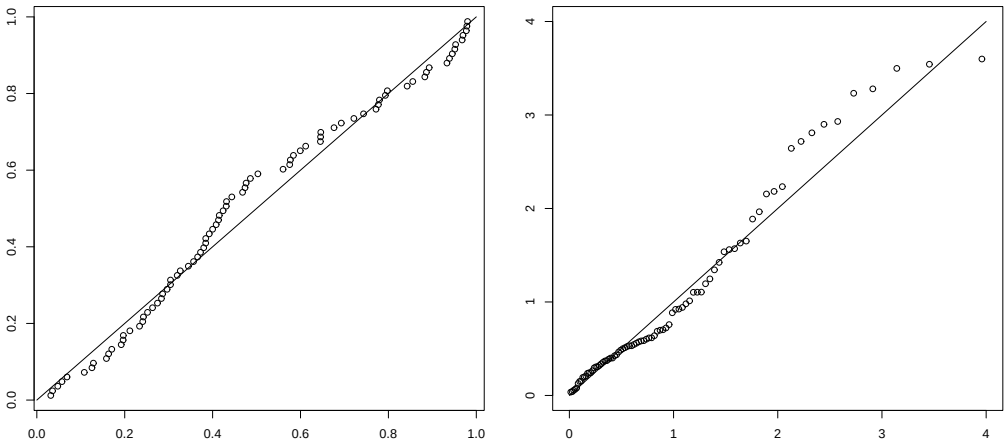
Ligikaudse võrduse (3.10) põhjal koondub x_m negatiivse ε korral suuruseks $-\tilde{\sigma}/\varepsilon$. Seega seab leitud mudel maksimaalse võimaliku kaotuse tasemele $u + \lim x_m \approx 12.5\%$.

Sobitatud jaotuse keskväärtus on seose (3.2) põhjal ligikaudu 1.02% ehk siis halva börsipäeva kohta, mil Hansapanga aktsia kaotab enam kui 2% oma väärtusest, võime hinnanguliselt öelda, et tegelik kaotus on keskmiselt 3.02%.



Joonis 4.3 *Hansapanga aktsia muudetud märgiga tulususte realiseerumisgraafik, mis on antud juhul suuruse VaR hinnang.*

Seni oleme üles lugenud olulisemad mudelist saadud järeldused, ent pole veel hinnanud (realiseerumisgraafik välja arvatud), kui hästi sobitatud jaotus tegelikult andmetega klapib. Joonisel 4.4 kujutatud graafikud osutavad teatavatele puudujääkidele, mis võivad viidata eelduste ligikaudsele täidetusele. On küsitav, kas kasutatud vaatlused ikka on sõltumatud – börsil järgnevad ekstremaalsete tulusustega kauplemispäevad sageli üksteisele, moodustades nii eraldi gruppe.



Joonis 4.4 *Sobitatud mudeli diagnostika: a) tõenäosuspaber, b) kvantiil-kvantiil graafik.*

Kokkuvõte

Lävendimudelite teooria on üks klassikalisest ekstremaalväärtuste teooriast välja kasvanud harusid, mis tegeleb jaotuste parempoolse saba (minimaalsete väärtuste uurimisel jaotuste vasakpoolse saba) omaduste uurimisega.

Kui juhusliku suuruse jaotuse saba on sedavõrd sile, et sõltumatute ja sama jaotusega juhuslike suuruste jada sobivalt normeeritud maksimum (miinimum) koondub üldistatud ekstremaalväärtuste jaotuseks, siis koondub antud jaotusest pärit juhusliku suuruse jaotus, tingimusel et ta ületab fikseeritud lävendit, lävendi kasvamisel üldistatud Pareto jaotuseks. Piirjaotuse parameetrid on seejuures üldistatud ekstremaalväärtuste jaotuse parameetreid teades vahetult leitavad.

Töötades reaalsete andmetega, mis ligilähedaselt rahuldavad sõltumatuse ja sama jaotuse eeldusi, kasutatakse sobiva lävendi määramiseks keskmise jääkeluea graafikut. Tulemuste tõlgendamiseks on lisaks hinnatud mudelile kasutatav ka realiseerumisgraafik, millelt on võimalik välja lugeda teatava vaatlusperioodi jooksul keskmiselt korra esineva ekstremaalse vaatluse väärtust.

Threshold models for extreme values

Bachelor thesis

Ants Kaasik

Abstract

The purpose of this thesis was to present the basic results concerning threshold models in extreme value theory and apply them in practice.

This paper studies the relations between generalized extreme value distribution (GEVD) and generalized Pareto distribution (GPD) and explains the motivation behind threshold apporach. Distribution of a random variable which belongs to the domain of attraction of an extreme value distribution, on the condition that it exceeds a high threshold, converges to the generalized Pareto distribution. The distribution parameters can be found for GPD when they are already known for the extreme value distribution.

Issues of selecting the proper threshold to assure the applicability of GPD as the limit distribution are also dealt with and a R based program (given in the appendix of the paper) can be used in practice.

In the last chapter extremal values of real data (daily rate of returns for the Hansapank share acquired from OMX Exchanges Tallinn) are modeled with a generalized Pareto distribution.

Lisa A. Programm jääkeluea graafiku leidmiseks

```
#sisendid on andmevektor ja olulisusnivoo
mrlp=function(andmed,alpha){
c=qnorm(1-alpha/2)
data=sort(andmed)
pikkus=length(andmed)
u=rep(NA,pikkus-1)
y=rep(NA,pikkus-1)
usaldus=rep(NA,pikkus-1)
for (i in 1:pikkus){
u[i]=data[1]
y[i]=mean(data)-data[1]
usaldus[i]=c*sd(data)/sqrt(length(data))
data=data[-1]
}
tulemus=new.env()
tulemus$u=u
tulemus$ex=y
tulemus$uci=y+usaldus
tulemus$lci=y-usaldus
as.list(tulemus)
return(tulemus)
}
r=mrlp(andmed,0.05)
plot(r$u,r$ex,lty=1,lwd=2,xlab="u",type="l",ylab="keskmine ületus")
lines(r$u,r$lci,lty=3)
lines(r$u,r$uci,lty=3)
```

Lisa B. Programm Fisheri informatsiooni pöördmaatriksi hindamiseks

```
#sisendid on lävendi ületamised ja parameetrite hinnangud
mlse=function(andmed,s,e){
n=length(andmed)
i11=(-n/e/s**2)+((1+e)/e)*sum(1/((s+e*andmed)**2))
i12=(-1/e)*sum(andmed/(s*(s+e*andmed)))
i22=(-2/e**3)*sum(log(1+e*andmed/s))+(2/e**2)*sum(andmed/(s+e*andmed))+
((1+e)/e)*sum((andmed**2)/((s+e*andmed)**2))
i=(matrix(c(-i11,-i12,-i12,-i22),nrow=2))**(-1)
return(i)
}
```

Lisa C. Programm realiseerumisgraafiku leidmiseks

```
#sisendid on andmevektor, parameetrite hinnangud ja lävendi väärtus
rlp=function(andmed,s,e,u){
n=length(andmed)
pu=sum(andmed>u)/n
minm=ceiling(1/pu)
m=seq(minm,length(andmed),1)
xm=s*((m*pu)**e-1)/e
data=sort(andmed)
emp=data[floor(n*(1-1/m))]
tulemus=new.env()
tulemus$m=m
tulemus$mudel=u+xm
tulemus$emp=emp
as.list(tulemus)
return(tulemus)
}
r=rlp(andmed,s,e,u)
plot(r$m,r$mudel,type="l",xlab="periood",ylab="ületatud väärtus")
lines(r$m,r$emp)
```

Lisa D. Väljavõte andmetabelist

Kuupäev	Sulgemishind (kr)	Muutus (%)	Kuupäev	Sulgemishind (kr)	Muutus (%)
2.01.1997	80,60	0,0000	21.01.1997	85,30	-0,2922
3.01.1997	80,85	0,3102	22.01.1997	86,25	1,1137
6.01.1997	83,50	3,2777	23.01.1997	92,50	7,2464
7.01.1997	85,50	2,3952	24.01.1997	91,00	-1,6216
8.01.1997	90,50	5,8480	27.01.1997	89,38	-1,7857
9.01.1997	92,25	1,9337	28.01.1997	87,88	-1,6783
10.01.1997	93,00	0,8130	29.01.1997	87,53	-0,3983
13.01.1997	93,00	0,0000	30.01.1997	88,73	1,3710
14.01.1997	92,53	-0,5108	31.01.1997	88,88	0,1691
15.01.1997	91,90	-0,6755	3.02.1997	88,43	-0,5063
16.01.1997	89,00	-3,1556	4.02.1997	88,25	-0,1979
17.01.1997	86,50	-2,8090	5.02.1997	91,00	3,1161
20.01.1997	85,55	-1,0983	6.02.1997	91,50	0,5495

Viited

[1] COLES, S. *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag, 2001.

[2] DREES, H., FERREIRA, A. & DE HAAN, L. On Maximum Likelihood Estimation of the Extreme Value Index. *The Annals of Applied Probability 3*, 2004.

[3] EMBRECHTS, P., KLÜPPELBERG, C. & MIKOSCH, T. *Modelling Extremal Events for Insurance and Finance.* Springer-Verlag, 1997.

[4] KAASIK, A. *Ekstremaalväärtuste jaotuste modelleerimine. Semestritöö.* Tartu Ülikool, 2004.

[5] PICKANDS, J. Statistical Inference Using Extreme Order Statistics. *Annals of Statistics 3*, 1975.

[6] SMITH, R. Statistics of Extremes with Applications in Environment, Insurance, and Finance. *Extreme Values in Finance, Telecommunications, and the Environment*, Chapman & Hall/CRC, 2004.