Tampereen yliopisto

Veera Kalliovalkama

STUDENTIN T-JAKAUMAN JA BETA-JAKAUMAN SOVITTAMINEN DATAAN

Tekniikan ja luonnontieteiden tiedekunta
Kandidaatintyö
Huhtikuu 2020

TIIVISTELMÄ

Veera Kalliovalkama: Studentin t-jakauman ja beta-jakauman sovittaminen dataan
Kandidaatintyö
Tampereen yliopisto
Tekniikka ja luonnontieteet, TkK
Huhtikuu 2020

Tässä kandidaatintyössä tutustutaan siihen, miten dataan saadaan sovitettua jakauma. Työssä on käytössä atsimuuttidata, jota käytetään äänilähteen paikantamiseen. Koska data ei noudata suoraan mitään todennäköisyysjakaumaa, sille muodostetaan tiheysfunktio yhdistämällä kahden eri jakauman tiheysfunktiot. Lopputuloksena saadaan dataa hyvin mukaileva tiheysfunktio.

Työn tarkasteluissa käytetään Beta-jakaumaa ja Studentin t-jakaumaa. Molemmat ovat jatkuvia todennäköisyysjakaumia, mutta niillä on kuitenkin hieman erilaiset ominaisuudet. Beta-jakauma sopii satunnaismuuttujien mallintamiseen äärellisellä välillä nollasta yhteen. Sen tiheysfunktion muodon määräävät parametrit α ja β, jotka ovat molemmat positiivisia reaalilukuja. Studentin t-jakauma taas sopii symmetrisen datan mallintamiseen ja on muodoltaan kellomainen. Studentin t-jakauman muotoon vaikuttaa sen vapausaste k, joka on myös positiivinen reaaliluku.

Työssä tutustutaan suurimman uskottavuuden estimointiin, jota käytetään jakauman sovittamiseen dataan. Menetelmän avulla pystytään etsimään parametrit, joiden myötä haluttu jakauma parhaiten sopii dataan. Suurimman uskottavuuden estimoinnissa siis pyritään maksimoimaan uskottavuusfunktio sen parametrien suhteen. Siispä mitä lähempänä uskottavuusfunktion arvot ovat sen maksimia, sitä paremmin haluttu jakauma sopii annettuun dataan.

Työssä on käytössä äänilähteen paikannukseen käytettyä dataa, eli atsimuuttidataa. Kuten aikaisemmin on mainittu, data ei noudata suoraan mitään todennäköisyysjakaumaa. Tästä syystä tarkasteluissa data jaetaan osiin kahteen osaan ja kumpaankin osaan sovitetaan sekä Studentin t-jakauma että Beta-jakauma. Näistä jakaumista valitaan kumpaankin parhaiten sopiva jakauma ja niistä muodostetaan yhdistetty tiheysfunktio. Tiheysfunktiot kerrotaan molemmat niiden painotuskertoimilla ja sen jälkeen summataan yhteen, jolloin lopputuloksena on koko käytössä olevaa dataa mukaileva tiheysfunktio.

Avainsanat: Studentin t-jakauma, Beta-jakauma, suurimman uskottavuuden menetelmä

SISÄLLYSLUETTELO

LYHENTEET JA MERKINNÄT

B_x	Etätäydellinen Beta-funktio
F_T	Studentin t-jakuman kertymäfunktio
I_x	Epätäydellisen ja täydellisen Beta-funktion suhde
L	Uskottavuusfunktio
M_β	Beta-jakauman momentit generoiva funktio, MGF
Γ	Gamma-funktio
$\bar{L}$	Normalisoitu uskottavuusfunktio
μ_T	Studentin t-jakauman odotusarvo
μ_β	Beta-jakauman odotusarvo
ω_{MLE}	Suurimman uskottavuuden estimaattori, ML-estimaattori
σ_T^2	Studentin t-jakauman varianssi
σ_β^2	Beta-jakauman varianssi
f_T	Studentin t-jakauman tiheysfunktio
f_β	Beta-jakauman tiheysfunktio
f_v	Vinon jakauman tiheysfunktio
MGF	Momentit generoiva funktio
ML	Suurin uskottavuus, maximum likelihood

1 JOHDANTO

Todennäköisyyslaskennan teoria pohjautuu tavallisiin uhkapeleihin, esimerkiksi korttipeleihin. Eräänä teorian alkutekijänä pidetään matemaatikkojen Pierre Fermat ja Blaise Pascal välistä kirjeenvaihtoa uhkapeleistä 1600-luvulla [1].

Todennäköisyyslaskennan keskeisiä asioita ovat diskreetit ja jatkuvat satunnaismuuttujat ja todennäköisyysjakaumat. Satunnaismuuttujien eri arvojen yleisyyttä kuvataan todennäköisyysjakauman avulla. Toisin sanoen todennäköisyysjakauman avulla voidaan arvioida jonkun tapahtuman lopputuloksien todennäköisyyttä. Diskreetti todennäköisyysjakauma sopii tilanteisiin, joissa mahdollisia tapahtumia on numeroituva määrä. Tästä hyvä esimerkki on kolikon heitto, jossa yhdellä heitolla lopputulokseksi voi saada joko kruunan tai klaavan. Jatkuva todennäköisyysjakauma taas sopii tilanteisiin, joissa tapahtuma tapahtuu jatkuvana. Lämpötilan muutos tietyn päivän aikana on eräs esimerkki tilanteesta.

Tässä työssä käsitellään kahta jatkuvaa jakaumaa. Näiden jakaumien avulla analysoidaan atsimuuttidataa, jonka alkiot ovat jatkuvia satunnaismuuttujia. Atsimuuttidatasta pyritään määrittämään äänilähteen paikkaa detektorin ympärillä. Kun dataan sijoitetaan todennäköisyysjakauman tiheysfunktio, äänilähteen paikan määrittäminen ei ole enää silmämääräistä tarkastelua, vaan tuloksille saadaan matemaattinen pohja. Tapahtuma ei kuitenkaan aina noudata yhtä tiettyä jakaumaa, ja sen takia työssä tutustutaan kahteen eri jakaumaan. Näistä jakaumista voidaan muodostaa yhdistetty jakauma, joka noudattaa dataa paremmin.

Ensimmäiseksi työssä tutustutaan lyhyesti jakaumiin, joita työn aikana käytetään. Jakaumiksi on valikoitunut Beta-jakauma ja Studentin t-jakauma, joista työn edetessä tullaan muodostamaan yhdiste. Tämä yhdistetty jakauma sovitetaan käytössä olevaan atsimuuttidataan. Seuraavaksi tutustutaan suurimman uskottavuuden estimointiin. Menetelmän avulla pystytään löytämään parametrien arvot, jotka vastaavat haluttua tiheysfunktiota. Kun jakaumat ja menetelmät ovat tutut, voidaan ryhtyä käsittelemään atsimuuttidataa. Tarkoituksena on muodostaa yhdistetty tiheysfunktio ja sovittaa se dataan. Työn lopussa on yhteenveto työn käsittelemistä aiheista.

2 BETA-JAKAUMAN JA STUDENTIN T-JAKAUMAN ESITTELY

Tiheysfunktiot ovat satunnaismuuttujan reaaliarvoisia funktioita. Niillä on useimmiten yhdestä kolmeen parametria, joiden avulla määritellään funktion paikka, asteikko ja muoto. Tässä luvussa esitellään kaksi jakaumaa, Beta-jakauma ja Studentin t-jakauma, joita käytetään työssä käytettävän datan analysointiin.

2.1 Beta-jakauma

Beta-jakauma on yksi monista jatkuvista jakaumista ja se sopii satunnaismuuttujien mallintamiseen äärellisellä välillä. Sen tiheysfunktio on

$$f_\beta(t;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1}, \quad t \in [0,1], \tag{2.1}$$

missä α ja β ovat funktion parametrejä, jotka kertovat jakauman muodon ja ovat positiivisia reaalilukuja. Merkintä Γ viittaa Gamma-funktioon, joka on muotoa

$$\Gamma(n) = \int_0^\infty x^{n-1}\mathrm{e}^{-x}\mathrm{dx},$$

missä n voi olla mikä tahansa positiivinen reaaliluku. [2]

Beta-jakauman tiheysfunktio voidaan ilmoittaa myös Beta-funktion avulla, joka voidaan kirjoittaa täydellisessä tai epätäydellisessä muodossa. Kuitenkin tiheysfunktio voidaan kirjoittaa suljettuun muotoon ainoastaan epätäydellisen Beta-funktion avulla.

Epätäydellinen Beta-funktio on muotoa

$$B_x(\alpha,\beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{dt}. \tag{2.2}$$

Funktiosta (2.2) saadaan täydellinen, $B(\alpha,\beta)$, kun integraali tehdään välillä $[0,1]$ välin $[0,x]$ sijaan. [3] Epätäydellisen ja täydellisen Beta-funktion suhdetta merkitään

$$I_x(\alpha,\beta) = \frac{B_x(\alpha,\beta)}{B(\alpha,\beta)} \quad [4].$$

Tästä usein kirjallisuudessa unohtuu sana ’suhde’, jolloin Beta-funktio ja Beta-funktioiden suhde saattaa sekoittua toisiinsa.

Beta-funktio ja Gamma-funktio liittyvät toisiinsa seuraavasti:

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \text{ [3]}$$

Huomataan, että jakauman tiheysfunktiosta saadaan yksinkertaisemman näköinen, jos funktioon (2.1) sijoitetaan $\frac{1}{B(\alpha,\beta)}$ Gamma-funktioiden tilalle.

Beta-jakauman odotusarvo μ_β ja varianssi σ_β^2 voidaan ilmoittaa parametrien α ja β avulla

$$\mu_\beta = \frac{\alpha}{\alpha+\beta} \tag{2.3a}$$

$$\sigma_\beta^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \text{ [2]} \tag{2.3b}$$

Todistus. Forbes et al. kirjoittamassa kirjassa on taulukko 2.1, josta löytyy kaavat sekä odotusarvolle että varianssille. Käytetään näitä kaavoja todistuksessa.

Lähdetään liikkeelle odotusarvon yleisestä muodosta

$$\mu = \int t f(t)\mathrm{dt}, \tag{2.4}$$

missä f on jakauman tiheysfunktio [5]. Sijoitetaan kaavaan (2.4) funktion $f(t)$ paikalle Beta-jakauman tiheysfunktio (2.1). Beta-jakauman odotusarvo on muotoa

$$\begin{aligned}\mu_\beta &= \int_0^1 t\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{dt}\\ &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}t^{\alpha}(1-t)^{\beta-1}\,\mathrm{dt}.\end{aligned}$$

Beta-jakaumassa $t \in [0,1]$, jolloin tiheysfunktio on kyseisellä välillä nollasta poikkeava. Siispä integraali tehdään myös välillä $[0,1]$. Gamma-funktio on muuttujan t suhteen vakio, joten se voidaan siirtää integraalin ulkopuolelle. Beta-funktion odotusarvo on nyt

$$\begin{aligned}\mu_\beta &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 t^{\alpha}(1-t)^{\beta-1}\,\mathrm{dt}\\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(\alpha-1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)},\quad \alpha>-1 \;\; ja \;\; \beta>0\\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\cdot\frac{\Gamma(\alpha-1)}{\Gamma(\alpha+\beta+1)}.\end{aligned}$$

Gamma-funktiolla on ominaisuus $\Gamma(x+1) = x\Gamma(x)$, kun $x > 0$ [6]. Koska Beta-jakauman tiheysfunktion määritelmän mukaan $\alpha > 0$ ja $\beta > 0$, voidaan kyseistä ominaisuutta käyttää. Nyt odotusarvo saadaan muotoon

$$\mu_\beta = \frac{\Gamma(\alpha+\beta)\alpha\Gamma(\alpha)}{\Gamma(\alpha)(\alpha+\beta)\Gamma(\alpha+\beta)} = \frac{\alpha}{\alpha+\beta},$$

joka on samaa muotoa kuin kaava (2.3a).

Varianssi on yleistä muotoaan

$$\sigma^2 = \int (t-\mu)^2 f(t)\mathrm{d}t, \tag{2.5}$$

missä f on jakauman tiheysfunktio ja μ sen odotusarvo [5]. Sijoitetaan kaavaan (2.5) Beta-jakauman tiheysfunktio (2.1) sekä edellä todistettu Beta-jakauman odotusarvo μ_β. Nyt varianssi on muotoa

$$\sigma_\beta^2 = \int_0^1 (t-\mu_\beta)^2 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t.$$

Edelleen koska Gamma-funktio on muuttujasta t riippumaton, se voidaan ottaa integraalin ulkopuolelle.

$$\begin{aligned}\sigma_\beta^2 &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \left(t - \frac{\alpha}{\alpha+\beta}\right)^2 (1-t)^{\beta-1}\mathrm{d}t \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta+1}{(\alpha+\beta)\Gamma(\alpha+\beta+2)}, \;\; \alpha > 0 \;\; ja \;\; \beta > 0.\end{aligned}$$

Käytetään jälleen Gamma-funktion ominaisuutta $\Gamma(x+1) = x\Gamma(x)$ ja varianssi tulee muotoon

$$\begin{aligned}\sigma_\beta^2 &= \frac{\Gamma(\alpha+\beta)\alpha\Gamma(\alpha)\beta\Gamma(\beta)}{\Gamma(\alpha)\Gamma(\beta)(\alpha+\beta)(\alpha+\beta+1)\Gamma(\alpha+\beta+1)} \\ &= \frac{\Gamma(\alpha+\beta)\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)(\alpha+\beta)\Gamma(\alpha+\beta)} \\ &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.\end{aligned}$$

Varianssi on saatu samaan muotoon kuin kaavassa (2.3b).

□

Todistetaan yhtälöt (2.3) vaihtoehtoisella tavalla käyttäen momentit generoivaa funktiota (Moment generating function) Beta-jakaumalle.

Todistus. Momentit generoiva funktio Beta-jakaumalle on muotoa

$$M_\beta(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+1} \right) \frac{t^k}{k!}, \;\; \alpha,\beta > 0 \;\; [5].$$

Aloitetaan odotusarvosta, joka voidaan ilmoittaa MGF:n avulla $\mu_\beta = M'_\beta(0)$ [5].

Derivoidaan MGF muuttujan t suhteen kerran

$$M'_\beta(t) = \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{kt^{k-1}}{k!} = \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{t^{k-1}}{(k-1)!}. \tag{2.6}$$

Kun MGF:n derivaattaan sijoitetaan $t = 0$, saadaan

$$M'_\beta(0) = \left(\prod_{r=0}^{0} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{0^0}{0!} + \left(\prod_{r=0}^{1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{0^1}{1!} + \left(\prod_{r=0}^{2} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{0^2}{2!} + \ldots$$

Kun $k > 1$, summan termit ovat nolla. Lisäksi, kun hyödynnetään laskusääntöjä $0! = 1$ ja $0^0 = 1$, summasta saadaan Beta-jakauman odotusarvo

$$\mu_\beta = \left(\frac{\alpha + 0}{\alpha + \beta + 0}\right)\frac{1}{1} = \frac{\alpha}{\alpha + \beta},$$

joka vastaa haluttua muotoa (2.3a).

Beta-jakauman varianssi voidaan myös ilmoittaa MGF:n avulla $\sigma_\beta^2 = M_\beta''(0) - [M_\beta'(0)]^2 = M_\beta''(0) - \mu_\beta^2$ [5].

Muodostetaan MGF:n toinen derivaatta derivoimalla yhtälö (2.6) toiseen kertaan

$$M_\beta''(t) = \sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha + r}{\alpha + \beta + r}\right)\frac{(k-1)t^{k-2}}{(k-1)!} = \sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha + r}{\alpha + \beta + r}\right)\frac{t^{k-2}}{(k-2)!}.$$

Kun saatuun yhtälöön sijoitetaan $t = 0$ ja summa lasketaan auki, saadaan

$$M_\beta''(0) = \left(\prod_{r=0}^{0}\frac{\alpha + r}{\alpha + \beta + r}\right)\frac{0^{-1}}{(-1)!} + \left(\prod_{r=0}^{1}\frac{\alpha + r}{\alpha + \beta + r}\right)\frac{0^0}{(0)!} + \left(\prod_{r=0}^{2}\frac{\alpha + r}{\alpha + \beta + r}\right)\frac{0^1}{(1)!} + \ldots$$

Summan termit ovat nolla aina, kun $k \neq 2$. Beta-jakaumalle MGF:n toinen derivaatta pisteessä $t = 0$ tulee siis muotoon

$$\begin{aligned} M_\beta''(0) &= \left(\frac{\alpha + 0}{\alpha + \beta + 0}\cdot\frac{\alpha + 1}{\alpha + \beta + 1}\right)\frac{1}{1} \\ &= \frac{\alpha^2 + \alpha}{(\alpha + \beta)(\alpha + \beta + 1)}. \end{aligned}$$

Nyt Beta-jakauman varianssi saadaan muotoon

$$\begin{aligned} \sigma_\beta^2 &= \frac{\alpha^2 + \alpha}{(\alpha + \beta)(\alpha + \beta + 1)} - \left(\frac{\alpha}{\alpha + \beta}\right)^2 \\ &= \frac{(\alpha + \beta)^2(\alpha^2 + \alpha) - (\alpha + \beta)(\alpha + \beta + 1)\alpha^2}{(\alpha + \beta)(\alpha + \beta + 1)(\alpha + \beta)^2} \\ &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}, \end{aligned}$$

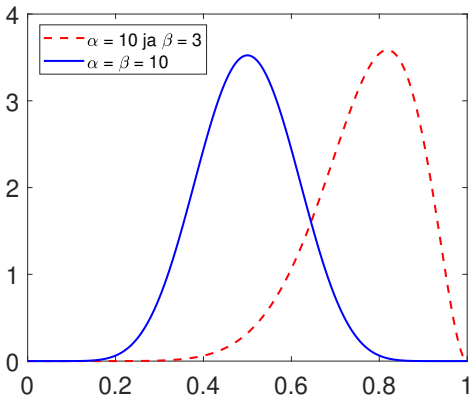
joka vastaa haluttua muotoa (2.3b). □

Määritelmä 2.1. (Symmetrisyys) Satunnaismuuttujan x todennäköisyysjakauma on symmetrinen luvun c suhteen, jos ja vain jos

$$P(x < c - k) = P(x > c + k), \quad k \geq 0. \quad [7]$$

Beta-jakauma on määritelmän mukaan symmetrinen luvun $0,5$ suhteen, kun sen parametrit α ja β ovat yhtä suuria. Kuvassa 2.1 on esimerkit symmetrisestä ja epäsymmetrisestä Beta-jakaumasta.

Kummallakin tiheysfunktiolla muuttujan t arvot vaihtelevat arvojen 0 ja 1 välillä. Kuvasta 2.1



Kuva 2.1. *Symmetrinen ja epäsymmetrinen beta-jakauma.*

huomataan, että epäsymmetrisen jakauman vasemmanpuoleinen häntä on huomattavasti paksumpi ja leveämpi kuin oikeanpuoleinen häntä.

2.2 Studentin t-jakauma

Studentin t-jakauma on jatkuva todennäköisyysjakauma, joka on muodoltaan symmetrinen ja kellomainen. Sen tiheysfunktio voidaan kirjoittaa muotoon

$$f_T(t;k) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{(\pi k)^{\frac{1}{2}}\Gamma\left(\frac{k}{2}\right)\left(1+\frac{t^2}{k}\right)^{\frac{1}{2}(k+1)}}, \quad t \in \mathbb{R}. \tag{2.7}$$

Parametri k on jakauman vapausaste, joka on positiivinen reaaliluku. Jakauman kertymäfunktio on

$$F_T(t;k) = 1 - \frac{1}{2} I_{\frac{k}{k+t^2}}\left(\frac{k}{2}, \frac{1}{2}\right) \quad [4].$$

Kun Studentin t-jakauman vapausaste k on suurempaa kuin yksi, jakauman odotusarvo μ_T on nolla. Samalla tavalla voidaan todeta, että kun jakauman vapausaste on suurempaa kuin kaksi, varianssi on

$$\sigma_T^2 = \frac{k}{k-2}.$$

Kun jakauman vapausaste lähestyy ääretöntä, varianssi lähestyy lukuarvoa yksi. Tällöin Studentin t-jakauma alkaa muistuttamaan normaalijakaumaa. Muulloin Studentin t-jakauman hännät ovat paksummat kuin normaalijakauman. [2]

2.3 Vino Studentin t-jakauma

Studentin t-jakauman heikkous on ollut se, ettei jakaumalla pystytä mallintamaan epäsymmetristä, eli vinoa dataa [8]. Viime vuosina on kuitenkin pystytty yleisesti muuttamaan symmetrinen jakauma vinoksi jakaumaksi ja tämä yleistys toimii myös Studentin t-jakaumalle.

Yleisesti symmetrinen tiheysfunktio $g(t)$ voidaan vinouttaa sitä vastaavan kertymäfunktion $G(t)$

avulla dataan sopivaksi määrittelemällä

$$f_v(t;\lambda) = 2g(t)G(\lambda t), \tag{2.8}$$

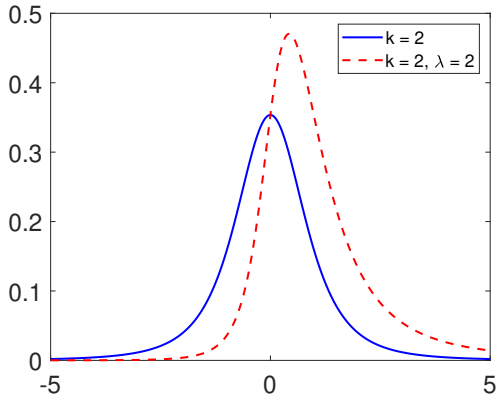
missä λ on kokonaisluku [3].

Funktiosta (2.8) saadaan vino Studentin t-jakauma, kun tiheysfunktion $g(t)$ paikalle sijoitetaan Studentin t-jakauman tiheysfunktio vapausasteella k. Samoin kertymäfunktion $G(t)$ paikalle sijoitetaan Studentin t-jakauman vastaava funktio.

Esimerkki 2.2. Kun Studentin t-jakauman vapausaste k on 2, sen vino tiheysfunktio on muotoa

$$2f_T(t;2)F_T(\lambda t;2) = \frac{1}{(2+t^2)^{\frac{3}{2}}}\left(1 + \frac{\lambda t}{\sqrt{2+\lambda^2 t^2}}\right), \quad t \in \mathbb{R} \quad [9].$$

Kuvassa 2.2 on esimerkit suorasta ja vinosta Studentin t-jakaumasta.



Kuva 2.2. *Suora ja vino Studentin t-jakauma.*

Kumpikin tiheysfunktio on tehty niin, että muuttujan t arvo on vaihdellut arvojen -5 ja 5 välillä. Kuvasta 2.2 huomataan, että vinon jakauman oikeanpuoleinen häntä on selkeästi paksumpi, kuin vasemmanpuoleinen. Lisäksi se on myös paksumpi kuin suoran jakauman hännät. Samoin vinon jakauman vasemmanpuoleinen häntä on ohuempi kuin suoran jakauman hännät.

3 SUURIMMAN USKOTTAVUUDEN ESTIMOINTI

Suurimman uskottavuuden estimointi (Maximum likelihood estimation) maksimoi uskottavuusfunktion parametrien suhteen. Toisin sanoen menetelmän avulla pyritään etsimään tiheysfunktio, joka parhaiten sopisi dataan. Tavoitteena on määrittää sellaisen jakauman parametrit, josta data on kaikkein uskottavimmin peräisin.

3.1 Uskottavuusfunktio

Uskottavuusfunktio on oleellinen osa suurimman uskottavuuden estimointia. Jotta estimointi on mahdollista tehdä, on uskottavuusfunktion oltava olemassa ja määritelty. Uskottavuusfunktio kertoo parametrivektorin $\omega = \{\omega_1, \omega_2, ..., \omega_k\}$ uskottavuuden annetun vektoriarvoisen datan $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ suhteen, kun $\mathbf{x}_i \in \mathbb{R}^p$. Toisin sanoen mitä lähempänä maksimiaan uskottavuusfunktion arvo on, sitä todennäköisemmin sovitettava jakauma sopii dataan. [10]

Uskottavuusfunktiota ei kuitenkaan pidä sekoittaa tiheysfunktioon. Kyseiset funktiot ovat määriteltyjä eri asteikoilla, joten ne eivät ole verrattavissa toisiinsa. Tiheysfunktion avulla voidaan laskea todennäköisyyksiä tapahtumille. Tällöin parametrin arvo on tunnettu, kun taas uskottavuusfunktiota käytettäessä data on tunnettu.

Parametrivektorin ω uskottavuusfunktiota datan $\mathbf{x}$ suhteen merkitään seuraavasti:

$$L(\omega; \mathbf{x}) = f(\mathbf{x}; \omega) = f(\mathbf{x}_1; \omega) f(\mathbf{x}_2; \omega) ... f(\mathbf{x}_n; \omega), \tag{3.1}$$

missä f on otoksen $\mathbf{x}$ tiheysfunktio parametreilla ω [11]. Kaikki termit, joissa ei ilmene parametriä ω, voidaan jättää huomiotta.

Uskottavuusfunktio pystytään normalisoimaan sen pienimmän ylärajan eli supremumin avulla

$$\bar{L}(\omega; \mathbf{x}) = \frac{L(\omega; \mathbf{x})}{\sup(L(\omega; \mathbf{x}))} \text{ [12].}$$

Normalisointi on yleistä erityisesti piirrettäessä, jolloin uskottavuusfunktion maksimiarvo on 1 ja sen logaritmi saa maksimiarvon 0.

Funktion (3.1) maksimi-pisteiden joukko muodostaa suurimman uskottavuuden estimaatin parametreille ω. Uskottavuusfunktion on oltava kahdesti derivoituva, jotta sen käsittely menetelmän myöhäisemmässä vaiheessa olisi mahdollista.

3.2 Uskottavuusyhtälö

Suurimman uskottavuuden estimaattori (ML-estimaattori) löydetään, kun derivoidaan uskottavuusfunktion luonnollista logaritmia eli log-uskottavuutta $\ln(L(\omega; \mathbf{x}))$. Uskottavuusyhtälö on muotoa

$$\frac{\partial \ln L(\omega; \mathbf{x})}{\partial \omega_j} = 0,$$

missä $j = \{1, 2, ..., k\}$. [10] Tämän yhtälön ratkaisu on suurimman uskottavuuden estimaattori $\omega_{MLE} = \{\omega_{1,MLE}, \omega_{2,MLE}, ..., \omega_{k,MLE}\}$.

Merkittävä ehto estimaattorin yksikäsitteisyydelle on uskottavuusfunktion kuperuus. Estimaattorin ainutlaatuisuus ei kuitenkaan ole taattavissa, joten yleisesti on yritettävä löytää normalisoidun uskottavuusfunktion (3.1) paikallinen maksimi. Toisin sanoen derivoidaan normalisoitu uskottavuusfunktio parametrien ω suhteen ja etsitään derivaatan nollakohta. [13]

On mahdollista, että uskottavuusyhtälön ratkaisu antaakin minimin log-uskottavuudelle maksimin sijaan. Log-uskottavuuden on siis oltava kupera ML-estimaattorin ympäristössä, ja tämä voidaan tarkistaa log-uskottavuuden toisen derivaatan avulla

$$\frac{\partial^2 \ln(L(\omega; \mathbf{x}))}{\partial \omega_j^2} < 0,$$

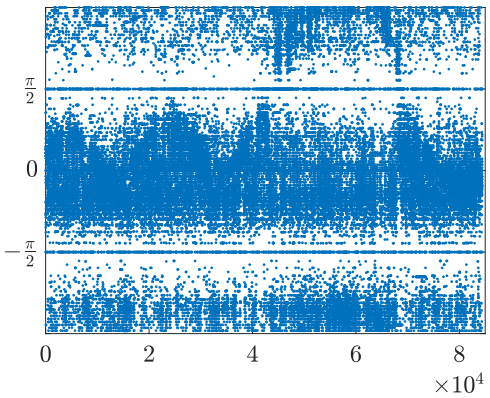
kun $\omega_j = \omega_{j,MLE}$. [10] Mikäli yllä olevan epäyhtälön ratkaisut ovat kaikki negatiivisia, kyseessä on log-uskottavuuden maksimi.

Ei kuitenkaan voida varmasti sanoa, onko kyseessä globaali vai lokaali maksimi. Lokaalin maksimin ongelma (local maxima problem) tulee vastaan, mikäli uskottavuusfunktiolla on useampia lokaaleja maksimeita. Tällöin menetelmä ei välttämättä palauta kaikkein optimaalisinta arvoa parametrille. Optimoinnin lopputulos riippuu pääasiassa optimoinnin alkuarvoista, jotka määrätään joko satunnaisesti tai arvaamalla. Lokaalin maksimin ongelmaan ei ole olemassa yleistä ratkaisua, vaikkakin erilaisia tekniikoita ongelman välttämiseen on kehitetty. Esimerkiksi menetelmää voi toistaa useilla eri alkuarvoilla ja verrata lopputuloksia. Tällöin, mikäli tulokset ovat samoja, on todennäköistä, että kyseessä on optimaalisin arvo parametrille.

4 JAKAUMAN SOVITTAMIEN ATSIMUUTTIDATAAN

Ihmiset käyttävät äänilähteen paikannusta luonnollisesti päivittäin paikantaessaan toisesta huoneesta kuuluvan kolahduksen tai auton, joka ei vielä ole nähtävissä. Äänilähteen paikannukselle on monia sovelluksia teollisuudessa ja uusia kehitellään koko ajan lisää. Esimerkiksi meren alla liikuttaessa voidaan vastaantulevia kulkuvälineitä paikantaa juuri äänen avulla [14]. Eräs sovellus on myös puheentunnistus. Lisäksi paikannuksen avulla pystytään erittelemään äänilähteitä toisistaan. Tätä sovellusta voidaan käyttää esimerkiksi erilaisiin robotteihin. [15]

Työssä käsitelläänkin atsimuuttidataa, jonka avulla pyritään määrittämään äänilähteen paikka. Signaalia on vastaanotettu 360° ympäri pyörivällä laitteella ja data kuvaa äänilähteen koordinaatteja. Datan arvot heilahtelevat arvojen $-\pi$ ja π välillä. Kuvassa 4.1 on esitetty data.



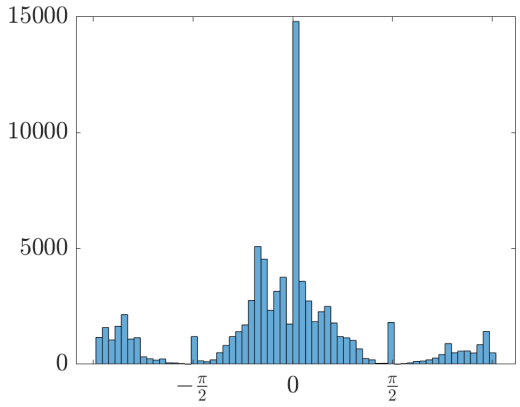
Kuva 4.1. *Alkuperäinen atsimuuttidata.*

Työn tarkoitus on muodostaa kahden esitellyn jakauman avulla yhdiste ja sovittaa se dataan. Jakaumien yhdiste kahdelle eri jakaumalle on muotoa

$$f(t) = \eta_1 f_1(t) + \eta_2 f_2(t), \quad \eta_1,\ \eta_2 \in [0,1], \tag{4.1}$$

missä η_1 ja η_2 ovat tiheysfunktioiden $f_1(t)$ ja $f_2(t)$ painotuskertoimet [5].

Kuvassa 4.2 on piirrettynä histogrammi annetusta datasta.

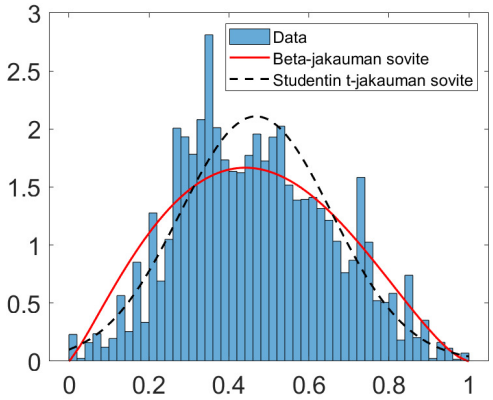


Kuva 4.2. *Histogrammi atsimuuttidatasta.*

Histogrammista voidaan huomata kolme osaa, joihin datan arvot ovat painottuneet. Datasta huomataan myös kolme selkeää piikkiä kohdissa $-\frac{\pi}{2}$, 0 ja $\frac{\pi}{2}$. Näitä piikkejä ei kuitenkaan huomioida työssä tarkastelun yksinkertaistamisen vuoksi. Koska äänilähde on kiertänyt laitetta ympäri, reunimmaiset kuvut ovat oikeastaan samaa kupua. Voidaankin siis lähteä käsittelemään dataa kahdessa osassa. Ensimmäinen osa koostuu välillä $(-\frac{\pi}{2}, \frac{\pi}{2})$ olevasta datasta. Toinen osa muodostuu, kun yhdistetään väleillä $[-\pi, -\frac{\pi}{2})$ ja $(\frac{\pi}{2}, \pi]$ oleva data siten, että siirretään arvosta $-\pi$ alkava data arvoon π päättyvän datan perään.

Dataa käsitellään Matlab-ohjelmalla. Liitteestä A löytyvää Matlab-koodia käytetään työn tekemiseen. Molemmat datan osat skaalataan avoimelle välille $(0, 1)$, jotta Beta-jakauman sovittaminen onnistuu. Skaalauksen jälkeen halutulle jakaumalle etsitään dataan sopivat parametrit käyttäen suurimman uskottavuuden menetelmää, joka esiteltiin kolmannessa luvussa. Saatujen parametrien avulla sovitteelle muodostetaan haluttu tiheysfunktio.

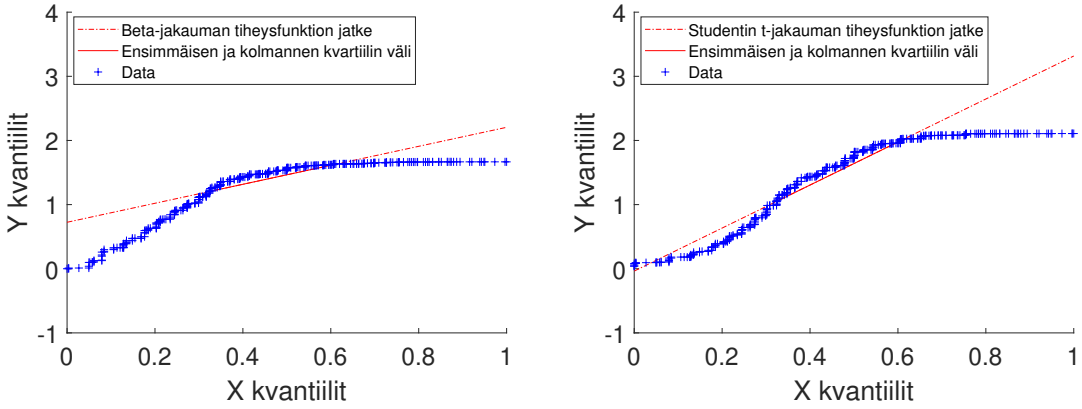
Datan keskiosaan on sovitettu sekä Studentin t-jakauman että Beta-jakauman tiheysfunktiot. Kuvassa 4.3 on datan keskiosa avoimelle välille $(0, 1)$ skaalattuna sekä siihen sovitetut Beta- ja Studentin t-jakauman tiheysfunktiot.



Kuva 4.3. *Datan keskiosa ja siihen tehdyt sovitteet Beta- ja Studentin t-jakaumalle.*

Silmämääräisesti kuvasta on hankala arvioida kumpi jakauma sopii datan osaan paremmin. Asiaa voidaan tutkia tekemällä datasta ja halutusta jakaumasta Quantile-quantile plot. Quantile-quantile

plot vertailee, kuinka hyvin datan ja tiheysfunktion kvantiilit vastaavat toisiaan. [16] Kuvassa 4.4 on esitettynä datan keskiosalle Quantile-quantile plot sekä Beta-jakaumalla että Studentin t-jakaumalla.

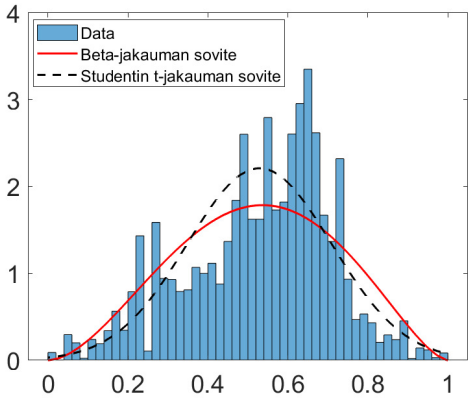


(a) *Beta-jakaumalle* ***(b)*** *Studentin t-jakaumalle*

Kuva 4.4. *Quantile-quantile plot datan keskiosalle.*

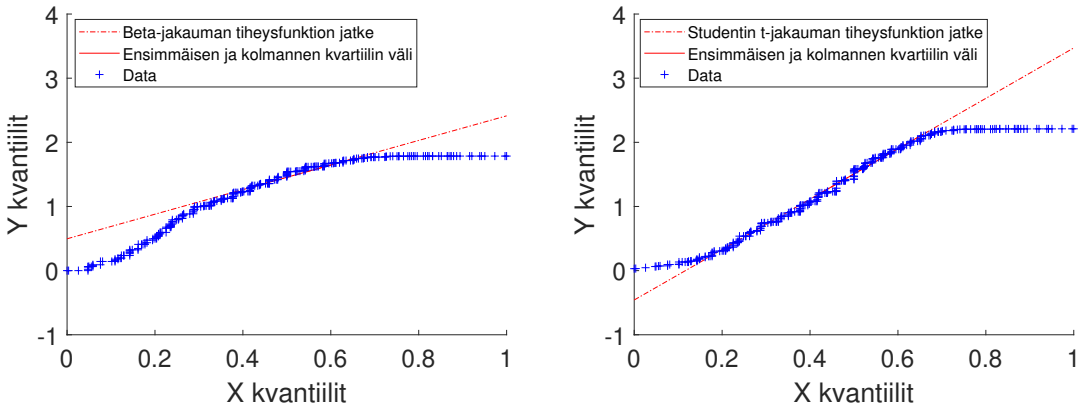
Kummassakin kuvassa huomataan poikkeamaa. Data näyttää kuitenkin seuraavan Studentin t-jakaumaa hieman paremmin pidemmän matkan ajan. Yhdistetyn tiheysfunktion muodostamiseen käytetään siis datan keskiosaan sovitettua Studentin t-jakauman tiheysfunktiota.

Sama tarkastelu on tehty myös datan reunaosille. Kuvassa 4.5 on datan reunaosat yhdistettynä ja skaalattuna avoimelle välille $(0, 1)$ sekä siihen sovitettu Beta-jakauman ja Studentin t-jakauman tiheysfunktiot.



Kuva 4.5. *Datan reunaosat yhdistettynä ja siihen tehdyt sovitte Beta- ja Studentin t-jakaumalle.*

Kuvasta on hankala arvioida, kumpi jakauma sopii datan osaan paremmin. Tässä kohtaa on myös tehty Quantile-quantile plot datalle kummallakin jakaumalla, ja ne on esitetty kuvassa 4.6.



(a) *Beta-jakaumalle* ***(b)*** *Studentin t-jakaumalle*

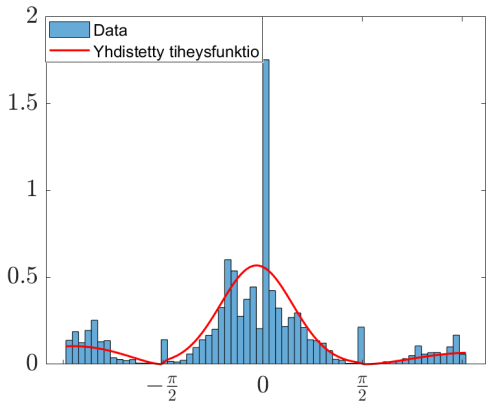
Kuva 4.6. *Quantile-quantile plot datan reunaosille.*

Kuvien perusteella data seuraa kumpaakin tiheysfunktiota melko hyvin. Kuitenkin kummassakin tapahtuu poikkeamaa sekä oikeassa että vasemmassa reunassa. Valitaan käyttöön Beta-jakauma, koska tapahtunut poikkeama on kyseisen jakauman kohdalla pienempää.

Yhdistetty tiheysfunktio koko datalle muodostetaan kaavan (4.1) mukaisesti. Nyt $f_1(t)$ on datan keskiosaan sijoitettu Studentin t-jakauman tiheysfunktio (2.7) ja $f_2(t)$ datan reunoihin sijoitettu Beta-jakauman tiheysfunktio (2.1). Painotuskertoimet saadaan jakamalla halutulla osuudella olevien datan arvojen lukumäärä koko datan arvojen lukumäärällä. Lopullinen tiheysfuntio on muotoa

$$
\begin{aligned}
f(t) &= 0,60 \cdot f_T(t; 3,41 \cdot 10^6) + 0,21 \cdot f_\beta(t; 2,86; 2,61) \\
&= 0,60 \cdot \frac{\Gamma\left(\frac{3,41\cdot 10^6+1}{2}\right)}{(\pi \cdot 3,41 \cdot 10^6)^{\frac{1}{2}} \Gamma\left(\frac{3,41\cdot 10^6}{2}\right)\left(1+\frac{t^2}{3,41\cdot 10^6}\right)^{\frac{1}{2}(3,41\cdot 10^6+1)}} \\
&\quad + 0,21 \cdot \frac{\Gamma(2,86+2,61)}{\Gamma(2,86)\Gamma(2,61)} t^{2,86-1}(1-t)^{2,61-1} \\
&= 0,60 \cdot \frac{\Gamma\left(1,71 \cdot 10^6\right)}{3,27 \cdot 10^3 \cdot \Gamma\left(1,71 \cdot 10^6\right)\left(1+\frac{t^2}{3,41\cdot 10^6}\right)^{1,71\cdot 10^6}} \\
&\quad + 0,21 \cdot \frac{\Gamma(5,47)}{\Gamma(2,86)\Gamma(2,61)} t^{1,86}(1-t)^{1,61}.
\end{aligned}
$$

Saatu tiheysfunktio piirretään kuvan 4.2 histogrammin kanssa samaan kuvaan 4.7.



Kuva 4.7. *Histogrammi datasta ja siihen sovitettu tiheysfunktio.*

Tiheysjakauma on normalisoitu noudattamaan datan histogrammia, kun piikit jätetään huomiotta. Kuvasta huomataankin, että data seuraa hyvin muodostettua tiheysfunktiota. Tästä voidaan päätellä, että jakauman sovittamien dataan on onnistunut.

Seuraava askel datan tutkinnassa olisi lähteä tutkimaan, miten jakauman sovittaminen onnistuisi dataan ilman, että sitä tarvitsisi jakaa osiin, kuten työssä on tehty. Eräs tapa voisi olla se, että muodostettaisiin haluttu tiheysfunktio (4.1) ensin. Toisin sanoen, etsittäisiin ensin halutut tiehysfunktiot ja niille sopivat painotuskertoimet. Näistä muodostettaisiin sitten yhdistetty tiheysfunktio, jolle lähdettäisiin kokonaisuudessaan etsimään sopivia parametreja suurimman uskottavuuden estimoinnin avulla. Tämä on kuitenkin haastavampi tapa lähteä sovittamaan jakaumaa, eikä sitä toteuteta tässä työssä.

5 YHTEENVETO

Tässä työssä tutustuttiin aluksi Beta-jakauman ja Studentin t-jakauman ominaisuuksiin ja sen jälkeen suurimman uskottavuuden menetelmään. Näiden jälkeen lähdettiin muodostamaan todennäköisyysjakaumaa käytössä olevaan atsimuuttidataan. Data ei kokonaisuudessaan noudattanut mitään todennäköisyysjakaumaa, joten sen tiheysfunktio muodostettiin eri tiheysfunktioiden yhdistelmänä.

Beta- ja Studentin t-jakaumasta esiteltiin molemmista tiheysfunktio, varianssi sekä odotusarvo. Lisäksi käytiin lyhyesti läpi, mitä ovat Beta- ja Gamma-funktio. Beta-jakauman odotusarvo ja varianssi todistettiin ensin lähtien liikkeelle odotusarvon ja varianssin määritelmistä. Näille toteutettiin myös vaihtoehtoinen todistus hyödyntäen momentit generoivaa funktiota. Jotta edellä mainituista todennäköisyysjakaumista pystyttiin muodostamaan dataan sopivia jakaumia, työssä esiteltiin suurimman uskottavuuden menetelmä. Tämän menetelmän avulla jakaumille pystyttiin määrittämään dataan sopivat parametrit.

Kuten edellä on mainittu, käytössä oleva atsimuuttidata ei noudattanut mitään todennäköisyysjakaumaa. Tästä syystä data jaettiin sopiviin osiin, joihin kuhunkin sovitettiin sitä parhaiten kuvaavan jakauman tiheysfunktio. Lopullinen tiheysfunktio datalle muodostettiin summaamalla yhteen osiin sopivat tiheysfunktiot painotuskertoimilla kerrottuna. Painotuskertoimet saatiin jakamalla halutulla osuudella olevien datan arvojen lukumäärä koko datan arvojen lukumäärällä. Lopputulokseksi saatiin melko hyvin dataa kuvaava tiheysfunktio.

LÄHTEET

[1] E. W. Packel. *The mathematics of games and gambling*. 2nd ed. Anneli Lax new mathematical library ; v. 28. Washington, DC: Mathematical Association of America, (2006).

[2] B. M. Ayyub ja R. H. McCuen. *Probability, Statistics, and Reliability for Engineers and Scientists*. Bosa Roca: CRC Press, (2011).

[3] N. Balakrishnan ja C.-D. Lai. *Continuous Bivariate Distributions*. 2. Aufl.;2nd; US: Springer-Verlag, (2009).

[4] N. L. Johnson ja N. Kotz Samuel Balakrishnan. *Continuous Univariate Distribution, Volume 2*. A Wiley-Interscience Publication, (1995).

[5] C. Forbes, M. Evans, N. Hastings ja B. Peacock. *Statistical distributions*. 4th ed. Hoboken, N.J: Wiley, (2011).

[6] F. J. Havil Julian Dyson. The Gamma Function. *Gamma: Exploring Euler's Constant*. Princeton; Oxford: Princeton University Press, (2010).

[7] G. Upton ja I. Cook. *symmetric distribution*. (2014). URL: http://www.oxfordreference.com/view/10.1093/acref/9780199679188.001.0001/acref-9780199679188-e-1607 (viitattu 28. 01. 2020).

[8] S. Nadarajah, S. Chan ja E. Afuecheta. On the characteristic function for asymmetric Student t distributions. *Economics Letters* 121.2 (2013), 271–274.

[9] M. Ahsanullah ja V. B. Nevzorov. Some Inferences on Skew-t Distribution of 2 Degrees of Freedom. *Journal of Statistical Theory and Applications (JSTA)* 16.4 (2017). URL: https://doaj.org/article/260219ff3dee4e6bb74cd7e4892138a9 (viitattu 07. 01. 2020).

[10] I. J. Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology* 47.1 (2003), 90–100.

[11] R. E. Walpole. *Probability statistics for engineers scientists*. 9th edition, global edition. Boston, Mass: Pearson, (2016).

[12] O. Barndorff-Nielsen. *Information and exponential families : in statistical theory*. 2nd ed. Wiley Series in Probability and Statistics. Chichester, England: John Wiley Sons, (2014).

[13] A. Ziegler. *Generalized Estimating Equations*. 1st ed. 2011. Lecture Notes in Statistics, 204. New York, NY: Springer New York, (2011).

[14] R. Duan, K. Yang, Y. Ma, Q. Yang ja H. Li. Moving source localization with a single hydrophone using multipath time delays in the deep ocean. *The Journal of the Acoustical Society of America* 136.2 (2014), 159–165.

[15] M. Lager. *Audio Source Positioning Based on Angle of Arrival Measurements*. eng. Tampere University, (2020).

[16] G. J. Myatt. *Making sense of data II a practical guide to data visualization, advanced data mining methods, and applications*. Hoboken, N.J: John Wiley Sons, (2009).

A MATLAB-KOODI JAKAUMIEN SOVITTAMISEEN

```
1  % Ladataan data excelistä ja muutetaan matriisiksi.
2  table = readtable('azimuth.xlsx');
3  data = table{:,:};
4
5  % Muutetaan data matriisista vektoriksi ja skaalataan avoimelle välille (0,1).
6  vdata = scaling(data(:));
7  % Datan jakamiseen käytetyt rajat.
8  limits = [0.76 inf -inf 0.24;0.26 0.49 0.5 0.74];
9  dist ={'Beta','tLocationScale'};% Tutkittavien jakaumien nimet.
10
11 % Otetaan datan osa ja skaalataan se halutulle välille. Skaalattuun datan
12 % osaan sovitetaan molemmat jakaumat fitdist-komennon avulla käyttäen
13 % suurimman uskottavuuden menetelmää [Luku 3].
14 for ii=1:size(limits,1)
15     for jj=1:numel(dist)
16         % Haluttu datan osa.
17         d = scaling([scaling(vdata(limits(ii,1)< vdata & vdata <
               limits(ii,2)));...
18             scaling(vdata(limits(ii,3)< vdata & vdata < limits(
                   ii,4)))+1]);
19         a = numel(d)/numel(vdata) % Datan osan painotuskerroin.
20         fitdist(d,dist{jj}) % Etsitään halutulle datan osalle jakaumaan sopivat
               parametrit.
21     end
22 end
23
24 function data = scaling(data)
25 % Funktio, joka skaalaa annetun datan arvot avoimelle välille (0,1).
26 % Ottaa vastaan parametrinä skaalattavan datan ja palauttaa datan
27 % skaalattuna.
28     alpha = 0.000001;
29     data = (1-alpha)*(data-min(data))./(max(data-min(data)))+(
           alpha/2);
30 end
```