Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Birgit Kadastik

Magnetresonantstomograafia (MRT) ja kompuutertomograafia (KT) andmete jaotuste võrdlemine

Matemaatika ja statistika eriala

Magistritöö (30 EAP)

Juhendaja Kristi Kuljus

Tartu 2018

Magnetresonantstomograafia (MRT) ja kompuutertomograafia (KT) andmete jaotuste võrdlemine

Magistritöö

Birgit Kadastik

Lühikokkuvõte. Magnetresonantstomograafia (MRT) ja kompuutertomograafia (KT) on kaks erinevat meetodit, mida kasutatakse inimkehast kihiliste ja ruumiliste kujutuste saamiseks meditsiinilises diagnostikas. Käesoleva magistritöö eesmärgiks on uurida ning võrrelda MRT ja KT andmete jaotusi erinevatele patsientidele kuuluvate peade korral. Esimeses peatükis seletatakse rakendatavaid statistilisi meetodeid koos sobivate näidetega. Teises peatükis leitakse K-keskmiste meetodi ja Gaussi segumudelite tulemuste abil Kullback-Leibleri kaugused, mida kasutatakse peadevaheliste jaotuste võrdlemiseks.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Märksõnad: K-keskmised, Kullback-Leibleri kaugus, Gaussi segumudelid, mitmemõõtmeline analüüs, klasterdamine, kompuutertomograafia, magnetresonantstomograafia

Comparing distributions of magnetic resonance imaging (MRI) and computed tomography (CT) measurements data

Master's thesis

Birgit Kadastik

Abstract. Magnetic resonance imaging (MRI) and computed tomography (CT) are two different methods that are used in medicine to image the anatomy of a human body. The aim of this Master's thesis is to investigate the differences in MRI and CT measurements from different patients. The data available are head measurements. Firstly, the statistical methods used are explained with appropriate examples. In the second part of the thesis the results of K-means and Gaussian mixture models are used to calculate the Kullback-Leibler divergences which are used for comparing the distributions.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and acturial mathematics

Keywords: K-means, Kullback-Leibler divergence, Gaussian mixture models, multivariate analysis, clustering, computed tomography, magnetic resonance imaging

Sisukord

Sissejuhatus

Magnetresonantstomograafia (MRT) ja kompuutertomograafia (KT) on kaks erinevat meetodit, mida kasutatakse inimkehast kihiliste ja ruumiliste kujutuste saamiseks meditsiinilises diagnostikas. Kompuutertomograafia põhineb röntgenkiirgusel, magnetresonantstomograafia aga magnetismil. MRT kujutiselt on parem eristada pehmet kudet ning seetõttu eelistatakse seda kasvajate identifitseerimiseks, samas aga on luu ja õhu eristamine raskendatud. Kompuutertomograafia on ajaliselt kiirem ning annab parema kujutuse luust, kuid protseduuri miinuseks on omandatud kiirgus, mida on seostatud kõrgendatud vähiriskiga.

Antud magistritööga sama andmestikku on kasutatud varasemas modelleerimisülesandes, mille eesmärk oli KT piltide prognoosimine MRT piltide põhjal [1]. Tulemused olid, et teatud pead käitusid modelleerimisel teistest väga erinevalt. Peade käitumise põhjal moodustus kaks gruppi. Esimesse gruppi kuuluvad pead olid mudeli jääkide käitumise põhjal homogeensed. Samas teise grupi pead erinesid homogeense grupi peadest ning erinevusi võis täheldada ka grupisiseselt. Nendest järeldustest tulenevalt on käesoleva magistritöö eesmärgiks uurida, kas erinevusi peade vahel on võimalik tuvastada MRT ja KT mõõtmistulemuste jaotuste võrdlemisel.

Töö koosneb kahest peatükist. Esimeses osas kirjeldatakse antud töös kasutatavaid statistilisi meetodeid, mida on illustreeritud sobivate näidetega. Teises peatükis kirjeldatakse andmeid ja tuuakse välja metoodika probleemi uurimiseks. Peade andmetele rakendatakse K-keskmiste meetodil põhinevat klasteranalüüsi ja leitakse igale peale sobivad Gaussi segumudelid. Mõlema meetodi tulemusi kasutatakse peadevaheliste jaotuste võrdlemiseks Kullback-Leibleri kauguse abil.

Töö on kirjutatud tekstitöötlusprogrammiga Microsoft Office Word 2016 ning analüüs on läbi viidud rakendustarkvaras R (versioon 3.3.2).

Autor tänab juhendajat Kristi Kuljust arvukate konsultatsioonide ja heade soovituste eest.

1. Statistilised meetodid

1.1 K-keskmiste meetod

K-keskmiste meetod on lisaks hierarhilisele klasterdamisele üks klasteranalüüsi meetoditest, mis üritab leida andmete seast alamgruppe ehk teisisõnu klastreid. Eesmärk on leida sellised klastrid, et vaatlused oleksid klastrisiseselt võimalikult sarnased, aga erinevad võrreldes teiste klastrite vaatlustega. K-keskmiste meetodi korral võrreldakse klastritesse paigutamisel Eukleidilist kaugust.

K-keskmiste meetod jagab vaatlused *K* erinevasse, mittekattuvasse klastrisse. Klastrite arv antakse algoritmile ette. Kasutades keskpunktide alglähendeid, põhineb K-keskmiste meetod kahel sammul, mida korratakse koondumiseni:

1. leitakse vaatluse kaugus iga klastri keskpunktist ja paigutatakse punkt klastrisse, mille korral on kaugus kõige väiksem,
2. arvutatakse uued keskpunktid.

Alglähendid valitakse algoritmis juhuslikult või antakse kasutaja poolt ette enne algoritmi rakendamist. [2]

Klasterdamisel on võimalike kombinatsioonide arv, kuidas vaatlusi *K* klastrisse paigutada, suur. Tähistagu *n* vaatluste arvu andmestikus. Kombinatsioonide arvu on võimalik arvutada valemiga

$$S(n,K) = \frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\binom{K}{k}k^n.$$

Juhul kui andmestikus on näiteks 9 vaatlust ja tahame need jaotada 3 gruppi, siis on võimalike kombinatsioonide arv 3025. Vaadates aga näiteks 16 vaatlusega andmestikku, kasvab võimalike kombinatsioonide arv $S(16,3) = 7\,141\,686$ peale. Seepärast vaatavad klasterdamise algoritmid väiksemat kombinatsioonide hulka ning valivad nende seast välja kõige parema. [2] Oluline on arvestada, et klasterdamise lõpptulemus sõltub alglähenditest ja seetõttu peab algoritmi rakendama mitmete erinevate alglähenditega. Üheks K-keskmiste meetodi miinuseks on ka klastrite arvu etteandmine. Klastrite arv määratakse tavaliselt kas eelnevate teadmiste põhjal või kasutatakse sobivaid statistikuid, näiteks Gap-statistikut või küünarnuki meetodit. [3] [2]

Olgu meil andmestik $\{x_1, \dots, x_n\}$, kus on *n* vaatlust ja iga vaatluse korral on mõõdetud *p* tunnust, seega x_i: $p \times 1$, $i = 1, \dots, n$. Eesmärk on paigutada iga vaatlus klastrisse nii, et klastrisiseste punktide kaugused on väikesed võrreldes teistes klastrites asuvate punktide kaugustega. Olgu vektor μ_k: $p \times 1$, $k = 1, \dots, K$, klastri k keskpunkt. Tähistagu $x_i \in k$, et vaatlus x_i kuulub klastrisse k. Peame leidma klastrite keskpunktid μ_k ja paigutama vaatlused klastritesse nii, et vaatluste kauguste ruutude summa oma klastri keskpunktist on minimaalne. [4]

Antud töös vaatleme K-keskmiste leidmiseks klassikalist ehk Lloydi algoritmi ning mõnevõrra kiiremat Hartigan-Wongi algoritmi.

1.1.1 Lloydi K-keskmiste algoritm

Selles peatükis näitame ära formaalselt, et parameetrite μ_k hinnanguks on vastava klastri keskpunkt. Antud peatükk põhineb [4], kui ei ole mainitud teisiti.

Olgu iga vaatluse x_i jaoks defineeritud binaarne tunnus $r_{ik} \in \{0,1\}$ nii, et

$$r_{ik} = \begin{cases} 1, & \text{kui } x_i \in k \\ 0, & \text{muidu} \end{cases}, \qquad k = 1, \dots, K.$$

Tähistagu *J* järgmist kauguste ruutude summat

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \|x_i - \mu_k\|^2.$$

Eesmärgiks on leida r_{ik} ja μ_k väärtused nii, et *J* oleks minimaalne. Seda saab teha kahesammulise iteratiivse protseduuriga. Esimeses sammus hoitakse μ_k väärtused fikseeritud ja leitakse optimaalsed r_{ik} väärtused. Teises sammus aga kasutatakse eelmise sammu r_{ik} väärtusi ning viiakse läbi minimiseerimine μ_k suhtes. Neid kahte sammu korratakse koondumiseni.

Antud μ_k väärtuste korral on selge, et vähim *J* väärtus saavutatakse, kui vaatlus x_i paigutatakse klastrisse, mille korral on vaatluse ja klastri keskpunkti vaheline kaugus kõige väiksem:

$$r_{ik} = \begin{cases} 1, & \text{kui } k = arg\,min_j \left\|x_i - \mu_j\right\|^2, \\ 0, & \text{mujal.} \end{cases}$$

Kui r_{ik} on fikseeritud, tuleb iga klastri keskpunkti μ_k leidmiseks lahendada võrrand

$$2 \sum_{i=1}^{n} r_{ik}(x_i - \mu_k) = 0, \qquad \text{seega} \quad \mu_k = \frac{\sum_i r_{ik} x_i}{\sum_i r_{ik}}, \qquad k = 1, \dots, K.$$

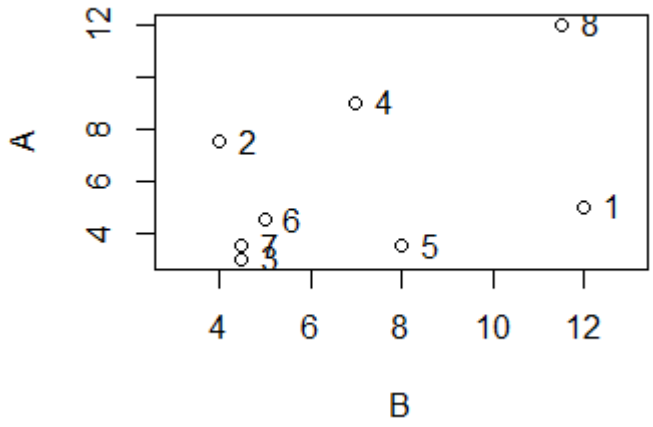
Kuna $\sum_i r_{ik}$ annab klastri *k* elementide arvu, siis on μ_k võrdne klastri *k* vaatluste keskmisega ehk $\mu_k = \bar{x}_k$. Seetõttu kutsutaksegi antud meetodit K-keskmiste meetodiks.

Punktide ümberpaigutamist ja keskmiste arvutamist korratakse kuni ümberpaigutamine funktsiooni *J* väärtust enam ei kahanda või maksimum arv iteratsioone on läbi tehtud.

Järgmisena toome lihtsa näite Lloydi algoritmi rakendamisest. Sama näite abil demonstreerime hiljem, et Hartigan-Wongi algoritm jõuab samade alglähendite korral lahendini kiiremini.

Näide 1.1

Olgu meil andmestik, kus on 2 tunnust A ja B ning andmed 8 inimese jaoks (joonis 1). Tunnuste väärtused on isikute jaoks vastavalt (5; 12), (7,5; 4), (3; 4,5), (9; 7), (3,5; 8), (4,5; 5), (3,5; 4,5) ja (12; 11,5).



Joonis 1. Näiteandmestiku hajuvusdiagramm

Soovime jagada andmed kolme klastrisse. Valime iga klastri jaoks juhuslikult esialgsed keskpunktid: esimese klastri keskpunktideks on (9; 7), teise ja kolmanda klastri keskpunktid on vastavalt (3,5; 8) ning (7,5; 4).

Järgnevalt arvutame iga isiku jaoks Eukleidilise kauguse ruudu iga klastri keskpunktidest ja klasterdame isikud kõige väiksema kauguse järgi. Pärast esimest klasterdamist arvutame uued keskmiste vektorid. Uued klastrid ja nende keskpunktid on toodud tabelis 1 tulbas iteratsioon 1.

Tabel 1. Klastrid ja nende keskpunktid kolme iteratsiooni jaoks Lloydi algoritmi korral

	Iteratsioon 1		Iteratsioon 2		Iteratsioon 3	
Klaster	Isikud	Keskpunktid	Isikud	Keskpunktid	Isikud	Keskpunktid
1	4, 8	(10,5; 9,25)	4, 8	(10,5; 9,25)	4, 8	(10,5; 9,25)
2	1, 3, 5, 7	(3,75; 7,25)	1, 3, 5	(3,83; 8,17)	1, 5	(4,25; 10)
3	2, 6	(6; 4,5)	2, 6, 7	(5,17; 4,5)	2, 3, 6, 7	(4,63; 4,5)

Pärast uute keskpunktide arvutamist teeme taaskord läbi isikute klastritesse paigutamise Eukleidilise kauguse ruudu põhjal ja arvutame välja uute klastrite keskpunktid (tabel 1, iteratsioon 2). Kasutame järgmisena viimasena leitud keskpunkte ja viime läbi klasterdamise kolmandat korda, tulemused on toodud tabelis 1 tulbas iteratsioon 3.

Järgmisel sammul ümberpaigutamist ei toimu ja seega oleme leidnud lahendi.

1.1.2 Hartigan-Wongi K-keskmiste algoritm

Nagu ikka K-keskmiste meetodi korral, jagab ka Hartigan-Wongi algoritm vaatlused klastritesse nii, et klastrisisene kauguste ruutude summa ei muutu väiksemaks, kui paigutada vaatlus ühest klastrist teise, see tähendab leitakse lokaalne optimum. Antud peatükk põhineb [5].

Tähistame klastri k elementide arvu tähisega n_k ning punkti x_i kaugust klastri k keskpunktist $d(x_i, \bar{x}_k)$. Lihtsustamise mõttes tähistame praegu kauguse ruudu d_k^2 (x_i on fikseeritud), mis esitub valemiga $d_k^2 = (x_i - \bar{x}_k)^T(x_i - \bar{x}_k)$.

Klassikalise K-keskmiste meetodi korral võrreldakse vaatluse kaugust erinevate klastrite keskpunktidest. Kui punkt x_i kuulus eelnevalt klastrisse k, aga $d_l^2 < d_k^2$ mingi $l \neq k$ korral, siis paigutatakse x_i ümber klastrisse l. Hartigan-Wongi meetodi korral on aga kasutusele võetud efektiivsem kriteerium ümberpaigutamise jaoks, nimelt võetakse kauguste võrdlemisel arvesse uusi võimalikke keskpunkte. Seega paigutatakse vaatlus x_i klastrist k ümber klastrisse l, kui

$$\frac{n_l}{n_l+1} d_l^2 < \frac{n_k}{n_k-1} d_k^2. \qquad (1.1)$$

Vaatame, millised on vaatluse x_i kaugused klastrite keskpunktidest, kui tõstame punkti x_i klastrist k klastrisse l. Klastrite kaugused pärast ümberpaigutamist on tähistatud vastavalt $d_{l,uus}^2$ ja $d_{k,uus}^2$. Kauguste leidmiseks arvutame esmalt klastri l uue keskpunkti:

$$\bar{x}_{l,uus} = \frac{1}{n_l + 1} \sum_{x_j \in l} x_j = \frac{n_l \bar{x}_l + x_i}{n_l + 1}.$$

Paneme tähele, et

$$\left(x_i - \bar{x}_{l,uus}\right) = \left(x_i - \frac{n_l \bar{x}_l + x_i}{n_l + 1}\right) = \left(\frac{n_l x_i + x_i - n_l \bar{x}_l - x_i}{n_l + 1}\right) = \frac{n_l}{n_l + 1}(x_i - \bar{x}_l).$$

Kuna $d_{l,uus}^2 = \left(x_i - \bar{x}_{l,uus}\right)^T \left(x_i - \bar{x}_{l,uus}\right)$, siis saame

$$d_{l,uus}^2 = \left(\frac{n_l}{n_l + 1}\right)^2 d_l^2.$$

Suuruse $d_{k,uus}^2$ leiame sarnaselt ülaltooduga, erinevus on see, et keskmine arvutatakse nii, et vaatlus on eemaldatud klastrist *k*, seega $\bar{x}_{k,uus} = \frac{n_k \bar{x}_k - x_i}{n_k - 1}$ ja $\left(x_i - \bar{x}_{k,uus}\right) = \frac{n_k}{n_k - 1}(x_i - \bar{x}_k)$.

Arvestame, et $\left(\frac{n_l}{n_l + 1}\right) < 1$, siis, kui kehtib (1.1), saame

$$d_{l,uus}^2 = \left(\frac{n_l}{n_l + 1}\right)^2 d_l^2 < \frac{n_l}{n_l + 1} d_l^2 < \frac{n_k}{n_k - 1} d_k^2 < \left(\frac{n_k}{n_k - 1}\right)^2 d_k^2 = d_{k,uus}^2.$$

Seega, kui kehtib (1.1), siis kehtib ka $d_{l,uus}^2 < d_{k,uus}^2$ ehk kasutades (1.1) on garanteeritud, et ümberpaigutamise korral on punkti x_i kaugus klastri *l* uuest keskpunktist väiksem.

Kirjeldame järgnevalt Hartigan-Wongi algoritmi. Kõigepealt valitakse klastrite arv ja määratakse algsed klastrite keskpunktid.

Algoritm:

1. Leitakse vaatlusele x_i kaks kõige lähemat klastrit, see tähendab need klastrid, mille korral keskpunktide ja vaatluse vahelised Eukleidiliste kauguste ruudud on kõige väiksemad. Olgu need tähistatud vastavalt $C_1(x_i)$ ja $C_2(x_i)$, punkt x_i paigutatakse klastrisse $C_1(x_i)$. Antud samm tehakse läbi kõikide punktide jaoks.
2. Uuendatakse klastrite keskpunkte, leides iga klastri vaatluste keskmise.
3. Algselt kuuluvad kõik klastrid aktiivsesse (inglise keeles *live*) hulka.
4. See on nn „optimaalse üleviimise“ (inglise keeles *optimal-transfer, OPTRA)* staadium. Vaadatakse järjest kõiki vaatluseid x_i, $i = 1,2,\dots,n$. Kui klastrit l ($l = 1,2,\dots,K$) uuendati viimases *QTRAN* staadiumis (vaata samm 6), siis kuulub see klaster aktiivsesse hulka terve selle staadiumi vältel. Kui klastrit *l* ei ole viimase *n* sammu jooksul

uuendatud, siis see klaster aktiivsesse hulka ei kuulu. Olgu punkt x_i klastris l_1. Kui l_1 on aktiivses hulgas, siis minnakse sammu (a) juurde, vastasel juhul sammu (b) juurde.

(a) Arvutatakse

$$R_2^{(l)} = \frac{n_l}{n_l + 1} d^2(x_i, \bar{x}_l) \qquad (1.2)$$

üle kõikide klastrite l ($l \neq l_1, l = 1,2,\dots,K$). Olgu l_2 klaster, millel on kõige väiksem $R_2^{(l)}$. Juhul kui $R_2^{(l_2)} \geq \frac{n_{l_1}}{n_{l_1}-1} d^2\left(x_i, \bar{x}_{l_1}\right)$, siis ümberpaigutamist ei toimu ja l_2 on uus $C_2(x_i)$. Kui $R_2^{(l_2)} \leq \frac{n_{l_1}}{n_{l_1}-1} d^2\left(x_i, \bar{x}_{l_1}\right)$, siis punkt x_i pannakse klastrisse l_2 ning l_1 on uus $C_2(x_i)$. Keskpunktid arvutatakse uuesti juhul, kui ümberpaigutamine toimus. Kaks klastrit, mida kasutati, on nüüd aktiivses hulgas.

(b) See samm on sama, mis samm (a), aga $R_2^{(l)}$ arvutatakse ainult nende klastrite *l* jaoks, mis asuvad aktiivses hulgas.

5. Algoritm lõpetab töö, kui aktiivne hulk on tühi. Muidu minnakse pärast andmestiku läbimist sammu 6 juurde.
6. See on nn „kiire üleviimise“ (inglise keeles *quick-transfer, QTRAN)* staadium. Vaadatakse järjest kõiki punkte x_i, $i = 1,2,\dots,n$. Olgu $l_1 = C_1(x_i)$ ja $l_2 = C_2(x_i)$. Juhul, kui l_1 ja l_2 pole viimase *n* sammu jooksul muutunud, siis liigutakse edasi järgmise vaatluse juurde. Arvutatakse $R_1^{(l_1)} = \frac{n_{l_1}}{n_{l_1}-1} d^2(x_i, \bar{x}_{l_1})$ ja $R_2^{(l_2)} = \frac{n_{l_2}}{n_{l_2}+1} d^2(x_i, \bar{x}_{l_2})$.

   Kui $R_1^{(l_1)} \leq R_2^{(l_2)}$, siis punkt x_i jääb klastrisse l_1. Vastasel juhul vahetatakse $C_1(x_i)$ ja $C_2(x_i)$ omavahel ära ning uuendatakse klastrite l_1 ja l_2 keskpunkte. Kaks klastrit on nüüd märgitud üleviimise sammus.
7. Kui viimase *n* sammu jooksul ühtegi ümberpaigutamist ei toimunud, siis liigutakse sammu 4 juurde, muidu aga sammu 6 juurde.

Kui enam punktide ümberpaigutamist ei toimu, on leitud lahend ning algoritm lõpetab töö.

Kuigi Hartigan-Wongi algoritm on kiirem, siis väga sarnaste andmete korral ei pruugi algoritm koonduda, see tähendab, et aktiivne hulk ei saa kunagi tühjaks. Üheks võimalikuks lahenduseks on väärtuste ümardamine.

Näide 1.2

Järgnevalt demonstreerime Hartigan-Wongi algoritmi sama näiteandmestiku (joonis 1) põhjal, valime ka samad klastrite keskpunktide alglähendid. Arvutame iga vaatluse jaoks kaugused

klastrite keskpunktidest, paneme kirja kaks kõige väiksema kaugusega klastrit (tabel 2) ning määrame vaatluse klastrisse $C_1(x_i)$ ja arvutame seejärel uued keskpunktid.

Tabel 2. Kõige väiksemate kaugustega klastrid

Isik	1	2	3	4	5	6	7	8
$C_1(x_i)$	2	3	2	1	2	3	2	1
$C_2(x_i)$	1	1	3	3	1	2	3	3

Optimaalse üleviimise sammus vaatame iga isikut eraldi. Arvutame isiku 1 jaoks suuruse (1.2) klastri 1 ja klastri 3 korral ning suuruse $R_1^{(l_1)}$ klastri 2 korral. Need on vastavalt 25,2; 38,2 ja 32,2, seega kriteeriumi (1.1) põhjal tõstame vaatluse 1 ümber klastrist 2 klastrisse 1 ja arvutame uuesti klastrite 1 ja 2 keskpunktid (tabel 3).

Tabel 3. Klastrid ja nende keskpunktid pärast esimese isiku ümberpaigutamist

Klaster	Isikud	Keskpunktid
1	1, 4, 8	(8,67; 10,27)
2	3, 5, 7	(3,33; 5,67)
3	2, 6	(6; 4,5)

Isiku 2 jaoks kasutame kriteeriumite arvutamisel keskmistena tabeli 3 tulemusi. Optimaalse üleviimise sammus paigutasime ümber isikud 1, 4 ning 6 ning uued lähimad klastrid on toodud tabelis 4.

Tabel 4. Kõige väiksemate kaugustega klastrid pärast OPTRA sammu

Isik	1	2	3	4	5	6	7	8
$C_1(x_i)$	1	3	2	3	2	2	2	1
$C_2(x_i)$	2	2	3	1	3	3	3	3

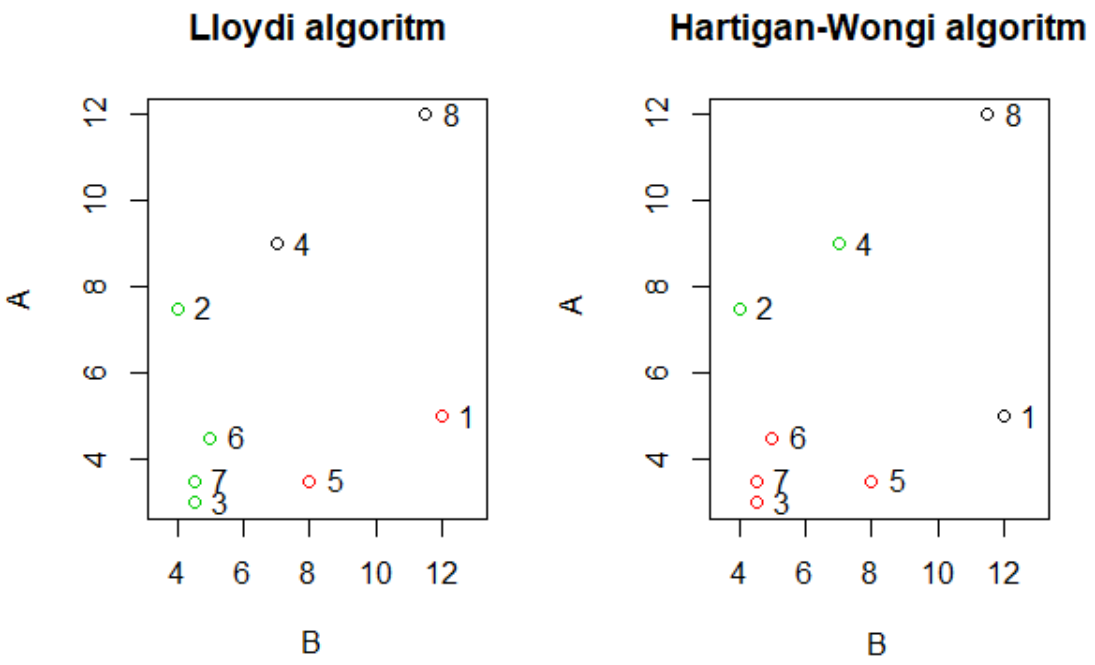
Järgmisena liigume kiire üleviimise sammu juurde, kus arvutame iga isiku jaoks $R_1^{(l_1)}$ kõige lähema klastri jaoks ning $R_2^{(l_2)}$ järgmise klastri jaoks. Isiku 1 jaoks $R_1^{(1)} = 24{,}63$ ning $R_2^{(2)} = 35{,}31$, kuna $R_1^{(1)} \leq R_2^{(2)}$, siis ümberpaigutamist ei toimu. Võrdleme $R_1^{(l_1)}$ ja $R_2^{(l_2)}$ järjest kõikide isikute jaoks. Selgub, et klastrite jaotus jääb samaks, seega liigume edasi sammu 4

juurde. Kuna isikute ümberpaigutamist ei toimunud QTRAN staadiumis, siis on aktiivne hulk tühi ning algoritm lõpetab töö. Lahend on toodud tabelis 5.

Tabel 5. Lõplikud klastrid ja nende keskpunktid

Klaster	Isikud	Keskpunktid
1	1, 8	(8,5; 11,75)
2	3, 5, 6, 7	(3,63; 5,5)
3	2, 4	(8,25; 5,5)

Näeme, et Hartigan-Wongi algoritm jõudis kiiremini lahendini, kuigi lahendid tulid meetodite korral erinevad. Joonisel 2 on toodud Lloydi ja Hartigan-Wongi algoritmi klasterdamise tulemused, klastrid on eristatavad värvidega.



Joonis 2. Klasterdamise tulemused näiteandmestiku jaoks Lloydi ja Hartigan-Wongi algortimi korral

Antud alglähendite korral tuli Lloydi algoritmi korral funktsiooni *J* väärtus 36,44 ja Hartigan-Wongi korral 39,94 ehk Lloydi algoritm annab parema lahendi. Viies aga läbi klasterdamise mõlema algoritmi jaoks 20 erineva alglähendite komplekti korral jõuavad mõlemad algoritmid sama tulemuseni, aga Hartigan-Wongi korral kulub parima lahendi leidmiseks 2 korda vähem

aega (vastavalt 0,014 sekundit ja 0,006 sekundit). Parim lahend tuli sama, mille saime etteantud keskväärtuste alglähenditega Lloydi algoritmi korral (tabel 1, iteratsioon 3).

Viisime läbi ka näite pea andmetega rakendustarkavaras R Hartigani-Wongi ja Lloydi algoritmide kiiruse võrdlemiseks. Andmestikus oli 262 144 rida ja 5 tunnust. Mõlema algoritmi korral olid alglähendid valitud juhuslikult. Lloydi K-keskmiste meetodi rakendamiseks kulus 20,03 sekundit ning Hartigan-Wongi meetodi programmi läbimiseks kulus 15,72 sekundit. Suurte andmestikega töötamise korral on seega ajalises mõttes efektiivsem kasutada Hartigan-Wongi algoritmi.

1.2 Kullback-Leibleri kaugus

Antud peatükk põhineb viidetel [4] ja [6].

Kullback-Leibleri kaugust kasutatakse statistikas kahe jaotuse vahelise seose mõõtmiseks. Sageli mõeldakse Kullback-Leibleri kauguse all informatsiooni hulka, mis läheb kaduma, kui nii-öelda „tõelise“ jaotuse asemel eeldame teistsuguste parameetritega jaotust.

Olgu meil kaks jaotust *P* ja *Q*. Tavaliselt tähistab *P* tõelist jaotust, *Q* aga jaotust, millega soovitakse lähendada jaotust *P*.

Kahe diskreetse jaotuse vaheline Kullback-Leibleri kaugus defineeritakse järgmiselt. Olgu meil kaks diskreetset tõenäosusjaotust P ja Q hulgal $X = \{x_1, x_2, \dots\}$. Tähistagu $p_i = P(x_i)$ ja $q_i = Q(x_i)$ elemendi x_i tõenäosusi vastavalt jaotuste *P* ja *Q* korral. Kullback-Leibleri kaugus jaotuste *P* ja *Q* vahel on defineeritud valemiga

$$KL(P \parallel Q) = \sum_i p_i \ln \frac{p_i}{q_i},$$

seejuures $p_i \ln \frac{p_i}{0} = \infty$, kui $p_i > 0$, ja $0 \ln \frac{0}{q_i} = 0$, kui $q_i \geq 0$. Seega, kui eeldame, et mingi sündmuse esinemise tõenäosus on 0 (st $q_i = 0$ mingi *i* korral), aga tegelikult on vastav tõenäosus positiivne ($p_i > 0$), siis on Kullback-Leibleri kaugus võrdne lõpmatusega, sest $p_i \ln \frac{p_i}{0} = \infty$, kui $p_i > 0$. Juhul, kui mingi sündmuse esinemise tegelik tõenäosus on 0 ($p_i = 0$ mingi *i* korral), aga meie eeldame, et sündmus siiski esineb ($q_i > 0$), siis $0 \ln \frac{0}{q_i} = 0$, kui $q_i \geq 0$, ehk selline sündmus Kullback-Leibleri kaugust ei suurenda.

Järgnevalt defineerime Kullback-Leibleri kauguse kahe pideva jaotuse korral. Olgu $p(x)$ ja $q(x)$ jaotustele *P* ja *Q* vastavad tihedusfunktsioonid. Kullback-Leibleri kaugus pidevate jaotuste vahel on defineeritud valemiga

$$KL(P \parallel Q) = \int p(x) \ln \frac{p(x)}{q(x)} dx.$$

Kuigi suurust $KL(P \parallel Q)$ nimetatakse kauguseks, siis pole tegelikult tegemist siiski meetrikaga, kuna Kullback-Leibleri kaugus on ebasümmeetriline ja kolmnurga võrratus ei kehti. Järgnevalt toomegi välja Kullback-Leibleri kauguse põhiomadused:

- mittenegatiivsus ehk $KL(P \parallel Q) \geq 0$, kusjuures $KL(P \parallel Q) = 0$ siis ja ainult siis, kui $P = Q$. Mittenegatiivsus järeldub Jenseni võrratusest, kuna $g(x) = -\ln x$ on rangelt kumer funktsioon,
- ebasümmeetrilisus ehk $KL(P \parallel Q) \neq KL(Q \parallel P)$.

Olgu meil andmed $x_i,\ i = 1, \dots, n,$ tundmatust jaotusest $p(x)$. Oletame, et meie arvates on sobiv jaotust $p(x)$ lähendada jaotusega $q(x|\theta)$. Parameetrite θ leidmiseks minimiseerime jaotuste *P* ja *Q* vahelise Kullback-Leibleri kauguse θ suhtes. Kuna Kullback-Leibleri kaugus on keskväärtus jaotuse *P* suhtes, siis saame:

$$KL(P \parallel Q) \simeq \frac{1}{n} \sum_{i=1}^{n} \{\ln p(x_i) - \ln q(x_i|\theta)\}. \qquad (1.3)$$

Võrduse parema poole esimene liidetav ei sõltu parameetritest θ ja teine liidetav $\frac{1}{n}\sum_{i=1}^{n} \ln q(x_i|\theta)$ on jaotuse $q(x|\theta)$ logaritmiline tõepärafunktsioon. Seetõttu on Kullback-Leibleri kauguse minimiseerimine samaväärne tõepärafunktsiooni maksimiseerimisega.

Näide 1.3

Järgnevalt demonstreerime Kullback-Leibleri kauguse leidmist mitmemõõtmeliste normaaljaotuste korral.

Tuletame meelde, et *d*-mõõtmelise normaaljaotuse korral avaldub tihedusfunktsioon kujul

$$f(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

Kullback-Leibleri kauguse kahe normaaljaotuse P: $\mathcal{N}_d(\mu_0, \Sigma_0)$ ja Q: $\mathcal{N}_d(\mu_1, \Sigma_1)$ võrdlemiseks saab arvutada järgmiselt:

$$
\begin{aligned}
KL(P \parallel Q) &= E_P[\ln P - \ln Q] \\
&= \frac{1}{2} E_P[-\ln|\Sigma_0| - (x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0) + \ln|\Sigma_1| + (x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)] \\
&= \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_0|} + \frac{1}{2} E_P[-tr(\Sigma_0^{-1}(x-\mu_0)(x-\mu_0)^T) + tr(\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^T)] \\
&= \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_0|} + \frac{1}{2}[-tr(\Sigma_0^{-1}\Sigma_0)] + \frac{1}{2} E_P[tr(\Sigma_1^{-1}(xx^T - 2x\mu_1^T + \mu_1\mu_1^T)] \\
&= \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_0|} - \frac{1}{2}d + \frac{1}{2} tr\left(\Sigma_1^{-1}(\Sigma_0 + \mu_0\mu_0^T - 2\mu_1\mu_0^T + \mu_1\mu_1^T)\right) \\
&= \frac{1}{2}\left(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - d + \ln\frac{|\Sigma_1|}{|\Sigma_0|}\right),
\end{aligned}
$$

kus d on normaaljaotuste dimensioon.

Olgu meil tegelik jaotus P: $\mathcal{N}_2(\mu_0, \Sigma_0)$, kus $\mu_0 = (0 \quad 2)^T$ ja $\Sigma_0 = \begin{pmatrix} 10 & 3 \\ 3 & 2 \end{pmatrix}$. Soovime tegelikku jaotust lähendada jaotustega Q_1: $\mathcal{N}_2(\mu_1, \Sigma_1)$ ja Q_2: $\mathcal{N}_2(\mu_2, \Sigma_2)$, kus $\mu_1 = (1 \quad 2)^T$, $\mu_2 = (5 \quad 2)^T$ ja $\Sigma_0 = \Sigma_1 = \Sigma_2$. Leiame Kullback-Leibleri kaugused, mis on vastavalt $KL(P \parallel Q_1) = 0{,}091$ ja $KL(P \parallel Q_2) = 2{,}273$. Näeme, et tõelise jaotuse lähendamiseks on parem jaotus Q_1, kuna Q_2 korral on Kullback-Leibleri kaugus suurem.

1.3 Gaussi segumudelid

See peatükk põhineb viidetel [4] ja [2].

Segujaotus on jaotus, mis on esitatav mitme jaotuse lineaarkombinatsioonina. Praktikas on tihti komponendid ühest jaotuse klassist, näiteks normaaljaotusest, Poissoni jaotusest jt. Samas võivad komponendid olla ka erinevatest jaotusest, kuid sellisel juhul on mudeli hindamine keerulisem. Kui juhuslik vektor X on K komponendi segu, siis saab selle vektori tihedusfunktsiooni $f(x)$ kirja panna järgmiselt:

$$
f(x) = \sum_{k=1}^{K} \pi_k f_k(x|\theta_k),
$$

kus $\pi_1, \dots, \pi_K$ on komponentide kaalud, $0 \le \pi_k \le 1$, $\sum_k \pi_k = 1$, ja θ_k on komponendi k tihedusfunktsiooni f_k parameetrite vektor, $k = 1, \dots, K$.

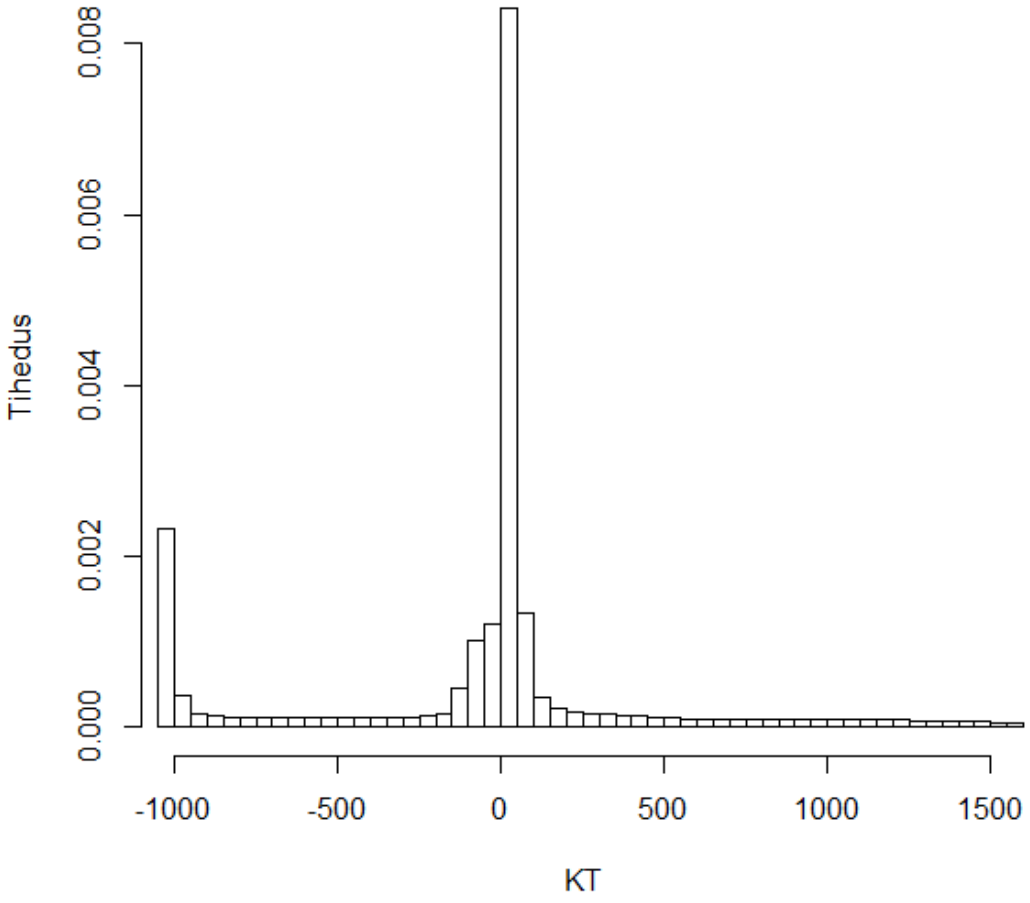
Gaussi segumudeliks kutsutakse segujaotust, mille komponentideks on K normaaljaotust. Jaotuse tihedusfunktsioon on kujul

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x|\mu_k, \Sigma_k),$$

kus

$$f_k(x|\mu_k, \Sigma_k) = \frac{1}{|2\pi\Sigma_k|^{1/2}} exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right).$$

Joonisel 3 on toodud ühe pea kompuutertomograafia mõõtmiste tulemused. Histogrammi põhjal näeme, et kompuutertomograafia tunnuse jaotust saab kirjeldada segujaotuse abil. Väärtuse -1000 ümbruses, mis vastab õhu piirkonnale, võiks vaatlustele sobitada äralõigatud normaaljaotuse. Lähemal uurimisel selgub, et väärtuse 0 ümbruses sobib andmete kirjeldamiseks kasutada kahte erinevat normaaljaotusega komponenti. Paremale sabale (luu piirkond) vastab suure dispersiooniga normaaljaotus. Väärtuse -500 ümbruses võime täheldada ühtlast jaotust. Seega võiks KT tunnusele sobitada viiekomponendilise segujaotuse.



Joonis 3. Kompuutertomograafia segujaotus

Toome sisse latentse juhusliku suuruse Z, mille võimalikud väärtused on $1, \dots, K$. Seejuures $P(Z = k) = \pi_k, \ k = 1, \dots, K$. Vaatame tinglikku tõenäosust, et $X = x$ korral on tegemist komponendi *k* vaatlusega:

$$P(Z = k|X = x) = \frac{P(Z = k)f(x|Z = k)}{\sum_{j=1}^{K} P(Z = j)f(x|Z = j)} = \frac{\pi_k f_k(x|\theta_k)}{\sum_{j=1}^{K} \pi_j f_j\left(x|\theta_j\right)} := \gamma_k.$$

Suurus γ_k näitab, kui suure osa vastutusest (inglise keeles *responsibility)* võtab k-s komponent andmepunkti x kirjeldamisel.

Segumudeli parameetrite hindamiseks kasutatakse suurima tõepära meetodit, see tähendab, et maksimiseeritakse tõepärafunktsioon parameetrite (keskväärtuste vektorid, kovariatsioonimaatriksid ja kaalude vektor) suhtes. Mudeli suurima tõepära hinnangud leitakse EM (inglise keeles *expectation–maximization*) algoritmiga, mis on iteratiivne protseduur.

EM algoritm põhineb kahel sammul, mida korratakse koondumiseni. Esmalt valitakse keskväärtustele, kovariatsioonimaatriksitele ja kaaludele alglähendid. Seejärel:

1. **E-samm:** olemasolevate parameetrite väärtuste põhjal arvutatakse nn vastutused,
2. **M-samm:** leitakse uued parameetrite hinnangud kasutades E-sammus leitud vastutusi.

Iga iteratsiooni korral on garanteeritud, et logaritmilise tõepärafunktsiooni väärtus ei kahane. Algoritm koondub, kui muutus logaritmilises tõepärafunktsioonis või parameetrite hinnangutes on piisavalt väike.

Järgnevalt näitlikustamegi EM algoritmi ideed kahekomponendilise ja ühemõõtmelise Gaussi segumudeli korral. Olgu *X* kahe normaaljaotuse segu nii, et $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ ja $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ ning $P(Z = 1) = \pi_1$ ja $P(Z = 2) = \pi_2$. Tahame sobitada mudelit andmetele $\{x_1, \dots, x_n\}$ suurima tõepära meetodiga. Parameetrid, mida soovime hinnata, on $\theta = (\pi_1, \pi_2, (\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2))$, seejuures $\pi_1 + \pi_2 = 1$. Paneme kirja tõepärafunktsiooni:

$$L(\theta|x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} [\pi_1 f_1(x_i|\mu_1, \sigma_1^2) + \pi_2 f_2(x_i|\mu_2, \sigma_2^2)],$$

seega logaritmiline tõepärafunktsioon on kujul:

$$l(\theta|x_1, \dots, x_n) = \ln L(\theta|x_1, \dots, x_n) = \sum_{i=1}^{n} \ln\{\pi_1 f_1(x_i|\mu_1, \sigma_1^2) + \pi_2 f_2(x_i|\mu_2, \sigma_2^2)\}.$$

Olgu $\gamma_{ik} = P(Z = k|X = x_i)$, $k = 1,2$, vastutus, mille komponent k võtab andmepunkti x_i kirjeldamisel. Võtame funktsioonist $l(\theta|x_1, \dots, x_n)$ osatuletised hinnatavate parameetrite suhtes ja võrdsustame need nulliga. Saame järgmised avaldised parameetrite hindamiseks:

$$\mu_1 = \frac{\sum_{i=1}^{n} \gamma_{i1} x_i}{\sum_{i=1}^{n} \gamma_{i1}}, \qquad \mu_2 = \frac{\sum_{i=1}^{n} \gamma_{i2} x_i}{\sum_{i=1}^{n} \gamma_{i2}},$$

$$\sigma_1^2 = \frac{\sum_{i=1}^{n} \gamma_{i1}(x_i - \mu_1)^2}{\sum_{i=1}^{n} \gamma_{i1}}, \qquad \sigma_2^2 = \frac{\sum_{i=1}^{n} \gamma_{i2}(x_i - \mu_2)^2}{\sum_{i=1}^{n} \gamma_{i2}},$$

$$\pi_1 = \sum_{i=1}^{n} \frac{\gamma_{i1}}{n}, \qquad \pi_2 = 1 - \pi_1.$$

Paneme tähele, et iga andmepunkt mõjutab kõiki parameetrite hinnanguid, kuid tema panus parameetri hinnangusse on määratud kaaluga γ_{ik}. Summad $\sum_{i=1}^{n} \gamma_{i1}$ ja $\sum_{i=1}^{n} \gamma_{i2}$ näitavad, kui palju vaatlusi keskmiselt kuulub komponenti 1 ja komponenti 2. Kuna vastutused γ_{ik} sõltuvad parameetritest, siis pole tegemist ülesande lõpliku lahendusega, vaid antud ülesanne annab aimu, kuidas leida lahendus iteratiivselt: vastutuste γ_{ik} hinnanguid teades saame leida ka ülejäänud parameetrite hinnangud.

Järgnevalt kirjeldame üldist EM algoritmi Gaussi segumudeli parameetrite hindamiseks. Olgu meil andmestik $\{x_1, \dots, x_n\}$. Tahame hinnata *K*-komponendilise Gaussi segumudeli parameetreid $\pi = (\pi_1, \dots, \pi_K)$, $\mu = (\mu_1, \dots, \mu_K)$ ja $\Sigma = (\Sigma_1, \dots, \Sigma_K)$.

Algoritm:

1. Valitakse parameetrite alglähendid. Olgu selguse mõttes keskväärtuse vektorid tähistatud μ_k^{vana}, kovariatsioonimaatriksid Σ_k^{vana} ja kaalud π_k^{vana}, $k = 1, \dots, K$. Arvutatakse esialgne logaritmiline tõepärafunktsioon:

$$\ln f(x_1, \dots, x_n | \theta^{vana}) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k^{vana} f_k(x_i | \mu_k^{vana}, \Sigma_k^{vana}) \right\}.$$

2. **E-samm**: arvutatakse suurused γ_{ik} (vastutused) kasutades olemasolevaid parameetrite väärtusi:

$$\gamma_{ik} = \frac{\pi_k^{vana} f_k(x_i | \mu_k^{vana}, \Sigma_k^{vana})}{\sum_{j=1}^{K} \pi_j^{vana} f_j\left(x_i \middle| \mu_j^{vana}, \Sigma_j^{vana}\right)}, \qquad i = 1, \dots, n, \qquad k = 1, \dots, K.$$

3. **M-samm**: arvutatakse uued parameetrite hinnangud kasutades E-sammus leitud vastutusi:

$$\mu_k^{uus} = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{ik} x_i,$$

$$\Sigma_k^{uus} = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{ik} (x_i - \mu_k^{uus})(x_i - \mu_k^{uus})^T,$$

$$\pi_k^{uus} = \frac{n_k}{n},$$

kus $n_k = \sum_{i=1}^{n} \gamma_{ik}, \quad k = 1, \dots, K.$

4. Arvutatakse logaritmiline tõepärafunktsioon

$$\ln f(x_1, \dots, x_n | \theta^{uus}) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k^{uus} f_k(x_i | \mu_k^{uus}, \Sigma_k^{uus}) \right\}$$

ja kontrollitakse parameetrite või logaritmilise tõepärafunktsiooni koonduvust. Juhul, kui koondumiskriteerium ei ole rahuldatud, minnakse tagasi sammu 2 juurde.

EM algoritm vajab koondumiseks palju rohkem iteratsioone kui K-keskmiste algoritm. Tihti kasutatakse Gaussi segumudelite alglähenditeks K-keskmiste meetodi tulemusi. Selle jaoks arvutatakse K-keskmiste meetodi abil saadud klastrite suuruste põhjal segumudelite kaalude alglähendid, keskpunkte kasutatakse keskväärtuste parameetritena ning kovariatsioonimaatriksid arvutatakse K-keskmiste klastrite põhjal. Sarnaselt K-keskmiste algoritmiga on ka EM algoritmi vajalik rakendada mitu korda erinevate algväärtustega, kuna pole garanteeritud, et leitud lahend on globaalne ekstreemum.

2. Peadevaheliste erinevuste uurimine K-keskmiste meetodi ja Gaussi segumudelite abil

Antud peatüki eesmärk on uurida üheksa pea (viis meest ja neli naist) erinevusi KT ja MRT vaatluste korral.

Varasema modelleerimisülesande [1] tulemuste järgi jagunesid pead vastavalt mudeli jääkide käitumisele kaheks grupiks: homogeensed pead ja jääkide poolest erinevalt käitunud ehk mittehomogeensed pead. Vastavalt modelleerimisel saadud jääkide käitumisele on pead edaspidi jaotatud nii-öelda „headeks“ peadeks (pea 1, pea 2, pea 4, pea 8 ja pea 9) ning „halbadeks“ peadeks (pea 3, pea 5, pea 6 ja pea 7).

Antud ülesandes sõltuvad tunnuste väärtused iga vaatluse korral otseselt sellest, millise koe või kudede seguga on antud vaatluse puhul tegemist. Koe tüüpi on võimalik kindlaks teha KT väärtuste põhjal (vaata lisa 1). Näiteks õhu klassi kuuluva vaatluse väärtused on väiksemad kui -1000, valge- ja hallolluse korral on väärtused aga vastavalt vahemikus 20 kuni 35 või 30 kuni 40. KT väärtuste põhjal teame, kui palju klasse oleks mõistlik klasterdamisel või segukomponentide arvu määramisel vaadelda. Koe klasside ja varasema modelleerimisülesande [1] põhjal jagasime andmed K-keskmiste meetodi korral viide klastrisse ja vaatlesime segumudelite jaoks viit segukomponenti..

Peadevaheliste erinevuste uurimiseks viisime läbi klasterdamise K-keskmiste meetodil. Ühendasime homogeense grupi andmestikud üheks andmestikuks ning leidsime klastrid juhuslike alglähenditega. Leitud klastrite keskmisi kasutasime edaspidi ülejäänud peade jaoks, et leida vaatluste klastritesse paigutus. Leidsime klastrite suurused homogeense grupi korral, eeldades, et tegemist on nii-öelda „tõese“ jaotusega, ning samadele keskpunktidele vastavate klastrite suurused ka iga pea korral eraldi. Vaatluste osakaalu kasutasime homogeensete peade jaotuse võrdlemiseks iga üksiku pea jaotusega, selleks leidsime Kullback-Leibleri kaugused diskreetsete jaotuste korral.

Teises osas soovisime võrrelda peade jaotusi segumudelite abil. Kasutasime tunnuste jaotuste kirjeldamiseks Gaussi segujaotust, see tähendab eeldasime, et iga üksikut kudede klassi on sobilik kirjeldada mitmemõõtmelise normaaljaotuse abil. Sarnaselt K-keskmistega vaatasime viie homogeense pea koondandmestikku ning hindasime nende peade korral nn „tegeliku“ mudeli. Seejärel võrdlesime iga pea jaotust tegeliku ehk viie hea pea korral leitud jaotusega. Selleks arvutasime Kullback-Leibleri kaugused vastavate pidevate jaotuste korral.

2.1 Andmete kirjeldus

Iga pea jaoks on andmed kujutatud kuubi sees, mis on jaotatud $192 \times 192 \times 192$ väiksemaks kuubiks ehk voksliks, kusjuures ühe voksli küljepikkus on 1,33 mm. Andmestikus on iga voksli jaoks antud x-, y- ja z-koordinaadid, mis näitavad voksli asukohta kuubis. Iga voksli jaoks on andmestikus viie tunnuse väärtused, üks väärtus KT jaoks ja neli erinevat MRT mõõtmist. MRT mõõtmised (ute10, ute30, ge10, ge30) vastavad neljale erinevale MRT pildile, mis on saadud nelja erineva magnetvälja parameetrite komplektiga magnetkaameras. Kuna üldiselt on MRT piltidel väga raske eristada luud ja õhku, siis on antud MRT parameetrid valitud nii, et saada võimalikult head informatsiooni luu kohta. Lisaks on iga voksli jaoks antud binaarne tunnus „indeks“, mille väärtus on 1, kui tegemist on pea ehk vaatluste voksliga, ja 0, kui tegemist on pead ümbritseva õhuga.

2.2 K-keskmiste meetodi rakendamine peadevaheliste erinevuste uurimiseks

Antud peatükis arvutame K-keskmiste tulemuste põhjal Kullback-Leibleri kaugused, mida kasutame peadevaheliste jaotuste võrdlemiseks. Esmalt toome sisse selgitavad tähistused. Teame, et iga voksli korral sõltub mõõtmiste väärtus sellest, millisesse koe klassi ta kuulub. Tuletame meelde, et vaatleme viit klassi. Olgu P tõenäosusjaotus hulgal $X = \{1, 2, 3, 4, 5\}$ tegeliku jaotuse korral ehk p_i on tõenäosus, et voksel kuulub klassi *i* homogeensete peade korral. K-keskmiste meetodi abil leiame viie hea pea koondandmestiku klastrite keskpunktid ja nende suurused (osakaalud), mis on tähistatud vastavalt $p_1, \dots, p_5$. Olgu $Q_j,\ j = 1, \dots, 9,$ vastavad tõenäosusjaotused iga üksiku pea korral, $q_i^{(j)}$ näitab voksli kuulumise tõenäosust klassi *i* pea *j* korral. Tõenäosused $q_i^{(j)}$ kirjeldavad klastrite suurusi, mis on leitud viie homogeense pea K-keskmiste keskpunkte kasutades. Kui kudede jaotus on iga pea korral sarnane, siis peaksid ka suurused $p_1, \dots, p_5$ ja $q_1^{(j)}, \dots, q_5^{(j)}$ vähe erinema ning vastavad Kullback-Leibleri kaugused peaksid olema väikesed. Juhul, kui kudede jaotus ei sarnane erinevate peade korral, siis peaksid seda väljendama ka suurused $p_1, \dots, p_5$ ja $q_1^{(j)}, \dots, q_5^{(j)}$ ning Kullback-Leibleri kaugused peaksid olema suuremad.

2.2.1 K-keskmiste meetodi funktsiooni kirjeldus

Rakendustarkvaras R on klasterdamise jaoks K-keskmiste meetodil funktsioon *kmeans(x,centers,iter.max,nstart,algorithm).* Antud käsus tähistab *x* andmestikku, mida soovitakse klasterdada; *centers* on klastrite arv, mitmeks soovitakse andmestik jagada, või keskpunktide alglähendid; *nstart* väärtus määrab, mitme erineva alglähendite komplekti korral algoritmi rakendatakse; *iter.max* on maksimaalne iteratsioonide arv; parameetriga *algorithm* saab valida, millist algoritmi klasterdamisel kasutatakse. Töös kasutatud andmed olid suuremahulised ja seetõttu pidime esmalt parameetritele *iter.max, nstart* ja *algorithm* sobivate väärtuste valimiseks läbi viima katsetusi.

Parameetri *nstart* valimiseks testisime algoritmi pea 1 korral väärtustega 20, 30, 40 ja 50 ning tegime järeldused, et erinevate väärtuste korral tulemused oluliselt ei muutunud. Seega edaspidises töös jäime väärtuse 20 juurde.

Algoritmi valikuteks on Hartigan-Wong, Lloyd, Forgy ja MacQueen, kus Lloyd ja Forgy viitavad tegelikult samale algoritmile. Vaikimisi soovitab R kasutada Hartigan-Wongi algoritmi, mis on kiirem kui teised algoritmid. Väga sarnaste vaatluste korral aga ei pruugi algoritm koonduda kiire üleviimise sammus ning programm annab siis hoiatusi. Antud probleem esines ka meie andmete korral. Võrdlesime hoiatusteta ja hoiatustega tulemusi ning nägime, et lõpuks olid keskpunktid viie kümnendkohaga samad.

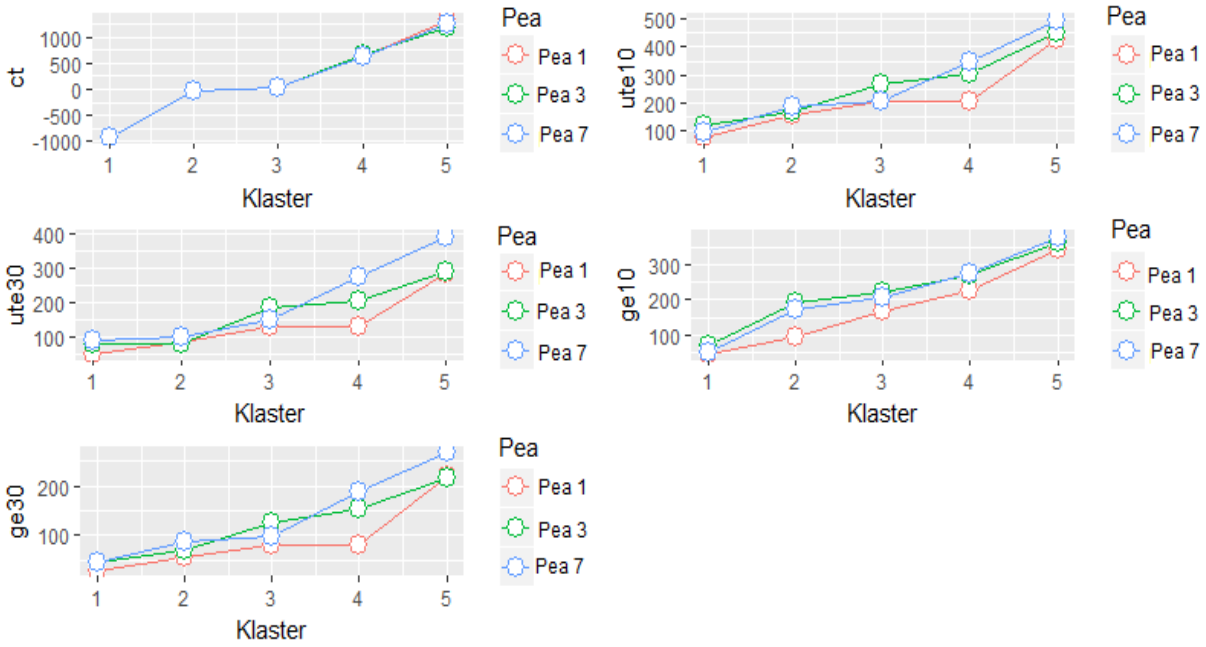
Hoiatuste tõttu katsetasime ka Lloydi algoritmi. Suurte andmestike korral oli vajalik leida iteratsioonide arv, mille korral algoritm koonduks. Pärast katsetamist valisime iteratsioonide arvu ülempiiriks 500.

Võrdlesime K-keskmiste meetodi Hartigan-Wongi ja Lloydi algoritmide tulemusi, et näha, kas algoritmid annavad erinevaid väärtusi – tulemused olid samad kuni kolmanda kümnendkohani. Kuna Hartigan-Wong siiski andis hoiatusi, rakendasime töös edaspidi Lloydi algoritmi.

2.2.2 K-keskmiste meetod kolme pea korral

Kõigepealt soovisime uurida, kuidas käitub K-keskmiste meetod heade ja halbade peade korral. Selleks viisime läbi esmase analüüsi ühe hea (pea 1) ning kahe halva pea (pea 3 ja pea 7) jaoks.

Kasutasime nii MRT kui ka KT mõõtmisi ning leidsime klastrid ja nende keskpunktid K-keskmiste meetodil, tulemused on toodud joonisel 4.

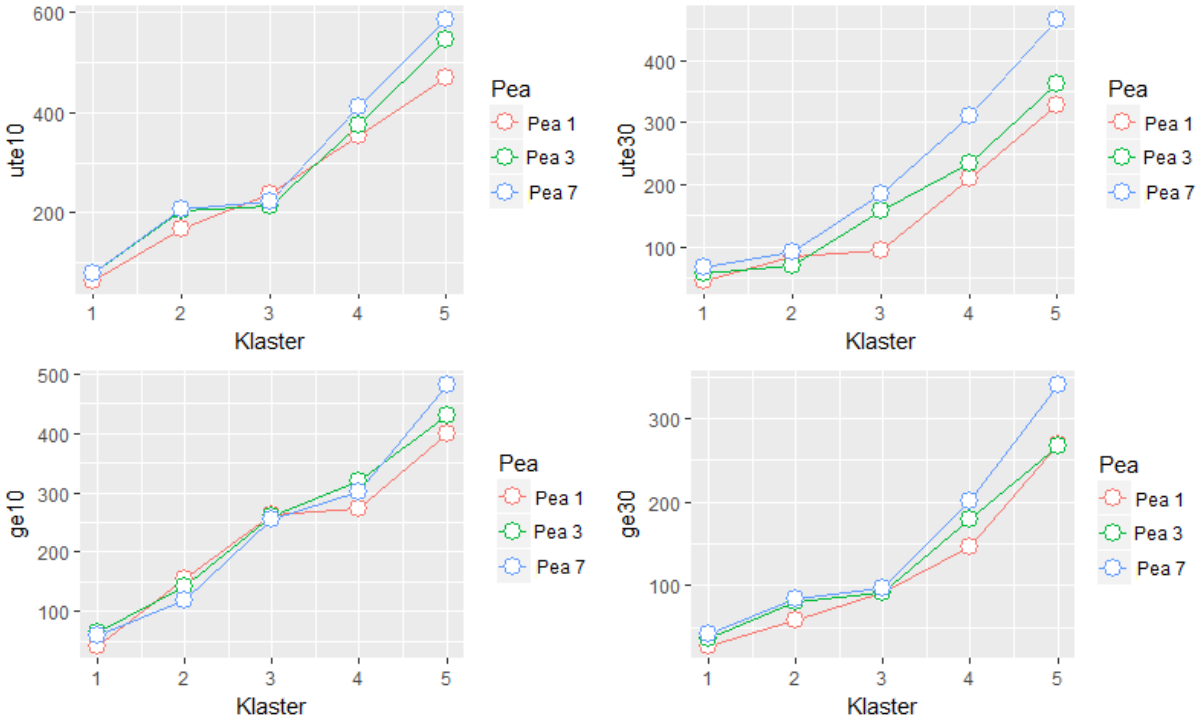


Joonis 4. Viie klastri keskpunktid pea 1, pea 3 ja pea 7 jaoks

Paneme tähele, et kõigi kolme pea jaoks KT (tunnus ct) keskmised ei erine oluliselt. Samas MRT tunnuste jaoks näeme suuremaid erinevusi kõrgemate väärtuse juures, samuti võib täheldada, et pea 1 erineb peadest 3 ja 7.

Joonisel 5 on toodud tulemused MRT tunnuste jaoks. Näeme, et kui klastrites 1 ja 2 on kolme pea jaoks tulemused suhteliselt sarnased, siis ülejäänud klastrite keskmiste korral on märgata suuremaid erinevusi. Kõige suuremad erinevused on tunnuse ute30 jaoks.

Järeldasime, et erinevused peade jaoks võivad olla tingitud just MRT mõõtmistest. Peadevahelist erinevust on uurinud ka varasem bakalaureusetöö [7], kus visualiseeriti ühte head ja ühte halba pead. Antud töös täheldati samuti peadevahelisi erinevusi MRT tunnuste jaoks. Seega edasises töös otsustasime kasutada ainult MRT tunnuseid.



Joonis 5. MRT tunnuste viie klastri keskpunktid pea 1, pea 3 ja pea 7 jaoks

2.2.3 K-keskmiste meetod kogu pea andmete korral

Antud peatüki eesmärk on võrrelda viie hea pea jaotust ülejäänud peade jaotusega kogu pea andmete korral K-keskmiste meetodi abil. Selleks rakendasime esiteks K-keskmiste klasterdamist andmestikule, kus olid kokku pandud viis homogeenset pead, ja hindasime suurused $p_1, \dots, p_5$. Viie pea klasterdamisel saadud klastrite keskpunkte kasutades viisime seejärel läbi vokslite klastritesse paigutamise iga pea korral ja hindasime suurused $q_1^{(j)}, \dots, q_5^{(j)}, j = 1, \dots, 9$.

Heade peade andmestiku jaoks ühendasime pea 1, pea 2, pea 4, pea 8 ja pea 9 andmed (edaspidi tabelites toodud paksus kirjas) ning see andmestik sisaldas mõõtmiseid 9 701 656 voksli jaoks. Viisime klasterdamise läbi viie klastri korral, tulemused on toodud tabelis 6.

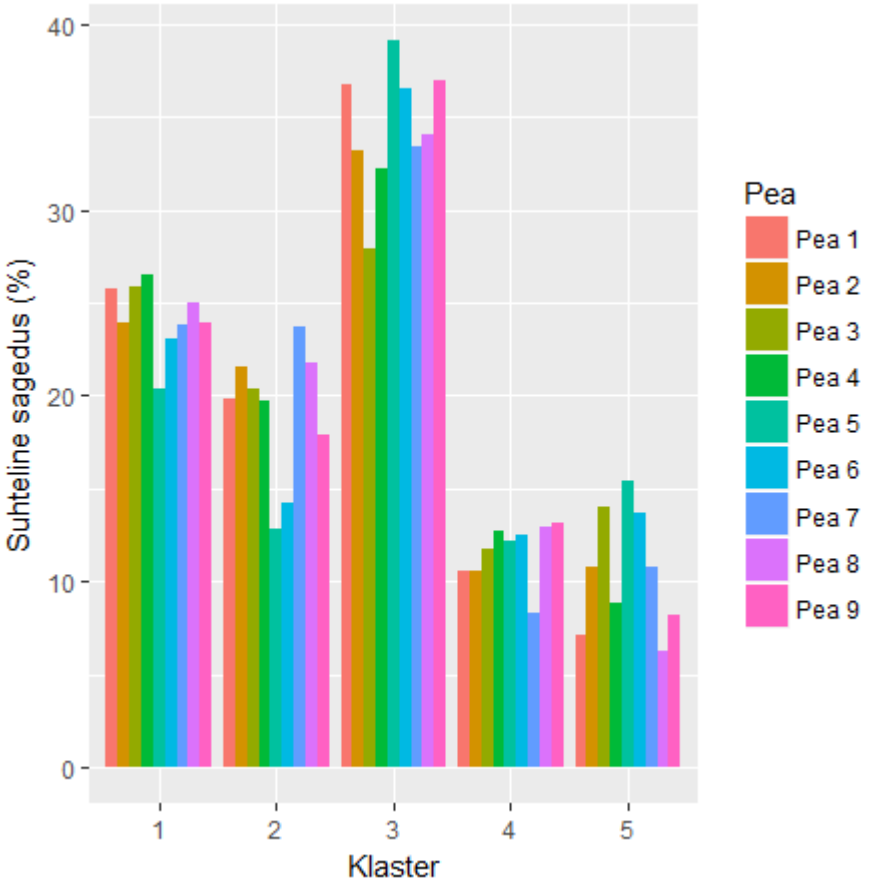
Tabel 6. MRT tunnuste viie klastri keskpunktid heade peade korral

Klaster (vokslite arv)	ute10	ge10	ute30	ge30
1 (2 426 721)	69,36	43,96	49,43	28,68
2 (1 951 700)	170,88	150,19	88,24	58,75
3 (3 359 355)	228,01	272,83	84,92	88,12
4 (1 170 298)	349,83	257,41	205,58	138,66
5 (793 582)	469,16	402,36	312,59	255,27

Kasutades tabelis 6 toodud klastrite keskpunkte, leidsime vastavad klastrid ja klastrite suurused iga pea jaoks. Klastritesse kuuluvate punktide suhtelised sagedused iga pea korral on toodud tabelis 7. Parema ülevaate saamiseks peade võrdlemisel on suhtelised sagedused esitatud ka joonisel 6.

Tabel 7. Heade peade andmestiku põhjal leitud klastrite keskpunktidel põhinevate klastrite suurused iga pea korral (suhtelised sagedused %)

Pea (vokslite arv)	Klaster 1	Klaster 2	Klaster 3	Klaster 4	Klaster 5
1 (1 853 702)	25,72	19,84	36,77	10,54	7,13
2 (1 747 700)	23,87	21,58	33,17	10,62	10,76
3 (1 597 634)	25,90	20,41	27,93	11,71	14,05
4 (2 020 028)	26,52	19,68	32,19	12,72	8,88
5 (1 654 901)	20,39	12,84	39,11	12,20	15,46
6 (1 722 630)	23,01	14,25	36,57	12,48	13,69
7 (1 325 922)	23,84	23,69	33,37	8,29	10,81
8 (2 020 901)	24,97	21,79	34,03	12,98	6,24
9 (2 059 325)	23,91	17,91	36,91	13,12	8,16



Joonis 6. Heade peade andmestiku põhjal leitud klastrite keskpunktidel põhinevate klastrite suurused iga pea korral (suhtelised sagedused %)

Suhtelisi sagedusi vaadates näeme, et kõige rohkem erinevad halvad pead homogeensetest peadest klastris viis. Heade peade korral kuulub pigem alla 10% vokslitest viiendasse klastrisse, samas mittehomogeensete peade korral aga üle 10%. Pea 5 ja pea 6 korral tuleb selgelt välja, et väiksemate MRT väärtustega klastrites on vähem voksleid, samas viiendas klastris on aga ligi kaks korda rohkem voksleid võrreldes heade peadega. Sarnast trendi näeme ka teiste mittehomogeensete peade korral. Seega üheks peadevaheliste erinevuste põhjuseks võib olla, et mittehomogeensete peade grupi korral on MRT tunnustel üldiselt suuremad väärtused.

Homogeensete ja mittehomogeensete peade erinevust kinnitavad ka Kullback-Leibleri kaugused (tabel 8). Kullback-Leibleri kaugused $KL(P \parallel Q_j) = \sum_i p_i \ln \frac{p_i}{q_i^{(j)}}$ on leitud eeldusel, et heade peade jaotus on tõene jaotus ehk heade peade korral leitud klastrite suurused vastavad tegelikele klastrite suurustele. Iga pea klasterdamisel saadud klastrite suurusi on seejärel võrreldud viie pea korral saadud klastrite suurustega.

Tabel 8. K-keskmiste klastrite suurusi kasutades leitud Kullback–Leibleri kaugused iga pea jaoks ($\times 10^{-3}$)

Klastrite arv	**Pea 1**	**Pea 2**	Pea 3	**Pea 4**	Pea 5	Pea 6	Pea 7	**Pea 8**	**Pea 9**
5	0,255	0,539	2,212	0,184	4,586	2,511	1,437	0,379	0,272
4	0,227	0,462	2,084	0,203	3,663	2,324	0,690	0,313	0,180
3	0,173	0,197	1,910	0,169	2,368	1,244	0,019	0,143	0,155

Kullback-Leibleri kaugused on kõigi homogeensete peade korral alla 0,6, halbade peade korral aga näeme, et kaugused on mitu korda suuremad. Seega kinnitab ka Kullback-Leibleri kaugus, et halbade peade jaotus erineb heade peade jaotusest. Tegime võrdlemise eesmärgil läbi klasterdamise ka 3 ja 4 klastri korral ning on näha, et ka väiksemate klastrite arvu korral on erinevused olemas, teistest eristuvad alati selgelt pea 3, pea 5 ja pea 6.

2.2.4 K-keskmiste meetod peade sisemise osa korral

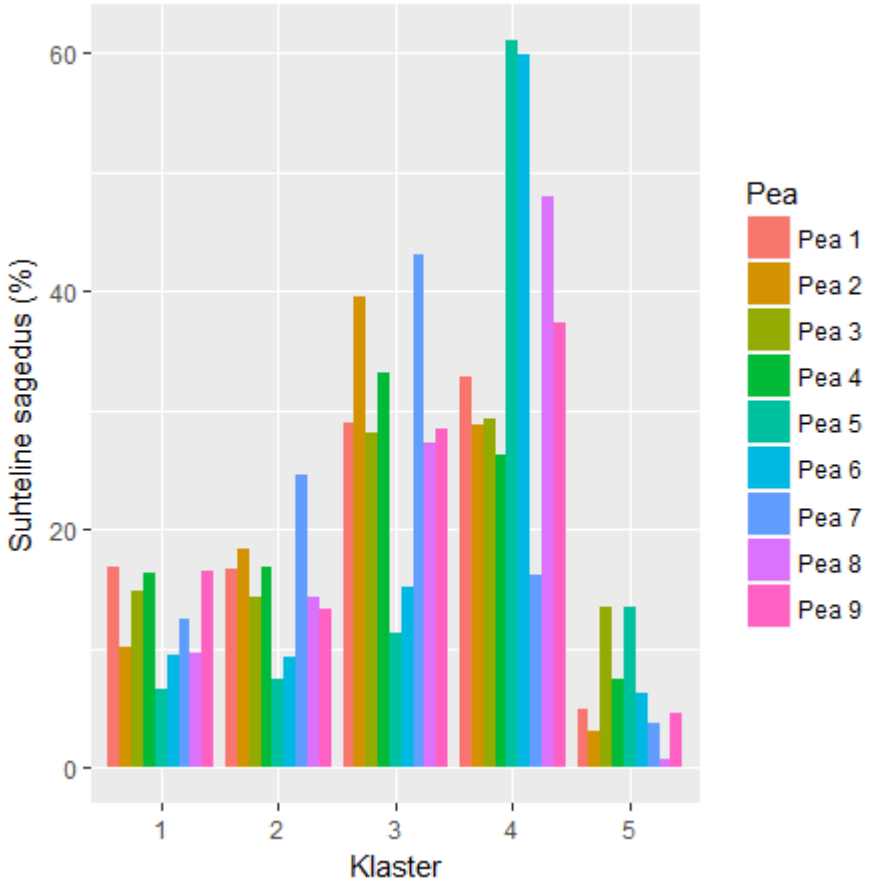
Bakalaureusetöös [7] uuriti peadevahelisi erinevusi visualiseerimismeetodi t-SNE abil ja täheldati MRT tunnuste erinevust peade sisemises osas, kus on kudedest peamiselt esindatud valge- ja hallollus. Seega soovisime uurida, kas ka meie lähenemist kasutades tulevad erinevused Kullback-Leibleri kauguste abil välja samas piirkonnas. Pea sisemise osa saamiseks jagasime pea 216 väiksemaks kuubiks, kus ühe kuubi küljepikkus oli 32 vokslit. Analüüsiks kasutasime 8 sisemist kuupi ehk tükke, mille vokslite x-, y- ja z-koordinaatide väärtused jäid vahemikku 65-96 ja 97-128. Maksimaalne vokslite arv sisemise osa jaoks oli 262 144 vokslit. Kõikide heade peade korral olid andmed olemas kõigi vokslite jaoks, samas aga halbadest peadest oli peal 3 (230 721 vokslit) ja peal 7 (224 337 vokslit) osa mõõtmisi puudu.

Peade sisemise osa erinevuste uurimisele lähenesime samamoodi nagu kogu pea andmete korral: leidsime viie homogeense pea klastrite keskpunktid ning kasutasime saadud keskpunkte klastrite suuruse määramiseks iga üksiku pea korral. Klastrite suurused on toodud tabelis 9.

Tabel 9. Heade peade andmestiku põhjal leitud klastrite keskpunktidel põhinevate klastrite suurused iga pea sisemise osa korral (suhtelised sagedused %)

Pea (vokslite arv)	Klaster 1	Klaster 2	Klaster 3	Klaster 4	Klaster 5
1 (262 144)	16,76	16,61	28,89	32,79	4,95
2 (262 144)	10,12	18,41	39,52	28,82	3,13
3 (230 721)	14,84	14,37	28,11	29,25	13,43
4 (262 144)	16,35	16,83	33,18	26,18	7,45
5 (262 144)	6,61	7,47	11,28	61,08	13,55
6 (262 144)	9,48	9,27	15,17	59,85	6,22
7 (224 337)	12,5	24,5	43,06	16,23	3,72
8 (262 144)	9,65	14,28	27,35	47,93	0,78
9 (262 144)	16,43	13,3	28,44	37,29	4,54

Joonisel 7 on toodud klastrite suurused ning pea 5 ja pea 6 erinevused on selgelt näha. Esimeses ja teises klastris on peadel 5 ja 6 umbes 7-10% vähem voksleid kui homogeensetel peadel. Heade peade korral on suurem osa voksleid jaotunud kolmanda ja neljanda klastri vahel, samas pea 5 ja pea 6 korral on 60% vokslitest ainult neljandas klastris. Seega ka sisemiste osade korral paistab, et peal 5 ja peal 6 on suuremad MRT väärtused. Puuduvate andmetega pea 7 korral on voksleid headest peadest rohkem klastrites kaks ja kolm, pea 3 korral aga klastris viis.



Joonis 7. Heade peade andmestiku põhjal leitud klastrite keskpunktidel põhinevate klastrite suurused iga pea sisemise osa korral (suhtelised sagedused %)

Sarnaselt tervete peadega arvutasime ka siin välja Kullback-Leibleri kaugused, see tähendab võrdlesime homogeensete peade klasterdamisel saadud klastrite suuruseid iga üksiku pea klastrite suurustega. Kullback-Leibleri kaugused on toodud tabelis 10. Kullback-Leibleri kaugused on suuremad just peade 5 ja 6 jaoks ehk need pead erinevad rohkem tõesest jaotusest, mida nägime ka klastrite suurustest. Ka pea 7 korral on Kullback-Leibleri kaugus viie klastri korral suurem kui heade peade korral. Võrdlesime ka Kullback-Leibleri kaugusi 4 ja 3 klastri korral ning pea 5 erinevus tuli ka nende korral välja.

Tabel 10. K-keskmiste klastrite suurusi kasutades leitud Kullback–Leibleri kaugused iga pea sisemise osa jaoks

Klastrite arv	**Pea 1**	**Pea 2**	Pea 3	**Pea 4**	Pea 5	Pea 6	Pea 7	**Pea 8**	**Pea 9**
5	0,005	0,024	0,051	0,024	0,300	0,162	0,114	0,068	0,007
4	0,005	0,017	0,052	0,019	0,246	0,124	0,073	0,064	0,005
3	0,004	0,004	0,052	0,015	0,135	0,062	0,003	0,056	0,001

2.3 Gaussi segumudelid

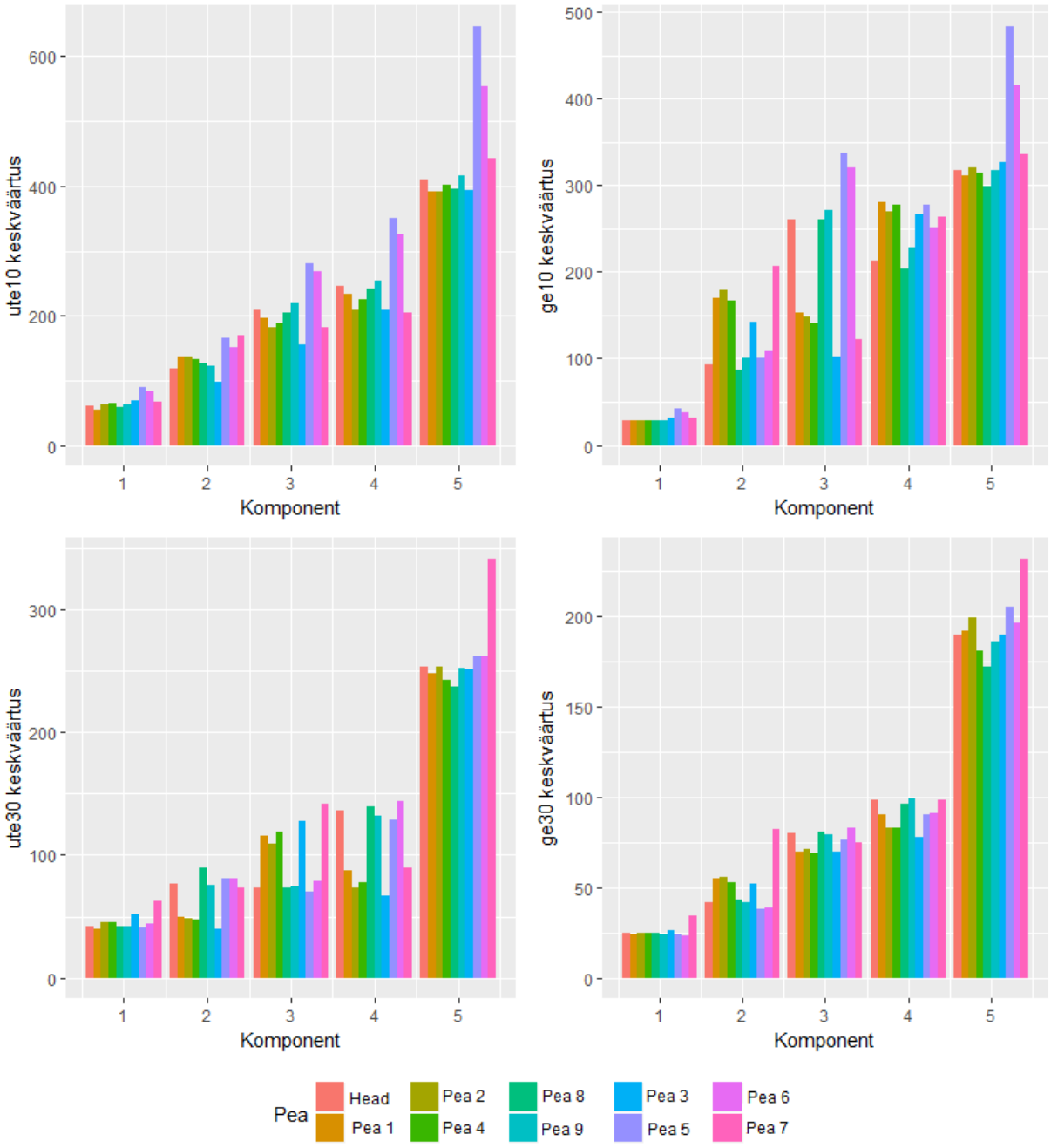
Peade erinevuste uurimiseks kasutasime lisaks K-keskmiste meetodile Gaussi segumudeleid. Rakendustarkvaras R on tööks Gaussi segumudelitega olemas pakett *Mclust*. Pakett sisaldab funktsioone, et vastavalt vajadusele hinnata kas juhuslike alglähenditega või etteantud parameetrite alglähenditega Gaussi segumudel. Programmikoodi näide on toodud lisas 2. Juhuslike alglähendite korral määratakse esialgsed komponendid hierarhilise klasterdamise teel. Üheks oluliseks aspektiks antud paketis on kovariatsioonimaatriksite stuktuur. Sobiva struktuuri kovariatsoonimaatriksitele saab ette anda varasemate teadmiste põhjal või lasta algoritmil andmete põhjal leida sobiv struktuur. Antud magistritöös kasutasime kitsendusteta mudelit, st me ei seadnud kovariatsioonimaatriksitele mingeid piiranguid.

Peadevaheliste erinevuste uurimiseks hindasime esmalt viiest homogeensest peast koosneva koondandmestiku korral tõese ehk tegeliku jaotuse. Selle jaoks leidsime komponentide kaalude, keskväärtuste ja kovariatsioonimaatriksite hinnangud viie hea pea korral. Olgu vastava segujaotuse tihedus $p(x)$. Seejärel hindasime Gaussi segujaotuse iga pea korral eraldi, see tähendab leidsime vastavad parameetrite hinnangud. Olgu peadele $1, \dots, 9$ vastavad segujaotuse tihedusfunktsioonid tähistatud $q_1(x), \dots, q_9(x)$.

Kuna tõepära funktsioonil võib olla mitmeid lokaalseid maksimume, siis pole garanteeritud, et EM algoritm leiab nendest suurima. Sellepärast on tarvis mudelit hinnata erinevate alglähendite komplektide korral, et jõuda võimalikule globaalsele tõepära maksimumile võimalikult lähedale. Alglähenditena kasutasime kolme erinevat varianti: iga pea K-keskmiste tulemusi (keskpunktid, klastrite suurused ning klastritesse kuuluvad andmepunktid) kasutades leitud Gaussi segumudelite algparameetrid, viie hea pea andmestiku K-keskmiste tulemusi kasutades leitud algparameetrid ja juhuslikud algparameetrid. Juhuslike alglähendite korral hindasime kolm erinevat mudelit. Iga pea korral võrdlesime mudelite logaritmilise tõepära väärtusi ja sobivaimaks mudeliks valisime kõige suurema väärtusega mudeli. Iga pea parima mudeli logaritmiline tõepära on toodud lisas 3 paksus kirjas. Paneme tähele, et neljal korral (pea 1, pea 2, pea 3 ja pea 4) osutus parimaks juhuslike alglähenditega mudel. Kovariatsioonimaatriksi singulaarsuse tõttu ei saanud nelja pea korral hinnata sobivat mudelit, kui alglähendid olid leitud pea 5 K-keskmiste klastrite põhjal.

Soovisime võrrelda iga pea mudeli korral parameetreid. Joonisel 8 on toodud saadud mudelite komponentide keskväärtused (homogeensed pead on toodud eespool paremaks võrdlemiseks).

Sarnaselt K-keskmiste peatükiga eeldame ka edaspidi, et heade peade jaotus on tõene ehk tegelik jaotus.



Joonis 8. Peade 1-9 ja heade peade Gaussi segumudeli komponentide keskväärtused

K-keskmiste tulemuste peatükis järeldasime, et mittehomogeensete peade erinevus võib seisneda selles, et nende MRT tunnuste väärtused on suuremad - antud järeldust kinnitavad ka komponentide keskväärtused. Peal 7 on mitmete komponentide korral suuremad keskväärtused kui homogeensetel peadel tunnuste ute30 ja ge30 korral. Pea 5 ja pea 6 erinevus mitmete

komponentide puhul tuleb välja aga tunnuste ute10 ja ge10 korral. Keskväärtuste tabel on toodud lisas 4.

Lisaks keskväärtustele vaatasime ka kõigi peade jaoks erinevate komponentide kaalude ja standardhälvete hinnanguid (vaata lisa 4). Suuremat standardhälvete erinevust võib samuti täheldada just suuremate MRT väärtustega komponentides, st komponentide neli ja eriti komponendi viis korral. Tunnuse ute10 standardhälbed erinevad heade peade mudeli standardhälbest komponendis viis märgatavalt kõigi nelja halva pea korral. Heade peade mudeli korral on standardhälve 86, peade 3, 6 ja 7 korral on standardhälve ~140. Pea 5 korral on standardhälve aga ligi kaks korda suurem (167). Erinevusi heade ja halbade peade vahel esineb ka teiste tunnuste korral, aga eriti kerkib esile pea 7, mille korral erineb standardhälve kõigi tunnuste korral komponendis viis. Seega üheks põhjuseks, miks peade mudelite jäägid käitusid erinevalt, võivad olla just suured standardhälvete erinevused.

Paneme tähele, et kõikide mudeli parameetrite korral on viie homogeense pea koondandmestiku segumudeli parameetrid väga sarnased pea 8 ja pea 9 segumudelite parameetritega, samas esinevad väikesed erinevused teiste homogeensete peadega. Sellest võib järeldada, et pead 8 ja 9 domineerivad heade peade mudeli sobitamisel.

Peade jaotuste erinevuste võrdlemiseks kasutasime taaskord Kullback-Leibleri kaugusi. Selle jaoks genereerisime kümme andmestikku tõesest jaotusest, kus parameetrite hinnangud olid leitud heade peade koondandmestiku põhjal. Iga andmestiku jaoks leidsime Kullback-Leibleri kaugused, kasutades varasemalt toodud valemit (1.3) tihedustega $p(x)$ ja $q_j(x),\ j = 1, \dots, 9$. Tabelis 11 esitatud Kullback-Leibleri kaugused $KL(P \parallel Q_j)$ on kümne kauguse keskmised.

Tabel 11. Gaussi segumudeleid kasutades leitud Kullback-Leibleri kaugused

	Pea 1	**Pea 2**	Pea 3	**Pea 4**	Pea 5	Pea 6	Pea 7	**Pea 8**	**Pea 9**
KL	0,228	0,199	0,423	0,154	2,168	1,081	1,228	0,057	0,060

Paneme tähele, et kõigi heade peade korral on Kullback-Leibleri kaugus väiksem kui 0,25, samas halbade peade korral on kaugused suuremad kui 1, välja arvatud pea 3 korral. Kõige rohkem erineb Kullback-Leibleri kauguse põhjal heade peade jaotusest pea 5 jaotus.

Kokkuvõte

Käesolevas magistritöös uuriti erinevusi KT ja MRT tunnuste jaotuses üheksa patsiendi pea andmete korral. Varasemalt läbi viidud modelleerimisülesande [1] põhjal tulid välja viis homogeenset pead, mis olid mudeli jääkide käitumise põhjal sarnased, ja neli pead, mis olid erinevad homogeensetest peadest. Eesmärk oli uurida, kas peadevahelisi erinevusi on võimalik tuvastada Kullback-Leibleri kauguse abil.

Esmalt võrreldi K-keskmiste meetodiga ühe hea ja kahe halva pea korral klastrite keskpunkte KT ja MRT tunnuste jaoks. Osutus, et erinevused ei olnud selgelt märgatavad KT mõõtmiste korral ja seega vaadeldi edaspidi MRT mõõtmistulemusi. Kasutades viie homogeense pea klastrite keskpunkte, paigutati vokslid klastritesse iga pea korral. Eeldati, et viie homogeense pea klastrite suurused esindasid tegelikku jaotust ja seda võrreldi iga pea jaotusega leides vastavad Kullback-Leibleri kaugused. Antud metoodikat rakendati kogu pea andmete ja peade sisemise osa jaoks. Kogu pea andmete korral erinesid viie homogeense pea jaotusest kõik neli mittehomogeenset pead. Peade sisemise osa korral täheldati peade 5, 6 ja 7 jaotuste erinevust heade peade jaotusest. Klastrite suuruste põhjal järeldati, et mittehomogeensete peade korral on MRT tunnustel üldiselt suuremad väärtused.

Järgmisena hinnati viie homogeense pea koondandmestiku ja iga pea jaoks eraldi Gaussi segumudelid. Leitud komponentide keskväärtused kinnitasid samuti pea 5, pea 6 ja pea 7 jaoks suuremate MRT tunnuste väärtuste olemasolu. Seejärel leiti Kullback-Leibleri kaugused heade peade korral hinnatud tegeliku jaotuse ja iga pea jaotuse vahel. Tõesest jaotusest erinesid pead 5, 6 ja 7. Lisaks selgus segumudelite parameetrite võrdlemisel, et suuremate MRT väärtustega komponendis viis olid mittehomogeensetel peadel märgatavalt suuremad standardhälbed. Antud tulemus võib selgitada modelleerimisülesandes [1] välja tulnud mudeli jääkide erinevusi. Standardhälvete erinevuste põhjusi võiks edaspidi lähemalt uurida.

K-keskmiste meetod ja segumudelid on väga erinevad statistilised meetodid. Esimene põhineb ainult vaatlustevahelistel kaugustel, teise korral sobitatakse andmetele konkreetne jaotus. Mõlema meetodi korral küll ilmnes, et halbade peade korral on suuremad MRT väärtused, samas aga erinesid homogeensete peade keskväärtused ehk nn tegelikud keskväärtused meetodite korral. Seega ei pruugi olla K-keskmiste tulemuste kasutamine alglähenditena kõige parem valik ning tuleks kindlasti vaadata juhuslike alglähenditega hinnatud mudeleid. Mõlema meetodi korral ilmnesid jaotuste erinevused Kullback-Leibleri kauguste põhjal. Samas tuleb tähele panna, et Kullback-Leibleri kaugus on suhteline, seega on keeruline täpselt määratleda, kui palju

erinesid halbade peade jaotused homogeensetest peadest. Üheks võimalikuks töö edasiarenduseks on uurida jaotustevahelisi erinevusi ka teiste kaugusmõõtude abil.

Antud töös ei võetud K-keskmiste meetodit ja Gaussi segumudeleid kasutades arvesse vokslitevahelist sõltuvust. Kuna erinevusi täheldati sõltumatute vaatluste korral, siis järgmisena saaks erinevuste uurimiseks rakendada meetodeid, mille korral võetakse sõltuvust arvesse. Segumudelite korral uuriti paketi *Mclust* abil, milline mudel oleks kõige parem sobitada pea andmetele. Bayesi informatsiooni kriteeriumi järgi osutusid parimateks mudeliteks kitsendusteta mudelid 8 või 9 komponendiga. Seega edaspidi võiks uurida ka suuremat komponentide arvu MRT tunnuste korral.

Viited

[1] K. Kuljus, F. L. Bayisa, D. Bolin, J. Lember & J. Yu, „Comparison of hidden Markov chain models and hidden Markov random field models in estimation of computed tomography images,“ *Arxiv: 1705.01727.*, 2017. Avaldamiseks vastu võetud ajakirjas *Communications in Statistics: Case Studies, Data Analysis and Applications.*

[2] T. Hastie, R. Tibshirani & J. Friedman, The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition, Springer, 2008.

[3] A. Kassambara, Practical Guide to Cluster Analysis in R, Unsupervised Machine Learning, STHDA, 2017.

[4] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[5] J. Hartigan & M. Wong, „Algorithm AS 136: A K-Means Clustering Algorithm,“ *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, kd. 28, lk. 100-108, 1979.

[6] J. Lember, „Informatsiooniteooria. Loengukonspekt ja ülesanded,“ 2013. [Võrgumaterjal]. https://www.math.ut.ee/sites/default/files/ms/informatsiooniteooria_kevad_2013.pdf. [Kasutatud 23.04.2018].

[7] P. Kaasik, „Mitmemõõtmeliste andmete visualiseerimine meetodi t-SNE abil,“ [bakalaureusetöö], Tartu, 2017.

[8] T. Naidich, M. Castillo, C. Soonmee & J. Smirniotopoulos, Imaging of the Brain, 1st Edition, China: Saunders, 2012.

Lisad

Lisa 1. KT väärtused tavalisemate kudede korral

Tabel 12. KT väärtused tavalisemate kudede korral [8]

Kude	KT väärtus
Õhk	< -1000
Rasv	-20 kuni -100
Vesi	-20 kuni 20
Valgeollus	20 kuni 35
Hallollus	30 kuni 40
Lihaskude	20 kuni 40
Akuutne verejooks	50 kuni 100
Kaltsifikatsioon	> 150
Luu	800 kuni 1200

Lisa 2. Gaussi segumudelite parameetrite hindamise programmikood rakendustarkvaras R

```
# Pea 1 Gaussi segumudeli leidmine, kasutades pea 5 K-keskmiste tulemuste põhjal leitud parameetrite alglähendeid

library(Mclust)

g <- 5   #komponentide arv

data  <- pea1    #andmestik Gaussi segumudelite jaoks

p <- ncol(pea5);  n <- nrow(pea5)

km=kmeans(pea5,centers=5,algorithm="Lloyd",nstart=20,iter.max= 500)  #pea 5 klasterdamine

par <- vector("list", g)    #algparameetrite vektor

par$pro <- km$size/n    #kaalud

par$mean <- t(km$centers) #keskmised

sigma <- array(NA, c(p, p, g)) #kovariatsioonimaatriksid

new <- as.data.frame(cbind(pea5, km$cluster))
```

```
for (i in 1 : g) {
  subdata <- subset(new[, 1 : p], new[, (p+1)]==i)
  sigma[,, i] <- cov(subdata) }
variance <- mclustVariance("VVV", d = p, G = g)
par$variance <- variance
par$variance$sigma <- sigma
gmm <- em(modelName = "VVV", data = data, parameters = par)

#Pea 1 Gaussi segumudel juhusliku alglähendiga
gmm_juh <- Mclust(modelName = "VVV",G=5, data = pea1)
```

Lisa 3. Erinevate alglähenditega hinnatud Gaussi segumudelite logaritmilise tõepära väärtused

Tabel 13. Erinevate alglähenditega hinnatud Gaussi segumudelite logaritmilise tõepära väärtused

Alglähend	Head pead	Pea 1	Pea 2	Pea 3	Pea 4	Pea 5	Pea 6	Pea 7	Pea 8	Pea 9
Pea 1	-189337188	-35530761	-34105406	-31815809	-39550032	-33505378	-34913010	-26856657	-39127207	-40455029
Pea 2	-189519525	-35530805	-34152970	-31815869	-39549583	-33505329	-34898268	-26855643	-39127029	-40435478
Pea 3	-189685920	-35601945	-34153614	-31823022	-39553546	-33505196	-34931841	-26867198	-39170433	-40482370
Pea 4	**-189336787**	-35530760	-34142005	-31815964	-39549656	-33505437	-34912786	-26855696	**-39104806**	-40455325
Pea 5	-189850013	NA	NA	-31905669	NA	-33541446	-34935240	**-26843994**	NA	-40496887
Pea 6	-189847948	-35603760	-34177089	-31880985	-39553708	-33505858	-34934807	-26873207	-39115260	-40497057
Pea 7	-189732374	-35611513	-34159039	-31822970	-39594559	**-33505115**	-34931707	-26872896	-39126540	-40484086
Pea 8	-189514132	-35530699	-34152694	-31852315	-39548928	-33505700	-34918406	-26855942	-39126830	**-40435264**
Pea 9	-189518828	-35530697	-34154328	-31816053	-39549707	-33505232	-34898236	-26855801	-39127226	-40496526
Head pead	-189338513	-35530854	-34153005	-31815699	-39549976	-33505446	**-34898211**	-26856071	-39127046	-40435421
Juhuslik 1	-189522664	**-35484840**	**-34083159**	-31805919	-39548981	-33506020	-34902420	-26848240	-39123437	-40508901
Juhuslik 2	-189439865	-35485366	-34119591	-31848258	-39547946	-33536998	-34984296	-26848329	-39108566	-40449887
Juhuslik 3	-189521697	-35484859	-34164416	**-31805877**	**-39534274**	-33536544	-34930174	-26848183	-39108874	-40435496

Lisa 4. Gaussi segumudelite komponentide parameetrite hinnangud

Tabel 14. Peade 1-9 ja heade peade Gaussi segumudeli komponentide kaalude hinnangud

Pea	1	2	3	4	5
Head pead	0,165	0,152	0,357	0,15	0,176
Pea 1	0,181	0,149	0,206	0,29	0,173
Pea 2	0,163	0,151	0,196	0,262	0,228
Pea 4	0,168	0,151	0,216	0,254	0,211
Pea 8	0,162	0,166	0,358	0,146	0,167
Pea 9	0,17	0,133	0,336	0,189	0,172
Pea 3	0,16	0,11	0,142	0,281	0,306
Pea 5	0,17	0,123	0,372	0,168	0,167
Pea 6	0,175	0,156	0,351	0,142	0,176
Pea 7	0,181	0,194	0,172	0,267	0,186

Tabel 15. Gaussi segumudelite komponentide keskväärtuste hinnangud

Pea	MRT tunnus	1	2	3	4	5
Head pead	ute10	60	119	209	245	409
	ge10	28	93	260	213	318
	ute30	42	77	73	136	254
	ge30	25	42	80	98	190
Pea 1	ute10	55	136	197	233	391
	ge10	28	170	153	280	312
	ute30	40	49	116	87	248
	ge30	24	55	70	90	192
Pea 2	ute10	62	137	182	209	392
	ge10	29	179	149	270	320
	ute30	45	48	109	73	254
	ge30	25	56	71	83	199
Pea 4	ute10	65	133	189	226	401
	ge10	29	167	141	278	314
	ute30	45	47	119	78	243
	ge30	25	53	69	83	181
Pea 8	ute10	59	126	205	241	395
	ge10	29	87	261	204	299
	ute30	42	90	73	140	237
	ge30	25	43	81	96	172
Pea 9	ute10	62	122	218	253	416
	ge10	29	100	271	228	317
	ute30	42	75	74	132	253
	ge30	24	42	79	99	186
Pea 3	ute10	69	97	155	208	394
	ge10	32	142	103	266	327
	ute30	52	40	128	67	251
	ge30	26	52	70	78	190
Pea 5	ute10	89	165	281	350	647
	ge10	43	101	337	277	484
	ute30	41	81	70	129	262
	ge30	24	38	76	90	205
Pea 6	ute10	84	152	269	325	554
	ge10	38	108	321	252	415
	ute30	44	81	79	144	262
	ge30	23	39	83	91	196
Pea 7	ute10	66	170	182	204	443
	ge10	32	207	122	264	336
	ute30	62	73	142	90	342
	ge30	34	82	75	98	232

Tabel 16. Gaussi segumudelite komponentide kovariatsioonimaatriksite diagonaalide ruutjuure hinnangud (st tunnuste standardhälvete hinnangud erinevates komponentides)

Pea	MRT tunnus	1	2	3	4	5
Head pead	ute10	27	49	41	54	86
	ge10	7	41	47	71	101
	ute30	17	39	19	45	75
	ge30	6	12	17	35	79
Pea 1	ute10	23	56	66	32	90
	ge10	7	54	78	31	91
	ute30	15	19	39	18	82
	ge30	6	15	32	14	79
Pea 2	ute10	28	53	61	28	97
	ge10	7	54	79	32	112
	ute30	19	16	39	15	81
	ge30	6	15	33	14	82
Pea 4	ute10	28	58	71	30	97
	ge10	8	57	79	32	99
	ute30	18	17	43	17	77
	ge30	7	15	35	14	74
Pea 8	ute10	28	48	40	53	76
	ge10	7	40	47	60	98
	ute30	17	40	17	50	67
	ge30	5	13	16	31	74
Pea 9	ute10	28	53	39	58	83
	ge10	7	44	46	89	100
	ute30	17	40	19	43	73
	ge30	6	12	16	37	78
Pea 3	ute10	30	42	60	31	135
	ge10	8	46	50	38	111
	ute30	21	13	43	17	100
	ge30	5	15	29	16	75
Pea 5	ute10	41	64	57	90	167
	ge10	12	46	59	102	143
	ute30	16	35	18	46	72
	ge30	6	11	15	34	73
Pea 6	ute10	36	68	53	86	138
	ge10	10	48	57	86	131
	ute30	17	43	20	61	78
	ge30	6	12	18	33	82
Pea 7	ute10	31	70	59	36	145
	ge10	8	67	66	33	147
	ute30	28	26	50	21	120
	ge30	8	25	31	19	114

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Birgit Kadastik,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Magnetresonantstomograafia (MRT) ja kompuutertomograafia (KT) andmete jaotuste võrdlemine,

mille juhendaja on Kristi Kuljus,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **15.05.2018**