Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Kristiina Uusna

Mudelipõhise klasteranalüüsi rakendamine Eesti Haigekassa andmetele

Matemaatika ja statistika õppekava

Matemaatilise statistika eriala

Magistritöö (30 EAP)

Juhendaja: vanemteadur Kristi Kuljus

Tartu 2020

Mudelipõhise klasteranalüüsi rakendamine Eesti Haigekassa andmetele

Magistritöö
Kristiina Uusna

Lühikokkuvõte. Magistritöös analüüsitakse Eesti Haigekassale saadetud raviarvete põhjal moodustatud patsientide ja neile määratud diagnooside andmekogu. Eesmärk on anda ülevaade kvalitatiivsete tunnustega andmetele rakendatud mudelipõhisest klasteranalüüsist. Töös tuuakse välja andmete klasterdamiseks kasutatud segumudeli kuju. Kirjeldatakse ära EM-algoritm, mida rakendatakse mudeli parameetrite hindamiseks. Lisaks antakse ülevaade integreeritud klssifitseerimistõepära (ICL) kriteeriumist, mille abil leitakse sobivaim segumudel klasterdamiseks.

Uurimise alla on võetud psüühika- ja käitumishäiretega ning vereringeelundite haigustega patsiendid. Klasteranalüüs viiakse eraldi läbi iga valitud vanusegrupi jaoks. Tulemustest selgub näiteks, et psühhoaktiivsete ainete tarvitamisest tingitud psüühika- ja käitumishäiretega patsientidest enamuse moodustavad mehed. Vereringeelundite haigusi uurides aga selgub, et kõige rohkem on patsiente kõrgvererõhkhaigustega, mida seejuures nooremas eas isikutel (vanuses 20-49) esineb rohkem meestel kui naistel. Samuti on esinenud noorematel meestel südame isheemiatõvesid ligi kaks korda rohkem kui naistel.

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

Märksõnad: mudelipõhine klasteranalüüs, segujaotused, EM-algoritm, ICL, R (programmeerimiskeel)

Application of model based clustering to the data of Estonian Health Insurance Fund

Master's thesis

Kristiina Uusna

Abstract. This Master's thesis analyses the dataset of patients and their diagnoses based on treatment bills which have been sent to the Estonian Health Insurance Fund. The purpose of the thesis is to give an overview of model-based clustering method for qualitative data. The description of the mixture model used for data clustering is presented. Also, the EM algorithm which is used for estimating the model parameters is given. In this work, the Integrated Completed Likelihood (ICL) criterion is used for choosing the most suitable mixture model for clustering. The thesis is focused on patients with mental and behavioural disorders and patients who have diseases of the circulatory system. Cluster analysis is performed separately for certain age groups. The results show that there are more men than women with mental and behavioural disorders due to psychoactive substance use. Also, it turns out that the number of patients with hypertensive diseases is biggest compared to other diseases of the circulatory system. There are more men than women with hypertensive diseases when we look at the age groups 20-29 and 40-49. In addition, ischaemic heart diseases are nearly twice as common for men of age 20-29 compared to women of the same age.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics

Keywords: model-based cluster analysis, mixture distribution, EM algorithm, ICL, R (programming language)

Sisukord

Sissejuhatus

Eesti Haigekassa on avalik-õiguslik organisatsioon, mille tähtsaim ülesanne on korraldada riiklikku ravikindlustust. Haigekassa võimaldab kindlustatud inimestele hüvitisi ning arstiabi. Teenuste eest tasumiseks peab tervishoiuteenuse osutaja haigekassale esitama raviarve, kus on kajastatud ühele kindlustatud isikule kogu haigusjuhu käigus tehtud terviseuuringud ning osutatud teenused. Sedasi moodustub ravikindlustuse andmekogu, mille põhjal on võimalik saada ka terviklik ülevaade Eesti patsientide terviseprobleemidest.

Haiguste ennetamiseks, sealhulgas kaasuvate terviseprobleemide ärahoidmiseks ja vastava ravi leidmiseks, tuleb eelnevalt samade haigusnähtudega patsiente lähemalt analüüsida. Esile kerkivad küsimused, kas mõni diagnoos esineb naistel rohkem kui meestel või vastupidi. Kuidas muutub erinevate diagnooside esinemise sagedus ajas, vaadates seejuures eri vanuses patsiente? Kui suur on tõenäosus, et patsiendid, kes kuuluvad vaadeldavasse diagnoosigruppi, on ära märgitud ka mõnes teises konkreetses haigusgrupis? Käesolevas töös on keskendutud psüühika- ja käitumishäiretega ning vereringeelundite haigustega patsientide uurimisele, kasutades selleks klasteranalüüsi meetodit.

Klasteranalüüs on statistiline meetod, mida peamiselt kasutatakse andmeanalüüsis, kuid on leidnud rakendust ka pilditöötluses, turu-uuringutes ning paljudes teistes erinevates valdkondades. Klasterdamise eesmärk on vaatluste grupeerimine. Vaatlused, mis kuuluvad samasse gruppi, peaksid olema omavahel mingis mõttes sarnased. Vaatlused, mis on aga määratud erinevatesse gruppidesse, peaksid olema võimalikult erinevate omadustega. Tuntumad klasteranalüüsi meetodid on hierarhiline klasterdamine ning K-keskmiste meetod. Need meetodid põhinevad erinevusmõõtudel, kus täpsemalt K-keskmiste meetodi korral kasutatakse Eukleidilise kauguse mõõtu. Tulemused aga võivad suuresti sõltuda sellest, kuidas on tunnuseid enne meetodite rakendamist teisendatud või standardiseeritud. Alternatiiv on kasutada mudelipõhist klasteranalüüsi, mis eeldab, et klasterdatavate vaatluste jaotuse kirjeldamiseks sobib mitmemõõtmeline segujaotus.

Käesolevas töös on mudelipõhist klasteranalüüsi rakendatud andmetele, mille kõik tunnused on kvalitatiivsed. Töö eesmärk on anda ülevaade nimetatud klasterda-

mise meetodist ning selle rakendamisest mittearvuliste tunnustega andmetele. Töö on jagatud nelja peatükki. Esimeses peatükis antakse üldisem ülevaade andmestikust, kasutatavatest tunnustest ning nende tähendustest. Teises osas viiakse lugeja kurssi kasutatava klasteranalüüsi meetodiga, kus kirjeldatakse ära mudeli kuju, parameetrite hindamise protsess ning kriteerium, mille abil valitakse välja parim mudel. Kolmandas peatükis tehakse kirjeldav analüüs, kus antakse põhjalikum ülevaade töös kasutatavast andmestikust. Neljandas peatükis on läbi viidud käesoleva magistritöö praktiline osa, kus võib näha haigekassa andmetele rakendatud klasteranalüüsi tulemusi.

Töö on koostatud tekstikujundustarkvaraga LaTeX. Praktilise osa läbiviimiseks on kasutatud statistikatarkvara R ja selle lisapaketti „Rmixmod“.

Autor soovib tänada töö juhendajat Kristi Kuljust kasulike nõuannete, paranduste ja suunamise eest ning lisaks Sven Lauri, kes aitas muretseda vajaliku andmestiku töö praktilise osa läbiviimiseks.

1 Andmestiku ülevaade

Töös uuritav andmestik koosneb patsientidest, kelle kohta on esitatud haigekassa raviarve aastatel 2008-2018. Iga rida kirjeldab patsiendile määratud diagnoose mingil kindlal aastal. Seega, kui patsiendile on väljastatud raviarve mitmel aastal, siis on tema kohta ka vastav arv ridu andmestikus. Aastaid arvestatakse patsiendi sünnikuupäevast lähtudes ehk kui patsiendi esimene sissekanne on tehtud tema 30. sünnipäeval, siis viimane sissekanne saab olla kui ta just sai 40-aastaseks. Seega iga patsiendi kohta on andmestikus maksimaalselt 11 rida. Välja on toodud kõik diagnoosid, millega seotud raviteenuse osutamise eest on sellel aastal väljastatud haigekassale tasumiseks arve.

Andmestiku esialgne maht on 10 675 801. Esindatud on 1 480 012 erinevat patsienti, kellest 694 545 on mehed ja 785 467 on naised, mis näitab, et tegemist on praktiliselt kogu Eesti rahvastikuga.

Kirjeldatud andmestik sisaldab viit erinevat tunnust: patsiendi id, sugu, vanus, põhidiagnoos ja kõrvaldiagnoos. Põhidiagnoos on patsiendile määratud diagnoos, millega seoses on vaadeldaval aastal haigekassale saadetud arve. Diagnoos näitab patsiendi peamist põhjust/kaebust, mille pärast arsti poole pöörduti. Kõrvaldiagnoos on põhidiagnoosiga kaasnev mure ehk diagnoos, mis võib mõjutada ravi, aga mis ei ole fookuses. Näiteks on ühel patsiendil peadiagnoosiks määratud „Soole viirus- ja muud täpsustatud nakkused“ ning kõrvaldiagnoosiks on märgitud „Immuniseerimise vajadus nakkushaiguste kombinatsiooni vastu“. Ühel teisel juhul aga on näiteks patsient tulnud arsti juurde seljavalu murega ning lisaks on tal diagnoositud mao-söögitoru tagasivooluhaigus, mis kuulub kõrvaldiagnoosi alla. Käesolevas töös pole kõrvaldiagnoose klasteranalüüsi kaasatud.

Patsientide diagnooside määramisel kasutatakse Rahvusvahelise Haiguste Klassifikatsiooni RHK-10 (inglise keeles ICD-10 *International Statistical Classification of Diseases and Related Health Problems*)(World Health Organization (2016)). Igal haigusel, häirel, vigastusel või seisundil on oma kood, mis koosneb tähest ja kahest numbrist, millele võib veel järgneda punktiga eraldatud numbreid. Mida pikem on kood, seda täpsemalt see diagnoosi kirjeldab.

2 Mudelipõhine klasteranalüüs kvalitatiivsete tunnustega andmete korral

Antud töös uurimise alla võetud andmestik sisaldab ainult kategoorilisi tunnuseid. Seda tüüpi andmetega analüüsi läbiviimiseks kasutatakse käesolevas töös mudelipõhist klasteranalüüsi kvalitatiivsete tunnustega andmete jaoks. Selles peatükis kirjeldame lähemalt mudelipõhise klasteranalüüsi metoodikat, mudeli parameetrite hindamist ning kitsenduste rakendamist mudelile. Järgnevas klasteranalüüsi arutelus on toetutud peatükile 9 raamatus Hennig jt (2016).

2.1 Mitmemõõtmeline segumudel kvalitatiivsete tunnuste korral

Olgu meil vaja klasterdada vaatlused $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, mida iseloomustavad d kategoorilist tunnust. Igal tunnusel j on m_j võimalikku väärtust, $j = 1, ..., d$. Tähistame h-ga selle, mitmendat tunnuse j väärtust parasjagu vaatame ($h = 1, ..., m_j$). Iga vaatlus $\boldsymbol{x}_i$ on esitatud kujul $(x_i^j, j = 1, ..., d)$, mis on samaväärne esitusviisiga $(x_i^{jh}; j = 1, ..., d; h = 1, ..., m_j)$, nii et

$$\begin{cases} x_i^{jh} = 1, & \text{kui } x_i^j = h, \\ x_i^{jh} = 0, & \text{mujal,} \end{cases}$$

kus $x_i^j = h$ tähendab, et selle vaatluse j-ndal tunnusel on h-s võimalik väärtus. Seega on iga vaatlus $\boldsymbol{x_i}$ esitatav binaarse vektorina

$$\boldsymbol{x_i} = (x_i^{11}, ..., x_i^{1m_1}; x_i^{21}, ..., x_i^{2m_2}; ...; x_i^{d1}, ..., x_i^{dm_d}).$$

Eeldame, et iga vaatlus $\boldsymbol{x}_i$ pärineb K mitmemõõtmelise multinominaaljaotuse segust, mille tihedusfunktsioon avaldub järgmiselt:

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \mathcal{M}_k(\boldsymbol{x}; \boldsymbol{\alpha_k}) = \sum_{k=1}^{K} p_k \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x^{jh}}, \tag{1}$$

kus $\boldsymbol{x} = (x^{11}, ..., x^{1m_1}; x^{21}, ..., x^{2m_2}; ...; x^{d1}, ..., x^{dm_d})$ ja α_k^{jh} on tõenäosus, et tunnusel j on h-s võimalik väärtus, kui $\boldsymbol{x}$ on selle segujaotuse k-nda komponendi realisatsioon ning $\boldsymbol{\alpha_k} = (\alpha_k^{jh}; j = 1, ..., d; h = 1, ..., m_j)$. Komponentide kaalude vektorile anname tähistuse $\boldsymbol{p} = (p_1, ..., p_K)$. Paneme tähele, et $p_k \geqslant 0$ ja $\sum_{k=1}^{K} p_k = 1$. Vektor $\boldsymbol{\theta} = (p_k, \boldsymbol{\alpha_k}, k = 1, ..., K)$ sisaldab segujaotuse kõigi parameetrite hulka, mida soovime hinnata.

Segujaotuse mudel (1) eeldab, et vaatlused on mitmemõõtmelise segujaotusega ning mõõdetud d kvalitatiivset tunnust on iga komponendi k sees sõltumatud. Sellist tinglikku sõltumatust nimetatakse ka lokaalseks sõltumatuseks. Lokaalse sõltumatuse eeldus võimaldab meil tihedusfunktsiooni (1) välja kirjutada selliselt, et iga komponent sisaldab d multinominaalse jaotuse korrutist. Iga sellist korrutist tähistame sümboliga $\mathcal{M}_k(\boldsymbol{x}; \boldsymbol{\alpha_k})$. Lokaalse sõltumatuse mõiste on lahti seletatud ka magistritöös Mirski (2019), kus on välja toodud, et peamine põhjus lokaalse sõltumatuse eelduse tegemiseks seisneb selles, et diskreetsete tunnuste vahelise sõltuvuse iseloomustamiseks ei leidu tavalist näitajat nagu seda on lineaarne korrelatsioonikordaja kahe kvantitatiivse tunnuse vahel. Näiteks normaaljaotuse segu korral modelleerivad komponentide kovariatsioonimaatriksid kvantitatiivsete tunnuste vahelist sõltuvusstruktuuri, kuid kvalitatiivsete tunnuste korral pole see võimalik. Raamatus Hennig jt (2016) leheküljel 174 on välja toodud mitu tööd, kus on ära näidatud viisid, mille korral ei nõuta lokaalse sõltumatuse eeldust. Sellised mudelid on aga keerulisemad ja nõuavad rohkemate parameetrite hindamist. Mitmed allikad, nagu näiteks Anderlucci ja Hennig (2014), väidavad, et kuigi lokaalse sõltumatuse eeldus võib näida väga kitsendav, siis praktikas annab see eeldus üsna häid tulemusi, kus klastrid on hästi interpreteeritavad.

2.2 EM-algoritm

Multinominaalsete jaotuste segu parameetrite hindamiseks kasutatakse EM-algoritmi (*expectation-maximization algorithm*). EM-algoritm on iteratiivne meetod parameetrite hindamiseks, mille abil saab leida suurima tõepära hinnangud latentsete (varjatud) tunnustega mudeli parameetritele. Selles peatükis kirjeldamegi EM-algoritmi üldist ideed (allikana on kasutatud raamatut Bishop (2006), kui ei ole viidatud teisiti) ja selle kasutamist multinominaalsete jaotuste segu parameet-

rite hindamiseks (allikaks on raamat Hennig jt (2016)).

Olgu $\boldsymbol{x} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ kogu vaatluste hulk (hulk, mis sisaldab kõigi nende tunnuste väärtuseid, mida on konkreetselt mõõdetud iga vaatluse korral) ning vastavate latentsete tunnuste väärtused olgu määratud hulgaga $\boldsymbol{z} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_n\}$, kus $\boldsymbol{z}_i = (z_{i1}, ..., z_{iK})$, $i = 1, .., n$, ja $z_{ik} = 1$, kui $\boldsymbol{x}_i$ on segujaotuse k-nda komponendi realisatsioon ning $z_{ik} = 0$ vastasel juhul. Seega $\sum_{k=1}^{K} z_{ik} = 1$. Vaatluste hulka $\{\boldsymbol{x}, \boldsymbol{z}\}$ nimetatakse täielikuks andmestikuks, kusjuures $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ väärtusi me praktikas ei tea. Järgnev EM-algoritmi idee on selgitatud ära ka magistritöös Mirski (2019), mis põhineb raamatul Bishop (2006). Tähistame täieliku andmestiku $\{\boldsymbol{x}, \boldsymbol{z}\}$ logaritmilise tõepärafunktsiooni $\ell(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) := \ln p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$, kus $\boldsymbol{\theta}$ viitab segujaotuse kõigi parameetrite hulgale. Kuna vektorid $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ on tundmatud, ei saa me funktsiooni $\ell(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$ praktikas maksimiseerida. Kui aga maksimiseerida vaatluste $\boldsymbol{x}$ logaritmilist tõepärafunktsiooni kujul $\ell(\boldsymbol{x}; \boldsymbol{\theta})$, sisaldavad leitud parameetrite hinnangud kaudselt varjatud tunnuseid. Selle probleemi lahendamiseks kasutatakse vektorite $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ kohta teadaolevat informatsiooni, mis on kirjeldatav tingliku jaotuse $p(\boldsymbol{z}_1, ..., \boldsymbol{z}_n \mid \boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta})$ abil. Funktsiooni $\ell(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$ asemel töötatakse tingliku keskväärtusega $E[\ell(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta}) \mid \boldsymbol{x}]$, kus $\boldsymbol{Z} = \{\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n\}$ on indikaatoritele $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ vastavate multinominaalse jaotusega juhuslike vektorite hulk. Keskväärtuse arvutamine on EM-algoritmi keskväärtuse leidmise samm (E-samm). Sellele järgneb maksimeerimise samm (M-samm), kus leitakse parameetrid, mis maksimeerivad tingliku keskväärtuse. Neid samme korratakse nii kaua kuni algoritm koondub.

Tähistagu $\boldsymbol{\theta}^{(r)}$ parameetrite esialgseid hinnanguid ning EM-algoritmi abil saadud uued hinnangud olgu tähistatud $\boldsymbol{\theta}^{(r+1)}$-ga. Võtame kokku EM-algoritmi sammud:

1. Fikseerida algväärtus $\boldsymbol{\theta}^{(r)} = \boldsymbol{\theta}^{(0)}$ ja arvutada logaritmilise tõepärafunktsiooni $\ell(\boldsymbol{x}; \boldsymbol{\theta}^{(r)})$ väärtus.

2. **E-samm**: Leida tundmatute indikaatorvektorite tinglik jaotus $p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}^{(r)})$ ning funktsiooni $\ell(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta})$ tinglik keskväärtus

$$
\begin{aligned}
Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) :&= E[\ell(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta}) \mid \boldsymbol{x}; \boldsymbol{\theta}^{(r)}] = \\
&= \sum_{\boldsymbol{z}_1, ..., \boldsymbol{z}_n} \ell(\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{z}_1, ..., \boldsymbol{z}_n; \boldsymbol{\theta}) p(\boldsymbol{z}_1, ..., \boldsymbol{z}_n \mid \boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta}^{(r)}).
\end{aligned}
$$

3. **M-samm**: Leida parameetritele uued hinnangud $\boldsymbol{\theta}^{(r+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)})$.

4. Arvutada $\ell(\boldsymbol{x}; \boldsymbol{\theta}^{(r+1)})$ ning kontrollida algoritmi koondumist logaritmilise tõepärafunktsiooni $\ell(\boldsymbol{x}; \boldsymbol{\theta})$ või parameetrite hinnangute koondumise kaudu. Kui algoritm ei ole koondunud, siis võtta $\boldsymbol{\theta}^{(r)} \longleftarrow \boldsymbol{\theta}^{(r+1)}$ ja naaseda sammu 2 juurde.

Logaritmilise tõepärafunktsiooni $\ell(\boldsymbol{x}; \boldsymbol{\theta}^{(r+1)})$ väärtus siinkohal kasvab või jääb samaks ($\ell(\boldsymbol{x}; \boldsymbol{\theta}^{(r+1)}) \geqslant \ell(\boldsymbol{x}; \boldsymbol{\theta}^{(r)})$). Vaatame täpsemalt, miks see väide kehtib, toetudes raamatule Izenman (2008) lk 455-456.

Paneme tähele, et kuna $p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}) = p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})/p(\boldsymbol{x}; \boldsymbol{\theta})$, siis saame logaritmilise tõepärafunktsiooni $\ell(\boldsymbol{x}; \boldsymbol{\theta})$ kirjutada kujul

$$\ell(\boldsymbol{x}; \boldsymbol{\theta}) = \ln p(\boldsymbol{x}; \boldsymbol{\theta}) = \ell(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) - \ln p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}).$$

Võttes nüüd võrduse mõlemast poolest tingliku keskväärtuse jaotuse $p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}^{(r)})$ suhtes ning tähistades $H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)}) := E[\ln p(\boldsymbol{Z} \mid \boldsymbol{x}; \boldsymbol{\theta}) \mid \boldsymbol{x}, \boldsymbol{\theta}^{(r)}]$ saame, et

$$\begin{aligned}\ell(\boldsymbol{x}; \boldsymbol{\theta}) &= E[\ell(\boldsymbol{x}; \boldsymbol{\theta}) \mid \boldsymbol{x}, \boldsymbol{\theta}^{(r)}] = E[\ell(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta}) \mid \boldsymbol{x}, \boldsymbol{\theta}^{(r)}] - E[\ln p(\boldsymbol{Z} \mid \boldsymbol{x}; \boldsymbol{\theta}) \mid \boldsymbol{x}, \boldsymbol{\theta}^{(r)}] \\ &= Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) - H(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}).\end{aligned}$$

Tähistades nüüd $h(\boldsymbol{z}) := p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta})/p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}^{(r)})$, saame, et $H(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) - H(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)})$ $= E[\ln h(\boldsymbol{Z}) \mid \boldsymbol{x}, \boldsymbol{\theta}^{(r)}] \leqslant E[h(\boldsymbol{Z}) \mid \boldsymbol{x}, \boldsymbol{\theta}^{(r)}] - 1 = 0$, kus on kasutatud teadmist, et kehtib $\ln x \leqslant x - 1$. Seega $H(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) \leqslant H(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)})$. Siit saamegi, et kehtib võrratus

$$\begin{aligned}\ell(\boldsymbol{x}; \boldsymbol{\theta}^{(r+1)}) - \ell(\boldsymbol{x}; \boldsymbol{\theta}^{(r)}) &= Q(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}) - H(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}) \\ &\quad - (Q(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)}) - H(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)})) \\ &= (Q(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}) - Q(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)})) \\ &\quad + (H(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)}) - H(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)})) \\ &\geqslant Q(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}) - Q(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)}) \geqslant 0,\end{aligned}$$

kuna EM-algoritmi kohaselt $\boldsymbol{\theta}^{(r+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)})$, siis $Q(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}) \geqslant Q(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)})$. Seega oleme saanud, et $\ell(\boldsymbol{x}; \boldsymbol{\theta}^{(r+1)}) \geqslant \ell(\boldsymbol{x}; \boldsymbol{\theta}^{(r)})$.

2.2.1 EM-algoritm multinominaalse segujaotuse parameetrite hindamiseks

Vaatame nüüd EM-algoritmi samme olukorras, kus soovime leida parameetrite hinnangud segujaotusele (1). Kirjutame välja logaritmilise tõepärafunktsiooni

$$\ell(\boldsymbol{x};\boldsymbol{\theta}) = \ln\left(\prod_{i=1}^{n} f(\boldsymbol{x}_i;\boldsymbol{\theta})\right) = \sum_{i=1}^{n} \ln\left(\sum_{k=1}^{K} p_k \mathcal{M}_k(\boldsymbol{x_i};\boldsymbol{\alpha_k})\right).$$

Vaatluste $\{\boldsymbol{x},\boldsymbol{z}\}$ logaritmiline tõepärafunktsioon omab kuju

$$\ell(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) = \ln\left(\prod_{i=1}^{n}\prod_{k=1}^{K}\Big(p_k \mathcal{M}_k(\boldsymbol{x_i};\boldsymbol{\alpha_k})\Big)^{z_{ik}}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln\left(p_k \mathcal{M}_k(\boldsymbol{x_i};\boldsymbol{\alpha_k})\right), \tag{2}$$

sest $(\boldsymbol{x}_1,\boldsymbol{z}_1),...,(\boldsymbol{x}_n,\boldsymbol{z}_n)$ on sõltumatud ning $p(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta}) = p(\boldsymbol{x}_i \mid \boldsymbol{z}_i;\boldsymbol{\theta})p(\boldsymbol{z}_i;\boldsymbol{\theta})$, kus $p(\boldsymbol{z}_i;\boldsymbol{\theta}) = \prod_{k=1}^{K} p_k^{z_{ik}}$ ja $p(\boldsymbol{x}_i \mid \boldsymbol{z}_i;\boldsymbol{\theta}) = \prod_{k=1}^{K}\Big(\mathcal{M}_k(\boldsymbol{x}_i;\boldsymbol{\alpha_k})\Big)^{z_{ik}}$.

Avaldise (2) kujust näeme, et tingliku keskväärtuse $E[\ell(\boldsymbol{x},\boldsymbol{Z};\boldsymbol{\theta}) \mid \boldsymbol{x},\boldsymbol{\theta}^{(r)}]$ arvutamiseks on vaja leida indikaatorile z_{ik} vastava juhusliku suuruse Z_{ik} tinglik keskväärtus:

$$E[Z_{ik} \mid \boldsymbol{x};\boldsymbol{\theta}^{(r)}] = P(Z_{ik}=1 \mid \boldsymbol{x};\boldsymbol{\theta}^{(r)}) = \frac{P(Z_{ik}=1) f(\boldsymbol{x}_i;\boldsymbol{\theta}^{(r)} \mid Z_{ik}=1)}{\sum_{l=1}^{K} P(Z_{il}=1) f(\boldsymbol{x}_i;\boldsymbol{\theta}^{(r)} \mid Z_{il}=1)} =$$
$$= \frac{p_k^{(r)} \mathcal{M}_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k^{(r)})}{\sum_{l=1}^{K} p_l^{(r)} \mathcal{M}_l(\boldsymbol{x}_i;\boldsymbol{\alpha}_l^{(r)})} := t_{ik}^{(r)}.$$

Suurus $t_{ik}^{(r)}$ on tinglik tõenäosus, et vaatlus $\boldsymbol{x}_i$ on parameetritega $\boldsymbol{\theta}^{(r)}$ segujaotuse (1) k-nda komponendi realisatsioon. Seega

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}) = E[\ell(\boldsymbol{x},\boldsymbol{Z};\boldsymbol{\theta}) \mid \boldsymbol{x},\boldsymbol{\theta}^{(r)}] = \sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}^{(r)} \ln\Big(p_k \mathcal{M}_k(\boldsymbol{x_i};\boldsymbol{\alpha_k})\Big). \tag{3}$$

Segujaotuse (1) komponentide kaalud ja iga multinominaalse jaotuse parameetrid summeeruvad üheks. Selleks, et neid kitsendusi arvesse võtta, on parameetrite hinnangute leidmisel abiks Lagrange'i kordajate meetod. Sellisel juhul maksimiseerime

EM-algortimi M-sammul tingliku keskväärtuse (3) asemel funktsiooni

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}) + \lambda\Big(\sum_{k=1}^{K} p_k - 1\Big) + \sum_{k=1}^{K}\sum_{j=1}^{d} \lambda_{kj}\Big(\sum_{h=1}^{m_j} \alpha_k^{jh} - 1\Big),$$

kus lisaparameetrid λ ja $\lambda_{11}, ..., \lambda_{Kd}$ garanteerivad, et parameetrite uued hinnangud rahuldavad eelnimetatud kitsendusi.

Järgmisena vaatamegi parameetrite $\boldsymbol{\theta} = (p_k, \boldsymbol{\alpha_k}, k = 1, ..., K)$ hinnangute leidmise protsessi. Selleks leiame tuletise nii parameetri p_k kui ka α_k^{jh} järgi ning võrdsustame saadud avaldised nulliga. Esmalt vaatame parameetri p_k hinnangu leidmist. Leiame tuletise p_k järgi ning võrdsustame selle nulliga:

$$\begin{aligned}
&\frac{\partial}{\partial p_k}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}^{(r)} \ln\Big(p_k \mathcal{M}_k(\boldsymbol{x_i};\boldsymbol{\alpha_k})\Big) + \lambda\Big(\sum_{k=1}^{K} p_k - 1\Big) + \sum_{k=1}^{K}\sum_{j=1}^{d} \lambda_{kj}\Big(\sum_{h=1}^{m_j} \alpha_k^{jh} - 1\Big)\right] \\
&= \sum_{i=1}^{n} t_{ik}^{(r)} \frac{1}{p_k \mathcal{M}_k(\boldsymbol{x_i};\boldsymbol{\alpha_k})} \mathcal{M}_k(\boldsymbol{x_i};\boldsymbol{\alpha_k}) + \lambda = \sum_{i=1}^{n} \frac{t_{ik}^{(r)}}{p_k} + \lambda,
\end{aligned}$$

$$\sum_{i=1}^{n} \frac{t_{ik}^{(r)}}{p_k} + \lambda = 0, \quad k = 1, ..., K. \tag{4}$$

Paneme tähele, et meil on K võrrandit, mida kõiki vaatame hetkel korraga. Korrutades võrduse (4) mõlemad pooled läbi p_k-ga ning leides summa üle k, saame

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \frac{t_{ik}^{(r)}}{p_k} + \lambda = \sum_{k=1}^{K}\Big(\sum_{i=1}^{n} p_k \frac{t_{ik}^{(r)}}{p_k} + p_k\lambda\Big) = \sum_{k=1}^{K}\Big(\sum_{i=1}^{n} t_{ik}^{(r)} + p_k\lambda\Big) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}^{(r)} + \lambda\sum_{k=1}^{K} p_k = n + \lambda.
\end{aligned}$$

Seega $\lambda = -n$. Asendades λ valemisse (4), saame, et

$$\hat{p}_k = \frac{\sum_{i=1}^{n} t_{ik}^{(r)}}{n}$$

Teiseks vaatame parameetri α_k^{jh} hinnangu leidmist. Leiame esmalt tuletise α_k^{jh} järgi ning võrdsustame saadud avaldised nulliga:

$$\frac{\partial}{\partial \alpha_k^{jh}}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}^{(r)} \ln\Big(p_k \prod_{j=1}^{d}\prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}\Big) + \lambda\Big(\sum_{k=1}^{K} p_k - 1\Big) + \sum_{k=1}^{K}\sum_{j=1}^{d} \lambda_{kj}\Big(\sum_{h=1}^{m_j} \alpha_k^{jh} - 1\Big)\right]$$
$$= \frac{\partial}{\partial \alpha_k^{jh}}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}^{(r)}\Big(\ln p_k + \sum_{j=1}^{d}\sum_{h=1}^{m_j} x_i^{jh} \ln \alpha_k^{jh}\Big) + \lambda\Big(\sum_{k=1}^{K} p_k - 1\Big)\right.$$
$$\left. + \sum_{k=1}^{K}\sum_{j=1}^{d} \lambda_{kj}\Big(\sum_{h=1}^{m_j} \alpha_k^{jh} - 1\Big)\right] = \sum_{i=1}^{n} t_{ik}^{(r)} x_i^{jh} \frac{1}{\alpha_k^{jh}} + \lambda_{kj},$$

$$\sum_{i=1}^{n} t_{ik}^{(r)} x_i^{jh} \frac{1}{\alpha_k^{jh}} + \lambda_{kj} = 0, \quad j = 1, ..., d; \quad h = 1, ..., m_j; \quad k = 1, ..., K. \tag{5}$$

Korrutades võrduse mõlemad pooled läbi α_k^{jh}-ga ning leides summa üle h, saame

$$\begin{aligned} 0 &= \sum_{i=1}^{n} t_{ik}^{(r)} x_i^{jh} \frac{1}{\alpha_k^{jh}} + \lambda_{kj} = \sum_{h=1}^{m_j}\Big(\sum_{i=1}^{n} \alpha_k^{jh} t_{ik}^{(r)} x_i^{jh} \frac{1}{\alpha_k^{jh}} + \alpha_k^{jh}\lambda_{kj}\Big) \\ &= \sum_{h=1}^{m_j}\Big(\sum_{i=1}^{n} t_{ik}^{(r)} x_i^{jh} + \alpha_k^{jh}\lambda_{kj}\Big) = \sum_{i=1}^{n} t_{ik}^{(r)} \sum_{h=1}^{m_j} x_i^{jh} + \lambda_{kj}\sum_{h=1}^{m_j} \alpha_k^{jh} \\ &= \sum_{i=1}^{n} t_{ik}^{(r)} + \lambda_{kj}. \end{aligned}$$

Seega $\lambda_{kj} = -\sum_{i=1}^{n} t_{ik}^{(r)}$. Asendades λ_{kj} valemisse (5), saame, et

$$\hat{\alpha}_k^{jh} = \frac{\sum_{i=1}^{n} t_{ik}^{(r)} x_i^{jh}}{\sum_{i=1}^{n} t_{ik}^{(r)}}.$$

Võtame nüüd kokku EM-algoritmi sammud segujaotuse (1) parameetrite hindamiseks, võttes alglähenditeks $\boldsymbol{\theta}^{(0)} = (\boldsymbol{p}^{(0)}, \boldsymbol{\alpha}^{(0)})$:

- **E-samm**: Leida vektor $\boldsymbol{t}_i^{(r)} = (t_{ik}^{(r)}, k = 1, ..., K)$, $i = 1, ..., n$, kus $t_{ik}^{(r)}$ on tinglik tõenäosus, et $\boldsymbol{x}_i$ on k-nda komponendi realisatsioon:

$$t_{ik}^{(r)} = \frac{p_k^{(r)} \mathcal{M}_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{(r)})}{\sum_{l=1}^{K} p_l^{(r)} \mathcal{M}_l(\boldsymbol{x}_i; \boldsymbol{\alpha}_l^{(r)})}.$$

- **M-samm**: uuendame segujaotuse parameetrite hinnanguid:

$$p_k^{(r+1)} = \frac{\sum_{i=1}^n t_{ik}^{(r)}}{n}, \quad k = 1, ..., K,$$

$$(\alpha_k^{jh})^{(r+1)} = \frac{\sum_{i=1}^n t_{ik}^{(r)} x_i^{jh}}{\sum_{i=1}^n t_{ik}^{(r)}}, \quad j = 1, ..., d; \quad h = 1, ..., m_j; \quad k = 1, ..., K.$$

2.3 Kitsenduste seadmine mudelile

Mitmemõõtmeliste kategooriliste andmete analüüsimine võib osutuda keeruliseks, kui vaatluse all olevad andmed sisaldavad palju tunnuseid, mille võimalike väärtuste arv on suur. Nimelt standardses kitsendusteta segumudelis on hinnatavate parameetrite arv $(K-1) + K\sum_j (m_j - 1)$. Näiteks, kui $K = 6$, $d = 12$, $m_j = 5$ iga tunnuse korral, siis saame mudeli 293 parameetriga. Valimimahu n suhtes võib selline mudel osutuda liiga keeruliseks ning võib isegi tekitada arvulisi keerukusi (esineb „nulliga jagamist“ või saadakse parameetrite α_k^{jh} hinnanguteks nulle). Parameetritele kitsendusi seades saame aga parameetrite arvu vähendada. Selles peatükis vaatamegi lähemalt, kuidas rakendada kitsenduste seadmist mudelile, milliseid erinevaid kitsenduste seadmise võimalusi saame kasutada ning kuidas hinnata kitsendustega mudeli parameetreid.

Järgnev kitsenduste seadmise idee pärineb raamatust Hennig jt (2016) ja artiklist Lebret jt (2015). Tuletame meelde, et vektori $\boldsymbol{\alpha}_k^j = (\alpha_k^{jh}, h = 1, ..., m_j)$ abil tähistasime tõenäosused, kus α_k^{jh} on tõenäosus, et tunnusel j on h-s võimalik väärtus, kui $\boldsymbol{x}$ on segujaotuse k-nda komponendi realisatsioon. Selleks, et kitsenduste tõlgendamist lihtsustada, saame parameetri $\boldsymbol{\alpha}_k^j$ asendada paariga $(\boldsymbol{a}_k, \boldsymbol{\varepsilon}_k)$, kus $\boldsymbol{a}_k = (a_k^1, ..., a_k^d)$ ja $\boldsymbol{\varepsilon}_k = (\varepsilon_k^{11}, ..., \varepsilon_k^{dm_d})$, see tähendab, et iga klastri jaoks näidatakse ära moodi positsioon ja moodi tõenäosus. Siinkohal a_k^j näitab iga tunnuse j jaoks väärtuse positsiooni, millel on kõige suurem tõenäosus ehk $a_k^j = \text{argmax}_h\, \alpha_k^{jh}$ ja $\varepsilon_k^{jh} = 1 - \alpha_k^{jh}$, kui selle j-nda tunnuse h-s võimalik väärtus langeb kokku positsiooniga a_k^j ning $\varepsilon_k^{jh} = \alpha_k^{jh}$ vastasel juhul.

Toome siinkohal ka konkreetse näite. Olgu kahel tunnusel ($j = 2$) vastavalt $m_1 = 2$ ja $m_2 = 4$ võimalikku väärtust, mille tõenäosused k-ndas klastris on $\boldsymbol{\alpha}_k = (0{,}6; 0{,}4; \quad 0{,}2; 0{,}3; 0{,}1; 0{,}4)$, siis uuteks parameetriteks on $\boldsymbol{a}_k = (1; 4)$ ja $\boldsymbol{\varepsilon}_k =$

$(0{,}4; 0{,}4; \quad 0{,}2; 0{,}3; 0{,}1; 0, 6)$.

Kitsenduste seadmise idee seisneb selles, et vektorile $\boldsymbol{\alpha}_k^j$ seatakse tingimus, kus igas klastris k on kõigil tunnustel j üks väärtus, millele on määratud teistest suurem tõenäosus (ehk igal tunnusel on üheselt määratud mood) ning ülejäänud tõenäosusmass jagatakse tunnuse teiste väärtuste vahel võrdselt ära. Seega saame vektorile $\boldsymbol{\alpha}_k^j$ kuju $(\beta_k^j, ..., \beta_k^j, \gamma_k^j, \beta_k^j, ..., \beta_k^j)$, kus γ_k^j on moodi tõenäosus ning kehtib $\gamma_k^j > \beta_k^j$. Kuna $\sum_{h=1}^{m_j} \alpha_k^{jh} = 1$, siis kehtib võrdus $(m_j - 1)\beta_k^j + \gamma_k^j = 1$, millest $\beta_k^j = (1 - \gamma_k^j)/(m_j - 1)$. Seega $\gamma_k^j > 1/m_j$. Kuna kitsendustega mudeli korral

$$\varepsilon_k^{jh} = \begin{cases} \varepsilon_k^j, & \text{kui } h = a_k^j, \\ \dfrac{\varepsilon_k^j}{m_j - 1}, & \text{mujal,} \end{cases}$$

kus $\varepsilon_k^j = 1 - \gamma_k^j$, siis saame, et

$$\alpha_k^{jh} = \begin{cases} 1 - \varepsilon_k^j, & \text{kui } h = a_k^j, \\ \dfrac{\varepsilon_k^j}{m_j - 1}, & \text{mujal.} \end{cases} \tag{6}$$

Seega eelmise näite põhjal on uueks $\boldsymbol{\alpha}_k$-ks vektor väärtustega $\boldsymbol{\alpha}_k = (0{,}6; 0{,}4; 0{,}2; 0{,}2; 0{,}2; 0{,}4)$ ning $\boldsymbol{\gamma}_k = (0{,}6; 0{,}4)$. Selle kitsendustega mudeli $[\varepsilon_k^j]$ korral on parameetrite arv $(K - 1) + Kd$. Seega peatüki alguses toodud näite korral saame parameetrite arvuks 77, mis on üle kolme korra väiksem tulemus kui kitsendusteta mudeli korral.

2.3.1 Ülevaade erinevatest kitsendustega segumudelite klassidest

Eelmises peatükis toodud parametriseerimise abil saame defineerida viis erinevat segumudeli klassi. Kitsendusteta mudelit tähistatakse $[\varepsilon_k^{jh}]$. Lisaks sellele mudelile on neli kitsendustega mudelit defineeritud kujul:

- $[\varepsilon_k^{jh}]$: standardne segumudel (1).
- $[\varepsilon_k^j]$: mudel, mille parameetrid on kujul (6). Iga tunnuse korral on klastris k

vaja hinnata moodi tõenäosus ($\gamma_k^j = 1 - \varepsilon_k^j$) ja moodi positsioon a_k^j.

- $[\varepsilon_k]$: mudel, kus ε_k^j ja tunnus j on sõltumatud. Klastris k on iga tunnuse moodi tõenäosus sama ehk $\gamma_k^j = \gamma_k, \quad \forall j = 1, ..., d$.

- $[\varepsilon^j]$: mudel, kus ε_k^j ja komponent k on sõltumatud. Tunnuse j moodi tõenäosus on kõigis klastrites sama ehk $\gamma_k^j = \gamma^j, \quad \forall k = 1, ..., K$.

- $[\varepsilon]$: mudel, kus ε_k^j on sõltumatu nii tunnusest j kui ka komponendist k. Moodi tõenäosus γ_k^j ei sõltu vaadeldavast klastrist k ega tunnusest j.

Juhul, kui kõik tunnused on binaarsed, siis on mudel $[\varepsilon_k^j]$ ekvivalentne standardse mudeliga $[\varepsilon_k^{jh}]$.

2.3.2 Parameetrite hindamine mudelis $[\varepsilon_k^j]$

Toome siinkohal näite kitsendustega mudeli parameetrite hindamisest. Vaatame lähemalt mudelit $[\varepsilon_k^j]$ ning selle parameetrite hindamise protsessi. Mudeli $[\varepsilon_k^j]$ korral saab tõenäosustiheduse kirja panna kujul

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \prod_{j=1}^{d} (1 - \varepsilon_k^j) \Big(\frac{\varepsilon_k^j}{(m_j - 1)(1 - \varepsilon_k^j)} \Big)^{1 - \delta(x^j, a_k^j)},$$

kus δ-ga on märgitud Kroneckeri deltafunktsioon. Logaritmilise tõepärafunktsiooni kuju on seega

$$\ell(\boldsymbol{x}; \boldsymbol{\theta}) = \ln \left(\prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}) \right) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} p_k \prod_{j=1}^{d} (1 - \varepsilon_k^j) \Big(\frac{\varepsilon_k^j}{(m_j - 1)(1 - \varepsilon_k^j)} \Big)^{1 - \delta(x_i^j, a_k^j)} \right)$$

ning vaatluste $\{\boldsymbol{x}, \boldsymbol{z}\}$ logaritmiline tõepärafunktsioon avaldub kujul

$$\begin{aligned} \ell(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) &= \ln \left(\prod_{i=1}^{n} \prod_{k=1}^{K} \left(p_k \prod_{j=1}^{d} (1 - \varepsilon_k^j) \Big(\frac{\varepsilon_k^j}{(m_j - 1)(1 - \varepsilon_k^j)} \Big)^{1 - \delta(x_i^j, a_k^j)} \right)^{z_{ik}} \right) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln \left(p_k \prod_{j=1}^{d} (1 - \varepsilon_k^j) \Big(\frac{\varepsilon_k^j}{(m_j - 1)(1 - \varepsilon_k^j)} \Big)^{1 - \delta(x_i^j, a_k^j)} \right). \end{aligned}$$

Funktsiooni $\ell(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta})$ tinglik keskväärtus $E[\ell(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta}) \mid \boldsymbol{x}; \boldsymbol{\theta}^{(r)}]$ on seega kujul

$$\begin{aligned}
Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}^{(r)} \ln\left(p_k \prod_{j=1}^{d} (1-\varepsilon_k^j) \Big(\frac{\varepsilon_k^j}{(m_j-1)(1-\varepsilon_k^j)} \Big)^{1-\delta(x_i^j, a_k^j)} \right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}^{(r)} \left(\ln p_k + \sum_{j=1}^{d} \left(\ln(1-\varepsilon_k^j) + (1-\delta(x_i^j, a_k^j)) \ln\Big(\frac{\varepsilon_k^j}{(m_j-1)(1-\varepsilon_k^j)} \Big) \right) \right) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}^{(r)} \ln p_k + \sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}^{(r)} \sum_{j=1}^{d} \ln(1-\varepsilon_k^j) \\
&\quad + \sum_{k=1}^{K}\sum_{j=1}^{d} \Big(\sum_{i=1}^{n} t_{ik}^{(r)} - \sum_{i=1}^{n} t_{ik}^{(r)} \delta(x_i^j, a_k^j) \Big) \ln\Big(\frac{\varepsilon_k^j}{(m_j-1)(1-\varepsilon_k^j)} \Big),
\end{aligned}$$

sest $E[Z_{ik} \mid \boldsymbol{x}; \boldsymbol{\theta}^{(r)}] = t_{ik}^{(r)}$. Asendades nüüd $n_k^{(r)} = \sum_{i=1}^{n} t_{ik}^{(r)}$, saame, et

$$\begin{aligned}
Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) &= \sum_{k=1}^{K} n_k^{(r)} \ln p_k + \sum_{k=1}^{K} n_k^{(r)} \sum_{j=1}^{d} \ln(1-\varepsilon_k^j) \\
&\quad + \sum_{k=1}^{K}\sum_{j=1}^{d} \ln\Big(\frac{\varepsilon_k^j}{(m_j-1)(1-\varepsilon_k^j)} \Big) \Big(n_k^{(r)} - \sum_{i=1}^{n} t_{ik}^{(r)} \delta(x_i^j, a_k^j) \Big).
\end{aligned} \tag{7}$$

Sellega oleme ära näidanud EM-algoritmi E-sammu. Nagu juba eelnevalt on tuttav, siis M-sammuga leitakse parameetrid, mis maksimiseerivad tingliku keskväärtuse (7). Parameetri p_k uuendamine käib nii nagu kitsendusteta mudeli korral sai kirjeldatud. Parameetri ε_k^j hinnangu saamiseks aga leiame esmalt $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$ tuletise ε_k^j järgi ning seejärel võrdsustame saadud tuletise nulliga ja lahendame saadud võrrandi

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})}{\partial \varepsilon_k^j} &= -n_k^{(r)} \frac{1}{1-\varepsilon_k^j} + \Big(\frac{1}{\varepsilon_k^j} + \frac{1}{1-\varepsilon_k^j} \Big) \Big(n_k^{(r)} - \sum_{i=1}^{n} t_{ik}^{(r)} \delta(x_i^j, a_k^j) \Big) \\
&= -\frac{n_k^{(r)}}{1-\varepsilon_k^j} + \frac{1}{\varepsilon_k^j(1-\varepsilon_k^j)} \Big(n_k^{(r)} - \sum_{i=1}^{n} t_{ik}^{(r)} \delta(x_i^j, a_k^j) \Big),
\end{aligned}$$

$$0 = -\frac{n_k^{(r)}}{1-\varepsilon_k^j} + \frac{1}{\varepsilon_k^j(1-\varepsilon_k^j)}\Big(n_k^{(r)} - \sum_{i=1}^{n} t_{ik}^{(r)}\delta(x_i^j, a_k^j)\Big) \quad \Big| \cdot \varepsilon_k^j(1-\varepsilon_k^j)$$

$$0 = -n_k^{(r)}\varepsilon_k^j + n_k^{(r)} - \sum_{i=1}^{n} t_{ik}^{(r)}\delta(x_i^j, a_k^j),$$

$$n_k^{(r)}\varepsilon_k^j = n_k^{(r)} - \sum_{i=1}^{n} t_{ik}^{(r)}\delta(x_i^j, a_k^j),$$

$$\varepsilon_k^j = \frac{n_k^{(r)} - \sum_{i=1}^{n} t_{ik}^{(r)}\delta(x_i^j, a_k^j)}{n_k^{(r)}}.$$

Kuna a_k^j hinnang on $(a_k^j)^{(r+1)} = \text{argmax}_h(\sum_{i=1}^{n} t_{ik}^{(r)} x_i^{jh})$, siis saame ε_k^j hinnangu uuendamiseks reegli

$$(\varepsilon_k^j)^{(r+1)} = \frac{n_k^{(r)} - \sum_{i=1}^{n} t_{ik}^{(r)}\delta(x_i^j, (a_k^j)^{(r+1)})}{n_k^{(r)}}.$$

2.4 Integreeritud klassifitseerimistõepära kriteerium

Hinnates erinevate kitsenduste ja komponentide arvuga segumudeleid, kerkib üles küsimus, kuidas nende hulgast leida parim. Kuna segujaotuse parameetrite hinnangud leitakse tõepära $p(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i;\boldsymbol{\theta})$ maksimiseerimise abil, siis saab sobiva segumudeli valimiseks kasutada erinevaid informatsioonikriteeriume. Selleks vaatame lähemalt Bayesi informatsioonikriteeriumi (BIC) ja integreeritud klassifitseerimistõepära kriteeriumi (ICL).

Valides mudeli maksimaalse tõepäraga on samaväärne, kui valime mudeli mille integreeritud tõepära on suurim. Siinkohal BIC pakub usaldusväärset lähendit integreeritud tõepärale. Samal ajal aga BIC kriteerium kipub mudeli õiget komponentide arvu üle hindama võtmata arvesse, kas klastrid on hästi eraldatud (Biernacki jt (2000)). Selle tulemusel BIC kaldub eelistama tegelikust suurema komponentide arvuga mudeleid. See puudujääk aga ei esine ICL-kriteeriumi puhul. Selle tõttu on käesoleva töö praktilises osas parima mudeli leidmiseks just kasutatud ICL-kriteeriumi. Kuna viimast on võimalik aga defineerida kui Bayesi kriteeriumi, millele on liidetud karistusliige, siis anname siin ülevaate ka BIC-st.

Bayesi informatsioonikriteeriumi on defineeritud järgmiselt:

$$\mathrm{BIC} = -2\ln p(\boldsymbol{x};\hat{\boldsymbol{\theta}}) + \nu \ln n,$$

kus $\hat{\boldsymbol{\theta}}$ on suurima tõepära hinnang mudeli parameetritele ja ν on mudeli hinnatavate parameetrite arv (Lebret jt (2015)).

Järgnevas ICL-kriteeriumi ülevaates on toetutud artiklile Biernacki jt (2000), kui pole märgitud teisiti. Vaatame täielikku andmestikku $\{\boldsymbol{x},\boldsymbol{z}\} = \{(\boldsymbol{x_1},\boldsymbol{z_1}),...,(\boldsymbol{x_n},\boldsymbol{z_n})\}$, mille tõepärafunktsioon on kujul

$$p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{k=1}^{K}\Big(p_k \prod_{j=1}^{d}\prod_{h=1}^{m_j}(\alpha_k^{jh})^{x_i^{jh}}\Big)^{z_{ik}}.$$

Asendades tundmatud $\boldsymbol{z}$ nende hinnangutega $\hat{\boldsymbol{z}}$, kus

$$\hat{z}_{ik} = \begin{cases} 1, & \text{kui } \operatorname*{argmax}_l t_{il}(\hat{\boldsymbol{\theta}}) = k, \\ 0, & \text{mujal,} \end{cases}$$

saame ICL-kriteeriumi defineerida kujul

$$\mathrm{ICL} = -2\ln p(\boldsymbol{x},\hat{\boldsymbol{z}};\hat{\boldsymbol{\theta}}) + \nu \ln n.$$

ICL-kriteeriumi on võimalik avaldada ka BIC kaudu. Järgneva võrduseni on jõutud artiklis Biernacki jt (2000):

$$\mathrm{ICL} = \mathrm{BIC} - 2\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}\ln\hat{t}_{ik}. \tag{8}$$

Valemis (8) on BIC-le liidetud veel kahekordne karistusliige kujul $-\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik} \cdot \ln\hat{t}_{ik} \geqslant 0$, mis iseloomustab hinnatud segumudeli võimet vaatluseid $\boldsymbol{x}_1,...,\boldsymbol{x}_n$ klasterdada. Kuna $E[Z_{ik} \mid \boldsymbol{x}] = P(Z_{ik} = 1 \mid \boldsymbol{x}) = t_{ik}$, siis kasutatakse karistusliikmena ka suurust $-\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{t}_{ik}\ln\hat{t}_{ik} \geqslant 0$. Kui hinnatud segumudeli komponendid $1,...,K$ on üksteisest hästi eraldatud, siis määravad tinglikud tõenäosused $\hat{t}_{ik}$ vaatluste $\boldsymbol{x}_1,...,\boldsymbol{x}_n$ selge klasterduse ning karistusliige on sellisel juhul ligikaudu null.

Kui aga segumudeli komponendid kattuvad suurel määral, siis on karistusliikme väärtus suur ning seega on ka ICL väärtus suurem. Klastrite eristamine on kõige raskem, kui iga vaatluse korral $\hat{t}_{ik} = 1/K$. Sellisel juhul on karistusliikme väärtus $n \ln K$, mis on suurim võimalik väärtus. Seega pooldab ICL-kriteerium segumudeleid, mille komponentide poolt määratud klastrite kattuvus on võimalikult väike. Parim mudel on see, millel on väikseim informatsioonikriteeriumi väärtus.

Kvalitatiivsete tunnuste korral on kasutusel täpne ICL-kriteerium, mille saab välja arvutada järgmise valemiga (Biernacki jt (2010))

$$\begin{aligned}\mathrm{ICL}_{\mathcal{M}} = \ln p(\boldsymbol{x}, \hat{\boldsymbol{z}}) &= \sum_{k=1}^{K} \ln \Gamma\Big(\hat{n}_k + \frac{1}{2}\Big) \\ &+ \sum_{k=1}^{K}\sum_{j=1}^{d}\Bigg\{\sum_{h=1}^{m_j} \ln \Gamma\Big(\hat{n}_k^{jh} + \frac{1}{2}\Big) - \ln \Gamma\Big(\hat{n}_k + \frac{m_j}{2}\Big)\Bigg\} \\ &+ \ln \Gamma\Big(\frac{K}{2}\Big) - K \ln \Gamma\Big(\frac{1}{2}\Big) - \ln \Gamma\Big(n + \frac{K}{2}\Big) \\ &+ K \sum_{j=1}^{d}\Big\{\ln \Gamma\Big(\frac{m_j}{2}\Big) - m_j \ln \Gamma\Big(\frac{1}{2}\Big)\Big\},\end{aligned}$$

kus Γ tähistab gammafunktsiooni, $\hat{n}_k = \sum_{i=1}^{n} \hat{z}_{ik}$ ning $\hat{n}_k^{jh} = \sum_{i=1}^{n} \hat{z}_{ik} x_i^{jh}$. Täpse ICL-kriteeriumi korral on parimaks mudeliks see, mille korral $\mathrm{ICL}_{\mathcal{M}}$ väärtus on suurim.

Lisapaketis „Rmixmod“, millest toome ülevaate järgmises peatükis, on implementeeritud ainult asümptootiline ICL-kriteerium (8). Kuna kõigi hinnatud segumudelite klasterdused salvestatakse, siis on võimalik iga mudeli korral ise välja arvutada $\mathrm{ICL}_{\mathcal{M}}$ väärtus. Käesolevas töös vaatame parima mudeli leidmiseks eelkõige asümptootilist ICL-kriteeriumi.

2.5 Tarkvara R lisapakett „Rmixmod“

Praktilise osa läbiviimisel on kasutatud statistikatarkvara R lisapaketti „Rmixmod“, mille abil on võimalik teostada mudelipõhist klasteranalüüsi. Selles peatükis vaatame lähemalt „Rmixmod“ funktsioone. Ülevaade põhineb paketi dokumentat-

sioonil Langrognet jt (2019).

Mudelipõhist klasteranalüüsi on võimalik läbi viia funktsiooniga `mixmodCluster()`, mis leiab optimaalse komponentide arvuga segumudeli võttes arvesse etteantud kriteeriumeid. Olulisemad argumendid, mida saame funktsioonile ette anda on `data`, `nbCluster`, `dataType`, `models`, `strategy` ja `criterion`. Neist parameetritest kohustuslikud on `data` ja `nbCluster`.

Andmestikuks (`data`) on klasterdavaid vaatluseid sisaldav andmetabel. Parameetrile `nbCluster` antakse ette hinnatavate klastrite arv, mis ühtlasi vastab segumudeli komponentide arvule). Argumendi `dataType` abil määratakse andmete tüüp ehk kas vaatluse all olev andmestik sisaldab ainult kvantitatiivseid (`quantitative`), kvalitatiivseid (`qualitative`) või segatüüpi (`composite`) tunnuseid. Vaikimisi antakse parameetrile väärtus `NULL` ning programm proovib andmestiku põhjal ise ära arvata, mis tüüpi tunnustega on tegemist.

Argumendiga `models` on võimalik määrata mudelite loetelu, mida funktsioon hindab. Kvalitatiivsete tunnuste korral antakse sellele argumendile väärtus funktsiooni `mixmodMultinomialModel()` abil, kus argumendiga `listModels` on võimalik ette anda loetelu mudelitest, mida soovitakse hinnata. Vaikimisi hinnatakse kõik peatükis 2.3.1 kirjeldatud mudelid $[\varepsilon_k^{jh}]$, $[\varepsilon_k^j]$, $[\varepsilon_k]$, $[\varepsilon^j]$ ja $[\varepsilon]$. Lisaks on võimalik võtta komponentide kaalud võrdseteks muutes parameetri `equal.proportions` väärtus `TRUE`-ks. Seega on iga fikseeritud komponentide arvu k korral võimalik hinnata kümmet erinevat mudelit.

Argumendi `strategy` abil saab määrata, millist strateegiat kasutatakse segujaotuse parameetrite hindamiseks. Selle jaoks kutsutakse välja funktsioon `mixmodStrategy()`, mille olulisemateks argumentideks on `algo`, `nbTry` ja `initMethod`. Argumendile `algo` saab ette anda algoritmi, mida kasutatakse parameetrite hindamiseks. Vaikimisi on selleks EM-algoritm, mida on kasutatud ka käesolevas töös. Parameetriga `nbTry` saab määrata mitu korda rakendatakse valitud algoritmi. Selle vaikeväärtuseks on 1. Maksimaalne iteratsioonide arv antakse ette argumendile `nbIterationInAlgo` (vaikimisi 200). Valitud algoritmile leitakse alglähendid argumendi `initMethod` väärtuseks valitud meetodi abil. Vaikimisi määratakse parameetrite alglähendid niinimetatud lühikese EM-algoritmi (`smallEM`) abil. Sel-

le meetodi korral valitakse alglähendid juhuslikult ning seejärel teostatakse väike arv EM-algoritmi iteratsioone. Iteratsioonide arv on määratud argumendiga `nbIterationInInit`, mis saab vaikimisi ette arvu 5. Lühikest EM-algoritmi korratakse seni kuni teostatud iteratsioonide arv on argumendi `nbTryInInit` väärtusest väiksem, milleks vaikimisi on 50. Lõplikeks alglähenditeks valitakse need parameetrite hinnangud, mille korral logaritmilise tõepärafunktsiooni väärtus on suurim.

Parim mudel valitakse välja kriteeriumi abil, mis on määratud argumendi `criterion` all. Selle võimalikud väärtused on `BIC` (Bayesi informatsioonikriteerium), `ICL` (integreeritud klassifitseerimistõepära kriteerium) ja `NEC` (niinimetatud entroopia kriteerium). Käesolevas töös on kasutusele võetud ICL kriteerium, mis on ära kirjeldatud peatükis 2.4. Parim mudel on see, millel on väikseim ICL väärtus.

Funktsiooni `mixmodCluster()` rakendamise tulemusel saadud väljund `results` sisaldab kõigi hinnatud mudelite tulemuste loendit, kus mudelid on sorteeritud kasvavalt vastavalt informatsioonikriteeriumi (`criterion`) väärtuse järgi. Lisaks kriteeriumi väärtusele väljastatakse veel näiteks vaatluste klasterdus, parameetrite hinnangud ja maksimeeritud logaritmiline tõepära. Lisaks on võimalik väljastada ka ainult parima mudeli abil saadud tulemus, mis on salvestatud objekti `bestResult` alla. Tulemuste reprodutseerimiseks on funktsioonile `mixmodCluster()` võimalik argumendiga `seed` ette anda kasutatav seemne väärtus.

2.6 Näited

Enne veel, kui asume haigekassa andmete peal klasteranalüüsi tegema, anname paar näidet funktsiooni `mixmodCluster()` kasutamisest. Esmalt toome näite simuleeritud andmete peal ning seejärel vaatame lähemalt R paketi andmestikku `birds`.

2.6.1 Näide simuleeritud andmetega

Kvalitatiivsete tunnustega andmete korral määravad mudelipõhisel klasterdamisel klastrite kattuvuse tunnuste jaotused. Kui tunnuse $j = 1, ..., d$ iga väärtus $h = 1, .., m_j$ esineb suure tõenäosusega ühes grupis, siis on klastrid üksteisest selgesti

eraldatud ning seega on nende kattuvus väike. Kui aga mõned väärtused esinevad mitmes grupis sarnase tõenäosusega, siis on klastrite kattuvus suurem. Selleks, et genereerida andmeid, kus saame ise reguleerida klastrite kattuvust, kasutame järgmist parametrisatsiooni (Biernacki jt (2010)):

$$\alpha_k^{jh} = \begin{cases} \dfrac{1}{m_j} + (1-\delta)\dfrac{m_j - 1}{m_j}, & \text{kui } h = [(k-1) \mod m_j] + 1 \\ \dfrac{\left(1 - \frac{1}{m_j} - (1-\delta)\frac{m_j - 1}{m_j}\right)}{m_j - 1} & \text{mujal.} \end{cases} \tag{9}$$

Konstandiga $\delta \in [0,1]$ saab määrata klastrite kattuvuse ning mod tähistab jääga jagamist. Kui $\delta = 0$, siis on klastrite kattuvus minimaalne, sest sellisel juhul $\alpha_k^{jh} = 0$ või 1. Samal ajal aga konstandi väärtus $\delta = 1$ määrab klastrite maksimaalset kattuvust, sest $\alpha_k^{jh} = 1/m_j$.

Genereerime nüüd $n = 3\ 000$ vaatlust $K = 2$ komponendiga segujaotusest (1). Olgu meil vaatluse all $d = 4$ kvalitatiivset tunnust, millel on vastavalt $m_1 = 2$, $m_2 = m_3 = 3$, $m_4 = 4$ võimalikku väärtust. Komponentide kaalud olgu vastavalt $p_1 = 0{,}2$ ja $p_2 = 0{,}8$. Teostame klasteranalüüsi, kus klastrite arvuks määrame $K = 2, ..., 4$. Parameetrite hinnangud leiame EM-algoritmi abil, mida rakendame 20 korda erinevate alglähenditega ja iteratsioonide arvuks võtame 1 000. Rakendame kõiki peatükis 2.3.1 kirjeldatud mudeleid ($[\varepsilon_k^{jh}]$, $[\varepsilon_k^j]$, $[\varepsilon_k]$, $[\varepsilon^j]$ ja $[\varepsilon]$). Kasutades kattuvusmäära $\delta = 0{,}6$ saame valemiga (9) segujaotuse kahe komponendi parameetrite vektoriteks

$$\boldsymbol{\alpha}_1 = \Big(0{,}7; 0{,}3; \quad 0{,}6; 0{,}2; 0{,}2; \quad 0{,}6; 0{,}2; 0{,}2; \quad 0{,}55; 0{,}15; 0{,}15; 0{,}15\Big),$$

$$\boldsymbol{\alpha}_2 = \Big(0{,}3; 0{,}7; \quad 0{,}2; 0{,}6; 0{,}2; \quad 0{,}2; 0{,}6; 0{,}2; \quad 0{,}15; 0{,}55; 0{,}15; 0{,}15\Big).$$

Asümptootilise ICL kriteeriumi põhjal on kolm parimat segumudelit $[\varepsilon^j]$,$[\varepsilon]$ ja $[\varepsilon_k^j]$, mis kõik on kahekomponendilised. ICL väärtused on vastavalt 25 285,63, 25 383,55 ja 25 439,77. Valesti on klasterdatud 356 vaatlust, mis annab meile veamäära 11,87%. Moodide tõenäosused on $\hat{\boldsymbol{\gamma}}_k \approx (0{,}69; 0{,}6; 0{,}61; 0{,}55)$, $k = 1, 2$, mis on mõlemas klastris samad, sest meil on vaatluse all kitsendustega mudel $[\varepsilon^j]$. Komponentide hinnatud kaalud on vastavalt $\hat{p}_1 = 0{,}1968$ ja $\hat{p}_2 = 0{,}8032$. Parameetrite

$\boldsymbol{\alpha}_1$ ja $\boldsymbol{\alpha}_2$ hinnangud on vastavalt

$$\hat{\boldsymbol{\alpha}}_1 \approx (0{,}69; 0{,}31; \quad 0{,}6; 0{,}2; 0{,}2; \quad 0{,}61; 0{,}195; 0{,}195; \quad 0{,}55; 0{,}15; 0{,}15; 0{,}15),$$

$$\hat{\boldsymbol{\alpha}}_2 \approx (0{,}31; 0{,}69; \quad 0{,}2; 0{,}6; 0{,}2; \quad 0{,}195; 0{,}61; 0{,}195; \quad 0{,}15; 0{,}55; 0{,}15; 0{,}15).$$

Näeme, et tõenäosuste hinnangud on tegelikele väärtustele väga lähedal.

Täpse ICL kriteeriumi kohaselt on parimateks kahekomponendilised mudelid $[\varepsilon^j]$, $[\varepsilon]$ ja $[\varepsilon_k^{jh}]$. Vatavad ICL kriteeriumi väärtused on $-11\ 933{,}63$, $-11\ 933{,}63$ ja $-12\ 158{,}41$. Kuna esimesed kaks $\mathrm{ICL}_{\mathcal{M}}$ väärtust on samad, siis parem mudel on $[\varepsilon]$, sest parameetrite arv selles mudelis on väiksem. Mudeli veamäär on 19,5% (vaatlustest on 584 klasterdatud valesti), mis on veidi suurem kui mudeli korral, mis osutus parimaks asümptootilise ICL põhjal. Seega täpne ICL antud näite korral meile paremat mudelit ei vali. Moodide tõenäosused on $\hat{\gamma}_k^j = 0{,}6033$ iga $j = 1, ..., 4$ ja $k = 1, 2$ korral. Parameetrite hinnangud saame vastavalt

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_1 = (&0{,}6033; 0{,}3967; \quad 0{,}6033; 0{,}1984; 0{,}1984; \\ &0{,}6033; 0{,}1984; 0{,}1984; \quad 0{,}6033; 0{,}1322; 0{,}1322; 0{,}1322).\end{aligned}$$

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_2 = (&0{,}3967; 0{,}6033; \quad 0{,}1984; 0{,}6033; 0{,}1984; \\ &0{,}1984; 0{,}6033; 0{,}1984; \quad 0{,}1322; 0{,}6033; 0{,}1322; 0{,}1322).\end{aligned}$$

2.6.2 Näide andmestikuga `birds`

Andmestik `birds` kirjeldab lindude (täpsemalt on tegemist linnuliigiga lunn) välimust nelja kvalitatiivse tunnuse abil - kulmude kuju (4 väärtust), krae välimus (5 väärtust), sabaaluse värvus (5 väärtust), rantide esinemine (3 väärtust). Iga linnu kohta on välja toodud ka tema sugu. Kokku on andmestikus 69 lunni.

Klasteranalüüsi läbiviimiseks on eemaldatud tunnuste väärtused, millele ei vastanud ühtegi lindu. Nendeks olid krae välimus kriipsudega, pikemate kriipsudega või ühtlase värvusega ning sabaaluse must-valge värv, kus must domineeris. Lisaks sai üheks tunnuse väärtuseks kokku tõstetud sabaaluse must-valge värvuse erinevad variatsioonid ning tunnuse rantide esinemine sai muudetud kaheväärtuseliseks vastavalt sellele, kas esineb rante või mitte. Viies läbi selle andmestiku peal klaste-

ranalüüsi, osutub parimaks mudeliks asümptootilise ICL (ICL = 453,1357) põhjal $[\varepsilon_k^j]$, mis on kahekomponendiline. ICL väärtus on 486,8 ning selle mudeliga saadud klastrite suurused on $\hat{n}_1 = 45$ ja $\hat{n}_2 = 24$. Tabelis 1 on antud ülevaade tunnuste sagedustest saadud kahes klastris.

Tabel 1. Andmestiku `birds` tunnuste sagedused klastrites

Klaster / Tunnuste väärtused	1	2
Mees	23	10
Naine	22	14
Kulmud (puuduvad)	4	2
Kulmud (halvasti näha)	0	21
Kulmud (ilmekad)	38	0
Kulmud (tugevasti esiletungivad)	3	1
Krae (puudub)	38	2
Krae (täpiline)	7	22
Sabaalune (valge)	45	8
Sabaalune (must)	0	11
Sabaalune (must-valge)	0	5
Randid (puuduvad)	43	22
Randid (esinevad)	2	2

Mõlemas klastris on isaslinde ja emaslinde üsna võrdselt. Esimese klastri moodustavad lunnid, kellel on kulmud ilmekalt välja joonistunud, kellel puudub iseäralik krae, kelle sabaalune on valge ning rante sulestikus pole. Teises klastris on linnud, kelle kulmud on halvasti näha ning kellel on täpiline krae. Sabaaluse värvus nendel lunnidel varieerub, kuid must värv domineerib. Sellest klasterdusest ei tule välja kindlat erinevust isaslinnu ja emaslinnu välimuses. ICL täpse põhjal osutus parimaks mudeliks aga kahekomponendiline standartne segumudel $[\varepsilon_k^{jh}]$, mis andis praktiliselt sama klasterduse nagu eelnevalt juba kirjeldasime.

3 Haigekassa andmete kirjeldav analüüs

Kirjeldava analüüsi idee seisneb seaduspärasuste otsimises andmetes ilma, et eelnevalt oleks püstitatud hüpoteese või eeldusi. Kirjeldav analüüs on vajalik ülevaate saamiseks andmetest, mis võimaldab meil uurida, kas teisenduste tegemine on vajalik või mitte. Lisaks aitab analüüsi läbitegemine edasisi uuringuid paremini kavandada.

Puuduvate väärtuste kontrollimisel avastati andmestikust 105 rida, kus vanus pole ära märgitud. Need kirjed eemaldati, mille tulemusel uue andmestiku suurus on 10 675 696, kus on 1 479 962 patsienti, kellest 694 517 on mehed ja 785 445 on naised.

Kõigil patsientidel ei ole andmestikus sama arv kirjeid. Põhjusteks võib olla näiteks, et paljud patsiendid pole igal aasta pidanud arsti poole pöörduma või polnud patsient 2008. aastal veel sündinudki või hoopis patsient suri ära enne 2018. aastat. Järgnevalt on esitatud ülevaatlik tabel, kus on näha patsientide arv vastavalt sellele, mitmel aasta patsient arsti poole pöördus või ravi vajas.

Tabel 2. Patsientide arv aastaste sissekannete arvu kohta

Aastaid	1	2	3	4	5	6	7	8	9	10	11
Patsientide arv	73 041	73 864	76 683	81 637	88 451	100 543	120 586	162 990	317 684	383 674	809

Kõige vähem on patsiente, kellel on diagnoos määratud kõigil vaadeldavatel aastatel. Suurem osa patsiente kuulub gruppi, kus ravi on vajatud 9 või 10 aastal.

Vaatame täpsemalt, kuidas patsiendid jagunevad vanusegruppidesse. Vanusegrupi määramisel on võetud aluseks iga patsiendi sissekannete keskmine vanus. Kui pooled patsiendi sissekanded sisaldavad vanuseid, mis kuuluvad madalamasse gruppi ja ülejäänud vanuseid, mis kuuluvad järgmisesse vanusegruppi, siis patsient on määratud nooremasse gruppi. Ülevaadet gruppide suurustest on võimalik näha tabelist 3.

Tabel 3. Patsientide arv vanusegruppide ja soo kaupa

Vanusegrupp	**Sugu**	**Patsientide arv**	
0-9	M	114 730	222 741
	N	108 011	
10-19	M	68 132	131 473
	N	63 341	
20-29	M	95 507	189 476
	N	93 969	
30-39	M	97 070	190 354
	N	93 284	
40-49	M	89 762	180 063
	N	90 301	
50-59	M	88 429	187 218
	N	98 789	
60-69	M	70 198	160 163
	N	89 965	
70-79	M	48 450	131 688
	N	83 238	
80-89	M	20 445	76 415
	N	55 970	
90-99	M	1 757	10 137
	N	8 380	
100+	M	37	234
	N	197	

Tabelist 3 võime näha, et kõige rohkem on patsiente vanuses 0-9. Meessoost ja naissoost patsientide arv on grupiti võrdlemisi tasakaalus, kui vaatame vanusegruppe kuni 49-aastasteni. Alates vanusest 50 hakkab naiste osakaal vanusegruppides kasvama.

Vaatame nüüd täpsemalt psüühika- ja käitumishäiretega seotud diagnoose, mis andmestikus on tähistatud F-tähega koodi alguses. Seda tüüpi probleeme on esinenud 402 524 patsiendil, kellest 166 644 on mehed ja 235 880 on naised. Seega on 27% patsientidele mingil aastal diagnoositud psüühika- või käitumishäire. Järgnevalt on toodud ülevaatlik tabel patsientide arvust, kellel on põhidiagnoosiks märgitud psüühika- ja käitumishäired.

Tabel 4. Psüühika- ja käitumishäirega patsientide arv vanusegruppide ja soo kaupa

Vanusegrupp	**Sugu**	**Patsientide arv**		**Protsent patsientide koguarvust vanusegrupis**
0-9	M	26 187	41 252	18,52%
	N	15 065		
10-19	M	19 791	37 414	28,46%
	N	17 623		
20-29	M	23 500	50 736	26,78%
	N	27 236		
30-39	M	23 012	53 368	28,04%
	N	30 356		
40-49	M	21 308	52 645	29,24%
	N	31 337		
50-59	M	21 180	56 112	29,97%
	N	34 932		
60-69	M	15 276	44 432	27,74%
	N	29 156		
70-79	M	10 652	37 864	28,75%
	N	27 212		
80-89	M	5 251	25 254	33,05%
	N	20 003		
90-99	M	479	3 381	33,35%
	N	2 902		
100+	M	8	66	28,2%
	N	58		

Suuri erinevusi vanusegruppide vahel pole näha. Näeme, et patsientide osakaal, kellel on psüühika-ja käitumishäired, jääb igas grupis 30% lähedale, välja arvatud laste seas.

Vaatame nüüd täpsemalt, milliseid diagnoose psüühika- ja käitumishäiretega patsientidele on määratud. Järgmises tabelis 5 on välja toodud patsientide arv vastavalt psüühika- ja käitumishäirete klassifikatsioonidele. Näiteks esimeses lahtris on välja toodud patsientide arv (48 946), kellel on esinenud psüühika-ja käitumishäire, mille diagnoosikood kuulub gruppi F00-F09. Esimese rea järgmises lahtris on märgitud need patsiendid (3 622), kellel lisaks diagnoosile, mis kuulub gruppi F00-F09, on ka esinenud diagnoos, mis kuulub gruppi F10-F19.

Tabel 5. Psüühika- ja käitumishäiretega patsientide arv diagnoosikoodide kaupa

	F00-F09	F10-F19	F20-F29	F30-F39	F40-F49	F50-F59	F60-F69	F70-F79	F80-F89	F90-F98	F99
F00-F09	48 946	3 622	2 494	11 640	12 275	4 591	675	1 139	463	619	117
F10-F19	3 622	34 034	1 218	7 551	9 100	3 147	792	760	245	705	48
F20-F29	2 494	1 218	16 292	3 774	3 817	954	495	692	255	242	89
F30-F39	11 640	7 551	3 774	128 266	52 955	13 841	2 359	1 110	1 151	2 340	134
F40-F49	12 275	9 100	3 817	52 955	191 416	16 637	2 561	2 050	3 335	6 110	167
F50-F59	4 591	3 147	954	13 841	16 637	53 167	516	292	450	780	50
F60-F69	675	792	495	2 359	2 561	516	5 238	164	160	241	20
F70-F79	1 139	760	692	1 110	2 050	292	164	12 987	2 651	1 334	32
F80-F89	463	245	255	1 151	3 335	450	160	2 651	30 624	8 317	32
F90-F98	619	705	242	2 340	6 110	780	241	1 334	8 317	33 086	46
F99	117	48	89	134	167	50	20	32	32	46	502

Tabelist võime näha, et kõige rohkem on patsiente (191 416), kellele on diagnoositud psüühika- ja käitumishäire, mis kuulub gruppi F40-F49 (*Neurootilised, stressiga seotud ja somatoformsed häired*). Kõige vähem patsiente on gruppides F99 (*Täpsustamata psüühikahäire*) ja F60-F69 (*Täiskasvanu isiksus- ja käitumishäired*). Kuna nii grupis F30-F39 (*Meeleoluhäired*) kui ka grupis F40-F49 esineb kõige rohkem patsiente, siis paistab kohe silma ka suur patsientide arv, kellel on esinenud mõlemasse gruppi kuuluvaid diagnoose. Täpsemalt on neid patsiente 52 955.

Andmestikus on päris palju isikuid, kellele pole määratud ainult üks psüühika- ja käitumishäirete kategooria alla kuuluv diagnoos (vaatame samu gruppe, mis tabelis 5 on märgitud). Tabel 6 annab ülevaate sellest, kui paljudele patsientidele on määratud F-diagnoose, mis kuuluvad erinevatesse psüühika- ja käitumishäirete gruppidesse.

Tabel 6. Psüühika- ja käitumishäirega patsientidele määratud erinevate F-diagnooside arv

F-diagnoosi gruppide arv	1	2	3	4	5	6	7
Patsientide arv	283 120	92 478	22 075	4 121	617	103	10

Näeme, et enamus patsientide puhul on neile siiski määratud ühte kindlasse psüühika- ja käitumishäirete gruppi kuuluv diagnoos. Ligi kolm korda vähem on patsiente, kes on saanud diagnoose, mis kuuluvad kahte erinevasse F-koodiga diagnoosigruppi. Edasi näeme, et patsientide arv, kellel esineb kolm või enam diagnoosi, kahaneb.

Vaatame nüüd vereringeelundite haiguseid, mis andmestikus on tähistatud tähega I koodi alguses. Sellesse gruppi kuulub 623 883 patsienti, kellest 260 563 on mehed

ja 363 320 on naised. Seega on 42% patsientidel mingil aastal diagnoositud vereringeelunditega seotud haigus. Järgnevalt on toodud ülevaatlik tabel patsientide arvust, kellel on põhidiagnoosiks märgitud vereringeelunditega seotud haigus.

Tabel 7. Vereringeelundite haigustega patsientide arv vanusegruppide ja soo kaupa

Vanusegrupp	**Sugu**	**Patsientide arv**		**Protsent patsientide koguarvust vanusegrupis**
0-9	M	3 144	5 706	3%
	N	2 562		
10-19	M	7 062	12 684	10%
	N	5 622		
20-29	M	18 970	37 672	20%
	N	18 702		
30-39	M	25 710	53 505	28%
	N	27 795		
40-49	M	36 806	75 359	42%
	N	38 553		
50-59	M	53 699	116 022	62%
	N	62 323		
60-69	M	53 657	125 592	67%
	N	71 935		
70-79	M	41 518	117 041	89%
	N	75 523		
80-89	M	18 373	70 867	93%
	N	52 494		
90-99	M	1 592	9 250	91%
	N	7 658		
100+	M	32	185	79%
	N	153		

Sagedustabelist 7 võime näha, et laste ja noorukite hulgas vereringeelundite haigusi eriti ei esine. Seda tüüpi haiguste hulk on suurem vanuses üle 50 eluaasta. Alates sellest vanusest on vanusegruppides vereringeelunditega seotud haigeid üle 50 protsendi. Kõige suurem on see protsent vanusegrupis 80-89, kus koguni 93 protsendil patsientidest on määratud diagnoos, mis kuulub vereringeelundite haiguste gruppi.

4 Haigekassa andmete klasteranalüüs

Esmalt uurime psüühika- ja käitumishäiretega patsiente. Kirjeldavas analüüsis välja toodud tabeli 4 põhjal nägime, et seda tüüpi diagnoose esineb vanusegrupiti üsna võrdselt. Eesmärk on välja selgitada, milliseid psüühikahäireid võime näha erinevates vanusegruppides rohkem. Lisaks oleks huvitav vaadata meeste ja naiste osakaalu nendes diagnoosigruppides.

Teises klasterdamisülesandes keskendume vereringeelundite haigustega patsientidele. Kui uurida tervisestatistika ja terviseuuringute andmebaasist surmapõhjuste tabelit (Surma põhjuste register (2011-2019)), võime näha, et peamine surma põhjus aastate 2010-2018 jooksul on olnud haigestumine just vereringeelundite haigustesse. Igal aastal on mainitud surmapõhjus märgitud umbes 8 000 inimesel. Võrdluseks järgmine kõige levinum surmapõhjus, milleks on kasvaja, on esinenud 4 000 juhul, mis on poole vähem. Eelnevas peatükis välja toodud tabelist 7 võime samuti näha, et väga suur osa vanemaealistest patsientidest põevad just vereringeelundite haigusi. Eesmärk on välja selgitada, milliseid seda tüüpi haigusi esineb vanusegrupiti kõige rohkem. Sellega seoses uurime domineerivate diagnoosigruppide muutumist ajas. Lisaks oleks vaja teada, kas ja millised vereringeelundite haigused on levinud rohkem meestel ja millised naistel.

4.1 Segumudeli parameetrite hindamine kasutades alamandmestikke

Esmalt viime läbi klasteranalüüsi ühe konkreetse alamandmestiku peal. Soov on saada ülevaade klasterdamise stabiilsusest ning võimalusest kasutada mudeli parameetrite hindamiseks väiksema mahuga alamandmestikku. Vaatluse alla on võetud 10-19-aastased patsiendid, kellel kõigil on diagnoositud psüühika- ja käitumishäirete kategooria alla kuuluv haigus. Kokku on selliseid patsiente 37 414. Parameetrite hindamine on läbi viidud kuue erineva suurusega andmestiku peal. Esimesel kahel korral on antud andmestikust juhuslikult välja võetud 5 000 patsienti. Järgmisel kahel korral 10 000, edasi 20 000 ning viimaseks on vaadatud tervet andmestikku. Mudeli parameetrite hindamist on iga andmestikuga läbi tehtud kaks korda.

Kuna antud juhul on kõik tunnused binaarsed, siis hindame kitsendusteta mude-

li. Andes klasteranalüüsi läbiviimisel funktsioonile `mixmodCluster` ette erinevaid argumentide väärtuseid, on funktsioonile leitud sobivad lähendid võimalikult stabiilsete tulemuste saavutamiseks. Segumudeli komponentide arvuks on määratud $K = 5, ..., 11$. Parameetrite hinnangud leitakse EM-algoritmi abil, mida rakendatakse 50 korda erinevate alglähenditega ning iteratsioonide arvuks on võetud 1 000. Nii EM-algoritmi rakendamiseks kui ka parameetrite alglähendite määramiseks kasutatud lühikese EM-algoritmi korral on logaritmilise tõepärafunktsiooni suhtelise muutuse koondumiskriteeriumiks võetud $\varepsilon = 0{,}0001$.

Tabelis 8 on ülevaade tulemustest. Iga katse korral on välja toodud kolm parimat mudelit nii asümptootilise ICL kui ka täpse ICL põhjal. Lisaks on esimeses veerus ära märgitud, kui palju aega parameetrite hindamiseks erinevate alamandmestike korral kulus.

Tabel 8. Segumudeli parameetrite hindamine erinevate suurustega andmestike korral psüühika- ja käitumishäiretega patsientide näitel.

Andmestik	Katse	Järjestus	ICL			$\text{ICL}_{\mathcal{M}}$		
			K	Tõepära	ICL	K	Tõepära	$\text{ICL}_{\mathcal{M}}$
Alamandmestik 1 $n = 5\,000$ Aeg: 45 min	1	1	**6**	−17 721,00	37 352,84	**11**	−17 497,65	−73 262,82
		2	**7**	−17 636,99	37 363,67	**10**	−17 506,25	−74 581,16
		3	**5**	−17 862,51	37 481,97	**9**	−17 520,08	−74 678,09
	2	1	**6**	−17 720,98	37 349,12	**10**	−17 507,70	−73 390,53
		2	**7**	−17 636,99	37 365,05	**11**	−17 497,63	−74 352,09
		3	**5**	−17 862,51	37 482,77	**9**	−17 520,08	−74 679,34
Alamandmestik 2 $n = 5\,000$ Aeg: 42 min	1	1	**7**	−17 943,11	37 901,25	**11**	−17 813,43	−72 781,16
		2	**6**	−18 054,11	38 114,12	**10**	−17 821,62	−73 166,42
		3	**5**	−18 185,30	38 165,03	**9**	−17 868,67	−74 533,93
	2	1	**7**	−17 950,62	38 021,45	**11**	−17 813,43	−72 978,12
		2	**6**	−18 054,08	38 103,72	**10**	−17 821,88	−74 228,18
		3	**5**	−18 185,30	38 164,67	**9**	−17 868,63	−74 533,93
Alamandmestik 3 $n = 10\,000$ Aeg: 1 h 41 min	1	1	**7**	−35 585,41	74 027,08	**11**	−35 343,34	−74 514,73
		2	**6**	−35 799,79	74 993,35	**10**	−35 339,28	−79 652,34
		3	**5**	−36 030,14	75 119,69	**9**	−35 375,67	−79 929,16
	2	1	**7**	−35 587,49	74 130,10	**11**	−35 337,00	−78 486,49
		2	**6**	−35 799,78	74 995,41	**10**	−35 354,63	−78 656,51
		3	**5**	−36 030,14	75 115,53	**9**	−35 465,37	−79 131,37
Alamandmestik 4 $n = 10\,000$ Aeg: 1 h 39 min	1	1	**7**	−35 873,25	75 205,24	**11**	−35 599,54	−77 155,53
		2	**6**	−36 053,29	75 592,87	**10**	−35 635,20	−77 812,18
		3	**5**	−36 294,17	75 642,67	**9**	−35 742,66	−79 713,76
	2	1	**7**	−35 873,25	75 205,78	**11**	−35 606,46	−77 976,75
		2	**6**	−36 053,32	75 596,58	**10**	−35 634,31	−78 234,60
		3	**5**	−36 294,17	75 641,26	**9**	−35 741,41	−79 724,98
Alamandmestik 5 $n = 20\,000$ Aeg: 4 h 38 min	1	1	**7**	−72 106,86	148 765,5	**11**	−71 588,66	−167 635,3
		2	**6**	−72 477,64	150 260,7	**10**	−71 614,21	−169 748,2
		3	**5**	−72 967,41	150 969,0	**9**	−71 675,83	−173 876,3
	2	1	**7**	−72 106,86	148 765,0	**11**	−71 586,74	−165 971,6
		2	**6**	−72 477,64	150 260,6	**10**	−71 614,20	−169 748,2
		3	**5**	−72 967,43	150 982,3	**9**	−71 826,57	−173 114,9
Alamandmestik 6 $n = 20\,000$ Aeg: 4 h 33 min	1	1	**7**	−71 748,14	149 128,4	**11**	−71 228,15	−170 527,2
		2	**6**	−72 141,20	149 565,0	**10**	−71 254,52	−171 232,3
		3	**5**	−72 653,68	150 396,7	**9**	−71 295,40	−175 264,7
	2	1	**6**	−72 141,24	149 583,6	**11**	−71 228,89	−170 553,0
		2	**5**	−72 653,68	150 399,8	**10**	−71 254,54	−171 215,1
		3	**9**	−71 295,40	152 096,5	**9**	−71 295,40	−175 264,7
Terve andmestik $n = 37\,414$ Aeg: 14 h 56 min	1	1	**7**	−134 584,6	276 690,1	**11**	−133 662,6	−337 176,0
		2	**6**	−135 330,4	279 887,9	**10**	−133 659,9	−350 751,5
		3	**5**	−136 277,9	281 187,4	**9**	−133 802,4	−351 962,9
	2	1	**7**	−134 584,6	276 690,2	**11**	−133 662,2	−335 848,3
		2	**6**	−135 330,4	279 879,1	**10**	−133 706,7	−341 409,0
		3	**5**	−136 277,9	281 187,3	**9**	−133 800,6	−351 036,6

Esmalt paneme tähele, et parameetrite hinnangute leidmine alamandmestiku kor-

ral hoiaks väga palju aega kokku. Nimelt 5 000 vaatlusega alamandmestiku korral võttis algoritmi töö aega rohkem kui 14 tundi vähem kui terve andmestiku korral ning 10 000 ja 20 000 patsiendiga andmestike korral vastavalt 13 ja 10 tundi vähem. Lisaks võime tabelist 8 näha, et segumudeli hindamisel osutus enamus juhtudel täpse ICL põhjal parimaks 11-komponendiline mudel ning asümptootilise ICL põhjal 7-komponendiline mudel. Ainult esimene alamandmestik andis mõlemal katsel parimaks mudeliks 6-komponendilise segumudeli, mis võib tähendada, et andmestiku maht 5 000 on liiga väike, et anda parameetrite hindamisel stabiilselt samu tulemusi nagu terve andmestiku korral. Üldiselt on aga mudeli komponentide arvu määramine olnud üsna püsiv ning ei sõltu sellest, kui suure andmestiku oleme parameetrite hindamiseks valinud. Siiski tekib küsimus, millist kriteeriumi peaksime arvesse võtma parima mudeli valimisel? Tulemustest selgus, et mudel, mis osutus parimaks täpse ICL kohaselt, oli igal korral märgitud kõige kehvemaks asümptootilise ICL põhjal. Uurides tunnuste väärtuste sagedusi klastrites olid tulemused 11 klastri korral raskemini interpreteeritavad. Sellest lähtudes määrame edaspidises analüüsis parima mudeli asümptootilise ICL põhjal.

Vaatame nüüd, kui erinevalt on patsiendid klastritesse määratud, kui on kasutatud erinevaid alamandmestikke parameetrite hinnangute leidmiseks võrreldes parameetrite hinnangute leidmisega kasutades kogu andmestikku. Alamandmestikega 2-6 ja terve andmestikuga saadud parameetrite hinnangud on välja toodud lisas 4.

Täpsemalt on võetud üle katsete vaatluse alla iga andmestiku korral parima mudeli parameetrite hinnangud. Parim mudel on valitud asümptootilise ICL põhjal. Rohelisega on märgitud need lahtrid, kus parameetri hinnang diagnoosi esinemisele on suurim. Tabelist 40 võime näha, et parameetrite hinnangud erinevate andmestike korral on sarnased. Mõnes kohas on näha, et alamandmestiku 6 korral erinevad need teiste andmestikega võrreldes veidi rohkem. Näiteks vaadates teist klastrit, kus parameetrite hinnangud tunnuse väärtuse F70-F79 jaoks on kõik 1 lähedal, kuid alamandmestiku 6 korral on hinnangu väärtuseks 0,687. Paneme tähele, et parameetrite hinnangud teiste tunnuste väärtuste jaoks erinevad samuti alamandmestiku 6 korral selles klastris. Sarnaselt on näha erinevust ka neljanda klastri korral, kui vaatame parameetrite hinnanguid tunnuse väärtuse F10-F19 jaoks. Sel-

les klastris aga parameetrite hinnangud alamandmestiku 6 korral teiste tunnuste väärtuste jaoks nii palju ei erine.

Järgnevalt analüüsime sagedustabeleid kahe klasteranalüüsi tulemuste lõikes. Esmalt võrdleme parimaid klasterdusi, mis on saadud alamandmestiku 1 korral hinnatud segumudeliga ja terve andmestiku korral hinnatud segumudeliga.

Tabel 9. Klastrite sagedused, kui segumudeli parameetrid on hinnatud alamandmestikuga 1 ja terve andmestikuga.

Terve andmestik / Alamandmestik 1	1	2	3	4	5	6	7	Kokku
1	0	7 642	0	325	0	0	0	7 967
2	0	10	0	7	1 482	0	0	1 499
3	11 291	0	2	149	0	0	0	11 442
4	0	1 553	4	1 183	24	3 817	100	6 681
5	0	0	1 197	1 001	0	0	1	2 199
6	0	19	8	30	0	0	7 569	7 626
Kokku	11 291	9 224	1 211	2 695	1 506	3 817	7 670	37 414

Alamandmestik 1 andis meile parimaks klastrite arvuks 6, mis erineb klastrite arvust, mis osutus parimaks terve andmestiku korral. Seega juba selle tõttu klasteranalüüsi tulemused erinevad ning seda näeme ka tabelist 9. Eriti suur erinevus lööb välja, kui vaatame terve andmestiku korral klastrit 4, kus sellesse klastrisse kuuluvad patsiendid on alamandmestiku parameetrite korral jagunenud mitme erineva klastri vahel. Samuti klaster 2 korral on suur osa vaatlustest paigutatud alamandmestiku parameetrite korral klastri 1 asemel klastrisse 4.

Tabel 10. Klastrite sagedused, kui segumudeli parameetrid on hinnatud alamandmestikuga 2 ja terve andmestikuga.

Terve andmestik / Alamandmestik 2	1	2	3	4	5	6	7	Kokku
1	0	8	1 209	0	1	0	2	1 220
2	0	0	0	0	0	0	7 553	7 553
3	0	9 214	0	333	0	0	75	9 622
4	11 291	1	0	0	5	1	0	11 298
5	0	0	1	0	34	3 816	0	3 851
6	0	1	0	0	1 466	0	0	1 467
7	0	0	1	2 362	0	0	40	2 403
Kokku	11 291	9 224	1 211	2 695	1 506	3 817	7 670	37 414

Alamandmestiku 2 korral osutus aga parimaks seitsmekomponendiline mudel ning selle tõttu võime ka kohe näha märksa paremat tulemust. Erinevalt on klasterdatud 503 vaatlust, seega suuri erinevusi pole näha ning enamus patsientidest on klasterdatud alamandmestiku parameetrite korral sama moodi nagu terve andmestiku puhul.

Tabel 11. Klastrite sagedused, kui segumudeli parameetrid on hinnatud alamandmestikuga 3 ja terve andmestikuga.

Terve andmestik / Alamandmestik 3	1	2	3	4	5	6	7	Kokku
1	0	0	0	0	24	3 817	0	3 841
2	0	6	2	2 694	0	0	1	2 703
3	11 291	0	1	0	0	0	0	11 292
4	0	0	1 208	0	0	0	1	1 209
5	0	0	0	0	1 482	0	0	1 482
6	0	4	0	1	0	0	7 668	7 673
7	0	9 214	0	0	0	0	0	9 214
Kokku	11 291	9 224	1 211	2 695	1 506	3 817	7 670	37 414

Tabelis 11 on ära toodud klasterdused, mis on saadud, kui parameetrid on hinnatud terve andmestikuga ja alamandestikuga 3, mille suurus oli 10 000 vaatlust. Tabelist võime näha, et patsiendid on klasterdatud väga sarnaselt, nimelt on ainult 40 neist klasterdatud erinevalt. Seega oleme saanud tunduvalt parema tulemuse olukorras, kus parameetrite hinnangud on leitud kaks korda suurema alamandmestiku abil, mis kinnitab teooriat, et võib-olla on alamandmestik 5 000 vaatlusega siiski liiga väike, et saada tervele andmestikule piisavalt lähedast klasterdust.

Tabel 12. Klastrite sagedused, kui segumudeli parameetrid on hinnatud alamandmestikuga 4 ja terve andmestikuga.

Terve andmestik / Alamandmestik 4	1	2	3	4	5	6	7	Kokku
1	11 205	0	0	0	0	0	0	11 205
2	0	0	0	0	1 409	0	0	1 409
3	0	1	0	0	0	0	7 608	7 609
4	86	13	1 211	10	1	1	13	1 335
5	0	9 207	0	0	70	0	5	9 282
6	0	0	0	0	24	3 816	0	3 840
7	0	3	0	2 685	2	0	44	2 734
Kokku	11 291	9 224	1 211	2 695	1 506	3 817	7 670	37 414

Tabelis 12 võime aga näha klastrite sagedusi, kui mudeli parameetrid on hinnatud alamandmestikuga 4 ja terve andmestikuga. Saadud tulemus on kehvem, kui saime olukorras, kus parameetrite hinnangud olid leitud alamandmestikuga 3. Erinevalt on klasterdatud lausa 273 patsienti, mis on ligi 6 korda rohkem kui alamandmestiku 3 korral. Siiski võrreldes, kui palju on meil patsiente kokku (37 414), pole erinevalt klasterdatud patsientide arv sugugi suur ning samuti on tulemus parem, kui saime olukorras, kus parameetrite hinnangud olid leitud andmestiku abil, mille suurus oli 5 000 patsienti. Seega võime antud olukorras klasteranalüüsi läbiviimisel kaaluda parameetrite hinnangute leidmist 10 000 vaatlusega alamandmestiku abil.

Järgmisena toome välja ülevaatlikud sagedustabelid alamandmestiku 5 ja 6 jaoks, mille suurused on 20 000 vaatlust. Kasutades suuremat alamandmestikku parameetrite hinnangute leidmiseks peaksime loogiliselt võttes saavutama veelgi lähedasema klasterduse olukorrale, kus parameetrid on hinnatud terve andmestikuga. Kontrollime, kas see nii ka on.

Tabel 13. Klastrite sagedused, kui segumudeli parameetrid on hinnatud alamandmestikuga 5 ja terve andmestikuga.

Terve andmestik / Alamandmestik 5	1	2	3	4	5	6	7	Kokku
1	0	0	0	0	0	3 817	0	3 817
2	0	9 216	0	0	0	0	3	9 219
3	0	1	0	0	1 506	0	0	1 507
4	0	1	0	0	0	0	7 622	7 623
5	11 290	0	0	0	0	0	0	11 290
6	1	0	1 211	0	0	0	0	1 212
7	0	6	0	2 695	0	0	45	2 746
Kokku	11 291	9 224	1 211	2 695	1 506	3 817	7 670	37 414

Tabelist 13 võime näha, et erinevalt on klasterdatud 57 vaatlust, mis on isegi suurem arv, kui saime alamandmestiku 3 pealt hinnatud parameetritega mudeli korral. Uurime täpsemalt ka alamandmestikku 6.

Tabel 14. Klastrite sagedused, kui segumudeli parameetrid on hinnatud alamandmestikuga 6 ja terve andmestikuga.

Terve andmestik / Alamandmestik 6	1	2	3	4	5	6	7	Kokku
1	0	0	2	1 132	0	3 804	0	4 938
2	81	0	772	0	0	13	0	866
3	0	9 220	0	373	0	0	4	9 597
4	11 210	0	1	151	0	0	0	11 362
5	0	4	9	37	0	0	7 665	7 715
6	0	0	427	996	0	0	1	1 424
7	0	0	0	6	1 506	0	0	1 512
Kokku	11 291	9 224	1 211	2 695	1 506	3 817	7 670	37 414

Tabelist 14 võime näha, et tulemus pole sugugi nii hea võrreldes tulemustega, kus parameetrite hinnangud olid leitud alamandmestikega 2-5. Täpsemalt võime näha, et patsiendid, kes kuuluvad klastrisse 4 terve andmestiku korral, on alamandmestikuga leitud parameetritega mudeli põhjal paigutatud ära mitme erineva klastri vahel. Sarnast olukorda näeme klastri 3 korral, kuhu kuuluvad patsiendid on alamandmestiku parameetritega saadud klasterduse korral jaotatud klastrite 2 ja 6 vahel. Kuna tabelist 40 nägime, et mõningal juhul alamandmestikuga 6 hinnatud parameetrid erinesid oluliselt teiste alamandmestikega leitud hinnangutest, siis oli tulemus ka ootuspärane. Arvatavasti pole EM-algoritmi rakendamisel jõutud parima lokaalse maksimumini ning algoritmi uuesti rakendamine oleks võib-olla andnud paremad tulemused.

Kokkuvõttes on alamandmestike 2-5 pealt saadud klasteranalüüsi tulemused lähedased tulemustele, mille korral viisime läbi parameetrite hindamise terve andmestikuga. Selle põhjal võime öelda, et alamandmestiku kasutamine parameetrite hindamisel on õigustatud. Viimane teeks ka meie töö tunduvalt lihtsamaks, sest nagu tabelist 8 võisime näha, siis aega parameetrite hindamiseks kuluks näiteks alamandmestiku 3 korral lausa 13 tundi vähem.

4.2 Psüühika- ja käitumishäiretega patsientide klasterdamine

Psüühika- ja käitumishäirete klasterdamise jaoks võtame vaatluse alla need patsiendid, kellel vähemalt ühel aastal on põhidiagnooside hulgas olnud F-diagnoosi

kood. Selliseid patsiente on 402 524, kes on kõik kirjeldatud ära 13 erineva tunnusega, mille hulka kuuluvad sugu, vanusegrupp ning F-diagnooside 11 gruppi, mis on ära toodud lisas 2.

Klasteranalüüsi viime läbi vanusegruppide 10-19, 30-39, 50-59, 70-79 ja 80+ jaoks. Kokku on igas alamandmestikus 12 erinevat tunnust, mis kõik on binaarsed. Antud juhul 0 näitab, et patsiendile pole seda diagnoosi määratud ning 1 tähistab diagnoosi esinemist. Hindame kõige üldisema kitsendusteta segumudeli klastrite arvu $K = 5, ..., 11$ korral. Parameetrite hinnangud leitakse EM-algoritmi abil, mida rakendatakse 50 korda erinevate alglähenditega. Iteratsioonide arvuks on võetud 1 000 ja $\varepsilon = 0{,}0001$.

Psüühika- ja käitumishäired vanuseklassis 10-19

Nagu eelneva analüüsi põhjal juba nägime, siis parimaks segumudeliks asümptootilise ICL põhjal osutus seitsmekomponendiline mudel. Vastavad klastrite suurused on järgmised:

$$\hat{n}_1 = 11\ 291, \hat{n}_2 = 7\ 670, \hat{n}_3 = 1\ 506, \hat{n}_4 = 9\ 224,$$
$$\hat{n}_5 = 2\ 695, \hat{n}_6 = 3\ 817, \hat{n}_7 = 1\ 211$$

ning tunnuste väärtuste sagedused kõigis seitsmes klastris on välja toodud tabelis 15.

Tabel 15. Psüühika- ja käitumishäirete sagedused klastrites (vanuseklass 10-19)

Klaster / Tunnused	1	2	3	4	5	6	7	Kokku
Naine	6 461 (57%)	4 958 (65%)	879 (58%)	2 812 (30%)	948 (35%)	1 164 (30%)	401 (33%)	17 623
Mees	4 830 (43%)	2 712 (35%)	627 (42%)	6 412 (70%)	1 747 (65%)	2 653 (70%)	810 (67%)	19 791
F00-F09	65	81	7	188	96	49	188	674
F10-F19	169	299	20	415	64	47	788	1 802
F20-F29	78	188	10	72	49	43	164	604
F30-F39	0	7 670	0	517	134	0	9	8 330
F40-F49	11 291	3 236	423	2 294	477	584	22	18 327
F50-F59	0	493	1 506	124	20	0	0	2 143
F60-F69	81	210	14	64	23	18	143	553
F70-F79	0	71	12	6	2 695	0	4	2 788
F80-F89	0	442	59	2 192	1 192	3 817	6	7 708
F90-F98	0	952	70	9 224	658	0	1	10 905
F99	7	18	1	17	5	4	8	60
Klastri suurus	11 291	7 670	1 506	9 224	2 695	3 817	1 211	

Rohelised lahtrid tabelis näitavad, et selles klastris on vaadeldav tunnuse väärtus esinenud kõigil või väga suurel osal patsientidest. Seega võime kohe esimese klastri puhul, mis on ühtlasi ka kõige suurem klaster, näha, et kõigil sinna kuuluvatel patsientidel on diagnoositud neurootilisi, stressiga seotud või somatoformseid häireid (F40-F49). Teise klastri moodustavad meeleoluhäiretega (F30-F39) patsiendid, kellest suurem osa on naised (65%). Lisaks on paljudel (42%) esinenud ka neurootilisi, stressiga seotud või somatoformseid häireid. Kolmandas klastris on patsiendid, kellel esineb füsioloogiliste funktsioonide häirete või füüsiliste teguritega seotud käitumissündroome (F50-F59), mille alla kuuluvad näiteks söömishäired ja mitteorgaanilised unehäired või seksuaaldüsfunktsioonid. Suuruselt teise klastri moodustavad aga patsiendid, kellel kõigil esineb lapseeas alanud käitumis- ja tundeeluhäireid (F90-F98). Lisaks on nendest 25%-il täheldatud ka neurootilisi, stressiga seotud või somatoformseid häireid (F40-F49) ning 24%-il psühholoogilise arengu häireid (F80-F89). Sellesse klastrisse kuuluvatest patsientidest moodustavad mehed 70%. Viiendas klastris võime taaskord näha meeste ülekaalu. Kõigile klastrisse kuuluvatele patsientidele on diagnoositud vaimne alaareng (F70-F79), lisaks on päris suurel osal (44%) diagnoositud ka psühholoogilise arengu häireid. Kuuendas klastris näeme olukorda, kus kõigile on diagnoositud psühholoogilise arengu häire. Paneme tähele, et suurem osa on nendest mehed (70%). Viimane klaster on segaklaster, kuhu kuuluvatele patsientidele on määratud mitmeid erinevaid diagnoose. Silma paistab veidi suurem hulk (788 ehk 65%) patsiente, kellel on esinenud psühhoaktiivsete ainete tarvitamisest tingitud psüühika- ja käitumishäireid (F10-F19). Klastrisse kuulub rohkem mehi kui naisi.

Kokkuvõttes võime öelda, et kõige rohkem esineb vanuseklassis 10-19 patsiente, kellel on neurootilisi, stressiga seotud või somatoformseid häireid. Seda diagnoosi on esinenud 49%-il sellesse vanusegruppi kuuluvatest psüühika-ja käitumishäiretega noortest. Lisaks on paljudel (29%) esinenud ka lapseeas alanud käitumis- ja tundeeluhäireid, mida neljanda klastri põhjal esineb rohkem meestel. Välja võiks tuua veel meeleoluhäiretega patsiendid, keda on 22% selles vanusegrupis.

Psüühika- ja käitumishäired vanuseklassis 30-39

Vanuseklassi 30-39 kuulub 53 368 psüühika- ja käitumishäiretega patsienti. Asümp-

tootiline ICL (ICL = 338 881,1) põhjal on parimaks mudeliks viiekomponendiline segumudel, mis annab klasterdamisel järgmise suurusega klastrid:

$$\hat{n}_1 = 20\ 918, \hat{n}_2 = 19\ 215, \hat{n}_3 = 3\ 099, \hat{n}_4 = 6\ 075, \hat{n}_5 = 4\ 061.$$

Tunnuste väärtuste sagedused kõigis viies klastris on ära toodud tabelis 16.

Tabel 16. Psüühika- ja käitumishäirete sagedused klastrites (vanuseklass 30-39)

Klaster / Tunnused	1	2	3	4	5	Kokku
Naine	13 851 (66%)	13 223 (69%)	836 (27%)	919 (15%)	1 527 (38%)	30 356
Mees	7 067 (34%)	5 992 (31%)	2 263 (73%)	5 156 (85%)	2 534 (62%)	23 012
F00-F09	243	462	42	426	969	2 142
F10-F19	0	518	0	6 075	17	6 610
F20-F29	220	462	37	210	1 349	2 278
F30-F39	0	19 215	0	1 030	107	20 352
F40-F49	20 918	8 824	510	1 654	214	32 120
F50-F59	521	1 552	3 099	465	51	5 688
F60-F69	111	396	17	149	281	954
F70-F79	116	145	0	189	1 843	2 293
F80-F89	9	29	0	2	93	133
F90-F98	85	97	8	17	169	376
F99	9	18	0	7	25	59
Klastri suurus	20 918	19 215	3 099	6 075	4 061	

Sarnaselt eelmisele vanuseklassile moodustavad ka 30-39-aastaste hulgas kõige suurema klastri patsiendid, kellel kõigil on esinenud neurootilisi, stressiga seotud või somatoformseid häireid ning kellest suurem osa on naised (66%). Mahult järgmine on teine klaster, kus võime näha meeloluhäiretega patsiente, kellest taaskord enamuse moodustavad naised (69%). Nagu eelmise vanuseklassi puhul, on võimalik ka siin näha, et paljudele (46%-il) klastrisse kuuluvatele patsientidele on diagnoositud ka neurootilisi, stressiga seotud või somatoformseid häireid. Kolmandasse klastrisse kuuluvatel patsientidel on kõigil esinenud füsioloogiliste funktsioonide häirete või füüsiliste teguritega seotud käitumissündroome. Erinevalt eelmisest vanuseklassist moodustavad selle klastri suurema osa mehed (73%). Vaadeldavas vanuseklassis on tulnud juurde uus klaster (neljas klaster), kuhu kuuluvatel patsientidel kõigil on täheldatud psühhoaktiivsete ainete tarvitamisest tingitud psüühika- ja käitumishäireid (F10-F19), mille hulka kuuluvad häired, mis on näiteks tingitud alkoholi, rahustite, kokaiini või tubaka tarvitamisest. Lisaks on selles klastris 17%-il

esinenud ka meeleoluhäireid ning 27%-il neurootilisi, stressiga seotud või somatoformseid häireid. Väga suur osa klastrisse kuuluvatest patsientidest on mehed (85%). Viimane ehk viies klaster jaguneb suuremalt osalt kahe haigusgrupi vahel: 45% patsientidest esineb vaimne alaareng ning 33%-il skisofreenia, skisotüüpsed ja luululised häired.

Kokkuvõttes võib öelda, et vanuseklassi 30-39 puhul on neurootiliste, stressiga seotud või somatoformsete häiretega patsientide hulk isegi suurem võrreldes vanuseklassiga 10-19. Nimelt on siin see protsent 60 (enne oli 49). Samuti on kasvanud meeleoluhäirete all kannatavate patsientide hulk. Vanuseklassi 10-19 korral oli see 22% ning nüüd 38%, kellest enamus on naised. Vanuseklassis 30-39 on ära kadunud suur patsientide arv, kes põevad lapseeas alanud käitumis- ja tundeeluhäireid.

Psüühika- ja käitumishäired vanuseklassis 50-59

Vanuseklassi 50-59 suurus on 56 112 patsienti. Asümptootilise ICL (ICL = 374 651,2) põhjal osutus parimaks kuuekomponendiline mudel, millega saadud klastrite mahud on järgmised:

$$\hat{n}_1 = 18\ 357, \hat{n}_2 = 21\ 182, \hat{n}_3 = 4\ 482,$$
$$\hat{n}_4 = 6\ 752, \hat{n}_5 = 2\ 745, \hat{n}_6 = 2\ 594.$$

Vaadeldavate diagnooside sagedused kõigis kuues klastris on ära toodud tabelis 17.

Tabel 17. Psüühika- ja käitumishäirete sagedused klastrites (vanuseklass 50-59)

Klaster / Tunnused	1	2	3	4	5	6	Kokku
Naine	13 557(74%)	16 183 (76%)	1 803(40%)	856(13%)	1 176(43%)	1 357(52%)	34 932
Mees	4 800(26%)	4 999 (24%)	2 679(60%)	5 896(87%)	1 569(57%)	1 237(48%)	21 180
F00-F09	884	1 713	0	857	2 745	0	6 199
F10-F19	217	681	0	6 752	6	7	7 663
F20-F29	304	664	43	140	159	1 734	3 044
F30-F39	0	21 182	0	1 014	3	5	22 204
F40-F49	18 357	9 211	0	1 184	43	12	28 807
F50-F59	1 440	2 437	4 482	709	153	4	9 225
F60-F69	96	337	7	69	47	142	968
F70-F79	101	116	15	72	144	768	1 216
F80-F89	4	4	1	0	0	13	22
F90-F98	42	30	3	2	2	54	133
F99	5	15	1	4	8	18	51
Klastri suurus	18 357	21 182	4 482	6 752	2 745	2 594	

Vanuseklassi 50-59 puhul moodustavad kõige suurema klastri meeleoluhäiretega (F30-F39) patsiendid, kus naiste osakaal küündib 76%-ni. Sarnaselt eelnevatele vanusegruppidele on ka siin paljudele patsientidele (44%-le) diagnoositud lisaks neurootilisi, stressiga seotud või somatoformseid häireid. Suuruselt järgmine on esimene klaster, mille moodustavad kõik need patsiendid, kellel on esinenud neurootilisi, stressiga seotud või somatoformseid häireid. Naisi on nende hulgas taaskord rohkem kui mehi, moodustades 74% klastrisse kuuluvatest patsientidest. Nagu ka eelnevalt moodustavad kolmanda klasteri patsiendid, kellele kõigile on määratud füsioloogiliste funktsioonide häirete ja füüsiliste teguritega seotud käitumissündroomide kategooria alla kuuluv haigus. Taaskord võime näha meeste ülekaalu klastris (60%). Vanuseklassile 30-39 sarnaselt näeme psühhoaktiivsete ainete tarvitamisest tingitud psüühika- ja käitumishäiretega patsientide poolt moodustatud klastrit (neljas klaster), kus meeste osakaal (87%) on isegi natukene suurenenud. Lisaks on vaadeldavas vanuseklassis lisandunud klaster, kuhu kuuluvatele patsientidele on kõigile määratud orgaaniliste psüühikahäirete kategooria alla kuuluv diagnoos (dementsuse erinevad vormid ning ajukahjustusest tingitud psüühika- ja käitumishäired). Kuuendas klastris võime näha samuti sarnasust eelmise vanuseklassi viimase klastriga. Siin on aga suuremale osale (67%-le) diagnoositud skisofreeniat, skisotüüpseid või luululisi häireid ning 30%-le vaimne alaareng.

Kokkuvõttes võime näha, et nagu ka teiste vanuseklasside puhul täheldasime, on ka vanusegrupis 50-59 kõige rohkem patsiente, kellele on diagnoositud neurootili-

si, stressiga seotud või somatoformseid häireid. Täpsemalt on neid patsiente 51%, mis on veidi vähem kui eelmise vanuseklassi korral. Meeleoluhäirete all vaevlevate patsientide arv on selle vanuseklassi korral tõusnud 40%-ni.

Psüühika- ja käitumishäired vanuseklassis 70-79

Vanuseklassi 70-79 kuulub 37 864 psüühika- ja käitumishäiretega patsienti, kellest 10 652 on mehed ja 27 212 on naised. Parimaks mudeliks asümptootilise ICL (ICL = 245 771,4) põhjal osutus kuuekomponendiline segumudel. Selle mudeliga saadud klastrite mahud on järgmised:

$$\hat{n}_1 = 10\ 758, \hat{n}_2 = 12\ 926, \hat{n}_3 = 4\ 019,$$
$$\hat{n}_4 = 1\ 255, \hat{n}_5 = 7\ 692, \hat{n}_6 = 1\ 214.$$

Tunnuste väärtuste sagedusi kõigis kuues klastris on võimalik näha tabelist 18.

Tabel 18. Psüühika- ja käitumishäirete sagedused klastrites (vanuseklass 70-79)

Klaster / Tunnused	1	2	3	4	5	6	Kokku
Naine	8 513(79%)	10 349(80%)	2 417(60%)	290(23%)	4 833(63%)	810(67%)	27 212
Mees	2 245(21%)	2 577(20%)	1 602(40%)	965(77%)	2 859(37%)	404(33%)	10 652
F00-F09	1 637	3 021	0	339	7 692	9	12 698
F10-F19	6	154	0	1 255	0	0	1 415
F20-F29	186	373	41	40	377	1 003	2 020
F30-F39	0	12 926	0	140	0	0	13 066
F40-F49	10 758	4 355	0	169	24	0	15 306
F50-F59	1 358	1 893	4 019	123	513	0	7 906
F60-F69	24	62	8	5	24	36	159
F70-F79	1	13	6	0	33	138	191
F80-F89	0	1	1	0	2	6	10
F90-F98	13	23	1	0	8	43	88
F99	3	11	1	4	22	18	59
Klastrite suurused	10 758	12 926	4 019	1 255	7 692	1 214	

Selle vanuseklassi puhul moodustavad kõige suurema klastri taas patsiendid, kellel kõigil on esinenud meeleoluhäireid ning lisaks 34%-le on diagnoositud ka neuroootilisi, stressiga seotud või somatoformseid häireid. Viimasena nimetatud haigusgrupile on eraldatud suuruselt järgmine klaster (esimene klaster). Naiste osakaal on jätkuvalt mõlemas klastris suurem, vastavalt 80% ja 79%. Sarnaselt eelmistele vanusegruppidele on ka siin kolmas klaster patsientide päralt, kellel kõigil on täheldatud füsioloogiliste funktsioonide häirete ja füüsiliste teguritega seotud

käitumissündroome. Naiste-meeste osakaal on siin aga muutunud naiste kasuks. Neljandasse klastrisse kuuluvad taas psühhoaktiivsete ainete tarvitamisest tingitud psüühika- ja käitumishäiretega patsiendid. Meeste protsent klastris on eelmise vanusegrupiga võrreldes vähenenud, kuid on siiski kõrge (77%). Viiendas klastris võime näha orgaaniliste psüühikahäiretega patsiente. Eelmise vanuseklassi 50-59 korral nägime sarnase sisuga klastrit, mille suurus (2 745 patsienti) oli tunduvalt väiksem kui nüüd. Viimase klastri puhul võime samuti näha sarnast mustrit eelmise vanusegrupi viimase klastriga, kus 83% põeb skisofreeniat, skisotüüpseid või luululisi häireid. Vaimse alaarenguga patsientide protsent klastris on vähenenud (enne oli 30%, nüüd 11%).

Kokkuvõttes näeme, et patsiendid on klastritesse paigutatud analoogselt eelmistele vanuseklassidele. Kõige rohkem on taas neurootiliste, stressiga seotud ja somatoformsete häirete käes kannatavaid patsiente. Kuid siiski näeme, et nende patsientide arv vanusegrupis on taaskord vähenenud (nüüd moodustavad nad vaid 40%). Samuti on langenud meeleoluhäirete all vaevlevate patsientide arv, kes nüüd moodustavad 35% kõigist vanuses 70-79 psüühika- ja käitumishäiretega patsientidest. Välja võib tuua, et psühhoaktiivsete ainete tarvitamisest tingitud psüühika- ja käitumishäiretega patsientide arv on vanuse kasvades samuti vähenenud. Kui vanuses 50-59 moodustasid need patsiendid 14%, siis vanusegrupis 70-79 on see protsent kõigest 4. See on põhjustatud ilmselt meeste lühemast elueast. Paneme tähele, et märkimisväärselt on suurenenud orgaaniliste psüühikahäiretega patsientide arv. Nimelt moodustavad seda tüüpi diagnoosi saanud patsiendid vanuseklassis 70-79 34% kui samal ajal vanuseklassis 50-59 oli see protsent kõigest 11.

Psüühika- ja käitumishäired vanuseklassis 80+

Kuna kolme viimasesse vanusegruppi kuulus võrreldes teiste gruppidega vähem patsiente, siis oli mõistlik vanuseklassid 80-89, 90-99 ja 100+ ühendada. Vanusegruppi 80+ kuulub 28 701 isikut. Ainult kahel patsiendil esines psühholoogilise arengu häireid, seega seda tunnust klasteranalüüsi pole kaasatud. Parimaks mudeliks asümptootilise ICL (ICL = 167 027,7) põhjal osutus kuuekomponendiline

segumudel, millega saadud klastrite suurused on järgmised:

$$\hat{n}_1 = 6\ 278, \hat{n}_2 = 7\ 579, \hat{n}_3 = 3\ 102,$$
$$\hat{n}_4 = 160, \hat{n}_5 = 10\ 280, \hat{n}_6 = 1\ 302.$$

Tabelist 19 võime näha tunnuste diagnooside sagedusi kõigis kuues klastris.

Tabel 19. Psüühika- ja käitumishäirete sagedused klastrites (vanuseklass 80+)

Klaster / Tunnused	1	2	3	4	5	6	Kokku
Naine	5 195(83%)	6 310(83%)	2 322(75%)	76(48%)	7 967(78%)	1 093(84%)	22 963
Mees	1 083(17%)	1 269(17%)	780(25%)	84(52%)	2 313(22%)	209(16%)	5 738
F00-F09	1 471	2 255	0	1	10 280	613	14 620
F10-F19	31	44	7	95	68	19	264
F20-F29	0	0	0	0	0	1 302	1 302
F30-F39	0	7 579	0	0	0	255	7 834
F40-F49	6 278	1 948	0	0	0	227	8 453
F50-F59	893	1 177	3 102	0	735	126	6 033
F60-F69	10	18	3	16	19	8	74
F70-F79	5	6	3	21	5	7	47
F90-F98	6	13	2	18	4	0	43
F99	9	5	4	11	21	6	56
Klastrite suurused	6 278	7 579	3 102	160	10 280	1 302	

Meeste lühemast elueast tulenevalt on vanusegrupis 80+, välja arvatud neljandas klastris, kõigis naiste osakaal tunduvalt suurem. Kõige suurema klastri (klaster 5) moodustavad patsiendid, kellel kõigil on diagnoositud orgaanilisi psüühikahäireid, mille alla kuuluvad näiteks dementsuse erinevad liigid. Esimese klastri moodustavad taaskord patsiendid, kellel kõigil on täheldatud neurootilisi, stressiga seotud või somatoformseid häireid. Teises klastris näeme meeleoluhäirete all kannatavaid patsiente, kellest 25%-il on diagnoositud ka neurootilisi, stressiga seotud või somatoformseid häireid, mis on väiksem arv kui eelnevate vanuseklasside korral täheldasime. Kolmanda klastri moodustavad taas patsiendid, kellel kõigil on esinenud füsioloogiliste funktsioonide häirete ja füüsiliste teguritega seotud käitumissündroome. Näeme, et nendele patsientidele pole eriti teisi diagnoose määratud. Neljandas klastris, kus ainsana meeste osakaal on veidi suurem, võime näha, et enamusel (60%-l) on täheldatud psühhoaktiivsete ainete tarvitamisest tingitud psüühika- ja käitumishäireid. Viimases klastris on need patsiendid, kellel kõigil on esinenud skisofreeniat, skisotüüpseid või luululisi häireid. Kui varasemalt nägime, et seda tüüpi klastris oli märkimisväärsel osal patsientidest diagnoositud ka vaim-

ne alaareng, siis siin jääb see pigem tahaplaanile.

Kokkuvõttes näeme, et patsientide arv, kes põeb neurootilisi, stressiga seotud või somatoformseid häireid, on endiselt üsna suur. Selline diagnoos on 29%-il patsientidest (28 701). Selle vanusegrupi korral on aga orgaaniliste psüühikahäiretega patsientide suhteline sagedus kasvanud märkimisväärselt. Nimelt on see protsent nüüd 51, kui eelmise vanusegrupi korral oli see 34. Ühtlasi on see ka kõige sagedasem diagnoos vanusegrupis 80+.

Psüühika- ja käitumishäiretega patsientide klasterdamise kokkuvõte

Kõigis analüüsitud viies vanusegrupis võisime näha kolme väga sarnaselt tõlgendatavat klastrit. Kõige sagedasem diagnoos oli neurootilised, stressiga seotud ja somatoformsed häired (F40-F49). Selle haigusega patsientide suhteline sagedus oli vaadeldud vanuseklasside järjestuses vastavalt 49%, 60%, 51%, 40% ja 29%. Seega kõige rohkem selle diagnoosi saanud patsiente oli vanuseklassis 30-39. Nägime, et nende häirete all vaevlevad enamasti naissoost patsiendid. Teisena kerkis esile klaster, kus kõigil patsientidel esines meeleoluhäireid (F30-F39) ning lisaks oli märkimisväärsel protsendil diagnoositud ka neurootilisi, stressiga seotud või somatoformseid häireid. Meeleoluhäiretega patsiente oli vanusegrupiti vastavalt 22%, 38%, 40%, 35% ja 27%, seega diagnoosi suhteline sagedus kasvas kuni vanusegrupini 50-59 ning seejärel hakkas langema. Ka meeleoluhäirete all kannatavatest patsientidest olid enamus naised. Kolmas klaster, mille tõlgendus iga vanusegrupi korral sarnanes, oli füsioloogiliste funktsioonide häirete ja füüsiliste teguritega seotud käitumissündroomide (F50-F59) klaster. See diagnoos esines vanusegrupiti vastavalt 6%, 11%, 16%, 21% ja 21% patsientidest. Näeme, et selle haiguse osatähtsus vanuse kasvades veidi suurenes, kuid mitte palju. Lisaks tasub ära märkida, et vanuseklassis 10-19 oli üsna suur osa (29%) patsientidest lapseeas alanud käitumis- ja tundeeluhäiretega (F90-F98), samas aga teistes vanuseklassides oli nende patsientide arv praktiliselt olematu. Alates vanuseklassist 30-39 kerkis esile klaster, kuhu kuulusid kõik psühhoaktiivsete ainete tarvitamisest tingitud psüühika- ja käitumishäiretega patsiendid. Sellise diagnoosi olid vanuseklassides 30-39, 50-59 ja 70-79 saanud vastavalt 12%, 14% ja 4% patsientidest. Tasub märkida, et meeste osakaal selles klastris oli kõigi nende vanusegruppide korral suurem kui naiste osa-

kaal, moodustades klastri sisust vastavalt 85%, 87% ja 77%. Kõrges vanuses (80+) patsientide hulgas domineeris orgaaniliste psüühikahäiretega (F00-F09) patsientide klaster, mille hulka kuuluvad erinevad dementsuse vormid.

4.3 Vereringeelundite haigustega patsientide klasterdamine

Järgmisena uurime täpsemalt vereringeelundite haigustega patsiente, keda oli 623 883. Kirjeldavaid tunnuseid on kokku 12, mille hulka kuuluvad sugu, vanusegrupp ning I-diagnooside 10 gruppi, mis on ära toodud lisas 3.

Klasteranalüüsi viime läbi vanusegruppide 0-9, 20-29, 40-49, 60-69 ja 80-89 jaoks. Kokku on igas alamandmestikus 11 erinevat tunnust, mis kõik on binaarsed. Lisaks viime iga vanusegrupi korral klasteranalüüsi läbi ka naiste ja meeste jaoks eraldi. Hindame standardse kitsendusteta segumudeli klastrite arvu $K = 5, ..., 11$ korral. Parameetrite hinnangud leitakse EM-algoritmi abil, mida rakendatakse 50 korda erinevate alglähenditega. Iteratsioonide arvuks on võetud 1000 ja $\varepsilon = 0{,}0001$. Parim mudel valitakse asümptootilise ICL põhjal.

Vereringeelundite haiguste diagnooside sagedused klastrites võib leida tabelitest $20 - 24$. Esmalt paneme tähele, et vanusegrupi 0-9 korral on esimesed neli klastrit väga selgelt ära määratud kindla diagnoosigrupiga. Viimasesse klastrisse kuuluvad patsiendid, kellel esinevad ülejäänud diagnoosid. Teiste vanusegruppide korral nii selget klasterdust meil näha pole võimalik. Eriti paistab see silma vanusegruppide 60-69 ja 80-89 korral, kus diagnoosid on jaotunud ära mitme erineva klastri vahel. Siiski võime näha ühiseid mustreid. Näiteks moodustavad iga vanusegrupi korral esimese klastri patsiendid, kellele kõigile on diagnoositud muude südamehaiguste kategooriasse (I30-I52) kuuluv haigus. Selles grupis on näiteks ägeda perikardiidiga või müokardiidiga, südameseiskumisega ja arütmiaga patsiente. Naiste osakaal nendes klastrites on üldjuhul suurem. Vanuseklassis 0-9 on mahult kõige suurem klaster kolmas, kuhu kuuluvad patsiendid, kellel kõigil on diagnoositud veenide, lümfisoonte või -sõlmede mujal klassifitseerimata haigusi (I80-I89). Selle kategooria alla kuuluvad näiteks veenipõletik, veenikomud, tromboos ja söögitoruvaariksid. Näeme, et klastris on rohkem poisse kui tüdrukuid. Klastrit, kuhu kuuluvad patsiendid, kellele kõigile on diagnoositud I80-I89, võime näha ka teiste vanusegruppide korral, kus vastupidiselt vanusegrupile 0-9 on naiste osakaal tunduvalt

suurem. Kõrgemas vanuses (80-89) patsientide hulgas aga kaob see klaster ära. Vanuseklassist 20-29 alates lisandub juurde teine klaster, kus kõigile patsientidele on diagnoositud kõrgvererõhkhaigus (I10-I15). Paneme tähele, et vanusegruppides 40-49, 60-69 ja 80-89 sisaldab see klaster kõige rohkem isikuid. Kui nooremas vanuses on see haigus pigem esinenud meestel, siis alates 60. eluaastast kuulub sinna gruppi rohkem naisi, mis pigem võib olla tingitud meeste lühemast elueast. Kui vaadata vanusegrupi 80-89 klasteranalüüsi tulemusel saadud teist klastrit lähemalt, siis näeme, et lisaks kõrgvererõhkhaigustele on 29%-il täheldatud ka muude südamehaiguste alla kuuluvaid probleeme (I30-I52) ning 28%-il peaajuveresoonte haigusi (I60-I69). Viimasele haigusgrupile on eraldatud ka klaster vanusklassides 0-9 (klaster 2), 40-49 (klaster 5) ja 80-89 (klaster 5). Nooremates vanuseklassides (0-9 ja 20-29) võime näha eraldi klastrit patsientide jaoks, kellel on esinenud arterite, arterioolide või kapillaaride haigusi (I70-I79). Naiste osakaal on nendes klastrites suurem. Vanuseklassi 40-49 ja vanemate korral võime näha klastrit, kuhu kuuluvad ainult südame isheemiatõvega patsiendid (I20-I25). Paneme tähele, et sinna kuuluvatel patsientidel on lisaks päris palju teisi diagnoose, mis seda klastrit iseloomustavad. Meeste osakaal on siinkohal suurem kui naiste.

Tabel 20. Vereringeelundite haiguste sagedused klastrites (vanuseklass 0-9)

Tunnused \ Klaster	1	2	3	4	5	Kokku
Naine	668	72	1 322	390	110	2 562
Mees	783	91	1 882	280	108	3 144
I00-I02	1	0	0	0	14	15
I05-I09	5	0	0	1	23	29
I10-I15	5	0	5	0	79	89
I20-I25	0	0	0	0	21	21
I26-I28	4	0	0	0	17	21
I30-I52	1 451	0	42	4	0	1 497
I60-I69	1	163	3	3	0	170
I70-I79	0	0	10	670	0	680
I80-I89	0	0	3 204	9	0	3 213
I95-I99	2	0	1	0	65	68
Klastrite suurused	1 451	163	3 204	670	218	5 706

Tabel 21. Vereringeelundite haiguste sagedused klastrites (vanuseklass 20-29)

Tunnused \ Klaster	1	2	3	4	5	6	Kokku
Naine	5 111	1 692	9 487	996	1 010	406	18 702
Mees	3 720	7 903	6 346	370	265	366	18 970
I00-I02	28	2	8	4	0	35	77
I05-I09	150	12	20	4	4	110	300
I10-I15	365	9 595	0	37	14	0	10 011
I20-I25	122	117	47	0	5	258	549
I26-I28	35	15	34	3	2	90	178
I30-I52	8 831	882	0	144	153	0	10 010
I60-I69	47	53	35	7	10	292	444
I70-I79	5	53	0	1 366	24	0	1 448
I80-I89	1 277	1 029	15 833	279	205	6	18 629
I95-I99	83	43	0	2	1 275	0	1 403
Klastrite suurused	8 831	9 595	15 833	1 366	1 275	772	37 672

Tabel 22. Vereringeelundite haiguste sagedused klastrites (vanuseklass 40-49)

Tunnused \ Klaster	1	2	3	4	5	6	Kokku
Naine	8 837	12 306	13 688	1 182	935	1 605	38 553
Mees	5 411	19 370	6 011	578	1 194	4 242	36 806
I00-I02	20	7	5	12	0	0	44
I05-I09	249	39	23	85	10	32	438
I10-I15	5 016	31 676	3 191	9	1 105	3 422	44 419
I20-I25	349	0	7	0	0	5 847	6 203
I26-I28	117	58	132	130	19	63	519
I30-I52	14 248	0	0	0	294	1 404	15 946
I60-I69	4	0	0	0	2 129	230	2 363
I70-I79	351	382	599	1 038	106	232	2 708
I80-I89	3 172	2 463	19 699	0	385	1 073	26 792
I95-I99	262	99	221	524	15	45	1 166
Klastrite suurused	14 248	31 676	19 699	1 760	2 129	5 847	75 359

Tabel 23. Vereringeelundite haiguste sagedused klastrites (vanuseklass 60-69)

Tunnused \ Klaster	1	2	3	4	5	Kokku
Naine	13 959	30 354	12 181	13 722	1 719	71 935
Mees	10 162	18 642	4 868	15 419	4 566	53 657
I00-I02	10	3	4	6	3	26
I05-I09	536	58	29	328	34	985
I10-I15	18 168	48 996	11 481	23 717	2 500	104 862
I20-I25	0	0	0	29 141	0	29 141
I26-I28	542	216	251	565	328	1 902
I30-I52	24 121	0	0	12 722	259	37 102
I60-I69	3 036	4 579	1 222	4 222	3 133	16 192
I70-I79	1 768	832	1 053	3 721	4 136	11 510
I80-I89	5 604	0	17 049	6 579	286	29 518
I95-I99	164	138	101	241	214	858
Klastrite suurused	24 121	48 996	17 049	29 141	6 285	125 592

Tabel 24. Vereringeelundite haiguste sagedused klastrites (vanuseklass 80-89)

Tunnused \ Klaster	1	2	3	4	5	6	Kokku
Naine	1 902	29 702	2 356	874	1 923	15 737	52 494
Mees	1 037	6 231	2 836	647	954	6 668	18 373
I00-I02	0	0	0	0	1	10	11
I05-I09	27	0	1	6	1	338	373
I10-I15	0	35 933	1 134	94	0	21 908	59 069
I20-I25	0	5 212	5 192	0	0	15 604	26 008
I26-I28	104	338	40	145	22	1 473	2 122
I30-I52	2 939	10 290	1 837	90	782	21 195	37 133
I60-I69	17	10 190	914	31	2 877	8 623	22 652
I70-I79	328	3 020	764	1 046	347	6 920	12 425
I80-I89	239	3 105	217	647	88	7 334	11 630
I95-I99	10	85	24	33	6	323	481
Klastrite suurused	2 939	35 933	5 192	1 521	2 877	22 405	70 867

Järgmisena võtame kokku klasteranalüüsi tulemused eraldi meeste ja naiste jaoks, mis on välja toodud tabelites $25-34$. Juba tabelist 20 võisime näha, et vanusegrupi 0-9 korral jagunesid mehed ja naised klastritesse üsna võrdselt. Seega pole eriti suuri erinevusi klasteranalüüsi tulemustes tabelites 25 ja 26 näha. Kui nüüd vaadata vanuseklassi 20-29, siis näeme, et meespatsientide klasterdamisel on kadunud üks klaster ära. Nimelt pole meeste jaoks eraldatud klastrit, kus kõik patsiendid kuuluksid kategooria I95-I99 (*Vereringeelundite muud ja täpsustamata haigusseisundid*) alla. Lisaks võime näha, et teine klaster, mis sisult on sarnane mõlema soo puhul, on meeste korral ligi 4 korda suurem. Kõigile klastrisse kuuluvatele patsientidele on diagnoositud kõrgvererõhkhaigus. Vanuseklassi 40-49 puhul võime näha, et klasterdades eraldi mehi ja naisi, saame tulemuseks viis klastrit. Nende hulgas pole enam klastrit, kuhu kõigile kuuluvatele patsientidele oleks diagnoositud peaajuveresoonte haigus (I60-I69) nagu nägime tabelist 22. Näeme, et esimesed kolm klastrit on sarnased. Erinevad kaks viimast klastrit, kus neljanda klastri moodustavad meeste korral kõik need patsiendid, kellel on esinenud südame isheemiatõvesid. Paneme tähele, et selle diagnoosiga mehi pole määratud ühtegi teise klastrisse. Naissoost patsientide klasterdamisel moodustavad neljanda klastri haiged, kellel on kõigil diagnoositud arterite, arterioolide või kapillaaride haigusi. Viimasesse klastrisse on läinud nii-öelda ülejääk, kus domineerivad kaks kuni kolm diagnoosigruppi. Kokkuvõttes tasub tähele panna, et vanusegrupis 40-49 esineb südame isheemiatõvesid meestel ligi kaks korda rohkem kui naistel. Sama tähelepanek kehtib tegelikult juba ka vanuseklassi 20-29 kohta. Klasterdades eraldi naised ja mehed vanuseklassis 60-69, saame tulemuseks väga sarnased klastrid. Kui tabelis 23 nägime, et peaajuveresoonte haigustega patsiendid olid jagunenud kõigi klastrite vahel ära, siis nüüd klasterdades mehi ja naisi eraldi, võime näha, et osadele nendele patsientidele on eraldatud viies klaster. Lisaks võime näha, et veenide, lümfisoonte ja -sõlmede mujal klassifitseerimata haigustega naiste suhteline sagedus on 10% suurem kui meestel (osakaal vanusegrupis vastavalt meestel 18%, naistel 28%). Vanuseklassi 80-89 puhul erineb naiste ja meeste klasteranalüüs veidi rohkem. Võime kohe näha, et meestel on kaks klastrit rohkem kui naistel. Nende klastrite suurused on aga väga väikesed ning nendes domineerinud haigusega patsiente võime näha mõnes teises klastris isegi rohkem. Paneme tähele, et ka teiste haigusgruppide korral on diagnoosid jagunenud ära mitme erineva klastri vahel,

mis teeb vanuseklassi 80-89 klasteranalüüsi tulemused eraldi meeste ja naiste jaoks raskesti interpreteeritavaks.

Tabel 25. Vereringeelundite haiguste sagedused klastrites (vanuseklassi 0-9 mehed)

Klaster / Tunnused	1	2	3	4	5	Kokku
I00-I02	0	0	0	0	9	9
I05-I09	0	0	0	0	12	12
I10-I15	3	0	2	0	42	47
I20-I25	0	0	0	0	11	11
I26-I28	1	0	0	0	7	8
I30-I52	784	0	26	0	0	810
I60-I69	0	91	2	1	0	94
I70-I79	1	0	10	279	0	290
I80-I89	0	0	1 882	0	0	1 882
I95-I99	2	0	1	0	27	30
Klastrite suurused	784	91	1 882	279	108	3 144

Tabel 26. Vereringeelundite haiguste sagedused klastrites (vanuseklassi 0-9 naised)

Klaster / Tunnused	1	2	3	4	5	Kokku
I00-I02	1	0	0	0	5	6
I05-I09	5	0	0	1	11	17
I10-I15	2	0	3	0	37	42
I20-I25	0	0	0	0	10	10
I26-I28	3	0	0	0	10	13
I30-I52	688	0	16	3	0	707
I60-I69	1	72	1	2	0	76
I70-I79	0	0	9	381	0	390
I80-I89	0	0	1 331	0	0	1 331
I95-I99	0	0	0	0	38	38
Klastrite suurused	668	72	1 331	381	110	2 562

Tabel 27. Vereringeelundite haiguste sagedused klastrites (vanuseklassi 20-29 mehed)

Klaster / Tunnused	1	2	3	4	5	Kokku
I00-I02	10	2	2	0	14	28
I05-I09	46	12	6	1	38	103
I10-I15	889	6 977	0	44	0	7 910
I20-I25	93	77	26	1	172	369
I26-I28	11	5	12	2	27	57
I30-I52	4 636	0	0	0	1	4 637
I60-I69	21	34	14	4	128	201
I70-I79	46	0	0	379	0	425
I80-I89	511	690	6 376	61	4	7 642
I95-I99	71	36	30	2	236	375
Klastrite suurused	4 636	6 977	6 376	379	602	18 970

Tabel 28. Vereringeelundite haiguste sagedused klastrites (vanuseklassi 20-29 naised)

Klaster / Tunnused	1	2	3	4	5	6	Kokku
I00-I02	20	0	6	2	0	21	49
I05-I09	104	0	14	3	3	73	197
I10-I15	367	1 692	0	28	14	0	2 101
I20-I25	54	13	21	0	0	92	180
I26-I28	29	6	22	0	1	64	122
I30-I52	5 373	0	0	0	0	0	5 373
I60-I69	32	14	21	3	7	166	243
I70-I79	118	0	0	905	0	0	1 023
I80-I89	903	217	9 487	198	138	4	10 947
I95-I99	170	0	0	19	839	0	1 028
Klastrite suurused	5 373	1 692	9 487	905	839	406	18 702

Tabel 29. Vereringeelundite haiguste sagedused klastrites (vanuseklassi 40-49 mehed)

Klaster / Tunnused	1	2	3	4	5	Kokku
I00-I02	1	3	1	1	3	9
I05-I09	70	10	8	24	23	135
I10-I15	2 474	17 391	2 547	0	42	22 454
I20-I25	0	0	0	4 244	0	4 244
I26-I28	57	30	61	40	69	257
I30-I52	5 563	0	0	1 072	0	6 635
I60-I69	154	488	66	167	486	1 361
I70-I79	124	200	185	166	433	1 108
I80-I89	958	57	8 536	664	19	10 234
I95-I99	56	45	39	23	139	302
Klastrite suurused	5 563	17 391	8 536	4 244	1 072	36 806

Tabel 30. Vereringeelundite haiguste sagedused klastrites (vanuseklassi 40-49 naised)

Klaster / Tunnused	1	2	3	4	5	Kokku
I00-I02	18	6	2	0	9	35
I05-I09	185	35	11	2	70	303
I10-I15	3 048	16 076	0	342	0	19 466
I20-I25	681	571	149	45	513	1 959
I26-I28	78	64	40	8	72	262
I30-I52	9 311	0	0	0	0	9 311
I60-I69	164	408	80	34	316	1 002
I70-I79	274	0	0	1 326	0	1 600
I80-I89	2 370	3 228	10 505	442	13	16 558
I95-I99	212	92	138	17	405	864
Klastrite suurused	9 311	16 076	10 505	1 326	1 335	38 553

Tabel 31. Vereringeelundite haiguste sagedused klastrites (vanuseklassi 60-69 mehed)

Klaster / Tunnused	1	2	3	4	5	6	Kokku
I00-I02	1	0	1	0	0	2	4
I05-I09	129	19	5	122	4	5	284
I10-I15	7 354	17 796	2 754	12 252	3 233	12	43 401
I20-I25	0	0	0	15 419	0	0	15 419
I26-I28	263	96	128	317	51	130	985
I30-I52	10 191	0	0	6 828	213	0	17 232
I60-I69	1 383	0	0	2 360	4 535	0	8 278
I70-I79	974	1 357	464	2 541	938	1 011	7 285
I80-I89	1 820	0	4 547	2 602	481	0	9 450
I95-I99	51	59	17	97	21	50	295
Klastrite suurused	10 191	17 796	4 547	15 419	4 535	1 169	53 657

Tabel 32. Vereringeelundite haiguste sagedused klastrites (vanuseklassi 60-69 naised)

Klaster / Tunnused	1	2	3	4	5	6	Kokku
I00-I02	9	3	3	6	0	1	22
I05-I09	407	40	23	206	9	16	701
I10-I15	10 826	27 991	7 755	11 465	3 424	0	61 461
I20-I25	0	0	0	13 722	0	0	13 722
I26-I28	299	127	116	248	55	72	917
I30-I52	13 976	0	0	5 894	0	0	19 870
I60-I69	1 692	0	0	1 862	4 360	0	7 914
I70-I79	839	844	605	1 180	344	413	4 224
I80-I89	3 796	0	11 296	3 977	999	0	20 068
I95-I99	118	94	81	144	27	99	564
Klastrite suurused	13 976	27 991	11 296	13 722	4 360	590	71 935

Tabel 33. Vereringeelundite haiguste sagedused klastrites (vanuseklassi 80-89 mehed)

Klaster / Tunnused	1	2	3	4	5	6	7	Kokku
I00-I02	1	0	1	0	0	0	0	2
I05-I09	13	4	0	1	0	2	48	68
I10-I15	4 171	4 061	1 288	0	0	21	4 460	14 000
I20-I25	2 319	0	2 112	0	0	5	3 320	7 756
I26-I28	108	69	6	7	0	74	328	592
I30-I52	6 009	0	0	0	0	96	4 106	10 211
I60-I69	1 051	1 096	564	601	91	2	2 117	5 522
I70-I79	130	564	129	0	458	8	2 730	4 019
I80-I89	0	311	44	26	45	231	1 974	2 631
I95-I99	11	10	5	3	3	13	110	155
Klastrite suurused	6 009	4 061	2 112	601	458	268	4 864	18 373

Tabel 34. Vereringeelundite haiguste sagedused klastrites (vanuseklassi 80-89 naised)

Klaster / Tunnused	1	2	3	4	5	Kokku
I00-I02	0	0	0	1	8	9
I05-I09	205	0	5	0	95	305
I10-I15	12 836	21 075	0	0	11 157	45 068
I20-I25	5 122	4 889	1 069	303	6 869	18 252
I26-I28	314	283	77	23	833	1 530
I30-I52	16 436	0	0	0	10 486	26 922
I60-I69	3 603	5 799	0	1 735	5 993	17 130
I70-I79	380	2 307	574	202	4 943	8 406
I80-I89	323	2 603	384	97	5 592	8 999
I95-I99	5	81	17	2	221	326
Klastrite suurused	16 436	21 075	1 858	1 735	11 390	52 494

Tabelitest 23 ja 24 võisime näha, et enamus diagnoose olid jagunenud mitme erineva klastri vahel, mis tegi vanusegruppide 60-69 ja 80-89 klastrite tõlgendamise keerulisemaks. Sellega seoses tekkis huvi just nende vanusegruppide korral vaadata, kui paljudele patsientidele on määratud vereringeelunditega seotud haigusi, mis kuuluvad mitmesse erinevasse I-diagnoosi gruppi. Vastavat ülevaadet on võimalik näha tabelitest 35 ja 36, kus on välja toodud ka protsendid, mis näitavad, kui suure osa moodustab vastavasse gruppi kuuluv patsientide arv sellesse vanusegruppi kuuluvatest vereringeelundite haigustega patsientide koguarvust.

Tabel 35. Vereringeelundite haigustega patsientidele vanuses 60-69 määratud erinevate I-diagnooside arv

I-diagnoosi gruppide arv	1	2	3	4	5	6	7
Patsientide arv	57 882(46%)	39 592(32%)	19 523(16%)	6 784(5%)	1 564(1%)	224(0.2%)	23(0.002%)

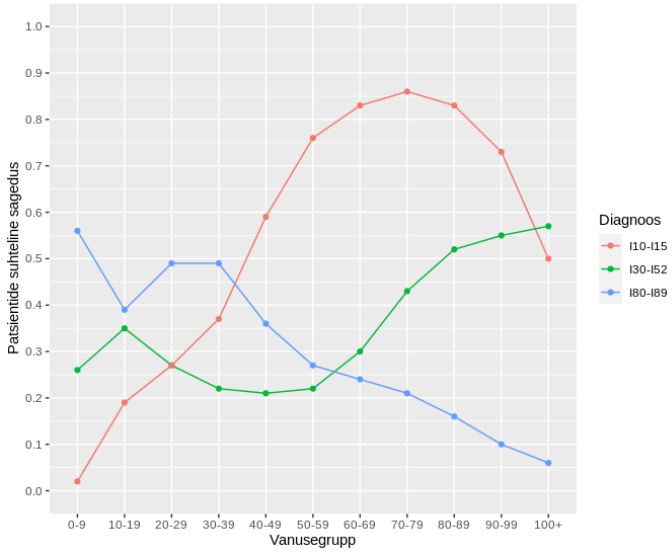
Tabel 36. Vereringeelundite haigustega patsientidele vanuses 80-89 määratud erinevate I-diagnooside arv

I-diagnoosi gruppide arv	1	2	3	4	5	6	7
Patsientide arv	17 505(25%)	22 486(32%)	18 077(26%)	9 382(13%)	2 859(4%)	533(0.8%)	25(0.04%)

Vanuseklassi 60-69 korral võime näha, et kõige rohkem on patsiente, kellele on diagnoositud vaid ühte kindlasse diagnoosigruppi kuuluv vereringeelundite haigus (46%). Patsientide arv, kellele on määratud kahte või enamasse diagnoosigruppi kuuluv diagnoos, kahaneb. Vanuseklassi 80-89 puhul võime aga näha, et kõige rohkem on patsiente, kellel on määratud diagnooside hulgas kahte erinevasse vereringeelundite haiguste gruppi kuuluvaid haigusi (32%). Üsna suurele osale (26%) patsientides on määratud lausa kolme erinevasse diagnoosigruppi kuuluv haigus. Patsiente, kellele määratud diagnoosid kuuluvad ainult ühte vereringeelundite haiguste gruppi, on 25%. Seega näeme, et selle vanusegrupi korral moodustavad üsna suure protsendi patsiendid, kellele on määratud kahte või lausa kolme erinevasse diagnoosigruppi kuuluvaid haigusi, mis antud juhul võibki olla põhjuseks miks diagnoosid nii mitme erineva klastri vahel olid ära jaotunud.

Parema ülevaate saamiseks toome välja graafikud, mis iseloomustavad kolme diagnoosigruppi kuuluvate patsientide suhtelist sagedust vanuseklassiti. Diagnoosigruppideks on valitud kolm kõige sagedamini esinenud gruppi: I10-I15 (*Kõrgvererõhkhaigused*), I30-I52 (*Muud südamehaigused*), I80-I89 (*Veenide, lümfisoonte ja*

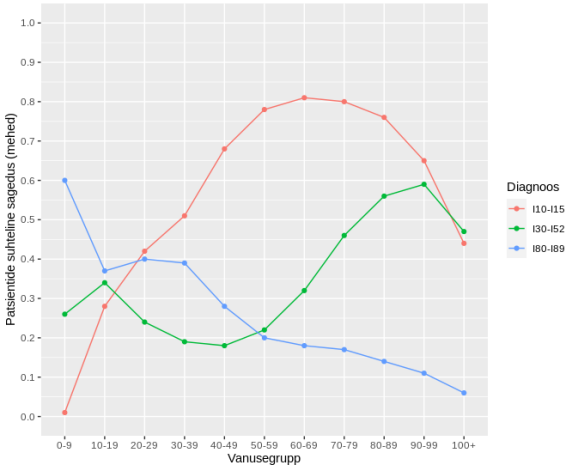
-sõlmede mujal klassifitseerimata haigused). Jooniselt 1 võime näha, et patsientide, kellele on diagnoositud veenide, lümfisoonte ja -sõlmede mujal klassifitseerimata haigusi, suhteline sagedus on olnud kõige kõrgem kuni vanuseni 39 ning seejärel hakkab nende patsientide osakaal langema. Kõrgvererõhkhaigustega patsientide suhteline sagedus kasvab sujuvalt kuni vanusegrupini 70-79 ning hakkab seejärel langema. Paneme tähele, et vaadeldavatest diagnoosidest esineb kõrgvererõhkhaigusi isikutel vanuses 40-99 kõige rohkem. Kui vaadata patsiente diagnoosiga, mis kuulub muude südamehaiguste alla, siis võime näha, et nende suhteline sagedus alates vanusest 60 on suurem kui diagnoosiga I80-I89 patsientide suhteline sagedus.



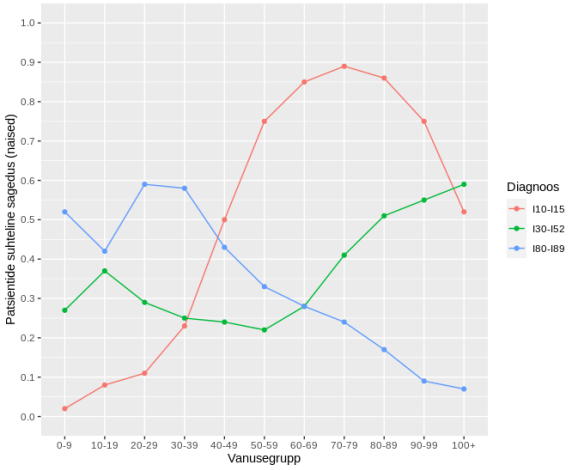
Joonis 1. Diagnoosiga I10-I15, I30-I52 ja I80-I89 patsientide suhteline sagedus igas vanusegrupis.

Kui võrrelda meessoost ja naissoost patsiente eraldi, siis võime üldpildis näha sarnase kujuga graafikuid (Joonis 2 ja Joonis 3). Siiski paneme tähele, et kõrgvererõhkhaiguste diagnoosi saanud patsientide suhteline sagedus vanuses 10-39 on meeste korral suurem kui naistel. Eriti hästi tuleb välja see erinevus vanuseklassides 10-19 ja 20-29, kus suhtelised sagedused meeste korral on 0,28 ning 0,42 ja naiste puhul kõigest 0,08 ning 0,11. Kui vaadata diagnoose I80-I89, siis võime näha, et nende diagnooside suhteline sagedus on enamus vanusegruppides naistel kõrgem kui meestel. Silma torkab suurem meeste suhteline sagedus vanuses 0-9, mis on 0,6 ning seejärel langeb vanuseklassis 10-19 järsult 0,37 peale. Diagnoosigrupi I30-I52

graafikud käituvad meeste ja naiste korral suhteliselt sarnaselt.



Joonis 2. Diagnoosiga I10-I15, I30-I52 ja I80-I89 meespatsientide suhteline sagedus igas vanusegrupis.



Joonis 3. Diagnoosiga I10-I15, I30-I52 ja I80-I89 naispatsientide suhteline sagedus igas vanusegrupis.

Kokkuvõte

Käesoleva magistritöö eesmärk oli anda ülevaade mudelipõhise klasteranalüüsi rakendamisest kvalitatiivsete tunnustega andmetele. Töö teoreetilises osas kirjeldati ära mudelipõhise klasteranalüüsi metoodika. Selgitati põhjalikult mudeli parameetrite hindamist EM-algoritmi abil ning vaadati ka kitsenduste rakendamist mudelile. Lisaks toodi kaks näidet parema ülevaate saamiseks klasteranalüüsi rakendamisest andmestiku `birds` ja simuleeritud andmete peal. Praktilises osas keskenduti teooria rakendamisele Eesti Haigekassa andmetele.

Töös viidi läbi segumudeli parameetrite hindamise protsess, võttes vaatluse alla erineva suurusega andmestikud. Eesmärk oli saada ülevaade klasterdamise stabiilsusest ning väiksema mahuga alamandmestiku kasutamisvõimalusest mudeli parameetrite hindamisel. Vaatluse all olev andmestik oli 37 414 kirjega. Parameetrite hindamise tulemusel nägime, et alamandmestik suurusega 5 000 võib osutuda liiga väikseks, et anda terve andmestikuga stabiilselt samaväärseid tulemusi. Leides parameetrite hinnangud alamandmestikega, mille suurused olid vastavalt 10 000 ja 20 000 vaatlust, olid saadud parimad tulemused väga lähedased klasterdusele, mis olid saadud terve andmestiku abil hinnatud mudeli parameetritega. Seega jõuti järeldusele, et alamandmestiku kasutamine parameetrite hinnangute leidmisel oli selle andmestiku korral õigustatud, ning selle tulemusel saaks klasterdamiseks vajalike segumudelite parameetrite hindamisel R-is aega kokku hoida lausa 13 tundi. Lisaks selgus, et täpse ICL-kriteeriumi põhjal parimaks tunnistatud mudeliga saadud klasterdused osutusid raskemini interpreteeritavaks ning olid märgitud alati kõige kehvemaks asümptootilise ICL puhul. Seda, miks asümptootiline ICL-kriteerium ja täpne ICL-kriteerium nii erinevalt käituvad, tuleks täpsemalt uurida veel edaspidises analüüsis. Selle tõttu oli käesolevas töös parima mudeli valimisel arvesse võetud asümptootilist ICL väärtust.

Antud töö praktilises osas uuriti patsiente, kellel esinesid psüühika- ja käitumishäirete kategooria alla kuuluvaid haigusi või vereringeelundite haigusi. Töös kasutatud haigekassa andmestiku patsientidest 27%-il esines psüühika- ja käitumishäireid. Samuti jäi see protsent 30 lähedale ka igas vanusegrupis. Vereringeelundite haigusi esines aga vanemates vanusegruppides rohkem. Näiteks kõige rohkem oli

protsentuaalselt selle diagnoosiga isikuid vanuses 80-89, moodustades 93% patsientide koguarvust selles vanusegrupis. Kokkuvõttes oli mingi vereringeelundite haigus diagnoositud 42%-il patsientidest haigekassa andmestikus.

Viies klasteranalüüsi läbi psüühika- ja käitumishäiretega patsientide peal nägime, et iga vanusegrupi korral olid esindatud kolm väga sarnaselt tõlgendatavat klastrit, kus vastavalt kõigile isikutele määratud diagnoosid kuulusid kas neurootiliste, stressiga seotud ja somatoformsete häirete, meeleoluhäirete või füsioloogiliste funktsioonide häirete hulka. Alates vanusegrupist 30-39 võisime näha klastrit, kuhu kuuluvatel patsientidel oli kõigil täheldatud psühhoaktiivsete ainete tarvitamisest tingitud psüühika- ja käitumishäireid, suurem osa nendest olid mehed. Kõrgemas eas (80+) isikute hulgas moodustasid kõige suurema klastri patsiendid, kellel kõigil oli diagnoositud orgaaniline psüühikahäire.

Vereringeelundite haigustega patsientide klasterdamisel võis samuti näha sarnaseid klastreid, mis olid esindatud iga vanusegrupi korral. Näiteks kuulusid ühte sellisesse klastrisse patsiendid, kellele kõigile oli määratud muude südamehaiguste kategooria alla kuuluv haigus. Alates vanusest 40 kerkis rohkem esile teistest suurema mahuga klaster, kuhu kuuluvatel patsientidel oli diagnoositud kõrgvererõhkhaigus. Vanusegruppides 20-29 ja 40-49 võis nendes klastrites rohkem näha mehi kui naisi. Täpsemalt selgus, et viies klasteranalüüsi läbi naiste ja meeste jaoks eraldi, esines vanuseklassis 20-29 kõrgvererõhkhaigusi meestel lausa ligi 4 korda rohkem kui naistel. Lisaks võis täheldada, et vanuseklassides 20-29 ja 40-49 esines meestel südame isheemiatõvesid umbes 2 korda rohkem kui naistel. Selle diagnoosiga patsientidele oli määratud ka päris palju teisi diagnoose. Üldiselt võis näha, et vanuseklassi 0-9 korral olid klastrid väga selgelt ära määratud kindla diagnoosigrupiga. Teiste vanusegruppide korral nii selget klasterdust näha polnud võimalik. Enamus juhtudel jagunesid diagnoosid mitme erineva klastri vahel ära ning eriti tuli see välja vanusgruppide 60-69 ja 80-89 korral.

Kasutatud kirjandus

Anderlucci, L. ja Hennig, C. (2014). Clustering of Categorical Data: A Comparison of a Model-Based and a Distance-Based Approach. Communications in Statistics-Theory and Methods 43, 704-721.

Biernacki, C., Celeux, G. ja Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(7), 719-725. doi: 10.1109/34.865189.

Biernacki, C., Celeux, G. ja Govaert, G. (2010). Exact and Monte Carlo Calculations of Integrated Likelihoods for the Latent Class Model. Journal of Statistical Planning and Inference, 140 (11), 2991-3002. doi: 10.1016/j.jspi.2010.03.042.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer, NewYork

Hennig, C., Meila, M., Murtagh, F. ja Rocci, R. (2016). Handbook of Cluster Analysis. Taylor & Francis, Boca Raton.

Izenman, A. J. (2008). Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer, NewYork.

Langrognet, F., Lebret, R., Poli, C., Iovleff, S., Auder, B., Bhatia, P. jt (2019). Package 'Rmixmod'. Classifcation with mixture modelling. `https://cran.r-project.org/web/packages/Rmixmod/Rmixmod.pdf` (05.04.2020)

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G. ja Govaert, G. (2015). Rmixmod: The R package of the Model-Based Unsupervised, Supervised, and Semi-Supervised Classification Mixmod Library. Journal of Statistical Software, 67 (6), 1-29. doi: 10.18637/jss.v067.i06.

Mirski, S. (2019). Mudelipõhine klasteranalüüs, Tartu Ülikool, Magistritöö. `https://dspace.ut.ee/handle/10062/64860`

Med24 (2016). RHK-10. [www]
https://www.med24.ee/andmebaasid/rhk10 (28.06.2020)

Surma põhjuste register (2011-2019), Tervise Arengu Instituut, Tervisestatistika ja terviseuuringute andmebaas, tabel SD21, 22.05.2019 seisuga. [www]
http://pxweb.tai.ee/PXWeb2015/

World Health Organization (2016). ICD-10 Classifications. [www]
http://www.who.int/classifications/icd/icdonlineversions/en/
(28.06.2020)

Lisad

Lisa 1. RHK-10 koodide tähendused

Tabel 37. RHK-10 (Rahvusvaheline Haiguste Klassifikatsioon) (Med24 (2016))

Indeks	Kood	Nimetus
1	A00-B99	Teatavad nakkus- ja parasiithaigused
2	C00-D48	Kasvajad
3	D50-D89	Vere- ja vereloomeelundite haigused ning teatavad immuunmehhanismidega seotud haigusseisundid
4	E00-E90	Sisesekretsiooni-, toitumis- ja ainevahetushaigused
5	F00-F99	Psüühika- ja käitumishäired
6	G00-G99	Närvisüsteemihaigused
7	H00-H59	Silma- ja silmamanuste haigused
8	H60-H95	Kõrva- ja nibujätkehaigused
9	I00-I99	Vereringeelundite haigused
10	J00-J99	Hingamiselundite haigused
11	K00-K93	Seedeelundite haigused
12	L00-L99	Naha- ja nahaaluskoe haigused
13	M00-M99	Lihasluukonna- ja sidekoehaigused
14	N00-N99	Kuse-suguelundite haigused
15	O00-O99	Rasedus, sünnitus ja sünnitusjärgne periood
16	P00-P96	Perinataal- e sünniperioodis tekkivad teatavad seisundid
17	Q00-Q99	Kaasasündinud väärarendid, deformatsioonid ja kromosoomianomaaliad
18	R00-R99	Mujal klassifitseerimata sümptomid, tunnused ja kliinilised ning laboratoorsete leidude hälbed
19	S00-T98	Vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed
20	V01-Y98	Haigestumise ja surma välispõhjused
21	Z00-Z99	Terviseseisundit mõjustavad tegurid ja kontaktid terviseteenistusega
22	U00-U99	Koodid spetsiifiliste eesmärkide jaoks

Lisa 2. Psüühika- ja käitumishäiretele vastavad RHK-10 koodid

Tabel 38. RHK-10, Psüühika- ja käitumishäired (Med24 (2016))

Kood	Nimetus	Näited
F00-F09	Orgaanilised - k.a sümptomaatilised - psüühikahäired	Dementsus Alzheimeri tõvest, isiksus- ja käitumishäired ajuhaigusest
F10-F19	Psühhoaktiivsete ainete tarvitamisest tingitud psüühika- ja käitumishäired	Näiteks alkoholi, tubaka, rahustite või kofeiini tarvitamisest tingitud psüühika- ja käitumishäired
F20-F29	Skisofreenia, skisotüüpsed ja luululised häired	
F30-F39	Meeleoluhäired	Maania, depressioon, bipolaarne meeleoluhäire
F40-F49	Neurootilised, stressiga seotud ja somatoformsed häired	Foobiad, obsessiiv-kompulsiivne häire, stressreaktsioonid ja kohanemishäired
F50-F59	Füsioloogiliste funktsioonide häirete ja füüsiliste e somaatiliste teguritega seotud käitumissündroomid	Söömishäired, mitteorgaanilised unehäired, mitteorgaanilised seksuaaldüsfunktsioonid
F60-F69	Täiskasvanu isiksus- ja käitumishäired	Harjumus- ja impulsihäired, sooidentsuse häired, seksuaalsuunitluse häired
F70-F79	Vaimne alaareng	
F80-F89	Psühholoogilise arengu häired e psüühilise arengu spetsiifilised häired	Kõne ja keele spetsiifilised arenguhäired, õpivilumuste spetsiifilised häired, motoorika spetsiifiline arenguhäire
F90-F98	Tavaliselt lapseeas alanud käitumis- ja tundeeluhäired	Hüperkineetilised häired, käitumishäired, lapse või nooruki suhtlemishäired, tikid
F99	Täpsustamata psüühikahäire	

Lisa 4. Mudeli parameetrite hinnangud psüühika- ja käitumishäiretega 10-19 aastaste näitel

Tabel 40. Mudeli parameetrite hinnangud (psüühika- ja käitumishäiretega patsiendid vanuses 10-19)

Klaster	Andmestik	Parameeter	Sugu		F00-F09		F10-F19		F20-F29		F30-F39		F40-F49		F50-F59		F60-F69		F70-F79		F80-F89		F90-F98		F99	
			naine	mees	ei	jah	ei	jah	ei	jah	ei	jah	ei	jah	ei	jah	ei	jah	ei	jah	ei	jah	ei	jah	ei	jah
1	Alamandmestik 2	$\hat{\alpha}_5$	0.2842	0.7158	0.9877	0.0123	0.9907	0.0093	0.9928	0.0072	0.9952	0.0048	0.852	0.148	0.9876	0.0124	0.9971	0.0029	0.9802	0.0198	0.0014	0.9986	0.984	0.016	0.9969	0.0031
	Alamandmestik 3	$\hat{\alpha}_1$	0.3182	0.6818	0.9887	0.0113	0.9872	0.0128	0.9872	0.0128	0.9926	0.0074	0.8504	0.1496	0.9923	0.0077	0.9955	0.0045	0.9923	0.0077	0.0007	0.9993	0.982	0.0180	0.9985	0.0015
	Alamandmestik 4	$\hat{\alpha}_6$	0.2882	0.7118	0.9921	0.0079	0.9892	0.0108	0.991	0.009	0.9882	0.0118	0.869	0.131	0.9947	0.0053	0.9931	0.0069	0.9789	0.0211	0.0006	0.9994	0.9759	0.0241	0.9995	0.0005
	Alamandmestik 5	$\hat{\alpha}_1$	0.2995	0.7005	0.9875	0.0125	0.9878	0.0122	0.9885	0.0115	0.9869	0.0131	0.8586	0.1414	0.9964	0.0036	0.9962	0.0038	0.9951	0.0049	0.0004	0.9997	0.9883	0.0117	0.9982	0.0018
	Alamandmestik 6	$\hat{\alpha}_1$	0.3136	0.6864	0.9833	0.0167	0.9916	0.0084	0.9876	0.0124	0.9766	0.0234	0.8393	0.1607	0.9958	0.0042	0.9946	0.0054	0.8188	0.1812	0.0004	0.9996	0.8539	0.1461	0.9988	0.0012
	Terve andmestik	$\hat{\alpha}_6$	0.301	0.699	0.9871	0.0129	0.9896	0.0104	0.9887	0.0113	0.9932	0.0068	0.8537	0.1463	0.9952	0.0048	0.9953	0.0047	0.9973	0.0027	0.0002	0.9998	0.9954	0.0046	0.9986	0.0014
2	Alamandmestik 2	$\hat{\alpha}_7$	0.3675	0.6325	0.9577	0.0423	0.9774	0.0226	0.9639	0.0361	0.9304	0.0696	0.8397	0.1603	0.9844	0.0156	0.9879	0.0121	0.0055	0.9945	0.6067	0.3933	0.887	0.113	0.9954	0.0046
	Alamandmestik 3	$\hat{\alpha}_2$	0.3629	0.6371	0.9612	0.0388	0.9725	0.0275	0.9831	0.0169	0.9476	0.0524	0.8218	0.1782	0.9881	0.0119	0.992	0.0080	0.0027	0.9973	0.5938	0.4062	0.7653	0.2347	0.9993	0.0007
	Alamandmestik 4	$\hat{\alpha}_7$	0.3682	0.6318	0.9742	0.0258	0.9808	0.0192	0.9724	0.0276	0.9327	0.0673	0.8003	0.1997	0.9893	0.0107	0.9845	0.0155	0.003	0.997	0.5764	0.4236	0.7837	0.2163	0.9992	0.0008
	Alamandmestik 5	$\hat{\alpha}_7$	0.3597	0.6403	0.9626	0.0374	0.9786	0.0214	0.9799	0.0201	0.94	0.06	0.8132	0.1868	0.9914	0.0086	0.9898	0.0102	0.0015	0.9985	0.5494	0.4506	0.7761	0.2239	0.9969	0.0031
	Alamandmestik 6	$\hat{\alpha}_6$	0.3286	0.6714	0.8541	0.1459	0.9817	0.0183	0.8893	0.1107	0.9904	0.0096	0.9277	0.0723	0.9944	0.0056	0.9065	0.0935	0.313	0.687	0.9657	0.0343	0.9787	0.0213	0.9912	0.0088
	Terve andmestik	$\hat{\alpha}_4$	0.3617	0.6383	0.9634	0.0366	0.98	0.02	0.9817	0.0183	0.9496	0.0504	0.8302	0.1698	0.9917	0.0083	0.9917	0.0083	0.0007	0.9993	0.5547	0.4453	0.7976	0.2024	0.9978	0.0022
3	Alamandmestik 2	$\hat{\alpha}_4$	0.6069	0.3931	0.9926	0.0074	0.9866	0.0134	0.9896	0.0104	0.9958	0.0042	0.0004	0.9996	0.9876	0.0124	0.9923	0.0077	0.9937	0.0063	0.9891	0.0109	0.9657	0.0343	0.999	0.001
	Alamandmestik 3	$\hat{\alpha}_3$	0.5785	0.4215	0.9932	0.0068	0.9864	0.0136	0.9963	0.0037	0.9976	0.0024	0.0004	0.9998	0.9936	0.0064	0.9937	0.0063	0.9978	0.0022	0.9973	0.0027	0.9684	0.0316	0.9995	0.0005
	Alamandmestik 4	$\hat{\alpha}_1$	0.5752	0.4248	0.9938	0.0062	0.9909	0.0091	0.9945	0.0055	0.998	0.002	0.0004	0.9998	0.9915	0.0085	0.9946	0.0054	0.9985	0.0015	0.9942	0.0058	0.9796	0.0204	0.9992	0.0008
	Alamandmestik 5	$\hat{\alpha}_5$	0.5721	0.4279	0.9951	0.0049	0.9888	0.0112	0.9926	0.0074	0.999	0.001	0.0001	0.9999	0.9985	0.0015	0.9908	0.0092	0.9989	0.0011	0.9932	0.0068	0.9851	0.0149	0.9993	0.0007
	Alamandmestik 6	$\hat{\alpha}_4$	0.5783	0.4217	0.9947	0.0053	0.993	0.007	0.9935	0.0065	0.999	0.001	0.0001	0.9999	0.9985	0.0015	0.994	0.006	0.9919	0.0081	0.997	0.003	0.987	0.013	0.9993	0.0007
	Terve andmestik	$\hat{\alpha}_1$	0.5757	0.4243	0.9945	0.0055	0.9886	0.0114	0.9932	0.0068	0.9995	0.0005	0	1	0.9987	0.0013	0.9931	0.0069	0.999	0.001	0.9971	0.0029	0.9794	0.0206	0.9994	0.0006
4	Alamandmestik 2	$\hat{\alpha}_1$	0.3282	0.6718	0.8524	0.1476	0.3391	0.6609	0.8905	0.1095	0.9662	0.0338	0.9095	0.0905	0.9917	0.0083	0.8992	0.1008	0.9922	0.0078	0.9895	0.0105	0.9686	0.0314	0.9918	0.0082
	Alamandmestik 3	$\hat{\alpha}_4$	0.3408	0.6592	0.848	0.152	0.3319	0.6681	0.8916	0.1084	0.9689	0.0311	0.9605	0.0395	0.9948	0.0052	0.9149	0.0851	0.9958	0.0042	0.9711	0.0289	0.985	0.015	0.9956	0.0044
	Alamandmestik 4	$\hat{\alpha}_4$	0.3146	0.6854	0.8537	0.1463	0.3966	0.6034	0.8574	0.1426	0.9386	0.0614	0.8584	0.1416	0.9875	0.0125	0.8974	0.1026	0.9671	0.0329	0.9679	0.0321	0.9459	0.0541	0.9848	0.0152
	Alamandmestik 5	$\hat{\alpha}_6$	0.3327	0.6673	0.8491	0.1509	0.3857	0.6143	0.8607	0.1393	0.9742	0.0258	0.9147	0.0853	0.9981	0.0019	0.8887	0.1113	0.9824	0.0176	0.9818	0.0182	0.9902	0.0098	0.9907	0.0093
	Alamandmestik 6	$\hat{\alpha}_2$	0.3685	0.6315	0.9829	0.0171	0.0014	0.9986	0.9802	0.0198	0.9932	0.0068	0.8824	0.1176	0.9962	0.0038	0.9967	0.0033	0.9909	0.0091	0.9678	0.0322	0.9889	0.0111	0.999	0.001
	Terve andmestik	$\hat{\alpha}_3$	0.3302	0.6698	0.8551	0.1449	0.3642	0.6358	0.8748	0.1252	0.9766	0.0234	0.9136	0.0864	0.9972	0.0028	0.8868	0.1132	0.9901	0.0099	0.9835	0.0165	0.9809	0.0191	0.9938	0.0062
5	Alamandmestik 2	$\hat{\alpha}_6$	0.5558	0.4442	0.9877	0.0123	0.9886	0.0114	0.9973	0.0027	0.9787	0.0213	0.8085	0.1915	0.0043	0.9957	0.9865	0.0135	0.9835	0.0165	0.9789	0.0211	0.9499	0.0501	0.9973	0.0027
	Alamandmestik 3	$\hat{\alpha}_5$	0.5574	0.4426	0.996	0.004	0.9808	0.0192	0.9906	0.0094	0.9924	0.0076	0.7395	0.2605	0.0023	0.9977	0.9958	0.0042	0.9897	0.0103	0.9777	0.0223	0.9532	0.0468	0.9986	0.0014
	Alamandmestik 4	$\hat{\alpha}_2$	0.569	0.431	0.9912	0.0088	0.9819	0.0181	0.9943	0.0057	0.9863	0.0137	0.7597	0.2403	0.0023	0.9977	0.9938	0.0062	0.9904	0.0096	0.9832	0.0168	0.9574	0.0426	0.9985	0.0015
	Alamandmestik 5	$\hat{\alpha}_3$	0.5807	0.4193	0.9935	0.0065	0.9865	0.0135	0.9913	0.0087	0.9948	0.0052	0.7241	0.2759	0.0011	0.9989	0.9902	0.0098	0.9894	0.0106	0.9707	0.0293	0.9367	0.0633	0.9975	0.0025
	Alamandmestik 6	$\hat{\alpha}_7$	0.5476	0.4524	0.9961	0.0039	0.9879	0.0121	0.9909	0.0091	0.9943	0.0057	0.7361	0.2639	0.0011	0.9989	0.99	0.01	0.9868	0.0132	0.9746	0.0254	0.9493	0.0507	0.9992	0.0008
	Terve andmestik	$\hat{\alpha}_5$	0.5704	0.4296	0.995	0.005	0.9854	0.0146	0.9932	0.0068	0.9974	0.0026	0.7248	0.2752	0.0006	0.9994	0.9905	0.0095	0.9922	0.0078	0.9696	0.0304	0.9536	0.0464	0.9991	0.0009
6	Alamandmestik 2	$\hat{\alpha}_2$	0.6395	0.3605	0.989	0.011	0.959	0.041	0.9714	0.0286	0.0015	0.9985	0.5808	0.4192	0.9358	0.0642	0.9787	0.0213	0.9956	0.0044	0.9437	0.0563	0.9202	0.0798	0.9995	0.0005
	Alamandmestik 3	$\hat{\alpha}_6$	0.6291	0.3709	0.9913	0.0087	0.9646	0.0354	0.9762	0.0238	0.0008	0.9992	0.5863	0.4137	0.9399	0.0601	0.9699	0.0301	0.9902	0.0098	0.9449	0.0551	0.886	0.114	0.9957	0.0043
	Alamandmestik 4	$\hat{\alpha}_3$	0.644	0.356	0.9875	0.0125	0.9708	0.0292	0.9822	0.0178	0.0007	0.9993	0.5925	0.4075	0.9344	0.0656	0.9809	0.0191	0.996	0.004	0.9408	0.0592	0.8816	0.1184	0.998	0.002
	Alamandmestik 5	$\hat{\alpha}_4$	0.6367	0.3633	0.9887	0.0113	0.9635	0.0365	0.9757	0.0243	0.0004	0.9996	0.5886	0.4114	0.9351	0.0649	0.9743	0.0257	0.995	0.005	0.9401	0.0599	0.8826	0.1174	0.9974	0.0026
	Alamandmestik 6	$\hat{\alpha}_5$	0.6377	0.3623	0.9866	0.0134	0.9551	0.0449	0.9718	0.0282	0.0004	0.9996	0.5877	0.4123	0.9405	0.0595	0.9745	0.0255	0.9867	0.0133	0.9481	0.0519	0.8883	0.1117	0.9973	0.0027
	Terve andmestik	$\hat{\alpha}_7$	0.6372	0.3628	0.989	0.011	0.9632	0.0368	0.9751	0.0249	0.0002	0.9998	0.5906	0.4094	0.9363	0.0637	0.9734	0.0266	0.9916	0.0084	0.9413	0.0587	0.8826	0.1174	0.9976	0.0024
7	Alamandmestik 2	$\hat{\alpha}_3$	0.2961	0.7039	0.9797	0.0203	0.9534	0.0466	0.9929	0.0071	0.9276	0.0724	0.7905	0.2095	0.9883	0.0117	0.9937	0.0063	0.9696	0.0304	0.759	0.241	0.0009	0.9991	0.9966	0.0034
	Alamandmestik 3	$\hat{\alpha}_7$	0.3026	0.6974	0.9789	0.0211	0.9546	0.0454	0.9925	0.0075	0.9338	0.0662	0.7817	0.2183	0.9879	0.0121	0.9935	0.0065	0.9921	0.0079	0.7686	0.2314	0.0005	0.9995	0.9986	0.0014
	Alamandmestik 4	$\hat{\alpha}_5$	0.3152	0.6848	0.9787	0.0213	0.9573	0.0427	0.9928	0.0072	0.9318	0.0682	0.7577	0.2423	0.9838	0.0162	0.9934	0.0066	0.9818	0.0182	0.7603	0.2397	0.0005	0.9995	0.9987	0.0013
	Alamandmestik 5	$\hat{\alpha}_2$	0.3131	0.6869	0.9808	0.0192	0.9534	0.0466	0.991	0.009	0.9367	0.0633	0.7526	0.2474	0.9885	0.0115	0.9915	0.0085	0.99	0.01	0.7601	0.2399	0.0002	0.9998	0.9982	0.0018
	Alamandmestik 6	$\hat{\alpha}_3$	0.3191	0.6809	0.9824	0.0176	0.9523	0.0477	0.9933	0.0067	0.9274	0.0726	0.7384	0.2616	0.9855	0.0145	0.9912	0.0088	0.955	0.045	0.8081	0.1919	0.0002	0.9998	0.997	0.003
	Terve andmestik	$\hat{\alpha}_2$	0.3095	0.6905	0.9799	0.0201	0.9548	0.0452	0.9923	0.0077	0.9345	0.0655	0.7573	0.2427	0.9851	0.0149	0.9925	0.0075	0.9836	0.0164	0.7599	0.2401	0.0001	0.9999	0.9981	0.0019

Lisa 5. Töös kasutatud R-koodi näited

Funktsioonide `klasterda`, `tulemus`, `ICLtapne`, `alfa` ja `andmed` moodustamisel on kasutatud magistriöös Mirski (2019) väljatoodud R-koodi näiteid.

```
library(Rmixmod)

#FUNKTSIOONID:

#1) Funktsioon, mis teostab kvalitatiivsete andmete mudelipõhise
#   klasteranalüüsi
klasterda <- function(andmed, mudelid, klastriteArv, katseteArv,
                      iterArv, epsilon, s) {
  mixmod=mixmodCluster(data=data.frame(andmed),nbCluster=klastriteArv,
                       dataType="qualitative",
                       models=mixmodMultinomialModel(listModels=mudelid),
                       strategy=mixmodStrategy(nbTry=katseteArv,
                       nbIterationInAlgo=iterArv, epsilonInInit=epsilon,
                       epsilonInAlgo=epsilon),
                       seed=s, criterion="ICL")
}

#2) Funktsioon, mis teeb Rmixmod tulemuste andmestiku
#   (tulemusF jaoks kutsutakse välja ICLtapneFDiagn)
tulemus <- function(valim,mudelid,m){
  tulemused=data.frame()
  for(i in 1:length(mudelid)){
    tulemused[i,1]=mudelid[[i]]@model
    tulemused[i,2]=mudelid[[i]]@nbCluster
    tulemused[i,3]=mudelid[[i]]@likelihood
    tulemused[i,4]=mudelid[[i]]@criterionValue #ICL
    tulemused[i,5]=ICLtapne(valim,mudelid[[i]],m=m) #ICL_M
  }
  colnames(tulemused)=c("Mudel","K","Toepara","ICL","ICLtapne")
  return(tulemused)
```

```
}

#3) Funktsioon, mis arvutab täpse ICL kriteeriumi väärtuse simuleetitud
#   andmetega näite jaoks
ICLtapne <- function(andmed,mudel,m){
  nk=table(mudel@partition) #klastrite mahud hat(n)_k
  K=length(nk) #klastrite arv K

  # Tunnuste (d=4) väärtuste sagedused klastrites
  uk_1=data.frame(xtabs(~mudel@partition+andmed$X1))$Freq
  uk_2=data.frame(xtabs(~mudel@partition+andmed$X2))$Freq
  uk_3=data.frame(xtabs(~mudel@partition+andmed$X3))$Freq
  uk_4=data.frame(xtabs(~mudel@partition+andmed$X4))$Freq

  #Täpse ICL kriteeriumi väärtuse arvutamine
  s1=sum(lgamma(uk_1+1/2))+sum(lgamma(uk_2+1/2))+sum(lgamma(uk_3+1/2))
    +sum(lgamma(uk_4+1/2))
  s2=0
  for (k in 1:K){
    s2=s2+sum(lgamma(nk[k]+m/2))
  }
  ICL=sum(lgamma(nk+1/2))+s1-s2+lgamma(K/2)-K*lgamma(1/2)
     -lgamma(sum(nk)+K/2)+K*sum(lgamma(m/2)-m*lgamma(1/2))
  return(round(ICL,3))
}

#4) Funktsioon, mis arvutab täpse ICL kriteeriumi väärtuse birds
#   andmestikuga näite jaoks
ICLtapneBirds <- function(andmed,mudel,m){
  nk=table(mudel@partition) #klastrite mahud hat(n)_k
  K=length(nk) #klastrite arv K

  # Tunnuste (d=5) väärtuste sagedused klastrites
```

```
  uk_1=data.frame(xtabs(~mudel@partition+andmed$gender))$Freq
  uk_2=data.frame(xtabs(~mudel@partition+andmed$eyebrow))$Freq
  uk_3=data.frame(xtabs(~mudel@partition+andmed$collar))$Freq
  uk_4=data.frame(xtabs(~mudel@partition+andmed$`sub-caudal`))$Freq
  uk_5=data.frame(xtabs(~mudel@partition+andmed$border))$Freq

  #Täpse ICL kriteeriumi väärtuse arvutamine
  s1=sum(lgamma(uk_1+1/2))+sum(lgamma(uk_2+1/2))+sum(lgamma(uk_3+1/2))
    +sum(lgamma(uk_4+1/2))+sum(lgamma(uk_5+1/2))
  s2=0
  for (k in 1:K){
    s2=s2+sum(lgamma(nk[k]+m/2))
  }
  ICL=sum(lgamma(nk+1/2))+s1-s2+lgamma(K/2)-K*lgamma(1/2)
     -lgamma(sum(nk)+K/2)+K*sum(lgamma(m/2)-m*lgamma(1/2))
  return(round(ICL,3))
}

#5) Funktsioon, mis arvutab täpse ICL kriteeriumi väärtuse
#   F-diagnoosidega patsientide jaoks
ICLtapneFDiagn <- function(andmed,mudel,m){
  nk=table(mudel@partition) #klastrite mahud hat(n)_k
  K=length(nk) #klastrite arv K

  # Tunnuste (d=12) väärtuste sagedused klastrites
  uk_1=data.frame(xtabs(~mudel@partition+andmed$gender))$Freq
  uk_2=data.frame(xtabs(~mudel@partition+andmed$F0))$Freq
  uk_3=data.frame(xtabs(~mudel@partition+andmed$F1))$Freq
  uk_4=data.frame(xtabs(~mudel@partition+andmed$F2))$Freq
  uk_5=data.frame(xtabs(~mudel@partition+andmed$F3))$Freq
  uk_6=data.frame(xtabs(~mudel@partition+andmed$F4))$Freq
  uk_7=data.frame(xtabs(~mudel@partition+andmed$F5))$Freq
  uk_8=data.frame(xtabs(~mudel@partition+andmed$F6))$Freq
```

```
  uk_9=data.frame(xtabs(~mudel@partition+andmed$F7))$Freq
  uk_10=data.frame(xtabs(~mudel@partition+andmed$F8))$Freq
  uk_11=data.frame(xtabs(~mudel@partition+andmed$F9))$Freq
  uk_12=data.frame(xtabs(~mudel@partition+andmed$F10))$Freq

  #Täpse ICL kriteeriumi väärtuse arvutamine
  s1=sum(lgamma(uk_1+1/2))+sum(lgamma(uk_2+1/2))+sum(lgamma(uk_3+1/2))
    +sum(lgamma(uk_4+1/2))+sum(lgamma(uk_5+1/2))+sum(lgamma(uk_6+1/2))
    +sum(lgamma(uk_7+1/2))+sum(lgamma(uk_8+1/2))+sum(lgamma(uk_9+1/2))
    +sum(lgamma(uk_10+1/2))+sum(lgamma(uk_11+1/2))+sum(lgamma(uk_12+1/2))
  s2=0
  for (k in 1:K){
    s2=s2+sum(lgamma(nk[k]+m/2))
  }
  ICL=sum(lgamma(nk+1/2))+s1-s2+lgamma(K/2)-K*lgamma(1/2)
     -lgamma(sum(nk)+K/2)+K*sum(lgamma(m/2)-m*lgamma(1/2))
  return(round(ICL,3))
}

#6) Funktsioon, mis arvutab multinomiaalsete jaotuste tõenäosused
alfa <- function(k,j,h,delta){
  if(h==((k-1)%%m[j])+1){
    tn=1/m[j] + (1-delta)*((m[j]-1)/m[j])
  } else {
    tn=(1-1/m[j] - (1-delta)*((m[j]-1)/m[j]))/(m[j]-1)
  }
  return(tn)
}

#7) Funktsioon, mis genereerib valimi kahe komponendiga
#   multinomiaalsete jaotuste segust
andmed <- function(n,p,d,m,delta) {
  #Multinomiaalsete jaotuste tõenäosuste arvutamine
```

```
alfa1=alfa2=rep(NA,times=sum(m)) #K=2
loendur=0
for (j in 1:d) { #d=4
  for (h in 1:m[j]) { #m1=2 m2=m3=3 ja m4=4
    loendur=loendur+1
    alfa1[loendur]=alfa(k=1,j=j,h=h,delta=delta)
    alfa2[loendur]=alfa(k=2,j=j,h=h,delta=delta)
  }
}

#Ühtlasest jaotusest genereeritud arvude abil leiame klastrite mahud
set.seed(12)
uhtlane=runif(n)
n1=sum(uhtlane < p[1])
n2=n-n1

#Esimene klaster
x1a=data.frame(t(rmultinom(n=n1,size=1,prob=alfa1[1:2])))
x1b=data.frame(t(rmultinom(n=n1,size=1,prob=alfa1[3:5])))
x1c=data.frame(t(rmultinom(n=n1,size=1,prob=alfa1[6:8])))
x1d=data.frame(t(rmultinom(n=n1,size=1,prob=alfa1[9:12])))
x1a2=data.frame("K1"=apply(x1a,1,function(x) which(x==max(x))))
x1b2=data.frame("K1"=apply(x1b,1,function(x) which(x==max(x))))
x1c2=data.frame("K1"=apply(x1c,1,function(x) which(x==max(x))))
x1d2=data.frame("K1"=apply(x1d,1,function(x) which(x==max(x))))
x1=cbind(x1a2,x1b2,x1c2,x1d2,1)
colnames(x1)=c("X1","X2","X3","X4","K")

#Teine klaster
x2a=data.frame(t(rmultinom(n=n2,size=1,prob=alfa2[1:2])))
x2b=data.frame(t(rmultinom(n=n2,size=1,prob=alfa2[3:5])))
x2c=data.frame(t(rmultinom(n=n2,size=1,prob=alfa2[6:8])))
x2d=data.frame(t(rmultinom(n=n2,size=1,prob=alfa2[9:12])))
```

```
  x2a2=data.frame("K2"=apply(x2a,1,function(x) which(x==max(x))))
  x2b2=data.frame("K2"=apply(x2b,1,function(x) which(x==max(x))))
  x2c2=data.frame("K2"=apply(x2c,1,function(x) which(x==max(x))))
  x2d2=data.frame("K2"=apply(x2d,1,function(x) which(x==max(x))))
  x2=cbind(x2a2,x2b2,x2c2,x2d2,2)
  colnames(x2)=c("X1","X2","X3","X4","K")

  #Kõik genereeritud vaatlused koos
  valim=rbind(x1,x2)
  for (i in 1:ncol(valim)) valim[,i]=as.factor(valim[,i])
  return(c(valim, alfa1, alfa2))
}

#-------------------------------------------------------------------------
#NÄIDE SIMULEERITUD ANDMETEGA
m=c(2,3,3,4)
andmedKoik=andmed(n=3000,p=c(0.2,0.8),d=4,m=m,delta=0.6)
valim<-andmedKoik[1:5]
alfa1<-andmedKoik[6:17]; alfa2<-andmedKoik[18:29]

mudelid<-c("Binary_pk_Ekjh", "Binary_pk_Ekj", "Binary_pk_Ek",
           "Binary_pk_Ej", "Binary_pk_E")
mudel <- klasterda(andmed = valim[1:4], mudelid = mudelid,
                   klastriteArv = 2:4, katseteArv = 20,
                   iterArv = 1000, epsilon = 0.001, s=12)
(tulemused=tulemus(valim=valim,mudelid=mudel@results, m=m))
(parim=mudel["results"][[which.min(tulemused$ICL)]])
table(valim$K,parim@partition)

#-------------------------------------------------------------------------

#NÄIDE ANDMESTIKU BIRDS PEAL
```

```
data(birds)
#Teisendused andmestikuga
birds$collar <- as.factor(as.character(birds$collar))
birds$`sub-caudal` <- as.factor(as.character(birds$`sub-caudal`))
birds$`sub-caudal`[birds$`sub-caudal` == "black & WHITE"] <- "black & white"
birds$border <- as.factor(as.character(birds$border))
birds$border[birds$border == "many"] <- "few"

mudelid<-c("Binary_pk_Ekjh", "Binary_pk_Ekj", "Binary_pk_Ek",
           "Binary_pk_Ej", "Binary_pk_E")
mudelBirds <- klasterda(andmed = birds, mudelid = mudelid,
                        klastriteArv = 2:4, katseteArv = 20,
                        iterArv = 1000, epsilon = 0.001, s=12)
mudelBirds@results
m<-c(2, 4, 2, 3, 2)
(tulemusedBirds=tulemus(valim=birds,mudelid=mudelBirds@results,m))
(parimBirds=mudelBirds["results"][[which.min(tulemusedBirds$ICL)]])
aggregate(birds,by=list(parimBirds@partition),FUN=table)
table(parimBirds@partition) #näitab klastrite suuruseid

#-------------------------------------------------------------------------

#SEGUMUDELI PARAMEETRITE HINDAMINE KASUTADES ALAMANDMESTIKKE
#Näitena on toodud kood ühe alamandmestiku jaoks

m=c(2,2,2,2,2,2,2,2,2,2,2,2)
mudelid<-c("Binary_pk_Ekjh")

#Alamandmestik 1; n= 5 000; Katse 1
start_time <- Sys.time()
mudelF1019_alam11 <- klasterda(andmed=patient_with_F_diagn_10_19_alam1,
                               mudelid = mudelid, klastriteArv = 5:11,
                               katseteArv = 50, iterArv = 1000,
```

```
                                epsilon = 0.0001, s=8)
(tulemusedF1019_alam01=tulemusF(patient_with_F_diagn_10_19_alam1,
                                 mudelF1019_alam11@results,m))
end_time <- Sys.time()
time0<-end_time - start_time #Algoritmi jooksutamiseks kulunud aeg

#Klastrite sagedused (Tabel 9)
pred0 <- mixmodPredict(data=patient_with_F_diagn_10_19,
                       classificationRule=parimF1019_alam01)
aggregate(patient_with_F_diagn_10_19,by=list(pred0@partition),FUN=table)
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kristiina Uusna,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Mudelipõhise klasteranalüüsi rakendamine Eesti Haigekassa andmetele", mille juhendaja on Kristi Kuljus, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kristiina Uusna
18.08.2020